Published in final edited form as:

Stat Med. 2018 March 30; 37(7): 1162–1177. doi:10.1002/sim.7575.

Using threshold regression to analyze survival data from complex surveys: With application to mortality linked NHANES III Phase II genetic data

Yan Li¹, Tao Xiao², Dandan Liao³, and Mei-Ling Ting Lee⁴

¹Joint Program for Survey Methodology, University of Maryland at College Park, College Park, MD, USA

²College of Mathematics and Statistics, Shenzhen University, Shenzhen, China

³Department of Measurement, Statistics and Evaluation, University of Maryland at College Park, College Park, MD, USA

⁴Department of Epidemiology and Biostatistics, University of Maryland at College Park, College Park, MD, USA

Abstract

The Cox proportional hazards (PH) model is a common statistical technique used for analyzing time-to-event data. The assumption of PH, however, is not always appropriate in real applications. In cases where the assumption is not tenable, threshold regression (TR) and other survival methods, which do not require the PH assumption, are available and widely used. These alternative methods generally assume that the study data constitute simple random samples. In particular, TR has not been studied in the setting of complex surveys that involve (1) differential selection probabilities of study subjects and (2) intracluster correlations induced by multistage cluster sampling. In this paper, we extend TR procedures to account for complex sampling designs. The pseudo-maximum likelihood estimation technique is applied to estimate the TR model parameters. Computationally efficient Taylor linearization variance estimators that consider both the intracluster correlation and the differential selection probabilities are developed. The proposed methods are evaluated by using simulation experiments with various complex designs and illustrated empirically by using mortality-linked Third National Health and Nutrition Examination Survey Phase II genetic data.

Keywords

cox proportional hazard; cure rate; intracluster correlation; pseudo-maximum likelihood estimation; stratified multistage sampling

1 | INTRODUCTION

Cox proportional hazards (PH) regression is a well-known methodology for analyzing survival data. PH regression focuses mainly on hazard ratios. Lee et al¹⁻⁴ proposed a series of threshold regression (TR) methodologies for time-to-event and survival analysis when data are collected by using simple random sampling (SRS). Compared to PH regression, TR provides a different perspective, giving attention to the underlying determinants of initial health status and differing patterns of survival during follow-up. For a simple TR illustration, a Wiener diffusion process $\{Y(t)\}$ can be postulated that corresponds to a latent degradation health status process. Censored survival data, combined with readings on degradation, can then be used to estimate initial health status and to predict residual survival time. Covariates, such as treatment variables, risk factors, and base-line conditions, are related to the model parameters through generalized linear regression functions. The TR method has greater potential for the study of treatment efficacy than do conventional survival models because it does not assume PH and focuses on underlying determinants of survival. It also easily handles a cure rate if one exists in the dataset.

In this paper, we suppose that the sample has been drawn from a finite population by using a complex sampling design, such as in the National Health and Nutrition Examination Survey Epidemiological Follow-up Study.⁵ The use of complex sampling (eg, stratified multistage cluster sampling) for the selection of subjects in social and biomedical studies is common. In addition to the cost-effectiveness and time-effectiveness, the use of complex sampling (eg, stratified multistage cluster sampling) yields representative samples from the study population and avoids biased selection. To realize the full potential of these advances, however, there are at least 2 complexities introduced by a complex sampling design. *First*, differential selection probabilities of study subjects as well as adjustments for nonresponse and noncoverage (that is, incomplete coverage of the population at risk) result in varying population weights (PWs) across surveyed subjects. Here, PWs reflect the number of population members represented by each participant in the study, providing for a capability of making inferences to the target population of interest. *Second*, with a multistage cluster sample design, variances can be inflated due to correlations within the hierarchical clusters of study subjects. Because of the above 2 complexities, the sample distribution can be different from the underlying population distribution from which the sample is selected. Ignoring the survey features may result in biased estimators, as well as erroneous confidence intervals and tests of hypotheses of the model parameters for the target population.^{6,7} In general, panel surveys or even cross-sectional health surveys with follow-up, such as linkage to a mortality database, are examples where this type of analysis may be suitable.

In complex sampling, the design parameters for the survey are often related to the true hazard function but are not explicitly part of the model being fitted. For example, the geographic regions controlled in sample designs are often related to environmental factors, ethnical/racial groups, or socioeconomic status, which may also be more or less related to human diseases. In addition, for complex disease such as cancer mortality, not all risk factors are known. In general, the model being fitted is at best an approximation. All these will affect the validity of model-based inferences. Design-based estimation methods,

however, estimate well-defined finite-population quantities, no matter whether the models hold or not.⁸

The estimation bias produced by ignoring complex designs in the analysis can be avoided by weighting sampled subjects by the inverse of their respective selection probabilities, as proposed by Horvitz and Thompson,⁹ and extended to the Cox PH setting for use in complex surveys by Binder,¹⁰ Lin,¹¹ and Boudreau and Lawless.¹² Weighted procedures yield unbiased estimators that are more “robust” against model misspecification at a cost in efficiency that is acceptable. For the estimation of the variance of weighted estimates, the Taylor linearization method is applied to obtain a finite population design-based variance estimator.⁸

In this paper, we develop a design-based procedure to estimate parameters of the TR model for data collected from complex surveys. Not requiring the PH assumption, the TR model accounts for design complexities and is appropriate for analyzing time-to-event data collected from complex survey designs. Innovative features of our methodology include use of (1) pseudo-maximum likelihood estimation (pseudo-MLE) in estimating TR model parameters and (2) computationally efficient Taylor linearization variance estimators that account for differential selection probabilities and intracluster correlations. We conducted an in-depth analysis of genetic data collected in phase II from the Third National Health and Nutrition Examination Survey (NHANES III). The dataset, linked with death certificate records (DCR) by using respondent social security number, provides an opportunity to apply the TR model to a study of time to death related to all types of causes.

The rest of this paper is organized as follows. We present the proposed TR estimator and its variance estimation in Section 2. The TR estimator and its variance estimator are applicable for studies with various complex sampling designs. Limited simulation studies, under both SRS and complex sampling, are conducted to evaluate the new estimators in unbiasedness and variances in Section 3. We apply the proposed extended TR model to assess the risk of all-cause mortality associated with blood lead level and the *ALADG177C* single-nucleotide polymorphism for NHANES III in Section 4. We wrap up the paper with closing remarks in Section 5.

