Statistical Analysis and Data Mining: An ASA Data Science Journal

WILEY

RESEARCH ARTICLE

Causal Mediation Analysis With Latent Subgroups for Survival Model

Yerong Sun[1] | Yuejin Zhou[2] | Tao Hu[3] | Tiejun Tong[4] | WenWu Wang[1]

[1]School of Statistics and Data Science, Qufu Normal University, Qufu, Shandong, China | [2]School of Mathematics and Big Data, Anhui University of Science and Technology, Huainan, Anhui, China | [3]School of Mathematical Sciences, Capital Normal University, Beijing, China | [4]Department of Mathematics, Hong Kong Baptist University, Hong Kong, China

Correspondence: WenWu Wang (wangwenwu@qfnu.edu.cn)

Received: 15 March 2025 | **Revised:** 30 January 2026 | **Accepted:** 14 March 2026

Keywords: accelerated failure time model | causal mediation analysis | heterogeneous mediation effect | subgroup identification | survival outcome

ABSTRACT

Causal mediation analysis is an effective method for understanding the mechanism between the exposure and the outcome, often assuming that the mediation model is consistent for each individual in the target population. In practice, however, the natural indirect effect (NIE) may vary across individuals due to their distinct characteristics. As a result, the population can be partitioned into subgroups according to the varying sizes of the NIEs. Distinguishing subgroups within the study population enables the development of more precise and targeted treatment strategies. In this paper, we propose an identifiable mixture mediation model with latent subgroups for the survival data, where the outcome follows an accelerated failure time model and the mediator is Gaussian distributed. We further employ three information criteria including the AIC, BIC, and singular BIC (sBIC) to select the number of subgroups, followed by the expectation–maximization (EM) algorithm to estimate the model parameters and NIEs. Simulation study shows that the sBIC is the most robust and efficient criterion for selecting the number of subgroups; therefore, we recommend the sBIC-EM algorithm for practical use. Lastly, we apply our algorithm to the lung cancer data and discover two latent groups with opposing NIEs.

1 | Introduction

Causal mediation analysis is an effective method for understanding the mechanism between the exposure and the outcome. Since the seminal paper by Baron and Kenny [1], casual mediation analysis has undergone continuous development and refinement by researchers [2–8]. Apart from the theory, causal mediation analysis has also been widely applied in various fields including, for example, psychology [9], medicine [10], and economics [11]. In mediation analysis, it is of great interest to estimate the natural indirect effect (NIE) of the exposure on the outcome variable through the mediator. In addition, mediation models often assume a constant NIE for all individuals; however, this assumption may not hold due to their distinct characteristics. Because of this, it is desirable to partition the study population into subgroups that reflect meaningful heterogeneity. Meanwhile, it is equally important to select the correct subgroups based on the latent NIEs.

In survival analysis, the outcome variable is the time until an event occurs. Survival data typically exhibit characteristics such as non-negativity and skewed distribution. For example, the survival times of cancer patients are often right-skewed [12–14], making the normal distribution unsuitable for modeling such data. In contrast, the Weibull distribution has relatively simple distribution functions and can accommodate various degrees of skewness by adjusting its shape parameter. It is also the only parametric distribution that possesses both a proportional

hazards representation and an accelerated failure time (AFT) representation [14]. These favorable properties make the Weibull distribution a commonly used parametric model in survival analysis. Furthermore, unlike the normal cumulative distribution function $\Phi(\cdot)$ that lacks a closed-form expression and requires numerical methods for maximum likelihood estimation, the Weibull distribution has an explicit likelihood function and thus enables more efficient computation.

Tein and MacKinnon [15] conducted a simulation study to estimate the mediation effects with survival data. They showed that, for uncensored data, the assumption that the mediation effects calculated by the product of coefficients method and the difference in coefficients method are identical also applies to the survival analysis. Based on the counterfactual framework, Lange and Hansen [16] analyzed the direct and indirect effects for the time-to-event data using an additive hazard model. VanderWeele [17] discussed the different effect decompositions in additive hazard, proportional hazard and AFT models. Since then, causal mediation analysis has been increasingly applied to survival models to examine the impact of mediators on patient survival. To name a few, Huang et al. [18] used the mediation model to discover the gene expression variables that mediate the influence of micro-RNA on survival time. Hornung et al. [19] revealed the common causal mechanisms of RUNX1 point mutations and RUNX1/RUNX1T1 fusions influencing survival of patients with acute myeloid leukemia through causal mediation analysis. Cui et al. [20] proposed a mediation model for survival data when the mediator variables are high-dimensional. They further applied this model to 379,330 DNA methylation markers in The Cancer Genome Atlas (TCGA) lung cancer cohort, and found four methylation sites that mediate the smoking and overall survival among lung cancer patients.

Heterogeneity refers to the variation in effects across different populations or subgroups. For example, a drug's overall effect often depends on intermediate processes, such as its interaction with specific genes or proteins. Since individuals differ in these mediating pathways, the drug's efficacy can vary among them. Similarly, in behavioral or social studies, the same intervention may operate through different mediating processes across groups, resulting in varied outcomes. This variation in mediation effects across groups is referred to as heterogeneity of mediation effects. There are few works for considering the mediation effect with heterogeneity. Park and Kaplan [21] proposed a fully Bayesian approach to address the problem of heterogeneous treatment or mediation effects. Wang et al. [22] proposed the mixture mediation model that both outcome and mediator follow the normal distribution and used the standard expectation–maximization (EM) algorithm to estimate the heterogeneous mediation effects. However, survival data often involve censoring, which makes the analysis more complex but also more realistic. Therefore, analyzing the heterogeneity in censored survival data is crucial for advancing personalized treatment strategies and holds significant practical value.

The purpose of this paper is to select the number of subgroups and estimate their causal mediation effects for survival data. Based on the linear structure equation model (LSEM) framework, we assume that the mediator follows the normal distribution and the outcome follows the log (T)-Weibull distribution. We then propose a mixture model approach in two steps. The model parameters of the mediator and outcome can be estimated using the standard EM algorithm and the EM gradient algorithm [23, 24], respectively. Inside, the EM gradient algorithm, which replaces the maximum likelihood estimation in the M-step with the Newton method, has the global convergence properties similar to the standard EM algorithm. To better select the number of latent subgroups, we also employ three information criteria including the Akaike information criterion (AIC) by Akaike [25, 26], the Bayesian information criterion (BIC) by Schwarz [27], and the singular Bayesian information criterion (sBIC) by Drton and Plummer [28].

The remainder of this paper is organized as follows. In Section 2, we introduce the related definitions, notations, and assumptions. We further discuss the identification conditions for subgroup-specific mediation effects, and propose the mixture model with latent subgroups in LSEM. In Section 3, we apply the AIC, BIC and sBIC to select the number of subgroups, and then use the EM algorithm to estimate the parameters and NIEs. In Section 4, we conduct the simulation study and provide practical recommendations. In Section 5, we apply our new method to the TCGA lung cancer cohort study. Lastly, we conclude the paper in Section 6 with discussion and future work.

2 | Causal Mediation Analysis for Survival Model

2.1 | Symbols and Assumptions

For a random sample $\mathcal{D} = (\mathcal{D}_1, \ldots, \mathcal{D}_n)$, we assume that each individuals $\mathcal{D}_i = \left(A_i, M_i, Y_i, \boldsymbol{X}_i^T\right)^T$ $(i = 1, \ldots, n)$ is mutually independent, where Y_i is the continuous outcome, M_i is the continuous mediator, $\boldsymbol{X}_i$ is the p-dimensional measured pre-exposure confounders, and A_i is the exposure of interest ($A = a/a^*$). The survival model can be expressed as

$$M_i = \gamma_0 + \gamma_A A_i + \gamma_{\boldsymbol{X}}^T \boldsymbol{X}_i + \xi_i,$$
$$Y_i = \beta_0 + \beta_A A_i + \beta_M M_i + \boldsymbol{\beta}_{\boldsymbol{X}}^T \boldsymbol{X}_i + \varepsilon_i,$$

where $Y_i = \log(T_i)$, T_i is the survival time of each individual, ε_i is a random variable following an extreme value distribution, and ξ_i follows a normal distribution.

In causal analysis, the value a of an exposure variable A denotes the actual observed exposure level of a study subject in the real-world scenario, while the counterfactual value a^* represents the hypothetical alternative exposure level that the same subject would have been exposed to if the real exposure condition were replaced by a contrasting counterfactual scenario.

In the potential outcome framework [29, 30], we need to specify some necessary assumptions in order to calculate the causal effect on the unit level. First, the consistency assumption is that, for the ith unit, $M_i(a)$ denotes the potential value of the mediator with the treatment value $A = a$, and $Y_i(a, m)$ denotes the potential outcome with the treatment value $A_i = a$ and the

mediator value $M_i = m$. That is, $M_i = M_i(a)$ when $A_i = a$; and $Y_i = Y_i(a, m)$ when $A_i = a$ and $M_i = m$. Second, we assume that there are no unmeasured post-treatment confounders; that is, $M_i(a) \perp\!\!\!\perp A_i \mid \boldsymbol{X}_i, Y_i(a, m) \perp\!\!\!\perp M_i \mid (A_i, \boldsymbol{X}_i)$, and $Y_i(a, m) \perp\!\!\!\perp A_i \mid \boldsymbol{X}_i$. Third, the cross-world independence assumption states that there is no association between the counterfactual outcomes under different exposure levels; that is, $M_i(a) \perp\!\!\!\perp Y_i(a^*, m) \mid \boldsymbol{X}_i$. Consequently, with the NIE for unit i becomes identifiable, the detailed derivation process being provided in Appendix A,

$$\begin{aligned}\mathrm{NIE}(a, a^*) &= E\{\log(T(a, M(a)))\} \\ &\quad - E\{\log(T(a, M(a^*)))\} = \beta_M \gamma_A (a - a^*).\end{aligned}$$

