

Article

A Semiparametric Probit Mixture Cure Model for Interval Censored Data with a Cure Fraction

Nan Zhang ¹ , Shishun Zhao ¹ and Dongmei Lu ^{2,*}

¹ School of Mathematics, Jilin University, Changchun 130012, China; zhangnan22@mails.jlu.edu.cn (N.Z.); zhaoss@jlu.edu.cn (S.Z.)

² Department of General Education, Changchun College of Electronic Technology, Changchun 130114, China

* Correspondence: ludongmei2025@126.com

Abstract

Accurately estimating the incubation period of emerging infectious diseases, defined as the time interval from infection to symptom onset, is essential for informing quarantine strategies and allocating public health resources. To address the challenges of interval-censored data commonly encountered in incubation period studies, as well as the possibility that some infected individuals may never develop symptoms, we propose a semiparametric probit mixture cure model. By introducing two latent variables to simplify the likelihood function, we develop a corresponding EM algorithm for parameter estimation. Under mild regularity conditions, we establish the consistency and asymptotic normality of the proposed estimators. Numerical simulation studies demonstrate that the proposed method performs well even with limited sample sizes. Furthermore, an empirical analysis of real-world COVID-19 data validates the practical applicability of the model. The proposed semiparametric probit mixed cure model offers methodological support for early prevention and control decision-making in response to sudden infectious disease outbreaks.

Keywords: semiparametric probit model; mixture cure model; EM algorithm; interval censoring; COVID-19

1. Introduction

In recent years, the recent emergence and spread of infectious diseases have globally challenged healthcare systems and destabilized social order worldwide. The COVID-19 pandemic, characterized by rapid transmission, unstable incubation period, and wide geographic spread, serves as a notable example. Accurate estimation of the incubation period, the interval between infection and symptom onset, is essential for understanding transmission dynamics. Precise estimation of the incubation period distribution facilitates the identification of infection sources and transmission chains, enables timely interventions to curb community transmission, and provides temporal clues for epidemiological traceability investigations, thereby helping to elucidate viral evolutionary pathways, mutational characteristics, and their complex interactions with the environment.

Several studies have estimated the incubation period distribution. Guan et al. [1] used descriptive statistics based on self-reported exposure histories. Backer et al. [2] constructed an interval likelihood using the duration of stay in Wuhan and symptom onset dates, fitting Weibull, gamma, and log-normal distributions. Li et al. [3] and Zhang et al. [4] assumed a log-normal distribution for patients with detailed exposure and symptom onset information. Lauer et al. [5] also assumed a log-normal distribution and applied a parametric accelerated



Received: 14 June 2026

Revised: 16 July 2026

Accepted: 17 July 2026

Published: 19 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

failure time model. For more complex data structures, Linton et al. [6] used a doubly interval-censored likelihood to model the exposure-to-onset interval. Deng et al. [7] and Qin et al. [8] treated the incubation period as a renewal process indexed by the departure time from Wuhan.

Given that the incubation period distribution may be non-standard, flexible semi-parametric and nonparametric methods have been explored. Wong et al. [9] assumed a semiparametric Cox proportional hazards model accommodating interval censoring, right truncation, and exponential epidemic growth, proposing a joint modeling approach based on sieved maximum likelihood estimation. Gressani et al. [10] developed a flexible semiparametric method using Laplace-P-splines to select between various parametric distributions and semiparametric fits in a data-driven manner. On the fully nonparametric side, Groeneboom [11] introduced a bandwidth parameter balancing noise and bias to obtain a smooth nonparametric estimator. Arntzen et al. [12] focused on robust estimation of the right-tail of the incubation period, demonstrating that parametric assumptions primarily affect right-tail percentiles and proposing an improved semiparametric method without such assumptions.

Despite these advances, all the above studies assume that all exposed individuals eventually develop symptoms. However, evidence suggests that some exposed individuals may not become infected or symptomatic. For example, Mizumoto et al. [13] found that approximately 18% of passengers on the Diamond Princess cruise ship were asymptomatic. Ignoring asymptomatic infections can bias the estimated median incubation period, and such bias may vary non-uniformly across age groups and immune statuses. In the context of mixture cure modeling for incubation period data, the “cured” fraction does not refer to clinically recovered patients, but rather to infected individuals who never develop symptoms, i.e., the asymptomatic subpopulation. This interpretation is consistent with the established usage of mixture cure models in infectious disease epidemiology, where the cured fraction represents the subpopulation not susceptible to the event of interest. For COVID-19, empirical evidence from multiple studies has confirmed the existence of a substantial asymptomatic proportion, with estimates ranging from approximately 18% among the Diamond Princess passengers to 8–43% depending on the population and testing strategy. Thus, treating asymptomatic individuals as “cured” is a statistically coherent approach that allows the incubation period distribution to be estimated without bias from those who will never experience symptom onset. Moreover, existing semiparametric methods predominantly rely on the Cox proportional hazards model or density smoothing techniques and have not yet incorporated a probit link function within a semiparametric mixture cure framework.

To accurately estimate the incubation time distribution in the presence of a cured (asymptomatic) fraction, statisticians have developed two-component mixture cure models [14–16]. These models assume that the target population consists of a mixture of susceptible and cured subgroups, offering intuitive interpretability. Let T denote a non-negative random variable representing the incubation period, let $\mathbf{X}$ and $\mathbf{Z}$ denote covariates, and the survival function is given by

$$S(t|\mathbf{X}, \mathbf{Z}) = 1 - \pi(\mathbf{Z}) + \pi(\mathbf{Z}) \cdot S_u(t|\mathbf{X})$$

where $\pi(\mathbf{Z})$ denotes the susceptible probability and $S_u(t|\mathbf{X})$ represents the conditional survival function for susceptible individuals. The relationship between the covariates $\mathbf{Z}$ and $\pi(\mathbf{Z})$ is typically modeled via a link function, with the most common link function being $\text{logit}(\pi(\mathbf{Z})) = \boldsymbol{\eta}^\top \mathbf{Z}$.

For the incidence component, the logistic link function remains the most widely adopted choice due to its interpretability in terms of odds ratios. Pal et al. [17] demon-

strated that the linear assumption of logistic/probit links for incidence may be restrictive in some settings, and proposed a support vector machine-based alternative. However, their critique targets the incidence component specifically; in our application, we retain that the logistic model provides a clinically interpretable baseline, and our primary innovation lies in the latency component rather than the incidence component. For the latency component, the proportional hazards (PH) structure has been the predominant choice. Foundational contributions by Kuk and Chen [18], Peng and Dear [19], and Sy and Taylor [20] established the foundational semiparametric PH mixture cure models for right-censored data, which were subsequently adapted to interval-censored settings. More recent extension include the joint semiparametric framework of Lou et al. [21] for interval-censored data with cure subgroups and nonignorable missing covariates, the doubly semiparametric mixture cure model of Liu and Xiang [22] permitting simultaneous linear and nonlinear covariate effects in both components, and the Bayesian PH mixture cure model for partly interval-censored data developed by Pan and Cai [23]. Beyond the PH framework, several alternative specifications for the latency component have been investigated, including the Box–Cox transformation cure model of Pal and Barui [24], which unifies mixture and promotion time cure formulations through a transformation approach. Semiparametric transformation models have also been explored as a flexible class that extends the Cox model. Additionally, additive hazard models and accelerated failure time models have been adapted to the interval-censored cure setting, though these remain less commonly used than the PH approach.

Despite these advances, the use of a semiparametric probit specification for the latency component of interval-censored mixture cure models has received relatively limited attention. Most existing studies have focused on proportional hazards, accelerated failure time, or transformation models for the latency part, leaving the probit specification largely unexplored in this setting. To fill this gap, we propose employing a semiparametric probit model for the survival probability of susceptible individuals, $S_u(t|x)$. The probit model was originally proposed by Bliss [25] to relate dose to response probability through a normality assumption, establishing its fundamental role in biology. It has since been widely applied in econometrics and biostatistics. In survival analysis, Lin and Wang [26] extended the probit model to a semiparametric framework for type II interval-censored data. Under this formulation, the conditional distribution of T given $\mathbf{X}$ as

$$F(t|\mathbf{X}) = P(T \leq t|\mathbf{X}) = \Phi(\alpha(t) + \boldsymbol{\beta}^\top \mathbf{X}).$$

where $\Phi(\cdot)$ represents a standard normal cumulative distribution function (CDF), $\boldsymbol{\beta}$ is a p -dimensional regression coefficient vector, and $\alpha(\cdot)$ is an unspecified non-decreasing baseline function satisfying $\alpha(0) = -\infty$, $\alpha(\infty) = \infty$.

The above formulation, combined with a logistic regression for the incidence component, constitutes the proposed logistic-probit mixture cure framework. To our knowledge, this represents the first integration of these two components in the context of interval-censored data with a cure fraction. The proposed model offers three specific contributions: (i) a novel logistic-probit mixture cure framework; (ii) a flexible B-spline estimation of the baseline function with monotonicity; (iii) the first application to interval-censored COVID-19 incubation data, contributing a new methodological toolkit for survival analysis with a cure fraction.

The remainder of this paper is organized as follows. Section 2 presents the mathematical formulation of the semiparametric probit mixture cure model, including its likelihood function and an EM algorithm. Section 3 establishes the asymptotic properties of the estimators. Section 4 evaluates the finite-sample performance through numerical simulation

studies. Section 5 applies the model to a real COVID-19 dataset. Section 6 discusses the model's applications and limitations.

2. Models and Likelihood

For each subject i , we observe a sequence of examination times $0 < V_{i1} < \dots < V_{i,k_i} < \infty$, with the event time T_i known only to lie within the interval $(L_i, R_i]$. The observed data are recorded as $\mathbf{O} = \{O_i = (L_i, R_i, \delta_{L,i}, \delta_{I,i}, \delta_{R,i}, \mathbf{X}_i, \mathbf{Z}_i); i = 1, \dots, n\}$, where $\delta_{L,i} = 1$ for left-censored observations ($L_i = 0$); $\delta_{I,i} = 1$ for interval-censored observations ($L_i < T_i \leq R_i$); and $\delta_{R,i} = 1$ for right-censored observations ($R_i = \infty$) with $\delta_{L,i} + \delta_{I,i} + \delta_{R,i} = 1$ for every i . Conditional on the covariate $\mathbf{X}_i$, the examination times V_{ij} are assumed to be independent of the event time T_i . Under this assumption, the observed likelihood function is given by

$$\begin{aligned} \mathcal{L} = & \prod_{i=1}^n \pi(\mathbf{Z}_i)^{1-\delta_{R,i}} \{\Phi(\alpha(R_i) + \boldsymbol{\beta}^\top \mathbf{X}_i)\}^{\delta_{L,i}} \\ & \times \{\Phi(\alpha(R_i) + \boldsymbol{\beta}^\top \mathbf{X}_i) - \Phi(\alpha(L_i) + \boldsymbol{\beta}^\top \mathbf{X}_i)\}^{\delta_{I,i}} \\ & \times \{1 - \pi(\mathbf{Z}_i) + \pi(\mathbf{Z}_i)(1 - \Phi(\alpha(L_i) + \boldsymbol{\beta}^\top \mathbf{X}_i))\}^{\delta_{R,i}}. \end{aligned} \quad (1)$$

Under the above likelihood framework, the population is decomposed into two distinct latent subgroups: the cured (non-susceptible) group, who will never experience the event of interest, and the susceptible group, who are at risk of experiencing the event. Accordingly, the model comprises two components with clearly differentiated roles, as detailed below:

(i) The incidence component models the susceptible probability $\pi(\mathbf{Z}_i)$. Specifically, a logistic regression is adopted:

$$\text{logit}(\pi(\mathbf{Z}_i)) = \frac{1}{1 + e^{-\boldsymbol{\eta}^\top \mathbf{Z}_i}}.$$

where $\mathbf{Z}_i$ is a covariates and $\boldsymbol{\eta}$ is the p -dimensional regression coefficient vector.

