Received: 30 December 2021 | Revised: 14 June 2022 | Accepted: 19 June 2022

DOI: 10.1002/sim.9513

RESEARCH ARTICLE

Statistics in Medicine WILEY

Controlled variable selection in Weibull mixture cure models for high-dimensional data

Han Fu[1] | **Deedra Nicolet**[2,3] | **Krzysztof Mrózek**[2] | **Richard M. Stone**[4] | **Ann-Kathrin Eisfeld**[2] | **John C. Byrd**[5] | **Kellie J. Archer**[1]

[1]Division of Biostatistics, College of Public Health, The Ohio State University, Columbus, Ohio, USA

[2]Clara D. Bloomfield Center for Leukemia Outcomes Research, The Ohio State University Comprehensive Cancer Center, Columbus, Ohio, USA

[3]Alliance Statistics and Data Management Center, The Ohio State University Comprehensive Cancer Center, Columbus, Ohio, USA

[4]Dana-Farber/Partners Cancer, Harvard University, Boston, Massachusetts, USA

[5]Department of Internal Medicine, University of Cincinnati, Cincinnati, Ohio, USA

Correspondence
Kellie J. Archer, Division of Biostatistics, College of Public Health, The Ohio State University, 240 Cunz Hall, 1841 Neil Avenue, Columbus, OH 43210, USA.
Email: archer.43@osu.edu

Funding information
American Society of Hematology; Coleman Leukemia Research Foundation; Leukemia and Lymphoma Society; Leukemia Research Foundation; National Cancer Institute, Grant/Award Numbers: P30CA016058, R35CA197734, U10CA180821, U10CA180882, U24CA196171, UG1CA233180, UG1CA233331, UG1CA233338; The D Warren Brown Foundation; U.S. National Library of Medicine, Grant/Award Number: R01LM013879

Medical breakthroughs in recent years have led to cures for many diseases. The mixture cure model (MCM) is a type of survival model that is often used when a cured fraction exists. Many have sought to identify genomic features associated with a time-to-event outcome which requires variable selection strategies for high-dimensional spaces. Unfortunately, currently few variable selection methods exist for MCMs especially when there are more predictors than samples. This study develops high-dimensional penalized Weibull MCMs, which allow for identification of prognostic factors associated with both cure status and/or survival. We demonstrated how such models may be estimated using two different iterative algorithms. The model-X knockoffs method was combined with these algorithms to control the false discovery rate (FDR) in variable selection. Through extensive simulation studies, our penalized MCMs have been shown to outperform alternative methods on multiple metrics and achieve high statistical power with FDR being controlled. In an acute myeloid leukemia (AML) application with gene expression data, our proposed approach identified 14 genes associated with potential cure and 12 genes with time-to-relapse, which may help inform treatment decisions for AML patients.

KEYWORDS
cure fraction, expectation-maximization, false discovery rate, forward stagewise, survival analysis

1 | INTRODUCTION

Medical breakthroughs in recent years have led to cures for various diseases including cancer. For example, improvement in outcomes has occurred in younger adults with acute myeloid leukemia (AML) during the past decades. Approximately

50% to 75% of adult AML patients treated with chemotherapy achieves complete remission (CR) and approximately 20% to 30% of the patients enjoy long-term disease-free survival.[1] Yanada et al[2] suggested that AML patients with 3-year relapse-free survival from the CR date can be considered "potentially cured." In practice, a long plateau in the Kaplan-Meier (K-M) estimated survival distribution with sufficient follow-up may indicate the presence of a cure fraction in the sample studied.

When a cure fraction is present, one assumption in regular survival models such as the Cox proportional hazards (PH) model, that all subjects will experience the event of interest, is violated. In this case, a Cox PH model tends to underestimate the hazard and overestimate the survival for the susceptible subjects.[3] A special type of survival model, namely, the mixture cure model (MCM), is frequently used in the presence of a cure fraction.[4-7] MCMs postulate that a fraction of the patients are cured with the survival probability of one and the other patients are susceptible to the event of interest with the failure time following a proper survival distribution, referred to as the latency distribution. The latency distribution can be depicted using parametric models[4] and semi-parametric PH models.[5-7] Parametric latency models have straightforward distributional assumptions and are easy to fit using various optimization methods for maximum likelihood estimation.

In order to better understand disease mechanisms, high-throughput genomic applications have been conducted to identify important biomarkers that are associated with some time-to-event outcomes. When the number of covariates exceeds the sample size, many traditional model-fitting approaches are not applicable. Effective variable selection procedures for high-dimensional data are needed to select a set of variables that are truly associated with the event of interest. Various penalization methods have been proposed and widely used during the past decades for variable selection in survival analysis, such as the least-absolute shrinkage and selection operator (LASSO)[8] and the adaptive LASSO.[9] However, only a few papers focused on the penalized MCMs when a cure fraction exists.

Liu et al[10] studied variable selection for the semi-parametric PH MCM, where both the LASSO and the smoothly clipped absolute deviation (SCAD) penalties were considered. The expectation-maximization (E-M) algorithm[11] was used to maximize the penalized likelihood. A recent paper[12] followed the work of Liu et al and extended the method to mixture and promotion time cure models[6] based on LASSO and adaptive LASSO.[13] Beretta and Heuchenne[14] generalized the PH MCM to accommodate time-varying covariates, using the SCAD penalty. Another approach by Scolas et al[15] adopted a parametric MCM with an accelerated failure time (AFT) regression model for latency and adaptive LASSO for variable selection on interval-censored data. However, a common limitation of these papers[10,12,14,15] is that only low-dimensional data, where the number of covariates is much smaller than the sample size, were considered.

Fan et al[16] proposed a penalization method for estimating the MCM where they explicitly considered the structural effects of covariates. That is, they postulated that the covariate effects on cure probability and those on survival function of susceptibles are potentially linearly related. Although the method is able to analyze high-dimensional data, their strong assumption of proportionality structure may constrain its applicability. They later proposed a less restrictive strategy which imposes a sign-based penalty to promote similarity in signs of the two-parts (referred to as SCinCRM in the remainder of this article).[17] Bussy et al[18] introduced a more general mixture model (C-Mix) which considered subgroups of patients with different prognosis and risks. The C-Mix model includes the MCM as a special case and applies to high-dimensional data, but it only allows subgroup membership, not the latency portion, to be driven by covariates. Additionally, the variable sets selected by the aforementioned methods are likely to contain redundant or noise variables, also known as false positives. Effectively controlling the false discovery rate (FDR), the expected proportion of false positives among all selected variables, without sacrificing power has been a major challenge in variable selection research.

This study aims to develop penalized parametric MCMs for high-dimensional datasets, which allow for identification of prognostic factors associated with both cure status and/or survival of susceptibles. Two different iterative algorithms, the generalized monotone incremental forward stagewise (GMIFS)[19] and the E-M algorithm,[11] are adopted for model estimation. These algorithms are further combined with the model-X knockoffs[20] which is a flexible selection framework that allows strict FDR control. The remainder of the article is organized as follows. Section 2 introduces the statistical models and algorithms, and presents a brief introduction of the model-X knockoffs framework. A simulation study designed to empirically compare our proposed methods with alternative approaches (including SCinCRM, C-Mix, and traditional survival models that do not consider cure) is reported in Section 3. In Section 4, we apply different methods to a high-throughput gene expression dataset for AML patients and present a list of genes that were identified as being associated with either the probability of cure or survival. Section 5 concludes the article with a discussion.

2 | STATISTICAL MODELS AND ALGORITHMS

2.1 | Weibull MCMs

Let $T \sim f(t)$ be a random non-negative continuous variable representing lifetime of interest (e.g., relapse-free survival time) with the cumulative distribution function (CDF) denoted by $F(t)$. A proportion π of subjects are susceptible ($Y = 1$) to the event of interest, and $1 - \pi$ are cured ($Y = 0$) who will not experience the event even after an extended follow-up, that is, $S(t|Y = 0) = 1$ for all t. Note the cure status Y is unknown for censored subjects. The survival function for the entire population is given by

$$S(t|\boldsymbol{x}, \boldsymbol{w}) = 1 - \pi(\boldsymbol{x}) + \pi(\boldsymbol{x})S(t|Y = 1, \boldsymbol{w}) = 1 - \pi(\boldsymbol{x})F(t|Y = 1, \boldsymbol{w}), \tag{1}$$

where $\boldsymbol{x}$ and $\boldsymbol{w}$ represent the covariates associated with incidence (whether one is cured) and latency (what is the survival function given one is uncured), respectively. Therefore, $F(t|\boldsymbol{x}, \boldsymbol{w}) = 1 - S(t|\boldsymbol{x}, \boldsymbol{w}) = \pi(\boldsymbol{x})F(t|Y = 1, \boldsymbol{w})$ and differentiating this with respect to t, the density is $f(t|\boldsymbol{x}, \boldsymbol{w}) = \pi(\boldsymbol{x})f(t|Y = 1, \boldsymbol{w})$. The likelihood L for right-censored survival data with a cured fraction is

$$\begin{aligned} L(\boldsymbol{\theta}) &\propto \prod_{i=1}^{N} [f(t_i|\boldsymbol{x}_i, \boldsymbol{w}_i)]^{\delta_i} [S(t_i|\boldsymbol{x}_i, \boldsymbol{w}_i)]^{1-\delta_i} \\ &= \prod_{i=1}^{N} [\pi(\boldsymbol{x}_i)f(t_i|Y_i = 1, \boldsymbol{w}_i)]^{\delta_i} [1 - \pi(\boldsymbol{x}_i)F(t_i|Y_i = 1, \boldsymbol{w}_i)]^{1-\delta_i}, \end{aligned} \tag{2}$$

where $\delta_i = 1$ indicates the failure time was observed while $\delta_i = 0$ indicates the observed time was censored. The log-likelihood is then proportional to

$$l(\boldsymbol{\theta}) \propto \sum_{i=1}^{N} [\delta_i \log(\pi(\boldsymbol{x}_i)f(t_i|Y_i = 1, \boldsymbol{w}_i)) + (1 - \delta_i) \log(1 - \pi(\boldsymbol{x}_i)F(t_i|Y_i = 1, \boldsymbol{w}_i))]. \tag{3}$$

In this article, a fully parametric likelihood function is considered. Specifically, the Weibull distribution, a popular parametric survival distribution that allows the formularization of both AFT and PH models, is assumed for the latency of the susceptibles. The Weibull density is $f(t|Y = 1) = \lambda\alpha(\lambda t)^{\alpha-1} \exp[-(\lambda t)^{\alpha}]$ where the shape α and scale λ parameters are both positive. We introduce the effects of the covariates $\boldsymbol{w}$ by replacing λ^{α} with $\lambda^{\alpha} \exp(\boldsymbol{\beta}^T\boldsymbol{w})$ such that

$$f(t|Y = 1, \boldsymbol{w}) = \lambda\alpha(\lambda t)^{\alpha-1} \exp(\boldsymbol{\beta}^T\boldsymbol{w}) \exp[-(\lambda t)^{\alpha} \exp(\boldsymbol{\beta}^T\boldsymbol{w})]. \tag{4}$$

The probability of being susceptible is most frequently modeled with logistic regression,[4,5,21,22] in which case we replace $\pi(\boldsymbol{x})$ with $\exp(b_0 + \boldsymbol{b}^T\boldsymbol{x})/(1 + \exp(b_0 + \boldsymbol{b}^T\boldsymbol{x}))$. Substituting these two expressions into Equation (3) yields the log-likelihood for the Weibull MCM, given by

$$\begin{aligned} l(\boldsymbol{\theta}) \propto \sum_{i=1}^{N} \Bigg\{ & \delta_i \log \left(\frac{\exp(b_0 + \boldsymbol{b}^T\boldsymbol{x}_i)}{1 + \exp(b_0 + \boldsymbol{b}^T\boldsymbol{x}_i)} \lambda\alpha(\lambda t_i)^{\alpha-1} \exp(\boldsymbol{\beta}^T\boldsymbol{w}_i) \exp\left[-(\lambda t_i)^{\alpha} \exp(\boldsymbol{\beta}^T\boldsymbol{w}_i)\right] \right) \\ & + (1 - \delta_i) \log \left(1 - \frac{\exp(b_0 + \boldsymbol{b}^T\boldsymbol{x}_i)}{1 + \exp(b_0 + \boldsymbol{b}^T\boldsymbol{x}_i)} \left(1 - \exp\left[-(\lambda t_i)^{\alpha} \exp(\boldsymbol{\beta}^T\boldsymbol{w}_i)\right]\right) \right) \Bigg\}. \end{aligned} \tag{5}$$

With high-dimensional covariate spaces, penalization is desired for both coefficient sets, $\boldsymbol{b}$ for incidence and $\boldsymbol{\beta}$ for latency. That being said, sometimes it is useful to coerce some covariates that are known risk factors into the model without penalty, such as baseline demographic and clinical characteristics. Hence, we further partition the two sets of predictors as $\boldsymbol{x} = (\boldsymbol{x}_u, \boldsymbol{x}_p)$ and $\boldsymbol{w} = (\boldsymbol{w}_u, \boldsymbol{w}_p)$, where the u subscript represents the unpenalized predictors that we wish to force into the model, while the p subscript represents the penalized predictors for which we seek a parsimonious model, such as genomic features. The parameters corresponding to the incidence portion of the model are given by $\boldsymbol{b} = (b_0, \boldsymbol{b}_u, \boldsymbol{b}_p)$ where b_0 is the intercept, while the parameters corresponding to the latency portion are $\boldsymbol{\beta} = (\boldsymbol{\beta}_u, \boldsymbol{\beta}_p)$, so that $[1, \boldsymbol{x}_i^T]\boldsymbol{b} = b_0 + \boldsymbol{b}_u^T\boldsymbol{x}_{u,i} + \boldsymbol{b}_p^T\boldsymbol{x}_{p,i}$ and $\boldsymbol{\beta}^T\boldsymbol{w}_i = \boldsymbol{\beta}_u^T\boldsymbol{w}_{u,i} + \boldsymbol{\beta}_p^T\boldsymbol{w}_{p,i}$. The unknown parameters are listed in $\boldsymbol{\theta} = (\alpha, \lambda, b_0, \boldsymbol{b}_u, \boldsymbol{b}_p, \boldsymbol{\beta}_u, \boldsymbol{\beta}_p)$, and the observed data include $\boldsymbol{O}_i = (t_i, \delta_i, \boldsymbol{x}_{u,i}, \boldsymbol{x}_{p,i}, \boldsymbol{w}_{u,i}, \boldsymbol{w}_{p,i})$, for $i = 1, \ldots, N$.

2.2 | Generalized monotone incremental forward stagewise

First, we used the GMIFS method[19] to estimate the penalized Weibull MCM for high-dimensional data. In the linear regression setting, the incremental forward stagewise (FS_ϵ) method is a version of boosting which produces a coefficient path strikingly similar to the L_1-penalized regression (LASSO) path along the penalization level.[23] FS_ϵ works by incrementing the coefficient of the variable most correlated with the current residuals by an amount $\pm\epsilon$ at each step. When the incremental amount $\epsilon \downarrow 0$, the algorithm (called FS_0) produces an identical path to the LASSO path under certain conditions.[24] FS_0 was later characterized as a monotone version of the LASSO with much smoother regularization paths.[19] The GMIFS method is a generalization of this characterization to problems involving other than squared error loss, such as the logistic regression model.[19] As long as the gradient functions can be derived, the GMIFS is theoretically applicable to any parametric models. In fact, it has been proven useful in a wide variety of high-dimensional settings for modeling discrete survival time,[25] ordinal,[26,27] or count responses.[28]

The GMIFS algorithm proceeds in an iterative fashion and updates one of the penalized coefficients by a small incremental amount at each iteration step. To determine which penalized covariate is to be updated, the algorithm adopts the steepest ascent method, that is, updating the coefficient associated with the largest gradient. In our case, one penalized incidence coefficient in $\boldsymbol{b}_p$ and one penalized latency coefficient in $\boldsymbol{\beta}_p$ were selected and updated at each step. To determine the direction of the update at each step, the expanded penalized design matrices $(\boldsymbol{X}_p, -\boldsymbol{X}_p)$ and $(\boldsymbol{W}_p, -\boldsymbol{W}_p)$ were used as input. Both expanded matrices were centered and scaled before entering the algorithm. The corresponding coefficients were then expanded to $(\boldsymbol{b}_p^+, \boldsymbol{b}_p^-)$ and $(\boldsymbol{\beta}_p^+, \boldsymbol{\beta}_p^-)$, and the Karush-Kuhn-Tucker condition ensures that at most one of $\boldsymbol{b}_j^+$ and $\boldsymbol{b}_j^-$ associated with the same covariate $\boldsymbol{x}_j$ (or $\boldsymbol{\beta}_j^+$ and $\boldsymbol{\beta}_j^-$ with $\boldsymbol{w}_j$) was greater than zero at the same time.[19] At each iteration, the selected coefficients were updated with a small incremental amount ϵ (set to be 0.001) so that the coefficient paths for both positive and negative parts are constrained to be monotonically nondecreasing. In the end of the algorithm, the solution paths for the original coefficients $\boldsymbol{b}_p$ and $\boldsymbol{\beta}_p$ were obtained by subtracting the coefficient estimates for the negative versions of the variables, from those for the positive counterparts.