2.2 | Linear Structural Equation Model With Latent Subgroup

In this section, we apply the linear structural equation model (LSEM) to account for the latent subgroup structure for the survival data. We assume that the population can be divided into K unknown subgroups, and π_k is the positive probability of the kth subgroup for $k = 1, \ldots, K$ with $\sum_{k=1}^{K} \pi_k = 1$. For each unit i, let G_i represent the membership of the latent subgroup that is distributed as $Pr(G_i = k) = \pi_k$ for $k = 1, \ldots, K$. Moreover, we assume that the mixing proportions are constant and do not depend on the other variables, yielding the linear structural equations with multiple subgroups as

$$\begin{aligned} M_i &= \gamma_{0,G_i} + \gamma_{A,G_i} A_i + \gamma_{\boldsymbol{X},G_i}^T \boldsymbol{X}_i + \varepsilon_{1i}, \\ Y_i &= \beta_{0,G_i} + \beta_{A,G_i} A_i + \beta_{M,G_i} M_i + \beta_{\boldsymbol{X},G_i}^T \boldsymbol{X}_i + \varepsilon_{2i}, \end{aligned} \tag{1}$$

where $\varepsilon_{1i} \sim N\left(0, \sigma_{1,G_i}^2\right)$, $\varepsilon_{2i} \sim \frac{1}{\sigma_{2,G_i}} \exp\left(\frac{y}{\sigma_{2,G_i}} - \exp\left(\frac{y}{\sigma_{2,G_i}}\right)\right)$, and $\varepsilon_{1i} \perp\!\!\!\perp \varepsilon_{2i}$ for $1 \le i \le n$. In other words, the mixture model in (1) consists of two regression models, one is the linear model and the other is the AFT model. Also for illustration, Figure 1 provides the graphics of the causal mediation analysis with two latent subgroups. And lastly, the conditional probability density functions of M_i and Y_i are given as

$$f_{G_i}(M_i | A_i, \boldsymbol{X}_i) = \phi\left(\frac{M_i - \gamma_{0,G_i} - \gamma_{A,G_i} A_i - \gamma_{\boldsymbol{X},G_i}^T \boldsymbol{X}_i}{\sigma_{1,G_i}}\right),$$

$$f_{G_i}(Y_i | A_i, M_i, \boldsymbol{X}_i) = \mathrm{LW}\left(\frac{Y_i - \beta_{0,G_i} - \beta_{A,G_i} A_i - \beta_{M,G_i} M_i - \beta_{\boldsymbol{X},G_i}^T \boldsymbol{X}_i}{\sigma_{2,G_i}}\right),$$

where $\phi(\cdot)$ is the standard normal distribution and $\mathrm{LW}(\cdot)$ is the probability density functions of the log-Weibull (also known as Gumbel) distribution, $\mathrm{LW}(\mu, \sigma) = \frac{1}{\sigma} \exp\left(\frac{y-\mu}{\sigma} - \exp\left(\frac{y-\mu}{\sigma}\right)\right)$.

2.3 | Model With Right-Censored Data

In reality, we may not be able to observe the occurrence of all events due to reasons such as limited follow-up time or patients dropping out midway. If we do not know when the event occurred, the data will be censored. We denote that T is the survival time, C is the censored time, and δ is a status variable where $\delta = 1$ ($\delta = 0$) denotes that we (do not) observe the survival time.

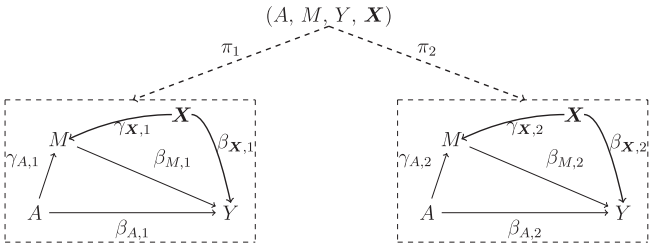

FIGURE 1 | Causal directed acyclic graph (DAG) diagrams of heterogeneous mediation effect model with $K = 2$ subgroups, where π_1 and π_2 represent the proportions of the two subgroups. In addition, $\gamma_{A,1}$ is the effect of A on M, $\gamma_{\boldsymbol{X},1}$ is the effect of $\boldsymbol{X}$ on M, $\beta_{A,1}$ is the direct effect of A on Y, $\beta_{M,1}$ is the effect of M on Y, $\beta_{\boldsymbol{X},1}$ is the effect of $\boldsymbol{X}$ on Y, and $\gamma_{A,2}$, $\gamma_{\boldsymbol{X},2}$, $\beta_{A,2}$, $\beta_{M,2}$, $\beta_{\boldsymbol{X},2}$ are defined similarly in subgroup 2.

The outcome variable is $Y = \min\{T, C\}$ with $Pr(T, \delta = 1) = f(t)$ and $Pr(T, \delta = 0) = S(t)$, where $S(T) = Pr(T > t)$ is the survival function.

Taking the most common right-censoring scenario as an example, for the ith observation, the likelihood function is given as $L_i = [f(y_i)]^{\delta_i} [S(y_i)]^{1-\delta_i}$. Given the n independent observations $\{(Y_i, \delta_i)\}_{i=1}^n$, the likelihood function is $L = \prod_{i=1}^n [f(y_i)]^{\delta_i} [S(y_i)]^{1-\delta_i}$. Following Kalbfleisch and Prentice [31], the likelihood function for the AFT model is

$$\begin{aligned} L &= \prod_{i=1}^{n} [f_Y(y_i)]^{\delta_i} [S_Y(y_i)]^{1-\delta_i} \\ &= \prod_{i=1}^{n} \left[\frac{1}{\sigma} \exp\left(\frac{y_i - \mu}{\sigma} - \exp\frac{y_i - \mu}{\sigma}\right)\right]^{\delta_i} \left[\exp\left(-\exp\frac{y_i - \mu}{\sigma}\right)\right]^{1-\delta_i}. \end{aligned}$$

Therefore, for the right-censored survival data, we have

$$f_{G_i}(Y_i | A_i, M_i, \boldsymbol{X}_i) = \mathrm{LW}^*\left(\frac{Y_i - \beta_{0,G_i} - \beta_{A,G_i} A_i - \beta_{M,G_i} M_i - \beta_{\boldsymbol{X},G_i}^T \boldsymbol{X}_i}{\sigma_{2,G_i}}\right),$$

where $\mathrm{LW}^*(\mu, \sigma) = \left[\frac{1}{\sigma} \exp\left(\frac{y-\mu}{\sigma} - \exp\left(\frac{y-\mu}{\sigma}\right)\right)\right]^{\delta} \left[\exp\left(-\exp\left(\frac{y-\mu}{\sigma}\right)\right)\right]^{1-\delta}$.

2.4 | Effect and Model Identification

Because of the mediation effect heterogeneity, we need to identify the average NIE for each of the K subgroups. The group-specific sequential ignorability assumptions [32] are as follows:

$$\{Y_i(a, m), M_i(a^*)\} \perp\!\!\!\perp A_i \mid G_i = k, \boldsymbol{X}_i = \boldsymbol{x}_i,$$

$$Y_i(a^*, m) \perp\!\!\!\perp M_i(a) \mid G_i = k, A_i = a, \boldsymbol{X}_i = \boldsymbol{x}_i,$$

where $G_i = k$ indicates that the ith individual comes from the kth subgroup. In addition, in the presence of censored data, the product method remains valid under standard assumptions, while the difference method may produce biased estimates [15, 33]. Thus, we can compute the average NIE for the kth subgroup by the formula $\mathrm{NIE}_k = \beta_{M,k} \gamma_{A,k}$.

Denote $Y = (Y_1, \ldots, Y_n)^T$, $T = (T_1, \ldots, T_n)^T$, $A = (A_1, \ldots, A_n)^T$, $M = (M_1, \ldots, M_n)^T$, and $\boldsymbol{X} = (\boldsymbol{X_1}, \ldots, \boldsymbol{X_n})^T$. Based on the conditional probability density functions of M_i

and Y_i, the conditional probability density function of the study population given $(A, \boldsymbol{X})$ is given as

$$f(Y, M|A, \boldsymbol{X}, \boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k f_k\left(Y, M|A, \boldsymbol{X}, \boldsymbol{\alpha_k}\right), \tag{2}$$

where $f_k\left(Y_i, M_i|A_i, \boldsymbol{X}_i, \boldsymbol{\alpha_k}\right) = f_k\left(M_i|A_i, \boldsymbol{X}_i\right) f_k\left(Y_i|A_i, M_i, \boldsymbol{X}_i\right)$ is the conditional probability density function of the k th subgroup, $\boldsymbol{\alpha}_k^T = \left(\boldsymbol{\gamma}_k^T, \boldsymbol{\beta}_k^T\right)$ with $\boldsymbol{\gamma}_k^T = \left(\gamma_{0,k}, \gamma_{A,k}, \boldsymbol{\gamma}_{\boldsymbol{X},k}^T, \sigma_{1,k}\right)$ and $\boldsymbol{\beta}_k^T = \left(\beta_{0,k}, \beta_{A,k}, \beta_{M,k}, \boldsymbol{\beta}_{\boldsymbol{X},k}^T, \sigma_{2,k}\right)$, $\boldsymbol{\pi} = \left(\pi_1, \ldots, \pi_K\right)^T$, and $\boldsymbol{\theta} = \left(\boldsymbol{\pi}, \boldsymbol{\alpha}_1^T, \ldots, \boldsymbol{\alpha}_K^T\right)^T$. Lastly, by Huang and Yao [34] and Wang et al. [22], we have the following model identifiability property.