(ii) The latency component models the incubation period distribution among susceptible individuals. The survival function $S_u(t_i|\mathbf{X}_i)$ for this subgroup is specified via a semiparametric probit model:

$$S_u(t_i|\mathbf{X}_i) = 1 - F(t_i|\mathbf{X}_i) = 1 - \Phi(\alpha(t_i) + \boldsymbol{\beta}^\top \mathbf{X}_i).$$

where $\Phi(\cdot)$ is the standard normal CDF, $\boldsymbol{\beta}$ is a p -dimensional regression coefficient vector for covariates $\mathbf{X}_i$, and $\alpha(\cdot)$ is an unspecified non-decreasing baseline function satisfying $\alpha(0) = -\infty$ and $\alpha(\infty) = \infty$. The nonparametric specification of $\alpha(\cdot)$ renders this component semiparametric, as it imposes no restrictive parametric form on the baseline time to event distribution.

To construct a complete likelihood function and simplify the derivation of the EM algorithm, we introduce two latent variables, u_i and Y_i .

Let $u_i \sim \text{Bernoulli}(\pi(\mathbf{Z}_i))$, where $u_i = 1$ if subject i is uncured (susceptible) and $u_i = 0$ if subject i is cured (non-susceptible). The probability $\pi(\mathbf{Z}_i)$ depends on the covariate vectors $\mathbf{Z}_i$ through the logistic regression. Denote $\mathbf{u} = (u_1, \dots, u_n)$ as the vector of latent susceptibility indicators. Under the standard assumption that cured subjects are necessarily right-censored, we have $(1 - u_i)\delta_{R,i} = (1 - u_i)$; that is, a cured individual

contributes only to the right-censored group. Then, the observed data likelihood function under parameter $\theta = (\beta, \eta, \alpha)$ is

$$\begin{aligned}\mathcal{L}(\theta|\mathbf{O}, \mathbf{u}) &= \prod_{i=1}^n \mathcal{L}_i(\theta|O_i, u_i) \\ &= \prod_{i=1}^n \pi(\mathbf{Z}_i)^{u_i} (1 - \pi(\mathbf{Z}_i))^{1-u_i} \{\Phi(\alpha(R_i) + \beta^\top \mathbf{X}_i)\}^{\delta_{L,i}} \\ &\quad \times \{\Phi(\alpha(R_i) + \beta^\top \mathbf{X}_i) - \Phi(\alpha(L_i) + \beta^\top \mathbf{X}_i)\}^{\delta_{L,i}} \\ &\quad \times \{1 - \Phi(\alpha(L_i) + \beta^\top \mathbf{X}_i)\}^{\delta_{R,i} u_i}.\end{aligned}\quad (2)$$

Define Y_i in relation to the event time T_i as $Y_i = \alpha(T_i) + \beta^\top \mathbf{X}_i + \epsilon_i$, where $\epsilon_i \sim N(0, 1)$. Since $\alpha(s)$ is a non-decreasing function, the event $\{T_i \leq s\}$ is equivalent to $\{Y_i \leq \alpha(s) + \beta^\top \mathbf{X}_i\}$ for any time point s , yielding the standard probit model:

$$P(T_i \leq s|\mathbf{X}_i) = \Phi(\alpha(s) + \beta^\top \mathbf{X}_i).$$

Because Y_i is not directly observable in practice, we use the censoring type of the observed interval $(L_i, R_i]$ to provide truncation information on Y_i , where $T_i \leq R_i$ when $Y_i \leq \alpha(R_i) + \beta^\top \mathbf{X}_i$; where $L_i < T_i \leq R_i$ when $\alpha(L_i) + \beta^\top \mathbf{X}_i < Y_i \leq \alpha(R_i) + \beta^\top \mathbf{X}_i$; and where $T_i > L_i$ when $Y_i > \alpha(L_i) + \beta^\top \mathbf{X}_i$. Thus, the observed data likelihood under parameter θ can be written as

$$\begin{aligned}\mathcal{L}(\theta|\mathbf{O}, \mathbf{u}, \mathbf{Y}) &= \prod_{i=1}^n \mathcal{L}_i(\theta|O_i, u_i, Y_i) \\ &= \prod_{i=1}^n \pi(\mathbf{Z}_i)^{u_i} (1 - \pi(\mathbf{Z}_i))^{1-u_i} \times \{P(Y_i > 0)\}^{\delta_{L,i}} \\ &\quad \times \{P(\alpha(L_i) - \alpha(R_i) < Y_i < 0)\}^{\delta_{L,i}} \times \{P(Y_i < 0)\}^{\delta_{R,i} u_i}.\end{aligned}\quad (3)$$

where $\delta_{L,i} = 1$ when $Y_i > 0$; where $\delta_{L,i} = 1$ when $\alpha(L_i) - \alpha(R_i) < Y_i < 0$; and where $\delta_{R,i} = 1$ when $Y_i < 0$.

The function $\alpha(t_i)$ is estimated using B-spline under a log-transformation strategy, taking the form

$$\alpha(t_i) = \alpha_{t_i} = \log\left\{\sum_{k=1}^K \gamma_k B_k(t_i)\right\}.$$

where $\gamma = (\gamma_1, \gamma_2, \dots, \gamma_K)$ are the corresponding coefficients and $B_k(t_i)$ denotes the B-spline basis functions defined at t_i , with $B_k(t_i) \geq 0$ for all t . The number of B-spline basis functions K must satisfy $K \geq 2$ and the knot locations should cover the observed time range. Let $\mathbf{B}(t_i) = (B_1(t_i), B_2(t_i), \dots, B_K(t_i))^\top$ be a $K \times 1$ column vector. To ensure the monotonicity of α_{t_i} , a required property for the baseline function in the semiparametric probit model, we impose a sufficient condition on the B-spline coefficients: $\gamma_1 \leq \gamma_2 \leq \dots \leq \gamma_K$. Since $B_k(t_i) \geq 0$ for all t , this condition guarantees that $\mathbf{B}(t_i)^\top \gamma$ is non-decreasing in t ; consequently, $\alpha_{t_i} = \log(\mathbf{B}(t_i)^\top \gamma)$ is also non-decreasing due to the monotonicity of the logarithmic transformation.

Finally, the completed data likelihood function under parameter $\Psi = (\eta, \beta, \gamma)$ is

$$\begin{aligned}\mathcal{L}_c(\Psi|\mathbf{O}, \mathbf{u}, \mathbf{Y}) &= \prod_{i=1}^n \pi(\mathbf{Z}_i)^{u_i} (1 - \pi(\mathbf{Z}_i))^{1-u_i} [f(Y_i|\alpha(t_i) + \beta^\top \mathbf{X}_i)]^{u_i} \\ &= \prod_{i=1}^n \pi(\mathbf{Z}_i)^{u_i} (1 - \pi(\mathbf{Z}_i))^{1-u_i} \\ &\quad \times \left[\frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}(Y_i - (\log\{\mathbf{B}(t_i)^\top \gamma\} + \beta^\top \mathbf{X}_i))^2\right\} \right]^{u_i}\end{aligned}\quad (4)$$

where $\delta_{L,i} = 1, Y_i > 0; \delta_{L,i} = 1, \alpha(L_i) - \alpha(R_i) < Y_i < 0; \delta_{R,i} = 1, Y_i < 0$. $Y_i \sim N(\alpha(t_i) + \beta^\top \mathbf{X}_i, 1)$, $Y_i = \alpha(t_i) + \beta^\top \mathbf{X}_i + \epsilon_i$, and $t_i = R_i 1_{(\delta_{L,i}=1)} + L_i 1_{(\delta_{L,i}=0)}$ is an auxiliary variable. When $u_i = 1$, Y_i follows the above normal density. When $u_i = 0$, Y_i does not contribute to the likelihood.

To estimate the parameters $\Psi = (\eta, \beta, \gamma)$, we further apply the EM algorithm, yielding the completed data log-likelihood by taking the logarithm of Equation (4):

$$\begin{aligned}l(\Psi|\mathbf{O}, \mathbf{u}, \mathbf{Y}) &= \sum_{i=1}^n l_i(\eta, \beta, \gamma|O_i, u_i, Y_i) \\ &= \sum_{i=1}^n [u_i \log \pi(\mathbf{Z}_i) + (1 - u_i) \log(1 - \pi(\mathbf{Z}_i))] \\ &\quad + \sum_{i=1}^n u_i \left[-\frac{1}{2} \log(2\pi) - \frac{1}{2} (Y_i - (\log\{\mathbf{B}(t_i)^\top \gamma\} + \beta^\top \mathbf{X}_i))^2 \right]\end{aligned}\quad (5)$$

Given the observed data $\mathbf{O}$ and current estimates $\Psi^{(d)}$, we decompose the expected completed data log-likelihood $Q(\Psi; \Psi^{(d)})$ into two components:

$$Q(\Psi; \Psi^{(d)}) = Q_1(\eta; \eta^{(d)}) + Q_2(\beta, \gamma; \beta^{(d)}, \gamma^{(d)}).$$

where

$$Q_1(\eta; \eta^{(d)}) = \sum_{i=1}^n E(u_i|\mathbf{O}, \Psi^{(d)}) \log(\pi(\mathbf{Z}_i)) + (1 - E(u_i|\mathbf{O}, \Psi^{(d)})) \log(1 - \pi(\mathbf{Z}_i)).$$

$$Q_2(\beta, \gamma; \beta^{(d)}, \gamma^{(d)}) = - \sum_{i=1}^n E_{u_i, Y_i|\mathbf{O}, \Psi^{(d)}} u_i \left(\frac{1}{2} (Y_i - (\log\{\mathbf{B}(t_i)^\top \gamma\} + \beta^\top \mathbf{X}_i))^2 \right) + L(\Psi^{(d)}).$$

where $L(\Psi^{(d)})$ is a function that does not depend on Ψ .

Let $w_i^{(d)} = E(u_i|\mathbf{O}, \Psi^{(d)})$, Q_1 simplifies to

$$Q_1(\eta; \eta^{(d)}) = \sum_{i=1}^n w_i^{(d)} \log(\pi(\mathbf{Z}_i)) + (1 - w_i^{(d)}) \log(1 - \pi(\mathbf{Z}_i)).$$

$$w_i = E(u_i|\mathbf{O}, \Psi) = 1 - \delta_{R,i} + \delta_{R,i} \frac{\pi(\mathbf{Z}_i)(1 - \Phi(\alpha(L_i) + \beta^\top \mathbf{X}_i))}{1 - \pi(\mathbf{Z}_i) + \pi(\mathbf{Z}_i)(1 - \Phi(\alpha(L_i) + \beta^\top \mathbf{X}_i))}.$$

$Q_2(\beta, \gamma; \beta^{(d)}, \gamma^{(d)})$ can be simplified into a square loss form:

$$\begin{aligned} Q_2(\beta, \gamma; \beta^{(d)}, \gamma^{(d)}) &= - \sum_{i=1}^n E_{u_i, Y_i | \mathbf{O}, \Psi^{(d)}} [u_i (Y_i - (\log\{\mathbf{B}(t_i)^\top \gamma\} + \beta^\top \mathbf{X}_i))^2] \\ &= - \sum_{i=1}^n E(u_i | \mathbf{O}, \Psi^{(d)}) E[(Y_i - (\alpha(t_i) + \beta^\top \mathbf{X}_i))^2 | \mathbf{O}, \Psi^{(d)}, u_i = 1] \\ &= - \sum_{i=1}^n E(u_i | \mathbf{O}, \Psi^{(d)}) [\text{Var}(Y_i | u_i = 1, \mathbf{O}, \Psi^{(d)}) \\ &\quad + (E(Y_i | u_i = 1, \mathbf{O}, \Psi^{(d)}) - (\log\{\mathbf{B}(t_i)^\top \gamma\} + \beta^\top \mathbf{X}_i))^2] \\ &= - \sum_{i=1}^n w_i^{(d)} \sum_{m \in \{L, I, R\}} \delta_{m,i} \{\tilde{v}_{\delta_{m,i}} + (\tilde{\mu}_{\delta_{m,i}} - \mu_i)^2\}. \end{aligned}$$

where $t_i = R_i 1_{(\delta_{L,i}=1)} + L_i 1_{(\delta_{L,i}=0)}$ and $w_i^{(d)} = E(Y_i | u_i, \mathbf{O}, \Psi^{(d)}) = P(Y_i | u_i = 1, \mathbf{O}, \Psi^{(d)})$. Define $\tilde{\mu}_{\delta_{m,i}} = E(Y_i | u_i = 1, \mathbf{O}, \Psi^{(d)})$, $\tilde{v}_{\delta_{m,i}}$ as the conditional variance of Y_i given that the individual is uncured, and the censored observation as type m , with $m \in \{L, I, R\}$.

