Received: 24 May 2022 | Accepted: 22 June 2023

DOI: 10.1111/biom.13898

BIOMETRIC PRACTICE

Biometrics A JOURNAL OF THE INTERNATIONAL BIOMETRIC SOCIETY WILEY

Analysis of length-biased and partly interval-censored survival data with mismeasured covariates

Li-Pang Chen | **Bangxu Qiu**

Department of Statistics, National Chengchi University, Taipei, Taiwan

Correspondence
Li-Pang Chen, Department of Statistics, National Chengchi University, Taipei, Taiwan. Email: lchen723@nccu.edu.tw

Funding information
National Science and Technology Council, Grant/Award Number: 110-2118-M-004-006-MY2

Abstract

In this paper, we analyze the length-biased and partly interval-censored data, whose challenges primarily come from biased sampling and interfere induced by interval censoring. Unlike existing methods that focus on low-dimensional data and assume the covariates to be precisely measured, sometimes researchers may encounter high-dimensional data subject to measurement error, which are ubiquitous in applications and make estimation unreliable. To address those challenges, we explore a valid inference method for handling high-dimensional length-biased and interval-censored survival data with measurement error in covariates under the accelerated failure time model. We primarily employ the SIMEX method to correct for measurement error effects and propose the boosting procedure to do variable selection and estimation. The proposed method is able to handle the case that the dimension of covariates is larger than the sample size and enjoys appealing features that the distributions of the covariates are left unspecified.

KEYWORDS

AFT model, biased sampling, boosting, Buckley–James formulation, incomplete data, measurement error correction, SIMEX, variable selection

1 | INTRODUCTION

Interval-censored data, known as that the failure time lies in some time interval, arise commonly in lifetime data analysis (e.g., Lawless, 2003, Section 2.3.1). That is, instead of collecting the failure time, one usually has censored intervals in datasets. Unlike interval-censored data that simply censored intervals can be observed, another general and attractive data structure is called partly interval-censored (PIC) data, which indicates that some of the failure times are exactly observed, while other failure times are subject to interval-censoring (e.g., Gao et al., 2017). In the past literature, many methods have been developed to handle those two types of data structures. To name a few, for interval-censored data, Cai & Betensky (2003) proposed the penalized spline method to construct semi-parametric regression models. Yavuz & Lambert (2011) developed the Bayesian penalized B-splines method to estimate survival functions. Wang et al. (2016) proposed a monotone spline representation to fit the Cox model. Fu & Simonoff (2017) adopted the survival tree method to analyze interval-censored data. Yao et al. (2019) employed an ensemble learning method for interval-censored and time-to-event data. A generous overview of interval-censored data analysis can be found in Du & Sun (2021a). Unlike estimation methods for interval-censored data that simply adopt interval-censoring times to construct estimating functions, the analysis of the PIC data requires to adjust interval-censoring effects and take observed failure times into account. For example, Huang (1999) proposed a non-parametric approach and developed asymptotic properties; Zhao et al. (2008) proposed a generalized log-rank test to

nonparametrically estimate survivor functions. With the covariates accommodated, Kim (2003) proposed nonparametric maximum likelihood estimation for the Cox model. Gao et al. (2017) considered the accelerated failure time (AFT) model and employed the Buckley–James formulation to construct an estimating function. Pan et al. (2020) adopted a Bayesian approach to model the Cox model under the PIC data.

With a complex sampling design, we usually have complicated data structure. Specifically, length-biased sampling has been an important issue in lifetime data analysis, and usually occurs due to prevalent cohort sampling approaches (e.g., Lawless, 2003, Section 2.4). In the presence of length-biased sampling, individuals with shorter survival times are less likely to be recruited for the study, thus resulting in a biased sample. In the past literature, a large body of methods have been available to deal with such a data structure. For example, Gao & Chan (2019) considered the Cox model and proposed the nonparametric maximum likelihood estimation. Wang et al. (2021) proposed a pairwise pseudo-likelihood approach for the Cox model. Pan & Chappell (2002) explored left-truncated and interval-censored data under the Cox model. Shen et al. (2022) considered length-biased and interval-censored data with a nonsusceptible fraction. However, those existing methods focus on interval-censored data instead of PIC data.

In addition to complex sampling design, the other challenges come from the covariates. The first feature is the involvement of irrelevant covariates. To address variable selection, several regularization strategies have been developed. In recent years, the boosting method is also adopted to deal with variable selection, including Wolfson (2011) and Brown et al. (2017). For the interval-censored survival data without length-biased sampling, existing regularization methods have been applied for the Cox model (e.g., Du et al., 2021; Zhao et al., 2020), the transformation model (e.g., Scolas et al., 2016), and the conditional cumulative hazard function (e.g., Sun et al. 2021). More comprehensive discussions can be found in Du & Sun (2021b). However, those methods cannot handle the PIC data. Moreover, to the best of our knowledge, there is no valid method to deal with variable selection for length-biased and interval-censored data.

The second feature of covariates is *measurement error*, which is ubiquitous when collecting data. It occurs due to wrong records of investigators or imprecise machines. In early developments, some methods have been explored for interval-censored data under the proportional odds model (e.g., Wen & Chen, 2014), the Cox model (e.g., Song & Ma, 2008), and the linear transformation model (e.g., Mandal et al., 2019). However, those methods focus on continuous covariates and do not explore length-biased sampling or PIC data. Regarding the framework of biased sampling induced by the prevalent cohort mechanism, several methods have been proposed for different models, such as the simulation and extrapolation (SIMEX) method (e.g., Chen, 2019, 2020) and the insertion method (e.g., Chen & Yi, 2021a). Moreover, relevant methods have also been extended to deal with variable selection, including the regularization method for the additive hazards model (e.g., Chen, 2021) and the focus information criterion for the Cox model (e.g., Chen & Yi, 2020). However, those strategies are based on the right-censored data, and the PIC data structure has not been carefully explored.

In this paper, we aim to fill out the research gap about dealing with measurement error and variable selection simultaneously for length-biased and PIC data. Specifically, we primarily consider the AFT model and adopt the Buckley–James formulation to develop an estimating function for length-biased and PIC data. After that, to correct for measurement error effects and address variable selection, we propose the SIMEXBoost method, which incorporates the SIMEX approach to the boosting algorithm. Our strategy has several advantages and differences from existing approaches. First, different from most developments that primarily focus on right-censored survival data, we develop a new approach to handle variable selection and measurement error for PIC data, which is rarely explored in the past literature. Second, unlike conventional regularization methods that aim to optimize penalized estimating functions, the SIMEXBoost method does not require nondifferentiable penalty functions; on the contrary, it enables us to deal with a general estimating function and obtain informative covariates as well as the corresponding estimator via finite iterations. For measurement error correction, the proposed method can correct for measurement error effects for error-prone binary and continuous covariates.

The remainder is organized as follows. In Section 2, we introduce the data structure and regression models. In Section 3, we propose the SIMEXBoost method, which aims to adopt the SIMEX method to correct for measurement error effect and apply the boosting method to do variable selection. In Section 4, we apply the proposed method to analyze a real dataset. Finally, a general discussion is presented in Section 5. Simulation studies and the corresponding numerical results are placed in the supporting information due to the limited space in the main text.

2 | NOTATION AND MODELS

2.1 | Length-biased and partly interval-censored data

Let $\mathbf{X}$ denote the p-dimensional vector of covariates. Let u be the initial event and let v denote the terminal event that is a primary interest. Let $\widetilde{T} > 0$ be the failure time that is defined as the length of the initial event u to the terminal event v. In the process of collecting data, it is possible that individuals have experienced the terminal event before being recruited in the study. Specifically, let ξ represent the calendar time of the recruitment and define $\widetilde{A} \triangleq \xi - u$ as the truncation time. Individuals cannot be recruited if $\widetilde{T} < \widetilde{A}$. On the contrary, the observable failure time and truncation time, denoted T and A, respectively, can be collected if $\widetilde{T} \geq \widetilde{A}$. Consequently, we define $(T, A) \equiv (\widetilde{T}, \widetilde{A})|\widetilde{T} \geq \widetilde{A}$, and it shows that this sampling scheme causes sampling bias.

Moreover, to describe $\widetilde{A}$, we follow the scenario in Chen & Yi (2021a) that the incidence of disease onset follows a stationary Poisson distribution, then the truncation time follows a uniform distribution with an interval $[0, \tau]$, where τ is the maximum support of $\widetilde{T}$. Under this situation, such a sampling scheme is called *length-biased sampling*.