The GMIFS algorithm for the penalized Weibull MCM is summarized in Algorithm 1. The algorithm starts with $\boldsymbol{b}_p^+ = \boldsymbol{b}_p^- = 0$, and $\boldsymbol{\beta}_p^+ = \boldsymbol{\beta}_p^- = 0$. The unpenalized parameters, α, λ, b_0, $\boldsymbol{b}_u$ and $\boldsymbol{\beta}_u$, are initialized using the Broyden-Fletcher-Goldfarb-Shanno (BFGS) optimization method with the method of moments (MOM) estimates for α and λ as starting values. The algorithm then iterates between updating the penalized parameters and updating the unpenalized parameters. To update the unpenalized parameters, the BFGS method is applied, considering the penalized parameters $(\boldsymbol{b}_p^+, \boldsymbol{b}_p^-)$ and $(\boldsymbol{\beta}_p^+, \boldsymbol{\beta}_p^-)$ fixed at the current estimates. If the difference between the log-likelihoods of two successive steps is less than a prespecified tolerance τ (set to be 10^{-5}), the iterative procedure is considered to be convergent. Otherwise, the algorithm stops after 10 000 iterations. To prevent over-fitting, cross-validation can be used to select the optimal step which yields the final estimates. The C-statistic designed for MCMs[29] (described in Section 3.3) is used as the cross-validation metric for the GMIFS as well as the E-M algorithm discussed in the next section.

2.3 | Expectation-maximization algorithm

The E-M algorithm is a natural choice for problems with hidden variables, such as the cure status Y_i in the MCMs.[10,12,16] In this section, we introduce the E-M and describe how it can be applied to the L_1 penalized MCMs.

Algorithm 1. GMIFS algorithm for penalized Weibull mixture cure models

1: Start with $\boldsymbol{b}_p^+ = \boldsymbol{b}_p^- = 0$ and $\boldsymbol{\beta}_p^+ = \boldsymbol{\beta}_p^- = 0$.
2: Initialize the unpenalized parameters, α, λ, b_0, $\boldsymbol{b}_u$, and $\boldsymbol{\beta}_u$, using a maximization algorithm of the log-likelihood.
3: Considering α, λ, b_0, $\boldsymbol{b}_u$, and $\boldsymbol{\beta}_u$ fixed, find the predictor $j = \arg\max\left(\frac{\partial l}{\partial \boldsymbol{b}_p}\right)$ and update $b_{p,j} \leftarrow b_{p,j} + \epsilon$ where $b_{p,j} \in \{\boldsymbol{b}_p^+, \boldsymbol{b}_p^-\}$; find the predictor $m = \arg\max\left(\frac{\partial l}{\partial \boldsymbol{\beta}_p}\right)$ and update $\beta_{p,m} \leftarrow \beta_{p,m} + \epsilon$ where $\beta_{p,m} \in \{\boldsymbol{\beta}_p^+, \boldsymbol{\beta}_p^-\}$.
4: Update α, λ, b_0, $\boldsymbol{b}_u$, and $\boldsymbol{\beta}_u$ by maximum likelihood given the current $\boldsymbol{b}_p^+$, $\boldsymbol{b}_p^-$, $\boldsymbol{\beta}_p^+$, and $\boldsymbol{\beta}_p^-$.
5: Repeat steps 3 and 4 until the difference between successive log-likelihoods is less than a prespecified tolerance τ.

Assuming that we have observed the cure status Y_i for $i = 1, \dots, N$, the complete-data likelihood function for the MCM is

$$L_C(\boldsymbol{\theta}) = \prod_{i=1}^{N} \pi(\boldsymbol{x}_i)^{Y_i}[1 - \pi(\boldsymbol{x}_i)]^{1-Y_i}[h(t_i|Y_i = 1, \boldsymbol{w}_i)^{\delta_i} S(t_i|Y_i = 1, \boldsymbol{w}_i))]^{Y_i}, \tag{6}$$

where $h(t_i|Y_i = 1, \boldsymbol{w}_i)$ is the hazard function of an uncured patient. In a Weibull MCM, the penalized complete-data log-likelihood with the LASSO penalty for both $\boldsymbol{b}_p$ and $\boldsymbol{\beta}_p$ can be written as

$$\begin{aligned} l_{\mathrm{CP}}(\boldsymbol{\theta}) = & \sum_{i=1}^{N} \left\{ Y_i \log \pi(\boldsymbol{x}_i) + (1 - Y_i) \log[1 - \pi(\boldsymbol{x}_i)] \right\} \\ & + \sum_{i=1}^{N} \left\{ \delta_i Y_i \log(\lambda\alpha) + (\alpha - 1)\delta_i Y_i \log(\lambda t_i) + \delta_i Y_i(\boldsymbol{\beta}^T \boldsymbol{w}_i) + Y_i[-(\lambda t_i)^{\alpha} \exp(\boldsymbol{\beta}^T \boldsymbol{w}_i)] \right\} \\ & - N \sum_{j=1}^{J} \mu_{1j} \left| b_{p,j} \right| - N \sum_{m=1}^{M} \mu_{2m} \left| \beta_{p,m} \right|, \end{aligned} \tag{7}$$

where J denotes the number of predictors in $\boldsymbol{b}_p$, M denotes the number of predictors in $\boldsymbol{\beta}_p$, and $\boldsymbol{\mu} = (\mu_{11}, \dots, \mu_{1J}, \mu_{21}, \dots, \mu_{2M})$ is a vector of tuning parameters tuned by cross-validation. For the sake of simplicity, the same value μ was used for all tuning parameters.

In the E-step, the algorithm calculated the expected $l_{\mathrm{CP}}(\boldsymbol{\theta})$ with respect to the conditional distribution of Y_i given the current parameter estimates $\hat{\boldsymbol{\theta}}$ and the observed data $\boldsymbol{O}_i$. Since the Y_i's are linear terms in $l_{\mathrm{CP}}(\boldsymbol{\theta})$, we only need to compute the expected value of Y_i given $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{O}_i$, denoted by $\hat{p}_i = E(Y_i|\hat{\boldsymbol{\theta}}, \boldsymbol{O}_i) = \mathbb{P}(Y_i = 1|\hat{\boldsymbol{\theta}}, \boldsymbol{O}_i)$. When the subject i was censored with $\delta_i = 0$,

$$\hat{p}_i = \frac{\hat{\pi}(\boldsymbol{x}_i)\hat{S}(t_i|Y_i = 1, \boldsymbol{w}_i)}{1 - \hat{\pi}(\boldsymbol{x}_i) + \hat{\pi}(\boldsymbol{x}_i)\hat{S}(t_i|Y_i = 1, \boldsymbol{w}_i)}, \tag{8}$$

according to the Bayes' theorem. When an event was observed for the subject i with $\delta_i = 1$, $\hat{p}_i$ equals to 1. We can integrate the two cases into one expression, given by

$$\hat{p}_i = \delta_i + (1 - \delta_i)\frac{\hat{\pi}(\boldsymbol{x}_i)\hat{S}(t_i|Y_i = 1, \boldsymbol{w}_i)}{1 - \hat{\pi}(\boldsymbol{x}_i) + \hat{\pi}(\boldsymbol{x}_i)\hat{S}(t_i|Y_i = 1, \boldsymbol{w}_i)}. \tag{9}$$

To obtain the expected $l_{\mathrm{CP}}(\boldsymbol{\theta})$, the E-step replaces Y_i in Equation (7) with $\hat{p}_i$ given above.

The M-step maximized the expected $l_{\mathrm{CP}}(\boldsymbol{\theta})$, which is equivalent to maximizing the logistic portion and the survival portion separately to obtain updated parameters. A trick was adopted to convert the L_1 penalization to a constrained optimization problem with a differentiable objective, by doubling the number of penalized parameters.[30] Specifically, for a real number a, we can write $|a| = a_+ + a_-$ and $a = a_+ - a_-$, where a_+ and a_- correspond to the positive and negative part of a, respectively, satisfying $a_+ \geq 0$ and $a_- \geq 0$. Then the limited-memory BFGS with bound constraints (L-BFGS-B) is applicable to efficiently solve the optimization problems in the M-step. The unpenalized parameters were updated using the regular BFGS method in each M-step.

The E-M algorithm is summarized in Algorithm 2. Similar to the GMIFS, α and λ were initialized using the MOM estimates, while $\boldsymbol{\beta}_u$ was initialized using estimates from a low-dimensional Cox model with covariates $\boldsymbol{W}_u$, and $\boldsymbol{b}_u$ was

Algorithm 2. E-M algorithm for penalized Weibull mixture cure models

1: Fix the tuning parameter μ and initialize $\boldsymbol{b}_u$, $\boldsymbol{b}_p$, α, λ, $\boldsymbol{\beta}_u$, and $\boldsymbol{\beta}_p$.
2: Execute the E-step by computing $\hat{p}_i = E(Y_i|\hat{\boldsymbol{\theta}}, \boldsymbol{O}_i)$ and replacing Y_i with $\hat{p}_i$ in $l_{\mathrm{CP}}(\boldsymbol{\theta})$.
3: Update $\boldsymbol{b}_p$ and $\boldsymbol{\beta}_p$ with the L-BFGS-B algorithm and the converting approach discussed above.
4: Update α, λ, b_0, $\boldsymbol{b}_u$, and $\boldsymbol{\beta}_u$, using a maximization algorithm of the log-likelihood given the current $\boldsymbol{b}_p$ and $\boldsymbol{\beta}_p$.
5: Repeat steps 2 to 4 until the difference between successive log-likelihoods is less than a prespecified tolerance τ.

initially set equal to $\boldsymbol{\beta}_u$ for simplicity. The algorithm then iterated between the E-step and the M-step until convergence. The E-M usually converged much quicker than the GMIFS thanks to the L-BFGS-B method, so a maximum iteration step of 100 was used. The optimal μ was selected using the cross-validated C-statistic for MCMs.[29]

2.4 | Model-X knockoffs

The model-X knockoffs approach[20] was recently developed as a flexible variable selection framework with exact finite-sample FDR control. Because the method places no restriction on data dimensionality or conditional distribution, it is known in theory to apply seamlessly to arbitrary response types including time-to-event data. In this framework, a set of "knockoff" variables $\tilde{\boldsymbol{Z}} = (\tilde{\boldsymbol{Z}}_1, \ldots, \tilde{\boldsymbol{Z}}_q)$ are constructed to mimic the covariance structure of the original covariates $\boldsymbol{Z} = (\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_q)$. Formally, for any subset of the covariates, $B \subset \{1, \ldots, q\}$, when the entries $\boldsymbol{Z}_j$ and $\tilde{\boldsymbol{Z}}_j$ are swapped for each $j \in B$, the joint distribution of $(\boldsymbol{Z}, \tilde{\boldsymbol{Z}})$ will remain invariant. Another property of the knockoff variables is that they are independent of the response conditionally on the original covariates. In this way, the knockoffs can be used as negative controls for the real covariates so that true signals can be teased apart from noise variables. Practically, the knockoff variables can be constructed using the second-order approximate approach[20] with the `knockoff` R package,[31] assuming that the covariates follow a multivariate normal distribution. When the covariates cannot be depicted by Gaussian distributions, deep generative models such as the deep knockoff machine[32] can be used to relax the distributional assumption and generate valid knockoffs in general settings. Since we have two sets of covariates $\boldsymbol{X}_p$ and $\boldsymbol{W}_p$ in our model, we constructed a group of knockoff variables for each set. The second-order approximate approach was used for the knockoff construction in the simulation studies in Section 3 and the deep knockoff machine was used in the application in Section 4 where the covariates were not normally distributed.

The knockoffs $\tilde{\boldsymbol{X}}_p$ and $\tilde{\boldsymbol{W}}_p$ were then augmented with the corresponding original covariates $\boldsymbol{X}_p$ and $\boldsymbol{W}_p$, respectively, to form two extended matrices, $[\boldsymbol{X}_p, \tilde{\boldsymbol{X}}_p]$ and $[\boldsymbol{W}_p, \tilde{\boldsymbol{W}}_p]$, with the number of predictors doubled. These extended matrices were used as new input matrices that entered the GMIFS or E-M algorithm. The absolute values of estimates of the penalized coefficients for the extended sets of covariates were obtained and used as the variable importance measurements in terms of explaining the cure status or the latency. For example, the importance measure for an incidence covariate $\boldsymbol{X}_{p,j} \in \{\boldsymbol{X}_p, \tilde{\boldsymbol{X}}_p\}$ is denoted by $U_j = |\hat{b}_{p,j}|$. We then calculated the difference between the importance measurement of the original predictor and that of the knockoff as the statistic used for variable selection, denoted by $V_j = U_j - \tilde{U}_j$. A data-dependent threshold for $\boldsymbol{b}_p$ is given by

$$\zeta = \min\left\{ v > 0 : \frac{1 + \#\{j : V_j \le -v\}}{\#\{j : V_j \ge v\}} \le \tau_{\text{FDR}} \right\}, \tag{10}$$

where τ_{FDR} denotes the target FDR level. The variables in $\boldsymbol{X}_p$ whose statistics exceed the threshold ζ are then selected. In this way, strict FDR control in variable selection is guaranteed,[20] meaning that on average, 80% of the variables selected by the framework are expected to be true signals if τ_{FDR} is set to 20%. A similar procedure is followed to select important variables from the latency covariates $\boldsymbol{W}_p$.

3 | SIMULATION STUDIES

3.1 | Competing approaches

Simulation studies were conducted to compare the performance of our methods with that of competing approaches. The SCinCRM[16] and C-Mix[18] are both relevant alternative methods with publicly available code and are thus included in the comparison. The SCinCRM aimed at promoting sign consistency between the incidence coefficients and the latency coefficients by imposing a sign-based penalty. The sign penalty term was governed by a tuning parameter which can be tuned to zero when the sign consistency assumption is not supported by the data. In that case, the method reduces to a penalized MCM. Besides the sign constraint, there are a few other differences between our method and the SCinCRM. They used the Cox PH model for the survival function of susceptible subjects, while we considered the Weibull model which is also a PH model. Instead of the L_1 penalty, the minimax concave penalty (MCP)[33] was adopted for the regression coefficients. They also applied the E-M algorithm to estimate the model but used coordinate descent in each M-step.

The C-Mix[18] method considers a general mixture model where different subgroups of patients have different prognosis and risks. The provided code supports the estimation of a MCM where one of the subgroups has zero risk. Compared with our method, they used a similar E-M procedure with the L-BFGS-B algorithm, but there are some distinctions between their model and ours. The most obvious distinction is that they only consider regression on the incidence part while we incorporate regression on both incidence and latency. They assume a simple geometric distribution in their code which allows for closed-form updates of a latency parameter that is identical for all subjects. Moreover, the elastic-net penalty[34] is used for the coefficient regularization instead of L_1.

Other than the aforementioned approaches, we would also like to investigate how a survival model without considering cure fraction would perform when a non-negligible subset of treated patients are cured. We thus included the L_1 penalized Weibull survival model (referred to as "non-cure Weibull" in the remainder of this article) and the L_1 penalized Cox PH model (referred to as "non-cure Cox") for comparison. Since no efficient implementation of the L_1 penalized Weibull model for high-dimensional data is available to our best knowledge (there is a Bayesian implementation[35] which is expected to be slow), we implemented the non-cure Weibull model via the GMIFS algorithm. The `glmnet` R package[36,37] was used to implement the L_1 penalized Cox model. These non-cure models were tuned using the cross-validated C-statistic based on only the latency coefficients (see Section 3.3 for a detailed description of the metric).

The R code for implementing the proposed estimation algorithms and conducting the simulation studies has been made available at https://github.com/hanfu-bios/curemodels.

3.2 | Simulation settings

In the simulations, we used the same set of covariates for incidence ($\boldsymbol{X}$) and latency ($\boldsymbol{W}$) with $N = 400$ observations, $J = 500$ penalized predictors and two unpenalized predictors. We randomly allocated 3/4 of the data to a training set and 1/4 to a testing set. The penalized covariates $\boldsymbol{X}_p = \boldsymbol{W}_p$ were generated from a J-dimensional Gaussian distribution MVN$(0, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma}$ was a block diagonal matrix with the block size of $B = 50$. Each block had an autoregressive structure with the $(i,j)^{\text{th}}$ element being $\sigma^2\rho^{|i-j|}$ where $\sigma = 0.5$ and $\rho = 0.2$ to reflect small correlations among predictors. The entries in the unpenalized covariate matrices $\boldsymbol{X}_u = \boldsymbol{W}_u$ were i.i.d. generated from a $N(0, \sigma^2)$ distribution with $\sigma = 0.5$.