In Q_2 , the term $\sum_{m \in \{L, I, R\}} \delta_{m,i} \tilde{v}_{\delta_{m,i}}$ is additive to the squared term and constant with respect to μ_i , thus it can be ignored when maximizing Q_2 . The final objective function becomes

$$Q_{2\text{new}}(\beta, \gamma; \beta^{(d)}, \gamma^{(d)}) = - \sum_{i=1}^n w_i^{(d)} \sum_{m \in \{L, I, R\}} \delta_{m,i} (\tilde{\mu}_{\delta_{m,i}} - \mu_i)^2.$$

Assume that where $\delta_{L,i} = 1$, $\tilde{\mu}_{R,i} = \log(\mathbf{B}(R_i)^\top \tilde{\gamma}) + \tilde{\beta}^\top \mathbf{X}_i$; where $\delta_{L,i} = 0$, $\tilde{\mu}_{L,i} = \log(\mathbf{B}(L_i)^\top \tilde{\gamma}) + \tilde{\beta}^\top \mathbf{X}_i$. By properties of the truncated normal distribution,

$$\begin{aligned} \tilde{\mu}_{\delta_{L,i}} &= E(Y_i | u_i = 1, Y_i > 0) = \tilde{\mu}_{R,i} + \frac{\phi(\tilde{\mu}_{R,i})}{\Phi(\tilde{\mu}_{R,i})}. \\ \tilde{\mu}_{\delta_{I,i}} &= E(Y_i | u_i = 1, a_i < Y_i < 0) = \tilde{\mu}_{L,i} + \frac{\phi(a - \tilde{\mu}_{L,i}) - \phi(\tilde{\mu}_{L,i})}{1 - \Phi(\tilde{\mu}_{L,i}) - \Phi(a - \tilde{\mu}_{L,i})}. \\ \tilde{\mu}_{\delta_{R,i}} &= E(Y_i | u_i = 1, Y_i < 0) = \tilde{\mu}_{L,i} - \frac{\phi(\tilde{\mu}_{L,i})}{1 - \Phi(\tilde{\mu}_{L,i})}. \\ \tilde{v}_{\delta_{L,i}} &= \text{var}_{\tilde{\theta}}(Y_i | u_i = 1, Y_i > 0) = 1 - \frac{\phi^2(\tilde{\mu}_{R,i})}{\Phi^2(\tilde{\mu}_{R,i})}. \end{aligned}$$

$$\begin{aligned} \tilde{v}_{\delta_{I,i}} &= \text{var}_{\tilde{\theta}}(Y_i | u_i = 1, a_i < Y_i < 0) \\ &= 1 + \frac{(a - \tilde{\mu}_{L,i})\phi(a - \tilde{\mu}_{L,i}) + \tilde{\mu}_{L,i}\phi(\tilde{\mu}_{L,i})}{1 - \Phi(\tilde{\mu}_{L,i}) - \Phi(a - \tilde{\mu}_{L,i})} - \left(\frac{\phi(a - \tilde{\mu}_{L,i}) - \phi(\tilde{\mu}_{L,i})}{1 - \Phi(\tilde{\mu}_{L,i}) - \Phi(a - \tilde{\mu}_{L,i})} \right)^2. \end{aligned}$$

$$\tilde{v}_{\delta_{R,i}} = \text{var}_{\tilde{\theta}}(Y_i | u_i = 1, Y_i < 0) = 1 - \frac{\phi^2(\tilde{\mu}_{L,i})}{(1 - \Phi(\tilde{\mu}_{L,i}))^2}.$$

where $a = \alpha(L_i) - \alpha(R_i) < 0$ (since $\alpha(\cdot)$ is a non-decreasing and $L_i < R_i$). $\phi(\cdot)$ and $\Phi(\cdot)$ denote the probability density function and cumulative distribution function of the standard normal distribution, respectively.

Let $\tilde{Y}_i = \delta_{L,i} \tilde{\mu}_{\delta_{L,i}} + \delta_{I,i} \tilde{\mu}_{\delta_{I,i}} + \delta_{R,i} \tilde{\mu}_{\delta_{R,i}}$. Consider the function $J(\beta, \gamma) = \sum_{i=1}^n w_i^{(d)} \{\tilde{Y}_i - \log\{\mathbf{B}(t_i)^\top \gamma\} - \beta^\top \mathbf{X}_i\}^2$. Maximizing $Q_{2\text{new}}$ is equivalent to minimizing $J(\beta, \gamma)$. Assume $\mathbf{W} = \text{diag}(w_1^{(d)}, w_2^{(d)}, \dots, w_n^{(d)}) \in \mathbb{R}^{n \times n}$, $\tilde{\mathbf{Y}} = (\tilde{Y}_1, \dots, \tilde{Y}_n)^\top \in \mathbb{R}^n$, and $\mathbf{B} \in \mathbb{R}^{n \times K}$ with the i -th row equal to $\mathbf{B}(t_i)^\top$, where K is the number of B-spline basis function and $\mathbf{X} \in \mathbb{R}^{n \times p}$ with the i -th row equal to $\mathbf{X}_i^\top$. Then

$$J(\beta, \gamma) = (\tilde{\mathbf{Y}} - \log\{\mathbf{B}\gamma\} - \mathbf{X}\beta)^\top \mathbf{W} (\tilde{\mathbf{Y}} - \log\{\mathbf{B}\gamma\} - \mathbf{X}\beta).$$

Taking the derivative with respect to β :

$$\frac{\partial J}{\partial \beta} = -2\mathbf{X}^\top \mathbf{W}(\tilde{\mathbf{Y}} - \log\{\mathbf{B}\gamma\} - \mathbf{X}\beta) = 0.$$

$$\tilde{\beta} = (\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{W}(\tilde{\mathbf{Y}} - \log\{\mathbf{B}\gamma\}).$$

Taking the derivative with respect to γ :

$$\frac{\partial J}{\partial \gamma} = -2\mathbf{B}^\top \text{diag}\left(\frac{1}{\mathbf{B}\gamma}\right) \mathbf{W}(\tilde{\mathbf{Y}} - \log\{\mathbf{B}\gamma\} - \mathbf{X}\beta) = 0.$$

where $\frac{1}{\mathbf{B}\gamma}$ denotes element-wise reciprocal (requiring $\mathbf{B}\gamma > 0$ and $\mathbf{B}\gamma$ is an $n \times 1$ vector, whose i -th element is $\mathbf{B}(t_i)^\top \gamma$).

We implement the EM algorithm through the following iterative steps:

Step 1 Set iteration counter $d = 0$ and initialize the parameters $\Psi^{(0)} = (\eta^{(0)}, \beta^{(0)}, \gamma^{(0)})$.

Step 2: Based on the current parameters $\Psi^{(d)}$, compute

$$w_i^{(d)} = E(u_i | \mathbf{O}, \Psi^{(d)}), i = 1, 2, \dots, n.$$

$$\tilde{\mu}_{\delta_{m,i}}^{(d)} = E(Y_i | u_i = 1, \mathbf{O}, \Psi^{(d)}, \delta_{m,i} = 1), m \in \{L, I, R\}.$$

$$\tilde{Y}_i^{(d)} = \delta_{L,i} \tilde{\mu}_{\delta_{L,i}}^{(d)} + \delta_{I,i} \tilde{\mu}_{\delta_{I,i}}^{(d)} + \delta_{R,i} \tilde{\mu}_{\delta_{R,i}}^{(d)}.$$

Step 3: Update β :

$$\beta^{(d+1)} = (\mathbf{X}^\top \mathbf{W}^{(d)} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{W}^{(d)} (\tilde{\mathbf{Y}}^{(d)} - \log\{\mathbf{B}\gamma^{(d)}\}).$$

where $\mathbf{W} = \text{diag}(w_1^{(d)}, w_2^{(d)}, \dots, w_n^{(d)})$, $\tilde{\mathbf{Y}}^{(d)} = (\tilde{Y}_1^{(d)}, \tilde{Y}_2^{(d)}, \dots, \tilde{Y}_n^{(d)})^\top$.

Step 4: Update γ . Compute the gradient

$$\frac{\partial J}{\partial \gamma} = -2\mathbf{B}^\top \text{diag}\left(\frac{1}{\mathbf{B}\gamma^{(d)}}\right) \mathbf{W}(\tilde{\mathbf{Y}}^{(d)} - \log\{\mathbf{B}\gamma^{(d)}\} - \mathbf{X}\beta^{(d+1)}).$$

where $\mathbf{B}\gamma^{(d)}$ is a vector, $\frac{1}{\mathbf{B}\gamma^{(d)}}$ is the element-wise reciprocal, and $\text{diag}(\frac{1}{\mathbf{B}\gamma^{(d)}})$ is an $n \times n$ diagonal matrix. The condition $\mathbf{B}\gamma^{(d)} > 0$ must hold for all i . To ensure numerical stability and avoid manual tuning of the step size, we determine the learning rate λ adaptively via an Armijo backtracking line search [27]. Starting from an initial value $\lambda_0 = 0.01$, we iteratively reduce $\lambda \leftarrow \tau \lambda$ with $\tau = 0.8$ until the Armijo condition is reached:

$$J(\gamma^{(d+1)}) \leq J(\gamma^{(d)}) - c\lambda \|\nabla J(\gamma^{(d)})\|^2,$$

satisfied with $c = 10^{-4}$. This adaptive strategy guarantees a sufficient decrease in the objective function $J(\gamma)$ at each gradient update while preventing oscillations and maintaining the feasibility of $\mathbf{B}\gamma > 0$. Perform a gradient descent update:

$$\tilde{\gamma}^{(d+1)} = \gamma^{(d)} - \lambda \frac{\partial J}{\partial \gamma} \Big|_{\gamma=\gamma^{(d)}, \beta=\beta^{(d+1)}}.$$

To enforce the constraint $B\gamma > 0$, we project $\tilde{\gamma}^{(d+1)}$ onto the feasible set $\mathcal{C} = \{\gamma : B_i^\top \gamma \geq \varepsilon, i = 1, \dots, n\}$, with a small tolerance of $\varepsilon > 0$, yielding

$$\gamma^{(d+1)} = \arg \min_{\gamma \in \mathcal{C}} \|\gamma - \tilde{\gamma}^{(d+1)}\|_2^2.$$

This projection guarantees that $B\gamma^{(d+1)} > 0$ holds for all individuals, ensuring the logarithmic transformation remains well-defined throughout the EM iterations.

Step 5: Update η :

$$\eta^{(d+1)} = \arg \max_{\eta} \sum_{i=1}^n [w_i^{(d)} \log \pi(\mathbf{Z}_i) + (1 - w_i^{(d)}) \log(1 - \pi(\mathbf{Z}_i))].$$

Step 6: If the relative change in the log-likelihood between successive iterations satisfies

$$\frac{|\ell(\Psi^{(d+1)}) - \ell(\Psi^{(d)})|}{|\ell(\Psi^{(d)})| + 1} < \epsilon_1,$$

or the maximum number of iterations $T_{\max} = 500$ is reached, then stop the iteration. Otherwise, set $d = d + 1$ and return to Step 2.

Step 7: Output the final estimated parameters $\hat{\Psi} = \Psi^{(d+1)}$.