In addition to the biased sampling, the failure time T can be incomplete due to the occurrence of censoring. In this paper, we primarily focus on the partly interval-censoring, which indicates that some of the failure times are exactly observed, while others are only known to lie within certain intervals (e.g., Gao et al., 2017). Specifically, let δ denote the binary indicator with $\delta = 1$ indicating that T is observed. When $\delta = 0$, there exists a sequence of examination times $A = U_0 < U_1 \cdots < U_K \leq \tau$ such that the censoring interval is given by (L, R), where $L = \max\{U_k : U_k < T, k = 0, \ldots, K\}$ and $R = \min\{U_k : U_k \geq T, k = 1, \ldots, K+1\}$. Consistent with Gao et al. (2017) and Gao & Chan (2019), $(U_1, \ldots, U_K)$ is assumed to be independent of T given $\mathbf{X}$.

Finally, for the sample with size n, let $\mathcal{D} \triangleq \{A_i, \delta_i, \delta_i T_i, (1-\delta_i)L_i, (1-\delta_i)R_i, \mathbf{X}_i : i = 1, \ldots, n\}$ denote the independent copy of $\{A, \delta, \delta T, (1-\delta)L, (1-\delta)R, \mathbf{X}\}$. In particular, if $\delta_i = 0$ for all $i = 1, \ldots, n$, then $\mathcal{D}$ is reduced to the length-biased and interval-censored data.

2.2 | Accelerated failure time models

In survival analysis, the primary interest is to characterize the relationship between the failure time and the covariates. In this paper, we mainly consider the well-known AFT model:

$$\log\widetilde{T} = \mathbf{X}^\top\boldsymbol{\beta} + \epsilon, \tag{1}$$

where $\boldsymbol{\beta}$ is a p-dimensional vector of primarily interested parameters, ϵ is the noise term with mean zero, and its density function $\varphi_\epsilon(\cdot)$ can be known or unknown.

If $\widetilde{T}$ is complete and fully observed, then the estimator of $\boldsymbol{\beta}$ can be naturally obtained by using the least-squares method. However, in the presence of length-biased sampling and interval-censoring, some necessary adjustments are required. To the end, we first adjust the effects induced by length-biased sampling, and then deal with interval-censoring.

Following the similar discussion in Chen & Yi (2021a), the joint density function of T and A given by $\mathbf{X} = \mathbf{x}$, denoted as $f(t, a|\mathbf{x})$, is formulated as

$$f(t, a|\mathbf{x}) \triangleq \frac{g(t|\mathbf{x})}{\mu(\mathbf{x})}\mathbb{I}(t > a), \tag{2}$$

where $\mathbb{I}(\cdot)$ is the indicator function and $\mu(\mathbf{x}) = \int_0^\infty tg(t|\mathbf{x})dt$ with $g(t|\mathbf{x})$ being the conditional density function of $\widetilde{T}$ given $\mathbf{X}$. By Equation (2), the observed failure time T has a length-biased conditional density function

$$f(t|\mathbf{x}) \triangleq \frac{tg(t|\mathbf{x})}{\mu(\mathbf{x})} \tag{3}$$

for $t > 0$. Moreover, under Equation (1), the conditional density function $g(t|\mathbf{x})$ is specified as

$$g(t|\mathbf{x}) = \frac{\varphi_\epsilon(\log t - \mathbf{x}^\top\boldsymbol{\beta})}{t} \tag{4}$$

for $t > 0$. Therefore, for $t > 0$, combining Equations (3) and (4) yields that

$$f(t|\mathbf{x}) = \frac{\varphi_\epsilon(\log t - \mathbf{x}^\top\boldsymbol{\beta})}{\mu(\mathbf{x})}. \tag{5}$$

We further define $\log T_\beta \triangleq \log T - \mathbf{X}^\top\boldsymbol{\beta}$. Then given $\mathbf{X}$, the expectation of $\frac{\log T_\beta}{T_\beta}$ based on Equation (5) yields that

$$\begin{aligned} E\left\{\frac{\log T_\beta}{T_\beta}\right\} &= \int_0^\infty \left\{\frac{\log t_0}{t_0} \times \frac{\varphi_\epsilon(\log t_0)}{\mu(\mathbf{x})}\right\} dt_0 \\ &= \int_0^\infty \left\{\frac{z_0\varphi_\epsilon(z_0)}{\mu(\mathbf{x})}\right\} dz_0 = 0, \end{aligned} \tag{6}$$

where the second step is obtained by the change of variable $z_0 \triangleq \log t_0$, and the last equality holds due to the zero expectation of ϵ. Consequently, under a sample

with size n, Equation (6) suggests an unbiased estimating function with length-biased sampling involved only: $\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\left\{\frac{\log T_i-\mathbf{X}_i^\top\boldsymbol{\beta}}{T_i\exp(-\mathbf{X}_i^\top\boldsymbol{\beta})}\right\}=0$.

In addition, in the presence of interval-censoring, $\frac{\log T_\beta}{T_\beta}$ is only observed when $\delta=1$; when $\delta=0$, the failure time T_β is unobservable. To address the interval-censoring effect for the subjects with $\delta=0$, we employ the Buckley–James method (e.g., Buckley & James, 1979), whose key strategy is to compute the conditional expectation of $\frac{\log T_\beta}{T_\beta}$, given $\delta=0$, and use it to adjust the censoring effects. Specifically, the idea of the Buckley–James method is to construct the *pseudo response* with the conditional expectation accommodated

$$Y_{\mathrm{BJ}}\triangleq\delta\frac{\log T_\beta}{T_\beta}+(1-\delta)E\left\{\frac{\log T_\beta}{T_\beta}\middle|\delta=0,L_\beta,R_\beta,A\right\},\quad(7)$$

where $R_\beta\triangleq R\exp(-\mathbf{X}^\top\boldsymbol{\beta})$ and $L_\beta\triangleq L\exp(-\mathbf{X}^\top\boldsymbol{\beta})$. In the case of $\delta=0$, we further express the following conditional expectation in Equation (7):

$$\begin{aligned}&E\left\{\frac{\log T_\beta}{T_\beta}\middle|\delta=0,L_\beta,R_\beta,A\right\}\\&=\frac{E\left\{\frac{\log T_\beta}{T_\beta}\mathbb{I}\left(L_\beta<T_\beta<R_\beta\right)\right\}}{P\left(L_\beta<T_\beta<R_\beta\right)}=\frac{\int_{L_\beta}^{R_\beta}\left(u^{-1}\log u\right)dG_0(u)}{G_0(R_\beta)-G_0(L_\beta)},\end{aligned}\quad(8)$$

where $G_0(t)$ is the cumulative distribution function of $\widetilde{T}_\beta\triangleq\widetilde{T}\exp(-\mathbf{X}^\top\boldsymbol{\beta})$. Therefore, together with Equations (7) and (8), a new estimating function for length-biased and PIC data is given by

$$\begin{aligned}\mathbf{U}(\boldsymbol{\beta})=\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\Bigg\{&\delta_i\frac{\log T_i-\mathbf{X}_i^\top\boldsymbol{\beta}}{T_i\exp(-\mathbf{X}_i^\top\boldsymbol{\beta})}\\&+(1-\delta_i)\frac{\int_{L_{\beta,i}}^{R_{\beta,i}}\left(u^{-1}\log u\right)dG_0(u)}{G_0(R_{\beta,i})-G_0(L_{\beta,i})}\Bigg\}\end{aligned}\quad(9)$$

with $R_{\beta,i}\triangleq R_i\exp(-\mathbf{X}_i^\top\boldsymbol{\beta})$ and $L_{\beta,i}\triangleq L_i\exp(-\mathbf{X}_i^\top\boldsymbol{\beta})$.

On the other hand, we observe from Equation (9) that $G_0(t)$ is a nuisance function and is implicitly affected by $\varphi_\epsilon(\cdot)$. If $\varphi_\epsilon(\cdot)$ is known, then $G_0(t)$ is simply the integral of $\varphi_\epsilon(t)$. If $\varphi_\epsilon(\cdot)$ and $G_0(t)$ are unknown, then we may use observed data to estimate it. Let $F_0(t)$ denote the cumulative distribution function of $T_{\beta,i}\triangleq T_i\exp(-\mathbf{X}_i^\top\boldsymbol{\beta})$. With $\boldsymbol{\beta}$ fixed at $\boldsymbol{\beta}^*$, we define $\widehat{F}_0(t;\boldsymbol{\beta}^*)$ as the estimator of $F_0(t)$ that satisfies the following self-consistency equation (e.g., Huang, 1999; Gao et al., 2017):

$$\begin{aligned}\widehat{F}_0(t;\boldsymbol{\beta}^*)=\frac{1}{n}\sum_{i=1}^{n}\Bigg\{&\delta_i\mathbb{I}\left(T_{\boldsymbol{\beta}^*,i}\le t\right)\\&+(1-\delta_i)\frac{\widehat{F}_0(R_{\boldsymbol{\beta}^*,i}\wedge t;\boldsymbol{\beta}^*)-\widehat{F}_0(L_{\boldsymbol{\beta}^*,i}\wedge t;\boldsymbol{\beta}^*)}{\widehat{F}_0(R_{\boldsymbol{\beta}^*,i};\boldsymbol{\beta}^*)-\widehat{F}_0(L_{\boldsymbol{\beta}^*,i};\boldsymbol{\beta}^*)}\Bigg\},\end{aligned}\quad(10)$$