2 | METHODS

2.1 | First-hitting-time-based threshold regression model

Let $\{Y(t); \text{ for } t \geq 0\}$ be a time-dependent latent health status stochastic process, with $Y(0) = y_0 > 0$ as the initial health status. The event occurrence time T is defined as

$$T = \inf\{t: Y(t) = Q\},$$

the first time the sample path of the health status process reaches threshold Q . The threshold Q is a critical health state, disease state, or other medical end point, such as death, the onset of disease, or hospital discharge. The first-hitting-time (FHT)-based TR model has been applied to a variety of health status processes. Probabilistic properties of the underlying process $\{Y(t)\}$, whether a standard Brownian motion, a gamma process, a Poisson process,

or others, can lead to various lifetime distributions. If $\{Y(t)\}$ follows a Wiener process, then the survival function $\bar{F}(t)$ of the event time T has an inverse Gaussian distribution. A review of examples of FHT-based TR models can be found in Lee and Whitmore.^{2,3}

2.2 | Threshold regression estimators for simple random sampling

Consider $\bar{F}(t|x; \eta)$ as the survival function of the event time T , given the covariate vector x , where η denote the vector of regression coefficients. The underlying process of time to event is connected to covariates (such as age and smoking history) by regression link functions. For a large simple random sample (or a finite population) of size N from an infinite population, the log-likelihood function is

$$\ln L(\theta) = \sum_{i=1}^{N_1} \ln f(t_i|x_i, \eta) + \sum_{j=N_1+1}^N \ln \bar{F}(t_j|x_j, \eta), \quad (1)$$

where $f(t_i|x_i, \eta)$ is the probability density function of the survival function, measuring the contribution made by the sampled subject i whose exact death time is observed and therefore provides information on the probability that the event is occurring at time t_i . For sampled subject j , who lives to the end of the study, however, all we know about this subject is that the event time is larger than the study time. Therefore, the information contributed by the surviving subject j in the sample likelihood function is the survival function evaluated at the corresponding study time t_j for this subject: $\bar{F}(t_j|x_j, \eta) = 1 - F(t_j|x_j, \eta)$. The censoring process is independent of the event process and thus noninformative. Among the N subjects in the sample, subjects with observed death times are indexed from 1 to N_1 and subjects with right-censored observations are indexed from $N_1 + 1$ to N . To maximize the above log-likelihood, we determine the finite population parameter η_N to be the value of model parameters maximizing the Equation 1 across N subjects in the finite population. The finite population parameter η_N is the parameter we wish to estimate from a complex sample of size n drawn from the finite population.

2.3 | Threshold regression estimators for complex sampling

Under complex sampling, n subjects in the sample are selected with differential probabilities, and each sampled subject has an associated PW, interpreted as the number of subjects in the population that can be represented by the sampled individual. To obtain design-consistent estimators, we need to incorporate the PW into the log-likelihood function. Pseudo-maximum likelihood estimation (MLE) methods¹³ are used to estimate the regression coefficients based on the pseudo log-likelihood function, given by

$$\ln L^w(\eta) = \sum_{i=1}^{n_1} w_i \ln f(t_i|x_i, \eta_N) + \sum_{j=n_1+1}^n w_j \ln \bar{F}(t_j|x_j, \eta_N), \quad (2)$$

where $w_i(\eta)$ denotes the PW for unit i (η) and n_1 denotes the number of subjects with observed death times in the sample. The first derivatives of Equation 2 with respect to η_N are referred to as pseudo-score functions $S^w(\eta_N)$, and the estimator $\hat{\eta}$ that maximizes Equation 2 with $S^w(\hat{\eta}) = 0$ is referred to as the TR estimator. By incorporating PWs in the log-likelihood, Equation 2 is estimating the full log-likelihood as if we have sampled the entire finite population. Li et al.¹⁴ elaborated on the motivation for considering PW when estimating model parameters and proved the desirable property of design-consistency of their proposed nonlinear estimator. The Newton-Raphson method is used to find the TR estimator ($\hat{\eta}$) that maximizes the pseudo log-likelihood Equation 2.

2.4 | Variance estimation of threshold regression estimators for complex sampling

There are various statistical methods proposed for the estimation of the variance of a nonlinear estimator $\hat{\eta}$. A computationally efficient approach is the Taylor linearization method, which approximates nonlinear estimators by a weighted total of the Taylor deviates for each observation.¹⁵ The Taylor deviate denotes the implicit influence of each observation on the value of the nonlinear estimator. In brief, the Taylor deviate for the u th unit ($u = 1, \dots, n$), which is obtained by taking the derivative of the nonlinear estimator $\hat{\eta}$ with respect to the weight for the u th unit, is a measure of the influence of the u th observation on the estimate $\hat{\eta}$. The computation of the variance of an estimator would be equivalent to computing the variance of the weighted Taylor deviates, considering sample designs. Specifically, we first derive the Taylor deviate of $\hat{\eta}$ by taking the derivative of the pseudo score functions $S^w(\hat{\eta}) = 0$ with respect to the weight of the u th individual w_u , given by

$$\frac{\partial S^w(\hat{\eta})}{\partial w_u} = \frac{\partial}{\partial w_u} \sum_{i=1}^n w_i s_i(\hat{\eta}) = s_u(\hat{\eta}) + \sum_{i=1}^n w_i \frac{\partial s_i(\hat{\eta})}{\partial \hat{\eta}} \frac{\partial \hat{\eta}}{\partial w_u} = 0. \quad (3)$$

Solving Equation 3 for the Taylor deviate $\frac{\partial \hat{\eta}}{\partial w_u}$, we have

$$z_u = \frac{\partial \hat{\eta}}{\partial w_u} = -s_u(\hat{\eta}) \left(\sum_{i=1}^n w_i \frac{\partial s_i(\hat{\eta})}{\partial \hat{\eta}} \right)^{-1} = -s_u(\hat{\eta}) (I^w)^{-1},$$

where $I^w = \sum_{i=1}^n w_i \frac{\partial s_i(\hat{\eta})}{\partial \hat{\eta}}$.

$\text{Var}(\hat{\eta})$ can be approximated by the variance of the weighted total of z_u , ie, $\text{Var}(\sum_{u=1}^n w_u z_u)$.