The EM algorithm is iterated until the relative change in the log-likelihood between successive iterations falls below $\epsilon_1 = 10^{-4}$, or a maximum of $T_{\max} = 500$ iterations is reached. Within the M-step, the gradient descent update for γ terminates when the ℓ_2 -norm of the gradient is less than $\epsilon_2 = 10^{-6}$. The monotone likelihood property of the EM algorithm Dempster et al. [28] ensures that the observed log-likelihood $\ell(\Psi)$ is non-decreasing after each iteration. Although the spline coefficients γ are updated via gradient descent within the M-step rather than through a closed-form maximization, the Armijo backtracking line search guarantees a sufficient decrease in the Q-function at each gradient update. Moreover, the projection step ensures that all iterates remain within the feasible region $\mathcal{C} = \{\gamma : B\gamma \geq 0\}$. Consequently, the overall monotonicity of the EM algorithm is preserved, and the algorithm converges to a stationary point of the observed log-likelihood under standard regularity conditions [29].

3. Asymptotic Properties

We discuss the asymptotic properties of the proposed estimators. Let $\theta = (\beta, \eta, \alpha)$, where β denotes the regression coefficients, η denotes the regression coefficients in the cure model, and α is an unspecified monotone increasing function. The corresponding estimators are $\hat{\theta} = (\hat{\beta}, \hat{\eta}, \hat{\alpha})$, and the true values $\theta_0 = (\beta_0, \eta_0, \alpha_0)$. Let $\ell^\infty[x, y]$ denote the space of bounded functions on $[x, y]$, and define

$$\mathcal{A} = \{\alpha(t) : \alpha(t) \text{ is monotone increasing, } -\infty < \alpha(t) < \infty, t \in [\rho, \tau], \rho > 0\}.$$

The parameter space is $\Theta = \mathcal{B} \times \mathcal{H} \times \mathcal{A}$, where $\mathcal{B} \subset \mathbb{R}^{p_\beta}$ and $\mathcal{H} \subset \mathbb{R}^{p_\eta}$ are compact. For $\theta_1, \theta_2 \in \Theta$, define

$$d(\theta_1, \theta_2) = (\|\beta_1 - \beta_2\|^2 + \|\eta_1 - \eta_2\|^2 + \|\alpha_1 - \alpha_2\|_{L_2}^2)^{1/2},$$

with $\|\alpha_1 - \alpha_2\|_{L_2} = \sqrt{\int_\rho^\tau \{\alpha_1(t) - \alpha_2(t)\}^2 dt}$. The symbols $\|\cdot\|$ and $\|\cdot\|_\infty$ denote the Euclidean and supremum norms, respectively. The following regularity conditions are imposed:

- (C1) The true values β_0 and η_0 lie in the interiors of $\mathcal{B}$ and $\mathcal{H}$; the cumulative baseline hazard function $\alpha_0 \in \mathcal{A}$ and its first derivative is strictly positive and continuous on $[\rho, \tau]$.
- (C2) The covariates X (corresponding to β) and Z (corresponding to η) are bounded almost surely, with positive definite covariance matrices.
- (C3) For some $\xi \geq 1$, the ξ -th derivative $\alpha_0^{(\xi)}(t)$ is Lipschitz on $[\rho, \tau]$: there exists a constant $L > 0$ such that for all $t_1, t_2 \in [\rho, \tau]$,

$$|\alpha_0^{(\xi)}(t_1) - \alpha_0^{(\xi)}(t_2)| \leq L|t_1 - t_2|.$$

- (C4) The number of knots $K = O(n^\nu)$ with $0 < \nu < 1/2$, and the maximum spacing $\Delta = \max_i |t_i - t_{i-1}| = O(n^{-\nu})$, with a bounded ratio of the maximum to the minimum spacing.
- (C5) The joint density function $f(t, x, z, \delta_L, \delta_I, \delta_R)$ of $(T, X, Z, \delta_L, \delta_I, \delta_R)$ has bounded ξ -th order partial derivatives in t .
- (C6) If $\eta^\top z + f(x) = 0$ almost surely on $\mathcal{X}$, then $\eta = 0$ and $f(x) = 0$.

Conditions (C1)–(C2) are standard boundedness assumptions for interval-censored survival models. Condition (C3) ensures the smoothness of the spline functions, providing theoretical support for spline approximation. Condition (C4) is needed for establishing the convergence rate and asymptotic normality. Condition (C5) ensures the score operator for the nonparametric component vanishes along least favorable directions, yielding asymptotic normality of the parametric estimators. Condition (C6) ensures the Fisher information matrix is nonsingular.

Under the above regularity conditions, we establish three theorems on the asymptotic properties of the estimators $\hat{\beta}$, $\hat{\eta}$, and $\hat{\alpha}$.

Theorem 1. Under (C1)–(C4), the sieve maximum likelihood estimators satisfy

$$\|\hat{\beta} - \beta_0\| = o_p(1), \quad \|\hat{\eta} - \eta_0\| = o_p(1), \quad \|\hat{\alpha} - \alpha_0\|_{L_2} = o_p(1), \quad \|\hat{\alpha} - \alpha_0\|_\infty = o_p(1).$$

Theorem 2. Under (C1)–(C4), if $1/(2\xi + 2) < \nu < 1/(2\xi)$ with $\xi > 1$, then

$$d(\hat{\theta}, \theta_0) = O_p(n^{-\min(\xi\nu, (1-\nu)/2)}).$$

In particular,

$$\|\hat{\alpha} - \alpha_0\|_{L_2} = O_p(n^{-\min(\xi\nu, (1-\nu)/2)}).$$

The optimal convergence rate $O_p(n^{-\xi/(1+2\xi)})$ is achieved, when $\nu = 1/(1 + 2\xi)$.

Theorem 3. Under (C1)–(C6),

$$\sqrt{n} \begin{pmatrix} \hat{\beta} - \beta_0 \\ \hat{\eta} - \eta_0 \end{pmatrix} \xrightarrow{d} N(\mathbf{0}, \Sigma),$$

where Σ is the inverse of the semiparametric efficient Fisher information matrix.

Proofs of the above theorems are deferred to the Appendix A. To perform statistical inference on the β and η , it is often necessary to estimate the asymptotic covariance matrix of $\hat{\beta}$ and $\hat{\eta}$. We adopt the nonparametric bootstrap ([30]) for this task, as it avoids the need for closed-form covariance expressions and offers straightforward implementation with demonstrated practical utility. Specifically, the bootstrap procedure consists of the following three steps. First, generate S resamples of size n from the original observed data

$\{O_i\}_{i=1}^n$, where S is a pre-specified value (e.g., $S = 200$). Second, for each s -th bootstrap sample, compute the corresponding estimators $\hat{\beta}_s^*$, $\hat{\eta}_s^*$, $\hat{\alpha}_s^*$, and $s = 1, 2, \dots, S$. Finally, use the empirical covariance matrix of $\{\hat{\beta}_s^*, \hat{\eta}_s^*\}_{s=1}^S$ as an approximation to the asymptotic covariance matrix of $\hat{\beta}$ and $\hat{\eta}$. For the function $\alpha(t)$, the 2.5% and 97.5% quantiles of the bootstrap replicates $\{\hat{\alpha}_s^*(t)\}_{s=1}^S$ at each fixed time point t . Simulation studies in the following section demonstrate the satisfactory performance of this bootstrap procedure in finite sample.

4. Simulation Studies

We generate data from the semiparametric probit mixture cure model with covariate matrices $\mathbf{X}$ and $\mathbf{Z}$:

$$S(t | \mathbf{X}, \mathbf{Z}) = 1 - \pi(\mathbf{Z}) + \pi(\mathbf{Z}) \times S_u(t | \mathbf{X}).$$

The covariate vectors are specified as $\mathbf{X} = (X_1, X_2)$ and $\mathbf{Z} = (1, Z_1, Z_2)$, with $X_1, Z_1 \sim U(0, 1)$ and $X_2, Z_2 \sim \text{Ber}(0.5)$. Multiple sets of true parameter values are considered for the coefficients β and η . The monotone function is set to $\alpha(t) = \log(t)$. For the cure indicator, we generate $\mathbf{u} = (u_1, \dots, u_n) \sim \text{Ber}(\pi(\mathbf{Z}))$, where $u_i = 1$ indicates a susceptible subject and $u_i = 0$ indicates a cured individual. For susceptible subjects, failure times are sampled from the probit model. For cured individuals, random observation intervals are generated, and the data are treated as right-censored. For susceptible individuals, censoring types are generated according to specific proportions, and the inspection times are generated from a uniform distribution. We conduct numerical experiments with sample sizes $n = 200$ and $n = 400$, each with 1000 replications. The simulation results for the EM algorithm under the semiparametric probit mixture cure model are presented. The reported results include sample size, Param, Bias, empirical standard deviations (StDev), average estimated standard errors (StErr), and (empirical coverage probability) CP.

Three different simulation settings for the regression coefficients are considered with fixed cure coefficients $\eta = (1, -1, 0)$. The simulation results are presented in Table 1. Comparing the results for sample sizes 200 and 400, we find that the empirical standard deviations and average estimated standard errors of η and β decrease significantly as the sample size increases. Similarly, with fixed regression coefficients $\beta = (1, -1)$, three different simulation settings for the cure coefficients are considered, and the results are presented in Table 2. As the sample size increases from 200 to 400, the biases, StDev, and StErr of η and β show a clear decreasing trend. In contrast, the empirical coverage probability for η_0 stays moderately below the nominal level regardless of the sample size. This may be due to the weak identifiability of the intercept term in the cure probability model under interval-censored data, or to the downward bias of the bootstrap standard errors under small samples. Nevertheless, as the sample size n increases from 200 to 400, the empirical standard deviation and the average estimated standard error of η_0 improve, indicating consistency of the estimator.

Table 1. Simulation results of the EM algorithm under finite samples with fixed cure coefficients.

Regression Coef.	Param	<i>n</i> = 200				<i>n</i> = 400			
		Bias	StDev	StErr	CP	Bias	StDev	StErr	CP
$\beta = (1, -1)$	η_0	−0.021	0.018	0.017	0.943	−0.021	0.017	0.016	0.933
	η_1	−0.067	0.203	0.238	0.942	−0.037	0.149	0.158	0.939
	η_2	0.063	0.201	0.235	0.946	0.035	0.148	0.156	0.942
	β_1	−0.013	0.168	0.168	0.961	−0.016	0.130	0.119	0.956
	β_2	0.015	0.167	0.167	0.954	0.020	0.126	0.119	0.950
$\beta = (1, 1)$	η_0	−0.021	0.019	0.017	0.943	−0.021	0.016	0.015	0.926
	η_1	−0.084	0.228	0.242	0.938	−0.038	0.156	0.156	0.950
	η_2	0.076	0.227	0.240	0.941	0.036	0.156	0.153	0.937
	β_1	0.020	0.186	0.171	0.943	0.022	0.125	0.120	0.953
	β_2	−0.021	0.178	0.169	0.958	−0.021	0.124	0.120	0.948
$\beta = (-1, -1)$	η_0	−0.022	0.018	0.017	0.939	−0.021	0.016	0.016	0.926
	η_1	−0.072	0.216	0.235	0.938	−0.038	0.155	0.155	0.937
	η_2	0.070	0.214	0.233	0.930	0.036	0.153	0.153	0.945
	β_1	0.022	0.182	0.167	0.954	0.013	0.121	0.116	0.956
	β_2	0.022	0.177	0.167	0.957	0.016	0.122	0.115	0.955
$\beta = (-1, 1)$	η_0	−0.021	0.019	0.017	0.940	−0.021	0.017	0.016	0.921
	η_1	−0.075	0.220	0.239	0.942	−0.036	0.151	0.156	0.949
	η_2	0.073	0.218	0.236	0.936	0.036	0.151	0.154	0.942
	β_1	−0.016	0.174	0.168	0.956	−0.019	0.128	0.120	0.953
	β_2	−0.013	0.164	0.166	0.956	−0.019	0.123	0.118	0.949

Furthermore, we set the parameters to $\beta = (1, -1)$ and $\eta = (0, 1, -1)$ and applied the EM algorithm to obtain the estimated function of $\alpha(t)$ under sample sizes 200 and 400, along with the corresponding 95% confidence intervals, as shown in Figures 1 and 2. The results indicate that under both sample sizes, the estimated mean curve of $\alpha(t)$ closely matches the true curve $\alpha(t) = \log(t)$. Compared to the results at sample size 200, the confidence intervals at sample size 400 are noticeably narrower, indicating improved estimation accuracy as the sample size increases.