where $a\wedge b\triangleq\min\{a,b\}$. In addition, according to the discussion in Ning et al. (2011), $G_0(t)$ can be expressed as

$$G_0(t)=\frac{\int_0^t s^{-1}dF_0(s)}{\int_0^\infty s^{-1}dF_0(s)}.\quad(11)$$

With the estimate $\widehat{F}_0(t;\boldsymbol{\beta}^*)$, the estimator of $G_0(t)$, denoted by $\widehat{G}_0(t;\boldsymbol{\beta}^*)$, can be obtained by Equation (11) with $F_0(t)$ replaced by $\widehat{F}_0(t;\boldsymbol{\beta}^*)$. Therefore, replacing $G_0(t)$ in Equation (9) by $\widehat{G}_0(t;\boldsymbol{\beta}^*)$ yields a workable estimating function:

$$\begin{aligned}\widehat{\mathbf{U}}(\boldsymbol{\beta},\boldsymbol{\beta}^*)=\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\Bigg\{&\delta_i\frac{\log T_i-\mathbf{X}_i^\top\boldsymbol{\beta}}{T_i\exp(-\mathbf{X}_i^\top\boldsymbol{\beta})}\\&+(1-\delta_i)\frac{\int_{L_{\beta,i}}^{R_{\beta,i}}\left(u^{-1}\log u\right)d\widehat{G}_0(u;\boldsymbol{\beta}^*)}{\widehat{G}_0(R_{\beta,i};\boldsymbol{\beta}^*)-\widehat{G}_0(L_{\beta,i};\boldsymbol{\beta}^*)}\Bigg\}.\end{aligned}\quad(12)$$

With fixed $\boldsymbol{\beta}^*$, one can solve an estimating equation $\widehat{\mathbf{U}}(\boldsymbol{\beta},\boldsymbol{\beta}^*)=\mathbf{0}_p$ with respect to $\boldsymbol{\beta}$, where $\mathbf{0}_p$ represents the p-dimensional zero vector. Finally, iteratively updating Equations (10) and (12) until convergence essentially gives the solution of $\widehat{\mathbf{U}}(\boldsymbol{\beta},\boldsymbol{\beta})=\mathbf{0}_p$, and the resulting estimator is given by $\widehat{\boldsymbol{\beta}}$. The detailed implementation as well as the procedure of solving Equation (12) is deferred to Section 3.

2.3 | Measurement error models

In applications, covariates are often subject to measurement error. Let $\mathbf{X}$ denote the unobserved covariates and it can be decomposed as $\mathbf{X}=(\mathbf{X}_{\mathrm{C}}^\top,\mathbf{X}_{\mathrm{D}}^\top)^\top$, where $\mathbf{X}_{\mathrm{C}}$ is the p_{C}-dimensional vector of continuous covariates, and $\mathbf{X}_{\mathrm{D}}$ represents the p_{D}-dimensional vector of discrete covariates. Let $\mathbf{X}^*$ denote the observed version of $\mathbf{X}$, and it can be expressed as $\mathbf{X}^*=(\mathbf{X}_{\mathrm{C}}^{*\top},\mathbf{X}_{\mathrm{D}}^{*\top})^\top$ with $\mathbf{X}_{\mathrm{C}}^*$ and $\mathbf{X}_{\mathrm{D}}^*$ being the observed versions of $\mathbf{X}_{\mathrm{C}}$ and $\mathbf{X}_{\mathrm{D}}$, respectively.

To describe the relationship between $\mathbf{X}^*$ and $\mathbf{X}$, we use the factorization

$$\left[\mathbf{X}_{\mathrm{C}}^*,\mathbf{X}_{\mathrm{D}}^*|\mathbf{X}_{\mathrm{C}},\mathbf{X}_{\mathrm{D}}\right]=\left[\mathbf{X}_{\mathrm{C}}^*|\mathbf{X}_{\mathrm{D}}^*,\mathbf{X}_{\mathrm{C}},\mathbf{X}_{\mathrm{D}}\right]\times\left[\mathbf{X}_{\mathrm{D}}^*|\mathbf{X}_{\mathrm{C}},\mathbf{X}_{\mathrm{D}}\right],\quad(13)$$

where $[\cdot|\cdot]$ represents the conditional distribution for the random variables indicated by the arguments. Making the

independence assumptions in $\mathbf{X}_{\mathrm{C}}^*$ and $\mathbf{X}_{\mathrm{D}}^*$ allows Equation (13) to be expressed as $[\mathbf{X}_{\mathrm{C}}^*, \mathbf{X}_{\mathrm{D}}^* | \mathbf{X}_{\mathrm{C}}, \mathbf{X}_{\mathrm{D}}] = [\mathbf{X}_{\mathrm{C}}^* | \mathbf{X}_{\mathrm{C}}] \times [\mathbf{X}_{\mathrm{D}}^* | \mathbf{X}_{\mathrm{D}}]$, suggesting that one can characterize $[\mathbf{X}_{\mathrm{C}}^* | \mathbf{X}_{\mathrm{C}}]$ and $[\mathbf{X}_{\mathrm{D}}^* | \mathbf{X}_{\mathrm{D}}]$ separately.

To characterize $[\mathbf{X}_{\mathrm{C}}^* | \mathbf{X}_{\mathrm{C}}]$, we consider the classical additive measurement error model (Carroll et al., 2006)

$$\mathbf{X}_{\mathrm{C}}^* = \mathbf{X}_{\mathrm{C}} + \mathbf{e}, \tag{14}$$

where $\mathbf{e}$ follows a normal distribution with mean zero and covariance $\boldsymbol{\Sigma}_e$. We assume that $\mathbf{e}$ is independent of $\mathbf{X}$ and ϵ in Equation (1). On the other hand, for discrete covariates, we first define possible values of $\mathbf{X}_{\mathrm{D}}$ as $\mathbf{x}_{(1)}, \mathbf{x}_{(2)}, ..., \mathbf{x}_{(m)}$, and let $\pi_{kl} \triangleq P(\mathbf{X}_{\mathrm{D}}^* = \mathbf{x}_{(k)} | \mathbf{X}_{\mathrm{D}} = \mathbf{x}_{(l)})$ denote the (mis)classification probability for $k, l = 1, ..., m$ (Chen & Yi, 2021b). We define $\boldsymbol{\Pi} = [\pi_{kl}]$ as a $m \times m$ (mis)classification matrix. In addition, by the law of total probability, we have that for $k = 1, ..., m$, $P(\mathbf{X}_{\mathrm{D}}^* = \mathbf{x}_{(k)}) = \sum_{l=1}^{m} \pi_{kl} P(\mathbf{X}_{\mathrm{D}} = \mathbf{x}_{(l)})$. Thus, we can model $[\mathbf{X}_{\mathrm{D}}^* | \mathbf{X}_{\mathrm{D}}]$ as (e.g., Chen & Yi, 2021b)

$$\begin{pmatrix} P(\mathbf{X}_{\mathrm{D}}^* = x_{(1)}) \\ \vdots \\ P(\mathbf{X}_{\mathrm{D}}^* = x_{(m)}) \end{pmatrix} = \boldsymbol{\Pi} \begin{pmatrix} P(\mathbf{X}_{\mathrm{D}} = x_{(1)}) \\ \vdots \\ P(\mathbf{X}_{\mathrm{D}} = x_{(m)}) \end{pmatrix}. \tag{15}$$

Therefore, the surrogate vector $\mathbf{X}_{\mathrm{D}}^*$ can be characterized by the true covariate vector $\mathbf{X}_{\mathrm{D}}$ through Equation (15). To ease notation, we let $\mathrm{MC}[\boldsymbol{\Pi}](\cdot)$ denote the misclassification operator indicated by Equation (15) and notationally write Equation (15) as $\mathbf{X}_{\mathrm{D}}^* = \mathrm{MC}[\boldsymbol{\Pi}](\mathbf{X}_{\mathrm{D}})$. Such a misclassification operator was also used by Carroll et al. (2006, p. 125) for a misclassified binary variable. In addition, as discussed in Carroll et al. (2006, p. 125), we assume that $\boldsymbol{\Pi}$ has the decomposition $\boldsymbol{\Pi} = \boldsymbol{\Omega}\mathbf{D}\boldsymbol{\Omega}^{-1}$, where $\mathbf{D}$ is the diagonal matrix with diagonal elements being the eigenvalues of $\boldsymbol{\Pi}$, and $\boldsymbol{\Omega}$ is the corresponding matrix of eigenvectors.