There were $R = 10$ penalized predictors having nonzero effects (called signals) on each regression part. One signal predictor ($b_{p,j}$ or $\beta_{p,j}$) was randomly selected from each block and assigned a coefficient value A (called signal amplitude) taking on random signs, that is, $b_{p,j}, \beta_{p,j} \in \{\pm A\}$. We varied the value of A from 0.4 to 1.8 in our simulations to cover common effect sizes of interest for gene expression data. All other predictors within that block were simply "null" noise variables with zero coefficients. The unpenalized coefficients $\boldsymbol{b}_u$ and $\boldsymbol{\beta}_u$ were sampled from $N(\mu_u, \sigma_u^2)$ where $\mu_u \in \{\pm 0.3\}$ and $\sigma_u = 0.1$. In the default simulation setting, the incidence covariates and latency covariates were independently sampled. In order not to favor our methods over SCinCRM because of their sign consistency assumption, we also performed a simulation where the coefficients from the two parts, $\boldsymbol{b}_u$ and $\boldsymbol{\beta}_u$, $\boldsymbol{b}_p$ and $\boldsymbol{\beta}_p$, had the same signs.

Next, the cure status Y_i for $i = 1, \ldots, N$ was generated from a Bernoulli distribution with the mean of $\pi(\boldsymbol{x}_i) = \exp(b_0 + \boldsymbol{b}^T\boldsymbol{x}_i)/(1 + \exp(b_0 + \boldsymbol{b}^T\boldsymbol{x}_i))$. The intercept b_0 was sampled from $N(\mu_{b_0}, 0.01)$. To control the cure rate in the simulated data, the expected value for b_0 was varied. Two values were chosen, $\mu_{b_0} \in \{0.5, 1.5\}$, to reflect different cure rates approximately at 40% and 25%, respectively. For susceptible subjects with $Y_i = 1$, the event time T_i (in years) followed a Weibull distribution with the shape parameter α and the scale parameter $\lambda \exp[(\boldsymbol{\beta}_u^T\boldsymbol{w}_{u,i} + \boldsymbol{\beta}_p^T\boldsymbol{w}_{p,i})/\alpha]$. We set $\alpha = 1$ and $\lambda = 2$. The censoring time followed a uniform distribution on $[0, C_{\max}]$. If the event occurred later than the censoring ($T_i > C_i$) or the subject was cured ($Y_i = 0$), then the censoring indicator $\delta_i = 0$; otherwise, $\delta_i = 1$ indicating the event was observed. For uncured subjects, the observed time t_i was the smaller value between T_i and C_i, and for cured subjects, $t_i = C_i$. We set $C_{\max} = 20$ so that the censoring rate was approximately 45% when the cure rate was around 40%, and 31% when the cure rate was around 25%.

To test robustness of our methods, we first applied an alternative data-generating distribution for the event time T_i of susceptible subjects. The generalized gamma (GG) family is known to be an extensive family containing many survival distributions as special cases, including the Weibull. It has three parameters, the location parameter μ, scale parameter σ and shape parameter Q, with the density of

$$f_{\text{GG}}(t) = \frac{|Q|}{\sigma t\Gamma(Q^{-2})}\left[Q^{-2}(e^{-\mu}t)^{\frac{Q}{\sigma}}\right]^{Q^{-2}}\exp\left[-Q^{-2}(e^{-\mu}t)^{\frac{Q}{\sigma}}\right]. \tag{11}$$

When $\sigma = \frac{1}{\alpha}$, $\mu = -\log\lambda - \frac{\beta^T \boldsymbol{w}}{\alpha}$ and $Q = 1$, the GG density translates into the Weibull density we used in Equation (4). In the alternative data-generating process, the event time T_i for uncured subjects followed a GG distribution with σ and μ described above but $Q = 2$ instead, which made it not a Weibull density anymore and further, violated the PH assumption. We therefore investigated the model performance of the different methods when the latency models were misspecified.

Further, we examined the robustness of our methods in the case where the underlying data generating process was a non-MCM, or more specifically, a promotion time cure model.[38] The promotion time cure model was constructed in the context of cancer recurrence which is assumed to be promoted by carcinogenic cells that remain active after treatment. The unobserved number of carcinogenic cells N_i is incorporated through a Poisson model. In the simulation, we let N_i follow a Poisson distribution with the mean of $\exp(\boldsymbol{\beta}^T \boldsymbol{w}_i)$. For each carcinogenic cell, the time to activation followed a standard exponential distribution. If $N_i > 0$ for subject i, the event time T_i was the time when the first carcinogenic cell became activated; if $N_i = 0$, subject i was cured with $T_i = \infty$. The covariates $\boldsymbol{w}_i$, the coefficients $\boldsymbol{\beta}$ and the censoring times were generated in the same way as in our default simulation setting. The signal amplitude A was set to be 1. Under this simulation setting, we assessed the performance of our penalized MCMs when the model was completely different from the data generating process.

3.3 | Metrics for performance evaluation

In this section, we describe multiple metrics for performance evaluation and method comparison. Since variable selection is our primary objective, the most important metrics in this study are false discovery proportion (FDP) and power in selecting variables. The FDP is the realized version of FDR, calculated by the proportion of false positives among all variables selected, that is, variables with nonzero estimated coefficients in the regularized models. Power can be estimated by the proportion of true signals being identified. A low FDP and a high power are usually desired but a tradeoff exists between the two. When the knockoff framework is applied so that FDR is controlled at a target level, a model with a higher power is preferred.

Besides variable selection, we also assessed performance in terms of prediction and estimation bias. The concordance index (C-index or C-statistic) is a frequently used metric for censored data which measures the probability of concordance between prediction and observation. Given a predicted risk score R_i ($R_i = \hat{\boldsymbol{\beta}}^T \boldsymbol{w}_i$ in our model), the C-statistic for a standard survival model is the proportion of concordant pairs divided by the total number of possible evaluation pairs, given by

$$\hat{C} = \frac{\sum_{i=1}^{N}\sum_{j=1,j\neq i}^{N} I[R_i > R_j] I_{i,j}}{\sum_{i=1}^{N}\sum_{j=1,j\neq i}^{N} I_{i,j}}, \tag{12}$$

where $I_{i,j} = I[t_i < t_j, \delta_i = 1] + I[t_i = t_j, \delta_i = 1, \delta_j = 0]$. In this formula, the cured patients are not differentiated from those uncured but censored. A refined version was proposed[29] to take the cure status into account. Specifically, they applied a prespecified cutoff to assume whether a censored subject was cured or not. If the censoring time was beyond the cutoff time point, the subject was assumed to be cured. Otherwise, the cure status was unknown. In their cure status weighting approach, they assigned the weight of 1 for subjects who experienced the event ($y_i = 1$), 0 for presumptive cured subjects ($y_i = 0$), and the estimated probability of non-cure for other censored subjects with unknown cure status (y_i missing). The C-statistic is defined as

$$\hat{C} = \frac{\sum_{i=1}^{N}\sum_{j=1,j\neq i}^{N} I[\hat{\boldsymbol{\beta}}^T \boldsymbol{w}_i > \hat{\boldsymbol{\beta}}^T \boldsymbol{w}_j] \left\{ v_j y_j + (1 - v_j)\hat{\pi}(\boldsymbol{x}_j) \right\} I_{i,j}}{\sum_{i=1}^{N}\sum_{j=1,j\neq i}^{N} \left\{ v_j y_j + (1 - v_j)\hat{\pi}(\boldsymbol{x}_j) \right\} I_{i,j}}, \tag{13}$$

where v_j is the indicator of whether y_i is known and $\hat{\pi}(\boldsymbol{x}_j) = \exp(\hat{b_0} + \hat{\boldsymbol{b}}^T \boldsymbol{x}_j)/(1 + \exp(\hat{b_0} + \hat{\boldsymbol{b}}^T \boldsymbol{x}_j))$. In this way, the estimated coefficients from both regression parts are incorporated into the metric. For approaches with both regression parts (our methods and SCinCRM), we used the definition in Equation (13), while the Equation (12) was used for approaches with only one regression part (C-Mix, non-cure Weibull, and non-cure Cox models), with R_i calculated using the corresponding regression coefficients.

In an earlier paper,[39] Asano et al proposed an AUC metric for susceptibility prediction based on a similar weighting scheme which they called the mean score imputation method. Given the estimated probability of non-cure $\hat{\pi}(\boldsymbol{x}_i)$ and

a cutoff value a ($0 \le a \le 1$), the true positive rate (TPR) and false positive rate (FPR) for susceptibility prediction were estimated by

$$\widehat{\text{TPR}}(a) = \frac{\sum_{i=1}^{N} I[\hat{\pi}(\boldsymbol{x}_i) \ge a]\,\{v_i y_i + (1-v_i)\hat{\pi}(\boldsymbol{x}_i)\}}{\sum_{i=1}^{N}\{v_i y_i + (1-v_i)\hat{\pi}(\boldsymbol{x}_i)\}},$$
$$\widehat{\text{FPR}}(a) = \frac{\sum_{i=1}^{N} I[\hat{\pi}(\boldsymbol{x}_i) \ge a]\,\{v_i(1-y_i) + (1-v_i)[1-\hat{\pi}(\boldsymbol{x}_i)]\}}{\sum_{i=1}^{N}\{v_i(1-y_i) + (1-v_i)[1-\hat{\pi}(\boldsymbol{x}_i)]\}}. \tag{14}$$

Having the TPR and FPR values, the AUC can be estimated using the trapezoidal method. In addition, we also assessed susceptibility prediction in terms of prediction accuracy. Specifically, we used the K-M estimated survival probability for the last observed event $\hat{c}$ as a cured proportion cutoff. Then, the subjects with the top $(1-\hat{c})$th percentile of the estimated susceptible probability $\hat{\pi}(\boldsymbol{x}_i)$ were predicted as being susceptible ($\hat{Y}_i = 1$) and the others being cured ($\hat{Y}_i = 0$). Since the true susceptibility status Y_i was available in the simulation studies, we calculated the susceptibility prediction accuracy with the proportion of $\hat{Y}_i = Y_i$ for $i = 1, \ldots, N$. The AUC and accuracy metrics assess the predictive performance for the cure status and thus only apply to the approaches which include the incidence regression part.

A common limitation of the C-statistic with cure status weighting and the AUC with mean score imputation is the requirement of a prespecified cutoff for cure, which may be considered subjective or arbitrary. The cutoff point was used to produce a proxy y_i for the unobserved real cure status Y_i, based on which the predictive performance for the incidence portion can be measured. In this article, a cutoff of 5 years was used since it is a commonly used time point in cancer prognosis to indicate potential cure after a patient achieves complete remission. Under the default simulation setting, the probability of being cured given one's observed time $t > 5$ was around 95.3%, suggesting that the cutoff of 5 was reasonable for the simulations presented here. Researchers are advised to select a cutoff value tailored to their own applications and data.

Two additional metrics were applied to measure estimation bias, the relative model error (RME) and estimation error (ERR), as described in the SCinCRM paper.[17] They are defined as

$$\text{RME} = \frac{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T \boldsymbol{\Sigma} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})}{(\hat{\boldsymbol{\beta}}^* - \boldsymbol{\beta})^T \boldsymbol{\Sigma} (\hat{\boldsymbol{\beta}}^* - \boldsymbol{\beta})},$$
$$\text{ERR} = \frac{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})}{(\hat{\boldsymbol{\beta}}^* - \boldsymbol{\beta})^T (\hat{\boldsymbol{\beta}}^* - \boldsymbol{\beta})}, \tag{15}$$

where $\boldsymbol{\Sigma}$ is the covariance matrix of the covariates, and $\hat{\boldsymbol{\beta}}^*$ represents the oracle estimates derived from the models where only the true signals were included and the coefficients for the other covariates were forced to be zero. These two metrics are basically comparing the distance between estimates and true values with that between oracle estimates and true values. They were calculated to assess the estimation for both penalized incidence coefficients $\boldsymbol{b}_p$ and penalized latency coefficients $\boldsymbol{\beta}_p$.

3.4 | Simulation results

In this section, simulation results are reported for different approaches using the metrics previously described. Due to space limitations, we only presented here the results under the default simulation setting where $\mu_{b_0} = 0.5$ corresponds to a cure rate of roughly 40%, $\boldsymbol{b}$ and $\boldsymbol{\beta}$ were independently simulated, and a Weibull distribution was used for latency. The results under alternative settings ($\mu_{b_0} = 1.5$, same signs of $\boldsymbol{b}$ and $\boldsymbol{\beta}$, or GG distribution for latency) were similar and reported in the Appendix.

Figure 1 presents the cure prediction accuracy and C-statistic for different methods on training and testing data. The x-axis is the signal amplitude A in the data generating process. Each point in the plots represents the averaged value among 100 repetitive experiments. From the figure, we can observe that our methods (MCMs estimated by GMIFS or E-M) generally achieved better performance than the competing approaches in terms of these two metrics. The C-statistic for the non-cure Cox model had a tiny advantage for the training data but a noticeable disadvantage for the testing data in comparison to our MCMs. Further, cure prediction cannot be assessed when using the non-cure models because of

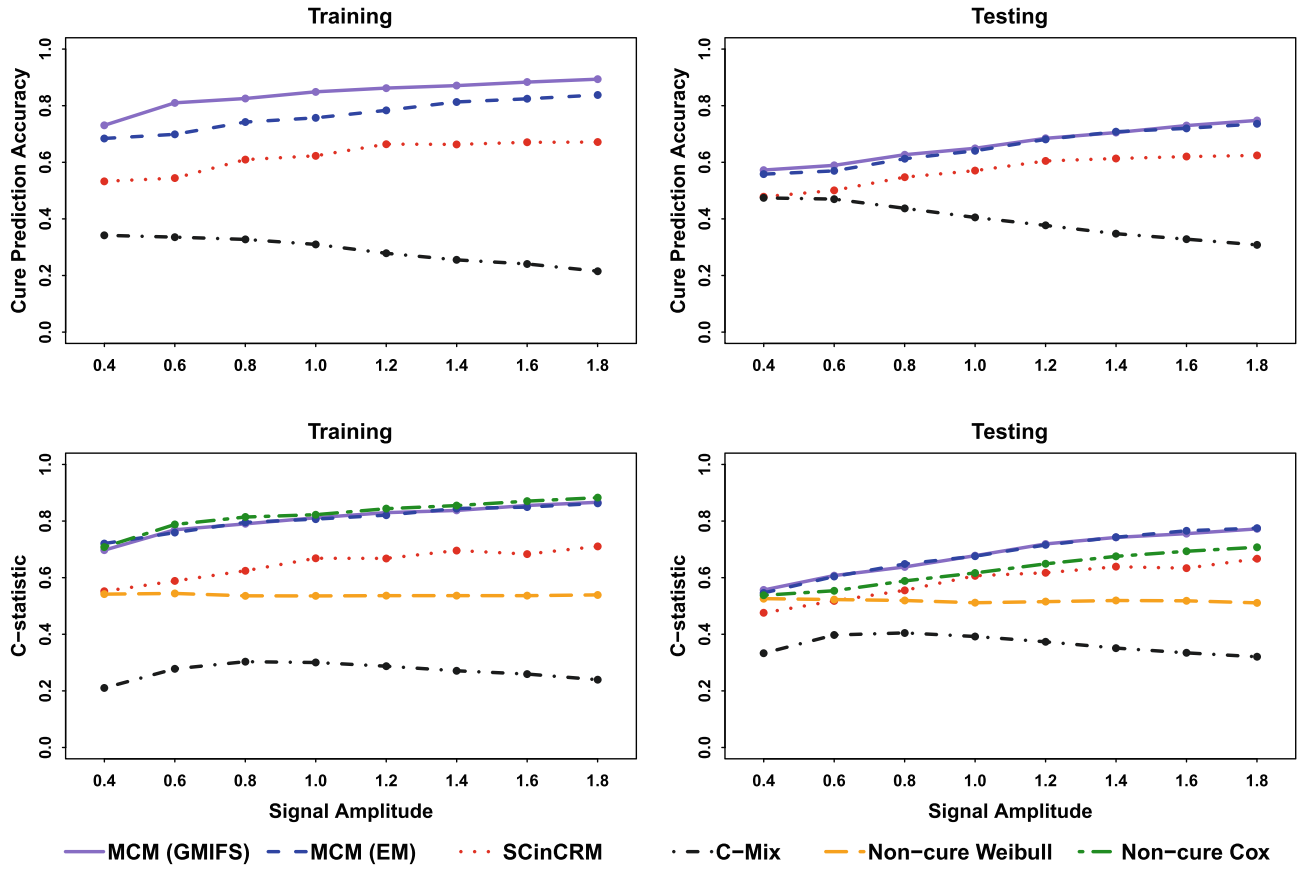

FIGURE 1 Cure prediction accuracy (top row) and C-statistic (bottom row) plotted by signal amplitude for the training (right) and testing (left) datasets for our mixture cure models (MCM) using the GMIFS and EM algorithm in comparison to SCinCRM, C-Mix, a non-cure Weibull and a non-cure Cox model. For our MCMs and SCinCRM, the C-statistic with cure status weighting was calculated, while the regular C-statistic was calculated for C-Mix and non-cure models

the lack of an incidence regression component. C-Mix performed the worst due to its simplified model assumption. As the signal became stronger, most methods tended to perform better except for C-Mix and the non-cure Weibull model. As observed, the testing performance was generally worse in comparison to the training performance, but again, that did not hold for C-Mix and the non-cure Weibull model.