Proposition 1. *Model (1) is identifiable up to a permutation of the group labels if the sequence of the parameter vector $\{\boldsymbol{\alpha}_i\}_{i=1}^K$ is different from each other.*

3 | Estimation Methods

3.1 | Estimation via the EM Algorithm

Given the number of subgroups K, we apply the EM algorithm [35] to estimate the parameter vector $\boldsymbol{\theta}$ in model (2). The EM algorithm is one of the most influential algorithms in statistics, which estimates the parameters by introducing the latent variable. Nevertheless, the standard EM algorithm may not be applicable to the AFT model due to the frequent failure of the maximum likelihood method. To overcome the problem, we further apply the Newton method in Lange [23] to replace the maximum likelihood method in the M-step for the AFT model. Let also the binary latent variables $z_{i,k} = I\left[G_i = k\right]$ with $E\left[z_{i,k}\right] = \pi_k$ for $k = 1, \ldots, K$, which satisfies $z_{i,k} \in \{0, 1\}$ and $\sum_{k=1}^K z_{i,k} = 1$ for $i = 1, \ldots, n$. The complete data log-likelihood is then given as

$$\ell_c = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{i,k} \log\left[\pi_k f_k\left(Y_i, M_i|A_i, \boldsymbol{X}_i; \boldsymbol{\alpha}_k\right)\right]. \tag{3}$$

3.1.1 | E-Step

The E-step of the EM algorithm is to compute the expectation of (3) for obtaining the Q-function. More specifically, by letting m be the current number of the iteration and $\boldsymbol{\theta}^{(m)}$ be the mth iterated estimation, we have

$$\begin{aligned} Q\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{(m)}\right) &= \mathbb{E}\left[\ell_c(\theta)|\mathcal{D}; \boldsymbol{\theta}^{(m)}\right] \\ &= \sum_{i=1}^{n} \sum_{k=1}^{K} \mathbb{E}\left[z_{i,k}|\mathcal{D}; \boldsymbol{\theta}^{(m)}\right] \log\left[\pi_k f_k\left(Y_i, M_i|A_i, \boldsymbol{X}_i; \boldsymbol{\alpha}_k\right)\right] \\ &= \sum_{i=1}^{n} \sum_{k=1}^{K} z_{i,k}^{(m+1)} \log\left[\pi_k f_k\left(Y_i, M_i|A_i, \boldsymbol{X}_i; \boldsymbol{\alpha}_k\right)\right], \end{aligned} \tag{4}$$

where

$$z_{i,k}^{(m+1)} = \frac{\pi_k^{(m)} f_k\left(Y_i, M_i|A_i, \boldsymbol{X}_i; \boldsymbol{\alpha}_k^{(m)}\right)}{\sum_{k=1}^K \pi_k^{(m)} f_k\left(Y_i, M_i|A_i, \boldsymbol{X}_i; \boldsymbol{\alpha}_k^{(m)}\right)} \tag{5}$$

is the posterior probability of the ith observation from the kth subgroup. Consequently, we can assign each individual to its most likely subgroup by comparing the size of this value in each subgroup.

3.1.2 | M-Step

The M-step of the EM algorithm is to get the estimates of parameters by maximizing the Q-function in (4). Let

$$Q\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{(m)}\right) = Q_1\left(\boldsymbol{\pi}; \boldsymbol{\theta}^{(m)}\right) + \sum_{k=1}^{K} Q_{2,k}\left(\boldsymbol{\alpha}_k; \boldsymbol{\theta}^{(m)}\right),$$

where

$$Q_1\left(\boldsymbol{\pi}; \boldsymbol{\theta}^{(m)}\right) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{i,k}^{(m+1)} \log \pi_k, \tag{6}$$

$$Q_{2,k}\left(\boldsymbol{\alpha}_k; \boldsymbol{\theta}^{(m)}\right) = \sum_{k=1}^{K} z_{i,k}^{(m+1)} \log\left[f_k\left(Y_i, M_i|A_i, \boldsymbol{X}_i; \boldsymbol{\alpha}_k^{(m)}\right)\right]. \tag{7}$$

Moreover, we can decompose (7) as follows:

$$\begin{aligned} Q_{2,k}\left(\boldsymbol{\alpha}_k; \boldsymbol{\theta}^{(m)}\right) &= \sum_{i=1}^{n} z_{i,k}^{(m+1)} \log\left[f_k\left(Y_i, M_i|A_i, \boldsymbol{X}_i; \boldsymbol{\alpha}_k^{(m)}\right)\right] \\ &= f_1^{(m)} + f_2^{(m)}, \end{aligned}$$

where

$$f_1^{(m)} = \sum_{i=1}^{n} z_{i,k}^{(m+1)} \log\left[\phi\left(M_i|A_i, \boldsymbol{X}_i; \boldsymbol{\gamma}_k^{(m)}\right)\right], \tag{8}$$

$$f_2^{(m)} = \sum_{i=1}^{n} z_{i,k}^{(m+1)} \log\left[\mathrm{LW}^*\left(Y_i|A_i, M_i, \boldsymbol{X}_i; \boldsymbol{\beta}_k^{(m)}\right)\right]. \tag{9}$$

Recall that the MLE for $\boldsymbol{\beta}_k$ cannot be obtained in closed form due to the complexity of the log-likelihood under the log-Weibull distribution. As an alternative, we use the Newton–Raphson method to numerically maximize the likelihood [23]. For the estimation of $\boldsymbol{\gamma}_k$, however, we continue to use the maximum likelihood method. Consequently, it yields that

$$\left(\gamma_{0,k}^{(m+1)}, \gamma_{A,k}^{(m+1)}, \boldsymbol{\gamma}_{\boldsymbol{X},k}^{(m+1)}\right)^T = \left(\tilde{\boldsymbol{X}}^T \boldsymbol{P}_k^{(m+1)} \tilde{\boldsymbol{X}}\right)^{-1} \tilde{\boldsymbol{X}}^T \boldsymbol{P}_k^{(m+1)} \boldsymbol{M},$$

$$\left(\sigma_{1,k}^2\right)^{(m+1)} = \frac{\sum_{i=1}^n z_{i,k}^{(m+1)} \left(M_i - \gamma_{0,k}^{(m)} - \gamma_{A,k}^{(m)} A_i - \boldsymbol{X}_i^T \boldsymbol{\gamma}_{\boldsymbol{X},k}^{(m)}\right)}{\sum_{i=1}^k z_{i,k}^{(m)}}$$

$$\pi_k^{(m+1)} = \frac{\sum_{i=1}^k z_{i,k}^{(m+1)}}{\sum_{k=1}^K \sum_{i=1}^k z_{i,k}^{(m)}},$$

$$\boldsymbol{\beta}_{M,k}^{(m+1)} = \boldsymbol{\beta}_{M,k}^{(m)} - \left[\nabla^2 f_2\left(\boldsymbol{\beta}_{M,k}^{(m)}\right)\right]^{-1} \nabla f_2\left(\boldsymbol{\beta}_{M,k}^{(m)}\right),$$

where $\mathbf{1} = (1, \ldots, 1)^T$, $\tilde{\boldsymbol{X}}^T = (\mathbf{1}, A, \boldsymbol{X})$, $\boldsymbol{P}_k^{(m+1)} = \mathrm{diag}\left(z_{1,k}^{(m+1)}, \ldots, z_{n,k}^{(m+1)}\right)$, and $\boldsymbol{\beta}_{M,k}^{(t)}$ is the parameters for the kth subgroup in the tth iteration. In addition, ∇f and $\nabla^2 f$ represent the gradient and the Hessian matrix with respect to the parameters of f, respectively.

3.2 | Determining the Number of Latent Subgroups by Information Criterion

Since the subgroup selection can be considered as a model selection problem, we thus apply three commonly used information criteria, including AIC, BIC, and sBIC, to select the number of subgroups. For AIC and BIC, we have:

$$\widehat{K}_{\mathrm{AIC}} = \arg\min_K \mathrm{AIC}(K) = \arg\min_K \{-2\log L_K + 2(2p+7)\},$$
$$\widehat{K}_{\mathrm{BIC}} = \arg\min_K \mathrm{BIC}(K) = \arg\min_K \{-2\log L_K + (2p+7)\log n\},$$

where L_K is the likelihood function when K is the number of subgroups, p is the dimension of the pretreatment confounders $\boldsymbol{X}$, and n is the size of the study population.

Next, to solve the singularity problem of the Fisher information matrix for the mixture models, we further apply the sBIC to select the optimal number of the latent subgroups [28]. More specifically, the sBIC is defined as follows:

$$\mathrm{sBIC}(K) = \log\big(L'(K)\big), \quad K \in I = \{1, 2, \ldots, K_{max}\},$$

where $\{L'(K) | K \in I\}$ is the positive and unique solution to the equation

$$\sum_{1\le J\le K} \Big\{ L'(K) - L'_{K,J} \Big\} L'(J) = 0, \quad K \in I,$$

in which $L'_{K,J} = \prod_{i=1}^n \Big[\sum_{k=1}^K \widehat{\pi}_k f_k\big(Y_i, M_i | A_i, \boldsymbol{X}_i; \widehat{\boldsymbol{\alpha}}_k\big)\Big]^{n^{-\lambda_{K,J}}}$ and $(\widehat{\pi}_k, \widehat{\boldsymbol{\alpha}}_k)$ are the EM estimates when the number of subgroups is fixed at K. Following Drton and Plummer [28], we let $\lambda_{K,J} = \frac{1}{2}\{K(7+2p) + J - 1\}$, which then yields the solution of $L'(K)$ as

$$L'(K) = \frac{1}{2}\left(-B_K + \sqrt{B_K^2 + 4C_K}\right), \tag{10}$$

where $B_K = L'_{K,K} + \sum_{J<K} L'(J)$ and $C_K = \sum_{J<K} L'_{K,J} L'(J)$. Finally, by the sBIC, the optimal K can be specified as

$$\widehat{K} = \arg\max_K \mathrm{sBIC}(K).$$

Moreover, by Drton and Plummer [28] and Wang et al. [22], we have $Pr\Big(\widehat{K} = K_0\Big) \to 1$ as $n \to \infty$ under some regular conditions, where K_0 is the true number of the latent subgroups.

3.3 | The sBIC-EM Algorithm

To conclude, each of the information criteria in Section 3.2 can be used to select the number of subgroups. Yet to explore which criterion performs the best, we will conduct a simulation study in Section 4 and show that the sBIC is among the most effective towards the target. We thus recommend using the sBIC-EM algorithm in practical applications.

To describe in more detail, the proposed sBIC-EM algorithm can be summarized into two steps. First, for a candidate set of K, we apply the EM algorithm to estimate the parameters and compute the value of the sBIC. Second, we choose the optimal K by maximizing the sBIC, and then assign each individual to one subgroup based on the largest posterior probability in (5). Besides, after the selection of K, the nonparametric bootstrap [36] will also be employed to obtain the standard errors and confidence intervals for testing the NIEs. We summarize the sBIC-EM algorithm as Algorithm (1).

ALGORITHM 1 | The sBIC-EM algorithm.

Input: $\mathcal{D}_i$, the ith observation sample; n, the sample size; K_{max}, a prespecified upper bound for K; δ, the tolerance level; p, the dimension of $\boldsymbol{X}_i$.