In this paper, to highlight the key idea of measurement error effects as well as measurement error correction, we assume that $\boldsymbol{\Sigma}_e$ and $\boldsymbol{\Pi}$ are known for now. Note that, in applications, $\boldsymbol{\Sigma}_e$ and/or $\boldsymbol{\Pi}$ are unknown, one may require additional data information, such as repeated measurements or validation sample, to estimate unknown parameters (e.g., Carroll et al., 2006, Section 2.3). In the implementation without availability of auxiliary information, one may conduct sensitivity analyses, whose purpose is to reasonably specify different values to $\boldsymbol{\Sigma}_e$ and $\boldsymbol{\Pi}$, and examine the impact of different magnitudes of measurement error effects. The demonstration of sensitivity analyses is placed in Section 4; detailed discussions of the estimation of unknown $\boldsymbol{\Sigma}_e$ and/or $\boldsymbol{\Pi}$ can be found in Section 5 and additional numerical results are placed in the supporting information.

3 | METHODOLOGY

3.1 | SIMEXBoost

In this section, we propose the *SIMEXBoost* method, the combination of the SIMEX method and the boosting procedure, to correct for measurement error effects and do variable selection simultaneously. Detailed descriptions are given below, and a pseudo code of the algorithm is summarized in Appendix A of the supporting information.

Stage 1: Setup

Let B be a given positive integer and let $\mathcal{Z} = \{\zeta_0, \zeta_1, ..., \zeta_M\}$ be a sequence of pre-specified values with $0 = \zeta_0 < \zeta_1 < \cdots < \zeta_M$, where M is a positive integer, and ζ_M is a prespecified positive number such as $\zeta_M = 1$ or 2. While B and $\mathcal{Z}$ are not uniquely specified, commonly, B is set as a value between 100 and 500, $\mathcal{Z}$ is taken as a collection of M points that equally cut the interval [0,1] or [0,2] with M set as 5 or 10 (e.g., Carroll et al., 2006, p. 106).

For a given subject i with $i = 1, ..., n$ as well as $\zeta \in \mathcal{Z}$ and $b = 1, ..., B$, we generate $\mathbf{V}_i$ from $N(\mathbf{0}_p, \boldsymbol{\Sigma}_e)$. Then for a vector $\mathbf{X}_{\mathrm{C},i}^*$, we define $\mathbf{W}_{\mathrm{C},i}(b, \zeta) = \mathbf{X}_{\mathrm{C},i}^* + \sqrt{\zeta}\mathbf{V}_i$. In addition, for the discrete error-prone covariates $\mathbf{X}_{\mathrm{D},i}^*$, we generate $\mathbf{W}_{\mathrm{D},i}(b, \zeta)$ by

$$\mathbf{W}_{\mathrm{D},i}(b, \zeta) = \mathrm{MC}\left[\boldsymbol{\Pi}^{\zeta}\right](\mathbf{X}_{\mathrm{D},i}^*), \tag{16}$$

where $\boldsymbol{\Pi}^{\zeta} = \boldsymbol{\Omega}\mathbf{D}^{\zeta}\boldsymbol{\Omega}^{-1}$, and $\mathbf{D}^{\zeta}$ is derived from $\mathbf{D}$ by replacing its diagonal elements, say d_{jj}, with d_{jj}^{ζ}.

We call

$$\mathbf{W}_i(b, \zeta) \triangleq \left(\mathbf{W}_{\mathrm{C},i}^{\top}(b, \zeta), \mathbf{W}_{\mathrm{D},i}^{\top}(b, \zeta)\right)^{\top} \tag{17}$$

the *working data* for $b = 1, ..., B$, $\zeta \in \mathcal{Z}$ and $i = 1, ..., n$.

Stage 2: Estimation

In this stage, we run the boosting algorithm with measurement error correction accommodated.

For given b and ζ, we first define $\boldsymbol{\beta}^{(0)}(b, \zeta) = \mathbf{0}_p$ as the initial value and define $\widehat{G}_0(t; \boldsymbol{\beta}^{(0)}(b, \zeta))$ by computing Equation (10) with $\mathbf{X}_i$ replaced

by $\mathbf{W}_i(b,\zeta)$. Let I denote the total number of iterations. For the $(r-1)$th iteration with $r = 1, 2, ..., I$ as well as fixed $b = 1, ..., B$ and $\zeta \in \mathcal{Z}$, let $\widehat{\mathbf{U}}_{\mathrm{SIM}}(\boldsymbol{\beta}, \boldsymbol{\beta}^{(r-1)}(b,\zeta); b, \zeta)$ denote (12) with $\mathbf{X}_i$ replaced by the working data (17).

Define $\boldsymbol{\Delta}^{(r-1)}(b,\zeta) \triangleq \widehat{\mathbf{U}}_{\mathrm{SIM}}(\boldsymbol{\beta}, \boldsymbol{\beta}^{(r-1)}(b,\zeta); b, \zeta)|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(r-1)}}$ at the $(r-1)$th iteration for $r = 1, ..., I$. After that, we take $\boldsymbol{\Delta}^{(r-1)}(b,\zeta)$ as signals and define an active set

$$\mathcal{J}_{r-1}(b,\zeta) \triangleq \left\{ j : \left|\Delta_j^{(r-1)}(b,\zeta)\right| \geq \varrho \cdot \max_{j'} \left|\Delta_{j'}^{(r-1)}(b,\zeta)\right| \right\}, \tag{18}$$

where $\Delta_j^{(r-1)}(b,\zeta)$ is the jth component of $\boldsymbol{\Delta}^{(r-1)}(b,\zeta)$ and $\varrho \in [0,1]$ is a thresholding constant. We then update values by the steepest descent

$$\beta_j^{(r)}(b,\zeta) = \beta_j^{(r-1)}(b,\zeta) + \kappa \cdot \mathrm{sign}\{\Delta_j^{(r-1)}(b,\zeta)\} \tag{19}$$

for all $j \in \mathcal{J}_{r-1}(b,\zeta)$ and $r = 1, ..., I$, where κ is a positive increment. Moreover, we further derive an updated function $\widehat{G}_0(t; \boldsymbol{\beta}^{(r)}(b,\zeta))$ that is obtained by solving Equation (10) with given $\boldsymbol{\beta}^{(r)}(b,\zeta)$.

Finally, continue this procedure based on the boosting algorithm until achieving the last iteration I, we have $\boldsymbol{\beta}^{(I)}(b,\zeta)$ for $b = 1, ..., B$ and $\zeta \in \mathcal{Z}$. For any fixed $\zeta \in \mathcal{Z}$, taking an average on $\boldsymbol{\beta}^{(I)}(b,\zeta)$ with respect to b yields that

$$\boldsymbol{\beta}^{(I)}(\zeta) = \frac{1}{B}\sum_{b=1}^{B} \boldsymbol{\beta}^{(I)}(b,\zeta). \tag{20}$$

Stage 3: Extrapolation

For a sequence $\{(\zeta, \boldsymbol{\beta}^{(I)}(\zeta)) : \zeta \in \mathcal{Z}\}$ obtained from (20), we fit a regression model to a sequence $\boldsymbol{\beta}^{(I)}(\zeta) = \psi(\zeta; \boldsymbol{\Gamma}) + \eta$, where $\psi(\cdot;\cdot)$ is the extrapolation function that is approximated by user-specific regression functions, $\boldsymbol{\Gamma}$ is the associated parameter, and η is the noise term. The parameter $\boldsymbol{\Gamma}$ can be estimated by the least-squares method; and we let $\widehat{\boldsymbol{\Gamma}}$ denote the resulting estimate of $\boldsymbol{\Gamma}$.

Finally, we calculate the predicted value $\widehat{\boldsymbol{\beta}} \triangleq \psi(-1; \widehat{\boldsymbol{\Gamma}})$ and take $\widehat{\boldsymbol{\beta}}$ as the final estimator.

The key idea of the proposed three-stage procedure is to artificially simulate surrogate measurements with varying magnitudes of mismeasurement to delineate the patterns of different degrees of mismeasurement on inference results. The first and third stages generalize the SIMEX method (Carroll et al., 2006, Chapter 5) which is applicable to error-contaminated continuous and binary covariates. In addition, similar to the discussion in Brown et al. (2017), taking the indices corresponding to the threshold values in the second stage undertakes the selection of important covariates with different magnitudes of mismeasurement. Variable selection in this stage is similar to the development in Chen & Yi (2021b) who proposed the adaptive least absolute shrinkage and selection operator (LASSO) method, and thus, it ensures that important covariates can be detected if the working data are implemented to adjust for measurement error effects.

To see the validity of the measurement error corrections, Stage 1 generates a sequence of surrogate covariates, say Equation (17), whose value of ζ reflects the degree of mismeasurement in the artificially generated surrogates $\mathbf{W}_i(b,\zeta)$. With a positive and increasing ζ, $\mathbf{W}_i(b,\zeta)$ incurs an increasing amount of mismeasurement whose effects on inferential procedures are recorded in Stage 2. When $\zeta = -1$, $\mathbf{W}_i(b,\zeta)$ reduces to $\mathbf{X}_i$, the ideal situation without measurement error. Using the patterns obtained from Stage 2 for different degrees of mismeasurement, in Stage 3, we employ a regression model to obtain estimators corresponding to the error-free scenario (i.e., $\zeta = -1$).