The standard analytical formulas from survey sampling theory can then be applied for estimating variances of a weighted total under different sample designs.¹⁶ In general, for complex survey designs, the target population is first divided into primary sampling units (PSUs), which can be clusters of individuals (eg, counties, small cities, or subdivisions of larger cities) or single individuals. The PSUs are then grouped into H strata, formed to be approximately homogeneous with respect to specific characteristics of the PSUs. At the first stage of sampling, m_h PSUs are randomly sampled from each stratum h in the population.

Within the sampled PSUs, stratification and clustered sampling can be applied to ultimately select individuals.

The $\text{Var}(\hat{\eta})$ for a general stratified multistage sample design is estimated by

$$\text{var}(\hat{\eta}) = \sum_{h=1}^H \frac{m_h}{m_h - 1} \sum_{l=1}^{m_h} (z^{hl} - \bar{z}^h)(z^{hl} - \bar{z}^h)^T \quad (4)$$

with $z^{hl} = \sum_{u \in (hl)} w_u z_u$ being the weighted total of z_u across all sampled subjects in PSU l of stratum h and $\bar{z}^h = \sum_{l \in (h)} z^{hl} / m_h$ being the mean of the PSU totals across all sampled PSUs in stratum h . The associated degrees of freedom (df) are number of first-stage sample units, ie, number of sampled PSUs minus the number of sampling strata.⁸

Variance estimator (4) is commonly used as an approximation for variance estimation when available design information is limited, assuming that the first-stage sampling is with replacement and ignores the finite population correction. To obtain an unbiased variance estimator, however, the computational effort is usually considerable and may require the second-order inclusion probabilities of all the remaining stages. It can be shown that the variance estimator (4) overestimates the true variance, and if the sampling fraction is small (which is often the case in real large-scale surveys), the amount of overestimation is unimportant.¹⁷ The variance estimator of Equation 4 simplifies the estimation by requiring only the PSU totals. For more complete discussion of design-based variance estimation, please see Wolter.¹⁸

3 | SIMULATIONS

In the simulation studies, we model the latent health process $\{Y(t)\}$ by a commonly used Wiener diffusion process. The distribution of the FHT, denoted by variable t , is then an inverse Gaussian distribution with the following probability density function

$$f(t|\mu, \sigma^2, y_0) = \frac{y_0}{\sqrt{2\pi\sigma^2 t^3}} \exp\left(-\frac{(y_0 + \mu t)^2}{2\sigma^2 t}\right)$$

and cumulative distribution function:

$$F(t|\mu, \sigma^2, y_0) = \Phi\left(-\frac{y_0 + \mu t}{\sqrt{\sigma^2 t}}\right) + \exp\left(-\frac{2y_0\mu}{\sigma^2}\right) \Phi\left(-\frac{y_0 - \mu t}{\sqrt{\sigma^2 t}}\right),$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution. Three parameters of the Wiener process are involved: μ , y_0 , and σ^2 . Parameter μ , called the drift of the Wiener process, is the mean change per unit of time in the level of the sample path. The sample path approaches the threshold if $\mu < 0$. Parameter y_0 is the initial value of the process and is taken as positive. Parameter σ^2 represents the variance per unit of time of the process.

Note that if the drift rate $\mu > 0$, the Wiener process may never hit the boundary at zero and hence there is a probability that the FHT is infinite, with $P(FHT = \infty) = 1 - \exp(-2y_0\mu/\sigma^2)$.¹⁹ Because the degradation process of health status is latent with undefined measurement scale, we choose to set the variance parameter $\sigma^2=1$ without loss of generality. Next, we set up simultaneous regressions for parameters μ and $\ln(y_0)$ on the covariate vector X with linear regression links:

$$\ln(y_0) = \gamma_0 + \gamma_1 X_1 + \dots + \gamma_k X_k = X' \gamma$$

and

$$\mu = \beta_0 + \beta_1 X_1 + \dots + \beta_k X_k = X' \beta$$

where $\gamma = (\gamma_0, \gamma_1, \dots, \gamma_k)'$ and $\beta = (\beta_0, \beta_1, \dots, \beta_k)'$ are regression coefficient vectors, and $\eta = (\beta', \gamma')'$. The regression coefficients measure the association of the covariates with the initial health status (y_0) and with how fast the health status drifts to the event (μ).

For variance estimation, we derive $S^w(\hat{\eta}) = (S^w(\hat{\beta}), S^w(\hat{\gamma}))^T$ by differentiating the $\ln L^w(\hat{\eta})$ with respect to $\hat{\eta}$, ie,

$$s^w(\hat{\beta}) = \sum_{i=1}^{n_1} w_i \left[-\exp(X_i' \gamma) - X_i' \beta t_i \right] X_i' + \sum_{j=n_1+1}^n w_j \left[\frac{2 \exp(X_j' \gamma)}{1 - F(t^{(j)} | \mu^{(j)}, y_0^{(j)})} \exp[-2 \exp(X_j' \gamma) X_j' \beta] \Phi \left(-\frac{\exp(X_j' \gamma) - X_j' \beta t_j}{\sqrt{t_j}} \right) \right] X_j'$$

and

$$s^w(\hat{\gamma}) = \sum_{i=1}^n w_i \left[1 - \frac{\exp(2X_i' \gamma)}{t_i} - X_i' \beta \exp(X_i' \gamma) \right] X_i' + \sum_{j=n_1+1}^n w_j \frac{2 \exp(X_j' \gamma)}{1 - F(t^{(j)} | \mu^{(j)}, y_0^{(j)})} \left[\frac{1}{\sqrt{2\pi t_j}} \exp \left[-\frac{(\exp(X_j' \gamma) + X_j' \beta t_j)^2}{2t_j} \right] + X_j' \beta \exp[-2 \exp(X_j' \gamma) X_j' \beta] \Phi \left(-\frac{\exp(X_j' \gamma) - X_j' \beta t_j}{\sqrt{t_j}} \right) \right] X_j'.$$

Taking the derivative of the pseudo score functions $S^w(\hat{\eta}) = 0$ with respect to the weight of the u th individual w_u by following Equation 3, the resulting Taylor deviates are then obtained and substituted into Equation 4 to estimate the variance of $\hat{\eta}$.