Figure 2 shows the changes of the metrics of RME (in the left panels) and ERR (in the right panels) over signal amplitude for both incidence and latency parts. The incidence error for C-Mix and the latency error for MCM(EM) and the non-cure Cox model were extremely high when the signal amplitude was small and thus omitted from the figure for better presentation. In most cases, our GMIFS approach outperformed SCinCRM even though SCinCRM uses the MCP penalty which is known to enjoy the oracle property while the L_1 penalty does not. With our variance-covariance matrix $\mathbf{\Sigma}$, the differences between the values of RME and those of ERR were unnoticeable.

To compare the performance in variable selection, we applied the model-X knockoffs framework to each of the methods and set the target FDR level set at 20%. The FDP and power results of different methods are presented in Figure 3. We can see the FDP of all methods was well controlled below the target FDR level, except for the non-cure Cox model. In the meantime, the power increased as the amplitude increased and the incidence power was generally lower than the latency power, but our GMIFS and E-M methods achieved higher power than the other approaches in both regression parts. The power for SCinCRM was pretty low all the time and the non-cure Weibull models performed the worst in latency variable selection given the power of almost zero.

The results under alternative simulation settings are presented in the Appendix. Figures A1 to A3 show the simulation results when $\mu_{b_0} = 1.5$ corresponding to a cure rate of roughly 25%. The results were very similar to what we have observed in Figures 1 to 3. From Figure A2, SCinCRM has achieved slightly better estimation error than our GMIFS approach when the signals were strong. The results when the latency followed a GG distribution are presented in Figures A4 to A6. With

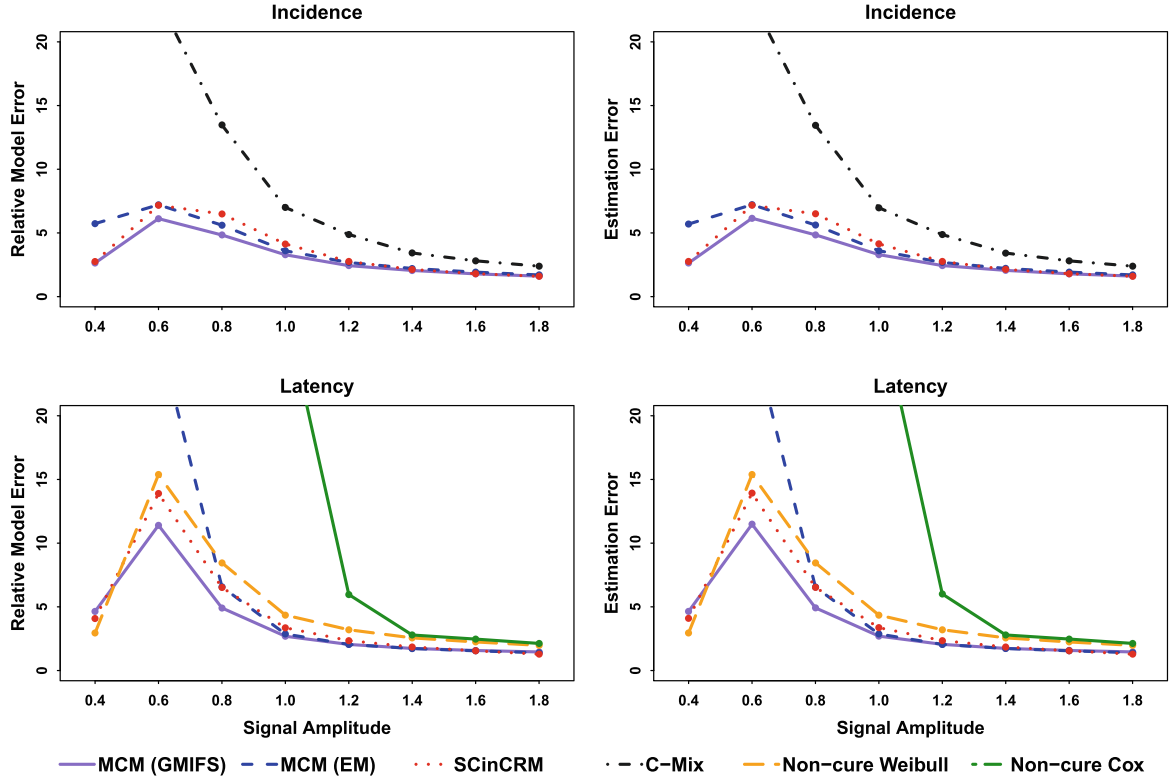

FIGURE 2 Relative model error (left) and estimation error (right) plotted by signal amplitude for incidence (top row) and latency (bottom row) for our mixture cure models (MCM) using the GMIFS and EM algorithm in comparison to SCinCRM, C-Mix, a non-cure Weibull, and a non-cure Cox model

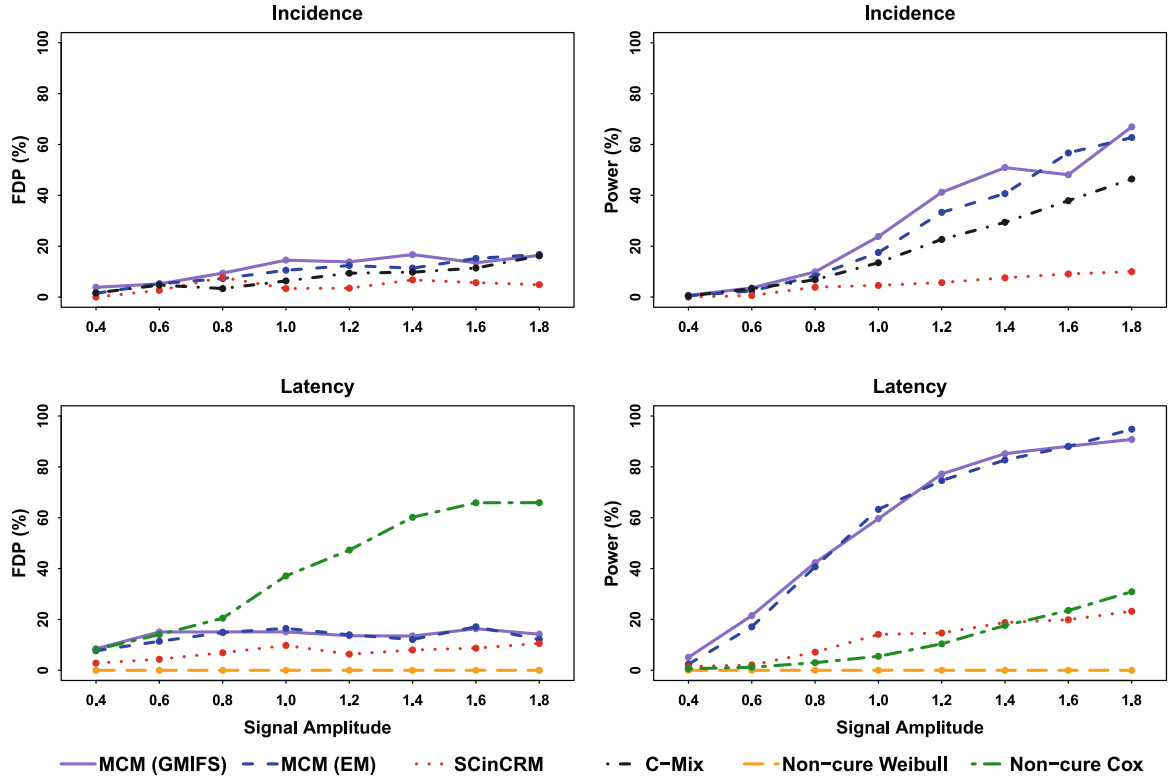

FIGURE 3 False discovery proportion (left) and power (right) plotted by signal amplitude for incidence (top row) and latency (bottom row) for our mixture cure models (MCM) using the GMIFS and EM algorithm in comparison to SCinCRM, C-Mix, a non-cure Weibull, and a non-cure Cox model

the misspecified latency distribution, the latency power of our Weibull MCMs was lower in comparison to our default simulation scenario, but still higher than the competitors. Other metrics were not noticeably hampered, which suggests our methods may be robust to latency model misspecification. Under the simulation setting where the event times were generated from a promotion cure model, only the regular C-statistic in Equation (12) was assessed as it was the only metric that was still meaningful when we fit a two-component MCM to data from a one-component data generating process. With 50 repetitive experiments where the signal amplitude was 1, our MCM using GMIFS achieved an average C-statistic of 0.778 in training sets and 0.572 in testing sets. In contrast, the non-cure Weibull model had an average C-statistic of 0.547 in training sets and 0.539 in testing sets. The results indicate that our MCM still achieves better predictive performance than the non-cure model even under severe deviations from the model assumptions, as long as a cure fraction exists.

Figures A7 to A9 display the results when the true penalized incidence and latency coefficients were identical ($\boldsymbol{b}_p = \boldsymbol{\beta}_p$), and the unpenalized incidence and latency coefficients ($\boldsymbol{b}_u$ and $\boldsymbol{\beta}_u$) had the same signs. The power for SCinCRM became higher under the same-sign settings, but was still lower than that for our GMIFS and E-M methods. Due to the two-part interconnections, the incidence power of our methods was boosted, so was the power for C-Mix and non-cure Cox model which only have one regression component. It is worth noting that the non-cure Cox model performed the best in terms of all metrics when the signal was strong enough under this simulation scenario. These results indicate that the non-cure Cox model may still be a good choice even when a cured fraction is present, provided the true signals from the two regression components are the same, having identical or at least strongly correlated coefficients. This is a rather limited scenario, though. In most cases, we would not expect the incidence and latency components to be comprised of exactly the same variables with the same effect sizes. Therefore, our two-component MCMs have clear advantages when different variables influence cure probability or survival of the susceptibles.

4 | APPLICATION ON AML

4.1 | Data

We applied the methods to a dataset containing 816 adult AML patients who were treated on frontline CALGB/Alliance protocols. Almost all of these patients received intensive cytarabine and daunorubicin or idarubicin-based induction treatment on CALGB/Alliance trials between 1986 and 2016. Institutional review board approval of all CALGB/Alliance protocols was obtained before any research was performed. In accordance with the Declaration of Helsinki, patients provided study-specific written informed consent to participate in treatment and companion cytogenetic (CALGB 8461), leukemia tissue bank (CALGB 9665), and molecular (CALGB 20202) studies, which involved collection of pretreatment bone marrow aspirates and blood samples. No patient received allogeneic stem cell transplantation (allo-SCT) in first complete remission (CR) on study protocols, and off-study patients who received an allo-SCT were excluded from the outcome analyses due to missing follow up data. Only those younger than 60 years old at enrollment were included into the analysis because for them, the treatment may be promising enough to lead to cures. We also limited the samples to those who had complete baseline and demographic data.

In the context of cure, relapse-free survival (RFS) is more relevant than overall survival (OS) because, obviously, one cannot be cured from death. We thus considered the patients who had attained a CR, which left us with 452 AML patients, and defined RFS as the duration between the date of CR and relapse or death, whichever was earlier. We used the traditional definition for RFS because patients who died prior to their visit may have relapsed, so this is a conservative estimate. The censoring rate was 34.5%. Out of the 296 events observed, most of the patients (87.1%) experienced AML relapse, indicating our definition of event is a good proxy for relapse. We performed a hypothesis test to detect whether there was a significant nonzero cure fraction for RFS,[40] which was significant (P-value $<10^{-4}$). Figure 4 depicts the RFS estimated using the K-M method. The long plateau in the RFS distribution demonstrates there is empirical evidence of sufficient follow-up as well as a fraction of long-term survivors at roughly 30%. The median follow-up among those censored for RFS was 8.95 years. Other than identifying a sufficient follow-up from a plateau of K-M curve, formal statistical tests[40,41] can be performed. Identifying a significant cure fraction and sufficient follow-up is usually the recommended first step before one applies cure models to survival data.[42] In fact, cure models may yield biased estimates when these assumptions are violated.[43]

Along with the time-to-event response, the dataset contains gene expression levels of 35 226 RNA transcripts captured using ribosomal RNA-depleted protocols, allowing for quantification of mRNA and non-coding RNAs. We filtered

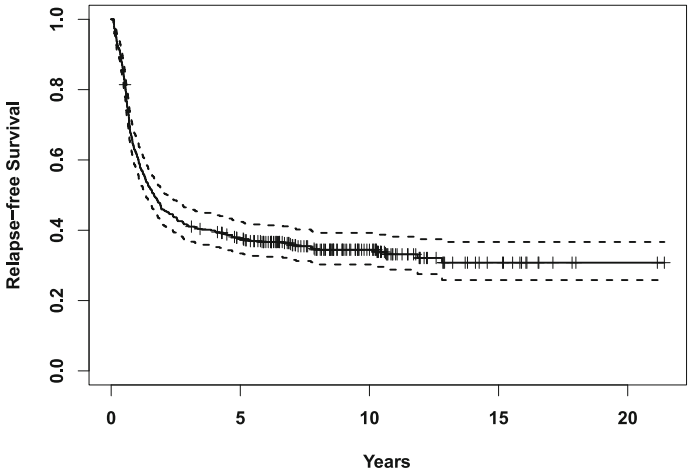

FIGURE 4 Relapse-free survival for 452 AML patients younger than 60 years at enrolment treated on frontline CALGB/Alliance protocols

TABLE 1 Descriptive statistics of unpenalized covariates in the Alliance data

Variables	Total, $N = 452$
Incidence covariates	
ELN 2017, no. (%)	
Favorable	279 (61.7%)
Intermediate	99 (21.9%)
Adverse	74 (16.4%)
WT1 mutation, no. (%)	35 (7.7%)
Latency covariates	
Platelets, mean (SD)	73.1 (63.6)
WBC, mean (SD)	41.7 (49.0)
FLT3-ITD, no. (%)	104 (23.0%)
TET2 mutation, no. (%)	36 (8.0%)
NRAS mutation, no. (%)	77 (17.0%)

Abbreviations: ELN 2017, European LeukemiaNet prognostic group;[45] *FLT3*-ITD, presence of *FLT3*-internal tandem duplication; SD, standard deviation; WBC, white blood cell count.

out the transcripts with low expression (mean value ≤5), leaving $J = 4887$ transcripts for the analyses. Some baseline/demographic characteristics were also available in the dataset, including age, race, sex, European Leukemia Net (ELN) risk group, cytogenetic abnormality (binary), white blood cell (WBC) count, hemoglobin, platelet count, percent of blasts in bone marrow and in peripheral blood, as well as mutation status for 18 known AML-associated genes. In the preliminary analysis, we performed a stepwise variable selection in a MCM[44] among the baseline/demographic variables. The cutoff of 0.05 was used to add or remove covariates based on *P*-values from likelihood ratio tests. The selected baseline variables for incidence and latency were considered as unpenalized covariates $\boldsymbol{X}_u$ and $\boldsymbol{W}_u$, respectively, whose descriptive statistics were displayed in Table 1. The gene expression data were used as penalized covariates $\boldsymbol{X}_p$ and $\boldsymbol{W}_p$ in the following analyses. We elected to include this initial screening process among baseline/demographic variables as we were interested in discovering genes associated with outcome of AML patients after controlling for commonly measured variables. Generally, researchers should appeal to existing literature, prior knowledge, or expert clinical opinion when determining whether to include unpenalized covariates vs enforcing penalties on all predictors.

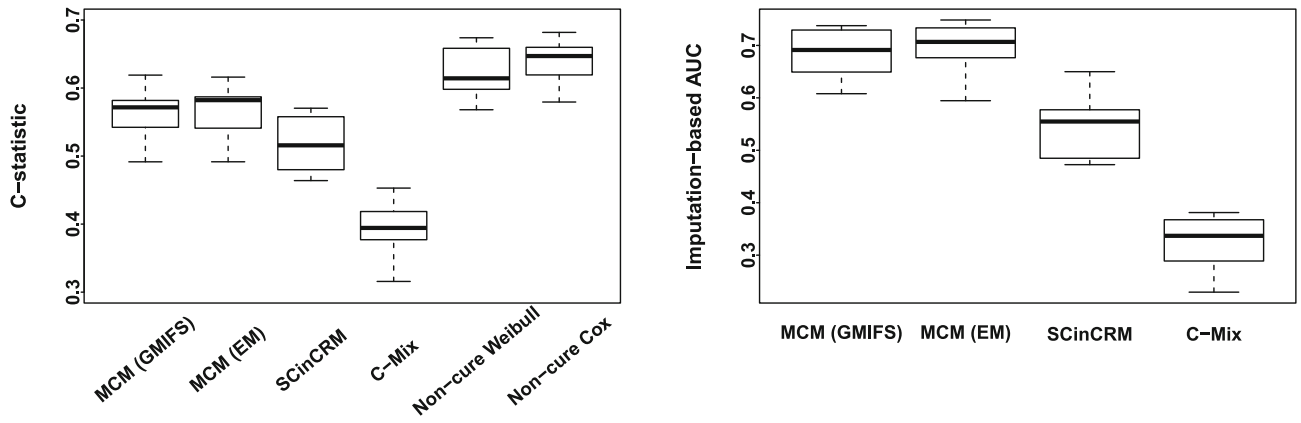

FIGURE 5 Boxplots of C-statistics (left) and AUC (right) from the testing datasets from 10 repeats of a data-splitting approach to evaluate the model performance on the AML dataset. Performance was compared among our mixture cure models (MCM) using the GMIFS and EM algorithm, SCinCRM, C-Mix, a non-cure Weibull, and a non-cure Cox model. For our MCMs and SCinCRM, the C-statistic with cure status weighting was calculated, while the regular C-statistic was calculated for C-Mix and non-cure models

4.2 | Method comparison

A data-splitting approach was used to evaluate the model performance on the real data. We split the dataset into a training set of size $3N/4$ and a testing set of size $N/4$. Then we fit different models on the training data and evaluated the performance on the testing data. Specifically, the C-statistic and the imputation-based AUC[39] described in Section 3.3 were measured for evaluation, since these two metrics do not rely on underlying true values of parameters (which are not available in real data). Again, the C-statistic with cure status weighting in Equation (12) was calculated for our MCMs and SCinCRM, while the regular C-statistic in Equation (13) was calculated for C-Mix and non-cure models which only have one regression component. The data-splitting process was repeated 10 times to reflect variability.