We further comment on some parameters in Stage 2. First of all, ϱ in Equation (18) controls the number of selected covariates. If $\varrho = 0$, all covariates will be included and variable selection is ignored. If $\varrho = 1$, then the whole algorithm reduces to the generic procedure by selecting one covariate only in each iteration (e.g., Wolfson, 2011), and it has been shown by Wolfson (2011) that the updated path obtained in Step 2 is (approximately) equivalent to the LASSO method. However, updating one component in each iteration may make computation cumbersome. Hence, since our goal is to do variable selection, to take a balance between reducing iterated times and selecting informative predictors simultaneously at each iteration, one can choose a value ϱ that is close to 1, and our numerical examinations find that $\varrho = 0.9$ gives satisfactory results.

Next, regarding the update in Equation (19), our approach follows the steepest descent for L_1-norm that takes the sign of the estimating function as an increment. In addition, similar to the discussion in Brown et al. (2017), the LASSO is approximately equivalent to Stage 2 with $I \cdot \kappa \to 0$ as $I \to \infty$ and $\kappa \to 0$, it suggests that the learning rate κ should be a small value. Finally, while a large value of I ensures a precise estimate, it may make over-fitting at the same time. As a result, early stopping of iteration can be implemented. One of commonly used criteria is to stop the iteration at step $r+1$ if $\|\widehat{\mathbf{U}}_{\mathrm{SIM}}(\boldsymbol{\beta}^{(r+1)}(b,\zeta), \boldsymbol{\beta}^{(r+1)}(b,\zeta); b, \zeta) -$

$\widehat{\mathbf{U}}_{\mathrm{SIM}}(\boldsymbol{\beta}^{(r+2)}(b,\zeta),\boldsymbol{\beta}^{(r+2)}(b,\zeta);b,\zeta)\|_\infty < \varepsilon_{\mathrm{tol}}$ is satisfied for some tolerated constant $\varepsilon_{\mathrm{tol}} > 0$ and given $b = 1,\ldots,B$ and $\zeta \in \mathcal{Z}$, where $\|\mathbf{a}\|_\infty = \max_i |a_i|$ is the infinity norm for a vector $\mathbf{a} = (a_1,\ldots,a_p)^\top$.

In summary, the proposed SIMEXBoost method has several advantages. First, unlike Brown et al. (2017) whose algorithm requires an assumption that covariates should be precisely measured, the SIMEXBoost method provides a flexible strategy to handle measurement error and variable selection simultaneously. Moreover, our setting explores biased and incomplete data induced by truncation and censoring, which is not studied in Brown et al. (2017). Second, since $\boldsymbol{\beta}$ in the AFT model (1) is based on the estimating function and its estimator is difficult to solve directly due to possible discontinuity of the estimating function (e.g., Gao et al., 2017), Stage 2 of the SIMEXBoost method is valid to handle variable selection and solve a constructed estimating function. In addition, unlike regularization methods (e.g., LASSO) that are required to handle non-differentiable penalty functions, Stage 2 simply adopts iterations to derive estimators with informative covariates detected.

3.2 | SIMEXBoost with collinearity in covariates

When constructing regression models, a crucial concern in covariates is *collinearity*, which shows high correlations among covariates. It is known that collinearity may produce misleading results, such as falsely concluding significant covariates to be insignificant.

To deal with variable selection and collinearity based on the SIMEXBoost method, we follow the idea of the elastic net method (e.g., Zou & Hastie, 2005) that incorporates L_1- and L_2-norm penalty functions. Specifically, for $b = 1,\ldots,B$ and $\zeta \in \mathcal{Z}$, we suggest replacing $\widehat{\mathbf{U}}_{\mathrm{SIM}}(\boldsymbol{\beta},\boldsymbol{\beta}^*;b,\zeta)$ in Equations (18) and (19) by

$$\widetilde{\mathbf{U}}_{\mathrm{SIM}}(\boldsymbol{\beta},\boldsymbol{\beta}^*;b,\zeta) = \widehat{\mathbf{U}}_{\mathrm{SIM}}(\boldsymbol{\beta},\boldsymbol{\beta}^*;b,\zeta) + 2\lambda\boldsymbol{\beta}, \qquad (21)$$

where λ is a tuning parameter. The optimal tuning parameter can be chosen by the cross-validation approach. To see why the proposed approach works, we observe that the elastic net based penalized likelihood function can be re-expressed as the LASSO method for the ridge regression function, and the derivative of the ridge regression function is equivalent to the sum of the estimating function and $2\lambda\boldsymbol{\beta}$. To connect with our approach, one can regard Equation (21) as the differentiation of the likelihood function with the L_2-norm penalty function, and the whole algorithm is taken as the LASSO approach to do variable selection.

4 | APPLICATION TO THE SIGNAL TANDMOBIEL STUDY

In this section, we apply the SIMEXBoost method to analyze the tooth data, which is a longitudinal prospective oral health study collected from the Signal Tandmobiel® study in the Flanders region of Belgium from 1996 to 2001. In this dataset, the cohort of 4468 randomly sampled children who attended the first year of the basic school at the beginning of the study was annually dental examined by trained dentists. Among them, we primarily consider the sample with size $n = 4,430$ because the remaining 38 sampled children did not come to any of the designed dental examinations.

In this paper, we are interested in the emergence times of 28 teeth that are divided by seven types of teeth, and each type contains four teeth, including permanent incisors (label: 11, 21, 31, 41), permanent central canines (label: 12, 22, 32, 42), permanent lateral canines (label: 13, 23, 33, 43), permanent first premolars (label: 14, 24, 34, 44), permanent second premolars (label: 15, 25, 35, 45), permanent first molars (label: 16, 26, 36, 46), and permanent second molars (label: 17, 27, 37, 47). We sort these labels as $1, 2, \ldots, 28$ in the following analysis. Each of the tooth emergence times can be taken as a response variable. However, as recorded in the dataset, the emergence time encounters interval-censoring. Moreover, for those 38 children, their tooth emergence times cannot be observed due to unavailability of designed dental examinations, yielding the biased sample. In this dataset, there are 41 covariates, including gender (`gender`, 0 for boy and 1 for girl), province (`province`, factor with code 0 for Antwerpen, 1 for Vlaams Brabant, 2 for Limburg, 3 for Oost Vlaanderen, 4 for West Vlaanderen), evidence of fluoride intake (`fluor`, binary with 0 for no and 1 for yes), type of educational system (`educ`, factor with code 0 for Free, 1 for Community school, 2 for Province/council school), starting age of brushing teeth (`startbr`, factor with code 1 for [0, 1] years, 2 for (1, 2] years, 3 for (2, 3] years, 4 for (3, 4] years, 5 for (4, 5] years, 6 for later than at the age of 5), deciduous tooth with label `xx` that were decayed or missing due to caries or filled (`Txx.DMF`, binary with 0 for no and 1 for yes), deciduous tooth with label `xx` that were removed because of orthodontic reasons (`BAD.xx`, binary with 0 for no and 1 for yes), deciduous tooth with label `xx` that were removed due to the orthodontical reasons or decayed on at most the last examination before the first examination when the emergency of the permanent successor was recorded (`Txx.CAR`, binary with 0 for no and 1 for yes) with `xx` being 53, 63, 73, 83 (deciduous lateral canines), 54, 64, 74, 84 (deciduous first molars), 55, 65, 75, 85 (deciduous second molars). Except for covariates `gender`, `province`, `educ`, `startbr`, however, other binary covariates may be

contaminated with measurement error (e.g., Küchenhoff et al., 2007). While this dataset was discussed by several authors (e.g., Fu & Simonoff, 2017; Yao et al., 2019), their approaches ignored the possible impact of biased sampling and assumed the covariates to be precisely measured. To address those concerns, we implement the SIMEXBoost method to analyze this dataset.

Since the dataset has no additional information, such as repeated measurements or validation data, to quantify the degree of measurement error, we conduct *sensitivity analyses* to investigate how the analysis results are affected by different magnitudes of measurement error. Since the error-prone covariates are binary random variables, we adopt Equation (16) with $\mathbf{\Pi}$ being specified as a 2×2 matrix $\begin{pmatrix} 1-\pi & \pi \\ \pi & 1-\pi \end{pmatrix}$ with π specified as 0.005 or 0.05 to characterize each error-prone covariate.

To assess the performance of estimation and prediction for each tooth, we adopt the H-fold cross-validation. Specifically, for a given tooth j with $j = 1, \ldots, 28$, we randomly split the original data into H roughly equal-sized subsets. For $h = 1, \ldots, H$, take the hth subset as the validation data, and let the remaining $(H-1)$ pooled subsets as the training data; let $\mathcal{V}_{j,h}$ and $\mathcal{T}_{j,h}$ represent the classes of the subject indexes for the jth tooth and the hth validation and training datasets, respectively. In our study, we take $H = 5$.