3.1 | Simulation setups

We generated a covariate x from a normal distribution with mean 5 and standard deviation 1. Given the parameter values $\gamma_0 = 0.1$, $\gamma_1 = 0.2$, $\beta_0 = -0.2$, and $\beta_1 = -0.3$, we generate y_0 and μ by the models

$$\ln(y_0) = \gamma_0 + \gamma_1 x, \mu = \beta_0 + \beta_1 x.$$

The outcome survival time, ie, FHT t , is then generated from the inverse Gaussian distribution, IG ($y_0, \mu, \sigma^2 = 1$). We set the right censoring time at values of 3 so that proportions of 8% of the population are right censored. We generated a finite population of size $N = 400\,000$, from which 2000 samples are independently selected with various sample designs (as described in Sections 3.2–3.4).

For comparison purposes, maximum likelihood estimator (MLE) and the proposed TR estimator are evaluated in 4 ways in (1) the relative bias (ie, the deviation of the average of the 2000 estimates from the true values divided by the true values of regression coefficients), (2) the empirical variance (emp.var) (ie, the variance of the 2000 estimates of regression coefficients), (3) the MLE variance estimator (mle.var) (ie, the average of the 2000 MLE variance estimates, derived from the inverse of the Hessian matrix for the pseudo-likelihood function), and (4) the proposed Taylor linearization variance estimator (taylor.var) (ie, the average of the 2000 estimated Taylor variances). The MLE is the estimator developed from Equation 2 but specifying sample weights common value of one, assuming that all the units are selected by SRS with replacement. Therefore, the MLE estimators do not consider differential selection probabilities, and the corresponding variance, mle.var, is estimated by assuming independence among sampled units. In addition to the relative bias and the empirical variance, we present the ratio taylor.var/emp.var in simulation tables. The closer the ratio of the variance estimator to the emp.var is to the value of 1, the better is the performance of the variance estimator.

In practice, complex sample designs involve (1) differential weighting (due to differential selection probabilities, nonresponse adjustment, or poststratification adjustment) and/or (2) hierarchical clustering. We conducted 3 simulation studies with different sampling designs to investigate how the differential weighting, clustering, and their combination affect the performance of the proposed estimators.

3.2 | One-stage sample designs with 3 different selection probabilities

In simulation results presented in Table 1, we study the performance of the proposed TR estimator under 1-stage sample designs with differential selection probabilities (that is, differential weighting). A sample of $n = 1000$ observations are randomly selected by 4 sampling schemes: SRS; sampling without replacement with selection probability proportional to the size (PPS), where the size is defined by $v_x = x + |e|$ with $e \sim N(0, 1)$. We denote the design by PPS _{x} ; PPS of size $v_t = t + |e|$, denoted by PPS _{t} ; and PPS of size $v_{tx} = x + 2t + |e|$, denoted by PPS _{tx} . Note that the sample weights for designs SRS and PPS _{x} are noninformative, whereas PPS _{t} and PPS _{tx} designs are related to the true hazard function, and therefore lead to informative sample weights.^{20–22} For sample designs with informative weights, we expect that the unweighted estimation would be badly biased.¹⁰ Table 1 shows the simulation results. It can be observed that across the designs, TR is approximately unbiased and the developed Taylor variance estimates approximate the true variances well

with the variance ratios, defined by $\frac{\text{taylor.var}}{\text{emp. var}}$, close to the value of 1, whereas the MLE estimator is biased when the sample weights are informative. The MLE variance estimator (not shown) considerably underestimates the true variance of the TR estimator across the designs.

3.3 | Two-stage sample designs with 4 different intracluster correlations

To study the clustering effect, we introduce the intracluster correlation within clusters by (1) sorting the population by the value of a random variable (RV) that is correlated with the FHT t , where RV follows normal distribution with the mean of t and the variance of σ_w^2 ; (2) grouping sequentially adjacent 100 units into $M = 4000$ clusters for a population of size $N = 400\,000$ units. We define the outcome survival time as the smaller value of the time-to-event (ie, the survival time generated in Section 3.1) or the time to a competing risk, which is a function of 1 plus a RV that follows an exponential distribution with the mean of 2 and the variance of 4. As a result, we have in the population approximately 31% censoring rate, including 28.4% of units who are censored due to the competing risk.

Two-stage cluster sampling is conducted with SRS without replacement at both stages to select a sample of size $n = 2000$. As a result, all the selected units have the same sample weight, and thus, the clustering effect is not confounded with the differential weighting effect in the performance of the TR estimators. In the simulation, we vary the number of clusters selected at the first stage $m (=25, 50, 100, \text{ and } 400)$ and the values of σ_w^2 so that the intracluster correlation $\rho_t = 0.05, 0.1, 0.2, \text{ and } 0.3$.¹⁶ It would be expected that the proposed Taylor linearization variance estimators, considering cluster sampling, produce estimates close to the true variance, while the variance based on an independence assumption (denoted by mle.var) would underestimate the true variance.

Table 2 shows the simulation results when $\rho_t = 0.3$ with the varying number of clusters m and the fixed total sample size $n = 2000$. Results from other ρ_t values show the similar patterns and therefore not shown. It can be observed that the relative biases are closer to zero and the empirical variances decrease as the number of clusters m increases. The MLE.var consistently underestimates the empirical variances, especially when the number of clusters is small with the mle.var/emp.var ranging from 0.32 to 0.65 when $m = 25$. This result is as our expectation because the MLE.var ignores the intracluster correlation within clusters, and the larger the cluster size ($=n/m$), the larger the design effects.

The proposed taylor.var , developed under asymptotic theorem, however, performs well when the number of clusters is sufficiently large, but tend to slightly underestimate the empirical variance with $\text{taylor.var/emp.var}$, ranging from 0.84 to 0.93 when the number of clusters is small $m = 25$.