Table 2. Simulation results of the EM algorithm under finite samples with fixed regression coefficients.

Cure Coef.	Param	<i>n</i> = 200				<i>n</i> = 400			
		Bias	StDev	StErr	CP	Bias	StDev	StErr	CP
$\eta = (0, 1, -1)$	η_0	−0.021	0.018	0.017	0.943	−0.021	0.017	0.016	0.933
	η_1	−0.067	0.203	0.238	0.942	−0.037	0.149	0.158	0.939
	η_2	0.063	0.201	0.235	0.946	0.035	0.148	0.156	0.942
	β_1	−0.013	0.168	0.168	0.961	−0.016	0.130	0.119	0.956
	β_2	0.015	0.167	0.167	0.954	0.020	0.126	0.119	0.950
$\eta = (0, -1, 1)$	η_0	−0.021	0.018	0.017	0.942	−0.021	0.016	0.016	0.928
	η_1	0.077	0.228	0.242	0.934	0.041	0.163	0.155	0.939
	η_2	−0.078	0.224	0.238	0.935	−0.039	0.161	0.153	0.944
	β_1	−0.016	0.182	0.169	0.956	−0.019	0.132	0.118	0.960
	β_2	0.020	0.175	0.167	0.963	0.023	0.128	0.117	0.954
$\eta = (0, -1, -1)$	η_0	−0.021	0.019	0.017	0.941	−0.020	0.016	0.016	0.933
	η_1	0.073	0.219	0.235	0.952	0.038	0.152	0.155	0.941
	η_2	0.076	0.216	0.231	0.936	0.038	0.150	0.153	0.948
	β_1	−0.014	0.172	0.167	0.957	−0.015	0.124	0.116	0.953
	β_2	0.020	0.176	0.168	0.957	0.017	0.125	0.116	0.956
$\eta = (0, 1, 1)$	η_0	−0.021	0.019	0.017	0.946	−0.021	0.016	0.016	0.927
	η_1	−0.073	0.218	0.241	0.937	−0.036	0.148	0.158	0.944
	η_2	−0.070	0.213	0.239	0.943	−0.035	0.146	0.156	0.946
	β_1	−0.011	0.163	0.168	0.968	−0.010	0.133	0.121	0.951
	β_2	0.012	0.162	0.166	0.969	0.013	0.129	0.119	0.958

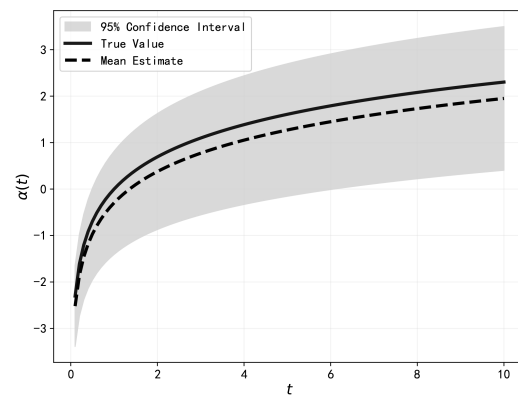


Figure 1. Estimation of $\alpha(t)$ from simulated interval-censored data (sample size $n = 200$).

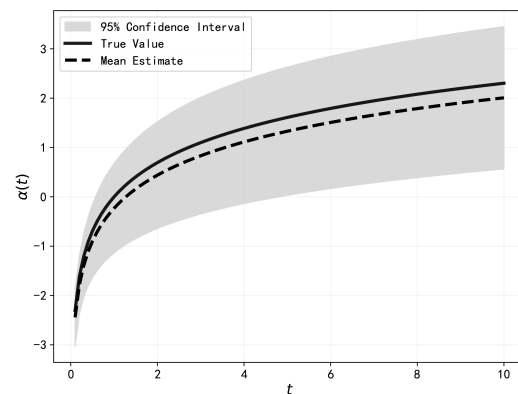


Figure 2. Estimation of $\alpha(t)$ from simulated interval-censored data (sample size $n = 400$).

5. Application

The real data used in this section were collected and continuously maintained by Xu et al. [31], who compiled a database of individual-level COVID-19 cases globally. This database aggregates publicly reported case information, including travel history, symptom onset dates, age, and gender. We extracted clinical information of COVID-19 patients in China from 1 December 2019 to 22 January 2020, setting the study starting date as 1 December 2019 (denoted as $t_0 = 0$). All dates were converted to days relative to this starting point. Since COVID-19 is transmitted through the respiratory tract, we define the patient's travel start time as $L_i = \text{date_L} - 2019/12/01$, and the time of symptom onset as $R_i = \text{date_R} - 2019/12/01$. Consequently, the time to symptom onset is the failure time T , which is the number of days from the study starting point to the first appearance of symptoms for individual i . Based on the recording of symptom information, the failure time is observed in the following forms:

1. If both L_i and R_i are known, the individual is interval-censored, and the exact time of symptom onset falls within $[L_i, R_i)$.
2. If L_i is unknown and R_i is known, the individual is left-censored, and the time of symptom onset falls within $[0, R_i)$.
3. If L_i is known and R_i is unknown, the individual is right-censored, and the time of symptom onset falls within $[R_i, \infty)$.

Data cleaning was performed as follows. First, rows with missing travel history (date_travel_history) and missing symptom onset date (date_onset_symptoms) were excluded. Then, rows with missing age or missing gender were removed. For the infection time, we ensured that it was either a fixed date or missing; the start of the travel period was taken as the infection time, and missing values were left as is. Furthermore, we ensured that both the infection time and the symptom onset date were greater than the study

starting date, and that the infection time preceded the symptom onset date; individuals not satisfying these conditions were excluded. To better fit the proposed semiparametric probit mixture cure model, we also required that the data exhibit a cure fraction. After the above screening, clinical data from 297 COVID-19 patients were finally collected. The Turnbull nonparametric survival curve in Figure 3 shows that the survival function approaches a non-zero level in the tail, indicating the presence of a cured subpopulation and satisfying the application conditions of the mixture cure model.

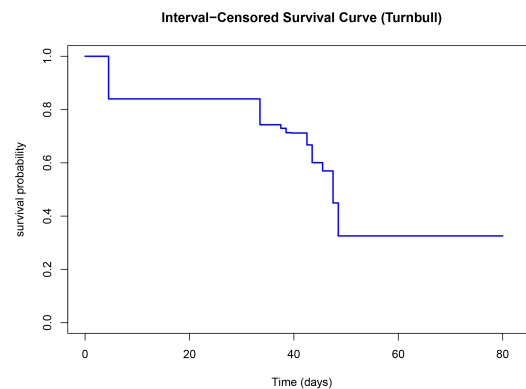


Figure 3. Turnbull nonparametric survival curve for the China patient dataset.

We applied the proposed semiparametric probit mixture cure model to the dataset of Chinese COVID-19 patients cohort. In this analysis, we assumed that the covariates for the latency period, X , and the covariates for the cure probability, Z , are identical, both including age (continuous variable) and gender (categorical variable). Specifically, $X_1 = Z_1$ denotes age, and $X_2 = Z_2$ denotes gender. Age is measured in years and used in its scale. Gender is coded as 1 for male and 0 for female. This specification allows us to simultaneously assess the dual effects of age and gender on both the time from infection to symptom onset and the cure probability. Table 3 summarizes the resulting parameter estimates for the COVID-19 patients.

The results show that $\beta_1 = 0.081$ (p value = 0.026), indicating a significant positive correlation between age and the time to symptom onset; that is, the older ages are associated with higher risk, based on the earlier time to symptom onset. This phenomenon may be related to age-related differences in immune responses, affecting the length of the incubation period. The estimate $\eta_1 = 1.645$ (p value = 0.021) indicates a positive correlation between age and cure probability. Both β_2 and η_2 are not significant (p value > 0.05), suggesting that gender has no significant effect on either the latency period or the cure probability. The estimate of η_0 is 4.844 ($p = 1.816 \times 10^{-5}$). Together with the fact that the survival curve in Figure 3 tends to a non-zero level in the tail, this further supports the existence of a cured subpopulation in the Chinese COVID-19 patient dataset. Subsequently, the semiparametric probit mixture cure model was used to obtain the estimated curve of $\alpha(t)$, as shown in Figure 4.

Table 3. Estimation results for the COVID-19 patient dataset.

Parameter	Covariate	Estimate	Std Error	p -Value
η_0	Intercept	4.844	1.130	1.816×10^{-5}
η_1	Age	1.645	0.715	0.021
η_2	Gender	0.938	0.611	0.125
β_1	Age	0.081	0.036	0.026
β_2	Gender	−0.014	0.031	0.653

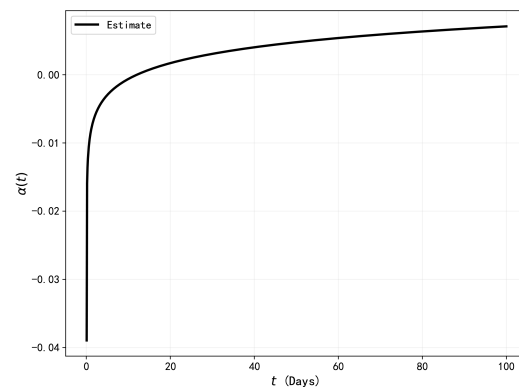


Figure 4. Estimation of $\alpha(t)$ for COVID-19 infected patients.

To further support the adequacy of this model specification, we compared the proposed probit model with the Weibull and log-normal models, for the survival function $S_u(t|x)$. The comparison results, presented in Table 4, show that the probit model yields substantially lower AIC and BIC values than both the Weibull and log-normal models, indicating that it provides a better fit to the observed interval-censored incubation data for the susceptible subpopulation.

Table 4. Model comparison for the survival function $S_u(t|x)$ of uncured patients.

Model	LogLike	AIC	BIC
probit	−4504.490	9020.979	9043.142
Weibull	−4559.117	9130.234	9152.397
log-norm	−8673.540	17,359.080	17,381.243

6. Discussion

This paper introduces a semiparametric probit mixture cure model for interval-censored data. By introducing two latent variables to simplify the likelihood function and employing B-splines with a logarithmic transformation to estimate the unspecified monotone function $\alpha(t)$, we developed a corresponding EM algorithm. The estimators enjoy the properties of consistency and asymptotic normality under the assumed regularity conditions. Numerical simulation studies demonstrated the favorable performance of the proposed method under finite-sample settings. Its practical value is further supported by an application to a real COVID-19 patient dataset from China.

Our empirical analysis of the Chinese COVID-19 patient cohort indicates that the advancing age is significantly associated with reduced incubation period, corroborating earlier reports in the literature. This finding is supported by related studies in the literature. For instance, Han et al. [32] compared the clinical characteristics of children and adults infected with COVID-19 and found that the incubation period in children was longer than that in adults, revealing that age factors affect the length of the incubation period. Hu et al. [33] inferred that the susceptibility to SARS-CoV-2 infection increases with age, further indicating that age plays an important role in the process of COVID-19 infection.

Several limitations should be acknowledged. First, the model assumes independence between examination times and failure times, which may not be fully satisfied in practice; the impact of violating this assumption should be assessed through sensitivity analyses. Second, during a finite observation period, right-censored individuals and cured individuals are difficult to distinguish completely in the likelihood function, and the identification of the cure probability relies to some extent on model assumptions. Third, this study only analyzed a dataset a cohort of Chinese COVID-19 patients with modest sample size; broader

validation across larger, multi-center populations is warranted to confirm the generalizability of our findings. Additionally, while a systematic simulation comparison with alternative mixture cure model specifications is beyond the scope of this work, previous studies, such as Pedrosa-Laza et al. [34], have shown that different model choices can lead to divergent conclusions, highlighting the need for such comparisons in future investigations.