Output: $\hat{K}$, the selected number of subgroups; $(\hat{\pi}_k, \hat{\boldsymbol{\alpha}}_k)$, the estimated parameters of the kth subgroup model.

for K=*1, 2, …, K_{max}* **do**
  initialize $Z^{(0)}$ and $\pi^{(0)}$;
  while $(|\pi_k^{(m+1)} - \pi_k^{(m)}| \ge \delta$ and $||\boldsymbol{\alpha}_k^{(m+1)} - \boldsymbol{\alpha}_k^{(m)}||_1 \ge \delta)$ **do**
    (E-step) compute $z_{i,k}^{(m+1)}$ according to (5);
    (M-step) update $\pi_k^{(m+1)}$ by maximizing (6); for $\boldsymbol{\alpha}_k^{(m+1)} = (\boldsymbol{\gamma}_{A,k}^{(m+1)}, \boldsymbol{\beta}_{M,k}^{(m+1)})$, update $\boldsymbol{\gamma}_{A,k}^{(m+1)}$ by maximizing (8) and $\boldsymbol{\beta}_{M,k}^{(m+1)}$ by maximizing (9) for $1 \le k \le K$;
  end
  Return the final EM estimates, denoted as $\{(\hat{\pi}_{k,K}, \hat{\boldsymbol{\alpha}}_{k,K}) \mid 1 \le k \le K\}$;
  Compute $L'(K)$ based on (10);
  Return sBIC(K);
end
Return $\hat{K} = \arg\max_{1\le K\le K_{max}} \mathrm{sBIC}(K)$ and $\{(\hat{\pi}_{k,\hat{K}}, \hat{\boldsymbol{\alpha}}_{k,\hat{K}}) \mid 1 \le k \le K\}$.

TABLE 1 | The model parameters for the four cases in the simulation study.

Case	k	π_k	$(\gamma_{0,k}, \gamma_{A,k})$	$(\beta_{0,k}, \beta_{A,k}, \beta_{M,k})$	$(\sigma_{1,k}, \sigma_{2,k})$	$\text{NIE}_k = \gamma_{A,k}\beta_{M,k}$
I	1	0.5	(1, 0.5)	(0.6, −1, 0.5)	(1.2, 1.2)	0.25
	2	0.5	(−1.2, 0.5)	(−0.7, 1.8, 1)	(0.8, 1.5)	0.5
II	1	0.4	(0, 0.5)	(0.5, 0.5, −1)	(1, 1)	−0.5
	2	0.6	(0, 0.5)	(0.5, 1, 1.5)	(0.8, 1.2)	0.75
III	1	0.3	(1, 0.5)	(−0.8, −0.7, −0.8)	(0.6, 0.6)	−0.4
	2	0.4	(0.2, 0.2)	(0.3, 0.5, 0)	(0.4, 0.7)	0
	3	0.3	(0.5, 0.8)	(1.2, 0.7, 1)	(0.5, 0.6)	0.8
IV	1	0.3	(1, 0.5)	(0, −0.8, 0.4)	(0.4, 0.5)	0.2
	2	0.4	(0.2, 0)	(−0.3, 0.4, 0)	(0.3, 0.6)	0
	3	0.3	(0.5, 0.8)	(1, 0.8, 0.5)	(0.6,0.3)	0.4

4 | Simulation Study

4.1 | Simulation Setup

There are two main tasks when classifying and selecting the potential subgroups of the research population: one is to determine the number of subgroups, and the other is to calculate the NIEs of subgroups. In this section, we conduct simulations to investigate whether the proposed method can effectively accomplish the above objectives. More specifically, we will identify the best criterion for selecting the correct number of subgroups, and meanwhile assess the accuracy of the EM algorithm in estimating the NIEs.

For $k = 1, \ldots, K_0$, we generate $\left(A_i, M_i, Y_i, X_i\right)$ for $n\sum_{l=1}^{k-1}\pi_l < i \leq n\sum_{l=1}^{k}\pi_l$ with $A_i \sim N(0,1)$, $X_i \sim N(1,4)$ and

$$M_i = \gamma_{0,k} + \gamma_{A,k}A_i + \gamma_{X,k}X_i + \varepsilon_{1i},$$
$$Y_i = \beta_{0,k} + \beta_{A,k}A_i + \beta_{M,k}M_i + \beta_{X,k}X_i + \varepsilon_{2i},$$

where $\varepsilon_{1i} \sim N\left(0, \sigma_{1,k}\right)$, $\varepsilon_{2i} \sim \frac{1}{\sigma_{2,k}}\exp\left(\frac{y}{\sigma_{2,k}} - \exp\left(\frac{y}{\sigma_{2,k}}\right)\right)$, and that they are independent of each other. Our simulations consider four different cases under two scenarios (uncensored data and 15% right censored rate). In Case I, the model consists of $K_0 = 2$ subgroups with equal probability, in which both subgroups have positive but different NIEs (0.25 vs. 0.5). In Case II, there are also $K_0 = 2$ subgroups, but with unequal probabilities and opposite direction for the two NIEs. In Case III, there are $K_0 = 3$ subgroups with unequal probabilities, and the NIEs in subgroups are negative, null and positive, respectively. In Case IV, there are also $K_0 = 3$ subgroups with unequal probabilities, but the NIEs in subgroups are positive, null and positive, respectively. For more specific settings of the four cases, see Table 1.

4.2 | Simulation Results

We generate 1000 datasets with $n = 100$, 300, or 500 for each case. For each $K \in \left\{1, \ldots, K_0 + 1\right\}$, we apply the EM algorithm to estimate the regression parameters, calculate the NIE estimate, and select the optimal subgroup number. For uncensored data, we apply the AIC, BIC and sBIC to select the optimal number of subgroups K_{opt}, and report the proportions of their correctly selected number, that is, $K_{opt} = K_0$, in Figure 2. Based on the simulation results, it is evident that the AIC does not provide a comparable performance for the first two cases with $K_0 = 2$, and the BIC fails to work in most cases. In contrast, the sBIC can effectively select the number of subgroups across all four cases, and its performance gets even better as the sample size n increases. Taken together, we recommend the sBIC for selecting the number of subgroups.

With the correctly selected number of subgroups, we obtain the parameter estimates including $\widehat{\pi}_k$, $\widehat{\gamma}_{A,k}$, $\widehat{\beta}_{M,k}$, and $\widehat{\gamma}_{A,k}\widehat{\beta}_{M,k}$. To assess estimation accuracy, we report their averaged bias and standard error (within parentheses) in Table 2. As a whole, most parameter estimates are stable with relatively small bias and standard error. As the sample size increases, the estimation accuracy becomes further improved. To conclude, the proposed sBIC-EM algorithm provides a high accuracy for the mediation effect estimation.

Lastly, for censored data, the simulation results are similar to those obtained from uncensored data, with most parameter estimates remaining stable with relatively small bias and standard errors. The main difference is that the estimates from uncensored data exhibit smaller bias and variability than those derived from censored data. To save space, we have presented the results for subgroup selection and parameter estimation in Appendix B.

5 | Application

Cancer is the most common type of malignant tumor consisting of three stages: initiation, promotion, and progression. Its onset is a complex process with multiple factors and steps, which are closely related to genetic factors, smoking, occupational exposure, environmental pollution, and unreasonable diets. With the development of industrialization and the aggravation of environmental pollution, the incidence rate of lung cancer has increased rapidly and has become the first cause of cancer death in the world. Nowadays, smoking may be a significant risk factor for inducing lung cancer. Long-term smoking may lead to the occurrence of lung squamous cell carcinoma.

The Cancer Genome Atlas (TCGA) was launched by the National Cancer Institute and the National Institute of Human Genome Research in 2006. This project collects and collates relevant clinical data of various cancers, including lung cancer, breast

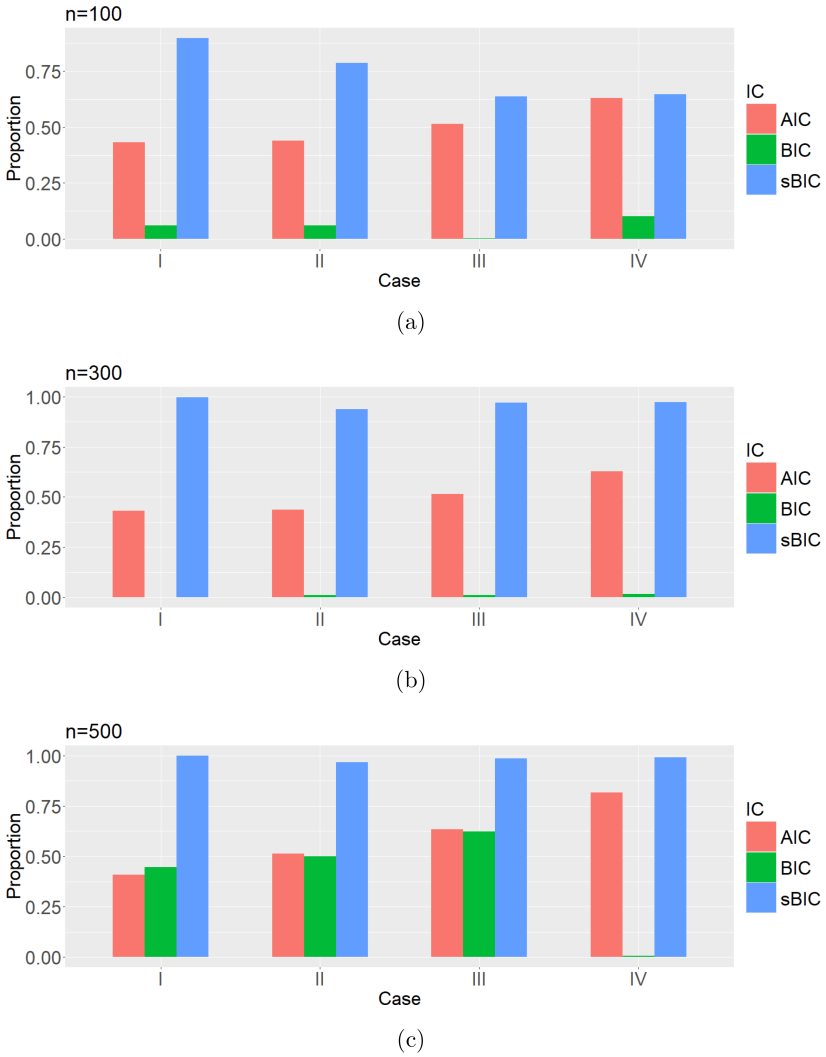

FIGURE 2 | The proportions of correctly selected number based on 1000 simulations for the four cases without censored time, where (a) to (c) represent the sample sizes from 100 to 500, respectively.

cancer, and other cancers. This project has greatly improved the level of the understanding, research, and prevention of cancer for cancer researchers. Among them, the study of the TCGA lung cancer cohort involves DNA methylation data (907 samples measured by the Illumina Infinium Human Method450 platform), phenotype data (1299 samples), and survival data (1145 samples) of lung squamous cell carcinoma and lung adenocarcinoma, including patient smoking status, survival time, gender, age, DNA methylation CpG site β values, and other related information, which helps to further study the relevant content of lung cancer. In order to determine whether smoking affects DNA methylation leading to lung cancer, Cui et al. [20] analyzed the complete data from 833 patients in the TCGA lung cancer cohort, including squamous cell carcinoma and adenocarcinoma, and found that multiple DNA methylation CpG sites (cg19757631, cg08636115, cg05147638, and cg24720672) have an impact on the causal pathway of smoking status on the survival risk of lung cancer patients. Among the four DNA methylation CpG sites mentioned above, the previous study [2] has found that cg19757631 is a related mediator in the causal pathway from smoking to lung cancer and has a negative impact on patients. Thus, we choose the CpG site cg19757631 as the mediator in the following mediation analysis.