For $j = 1, \ldots, 28$ and $h = 1, \ldots, H$, we adopt the estimating function (21) to fit the training data $\mathcal{T}_{j,h}$ due to potential collinearity in records of different deciduous teeth (e.g., Küchenhoff et al., 2007) and possibly strong effects between deciduous teeth (e.g., Fu & Simonoff, 2017; Aktan et al., 2012). To implement the SIMEXBoost method, we specify $B = 200$ and $\mathcal{Z} = \{0, 0.25, 0.5, 0.75, 1\}$ in Stage 1; $I = 1000$, $\varrho = 0.9$, and $\kappa = 0.02$ in Stage 2. The corresponding estimator of $\boldsymbol{\beta}$ under the training data $\mathcal{T}_{j,h}$ is denoted as $\widehat{\boldsymbol{\beta}}_h$. With the estimates $\widehat{\boldsymbol{\beta}}_h$ and the covariates in $\mathcal{V}_{j,h}$, we obtain the predicted failure time $\widehat{T}_{j,i}$ for $i \in \mathcal{V}_{j,h}$. Repeating the procedure H times gives the predicted failure time $\widehat{T}_{j,i}$ for the whole dataset.

For $j = 1, \ldots, 28$, let $n_{(j)}$ represent the number of interval-censored data for the jth tooth. Let $[L_{j,i}, R_{j,i}]$ denote the interval-censoring time recorded in the dataset for $j = 1, \ldots, 28$ and $i = 1, \ldots, n_{(j)}$. To conduct the performance of prediction, we follow an idea in Yao et al. (2019) to consider the predicted emergency times that are not within the observed emergence interval, and we calculate the proportion that the predicted emergence times lie outside censoring intervals $p_{\text{out}(j)} = \frac{n_{\text{out}(j)}}{n_{(j)}}$ and the average absolute prediction distance $\bar{d}_{\text{out}(j)} = \frac{100\%}{|\mathcal{N}_j|}\sum_{i\in\mathcal{N}_j} \min(|\frac{L_{j,i}-\widehat{T}_{j,i}}{L_{j,i}}|, |\frac{R_{j,i}-\widehat{T}_{j,i}}{R_{j,i}}|)$ for $j = 1, 2, \ldots, 28$, where $n_{\text{out}(j)} = \sum_{i=1}^{n_{(j)}} \mathbb{I}\{i \in \mathcal{N}_j\}$ with $\mathcal{N}_j = \{i : L_{j,i} > \widehat{T}_{j,i}$ or $R_{j,i} < \widehat{T}_{j,i}\}$ being a set that contains subjects whose predicted failure times lie outside censoring intervals. As commented by Yao et al. (2019), smaller values of $\bar{d}_{\text{out}}$ indicate better prediction of emergence times. In addition, smaller values of $\bar{p}_{\text{out}}$ reflect that more predicted failure times fall in censoring intervals.

In addition to the implementation of the proposed method, for the comparison, we also examine other approaches, including the boosting estimation in Section 3.1 without measurement error correction (naive), SIMEX estimation without variable selection (SIMEX(π) with $\pi = 0.005$ or 0.05), estimation from Equation (9) without variable selection and measurement error correction (LBPIC), and estimation proposed by Gao et al. (2017) without consideration of length-biased sampling (PIC(Gao et al.)). Moreover, we also examine two interval-censored based methods, denoted IC_Bayesian and IC_Par, that can be implemented by the R package `icenReg`. Throughout those comparisons, we wish to explore the impacts of length-biased sampling, variable selection, and measurement error correction in the analysis. Numerical results for 28 teeth are summarized in Table 1. Due to the smallest values of p_{out} and $\bar{d}_{\text{out}}$ derived by the SIMEXBoost method, we find that the prediction obtained by the proposed method is generally more precise than other methods that do not take noisy features into account. Among all teeth, all numerical results look similar, and values of p_{out} and $\bar{d}_{\text{out}}$ are comparable regardless of magnitude of measurement error effects, suggesting that the proposed method provides stable results. On the other hand, among other approaches with variable selection or measurement error ignored, we find that SIMEX(π) sometimes produces larger values of p_{out} and $\bar{d}_{\text{out}}$ than those from naive, suggesting that ignoring variable selection may provide worse prediction than ignoring measurement error correction. Finally, compared LBPIC with IC_Bayesian, IC_Par, and PIC(Gao et al.), we find that values of p_{out} and $\bar{d}_{\text{out}}$ derived by IC_Bayesian, IC_Par, and PIC(Gao et al.) are obviously larger than those from LBPIC. It indicates that, provided that variable selection and measurement error are not accommodated, length-biased effects indeed affect the estimation results.

Based on predicted failure times $\widehat{T}_{j,i}$, we estimate the survivor curves $\widehat{S}_j(t) = \frac{1}{n_{(j)}}\sum_{i=1}^{n_{(j)}} \mathbb{I}(\widehat{T}_{j,i} > t)$ for a tooth $j = 1, \ldots, 28$. Due to the limited space, we demonstrate the estimated survivor curves for label 11 based on the proposed method and its competitors in Figure 1; other figures can be found in the supporting information. In general, with measurement error correction, we find that the proposed method produces close curves regardless of

TABLE 1 Evaluation on the signal Tandmobiel® study datasets.

	Naive		SIMEXBoost (0.005)		SIMEXBoost (0.05)		SIMEX (0.005)		SIMEX (0.05)		LBPIC		PIC(Gao et al.)		IC_Bayesian		IC_Par	
Tooth	p_{out}	$\bar{d}_{\text{out}}$	p_{out}	$\bar{d}_{\text{out}}$	p_{out}	$\bar{d}_{\text{out}}$	p_{out}	$\bar{d}_{\text{out}}$	p_{out}	$\bar{d}_{\text{out}}$	p_{out}	$\bar{d}_{\text{out}}$	p_{out}	$\bar{d}_{\text{out}}$	p_{out}	$\bar{d}_{\text{out}}$	p_{out}	$\bar{d}_{\text{out}}$
11	0.950	0.455	0.939	0.451	0.942	0.452	0.972	0.511	0.971	0.513	0.969	0.513	0.983	0.528	0.997	0.809	0.998	0.807
21	0.948	0.462	0.947	0.452	0.943	0.451	0.967	0.503	0.971	0.506	0.973	0.513	0.983	0.521	0.999	0.831	0.999	0.830
31	0.951	0.435	0.937	0.423	0.935	0.432	0.960	0.498	0.954	0.501	0.958	0.509	0.989	0.542	0.997	38.340	1.000	32.808
41	0.941	0.442	0.937	0.445	0.946	0.435	0.967	0.506	0.969	0.514	0.955	0.506	0.973	0.579	0.990	418.480	0.990	418.484
12	0.961	0.479	0.959	0.467	0.960	0.467	0.969	0.521	0.967	0.525	0.969	0.520	0.974	0.526	1.000	0.809	1.000	0.809
22	0.962	0.470	0.960	0.462	0.963	0.466	0.971	0.523	0.970	0.524	0.971	0.528	0.975	0.552	0.999	0.785	0.999	0.785
32	0.951	0.462	0.952	0.459	0.943	0.464	0.966	0.511	0.968	0.512	0.976	0.513	0.984	0.549	1.000	0.873	1.000	0.873
42	0.953	0.470	0.947	0.464	0.944	0.463	0.969	0.519	0.968	0.521	0.972	0.517	0.975	0.556	0.999	0.825	0.999	0.825
13	0.972	0.521	0.971	0.507	0.967	0.509	0.971	0.509	0.969	0.511	0.978	0.519	0.982	0.535	1.000	0.877	1.000	0.877
23	0.972	0.497	0.969	0.494	0.970	0.496	0.972	0.507	0.971	0.516	0.963	0.515	0.975	0.515	1.000	0.873	1.000	0.873
33	0.964	0.519	0.962	0.506	0.963	0.509	0.966	0.513	0.973	0.516	0.970	0.509	0.979	0.549	1.000	0.820	1.000	0.820
43	0.965	0.499	0.963	0.499	0.963	0.494	0.964	0.505	0.967	0.505	0.969	0.510	0.972	0.772	0.990	0.762	0.990	0.762
14	0.969	0.509	0.968	0.490	0.963	0.494	0.969	0.502	0.966	0.504	0.978	0.505	0.981	0.691	0.999	0.829	0.999	0.829
24	0.968	0.503	0.954	0.494	0.954	0.495	0.962	0.503	0.963	0.519	0.980	0.541	0.984	0.644	1.000	0.811	1.000	0.810
34	0.965	0.514	0.961	0.511	0.962	0.508	0.966	0.506	0.966	0.514	0.961	0.551	0.993	0.756	1.000	0.868	1.000	0.868
44	0.966	0.512	0.961	0.497	0.960	0.501	0.969	0.507	0.968	0.509	0.964	0.506	0.970	0.796	1.000	0.871	1.000	0.871
15	0.971	0.515	0.960	0.497	0.960	0.502	0.965	0.506	0.967	0.489	0.970	0.553	0.976	0.731	1.000	0.860	1.000	0.860
25	0.965	0.520	0.962	0.500	0.960	0.494	0.965	0.522	0.964	0.521	0.965	0.512	0.973	0.566	1.000	0.814	1.000	0.816
35	0.974	0.516	0.962	0.495	0.962	0.501	0.974	0.512	0.976	0.513	0.962	0.503	0.985	0.756	1.000	0.877	1.000	0.877
45	0.973	0.513	0.963	0.492	0.961	0.488	0.970	0.515	0.971	0.517	0.967	0.517	0.975	0.740	1.000	0.880	1.000	0.879
16	0.958	0.452	0.947	0.448	0.943	0.439	0.960	0.503	0.965	0.503	0.962	0.504	0.965	0.537	0.987	0.884	0.987	0.880
26	0.941	0.465	0.938	0.458	0.940	0.462	0.964	0.502	0.966	0.504	0.969	0.502	0.974	0.554	0.994	1.650	0.992	1.649
36	0.953	0.453	0.943	0.457	0.945	0.460	0.966	0.503	0.961	0.510	0.965	0.508	0.966	0.508	0.994	0.894	0.994	0.894
46	0.950	0.459	0.950	0.459	0.948	0.457	0.968	0.513	0.963	0.521	0.972	0.523	0.973	0.618	0.996	0.819	0.998	0.811
17	0.980	0.517	0.960	0.506	0.957	0.495	0.963	0.505	0.963	0.504	0.963	0.509	0.980	0.538	1.000	0.910	1.000	0.910
27	0.975	0.518	0.952	0.506	0.950	0.497	0.965	0.511	0.963	0.513	0.979	0.522	0.983	0.528	1.000	0.896	1.000	0.896
37	0.977	0.519	0.973	0.502	0.975	0.511	0.975	0.503	0.979	0.504	0.979	0.509	0.983	0.540	1.000	0.835	1.000	0.839
47	0.980	0.516	0.967	0.505	0.966	0.490	0.971	0.505	0.976	0.503	0.980	0.515	0.985	0.532	1.000	0.879	1.000	0.879