In implementing the gradient descent step for updating γ , we recommend the following practical guidelines. First, when covariates are measured on substantially different scales, standardizing continuous covariates prior to analysis can improve the conditioning of the objective function and stabilize gradient descent. Second, if the log-likelihood exhibits non-monotonic behavior or parameter estimates oscillate during iterations, reducing the initial learning rate to 10^{-3} or 10^{-4} is advised. Third, if convergence is excessively slow (exceeding 200 iterations), increasing the initial learning rate to 0.05 or incorporating a momentum term may accelerate convergence. Finally, for large-scale datasets (e.g., $n \geq 1000$ or $K \geq 20$ basis functions), mini-batch gradient descent variants can be considered to reduce computational cost, though this is beyond the scope of the current work.

Future research directions include the following. First, the model can be extended to more flexible settings, such as relaxing the independence assumption between examination times and failure times or incorporating time-varying covariates to capture dynamic infection risks. Second, Bayesian computational methods can be explored to achieve efficient Gibbs sampling through latent variable augmentation, thereby improving estimation stability under small sample sizes. Third, the proposed model framework can be generalized to estimate the incubation periods of other infectious diseases, such as influenza, dengue fever, and other emerging infectious diseases that exhibit a cure structure.

In summary, the proposed semiparametric probit mixture cure model provides a flexible and theoretically rigorous tool for estimating the incubation period based on interval-censored incubation data. When applied to Chinese COVID-19 data, it uncovers a significant age-incubation relation, with older subjects exhibiting shorter delays. These findings add to the growing evidence on age-dependent COVID-19 manifestations and offer a useful framework for subsequent incubation period analyses in other infectious diseases.

Author Contributions: Conceptualization, N.Z. and S.Z.; Methodology, N.Z.; Software, N.Z.; Validation, N.Z.; Writing — original draft, N.Z.; Writing — review & editing, S.Z. and D.L.; Visualization, S.Z. and D.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data that support the findings in this paper are not publicly available due to privacy restrictions.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

EM	Expectation-maximization
COVID-19	novel coronavirus infection
BETS	Beginning End Transmission Symptom
Param	Parameter
StDev	empirical standard deviation
StErr	average estimated standard error
CP	empirical coverage probability

Appendix A

In this section, we provide detailed proofs of Theorems 1–3. First, define the empirical measure $\mathbb{P}_n$ as the empirical distribution of n independent observations, the true probability measure by $\mathbb{P}$, and the empirical process by $\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - \mathbb{P})$. For any measurable function f and a random variable $\mathbf{Z}$ with distribution function F , we have

$$\mathbb{P}f = \int f(\mathbf{z})dF(\mathbf{z}), \quad \mathbb{P}_nf = \frac{1}{n} \sum_{i=1}^n f(\mathbf{Z}_i), \quad \mathbb{G}_nf = \sqrt{n}(\mathbb{P}_nf - \mathbb{P}f),$$

where $\mathbf{Z}_1, \mathbf{Z}_2, \dots, \mathbf{Z}_n$ are n independent copies of $\mathbf{Z}$. For a measurable function class $\mathcal{F}$, define $\|\mathbb{G}_n\|_{\mathcal{F}} = \sup_{f \in \mathcal{F}} |\mathbb{G}_nf|$. Let $\mathcal{M} := \mathcal{M}(\mathcal{T}_n, d_M)$ denote the space of monotone spline functions based on spline degree d_M and interior knot set $\mathcal{T}_n$.

The observed data are $O_i = (L_i, R_i, \mathbf{X}_i, \mathbf{Z}_i, \delta_{L,i}, \delta_{I,i}, \delta_{R,i})$, which are independent and identically distributed. Write $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \boldsymbol{\eta}^\top, \alpha)^\top$ and let $\boldsymbol{\theta}_0 = (\boldsymbol{\beta}_0^\top, \boldsymbol{\eta}_0^\top, \alpha_0)^\top$ be the true parameter value satisfying Condition 1. The log-likelihood function is $\ell(\boldsymbol{\theta}) = \sum_{i=1}^n \ell_i(\boldsymbol{\theta})$. Denote $\mathbb{M}(\boldsymbol{\theta}) = \mathbb{P}\ell_i(\boldsymbol{\theta})$ as the expected log-likelihood and $\mathbb{M}_n(\boldsymbol{\theta}) = \mathbb{P}_n\ell_i(\boldsymbol{\theta})$ as the empirical log-likelihood.

For notational simplicity, define $\mu_{L,i} = \alpha(L_i) + \boldsymbol{\beta}^\top \mathbf{X}_i$, $\mu_{R,i} = \alpha(R_i) + \boldsymbol{\beta}^\top \mathbf{X}_i$, $\Phi_{L,i} = \Phi(\mu_{L,i})$, $\Phi_{R,i} = \Phi(\mu_{R,i})$, and $\pi_i = \pi(\mathbf{Z}_i) = \exp(\boldsymbol{\eta}^\top \mathbf{Z}_i) / (1 + \exp(\boldsymbol{\eta}^\top \mathbf{Z}_i))$, where $\Phi(\cdot)$ is the standard normal distribution function. The log-likelihood contribution from a single observation is

$$\begin{aligned} \ell(\boldsymbol{\theta}) &= \delta_L \log(\pi \cdot \Phi_R) + \delta_I \log(\pi[\Phi_R - \Phi_L]) \\ &\quad + \delta_R \log(1 - \pi + \pi(1 - \Phi_L)). \end{aligned}$$

Proof of Theorem 1. The consistency of $\hat{\boldsymbol{\theta}}$ is established by verifying the three conditions of Theorem 5.7 in van der Vaart [35].

Condition 2: $\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} |\mathbb{M}_n(\boldsymbol{\theta}) - \mathbb{M}(\boldsymbol{\theta})| = o_p(1)$.

By Condition 2, the covariates $\mathbf{X}$ and $\mathbf{Z}$ are bounded, and the sets $\{\boldsymbol{\beta}^\top \mathbf{X} : \boldsymbol{\beta} \in \mathcal{B}\}$ and $\{\boldsymbol{\eta}^\top \mathbf{Z} : \boldsymbol{\eta} \in \mathcal{H}\}$ are compact. Consequently, their covering numbers are $M(1/\varepsilon)^{p_\beta}$ and $M(1/\varepsilon)^{p_\eta}$, respectively. For the monotone function class $\mathcal{A} = \{\alpha(t) : \alpha \text{ is increasing, } t \in [\rho, \tau]\}$, by Theorem 2.7.5 of Van der Vaart and Wellner [36], its ε -bracketing number satisfies

$$N_{[]}(\varepsilon, \mathcal{A}, \|\cdot\|_{L_2}) \leq Me^{O(1/\varepsilon)},$$

hence $\mathcal{A}$ is a Donsker class.

For any $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \boldsymbol{\Theta}$, by the Lipschitz property of the log-likelihood function $\ell(\boldsymbol{\theta})$ (smoothness in parameters; bounded covariates; compact parameter space), there exists a constant $M > 0$ such that

$$|\ell(\boldsymbol{\theta}_1) - \ell(\boldsymbol{\theta}_2)| \leq M(\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\| + \|\boldsymbol{\eta}_1 - \boldsymbol{\eta}_2\| + \|\alpha_1 - \alpha_2\|_{L_2}).$$

Therefore, the ε -bracketing number of the class $\{\ell(\boldsymbol{\theta}) : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ is bounded by $M(1/\varepsilon)^{p_\beta + p_\eta} e^{O(1/\varepsilon)}$. This class is thus a Donsker class and, consequently, a Glivenko–Cantelli class. Hence,

$$\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} |\mathbb{M}_n(\boldsymbol{\theta}) - \mathbb{M}(\boldsymbol{\theta})| = o_p(1).$$

Condition 3: For any $\delta > 0$, $\sup_{\boldsymbol{\theta}: d(\boldsymbol{\theta}, \boldsymbol{\theta}_0) \geq \delta} \mathbb{M}(\boldsymbol{\theta}) \leq \mathbb{M}(\boldsymbol{\theta}_0)$.

Define $m(x) = x \log x - x + 1$, which attains its minimum at $x = 1$ and satisfies $(x - 1)^2 \geq m(x) \geq (x - 1)^2/4$. Then we have

$$\begin{aligned} \mathbb{M}(\boldsymbol{\theta}_0) - \mathbb{M}(\boldsymbol{\theta}) = E \left[\pi_0 \Phi_{R,0} m\left(\frac{\pi \Phi_R}{\pi_0 \Phi_{R,0}}\right) + \pi_0 (\Phi_R - \Phi_L) m\left(\frac{\pi(\Phi_R - \Phi_L)}{\pi_0(\Phi_{R,0} - \Phi_{L,0})}\right) \right. \\ \left. + (1 - \pi_0 + \pi_0(1 - \Phi_{L,0})) m\left(\frac{1 - \pi + \pi(1 - \Phi_L)}{1 - \pi_0 + \pi_0(1 - \Phi_{L,0})}\right) \right], \end{aligned}$$

where the subscript 0 denotes evaluation at the true parameter $\boldsymbol{\theta}_0$.

By Taylor expansion and Conditions 1-3, there exists a constant $M_1 > 0$ such that

$$\mathbb{M}(\boldsymbol{\theta}_0) - \mathbb{M}(\boldsymbol{\theta}) \geq M_1 E[(\pi - \pi_0)^2 + (\Phi_L - \Phi_{L,0})^2 + (\Phi_R - \Phi_{R,0})^2].$$

From the Lipschitz property of the standard normal distribution function Φ and $\mu = \alpha(t) + \boldsymbol{\beta}^\top \mathbf{X}$, there exist constants $M_2, M_3 > 0$ such that

$$E[(\Phi_L - \Phi_{L,0})^2] \geq M_2 E[(\mu_L - \mu_{L,0})^2] \geq M_3 \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|^2 + M_3 \|\alpha - \alpha_0\|_{L_2}^2.$$

From the properties of logistic regression, there exists a constant $M_4 > 0$ such that

$$E[(\pi - \pi_0)^2] \geq M_4 \|\boldsymbol{\eta} - \boldsymbol{\eta}_0\|^2.$$

Thus, there exists a constant $M_5 > 0$ such that

$$\mathbb{M}(\boldsymbol{\theta}_0) - \mathbb{M}(\boldsymbol{\theta}) \geq M_5 d^2(\boldsymbol{\theta}, \boldsymbol{\theta}_0).$$

Therefore, for any $\delta > 0$,

$$\mathbb{M}(\boldsymbol{\theta}_0) - \sup_{\boldsymbol{\theta}: d(\boldsymbol{\theta}, \boldsymbol{\theta}_0) \geq \delta} \mathbb{M}(\boldsymbol{\theta}) \geq M_5 \delta^2 > 0,$$

i.e.,

$$\sup_{\boldsymbol{\theta}: d(\boldsymbol{\theta}, \boldsymbol{\theta}_0) \geq \delta} \mathbb{M}(\boldsymbol{\theta}) \leq \mathbb{M}(\boldsymbol{\theta}_0) - M_5 \delta^2 \leq \mathbb{M}(\boldsymbol{\theta}_0).$$

Condition 4: There exists $\boldsymbol{\theta}_{0,n} \rightarrow \boldsymbol{\theta}_0$ such that $\mathbb{M}_n(\boldsymbol{\theta}_{0,n}) - \mathbb{M}_n(\boldsymbol{\theta}_0) = o_p(1)$.