Figure 5 displays the C-statistic and AUC results for different methods. Our GMIFS and E-M methods achieved better results than SCinCRM and C-Mix in this application dataset. The non-cure models obtained a high C-statistic but they were unable to provide incidence-based information to calculate AUC for cure prediction. We also recorded the running time for different approaches. The non-cure Cox model was the fastest (less than 1 minute per repetition) due to the effective implementation of the `glmnet` R package.[36,37] C-Mix was also relatively fast (30 minutes per repetition) because of its simple model assumptions. SCinCRM was quite slow for such high-dimensional data and spent around two days for a single repetition. The other three methods (GMIFS, E-M and non-cure Weibull model) took a few hours, among which the E-M algorithm appeared to be the fastest.

4.3 | Gene discovery

In this section, we fit the models using the data we have after filtering ($N = 452$). Before combining with the model-X knockoffs, the estimates for unpenalized coefficients in the model were obtained using GMIFS or E-M and reported in Table A1. The estimates from GMIFS and E-M were very similar and their magnitudes and signs were mostly consistent with previous findings. For example, when we look at the coefficients for the ELN risk group[45] which stratifies AML patients into genetic-risk categories using selected cytogenetic abnormalities and gene mutations, subjects in the Intermediate and Adverse risk groups were associated with higher probabilities of being susceptible in contrast to the favorable group (reference group).

The model-X knockoffs framework was then combined with different methods for controlled variable selection. Since the gene expression values typically do not strictly follow a Gaussian distribution, we used the deep knockoff machine[32] to generate the knockoff copies. With the target FDR level set to be 20% for both incidence and latency coefficients, the number of selected genes for different methods was displayed in the first two rows of Table 2. The specific selected genes are presented in Figures A10 and A11 in the form of Venn diagrams. Since GMIFS was the only method that successfully identified important genes from both regression parts, we only focused on the GMIFS findings. Table 3 and 4 report the

TABLE 2 Number of selected genes for models fit using the Alliance data and validation performance using three other publicly available AML datasets

		MCM (GMIFS)	MCM (E-M)	SCin CRM	C-Mix	Non-cure Weibull	Non-cure Cox
# selected in incidence		14	0	0	29	-	-
# selected in latency		12	9	16	-	0	39
GSE37642	AUC	0.944	-	-	0.961	-	-
	C-index	0.679	0.660	0.606	0.712	-	0.849
GSE12417	AUC	1	-	-	1	-	-
	C-index	0.844	0.668	0.849	0.783	-	0.889
TCGA	AUC	0.917	-	-	1	-	-
	C-index	0.751	0.691	0.715	0.744	-	0.808

TABLE 3 Fourteen selected genes in the incidence regression part using GMIFS combined with the model-X knockoffs framework

Gene	Description
COQ8A	SNP associated with leukocyte count in GWAS studies[52]
EIF4E3	Tumor suppressor, reduced expression in AML specimens[53]
FAM30A	High expression was an adverse risk factor in AML[54]
TNPO1	Potential direct target of mixed lineage leukemia fusion proteins[55]
PEX2	Overexpressed in hepatocellular carcinoma (HCC) tissues[56]
KDM2B	Oncogene in diverse cancers, inducing cell proliferation[57]
SRSF2	Mutations found in AML, associated with inferior RFS[58]
RUNX2	Homolog of RUNX1, a central player in hematopoiesis[59]
ADA	Increased levels associated with short survival in AML[60]
MXD1	Tumor suppressor, expression affected AML survival[61]
IDS	Mutations lead to Hunter syndrome[52]
STK17B	High expression associated with apoptosis and poor OS[62,63]
PFDN5	Candidate for a tumor suppressor gene[64]
TAF9B	Essential for cell viability, associated with apoptosis[52]

Abbreviation: GWAS, genome-wide association study.

selected genes by GMIFS for incidence and latency, respectively, as well as brief description of known associations with AML or oncology in general. Some genes are known oncogenes or tumor suppressor genes, and some have been shown to differentially express in AML or other cancers.

4.4 | Validation of the identified genes

Independent datasets were employed to validate our selected genes, including two datasets from Gene Expression Omnibus (GEO) and one from The Cancer Genome Atlas (TCGA).[46] It is worth noting that there exist two major disparities between our dataset and the validation datasets. First, relapse-free survival data were not available in these independent datasets, so overall survival was used instead as a proxy. Second, the validation datasets included patients who received allo-SCT while our dataset only included patients with intensive chemotherapy. The first GEO dataset (GSE37642) includes the gene expression for 136 AML patients treated in the German AMLCG 1999 trial. The gene expression levels were measured using the Affymetrix HG-U133Plus2 GeneChip and each observation corresponded to a

TABLE 4 Twelve selected genes in the latency regression part using GMIFS combined with the model-X knockoffs framework

Gene	Description
CALCOCO2	Up-regulated in high-risk myelodysplastic syndrome[65]
TAX1BP1	Protects from liver cancer development[66]
LMO2	Oncogene of T-cell acute lymphoblastic leukemia[67]
RNU5B-1	Differentially expressed in various cancers[68,69]
H2BC21	Expression associated with AML relapse and prognosis[70]
EREG	Promotes progression of various cancers[71]
ABCC4	Regulates leukemia cell proliferation and differentiation[72]
CAMK2G	Suppresses differentiation and stimulates proliferation[73]
PLXNC1	Part of immunological signature predictive of prognosis[74]
PITRM1	Regulating mitochondrial function in AML[75]
TMEM50A	Highly up-regulated in late stage cervical cancer[76]
TMCO3	Up-regulated in chronic lymphocytic leukemia[77]

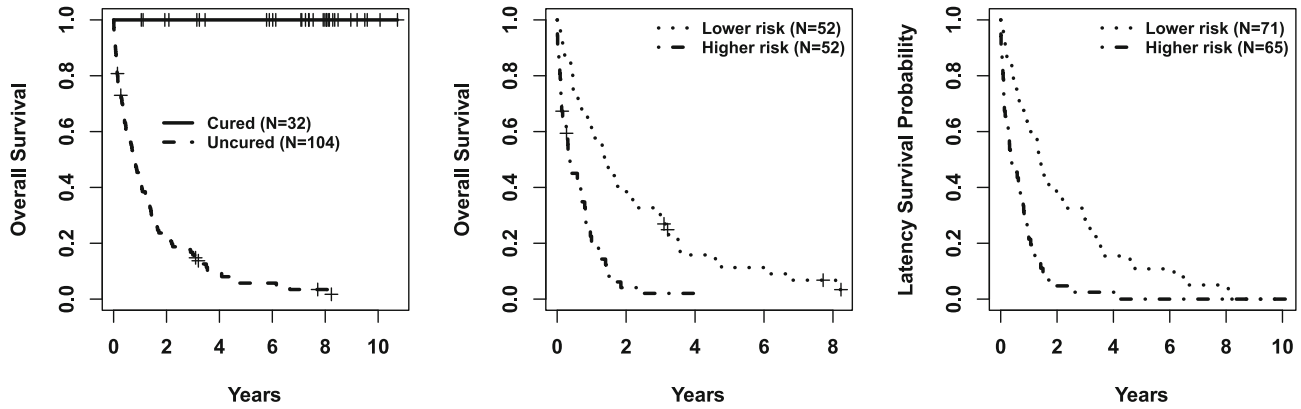

FIGURE 6 Using GSE37642 as a validation dataset and extracting Affymetrix probe sets that mapped to our selected genes from our training phase to assign subjects into groups based on our MCM, Kaplan-Meier estimates for predicted cured ($N = 32$) vs uncured ($N = 104$) groups (left panel), Kaplan-Meier estimates for subjects with lower risk ($N = 52$) vs higher risk ($N = 52$) among those that were predicted to be uncured (middle panel), and re-scaled latency survival functions for all 136 subjects with lower ($N = 71$) vs higher ($N = 65$) risk (right panel)

probe set instead of a RNA transcript as in the Alliance data. Thus, the predictors in the original and validation datasets are distinct and the coefficients cannot be directly applied. As a workaround, we linked RNA transcripts and probe sets through their corresponding genes and included all probe sets that mapped to our selected transcripts in Tables 3 and 4 into the Weibull MCM.

According to the estimated coefficients, we assigned the 136 subjects into different subgroups, including cured vs susceptible, and having low risk vs high risk of death. The predicted cured status was determined by the estimated incidence coefficients (see Section 3.3) and the high/low risk group was determined by the estimated latency coefficients and a cut-off of 0, that is, $I(\hat{\boldsymbol{\beta}}^T \boldsymbol{w} > 0)$ with $\boldsymbol{w}$ centered. The left panel in Figure 6 presents the K-M estimates for predicted cured ($N = 32$) vs uncured ($N = 104$) groups, and the middle panel in Figure 6 presents the K-M estimates for subjects with lower risk ($N = 52$) vs higher risk ($N = 52$) among those that were predicted to be uncured. From the figures, the predicted cured group had the survival probability of 1 as expected, while the predicted uncured group had an estimated survival function that descended toward 0. The two risk groups among those predicted to be susceptible were well separated, with a P-value of 2×10^{-7} for a log-rank test.

The above estimation method for latency survival probabilities only includes the subjects who were predicted to be uncured and thus relies on accurate predictions of each individual's cure status. A recent book on cure models[42] proposed

an alternative way to estimate latency. In this method, the latency is estimated using the information from all subjects and only depends on the cured proportion, which can be easily estimated. The latency regression part can be evaluated independently of the incidence part as a result. Specifically, the K-M estimates are rescaled through $\hat{S}' = (\hat{S} - \hat{c})/(1 - \hat{c})$ where $\hat{S}$ is the original K-M value and $\hat{c}$ is the estimated cured proportion, so that the latency survival probabilities range from 0 to 1. The right panel of Figure 6 displays the rescaled latency survival functions for all 136 subjects with lower ($N = 71$) vs higher ($N = 65$) risk. The survival curves were almost identical to the middle panel of Figure 6 which only included predicted uncured subjects, suggesting the relevance of our latency predictors as well as the accuracy of our cure status prediction.

Similar validation results for the other two datasets (GSE12417 and TCGA data) are shown in Figures A12 and A13. Due to the small sample sizes and potentially insufficient follow-up of these two datasets, the predicted cure status was not as accurate as that in GSE37642, but well-separated risk groups could still be observed in both datasets. Given the small sample sizes of the validation datasets, we did not further split the data to training and testing sets and thus the predictive performance may have been overestimated due to overfitting. However, since the main focus of the validation process is on gene selection rather than model fitting, we believe the current process can serve the purpose well. We validated the relevance and importance of our concise list of genes by using only the gene expression levels for these genes to predict cure status and survival probabilities in the new datasets. The results indicate that our identified gene list is useful in differentiating patient subgroups and may serve as a prognostic signature to better inform treatment decisions for AML patients with different genomic characteristics.

We also performed the same validation procedure for the genes selected by other approaches. For each method, we fit the corresponding model on different validation datasets including only the genes selected by that method, and calculated the AUC and C-index for the fitted model which are summarized in Table 2. Because the non-cure Weibull did not select any features with the FDR controlled, it was excluded in the validation. Our E-M approach and SCinCRM both only selected genes in the latency part, so the incidence-based AUC cannot be calculated. C-Mix has achieved pretty good AUC and C-index scores which is partly due to the long list of genes selected in the incidence part, similarly for the latency part of the non-cure Cox model. Overall, our GMIFS method is strongly recommended as it selected succinct lists of important genes for both regression components and had good predictive performance on independent datasets.

5 | DISCUSSION AND CONCLUSION

In this article, we proposed penalized Weibull MCMs to model censored survival outcomes in the presence of a cure fraction for high-dimensional data. The models allow us to estimate the effects of covariates on both the probability of cure and time-to-event of the susceptibles simultaneously. Two estimation algorithms (GMIFS and E-M) have been adapted to fit these models and combined with the model-X knockoffs framework for FDR control. The simulation results demonstrate that our proposed methods outperformed competing approaches in terms of variable selection, estimation, and prediction. In variable selection, we achieved high power with the FDR being controlled after adopting the knockoffs framework. In the AML application, important genes have been identified to be associated with cure and/or survival, which may have important implications for AML research or clinical practice paradigms.

One shortcoming of our study is that the incidence power is relatively low according to our simulation results. Although the incidence part has the form of a logistic regression, it is essentially a more difficult problem to tackle than a typical logistic regression. Unlike a regular binary problem, the response is hidden in the data and not explicitly observed. Besides, due to censoring and low cure rate, the effective sample size can be relatively small compared with the number of predictors, resulting in a poor signal-to-noise ratio. Any method that effectively boosts the incidence power may be interesting to explore in the future. One potential weakness in our application study is that we assume all AML patients who died prior to relapse died due to unobserved relapse and we used the death time as a surrogate for relapse time in this case. This assumption may not hold in all cases because some patients may have died of irrelevant events like car accidents. If we strictly consider relapse as the event of interest, then death is a competing risk for relapse and competing risk models may be of interest. Two semi-parametric regression models for competing risks data with a cured fraction based on finite-mixture models were introduced by Peng and Yu.[42]

In the future, we are interested in generalizing the proposed methods to penalized semi-parametric PH MCMs and other parametric MCMs incorporating distributions in the generalized gamma and generalized F families. Penalties other than LASSO, including the MCP[33] and the elastic-net penalty,[34] are interesting directions to work on. Besides penalized regression models, machine learning techniques including random survival forests[47] and gradient boosting machines[48] in

combination with the model-X knockoffs framework are promising approaches to high-dimensional variable selection in cure models. These techniques have been shown useful for controlled variable selection in other outcome types including continuous, binary,[49] and ordinal responses.[50] A previous paper[51] applied bagging survival trees to cure models based on the promotion time cure framework, but they only provided variable importance scores for latency covariates. How to extract variable importance and perform variable selection for both incidence and latency using machine learning techniques warrants further investigation.

ACKNOWLEDGEMENTS

The authors are grateful to the patients who consented to participate in clinical trials and the families who supported them; to Christopher Manring and the CALGB/Alliance Leukemia Tissue Bank at The Ohio State University Comprehensive Cancer Center, Columbus, OH, for sample processing and storage services; to Lisa J. Sterling for data management; and to Michael Pennell for helpful discussion about simulations and survival modeling. We would also like to thank the reviewers for valuable feedback and suggestions regarding an earlier version of this article.

FUNDING INFORMATION

Research reported in this publication was supported by the National Library of Medicine of the National Institutes of Health under Award Number R01LM013879. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. Research reported in this publication was also supported in part by the National Cancer Institute of the National Institutes of Health under Award Numbers U10CA180821, U10CA180882, U24CA196171, UG1CA233180, UG1CA233331, UG1CA233338, R35CA197734, P30CA016058; the Coleman Leukemia Research Foundation; ASH Junior Faculty Scholar Award and ASH Bridge Grant (A-KE), Leukemia Research Foundation (A-KE), Leukemia & Lymphoma Society (A-KE); The D Warren Brown Foundation; and by an allocation of computing resources from The Ohio Supercomputer Center and Shared Resources (Leukemia Tissue Bank). Support to Alliance for Clinical Trials in Oncology and Alliance Foundation Trials programs is listed at https://acknowledgments.alliancefound.org.

DATA AVAILABILITY STATEMENT

Data are available from Gene Expression Omnibus (GEO) using accession numbers GSE37642 and GSE12417 and from The Cancer Genome Atlas (TCGA-LAML). The R code for implementing the proposed estimation algorithms and conducting the simulation studies has been made available at https://github.com/hanfu-bios/curemodels.