3.4 | A complex design with 2-stage sample, different weighting, clustering effects, and censoring rates

Differential weighting and clustering effects are separately studied in simulations 1 to 2. In simulation results presented in Table 3, we study the performance of TR under a complex

design involving both differential weighting and clustering effects, like the design in NHANES III. This was carried out by sorting the population by the values of quantiles of t and continuous x from smallest to largest, and then grouping the sorted population into $M = 4000$ clusters sequentially with each cluster having size 100. We sample $m = 400$ clusters without replacement with selection probabilities proportional to the size of $\bar{v}$, the average value of v_{ix} within clusters, where v_{ix} is defined in Section 3.2 for simulation 1. The selection probability for cluster i is approximated by $p_i = \frac{m \times \bar{v}_i}{\sum_{i=1}^M \bar{v}_i}$ for $i = 1, \dots, M$, and the sample weight is the inverse of the selection probability, $w_i = 1/p_i$. Within each of the 400 clusters sampled, we selected 5 observations by SRS. Accordingly, the final sample weight for the j th selected individual in the i th selected cluster is $w_{ij} = 20 \times w_i$. Under this design, we further study the robustness of the proposed estimator to various censoring rates by varying the right censoring time so that proportions of 8%, 15%, and 25% of the population are right censored. We would expect that the proposed estimators are approximately unbiased and their Taylor variance estimators are close to the true variances. Table 3 shows the simulation results. It can be observed that MLE estimates are consistently biased across various censoring rates while the TR estimators are approximately unbiased. The mle.var (not shown) underestimates the true variances with ratios of mle.var/emp.var considerably smaller than 1, especially for the mle.var of TR. This result is consistent with our expectation because mle.var ignores both the clustering effect and differential weighting for TR, whereas only the clustering effect is ignored for mle.var of MLE estimates.

3.5 | Extension to multivariate regression modeling of μ and $\ln(y_0)$ in threshold regression methods

In real data examples, we would expect various numbers and types of covariates to be considered in the regression models of μ and $\ln(y_0)$. We further generate a continuous covariate X from a normal distribution with mean 5 and standard deviation 1, a dummy covariate (eg, gender) from a binary distribution with probability of .5, and a 4-category covariate with a values of 1, 2, 3, and 4 randomly assigned to segments representing 30%, 30%, 30%, and 10% of the population, respectively. For example, we have 4 age groups (= 1, 2, 3, or 4), and accordingly, 3 dummy variables of age_1 , age_2 , and age_3 can be created by specifying $\text{age}_i = 1$ if $\text{age} = i$ and 0 otherwise, for $i = 1$ to 3. With specified parameter values of $\gamma_0 = 0.3$, $\gamma_1 = 0.2$, and $\gamma_2 = 0.1$ and $\beta_0 = -0.6$, $\beta_1 = -0.3$, $\beta_2 = -0.1$, $\beta_3 = -0.2$, $\beta_4 = -0.1$, and $\beta_5 = -0.4$, we generate y_0 and μ by the models

$$\begin{aligned}\ln(y_0) &= \gamma_0 + \gamma_1 X + \gamma_2 \text{gender} \\ \mu &= \beta_0 + \beta_1 X + \beta_2 \text{gender} + \beta_3 \text{age}_1 + \beta_4 \text{age}_2 + \beta_5 \text{age}_3.\end{aligned}$$

The FHT, t , is then generated from the inverse Gaussian distribution, $\text{IG}(y_0, \mu, \sigma^2 = 1)$. We set the right censoring time to 2.84 so that 10% of the observations are right censored. Again, the results (not shown) are consistent with our expectation: the MLE estimators, which do not consider the differential selection probabilities, are badly biased under designs with informative weights. The variance estimators assuming independence among sampled units underestimate the true variance under designs involving clustering and/or differential

weighting. The proposed TR estimators, under simple random sample design and various complex designs, are approximately unbiased. The developed Taylor variance estimates are close to the empirical variances.

4 | REAL DATA ANALYSIS

The NHANESIII is a complex, multistage sample survey that was conducted by the National Center for Health Statistics (NCHS), Centers for Disease Control and Prevention from 1988 to 1994 to collect detailed information on the health and nutritional status from the civilian, noninstitutionalized population of the United States over the age of 2 months. The US population was first partitioned into first-stage PSUs, usually counties or contiguous counties. The PSUs were grouped into strata, formed to be approximately homogeneous in the population size of PSUs. At the first stage, PSUs were randomly sampled. At the second and further stages, stratification and cluster sampling were used to sample households. At the final stage, individuals within sampled households were randomly selected. During the sampling procedure, children aged 2 months to 5 years, persons aged 60 and over, non-Hispanic blacks, and Mexican Americans were oversampled. At Phase II of NHANES III (1991–1994), genetic samples were collected. Genetic laboratory results and the DCR from the National Death Index were then linked to NHANES data through the NCHS Research Data Center to ensure confidentiality of the study participants. Please refer to previous studies^{23–25} for details on the sample design of NHANES III, the methods of records linkage of NHANES III to the DCR, and the description of the NHANES III genetic data. In this paper, we assess the risk of all-cause mortality associated with blood lead level and *ALAD* genotype for a subset of NHANES III Phase II participants to illustrate the proposed TR estimator. A total of 3215 participants at least 40 years of old are analyzed by using both Cox PH regression and the TR regression estimator developed here to study the relationship of all-cause mortality with the *ALAD* genotype and blood lead level. The design-based PH method and TR method are compared in this real complex survey setting.

Using Cox PH regression, Van Bommel et al²⁶ estimated the relative risk of all-cause, cardiovascular disease, and cancer mortality by *ALAD* genotype and by blood lead levels. For illustration purposes, we assess the association of blood lead levels and the *ALAD* c177G polymorphism with the risk of mortality from all causes, including major cardiovascular disease and cancer, and compare results from the PH and TR analyses. In the TR analysis, we fit a model to the data based on the FHT of a Wiener diffusion process. Note that our dataset has a longer follow-up time than that used in Van Bommel et al²⁶ and uses the NHANES III linked mortality file updated to December 31, 2011.

Focusing on blood lead levels and adjusting for other covariates, we tested the PH assumption of the dataset and found that the PH assumption does not hold ($P\text{-value} < .1$). Therefore, the TR model provides a useful analytical method for this real data application.