We apply our proposed sBIC-EM algorithm to reanalyze these cancer data for investigating the potential number of subgroups and estimating the NIE of each subgroup. Specially, we focus on the 343 dead patients without censored data. This results in the mediation model as follows:

$$M = \gamma_0 + \gamma_A A + \gamma_{\boldsymbol{X}}^T \boldsymbol{X} + \varepsilon_1,$$
$$Y = \beta_0 + \beta_A A + \beta_M M + \beta_{\boldsymbol{X}}^T \boldsymbol{X} + \varepsilon_2,$$

where the exposure A is the smoking status of patients (smoking = 1, nosmoking = 0), the covariates $\boldsymbol{X}$ are the patient's gender and age, the outcome Y is the survival time in years, and the mediator is the CpG site cg19757631.

As shown in Figure 3, the number of latent subgroups for the CpG site cg19757631 is 2 using the sBIC method. We set $K = 2$ and apply the nonparametric bootstrap to construct the interval estimates for the NIE. The resampling is done with 1000 times replacement for confidence interval. The two latent subgroups have different NIE sizes: −0.14 (95% CI [−3.80,0.36]) and 0.31

TABLE 2 | The averaged bias and standard error (within parentheses) for the estimates of the model parameters based on 1000 simulations by the sBIC-EM algorithm with no censored time.

Case	n	k	bias$(\widehat{\pi}_k)$	bias$(\widehat{\gamma}_{A,k})$	bias$\left(\widehat{\beta}_{M,k}\right)$	bias$\left(\widehat{\gamma}_{A,k}\widehat{\beta}_{M,k}\right)$
I	100	1	−0.010 (0.087)	−0.068 (0.277)	−0.003 (0.559)	−0.092 (0.544)
		2	0.010 (0.087)	0.072 (0.213)	0.029 (0.523)	0.066 (0.321)
	300	1	0.000 (0.038)	−0.017 (0.117)	0.001 (0.124)	−0.010 (0.078)
		2	0.000 (0.038)	0.008 (0.089)	−0.009 (0.206)	0.000 (0.121)
	500	1	0.000 (0.022)	−0.004 (0.088)	−0.004 (0.080)	−0.004 (0.060)
		2	0.000 (0.022)	0.005 (0.064)	−0.004 (0.148)	0.002 (0.093)
II	100	1	0.016 (0.104)	0.002 (0.220)	−0.052 (0.563)	−0.049 (0.413)
		2	−0.016 (0.104)	0.007 (0.146)	−0.092 (0.404)	0.084 (0.303)
	300	1	0.002 (0.035)	−0.009 (0.106)	0.006 (0.160)	0.011 (0.132)
		2	−0.002 (0.035)	0.000 (0.073)	−0.008 (0.164)	−0.003 (0.127)
	500	1	0.001 (0.023)	0.005 (0.080)	0.000 (0.085)	−0.004 (0.089)
		2	−0.001 (0.023)	0.001 (0.052)	0.000 (0.101)	0.002 (0.092)
III	100	1	0.041 (0.141)	−0.036 (0.226)	−0.126 (1.067)	−0.201 (0.299)
		2	−0.021 (0.127)	0.083 (0.227)	0.163 (0.921)	0.080 (0.354)
		3	−0.020 (0.087)	−0.064 (0.188)	0.228 (2.600)	0.303 (2.720)
	300	1	0.024 (0.090)	−0.039 (0.089)	−0.026 (0.224)	0.026 (0.091)
		2	−0.017 (0.072)	0.043 (0.160)	0.082 (0.338)	0.044 (0.174)
		3	−0.007 (0.049)	−0.003 (0.100)	0.039 (0.400)	0.023 (0.283)
	500	1	0.017 (0.086)	−0.016 (0.081)	−0.086 (0.925)	−0.038 (0.661)
		2	−0.011 (0.071)	0.033 (0.133)	0.035 (0.281)	0.027 (0.141)
		3	−0.006 (0.037)	−0.015 (0.071)	0.057 (0.671)	0.001 (0.127)
IV	100	1	0.044 (0.112)	−0.024 (0.261)	−0.044 (0.654)	−0.028 (0.209)
		2	−0.029 (0.112)	0.199 (0.285)	−0.193 (0.678)	−0.081 (0.185)
		3	−0.015 (0.073)	−0.064 (0.225)	−0.047 (0.499)	0.029 (0.320)
	300	1	0.006 (0.042)	−0.027 (0.145)	0.003 (0.322)	0.003 (0.116)
		2	−0.007 (0.069)	0.052 (0.179)	−0.047 (0.334)	−0.018 (0.114)
		3	0.000 (0.042)	−0.024 (0.125)	−0.004 (0.117)	−0.015 (0.110)
	500	1	0.005 (0.047)	−0.005 (0.052)	0.000 (0.156)	0.000 (0.020)
		2	0.001 (0.032)	0.005 (0.066)	−0.002 (0.252)	0.000 (0.078)
		3	−0.006 (0.027)	−0.020 (0.074)	0.018 (0.101)	−0.002 (0.056)

(95% CI [−0.07,20.87]), with the mixing proportions 74% and 26%, respectively. Our study has the similar findings with Bakulski et al. [2] and Cui et al. [20]: in the causal pathway through which smoking affects patients' survival, gene site cg19757631 as the mediator exerts a negative impact. Nevertheless, we further note that the gene site as the mediator does not necessarily play an entirely negative role in all populations. We expect that this new result can help researchers to conduct reasonable prevention based on the characteristics of different populations in the future, as well as provide targeted treatment for different lung cancer patients.

We also analyze the censored data in Appendix C. The selected number of latent subgroups remains 2, which is the same as that for uncensored data. The two corresponding NIEs are −0.14 (95% CI [−3.92, −0.02]) and 0.18 (95% CI [−0.12,22.16]), with the mixing proportions 72% and 28%, respectively. This demonstrates that including censored data does not alter the overall conclusions.

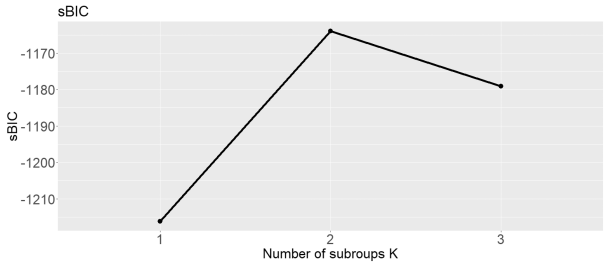

FIGURE 3 | The sBIC for the DNA methylation CpG sites cg19757631 which is maximized at $K = 2$.

6 | Discussion and Conclusion

In medical research, the outcomes of interest for survival data include, for example, death and disease recurrence. It is often desired to understand the causal mechanism of the occurrence of these unpleasant events and find the mediators that mediate the

treatment and the outcome. Note, however, that the same treatment may have different or even opposite causal effects on different units due to their unique characteristics. This is known as a heterogeneity problem for mediation analysis. Distinguishing different subgroups according to the mediation effect heterogeneity can avoid producing inconsistent effects and help units select the optimal strategy and treatment.

In this paper, we apply a mixture model to study the target population which consists of two regression models: one is the Gaussian model and the other is the AFT model. We first prove that this mixture model with potential subgroups is identifiable, and then develop the sBIC-EM algorithm to select the number of subgroups and estimate the heterogeneous causal mediation effects. More specifically, our new algorithm consists of two main steps. On one hand, considering that the number of subgroups is unknown, we compare three information criteria AIC, BIC and sBIC and conclude that the sBIC is the best to select the latent subgroups. On the other hand, to estimate the causal mediation of each subgroup, we use the classic EM algorithm to estimate the parameters of the Gaussian model and then use the EM gradient algorithm to estimate the parameters of the AFT model. Simulation studies show that our newly proposed sBIC-EM algorithm works robustly, and the effectiveness of our algorithm is confirmed. In addition, the real application to the TCGA data also shows that our proposed method can effectively distinguish the latent subgroups in survival data and estimate the corresponding causal mediation effect in the absence of censored data. This is of great significance in the medical studies for better targeted treatment of cancer patients with different characteristics.

Our study is general and can be readily applied to other models. For illustration, we take the AFT model as an example,

$$\log(T) = \beta_0 + \beta_A A + \beta_M M + \boldsymbol{\beta}_{\boldsymbol{X}}^T \boldsymbol{X} + \varepsilon,$$

where T can follow a logistic distribution, a normal distribution, or other distributions. When T follows the logistic distribution, by letting $Y = \log(T)$, we have

$$f_Y(y) = \frac{1}{\sigma}\exp\left(\frac{y-\mu}{\sigma}\right)/\left(1+\exp\left(\frac{y-\mu}{\sigma}\right)\right)^2,$$
$$S_Y(y) = 1/\left(1+\exp\left(\frac{y-\mu}{\sigma}\right)\right).$$

The log-logistic model is the only parametric model with both a proportional odds (PO) and an AFT representation. This enables it not only to serve as an alternative to the Weibull distribution in AFT models but also to play a unique role in PO models.