ORCID

Kellie J. Archer https://orcid.org/0000-0003-1555-5781

REFERENCES

1. Tallman MS, Gilliland DG, Rowe JM. Drug therapy for acute myeloid leukemia. *Blood*. 2005;106(4):1154-1163.
2. Yanada M, Garcia-Manero G, Borthakur G, Ravandi F, Kantarjian H, Estey E. Potential cure of acute myeloid leukemia: analysis of 1069 consecutive patients in first complete remission. *Cancer*. 2007;110(12):2756-2760.
3. Price DL, Manatunga AK. Modelling survival data with a cured fraction using frailty models. *Stat Med*. 2001;20(9-10):1515-1527.
4. Farewell VT. The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics*. 1982;38(4):1041-1046.
5. Kuk AY, Chen CH. A mixture model combining logistic regression with proportional hazards regression. *Biometrika*. 1992;79(3):531-541.
6. Peng Y, Dear KB. A nonparametric mixture model for cure rate estimation. *Biometrics*. 2000;56(1):237-243.
7. Sy JP, Taylor JM. Estimation in a Cox proportional hazards cure model. *Biometrics*. 2000;56(1):227-236.
8. Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc Ser B Stat Methodol*. 1996;58(1):267-288.
9. Zhang HH, Lu W. Adaptive Lasso for Cox's proportional hazards model. *Biometrika*. 2007;94(3):691-703.
10. Liu X, Peng Y, Tu D, Liang H. Variable selection in semiparametric cure models based on penalized likelihood, with application to breast cancer clinical trials. *Stat Med*. 2012;31(24):2882-2891.
11. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. *J R Stat Soc Ser B Stat Methodol*. 1977;39(1):1-22.
12. Masud A, Tu W, Yu Z. Variable selection for mixture and promotion time cure rate models. *Stat Methods Med Res*. 2018;27(7):2185-2199.
13. Zou H. The adaptive lasso and its oracle properties. *J Am Stat Assoc*. 2006;101(476):1418-1429.
14. Beretta A, Heuchenne C. Variable selection in proportional hazards cure model with time-varying covariates, application to US bank failures. *J Appl Stat*. 2019;46(9):1529-1549.
15. Scolas S, El Ghouch A, Legrand C, Oulhaj A. Variable selection in a flexible parametric mixture cure model with interval-censored data. *Stat Med*. 2016;35(7):1210-1225.

16. Fan X, Liu M, Fang K, Huang Y, Ma S. Promoting structural effects of covariates in the cure rate model with penalization. *Stat Methods Med Res*. 2017;26(5):2078-2092.
17. Shi X, Ma S, Huang Y. Promoting sign consistency in the cure model estimation and selection. *Stat Methods Med Res*. 2020;29(1):15-28.
18. Bussy S, Guilloux A, Ga¨ffas S, Jannot AS. C-mix: a high-dimensional mixture model for censored durations, with applications to genetic data. *Stat Methods Med Res* 2019; 28(5): 1523–1539.
19. Hastie T, Taylor J, Tibshirani R, Walther G. Forward stagewise regression and the monotone lasso. *Electron J Stat*. 2007;1:1-29.
20. Candes E, Fan Y, Janson L, Lv J. Panning for gold: "model-X" knockoffs for high dimensional controlled variable selection. *J R Stat Soc SerB Stat Methodol*. 2018;80(3):551-577.
21. Farewell VT. Mixture models in survival analysis: Are they worth the risk? *Can J Stat*. 1986;14(3):257-262.
22. Fang HB, Li G, Sun J. Maximum likelihood estimation in a semiparametric logistic/proportional-hazards mixture model. *Scand J Stat*. 2005;32(1):59-75.
23. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York, NY: Springer Science & Business Media; 2009.
24. Efron B, Hastie T, Johnstone I, Tibshirani R. Least angle regression. *Ann Stat*. 2004;32(2):407-499.
25. Ferber K, Archer KJ. Modeling discrete survival time using genomic feature data. *Cancer Inform*. 2015;14(Suppl 2):37-43.
26. Hou J, Archer KJ. Regularization method for predicting an ordinal response using longitudinal high-dimensional genomic data. *Stat Appl Genet Mol Biol*. 2015;14(1):93-111.
27. Archer KJ, Hou J, Zhou Q, Ferber K, Layne JG, Gentry AE. ordinalgmifs: an R package for ordinal regression in high-dimensional data settings. *Cancer Inform*. 2014;13:187-195.
28. Makowski M, Archer KJ. Generalized monotone incremental forward stagewise method for modeling count data: application predicting micronuclei frequency. *Cancer Inform*. 2015;14(Suppl 2):97-105.
29. Asano J, Hirakawa A. Assessing the prediction accuracy of a cure model for censored survival data with long-term survivors: application to breast cancer data. *J Biopharm Stat*. 2017;27(6):918-932.
30. Tsujii J. Evaluation and extension of maximum entropy models with inequality constraints. Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing; 2003:137-144.
31. Patterson E, Sesia M. *Knockoff: the knockoff filter for controlled variable selection*; 2018. R package version 0.3.2.
32. Romano Y, Sesia M, Candès E. Deep knockoffs. *J Am Stat Assoc*. 2020;115(532):1861-1872.
33. Zhang CH. Nearly unbiased variable selection under minimax concave penalty. *Ann Stat*. 2010;38(2):894-942.
34. Zou H, Hastie T. Regularization and variable selection via the elastic net. *J R Stat Soc Series B Stat Methodol*. 2005;67(2):301-320.
35. Newcombe PJ, Raza Ali H, Blows FM, et al. Weibull regression with Bayesian variable selection to identify prognostic Tumour markers of breast cancer survival. *Stat Methods Med Res*. 2017;26(1):414-436.
36. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw*. 2010;33(1):1.
37. Simon N, Friedman J, Hastie T, Tibshirani R. Regularization paths for Cox's proportional hazards model via coordinate descent. *J Stat Softw*. 2011;39(5):1.
38. Yakovlev AY, Asselain B, Bardou V, et al. A simple stochastic model of tumor recurrence and its application to data on premenopausal breast cancer. *Biometrie et analyse de donnees spatio-temporelles*. 1993;12:66-82.
39. Asano J, Hirakawa A, Hamada C. Assessing the prediction accuracy of cure in the Cox proportional hazards cure model: an application to breast cancer data. *Pharm Stat*. 2014;13(6):357-363.
40. Maller RA, Zhou X. *Survival Analysis with Long-Term Survivors*. New York: Wiley; 1996:525.
41. Klebanov LB, Yakovlev AY. A new approach to testing for sufficient follow-up in cure-rate analysis. *J Stat Plan Inference*. 2007;137(11):3557-3569.
42. Peng Y, Yu B. *Cure Models: Methods, Applications, and Implementation*. Boca Raton: CRC Press; 2021.
43. Othus M, Bansal A, Erba H, Ramsey S. Bias in mean survival from fitting cure models with limited follow-up. *Value Health*. 2020;23(8):1034-1039.
44. Asano J, Hirakawa A, Hamada C. A stepwise variable selection for a Cox proportional hazards cure model with application to breast cancer data. *Jpn J Biometr*. 2013;34(1):21-34.
45. Döhner H, Estey E, Grimwade D, et al. Diagnosis and management of AML in adults: 2017 ELN recommendations from an international expert panel. *Blood*. 2017;129(4):424-447.
46. Cancer Genome Atlas Research Network,Ley T, Miller C. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *N Engl J Med*. 2013;368(22):2059-2074.
47. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. *Ann Appl Stat*. 2008;2(3):841-860.
48. Friedman JH. Greedy function approximation: a gradient boosting machine. *Ann Stat*. 2001;29(5):1189-1232.
49. Shen A, Fu H, He K, Jiang H. False discovery rate control in cancer biomarker selection using knockoffs. *Cancers (Basel)*. 2019;11(6):744.
50. Fu H, Archer KJ. High-dimensional variable selection for ordinal outcomes with error control. *Brief Bioinform*. 2021;22(1):334-345.
51. Mbogning C, Broët P. Bagging survival tree procedure for variable selection and prediction in the presence of nonsusceptible patients. *BMC Bioinform*. 2016;17(1):1-21.
52. Stelzer G, Rosen N, Plaschkes I, et al. The GeneCards suite: from gene data mining to disease genome sequence analyses. *Curr Protoc Bioinform*. 2016;54(1):1-30.
53. Volpon L, Osborne MJ, Culjkovic-Kraljacic B, Borden KL. eIF4E3, a new actor in mRNA metabolism and tumor suppression. *Cell Cycle*. 2013;12(8):1159-1160.

54. Wang YH, Lin CC, Hsu CL, et al. Distinct clinical and biological characteristics of acute myeloid leukemia with higher expression of long noncoding RNA *KIAA0125*. *Ann Hematol*. 2021;100(2):487-498.
55. Li Z, Chen P, Su R, et al. PBX3 and MEIS1 cooperate in hematopoietic cells to drive acute myeloid Leukemias characterized by a core transcriptome of the *MLL*-rearranged disease. *Cancer Res*. 2016;76(3):619-629.
56. Cai M, Sun X, Wang W, et al. Disruption of peroxisome function leads to metabolic stress, mTOR inhibition, and lethality in liver cancer cells. *Cancer Lett*. 2018;421:82-93.
57. Nakamura S, Tan L, Nagata Y, et al. JmjC-domain containing histone demethylase 1B-mediated $p15^{Ink4b}$ suppression promotes the proliferation of leukemic progenitor cells through modulation of cell cycle progression in acute myeloid leukemia. *Mol Carcinog*. 2013;52(1):57-69.
58. Bamopoulos SA, Batcha AMN, Jurinovic V, et al. Clinical presentation and differential splicing of *SRSF2*, *U2AF1* and *SF3B1* mutations in patients with acute myeloid leukemia. *Leukemia*. 2020;34(10):2621-2634.
59. Okuda T, Nishimura M, Nakao M, Fujitaa Y. RUNX1/AML1: a central player in hematopoiesis. *Int J Hematol*. 2001;74(3):252-257.
60. Patmasiriwat P, Anukarahanonta T, Chinprasertsuk S. Purine degradative enzymes and terminal transferase in acute myelogenous leukemia: clinical relevance. *Ann Clin Lab Sci*. 1993;23(4):281-289.
61. Lacayo NJ, O'Brien M, Jain S, et al. Gene expression profiling predicts outcome in de novo acute myeloid Leukemia (AML) with normal karyotype: results of children's oncology group (COG) study POG# 9421. *Blood*. 2006;108(11):1915.
62. Ye P, Zhao L, Gonda TJ. The *MYB* oncogene can suppress apoptosis in acute myeloid leukemia cells by transcriptional repression of *DRAK2* expression. *Leuk Res*. 2013;37(5):595-601.
63. Stavropoulou V, Kaspar S, Brault L, et al. *MLL-AF9* expression in hematopoietic stem cells drives a highly invasive AML expressing EMT-related genes linked to poor outcome. *Cancer Cell*. 2016;30(1):43-58.
64. Fujioka Y, Taira T, Maeda Y, et al. MM-1, a c-Myc-binding protein, is a candidate for a tumor suppressor in Leukemia/Lymphoma and tongue cancer. *J Biol Chem*. 2001;276(48):45137-45144.
65. Colombo AR, Zubair A, Thiagarajan D, Nuzhdin S, Triche TJ, Ramsingh G. Suppression of transposable elements in leukemic stem cells. *Sci Rep*. 2017;7(1):7029.
66. Waidmann O, Pleli T, Weigert A, et al. Tax1BP1 limits hepatic inflammation and reduces experimental hepatocarcinogenesis. *Sci Rep*. 2020;10(1):16264.
67. El Omari K, Hoosdally SJ, Tuladhar K, et al. Structure of the leukemia oncogene LMO2: implications for the assembly of a hematopoietic transcription factor complex. *Blood*. 2011;117(7):2146-2156.
68. Koduru SV, Leberfinger AN, Ravnic DJ. Small non-coding RNA abundance in adrenocortical carcinoma: a footprint of a rare cancer. *J Genom*. 2017;5:99-118.
69. Serrano-Gómez SJ, Sanabria-Salas MC, Garay J, et al. Ancestry as a potential modifier of gene expression in breast tumors from Colombian women. *PloS One*. 2017;12(8):e0183179.
70. Ye C, Ma S, Xia B, Zheng C. Weighted gene coexpression network analysis identifies cysteine-rich intestinal protein 1 (CRIP1) as a prognostic gene associated with relapse in patients with acute myeloid leukemia. *Med Sci Monit*. 2019;25:7396-7406.
71. Sunaga N, Kaira K. Epiregulin as a therapeutic target in non-small-cell lung cancer. *Lung Cancer (Auckl)*. 2015;6:91-98.
72. Copsel S, Garcia C, Diez F, et al. Multidrug resistance protein 4 (MRP4/ABCC4) regulates cAMP cellular levels and controls human leukemia cell proliferation and differentiation. *J Biol Chem*. 2011;286(9):6979-6988.
73. Si J, Collins SJ. Activated Ca^{2+}/calmodulin-dependent protein kinase IIγ is a critical regulator of myeloid leukemia cell proliferation. *Cancer Res*. 2008;68(10):3733-3742.
74. Ragaini S, Wagner S, Marconi G, et al. A three-gene immune signature including *IDO1*, *BIN1* and *PLXNC1* predicts survival in acute Myeloid Leukemia. *Blood*. 2020;136:35-36.
75. Cole A, Wang Z, Coyaud E, et al. Inhibition of the mitochondrial protease ClpP as a therapeutic strategy for human acute myeloid leukemia. *Cancer Cell*. 2015;27(6):864-876.
76. Edgar R, Domrachev M, Lash AE. Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucl Acids Res*. 2002;30(1):207-210.
77. Fernàndez V, Jares P, Salaverria I, et al. Gene expression profile and genomic changes in disease progression of early-stage chronic lymphocytic leukemia. *Haematologica*. 2008;93(1):132-136.

How to cite this article: Fu H, Nicolet D, Mrózek K, et al. Controlled variable selection in Weibull mixture cure models for high-dimensional data. *Statistics in Medicine*. 2022;41(22):4340-4366. doi: 10.1002/sim.9513

APPENDIX

TABLE A1 Estimated unpenalized covariates using GMIFS and E-M

Incidence			Latency		
Covariate	**GMIFS**	**E-M**	**Covariate**	**GMIFS**	**E-M**
ELN 2017 Intermediate	0.62	0.64	Platelets	−0.26	−0.36
ELN 2017 Adverse	0.92	0.94	WBC	0.16	0.18
WT1	0.29	0.29	*FLT3*−ITD	0.23	0.22
			TET2	−0.12	−0.12
			NRAS	0.16	0.16

Abbreviations: ELN 2017, European LeukemiaNet prognostic group;[45] *FLT3*-ITD, presence of *FLT3*-internal tandem duplication; SD, standard deviation; WBC, white blood cell count.