The censored survival time is follow-up time or time to death after the NHANES interview (in years). Persons who were not matched to a death record were considered alive through the follow-up period and administratively censored at December 31, 2011. The TR model has 2 parameters: initial health level y_0 and process mean μ . The link functions are the

natural logarithm for y_0 and identity function for μ . Following Van Bemmelen et al.,²⁶ we include the same set of covariates in the linear regression TR models, namely, age (defined as the participant's age at the baseline examination), blood lead level (=1 if lead exposure <5; =2 if ≥ 5 $\mu\text{g/dL}$), ALAD (the referent category versus a combined category of heterozygous and homozygous variants, ie, ALAD = 1 if GG and =2 if CG/CC), urban/rural (=1 if central counties of metro areas of 1 million population or more or fringe counties of metro areas of 1 million population or more; =2 if all other areas), sex (=1 if male; =2 if female), race (non-Hispanic white, non-Hispanic black, Mexican-American, and other), education (<12 years, completed high school, and > high school), and smoking (never, former, and current smokers). Selected characteristics of participants genotyped in the NHANES III cohort, cross-classified by blood lead categories, are summarized in Table 4. In Figure 1, we also plot Kaplan-Meier product limit estimates of survival curves with blood lead level < 5 $\mu\text{g/dL}$ and ≥ 5 $\mu\text{g/dL}$ for all-cause mortality by age group at baseline. As shown in the panels of Figure 1, there are distinct differences in the survival curves for the 2 blood-lead categories for each age group. The KM plots also show a violation of the PH assumption because the pairs of survival curves are not fixed powers of one another. In the case of the curves in Figure 1A for age group [40,50) years of old, for example, the survival curves actually cross, which is not consistent with PH.

Threshold regression analysis was conducted by using the proposed method. Cox PH regression analysis with the same set of covariates was conducted by using SUDAAN proc survival procedure, incorporating complex designs of NHANES III, for comparison purposes.

Table 5 presents results from the Cox and TR regression analyses, respectively. We judge effects with P -values under .05 as significant. From the Cox PH regression analysis, it can be observed that subjects who are older at interview, male, the less educated, or current/former smokers have higher risks of all-cause mortality. In TR regression, covariates have potential effects through 2 regression functions, namely, effects associated with a patient's state of initial health at the interview and effects associated with the rate of change of health status after the interview. From the TR regression analysis, we observe that women, the highly educated, and those with ALAD CC or GC have better initial health status. For the drift rate μ , which describes how the process progresses, we observe that subjects who are female, older at interview, current smokers, or have ALAD GG tend to drift more rapidly toward the threshold for all-cause mortality. In female group, a comparison of effects in the 2 regressions shows that these subjects tend to have better initial health but also drift faster toward the threshold.

We look more closely at blood lead level effects. For the purpose of illustration and comparison with the Cox analyses, we are considering the same sets of covariates for modeling initial health status ($\ln y_0$) and drift rate (μ) in the TR analysis. We note that a higher level of blood lead is associated with better initial health but with a faster rate.

The mixed direction of effects is borne out by the paired survival curves seen in Figure 1. The slightly higher initial health level is quickly offset by declining health. Every panel in

Figure 1 shows that the high-lead survival curve quickly drops below the low-lead survival curve after a few years of follow-up.

We note that the constant term for the regression function of μ in Table 5 is positive (1.372), and for some combinations of covariates (that is, for some types of subjects), their estimates for the rate parameter μ are positive, indicating a drift away from the threshold. This situation suggests that these types of subjects have a cure rate. This result is consistent with the KM plots shown in Figures 1A–1C where survival curves for these age groups are not tending to 0. Since we are modeling all-cause mortality in this dataset, our models and our selection of covariates for the 20-year follow-up period are clearly not capturing some force of mortality that comes at later ages and, hence, explains the possibility of a positive μ .

While this case application has the purpose of illustrating the extension of TR methods to complex survey sampling, it shows that the FHT-based TR regression approach can provide additional insights and results not available from the Cox model. Development of model checking and diagnostics for TR models is an important need in future research.

5 | DISCUSSIONS

A noninformative censoring is assumed for the proposed methodology, which is motivated by our real data example—the NHANES linked mortality data, where the event is defined as all-cause mortality, all sample units have either an event or are still alive by the end of the study. There are no units who are lost to follow-up during the follow-up period. Thus, the censoring process is independent of the event process and noninformative. If the event is defined as cause-specific mortality, such as mortality due to cardiovascular disease, there would be 2 types of censoring: the deaths during the follow-up period due to causes other than cardiovascular disease and units who are alive at the end of study. Both types of censoring would be noninformative.

In the simulations, we accordingly generate noninformative censoring variable. As an independent process from censoring, the sampling is conducted related to the outcome t . As results, subjects with observed events (or smaller t) tend to be oversampled. This scenario can be practical in real studies. For example, units in geographic areas with high incidence of certain diseases tend to be overselected at the baseline of studies. Unweighted analysis, by assuming SRS, would produce biased estimates due to the confounding between the noninformative censoring and the informative sampling. Therefore, we need to account for the sample designs (differential weighting and multistage clustering) into the analysis. After incorporating design features in the proposed methods, the estimated survival process would be design consistent with noninformative censoring, as shown in Tables 1–3.

As discussed above, the missingness of units (subjects who are not selected in the sample) is related to the outcome by design. All design variables are known and may be related to the outcome variable t . Accordingly, the sample weights, a function of design variables, are also known and obtained by the inverse of inclusion probabilities. Weighted estimates (eg Horvitz-Thompson estimators) would be design consistent for finite population parameters, such as population mean. For regression coefficient estimation (eg logistic regression

coefficient estimators under case-control studies, where the outcome variable of case-control status is one of design variables), weighted estimators that account for multistage clustering design are design consistent even for the intercept.²⁷ By contrast, nonignorable missingness, different from the missingness of units by design (ie, sampling with the known design variables related to the outcome), depends on the unknown variables that are related to the outcome. Both complete record analysis and the methods that are valid under missing at random mechanism would produce biased estimates. This paper aims for developing design-based TR estimators accounting for various complex sample designs.

The proposed Taylor variance estimators are developed based on asymptotic theorem. As shown in Table 2, Taylor variance estimators perform well when df are sufficiently large but tend to underestimate the empirical variance when df is small. In real data example, however, the number of df ($=22$) is limited. Following the simulation setup described in Sections 3.1 and 3.3, we further conduct simulation studies to evaluate performance of the Taylor variance estimators with a limited number of clusters $m = 25$ ($df = m-1 = 24$) and the sample size $n = 2000$, varying intracluster correlations ρ_t ($=0.05, 0.1$, and 0.2). It can be observed from Table 6 that both `taylor.var` and `mle.var` underestimate empirical variances, especially when ρ_t is 0.2 with their ratios to empirical variances ranging 0.84 to 0.90 for `taylor.var/emp.var` and 0.39 to 0.76 for `mle.var/emp.var`. Although with the variance ratios closer to the value of one, the `taylor.var` confirms the concern regarding the limited df in real data analyses, which, however, consistently performs superior to the `mle.var` with varying intracluster correlations. Caveats should be especially put to the interpretation of results from the real data analyses with limited df for variance estimation.