Another extension is from the univariate case to the multivariate case with multiple treatments and mediators [20, 37]. Taking the two-dimensional case as an example, by letting $\boldsymbol{A}_i = (A_{1i}, A_{2i})$ and $\boldsymbol{M}_i = (M_{1i}, M_{2i})$ with a total of K latent subgroups, the regression equations are as follows:

$$M_{1i} = \gamma_{1,G_i} + \boldsymbol{\gamma}_{\boldsymbol{A},1,G_i}^T \boldsymbol{A}_i + \boldsymbol{\gamma}_{\boldsymbol{X},1,G_i}^T \boldsymbol{X}_i + \varepsilon_{1i},$$
$$M_{2i} = \gamma_{2,G_i} + \boldsymbol{\gamma}_{\boldsymbol{A},2,G_i}^T \boldsymbol{A}_i + \boldsymbol{\gamma}_{\boldsymbol{X},2,G_i}^T \boldsymbol{X}_i + \varepsilon_{2i},$$
$$Y_i = \beta_{0,G_i} + \boldsymbol{\beta}_{\boldsymbol{A},G_i}^T \boldsymbol{A}_i + \boldsymbol{\beta}_{\boldsymbol{M},G_i}^T \boldsymbol{M}_i + \boldsymbol{\beta}_{\boldsymbol{X},G_i}^T \boldsymbol{X}_i + \varepsilon_{3i},$$

where $\varepsilon_{1i} \sim N\left(0, \sigma_{1,G_i}^2\right)$, $\varepsilon_{2i} \sim N\left(0, \sigma_{2,G_i}^2\right)$, and $\varepsilon_{3i} \sim \frac{1}{\sigma_{3,G_i}}\exp\left(\frac{y}{\sigma_{3,G_i}} - \exp\left(\frac{y}{\sigma_{3,G_i}}\right)\right)$ for $i = 1, \ldots, n$ and $k = 1, \ldots, K$. The conditional density function of $(\boldsymbol{M}, Y)$ given $\boldsymbol{A}$ is

$$f(\boldsymbol{M}, Y|\boldsymbol{A}) = \sum_{k=1}^{K} \pi_k f_k(\boldsymbol{M}, Y|\boldsymbol{A}),$$

where $\sum_{k=1}^{K} \pi_k = 1$ and

$$f_k(\boldsymbol{M}, Y|\boldsymbol{A}) = \phi\left(\frac{M_{1i} - \gamma_{1,G_i} - \boldsymbol{\gamma}_{\boldsymbol{A},1,G_i}^T \boldsymbol{A}_i - \gamma_{\boldsymbol{X},1,G_i}^T \boldsymbol{X}_i}{\sigma_{1,G_i}}\right)$$
$$\phi\left(\frac{M_{2i} - \gamma_{2,G_i} - \boldsymbol{\gamma}_{\boldsymbol{A},2,G_i}^T \boldsymbol{A}_i - \gamma_{\boldsymbol{X},2,G_i}^T \boldsymbol{X}_i}{\sigma_{2,G_i}}\right)$$
$$\mathrm{LW}\left(\frac{Y_i - \beta_{0,G_i} - \boldsymbol{\beta}_{\boldsymbol{A},G_i}^T \boldsymbol{A}_i - \boldsymbol{\beta}_{\boldsymbol{M},G_i}^T \boldsymbol{M}_i - \beta_{\boldsymbol{X},G_i}^T \boldsymbol{X}_i}{\sigma_{3,G_i}}\right).$$

These extensions may warrant further research.

Author Contributions

Yerong Sun: conceptualization, data analysis, methodology, writing – original draft. **Yuejin Zhou:** software, data analysis, visualization. **Tao Hu:** validation, writing – review and editing. **Tiejun Tong:** validation, writing – review and editing. **WenWu Wang:** supervision, conceptualization, software, project administration, funding acquisition.

Acknowledgments

The authors sincerely thank the Editor, the Associate Editor, and two anonymous reviewers for their constructive comments that have led to a significant improvement of the paper.

Funding

Wang's work was supported by the Shandong Provincial Natural Science Foundation (ZR2024MA058), the Humanities and Social Sciences Fund of the Ministry of Education of China (25YJA910006) and the National Natural Science Foundation of China (12071248). Zhou's work was supported by the Grant of State Key Laboratory of Mining Response and Disaster Prevention and Control in Deep Coal Mines (SKLMRDPC22KF03). Hu's work was supported by Beijing Outstanding Young Scientist Program (JWZQ20240101027). Tong's work was supported by the General Research Fund of Hong Kong (HKBU12300123) and the Initiation Grant for Faculty Niche Research Areas of Hong Kong Baptist University (RC-FNRA-IG/23-24/SCI/03).

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data that support the findings of this study are openly available in The Cancer Genome Atlas at https://portal.gdc.cancer.gov/.

References

1. R. M. Baron and D. A. Kenny, "The Moderator–Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic, and Statistical Considerations," *Journal of Personality and Social Psychology* 51, no. 6 (1986): 1173–1182.

2. K. M. Bakulski, J. Dou, N. Lin, S. London, and J. Colacino, "DNA Methylation Signature of Smoking in Lung Cancer Is Enriched for Exposure Signatures in Newborn and Adult Blood," *Scientific Reports* 9, no. 1 (2019): 4576.

3. J. Berrevoets, K. Kacprzyk, Z. Qian, and M. van der Schaar, "Causal Deep Learning," ArXiv Preprint arXiv:2303.02186 (2023).

4. K. Imai, L. Keele, and D. Yamamoto, "Identification, Inference, and Sensitivity Analysis for Causal Mediation Effects," *Statistical Science* 25, no. 1 (2010): 51–71.

5. J. Pearl, "Direct and Indirect Effects. In: Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence. pp. 411–420" (2001).

6. J. M. Robins and S. Greenland, "Identifiability and Exchangeability for Direct and Indirect Effects," *Epidemiology* 3, no. 2 (1992): 143–155.

7. T. J. VanderWeele, *Explanation in Causal Inference: Methods for Mediation and Interaction* (Oxford University Press, 2015).

8. T. J. VanderWeele and S. Vansteelandt, "Conceptual Issues Concerning Mediation, Interventions and Composition," *Statistics and Its Interface* 2, no. 4 (2009): 457–468.

9. T. Q. Nguyen, I. Schmid, and E. A. Stuart, "Clarifying Causal Mediation Analysis for the Applied Researcher: Defining Effects Based on What We Want to Learn," *Psychological Methods* 26, no. 2 (2021): 255–271.

10. J. J. Rijnhart, S. J. Lamp, M. J. Valente, D. P. MacKinnon, J. W. Twisk, and M. W. Heymans, "Mediation Analysis Methods Used in Observational Research: A Scoping Review and Recommendations," *BMC Medical Research Methodology* 21 (2021): 226.

11. V. Celli, "Causal Mediation Analysis in Economics: Objectives, Assumptions, Models," *Journal of Economic Surveys* 36, no. 1 (2022): 214–234.

12. J. Hamfjord, T. Å. Myklebust, I. K. Larsen, et al., "Survival Trends of Right-and Left-Sided Colon Cancer Across Four Decades: A Norwegian Population-Based Study," *Cancer Epidemiology, Biomarkers & Prevention* 31, no. 2 (2022): 342–351.

13. G. Kirkebøen, ""The Median Isn't the Message": How to Communicate the Uncertainties of Survival Prognoses to Cancer Patients in a Realistic and Hopeful Way," *European Journal of Cancer Care* 28, no. 4 (2019): e13056.

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

15. J. Y. Tein and D. P. MacKinnon, "Estimating Mediated Effects With Survival Data," in *New Developments in Psychometrics: Proceedings of the International Meeting of the Psychometric Society IMPS2001. Osaka, Japan, July 15–19, 2001* (Springer, 2003), 405–412.

16. T. Lange and J. V. Hansen, "Direct and Indirect Effects in a Survival Context," *Epidemiology* 22, no. 4 (2011): 575–581.

17. T. J. VanderWeele, "Causal Mediation Analysis With Survival Data," *Epidemiology* 22, no. 4 (2011): 582–585.

18. Y. T. Huang, T. Hsu, K. T. Kelsey, and C. L. Lin, "Integrative Analysis of Micro-RNA, Gene Expression, and Survival of Glioblastoma Multiforme," *Genetic Epidemiology* 39, no. 2 (2015): 134–143.

19. R. Hornung, V. Jurinovic, A. M. Batcha, et al., "Mediation Analysis Reveals Common Mechanisms of RUNX1 Point Mutations and RUNX1/RUNX1T1 Fusions Influencing Survival of Patients With Acute Myeloid Leukemia," *Scientific Reports* 8, no. 1 (2018): 11293.

20. Y. Cui, C. Luo, L. Luo, and Z. Yu, "High-Dimensional Mediation Analysis Based on Additive Hazards Model for Survival Data," *Frontiers in Genetics* 12 (2021): 771932.

21. S. Park and D. Kaplan, "Bayesian Causal Mediation Analysis for Group Randomized Designs With Homogeneous and Heterogeneous Effects: Simulation and Case Study," *Multivariate Behavioral Research* 50, no. 3 (2015): 316–333.

22. W. Wang, J. Xu, J. Schwartz, A. Baccarelli, and Z. Liu, "Causal Mediation Analysis With Latent Subgroups," *Statistics in Medicine* 40, no. 25 (2021): 5628–5641.

23. K. Lange, "A Gradient Algorithm Locally Equivalent to the EM Algorithm," *Journal of the Royal Statistical Society, Series B* 57, no. 2 (1995): 425–437.

24. G. J. McLachlan and T. Krishnan, *The EM Algorithm and Extensions* (John Wiley & Sons, 2007).

25. H. Akaike, "Information Theory and an Extension of the Maximum Likelihood Principle," in *Second International Symposium on Information Theory* (Akademia Kiadom, 1973), 267–281.

26. H. Akaike, "A New Look at the Statistical Model Identification," *IEEE Transactions on Automatic Control* 19, no. 6 (1974): 716–723.

27. G. Schwarz, "Estimating the Dimension of a Model," *Annals of Statistics* 6, no. 2 (1978): 461–464.

28. M. Drton and M. Plummer, "A Bayesian Information Criterion for Singular Models," *Journal of the Royal Statistical Society. Series B, Statistical Methodology* 79, no. 6 (2017): 323–380.