Note: "Naive" is the naive method by adopting an algorithm in Section 3.1 without measurement error correction. "SIMEXBoost(π)" is the proposed SIMEXBoost method with $\pi = 0.05$ or 0.005. "SIMEX(π)" reflects the implementation of the SIMEX method for $\pi = 0.05$ or 0.005 without using the boosting method for variable selection. "LBPIC" indicates the implementation of Equation (12) without measurement error correction and variable selection. "PIC (Gao et al.)" gives the estimation without considerations of length-biased sampling, measurement error correction, and variable selection. "IC_Bayesian" and "IC_Par" are the IC-based methods derived by the R package `icenReg`.

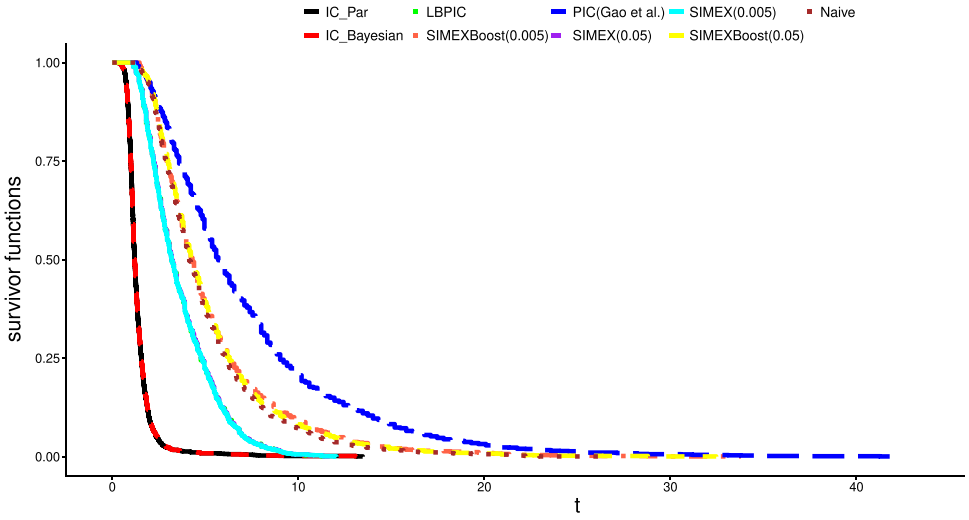

FIGURE 1 Survivor curves for permanent incisors with label 11. This figure appears in color in the electronic version of this paper, and any mention of color refers to that version.

different magnitudes of measurement error effects, which seems to show that the proposed method provides robust results. We observe that estimated survivor function based on the proposed method shows slightly higher possibility to emergence permanent teeth than that based on the naive method without measurement error correction. In addition, the estimated curves obtained by the proposed method are above of curves with all covariates accommodated, which might be caused by the involvement of non-informative covariates. It is also interesting to see that the curve determined by PIC(Gao et al.) is obviously higher than others, while two IC-based methods (IC_Bayesian and IC_Par) give the lowest curves among others, which might be the impacts induced by the ignorance of biased sampling and different methodologies.

Finally, to see the complexity of estimation methods, we examine the computation time by using the R function `proc.time()` to record the CPU time (in seconds). The runtimes are based on using Intel(R) Core(TM) i5-8250U CPU @ 1.60 GHz. Take a tooth label 11 as an example, with the SIMEX procedure accommodated, SIMEXBoost(π) and SIMEX(π) require approximate 5382.533 and 5400.742 s under $\pi = 0.05$ or 5356.53 and 5485.201 s under $\pi = 0.005$ to derive the estimators, respectively. Unsurprising, due to the involvement of B and $\mathcal{Z}$ in Stage 2, computation time based on measurement error correction by using the SIMEX method is longer than other methods without measurement error correction, such as naive (59.519 s), LBPIC (60.094 s), PIC (Gao et al.) (439.732 s), IC_Bayesian (178.810 s), and IC_Par (76.900 s). Consequently, we comment that the SIMEXBoost method provides the precise estimator and outperforms other methods, but longer computation time is the price that one should pay.

5 | SUMMARY AND FURTHER REMARKS

In this paper, we encounter survival data with irrelevant and error-prone covariates accommodated, where responses are biased and incomplete due to length-biased sampling and partly interval-censoring. To tackle those challenges, we first adopt the Buckley–James formulation to develop the estimating function based on the AFT model and address length-biased sampling as well as PIC. To deal with variable selection, we develop the SIMEX-Boost algorithm, which employs the SIMEX strategy and the boosting procedure to correct for measurement error and do variable selection simultaneously. Throughout simulation studies in the supporting information, we find that the proposed method successfully eliminates measurement error effects and correctly retains informative covariates. To see the robustness of the proposed method and the validity of Equation (7) for adjusting the censoring effects, we also examine different percentages of censoring. Numerical results in the supporting information show that biases, variances, and mean-squared errors are stable and do not have significant difference with the change of censoring rates. For the comparisons with the proposed method, we find that, without measurement error correction, simply employing the boosting method may falsely retain irrelevant covariates. Compared with Gao et al. (2017) who considered AFT models under the PIC data or the IC_Bayesian and IC_Par methods that treat PIC data as the IC data, we observe that ignoring length-biased sampling effect may induce biases in the estimators. To assess the performance of the estimator with the availability of auxiliary information, we conduct

a series of simulation settings and summarize numerical results in Appendix B.5 of the supporting information due to the limited space of the main text. In general, under the estimation from repeated measurements or validation data, the proposed method still produces satisfactory estimators. Finally, to show the flexibility of the proposed method, we examine the scenario $\delta_i = 0$, which refers to the conventional interval-censored data. According to the finding in Appendix B.6 of the supporting information, the proposed method is still valid to handle measurement error correction as well as variable selection. Although the proposed SIMEXBoost method has the best performance among other competitive methods listed in this paper, as summarized in Section 4 and simulation studies in the supporting information, SIMEXBoost requires longer computation time, which is due to the number of iterations I for the boosting procedure as well as a value B and a sequence $\mathcal{Z}$ for the SIMEX method. To reduce the computation time, one may require stronger computational tools or enhance programming algorithms to make computation faster. While we have not derived theoretical results in the current development, we examine the normality test for the estimators. Based on the simulation studies in Appendix B.2 of the supporting information, we adopt the Shapiro–Wilk normality test as well as the Q–Q plot to the proposed estimator under repeated simulations, and find that the proposed estimator $\widehat{\boldsymbol{\beta}}$ follows a normal distribution. Detailed description can be found in Appendix B.2 of the supporting information.