A resampling method, such as jackknife or balanced repeated replication, will be particularly attractive in the case of stratified multistage sampling because poststratification and unit nonresponse adjustment can be automatically taken into account through the use of appropriate replicate weights.²⁸ For example, a jackknife estimator of variance under stratified multistage sampling with I_h PSU's in the h^{th} stratum is given by

$$\hat{v}_J(\hat{S}_r) = \sum_{h=1}^H \frac{I_h - 1}{I_h} \sum_{i=1}^{I_h} \left(\hat{S}_r^{(hi)} - \hat{S}_r \right) \left(\hat{S}_r^{(hi)} - \hat{S}_r \right)^T,$$

where $\hat{S}_r^{(hi)}$ is obtained in the same manner as $\hat{S}_r$ when the data from the cluster- (hi) are deleted, but using jackknife weights.³³ The linearization method, however, has the advantage of being less computation intensive than these replications methods. Both the resampling and the linearization method provide consistent estimates of variance based on large sample theory; caution should be used when applying these methods to small samples.

6 | CLOSING REMARKS

The TR method is a relatively new approach for analyzing event time data but is gradually becoming popular in a wide assortment of medical and health studies.^{29–32} In this paper, we developed innovative TR estimators and their Taylor linearization variance estimators, accounting for differential weighting and multilevel clustering effects, for the analysis of

time-to-event data collected with complex sampling designs involving stratification and/or clustering. The developed method is numerically evaluated by simulation studies under various sample designs, and empirically illustrated using the NHANES III genetic data Linked Mortality File. In this example, we also show that TR can easily handle a cure rate if one exists in the dataset. The developed TR estimator assumes that FHT follows a Wiener diffusion process. Important extensions are needed for the development of diagnostics for FHT models, introducing models that allow time-varying covariates, and semi-parametric modeling of the covariate effects under the setting of complex surveys.

ACKNOWLEDGEMENT

This research was supported by the SEED grant as part of the campus ADVANCE Program at the University of Maryland for Inclusive Excellence (NSF award HRD1008117). The first author is also thankful for support from the NCHS. The work was done while the first author was an ASA/NCHS research fellow at the Division of Health and Nutrition Examination Surveys, NCHS, Centers for Disease Control and Prevention. The last author was also supported in part by R01EY02445.

Funding information

National Science Foundation, Grant/Award Number: HRD1008117; National Institute of Health, Grant/Award Number: R01EY02445

REFERENCE

1. Lee M-LT, DeGruttola V, Schoenfeld D. A model for markers and latent health status. *Journal of the Royal Statistical Society Series B*. 2000;62:747–762.
2. Lee M-LT, Whitmore GA. Threshold regression for survival analysis: modeling event times by a stochastic process reaching a boundary. *Statistical Science*. 2006;21:501–513.
3. Lee M-LT, Whitmore GA. Proportional hazards and threshold regression: their theoretical and practical connections. *Lifetime Data Anal*. 2010;16:196–214. [PubMed: 19960249]
4. Xiao T, Whitmore GA, He X, Lee M-LT. Threshold regression for time-to-event analysis: the stthreg package. *The STATA Journal*. 2012;12:257–283.
5. Folsom R, LaVange L, Williams RL. A probability sampling perspective on panel data analysis 1989 In: Kasprzyk D, Duncan G, Kalton G, Singh MP, eds. *Panel Surveys*. New York: Wiley:108–138.
6. Rubin-Bleuer S The proportional hazards model for survey data from independent and clustered super-population. *Journal of Multivariate Analysis*. 2011;102:884–895.
7. Pan Q, Schaubel DE. Proportional hazards models based on biased samples and estimated selection probabilities. *The Canadian Journal of Statistics*. 2008;36:111–127.
8. Korn EL, Graubard BI. *Analysis of Health Surveys*. New York: John Wiley & Sons; 1999.
9. Horvitz DG, Thompson DJ. A generalization of sampling without replacement from a finite universe. *Journal of the Royal Statistical Society Series B*. 1952;47:663–685.
10. Binder D Fitting Cox's proportional hazards models from survey data. *Biometrika*. 1992;79:139–147.
11. Lin DY. On fitting Cox's proportional hazards models to survey data. *Biometrika*. 2000;87:37–47.
12. Boudreau C, Lawless JF. Survey analysis based on the proportional hazards model and survey data. *The Canadian Journal of Statistics*. 2006;34:203–216.
13. Skinner CJ, Holt D, Smith TMF (Eds). *Analysis of Complex Surveys*. Chichester: Wiley; 1989.
14. Li Y Generalized regression estimators of a finite population total using the Box-Cox technique. *Survey Methodology*. 2008;34:79–89.
15. Shah B Comment on 'Linearization variance estimators for survey data' by A. Demnati and JNK Rao. *Survey Methodology*. 2004;30:29
16. Cochran WG. *Sampling Techniques*. Third ed. New York: Wiley; 1977.