29. D. B. Rubin, "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology* 66, no. 5 (1974): 688–701.

30. D. B. Rubin, "Causal Inference Using Potential Outcomes: Design, Modeling, Decisions," *Journal of the American Statistical Association* 100, no. 469 (2005): 322–331.

31. J. D. Kalbfleisch and R. L. Prentice, *The Statistical Analysis of Failure Time Data* (John Wiley & Sons, 2002).

32. C. Kim, M. Daniels, Y. Li, K. Milbury, and L. Cohen, "A Bayesian Semiparametric Latent Variable Approach to Causal Mediation," *Statistics in Medicine* 37, no. 7 (2018): 1149–1161.

33. I. R. Fulcher, E. J. T. Tchetgen, and P. L. Williams, "Mediation Analysis for Censored Survival Data Under an Accelerated Failure Time Model," *Epidemiology* 28, no. 5 (2017): 660–666.

34. M. Huang and W. Yao, "Mixture of Regression Models With Varying Mixing Proportions: A Semiparametric Approach," *Journal of the American Statistical Association* 107, no. 498 (2012): 711–724.

35. A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum Likelihood From Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society, Series B* 39, no. 1 (1977): 1–22.

36. D. P. MacKinnon, C. M. Lockwood, and J. Williams, "Confidence Limits for the Indirect Effect: Distribution of the Product and Resampling Methods," *Multivariate Behavioral Research* 39, no. 1 (2004): 99–128.

37. Z. Shao, T. Wang, M. Zhang, Z. Jiang, S. Huang, and P. Zeng, "IUSMMT: Survival Mediation Analysis of Gene Expression With Multiple DNA Methylation Exposures and Its Application to Cancers of TCGA," *PLoS Computational Biology* 17, no. 8 (2021): e1009250.

Appendix A

The Identifiability of the NIE

For the ith unit, $M_i(a)$ denotes the potential value of the mediator with the treatment value $A = a$, and $Y_i(a, m)$ denotes the potential outcome with the treatment value $A_i = a$ and the mediator value $M_i = m$. For exposure A, a is the exposure value and a^* is the counterfactual exposure value ($a \neq a^*$). The group-specific sequential ignorability assumptions [32] are as follows,

$$\left\{Y_i(a, m), M_i(a^*)\right\} \perp\!\!\!\perp A_i \mid \boldsymbol{G}_i = k, \boldsymbol{X}_i = \boldsymbol{x}_i,$$

$$Y_i(a^*, m) \perp\!\!\!\perp M_i(a) \mid \boldsymbol{G}_i = k, A_i = a, \boldsymbol{X}_i = \boldsymbol{x}_i,$$

where $\boldsymbol{G}_i = k$ indicates that the ith individual comes from the kth subgroup. Under these assumptions, the average NIE for each of the K

subgroups is identified. We consider the general mediation model as follows,

$$M_i = \gamma_{0,k} + \gamma_k A_i + \boldsymbol{\gamma}_{\boldsymbol{X},\boldsymbol{k}}^T \boldsymbol{X}_i + \varepsilon_{1i},$$
$$Y_i = \log(T_i) = \beta_{0,k} + \beta_{A,k} A_i + \beta_{M,k} M_i + \boldsymbol{\beta}_{\boldsymbol{X},\boldsymbol{k}}^T \boldsymbol{X}_i + \frac{1}{\sigma_{2,k}} \varepsilon_{2i},$$

where $\varepsilon_{1i} \sim N\left(0, \sigma_{1,k}^2\right)$, $\varepsilon_{2i} \sim exp\left(\frac{y}{\sigma_{2,k}} - exp\left(\frac{y}{\sigma_{2,k}}\right)\right)$, $\varepsilon_{1i} \perp \varepsilon_{2i}$, $P\left(G_i = k\right) = \pi_k$ with $k = 1, \ldots, K$ and $1 \le i \le n$. Then

$$\begin{aligned} E\left\{\log(T(a, M(a^*)))\right\} &= \sum_m E\{\log(T|(a,m))\} f(m|a^*, z) \\ &= \beta_{0,k} + \beta_{A,k} a + \beta_{M,k} E(M|a^*) + \boldsymbol{\beta}_{\boldsymbol{X},\boldsymbol{k}}^T \boldsymbol{x}_i + \frac{1}{\sigma_{2,k}} \varepsilon_{2i} \\ &= \beta_{0,k} + \beta_{A,k} a + \beta_{M,k} \gamma_{A,k} a^* + + \boldsymbol{\beta}_{\boldsymbol{X},\boldsymbol{k}}^T \boldsymbol{x}_i + \frac{1}{\sigma_{2,k}} \varepsilon_{2i}. \end{aligned}$$

which gives the following result:

$$\begin{aligned} \text{NIE}_k &= E\{\log(T(a, M(a)))\} - E\left\{\log(T(a, M(a^*)))\right\} \\ &= \beta_{M,k} \gamma_{A,k} (a - a^*). \end{aligned}$$

For binary A with $a = 1$ and $a^* = 0$, we have $\text{NIE}_k = \beta_{M,k} \gamma_{A,k}$.

Appendix B

The Additive Simulation Results

We provide the additive simulation results on censored data designed in Section 4.1. Figure B1 shows the proportions of the correctly selected subgroup number using the AIC, BIC and sBIC. Overall, the selection proportion for censored data is lower than those of uncensored data. Similar to the conclusion of uncensored data in Section 4.2, we have the similar findings: the AIC does not provide a comparable performance for the first two cases with $K_0 = 2$, the BIC fails to work in most cases, only the sBIC can effectively select the number of subgroups across all four case, and its performance gets even better as the sample size n increases.

The results of the parameter estimates are reported in Table B1 including $\widehat{\pi}_k$, $\widehat{\gamma}_{A,k}$, $\widehat{\beta}_{M,k}$, and $\widehat{\gamma}_{A,k}\widehat{\beta}_{M,k}$. Overall, estimates from censored data exhibit bigger bias and variability than those derived from uncensored data. Similar to the conclusion of uncensored data in Section 4.2, we have the similar findings. Most of the parameter estimates remain stable with relatively small bias and standard error. The estimation can be further improved as the sample size increases.

In addition, the estimation accuracy for mediation effect is a focus point. We present the variation trend of the estimation averaged bias in Figure B2 for uncensored data and in Figure B3 for censored data, which shows a monotonic downward trend as the sample size increases.

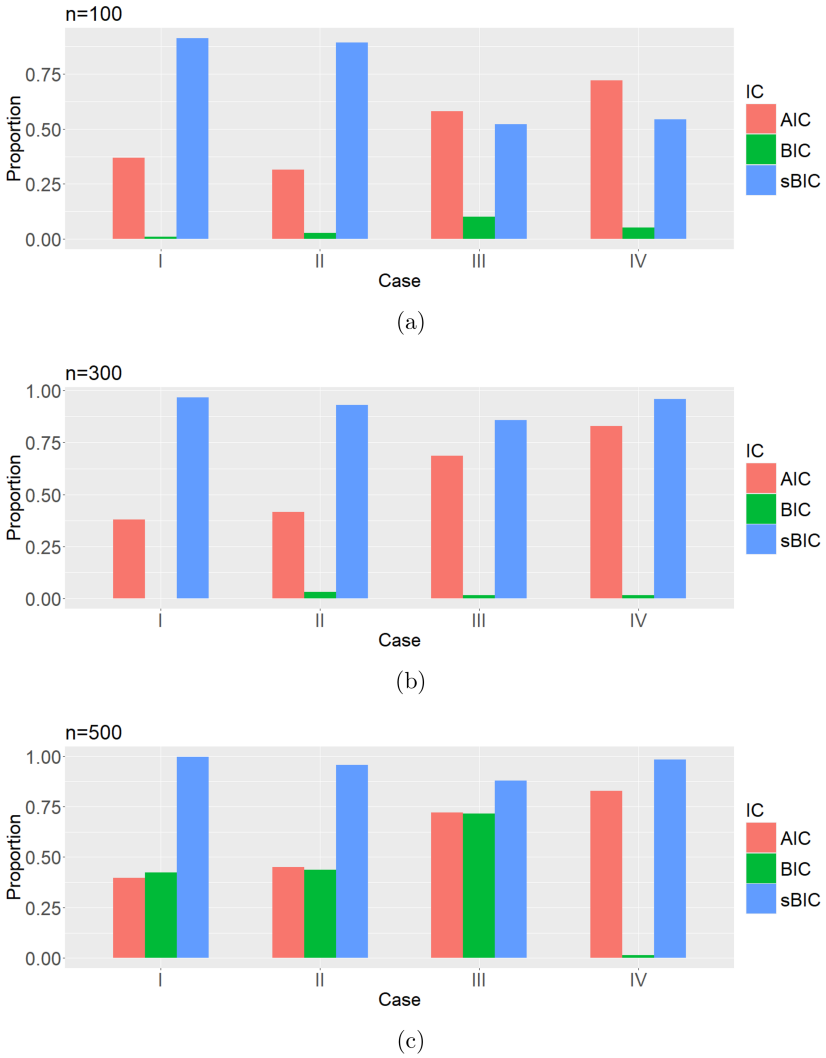

FIGURE B1 | The proportions of correctly selected number based on 1000 simulations for the four cases with 15% censoring rate, where (a–c) represent the sample sizes from 100 to 500, respectively.

TABLE B1 | The averaged bias and standard error (within parentheses) for the estimates of the model parameters based on 1000 simulations by the sBIC-EM algorithm with 15% censoring rate.