By Condition 4 and Corollary 6.21 of Schumaker [37] and Lemma 1 of Lu et al. [38], there exists a spline function $\alpha_{0,n} \in \mathcal{M}_n$ such that

$$\|\alpha_{0,n} - \alpha_0\|_\infty = O(n^{-\xi\nu}),$$

where ξ is defined in Condition 3 and ν satisfies Condition 4. Define $\boldsymbol{\theta}_{0,n} = (\boldsymbol{\beta}_0, \boldsymbol{\eta}_0, \alpha_{0,n})$. Clearly, $d(\boldsymbol{\theta}_{0,n}, \boldsymbol{\theta}_0) \rightarrow 0$ as $n \rightarrow \infty$. By the Lipschitz property of $\ell(\boldsymbol{\theta})$ with respect to α , there exists a constant $C > 0$ such that

$$|\ell(\boldsymbol{\theta}_{0,n}) - \ell(\boldsymbol{\theta}_0)| \leq C \|\alpha_{0,n} - \alpha_0\|_\infty = O(n^{-\xi\nu}).$$

Hence,

$$|\mathbb{M}_n(\boldsymbol{\theta}_{0,n}) - \mathbb{M}_n(\boldsymbol{\theta}_0)| \leq \frac{1}{n} \sum_{i=1}^n |\ell_i(\boldsymbol{\theta}_{0,n}) - \ell_i(\boldsymbol{\theta}_0)| \leq C \cdot O(n^{-\xi\nu}) = O(n^{-\xi\nu}) = o_p(1),$$

since $\xi\nu > 0$.

By Theorem 5.7 [35], $\hat{\theta}_n \xrightarrow{P} \theta_0$, i.e.,

$$d(\hat{\theta}_n, \theta_0) = o_p(1).$$

In particular, $\|\hat{\beta} - \beta_0\| = o_p(1)$, $\|\hat{\eta} - \eta_0\| = o_p(1)$, and $\|\hat{\alpha} - \alpha_0\|_{L_2} = o_p(1)$. From Condition 4, the maximum knot spacing is $\Delta = O(n^{-\xi\nu})$, and by properties of spline functions, we further obtain $\|\hat{\alpha} - \alpha_0\|_\infty = o_p(1)$.

Thus, Theorem 1 is proved. $\square$

Proof of Theorem 2. We prove the convergence rate $d(\hat{\theta}, \theta_0) = O_p(n^{-\min(\xi\nu, (1-\nu)/2)})$.

By Condition 1, $\mathcal{L}$ is a Donsker class, so

$$\sup_{\theta \in \Theta} |\mathbb{M}_n(\theta) - \mathbb{M}(\theta)| = O_p(n^{-1/2}).$$

By Condition 2, there exists a constant $M_5 > 0$ such that

$$\mathbb{M}(\theta_0) - \mathbb{M}(\theta) \geq M_5 d^2(\theta, \theta_0).$$

By Condition 3, there exists $\theta_{0,n}$ satisfying $d(\theta_{0,n}, \theta_0) = O(n^{-\xi\nu})$ and

$$\mathbb{M}_n(\theta_{0,n}) - \mathbb{M}_n(\theta_0) = O(n^{-\xi\nu}).$$

Since $\hat{\theta}_n$ maximizes $\mathbb{M}_n$, we have $\mathbb{M}_n(\hat{\theta}_n) \geq \mathbb{M}_n(\theta_{0,n})$. Therefore,

$$\mathbb{M}_n(\hat{\theta}_n) - \mathbb{M}_n(\theta_0) \geq \mathbb{M}_n(\theta_{0,n}) - \mathbb{M}_n(\theta_0).$$

Expanding the left-hand side and rearranging yields

$$M_5 d^2(\hat{\theta}_n, \theta_0) \leq O_p(n^{-\xi\nu}) + O_p(n^{-1/2}) = O_p(n^{-\min(\xi\nu, 1/2)}). \quad (4)$$

By consistency, there exists $\delta_n \rightarrow 0$ such that $d(\hat{\theta}_n, \theta_0) \leq \delta_n$ with probability tending to 1. In a local neighborhood, the dimension of the monotone spline function space $\mathcal{A}_n$ is $K = O(n^\nu)$. According to Theorem 1 of van der Vaart and Wellner (1996), the convergence rate δ_n satisfies $\phi_n(\delta_n) \asymp \delta_n^2$, where $\phi_n(\delta)$ is the modulus of the empirical process. For a finite-dimensional parameter space of dimension K ,

$$\phi_n(\delta) \asymp \frac{1}{\sqrt{n}} \int_{c\delta^2}^{\delta} \sqrt{K} d\varepsilon \asymp \sqrt{\frac{K}{n}} \cdot \delta.$$

Substituting $K \asymp n^\nu$ gives $\phi_n(\delta) \asymp n^{-(1-\nu)/2} \delta$. Plugging into the equation $n^{-(1-\nu)/2} \delta \asymp \delta^2$ yields

$$\delta \asymp n^{-(1-\nu)/2}.$$

The convergence rate is determined by the slower of the two rates:

$$d(\hat{\theta}, \theta_0) = O_p(n^{-\min(\xi\nu, (1-\nu)/2)}).$$

When $\xi\nu = (1-\nu)/2$, the two terms balance, giving $\nu = 1/(1+2\xi)$, and the optimal convergence rate $n^{-\xi/(1+2\xi)}$ is achieved.

Thus, Theorem 2 is proved. $\square$

Proof of Theorem 3. Let $\phi(\cdot) = \Phi'(\cdot)$ denote the standard normal density function. Define the following auxiliary functions:

$$\begin{aligned}
 A &= \frac{\phi(\mu_R)}{\Phi(\mu_R)}, \\
 B &= \frac{\phi(\mu_R) - \phi(\mu_L)}{\Phi(\mu_R) - \Phi(\mu_L)}, \\
 D &= \frac{-\pi\phi(\mu_L)}{1 - \pi + \pi(1 - \Phi(\mu_L))},
 \end{aligned}$$

where $\mu_L = \alpha(L) + \beta^\top X$, $\mu_R = \alpha(R) + \beta^\top X$, and $\pi = \exp(\eta^\top Z) / [1 + \exp(\eta^\top Z)]$ is the susceptibility probability. Based on these auxiliary functions, the score functions for a single observation are

$$\begin{aligned}
 \dot{\ell}_\beta(\theta) &= [\delta_L A + \delta_I B + \delta_R D] X, \\
 \dot{\ell}_\eta(\theta) &= [\delta_L(1 - \pi) + \delta_I(1 - \pi) + \delta_R \frac{-\pi(1 - \pi)\Phi_{L_I}}{1 - \pi + \pi(1 - \Phi(\mu_L))}] Z.
 \end{aligned}$$

Let $\phi = (\beta^\top, \eta^\top)^\top$. Consider the submodel $\alpha_\tau(t) = \alpha(t) + \tau q(t)$, where $q \in L_2([\rho, \tau])$ is a perturbation direction and τ is a scalar parameter. The score function in the direction of α is defined as the directional derivative:

$$\dot{\ell}_\alpha(\theta)[q] = [\delta_L A + \delta_I B + \delta_R D] \cdot q(t).$$

According to semiparametric theory (Bickel et al., 1993 [39]), the efficient score function for ϕ is

$$\ell_\phi^* = \dot{\ell}_\phi - \dot{\ell}_\alpha[q^*],$$

where $q^* \in L_2^{p_\beta + p_\eta}([\rho, \tau])$ is a vector-valued function (each component is an L_2 function) such that ℓ_ϕ^* is orthogonal to $\dot{\ell}_\alpha[q]$ for all q , i.e.,

$$E[\ell_\phi^* \dot{\ell}_\alpha[q]] = 0, \quad \forall q \in L_2([\rho, \tau]).$$

This orthogonality condition is equivalent to the Fredholm integral equation

$$E[\dot{\ell}_\phi \dot{\ell}_\alpha[q]] = E[\dot{\ell}_\alpha[q^*] \dot{\ell}_\alpha[q]], \quad \forall q \in L_2([\rho, \tau]).$$

By Conditions 3 (smoothness) and 6 (identifiability), this equation has a unique solution q^* . Analogous to the proof of Theorem 6.1 in Huang and Rossini [40], q^* has bounded variation and can therefore be approximated by spline functions. To apply Theorem 3 of Zhang et al. [41], we need to verify

$$\mathbb{P}_n\{\dot{\ell}_\alpha(\hat{\theta})[q^*]\} = o_p(n^{-1/2}).$$

By spline approximation, according to Condition 4 (knot number $K = O(n^\nu)$, maximum spacing $\Delta = O(n^{-\nu})$), there exists a spline function vector $q_n \in \mathcal{M}_n^{p_\beta + p_\eta}$ such that $\|q_n - q^*\|_{L_2} = O(n^{-\nu})$. Since $\hat{\theta}$ is the sieve maximum likelihood estimator, for any $q_n \in \mathcal{M}_n^{p_\beta + p_\eta}$ we have $\mathbb{P}_n\{\dot{\ell}_\alpha(\hat{\theta})[q_n]\} = 0$. Hence,

$$\mathbb{P}_n\{\dot{\ell}_\alpha(\hat{\theta})[q^*]\} = \mathbb{P}_n\{\dot{\ell}_\alpha(\hat{\theta})[q^* - q_n]\}.$$

Decompose this as

$$\begin{aligned}
 \mathbb{P}_n\{\dot{\ell}_\alpha(\hat{\theta})[q^* - q_n]\} &= (\mathbb{P}_n - \mathbb{P})\{\dot{\ell}_\alpha(\hat{\theta})[q^* - q_n]\} \\
 &\quad + \mathbb{P}\{(\dot{\ell}_\alpha(\hat{\theta}) - \dot{\ell}_\alpha(\theta_0))[q^* - q_n]\}.
 \end{aligned}$$

For the first term, the class $\{\dot{\ell}_\alpha[\mathbf{q}^* - \mathbf{q}_n]\}$ is Donsker (guaranteed by Conditions 1–4), and $\|\mathbf{q}_n - \mathbf{q}^*\|_{L_2} = O(n^{-\nu})$. By Lemma 19.24 [35],

$$(\mathbb{P}_n - \mathbb{P})\{\dot{\ell}_\alpha[\mathbf{q}^* - \mathbf{q}_n]\} = o_p(n^{-1/2}).$$

For the second term, by the Cauchy–Schwarz inequality,

$$|\mathbb{P}\{(\dot{\ell}_\alpha(\hat{\boldsymbol{\theta}}) - \dot{\ell}_\alpha(\boldsymbol{\theta}_0))[\mathbf{q}^* - \mathbf{q}_n]\}| \leq Md(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}_0)\|\mathbf{q}^* - \mathbf{q}_n\|_{L_2}.$$

From Theorem 1, $d(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}_0) = O_p(n^{-\xi/(1+2\xi)})$ and $\|\mathbf{q}^* - \mathbf{q}_n\|_{L_2} = O(n^{-\nu})$. For $\nu = 1/(1+2\xi)$,

$$\frac{\xi}{1+2\xi} + \nu = \frac{\xi}{1+2\xi} + \frac{1}{1+2\xi} = \frac{1+\xi}{1+2\xi} > \frac{1}{2} \quad \text{for all } \xi \geq 1.$$

Thus, $\mathbb{P}_n\{\dot{\ell}_\alpha(\hat{\boldsymbol{\theta}})[\mathbf{q}^*]\} = o_p(n^{-1/2})$.

By the definition of $\hat{\boldsymbol{\theta}}$, $\mathbb{P}_n\{\ell_\phi(\hat{\boldsymbol{\theta}})\} = 0$. Combining with the decomposition of the efficient score function,

$$0 = \mathbb{P}_n\{\dot{\ell}_\phi(\hat{\boldsymbol{\theta}})\} = \mathbb{P}_n\{\ell_\phi^*(\hat{\boldsymbol{\theta}})\} + \mathbb{P}_n\{\dot{\ell}_\alpha(\hat{\boldsymbol{\theta}})[\mathbf{q}^*]\},$$

so $\mathbb{P}_n\{\ell_\phi^*(\hat{\boldsymbol{\theta}})\} = o_p(n^{-1/2})$.