The current development in this paper has some potential extensions. First, we primarily focus on continuous and binary covariates since they can be modeled by measurement error models (14) and (15) in the current development, which are also frequently considered in measurement error analysis. The proposed method can be extended to handle error-prone ordinal or counted data and the mismeasurement can be adjusted by the same strategy as the generation of working data (17), provided that the corresponding measurement error models are well established. This issue should be carefully explored in our future work. Second, we primarily consider length-biased sampling in this paper, which specifies the truncation time to follow a uniform distribution. As commented by a referee, testing the length-biased assumption and checking relevant conditions are crucial issues. For example, one can check if the incidence of disease onset follows a stationary Poisson distribution and check if the truncation time follows the other distributions. In addition, one can further test if truncation time is independent of the failure time (e.g., Tsai, 1990). While those discussions are interesting and important, they require careful exploration based on our setting because of the involvement of measurement error and interval-censored mechanism. Those extensions can be our future research topic.

ACKNOWLEDGMENTS

The authors would like to thank the editorial team for useful comments to improve the initial manuscript. The authors specially thank Dr. Zhong-Lin Tsai, a dentist at Department of Dentistry, Wan Fang Hospital, Taipei Medical University, for sharing the dentistry knowledge to enhance the background of real-data analysis. Chen's research was supported by National Science and Technology Council with grant ID 110-2118-M-004-006-MY2.

DATA AVAILABILITY STATEMENT

The dataset of the Signal Tandmobiel Study, named `tandmob2.RData`, is available in the corresponding author's Github, whose link is placed in the supporting information.

ORCID

Li-Pang Chen https://orcid.org/0000-0001-5440-5036

REFERENCES

Aktan, A.M., Kara, I., Sener, I., Bereket, C., Celik, S., Kirtay, M., Ciftci, M.E. & Arici, N. (2012) An evaluation of factors associated with persistent primary teeth. *European Journal of Orthodontics*, 34, 208–212.

Brown, B., Miller, C.J. & Wolfson, J. (2017) ThrEEBoost: thresholded boosting for variable selection and prediction via estimating equations. *Journal of Computational and Graphical Statistics*, 26, 579–588.

Buckley, J. & James, I. (1979) Linear regression with censored data. *Biometrika*, 66, 429–436.

Cai, T. & Betensky, R.A. (2003) Hazard regression for interval-censored data with penalized spline. *Biometrics*, 59, 570–579.

Carroll, R.J., Ruppert, D., Stefanski, L.A. & Crainiceanu, C.M. (2006) *Measurement error in nonlinear model*. Boca Raton, FL: Chapman and Hall.

Chen, L.-P. (2019) Semiparametric estimation for cure survival model with left-truncated and right-censored data and covariate measurement error. *Statistics and Probability Letters*, 154, 108547.

Chen, L.-P. (2020) Semiparametric estimation for the transformation model with length-biased data and covariate measurement error. *Journal of Statistical Computation and Simulation*, 90, 420–442.

Chen, L.-P. (2021) Variable selection and estimation for the additive hazards model subject to left-truncation, right-censoring and measurement error in covariates. *Journal of Statistical Computation and Simulation*, 90, 3261–3300.

Chen, L.-P. & Yi, G.Y. (2020) Model selection and model averaging for analysis of truncated and censored data with measurement error. *Electronic Journal of Statistics*, 14, 4054–4109.

Chen, L.-P. & Yi, G.Y. (2021a) Semiparametric methods for left-truncated and right-censored survival data with covariate measurement error. *Annals of the Institute of Statistical Mathematics*, 73, 481–517.

Chen, L.-P. & Yi, G.Y. (2021b) Analysis of noisy survival data with graphical proportional hazards measurement error models. *Biometrics*, 77, 956–969.

Du, M. & Sun, J. (2021a) Statistical analysis of interval-censored failure time data. *Chinese Journal of Applied Probability and Statistics*, 37, 627–654.

Du, M. & Sun, J. (2021b) Variable selection for interval-censored failure time data. *International Statistical Review*, 1–23.

Du, M., Zhao, H., & Sun, J. (2021) A unified approach to variable selection for Cox's proportional hazards model with interval-censored failure time data. *Statistical Methods in Medical Research*, 30, 1833–1849.

Fu, W. & Simonoff, J.S. (2017) Survival trees for interval-censored survival data. *Statistics in Medicine*, 36, 4831–4842.

Gao, F., Zeng, D. & Lin, D.Y. (2017) Semiparametric estimation of the accelerated failure time model with partly interval-censored data. *Biometrics*, 73, 1161–1168.

Gao, F. & Chan, K.C. G. (2019) Semiparametric regression analysis of length-biased interval-censored data. *Biometrics*, 75, 121–132.

Huang, J. (1999) Asymptotic properties of nonparametric estimation based on partly interval-censored data. *Statistica Sinica*, 9, 501–519.

Kim, J.S. (2003) Maximum likelihood estimation for the proportional hazards model with partly interval-censored data. *Journal of the Royal Statistical Society, Series B*, 65, 489–502.

Küchenhoff, H., Lederer, W. & Lesaffre, E. (2007) Asymptotic variance estimation for the misclassification SIMEX. *Computational Statistics & Data Analysis*, 51, 6197–6211.

Lawless, J.F. (2003) *Statistical models and methods for lifetime data*. New York: Wiley.

Mandal, S., Wang, S. & Sinha, S. (2019) Analysis of linear transformation models with covariate measurement error and interval censoring. *Statistics in Medicine*, 38, 4642–4655.

Ning, J., Qin, J. & Shen, Y. (2011) Buckley–James-type estimator with right-censored and length-biased data. *Biometrics*, 67, 1369–1378.

Pan, C., Cai, B. & Wang, L. (2020) A Bayesian approach for analyzing partly interval-censored data under the proportional hazards model. *Statistical Methods in Medical Research*, 29, 3192–3204.

Pan, W. & Chappell, R. (2002) Estimation in the Cox proportional hazard model with left-truncated and interval censored data. *Biometrics*, 58, 64–70.

Scolas, S., Ghouch, A.E., Legrand, C. & Oulhaj, A. (2016) Variable selection in a flexible parametric mixture cure model with interval-censored data. *Statistics in Medicine*, 35, 1210–1225.

Shen, P.-S., Peng, Y., Chen, H.-J. & Chen, C.-M. (2022) Maximum-likelihoodestimation for length-biased and interval-censored data with a nonsusceptible fraction. *Lifetime Data Analysis*, 28, 68–88.

Song, X. & Ma, S. (2008) Multiple augmentation for interval-censored data with measurement error. *Statistics in Medicine*, 27, 3178–3190.

Sun, L., Li, S., Wang, L. & Song, X. (2021) Simultaneous variable selection in regression analysis of multivariate interval-censored data. *Biometrics*, 1–12.

Tsai, W.-Y. (1990) Testing the assumption of independence of truncation time and failure time. *Biometrika*, 77, 169–177.

Wang, L., McMahan, C.S., Hudgens, M.G. & Qureshi, Z.P. (2016) A flexible, computationally efficient method for fitting the proportional hazards model to interval-censored data. *Biometrics*, 72, 222–231.

Wang, P., Li, D. & Sun, J. (2021) A pairwise pseudo-likelihood approach for left-truncated and interval-censored data under the Cox model. *Biometrics*, 77, 1303–1314.

Wen, C.-C. & Chen, Y.-H. (2014) Functional inference for interval-censored data in proportional odds model with covariate measurement error. *Statistica Sinica*, 24, 1301–1317.

Wolfson, J. (2011) EEBOOST: a general method for prediction and variable selection based on estimating equation. *Journal of the American Statistical Association*, 106, 296–305.

Yao, W., Frydman, H. & Simonoff, J.S. (2019) An ensemble method for interval-censored time-to-event data. *Biostatistics*, 22, 198–213.

Yavuz, A. Ç. & Lambert, P. (2011) Smooth estimation of survival functions and hazard ratios from interval-censored data using Bayesian penalized B-splines. *Statistics in Medicine*, 30, 75–90.

Zhao, H., Wu, Q., Li, G. & Sun, J. (2020) Simultaneous estimation and variable selection for interval-censored data with broken adaptive ridge regression. *Journal of the American Statistical Association*, 115, 204–216.

Zhao, X., Zhao, Q., Sun, J. & Kim, J.S. (2008) Generalized log-rank tests for partly interval-censored failure time data. *Biometrical Journal*, 50, 375–385.

Zou, H. & Hastie, T. (2005) Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B*, 67, 301–320.

SUPPORTING INFORMATION

The supporting information contains simulation studies for the proposed method. Web Appendices, programming code, dataset, figures and tables referenced in Section 4 are available with this article at the Biometrics website on Wiley Online Library. All relevant information in this manuscript is available in the Github: https://github.com/lchen723/LBPIC-ME.git.

How to cite this article: Chen, L.-P. & Qiu, B. (2023) Analysis of length-biased and partly interval-censored survival data with mismeasured covariates. *Biometrics*, 1–12. https://doi.org/10.1111/biom.13898