Case	n	k	bias$(\hat{\pi}_k)$	bias$(\hat{\gamma}_{A,k})$	bias$\left(\hat{\beta}_{M,k}\right)$	bias$\left(\hat{\gamma}_{A,k}\hat{\beta}_{M,k}\right)$
I	100	1	−0.010 (0.086)	−0.059 (0.253)	−0.006 (0.378)	−0.059 (0.186)
		2	0.010 (0.086)	0.054 (0.215)	0.082 (0.546)	0.085 (0.377)
	300	1	−0.003 (0.035)	−0.010 (0.123)	0.000 (0.133)	−0.008 (0.082)
		2	0.003 (0.035)	0.009 (0.085)	−0.008 (0.241)	0.002 (0.139)
	500	1	−0.002 (0.023)	−0.003 (0.087)	0.003 (0.089)	−0.001 (0.060)
		2	0.002 (0.023)	0.000 (0.062)	0.006 (0.175)	0.001 (0.100)
II	100	1	0.015 (0.104)	0.018 (0.213)	0.089 (0.443)	0.015 (0.319)
		2	−0.015 (0.104)	−0.009 (0.141)	−0.074 (0.515)	−0.049 (0.314)
	300	1	−0.001 (0.039)	0.006 (0.104)	−0.007 (0.184)	−0.012 (0.177)
		2	0.001 (0.039)	0.002 (0.069)	0.007 (0.229)	0.006 (0.160)
	500	1	−0.001 (0.025)	0.001 (0.079)	0.000 (0.087)	−0.001 (0.095)
		2	0.001 (0.025)	0.002 (0.054)	0.005 (0.129)	0.006 (0.102)
III	100	1	0.024 (0.094)	−0.003 (0.203)	−0.363 (1.046)	−0.211 (0.730)
		2	−0.040 (0.100)	0.116 (0.250)	−0.035 (0.862)	−0.022 (0.353)
		3	0.016 (0.099)	−0.162 (0.282)	0.169 (1.254)	0.005 (0.812)
	300	1	0.009 (0.078)	−0.009 (0.087)	−0.017 (0.543)	−0.013 (0.187)
		2	−0.006 (0.062)	0.032 (0.127)	0.060 (0.394)	0.026 (0.175)
		3	−0.003 (0.049)	−0.015 (0.095)	0.013 (0.338)	−0.001 (0.304)
	500	1	0.006 (0.048)	−0.007 (0.069)	0.000 (0.148)	0.008 (0.088)
		2	−0.003 (0.047)	0.006 (0.078)	0.004 (0.223)	0.003 (0.075)
		3	−0.003 (0.023)	−0.009 (0.063)	−0.001 (0.221)	−0.002 (0.170)
IV	100	1	0.025 (0.085)	−0.013 (0.231)	0.194 (1.057)	−0.122 (0.311)
		2	−0.021 (0.091)	0.153 (0.231)	−0.045 (0.760)	−0.047 (0.183)
		3	−0.004 (0.070)	−0.067 (0.235)	0.028 (0.407)	0.024 (0.250)
	300	1	0.007 (0.053)	0.004 (0.122)	−0.002 (0.210)	0.007 (0.109)
		2	−0.005 (0.043)	0.004 (0.063)	0.036 (0.373)	0.000 (0.015)
		3	−0.002 (0.029)	−0.045 (0.116)	0.029 (0.175)	−0.016 (0.086)
	500	1	0.001 (0.025)	0.016 (0.049)	0.014 (0.154)	0.014 (0.083)
		2	−0.003 (0.026)	0.002 (0.026)	0.019 (0.246)	0.000 (0.006)
		3	0.002 (0.019)	−0.021 (0.076)	−0.007 (0.077)	−0.019 (0.065)

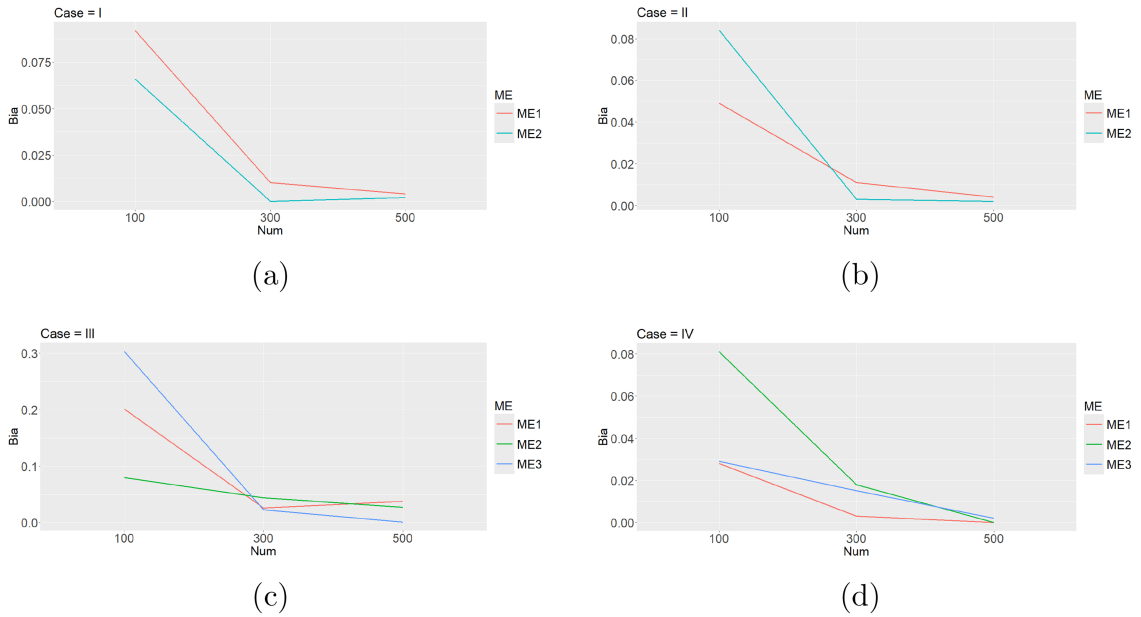

FIGURE B2 | (a–d) are the absolute value of the averaged bias for the estimates of the NIE based on 1000 simulations for four case with no censored time.

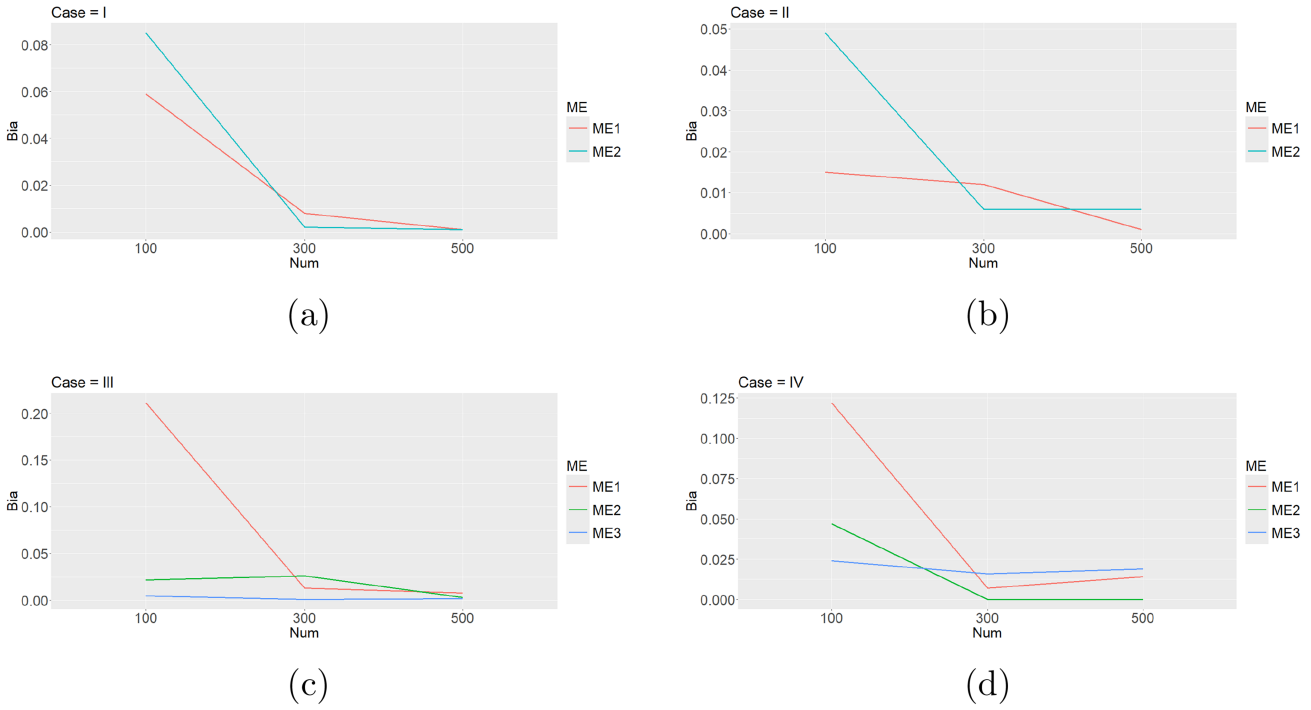

FIGURE B3 | (a–d) are the absolute value of the averaged bias for the estimates of the NIE based on 1000 simulations for four case with 15% censoring rate.

Appendix C

TCGA Reanalysis With Censored Data

In the simulation study, we set the censoring rate at 15%. For consistency, we select 428 patients, including all 343 deceased patients and 85 randomly selected surviving patients, thereby maintaining a similar censoring rate. For the same mediation model as in Section 5, we focus on selecting the optimal number of latent subgroups and estimating the NIE.

As shown in Figure C1, the number of latent subgroups for the CpG site cg19757631 was 2 using the sBIC method, which is the same as that for uncensored data. We set $K = 2$ and apply the nonparametric bootstrap to construct the interval estimates for the NIE. The resampling is done with 1000 times replacement for confidence interval. The two latent subgroups have different NIE sizes: −0.14 (95% CI [−3.92, −0.02]) and 0.18 (95% CI [−0.12,22.16]), with the mixing proportions 72% and 28%, respectively. Although the estimation accuracy decreases, the analysis results for the censored data are consistent with those for the uncensored data in Section 5. That is to say, smoking is harmful (−0.14) to most people (72%), but for a small subset of the population (28%), it may not be entirely detrimental (0.18).

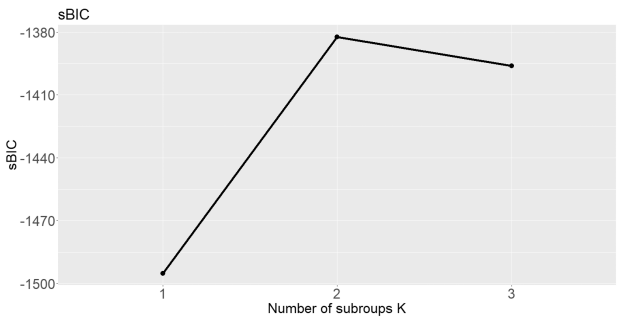

FIGURE C1 | The sBIC for the DNA methylation CpG sites cg19757631 which is maximized at $K = 2$.