Expand $\ell_\phi^*(\hat{\boldsymbol{\theta}})$ around $\boldsymbol{\theta}_0$ using a first-order Taylor expansion:

$$\ell_\phi^*(\hat{\boldsymbol{\theta}}) = \ell_\phi^*(\boldsymbol{\theta}_0) + \dot{\ell}_\phi^*(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + o_p(\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|).$$

Taking expectations,

$$\mathbb{P}_n\{\ell_\phi^*(\boldsymbol{\theta}_0)\} + \mathbb{P}\{\dot{\ell}_\phi^*(\boldsymbol{\theta}_0)\}(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}_0) + o_p(\|\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}_0\|) = o_p(n^{-1/2}),$$

where $\hat{\boldsymbol{\phi}} = (\hat{\boldsymbol{\beta}}^T, \hat{\boldsymbol{\eta}}^T)^T$.

By the definition of the information matrix, the efficient Fisher information matrix $I(\boldsymbol{\phi}_0) = E[\dot{\ell}_\phi^*(\boldsymbol{\theta}_0)\dot{\ell}_\phi^*(\boldsymbol{\theta}_0)^T]$ is positive definite, and the information equality gives

$$\mathbb{P}\{\dot{\ell}_\phi^*(\boldsymbol{\theta}_0)\} = -I(\boldsymbol{\phi}_0).$$

Furthermore, by the martingale central limit theorem (Hall and Heyde [42]), the score function satisfies

$$\sqrt{n}\mathbb{P}_n\{\ell_\phi^*(\boldsymbol{\theta}_0)\} \xrightarrow{d} N(0, I(\boldsymbol{\phi}_0)).$$

Thus,

$$\sqrt{n}(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}_0) = I(\boldsymbol{\phi}_0)^{-1}\sqrt{n}\mathbb{P}_n\{\ell_\phi^*(\boldsymbol{\theta}_0)\} + o_p(1).$$

By Slutsky's theorem,

$$\sqrt{n}(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}_0) \xrightarrow{d} N(0, I(\boldsymbol{\phi}_0)^{-1}).$$

Partition $I(\boldsymbol{\phi}_0)$ as

$$I(\boldsymbol{\phi}_0) = \begin{pmatrix} I_{\beta\beta} & I_{\beta\eta} \\ I_{\eta\beta} & I_{\eta\eta} \end{pmatrix}.$$

By the block matrix inversion formula, the submatrix of $I(\boldsymbol{\phi}_0)^{-1}$ corresponding to $\boldsymbol{\eta}$ is

$$\Sigma_\eta = (I_{\eta\eta} - I_{\eta\beta}I_{\beta\beta}^{-1}I_{\beta\eta})^{-1}.$$

Therefore, the marginal asymptotic distribution of η is

$$\sqrt{n}(\hat{\eta}_n - \eta_0) \xrightarrow{d} N(0, \Sigma_\eta).$$

Thus, Theorem 3 is proved. $\square$

References

1. Guan, W.J.; Ni, Z.Y.; Hu, Y.; Liang, W.H.; Ou, C.Q.; He, J.X.; Liu, L.; Shan, H.; Lei, C.L.; Hui, D.S.C.; et al. Clinical characteristics of coronavirus disease 2019 in China. *N. Engl. J. Med.* **2020**, *382*, 1708–1720.
2. Backer, J.A.; Klinkenberg, D.; Wallinga, J. Incubation period of 2019 novel coronavirus (2019-nCoV) infections among travellers from Wuhan, China, 20–28 January 2020. *Eurosurveillance* **2020**, *25*, 2000062.
3. Li, Q.; Guan, X.; Wu, P.; Wang, X.; Zhou, L.; Tong, Y.; Ren, R.; Leung, K.S.M.; Lau, E.H.Y.; Wong, J.Y.; et al. Early transmission dynamics in Wuhan, China, of novel coronavirus-infected pneumonia. *N. Engl. J. Med.* **2020**, *382*, 1199–1207.
4. Zhang, J.; Litvinova, M.; Wang, W.; Wang, Y.; Deng, X.; Chen, X.; Li, M.; Zheng, W.; Yi, L.; Chen, X.; et al. Evolving epidemiology and transmission dynamics of coronavirus disease 2019 outside Hubei province, China: A descriptive and modelling study. *Lancet Infect. Dis.* **2020**, *20*, 793–802.
5. Lauer, S.A.; Grantz, K.H.; Bi, Q.; Jones, F.K.; Zheng, Q.; Meredith, H.R.; Azman, A.S.; Reich, N.G.; Lessler, J. The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: Estimation and application. *Ann. Intern. Med.* **2020**, *172*, 577–582.
6. Linton, N.M.; Kobayashi, T.; Yang, Y.; Hayashi, K.; Akhmetzhanov, A.R.; Jung, S.M.; Yuan, B.; Kinoshita, R.; Nishiura, H. Incubation period and other epidemiological characteristics of 2019 novel coronavirus infections with right truncation: A statistical analysis of publicly available case data. *J. Clin. Med.* **2020**, *9*, 538.
7. Deng, Y.; You, C.; Liu, Y.; Qin, J.; Zhou, X.H. Estimation of incubation period and generation time based on observed length-biased epidemic cohort with censoring for COVID-19 outbreak in China. *Biometrics* **2021**, *77*, 929–941.
8. Qin, J.; You, C.; Lin, Q.; Hu, T.; Yu, S.; Zhou, X.H. Estimation of incubation period distribution of COVID-19 using disease onset forward time: A novel cross-sectional and forward follow-up study. *Sci. Adv.* **2020**, *6*, eabc1202.
9. Wong, K.Y.; Zhou, Q.; Hu, T. Semiparametric regression analysis of doubly-censored data with applications to incubation period estimation. *Lifetime Data Anal.* **2022**, *29*, 87.
10. Gressani, O.; Torneri, A.; Hens, N.; Faes, C. Flexible Bayesian estimation of incubation times. *Am. J. Epidemiol.* **2025**, *194*, 490–501.
11. Groeneboom, P. Estimation of the incubation time distribution for COVID-19. *Stat. Neerl.* **2021**, *75*, 161–179.
12. Arntzen, V.H.; Fiocco, M.; Leitzinger, N.; Geskus, R.B. Towards robust and accurate estimates of the incubation time distribution, with focus on upper tail probabilities and SARS-CoV-2 infection. *Stat. Med.* **2023**, *42*, 2341–2360.
13. Mizumoto, K.; Kagaya, K.; Zarebski, A.; Chowell, G. Estimating the asymptomatic proportion of coronavirus disease 2019 (COVID-19) cases on board the Diamond Princess cruise ship, Yokohama, Japan, 2020. *Eurosurveillance* **2020**, *25*, 2000180. Erratum in *Eurosurveillance* **2020**, *25*, 20200604c.
14. Boag, J.W. Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *J. R. Stat. Soc. Ser. B* **1949**, *11*, 15–53.
15. Berkson, J.; Gage, R.P. Survival curve for cancer patients following treatment. *J. Am. Stat. Assoc.* **1952**, *47*, 501–515.
16. Farewell, V.T. The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics* **1982**, 1041–1046.
17. Pal, S.; Peng, Y.; Aselisewine, W. A new approach to modeling the cure rate in the presence of interval censored data. *Comput. Stat.* **2023**, *39*, 2743.
18. Kuk, A.Y.; Chen, C.H. A mixture model combining logistic regression with proportional hazards regression. *Biometrika* **1992**, *79*, 531–541.
19. Peng, Y.; Dear, K.B.G. A nonparametric mixture model for cure rate estimation. *Biometrics* **2000**, *56*, 237–243.
20. Sy, J.P.; Taylor, J.M. Estimation in a Cox proportional hazards cure model. *Biometrics* **2000**, *56*, 227–236.
21. Lou, Y.; Du, M.; Wang, P.; Song, X. Semiparametric regression analysis of interval-censored failure time data with a cure subgroup and nonignorable missing covariates. *Stat. Methods Med. Res.* **2025**, *34*, 1994–2010.
22. Liu, X.; Xiang, L. Bayesian analysis of doubly semiparametric mixture cure models with interval-censored data. *Stat. Comput.* **2025**, *35*, 71.
23. Pan, C.; Cai, B.; Sui, X. A Bayesian proportional hazards mixture cure model for interval-censored data. *Lifetime Data Anal.* **2024**, *30*, 327.
24. Pal, S.; Barui, S. Likelihood-based inference for semi-parametric transformation cure models with interval censored data. *Commun. Stat.-Simul. Comput.* **2025**, *54*, 4780–4797.
25. Bliss, C.I. The method of probits. *Science* **1934**, *79*, 38–39.

26. Lin, X.; Wang, L. A semiparametric probit model for case 2 interval-censored failure time data. *Stat. Med.* **2010**, *29*, 972–981.
27. Armijo, L. Minimization of functions having Lipschitz continuous first partial derivatives. *Pac. J. Math.* **1966**, *16*, 1–3.
28. Dempster, A.P.; Laird, N.M.; Rubin, D.B. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B* **1977**, *39*, 1–22.
29. Wu, C.F.J. On the convergence properties of the EM algorithm. *Ann. Stat.* **1983**, 95–103.
30. Efron, B. Nonparametric estimates of standard error: The jackknife, the bootstrap and other methods. *Biometrika* **1981**, *68*, 589–599.
31. Xu, B.; Gutierrez, B.; Mekaru, S.; Sewalk, K.; Goodwin, L.; Loskill, A.; Cohn, E.L.; Hswen, Y.; Hill, S.C.; Cobo, M.M.; et al. Epidemiological data from the COVID-19 outbreak, real-time case information. *Sci. Data* **2020**, *7*, 106.
32. Han, Y.; Feng, Z.; Sun, L.; Ren, X.; Wang, H.; Xue, Y.; Wang, Y.; Fang, Y. A comparative-descriptive analysis of clinical characteristics in 2019-coronavirus-infected children and adults. *J. Med. Virol.* **2020**, *92*, 1596–1602.
33. Hu, S.; Wang, W.; Wang, Y.; Litvinova, M.; Luo, K.; Ren, L.; Sun, Q.; Chen, X.; Zeng, G.; Li, J.; et al. Infectivity, susceptibility, and risk factors associated with SARS-CoV-2 transmission under intensive contact tracing in Hunan, China. *Nat. Commun.* **2021**, *12*, 1533.
34. Pedrosa-Laza, M.; López-Cheda, A.; Cao, R. Cure models to estimate time until hospitalization due to COVID-19: A case study in Galicia (NW Spain). *Appl. Intell.* **2022**, *52*, 794–807.
35. Van der Vaart, A.W. *Asymptotic Statistics*; Cambridge University Press: Cambridge, UK, 2000.
36. Van der Vaart, A.W.; Wellner, J.A. *Weak Convergence and Empirical Processes: With Applications to Statistics*; Springer: New York, NY, USA, 1996.
37. Schumaker, L. *Spline Functions: Basic Theory*; Cambridge University Press: Cambridge, UK, 2007.
38. Lu, M.; Zhang, Y.; Huang, J. Estimation of the mean function with panel count data using monotone polynomial splines. *Biometrika* **2007**, *94*, 705–718.
39. Bickel, P.J.; Klaassen, C.A.; Bickel, P.J.; Ritov, Y.A.; Klaassen, J.; Wellner, J.A.; Ritov, Y.A. *Efficient and Adaptive Estimation for Semiparametric Models*; Johns Hopkins University Press: Baltimore, MD, USA, 1993.
40. Huang, J.; Rossini, A.J. Sieve estimation for the proportional-odds failure-time regression model with interval censoring. *J. Am. Stat. Assoc.* **1997**, *92*, 960–967.
41. Zhang, Y.; Hua, L.E.I.; Huang, J. A spline-based semiparametric maximum likelihood estimation method for the Cox model with interval-censored data. *Scand. J. Stat.* **2010**, *37*, 338–354.
42. Hall, P.; Heyde, C.C. *Martingale Limit Theory and Its Application*; Academic Press: Cambridge, MA, USA, 1980.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.