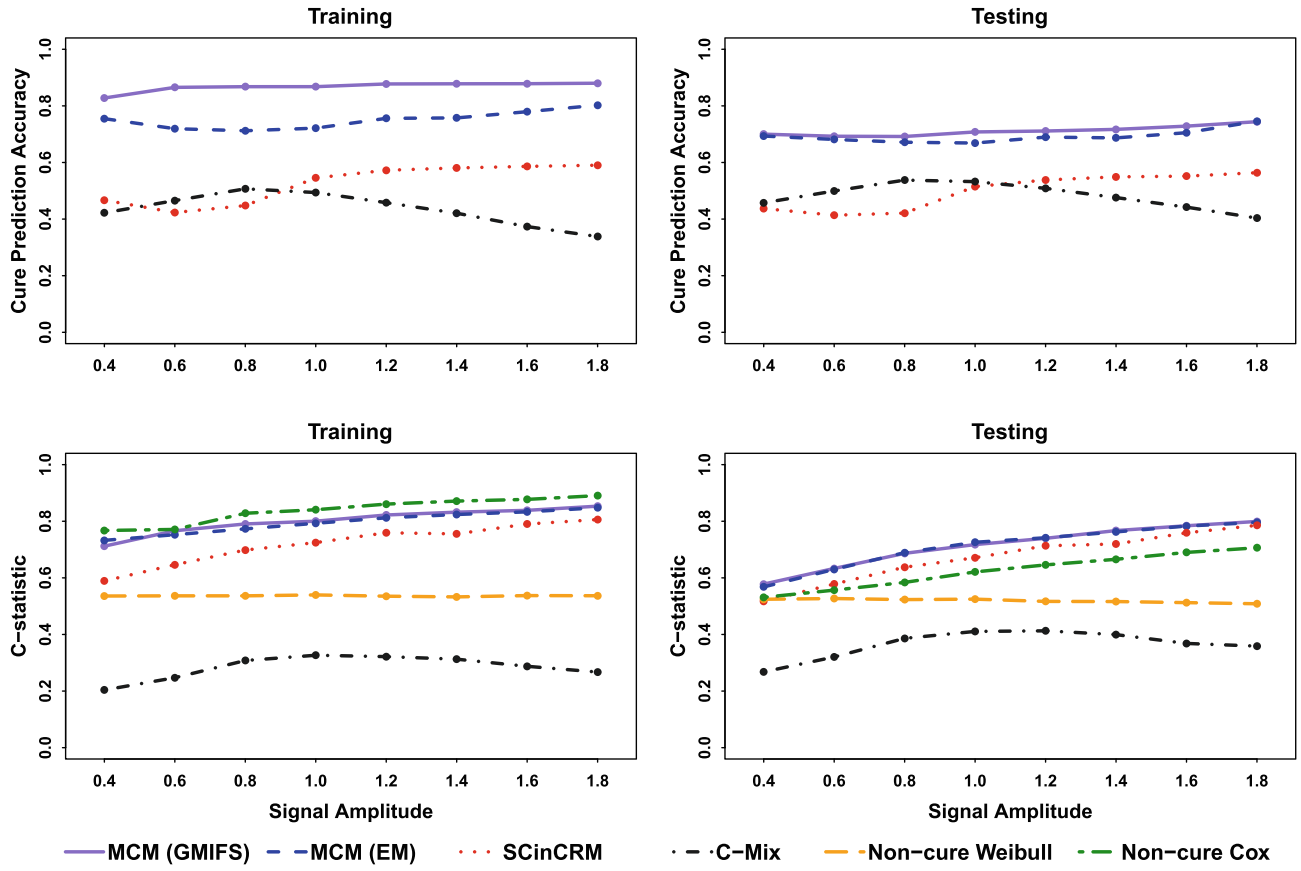

FIGURE A1 For a cure fraction of roughly 25%, cure prediction accuracy (top row) and C-statistic (bottom row) plotted by signal amplitude for the training (right) and testing (left) datasets for our mixture cure models (MCM) using the GMIFS and EM algorithm in comparison to SCinCRM, C-Mix, a non-cure Weibull and a non-cure Cox model. For our MCMs and SCinCRM, the C-statistic with cure status weighting was calculated, while the regular C-statistic was calculated for C-Mix and non-cure models

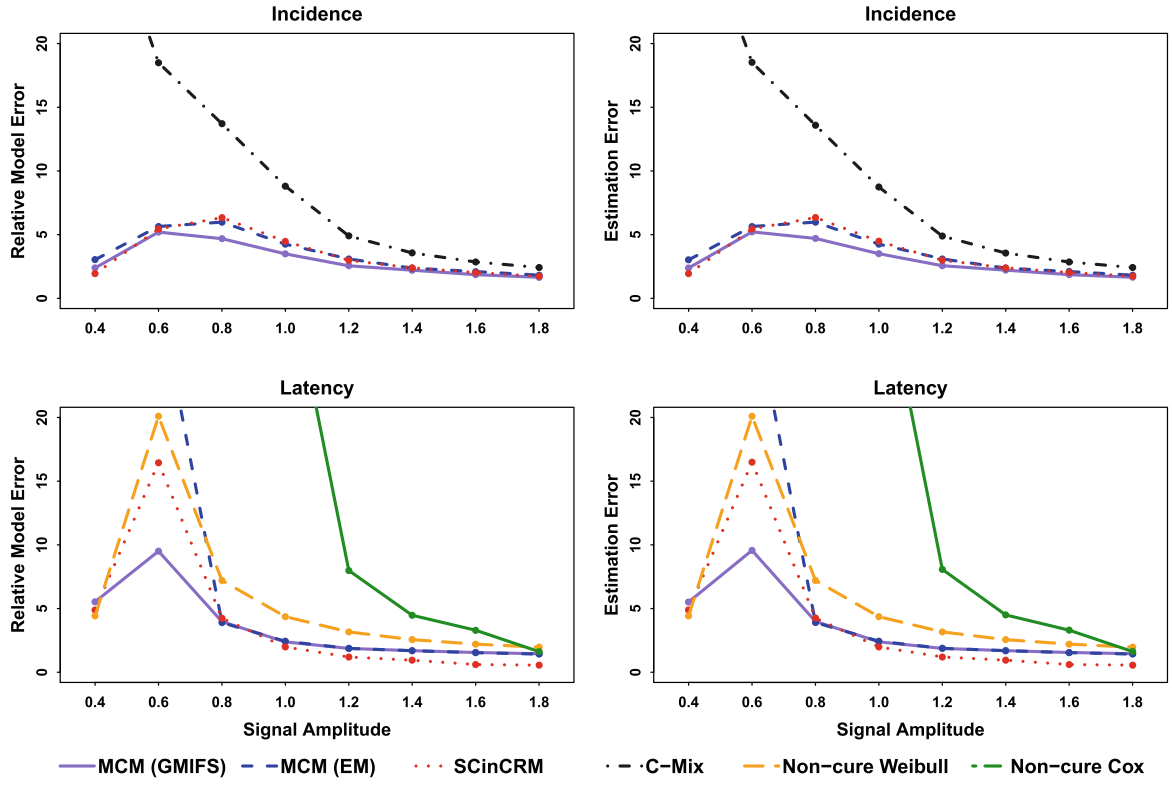

FIGURE A2 For a cure fraction of roughly 25%, relative model error (left) and estimation error (right) plotted by signal amplitude for incidence (top row) and latency (bottom row) for our mixture cure models (MCM) using the GMIFS and EM algorithm in comparison to SCinCRM, C-Mix, a non-cure Weibull and a non-cure Cox model

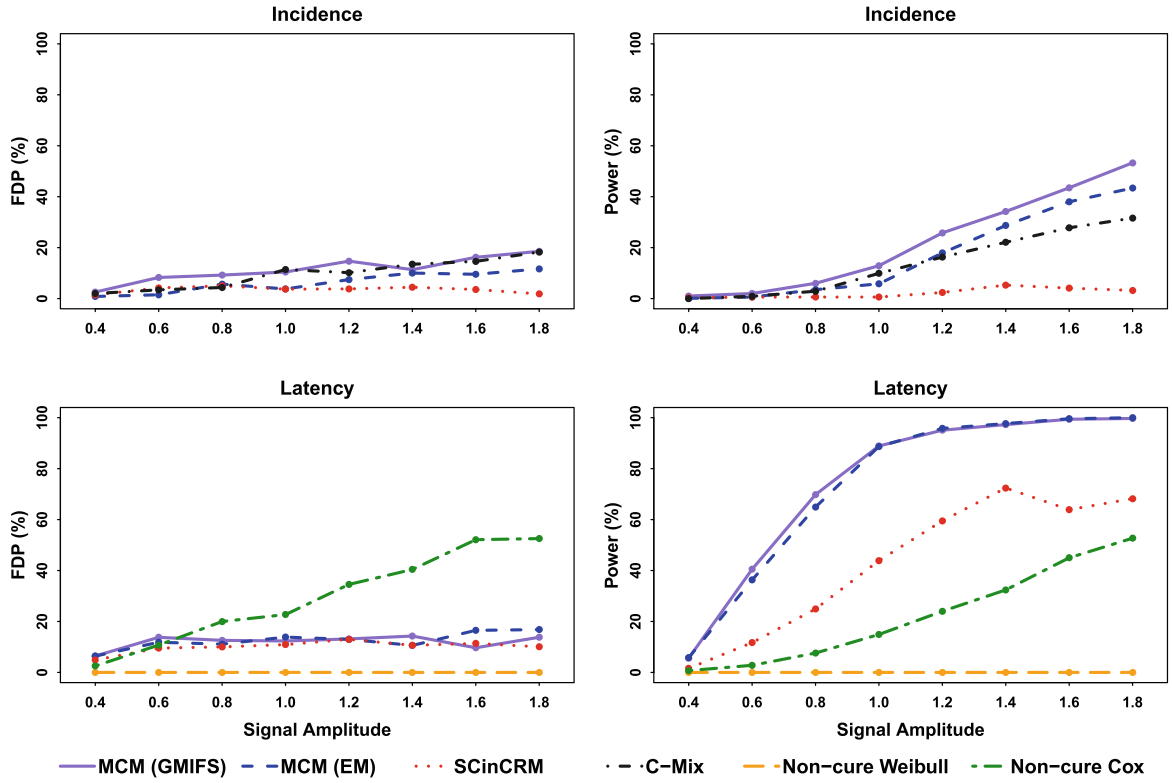

FIGURE A3 For a cure fraction of roughly 25%, false discovery proportion (left) and power (right) plotted by signal amplitude for incidence (top row) and latency (bottom row) for our mixture cure models (MCM) using the GMIFS and EM algorithm in comparison to SCinCRM, C-Mix, a non-cure Weibull, and a non-cure Cox model

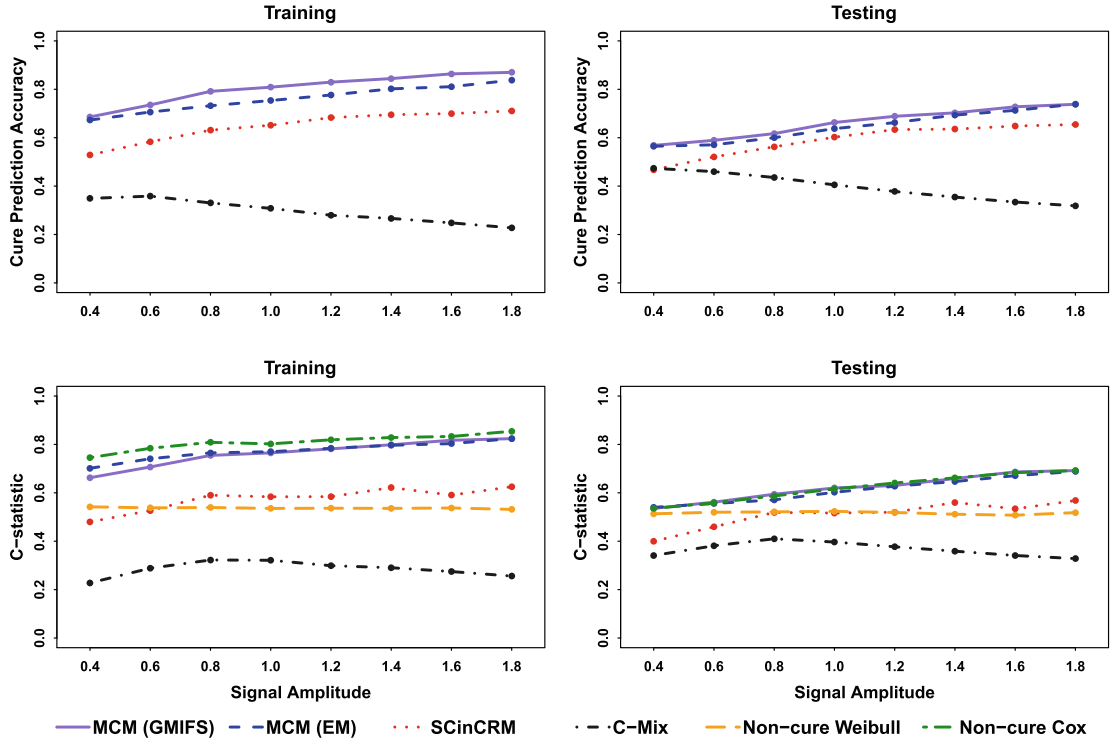

FIGURE A4 Results when the latency portion of the model was simulated from a generalized gamma distribution. Cure prediction accuracy (top row) and C-statistic (bottom row) plotted by signal amplitude for the training (right) and testing (left) datasets for our mixture cure models (MCM) using the GMIFS and EM algorithm in comparison to SCinCRM, C-Mix, a non-cure Weibull, and a non-cure Cox model. For our MCMs and SCinCRM, the C-statistic with cure status weighting was calculated, while the regular C-statistic was calculated for C-Mix and non-cure models

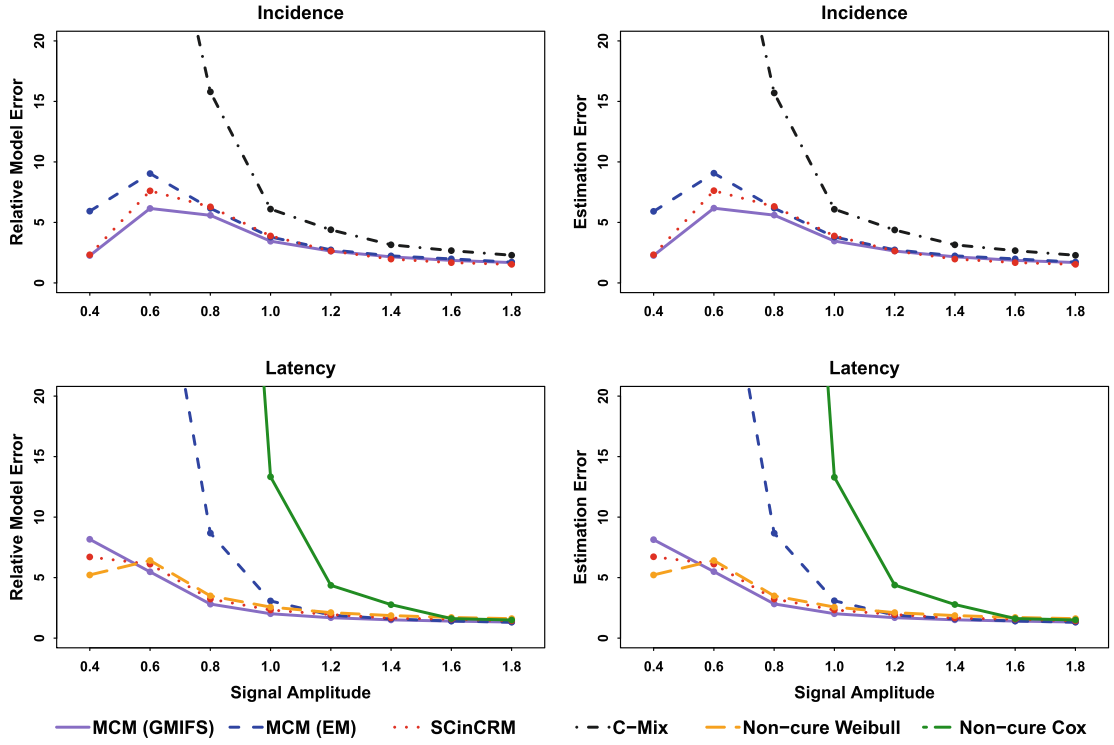

FIGURE A5 Results when the latency portion of the model was simulated from a generalized gamma distribution. Relative model error (left) and estimation error (right) plotted by signal amplitude for incidence (top row) and latency (bottom row) for our mixture cure models (MCM) using the GMIFS and EM algorithm in comparison to SCinCRM, C-Mix, a non-cure Weibull, and a non-cure Cox model

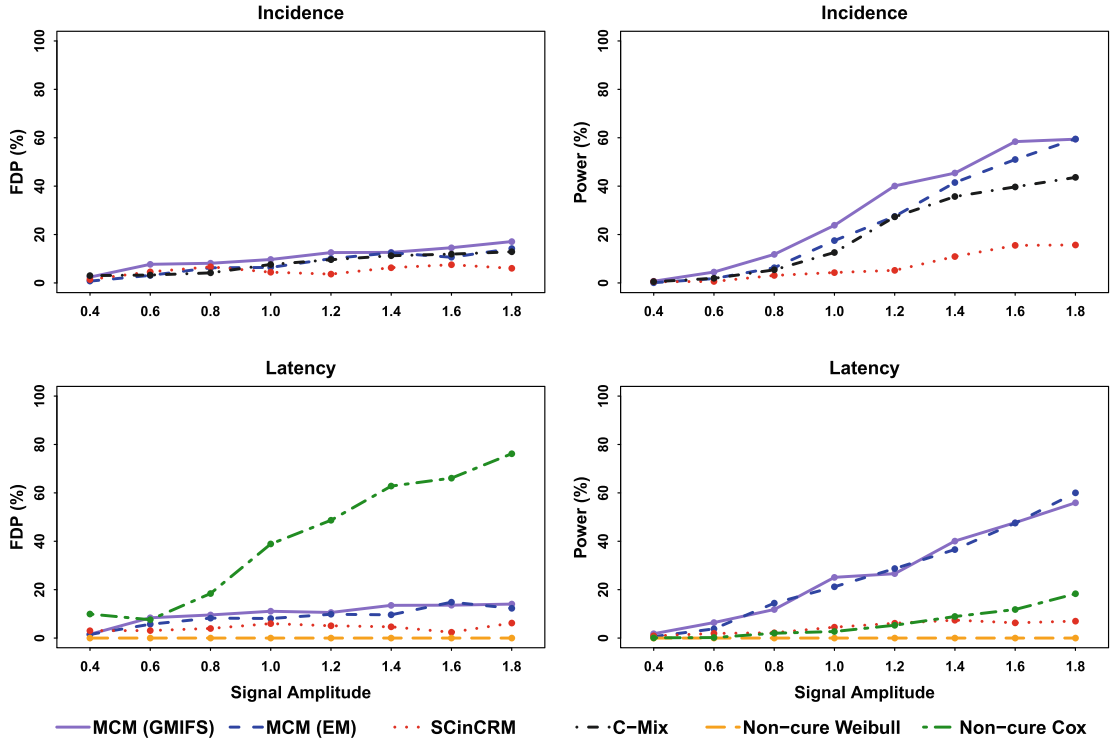

FIGURE A6 Results when the latency portion of the model was simulated from a generalized gamma distribution. False discovery proportion (left) and power (right) plotted by signal amplitude for incidence (top row) and latency (bottom row) for our mixture cure models (MCM) using the GMIFS and EM algorithm in comparison to SCinCRM, C-Mix, a non-cure Weibull, and a non-cure Cox model

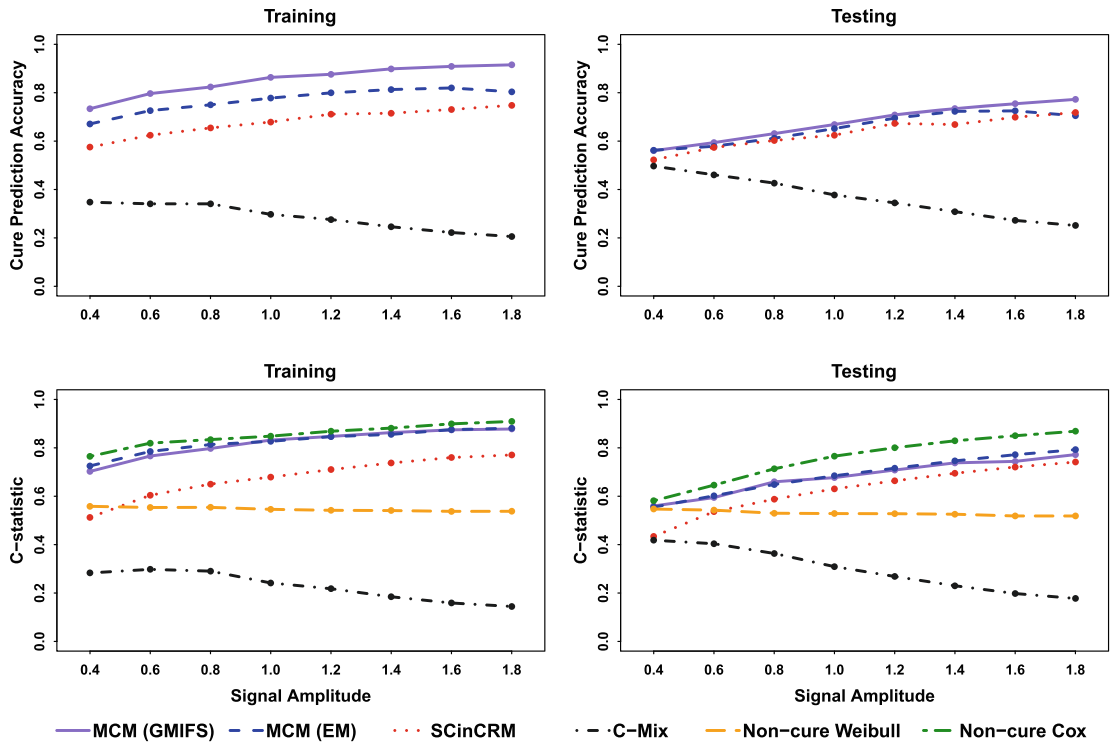

FIGURE A7 Simulation scenarios when the incidence coefficients and latency coefficients have the same signs. Cure prediction accuracy (top row) and C-statistic (bottom row) plotted by signal amplitude for the training (right) and testing (left) datasets for our mixture cure models (MCM) using the GMIFS and EM algorithm in comparison to SCinCRM, C-Mix, a non-cure Weibull, and a non-cure Cox model. For our MCMs and SCinCRM, the C-statistic with cure status weighting was calculated, while the regular C-statistic was calculated for C-Mix and non-cure models

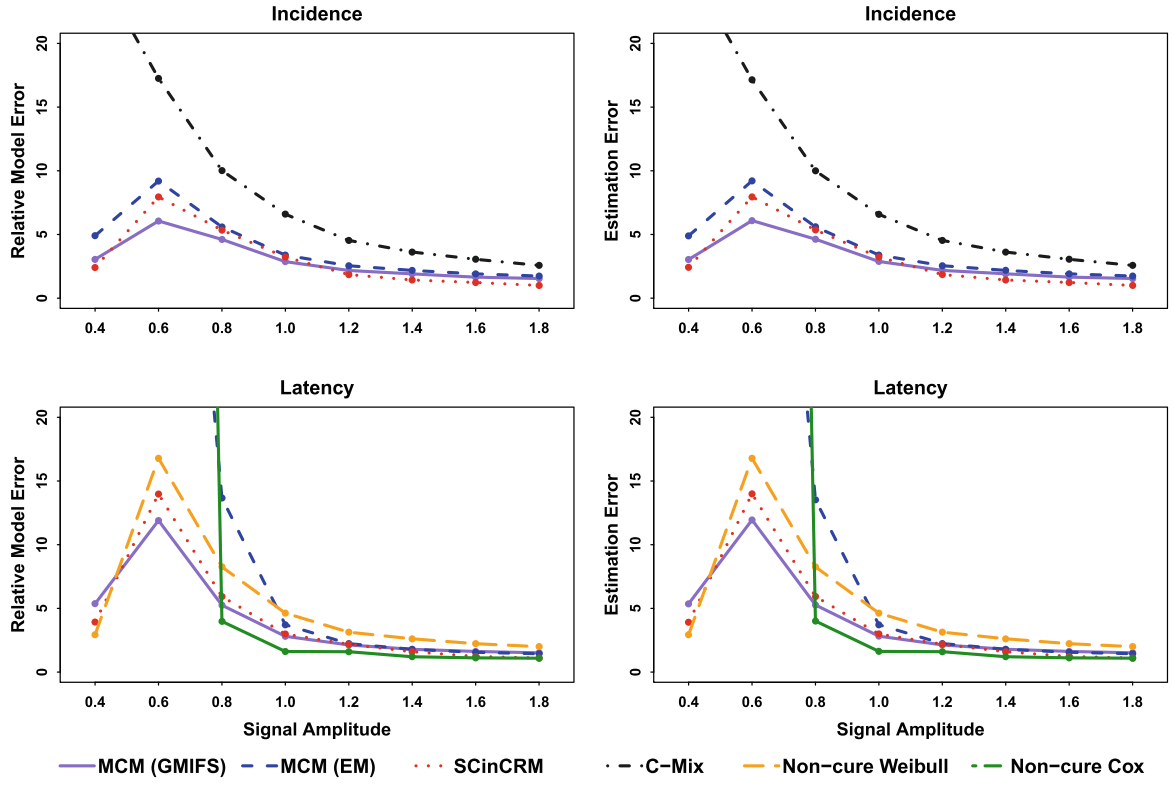

FIGURE A8 Simulation scenarios when the incidence coefficients and latency coefficients have the same signs. Relative model error (left) and estimation error (right) plotted by signal amplitude for incidence (top row) and latency (bottom row) for our mixture cure models (MCM) using the GMIFS and EM algorithm in comparison to SCinCRM, C-Mix, a non-cure Weibull, and a non-cure Cox model

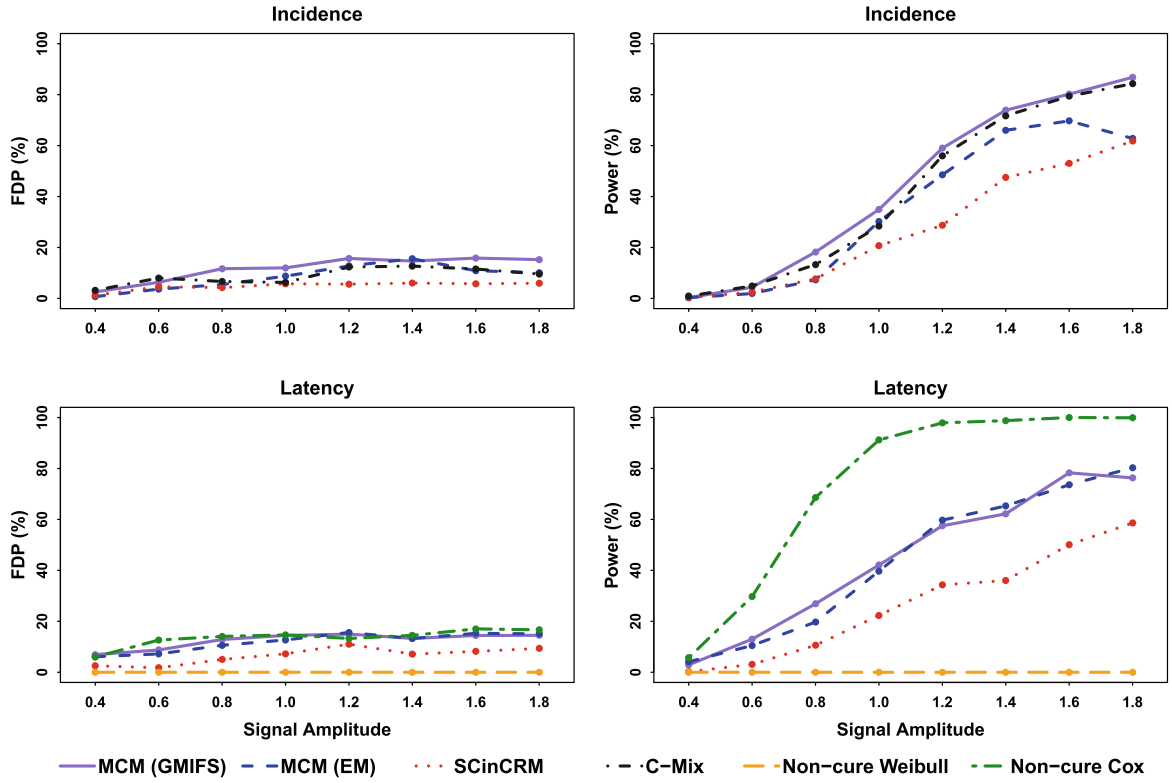

FIGURE A9 Simulation scenarios when the incidence coefficients and latency coefficients have the same signs. False discovery proportion (left) and power (right) plotted by signal amplitude for incidence (top row) and latency (bottom row) for our mixture cure models (MCM) using the GMIFS and EM algorithm in comparison to SCinCRM, C-Mix, a non-cure Weibull, and a non-cure Cox model

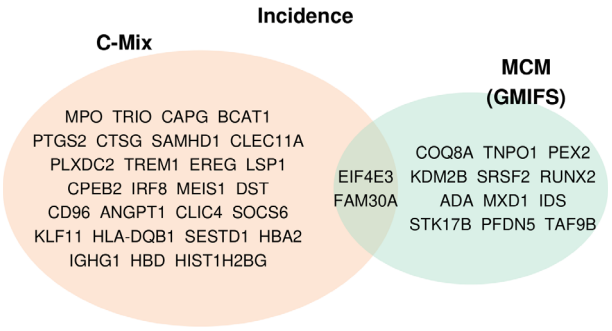

FIGURE A10 Venn diagram displaying the overlap with respect to genes identified as associated with incidence when using our mixture cure GMIFS algorithm and C-mix

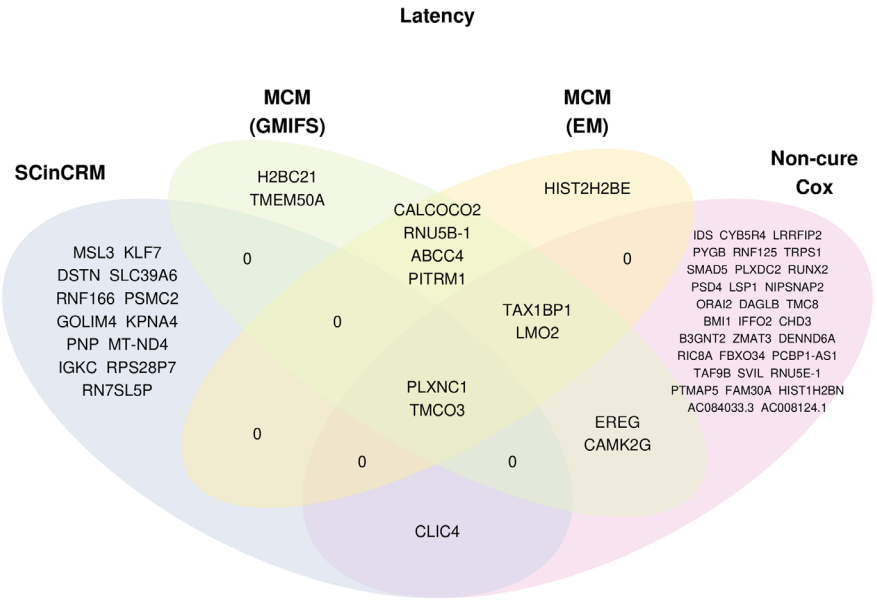

FIGURE A11 Venn diagram displaying the overlap with respect to genes identified as associated with latency when using our mixture cure GMIFS algorithm, mixture cure EM algorithm, SCinCRM, and a non-cure Cox model

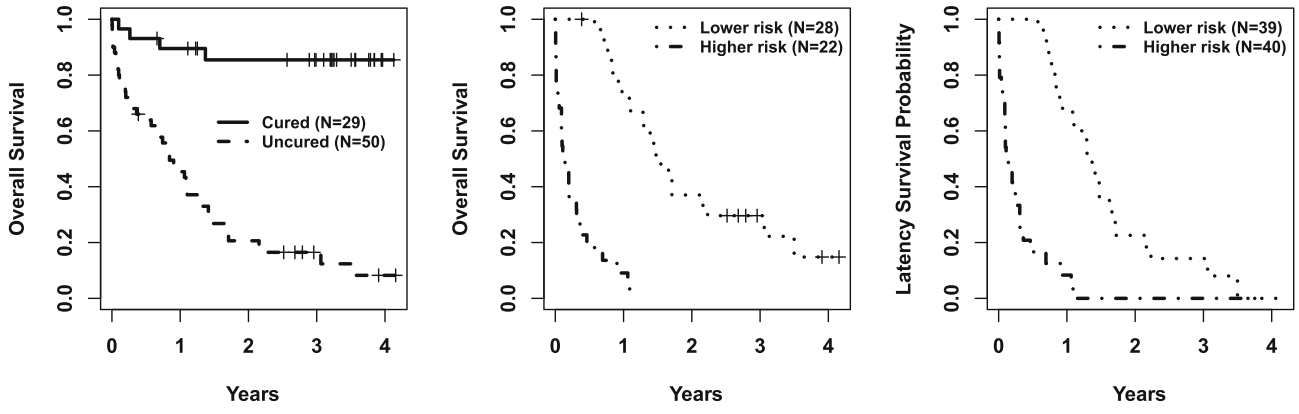

FIGURE A12 Using GSE12417 as a validation dataset and extracting Affymetrix probe sets that mapped to our selected genes from our training phase to assign subjects into groups based on our MCM, Kaplan-Meier estimates for predicted cured ($N = 29$) vs uncured ($N = 50$) groups (left panel), Kaplan-Meier estimates for subjects with lower risk ($N = 28$) vs higher risk ($N = 22$) among those that were predicted to be uncured (middle panel), and rescaled latency survival functions for all 79 subjects with lower ($N = 39$) vs higher ($N = 40$) risk (right panel)

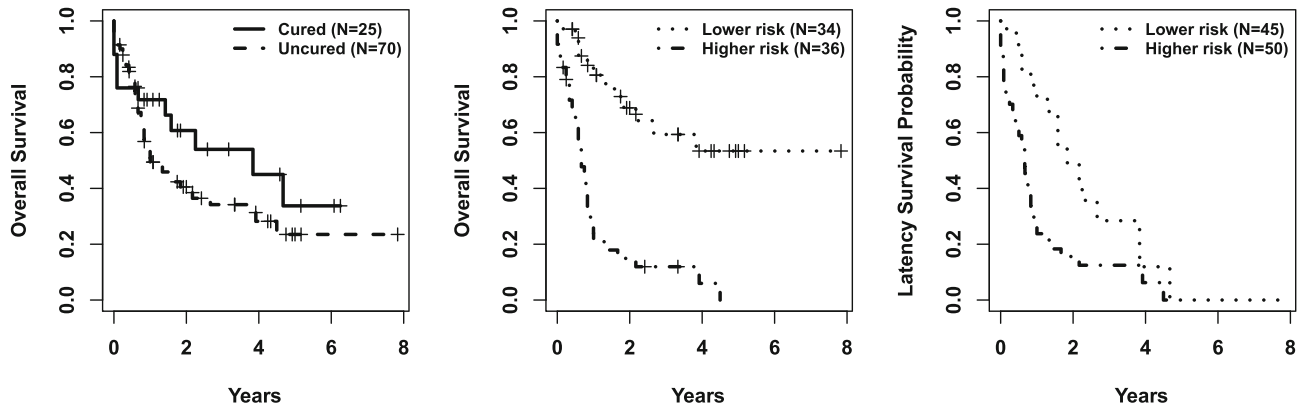

FIGURE A13 Using TCGA as a validation dataset, Kaplan-Meier estimates for predicted cured ($N = 25$) vs uncured ($N = 70$) groups (left panel), Kaplan-Meier estimates for subjects with lower risk ($N = 34$) vs higher risk ($N = 36$) among those that were predicted to be uncured (middle panel), and rescaled latency survival functions for all 95 subjects with lower ($N = 45$) vs higher ($N = 50$) risk (right panel)