17. Särndal CE, Swensson B, Wretman J. Model Assisted Survey Sampling. New York: Springer; 1992:153–154.
18. Wolter K Introduction to Variance Estimation. New York: Springer; 2003.
19. Cox DR, Millar HD. The Theory of Stochastic Processes. London: Chapman and Hall; 1965.
20. Pfeiffermann D The use of sampling weights for survey data analysis. *Stat Methods Med Res.* 1996;5:239–261. [PubMed: 8931195]
21. Rubin DB. Inference and missing data. *Biometrika.* 1976;63:581–592.
22. Little RJ. Models for non-response in sample surveys. *J Am Stat Assoc.* 1982;77:237–249.
23. NHANES, published 2012 http://www.cdc.gov/nchs/data/series/sr_02/sr02_113.pdf. Accessed April 06, 2016.
24. NHANES III Linked Mortality File, published 2012 http://www.cdc.gov/nchs/data_access/data_linkage/mortality/data_files_data_dicts-dictionaries.htm. Accessed April 06, 2016.
25. National Center for Health Statistics, CDC, NHANES III genetic data, published 2009 <http://www.cdc.gov/nchs/nhanes/genetics/genetic.htm>. Accessed February 11, 2015.
26. Van Bommel DM, Li Y, McLean J, et al. Blood lead levels, ALAD gene polymorphisms, and mortality. *Epidemiology.* 2011;22:273–278. [PubMed: 21293208]
27. Rust KF, Rao JNK. Variance estimation for complex surveys using replication methods. *Stat Methods Med Res.* 1996;5:283–310. [PubMed: 8931197]
28. Stogiannis D, Caroni C, Anagnostopoulos CE, Toupoulis IK. Comparing first hitting time and proportional hazards regression models. *Journal of Applied Statistics.* 2014;38:1483–1492.
29. Lee M-LT, Whitmore GA, Laden F, Hart JE, Garshick E. A case-control study relating railroad worker mortality to diesel exhaust exposure using a threshold regression model. *J Statist Plann Inference.* 2009;139:1633–1642.
30. Whitmore GA, Su Y. Modeling low birth weights using threshold regression: results for U.S. birth data. *Lifetime Data Anal.* 2007;13:161–190. [PubMed: 17286213]
31. Tong X, He X, Sun J, Lee M-LT. Joint analysis of current status and marker data: an extension of bivariate threshold model. *Int J Biostatistics.* 2008;4: Article 21
32. Aaron SD, Ramsay T, Vandemheen K, Whitmore GA. A threshold regression model for recurrent exacerbations in chronic obstructive pulmonary disease. *J Clinical Epidemiology.* 2010;63(12): 1324–1331.
33. Rust KF, Rao JNK. Variance estimation for complex surveys using replication methods. *Statistical Methods in Medical Research.* 1996;5:283–310. [PubMed: 8931197]

FIGURE 1.

Plots of Kaplan-Meier product limit estimates of survival with blood lead level $< 5 \mu\text{g/dL}$ ($\text{pbpc2} = 1$) and $> 5 \mu\text{g/dL}$ ($\text{pbpc2} = 2$) for all-cause mortality by age group at baseline. A, Age at baseline [40, 50) years of old. B, Age at baseline [50, 60) years of old. C, Age at baseline [60, 70) years of old. D, Age at baseline 70+ years of old [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 1

Simulations under 3 1-stage sample designs^a with different selection probabilities

	Initial health status $\ln(y_0)$		Drift Rate μ	
	Intercept	X	Intercept	X
<u>Design 1: Probability Proportional to Size of $(x + e)$</u>				
<u>MLE</u>				
Relative bias ($\times 100$)	-7.432	0.754	-7.287	0.859
emp.var ($\times 100$)	0.711	0.025	3.060	0.118
taylor.var/emp.var	1.006	1.014	1.012	1.015
<u>TR</u>				
Relative bias ($\times 100$)	-6.233	0.642	-7.047	0.827
emp.var ($\times 100$)	0.765	0.027	3.399	0.131
taylor.var/emp.var	1.020	1.026	0.997	1.002
<u>Design 2: Probability Proportional to Size of $(t + e)$</u>				
<u>MLE</u>				
Relative bias ($\times 100$)	129.973	-7.166	-79.712	4.826
emp.var ($\times 100$)	0.708	0.027	2.654	0.111
taylor.var/emp.var	1.003	0.994	0.969	0.957
<u>TR</u>				
Relative bias ($\times 100$)	8.307	-0.641	4.299	-0.349
emp.var ($\times 100$)	1.031	0.037	3.100	0.128
taylor.var/emp.var	0.953	0.959	0.927	0.925
<u>Design 3: Probability Proportional to Size of $(x + 2t + e)$</u>				
<u>MLE</u>				
Relative bias ($\times 100$)	63.268	-3.905	-74.077	5.633
emp.var ($\times 100$)	0.689	0.025	2.798	0.113
taylor.var/emp.var	1.059	1.036	1.012	1.001
<u>TR</u>				
Relative bias ($\times 100$)	4.994	-0.406	0.520	0.171
emp.var ($\times 100$)	0.763	0.028	2.810	0.113
taylor.var/emp.var	1.066	1.043	1.022	1.010

^a A sample of 1000 observations is randomly selected by probability proportional to the size (PPS) with size defined by (1) $x + |e|$ with $e \sim N(0, 1)$ for Design 1; (2) $t + |e|$ for Design 2; and (3) $x + 2t + |e|$ for Design 3.

TABLE 2

Simulations under 2-stage designs^a with different intraclass correlations

	Initial Health Status $\ln(y_0)$		Drift Rate μ	
	Intercept	X	Intercept	X
<u>The Number of Clusters $m = 400$</u>				
Relative bias ($\times 100$)	1.055	0.158	-1.309	0.701
emp.var ($\times 100$)	0.832	0.031	4.180	0.169
mle.var/emp.var	0.932	0.962	0.919	0.952
taylor.var/emp.var	0.984	0.968	1.011	0.985
<u>The Number of Clusters $m = 100$</u>				
Relative bias ($\times 100$)	7.689	-0.248	6.665	0.118
emp.var ($\times 100$)	1.042	0.034	5.873	0.203
mle.var/emp.var	0.746	0.874	0.660	0.796
taylor.var/emp.var	0.976	0.962	0.976	0.950
<u>The Number of Clusters $m = 50$</u>				
Relative bias ($\times 100$)	7.970	-0.245	7.681	0.180
emp.var ($\times 100$)	1.320	0.038	8.032	0.246
mle.var/emp.var	0.591	0.778	0.485	0.660
taylor.var/emp.var	0.940	0.924	0.928	0.897
<u>The Number of Clusters $m = 25$</u>				
Relative bias ($\times 100$)	15.340	-0.672	12.655	-0.266
emp.var ($\times 100$)	1.810	0.047	12.224	0.323
mle.var/emp.var	0.434	0.645	0.321	0.507
taylor.var/emp.var	0.928	0.874	0.870	0.838

^a A sample of 2000 observations is randomly selected by 2-stage sample designs with m out of 4000 clusters selected at the first stage and $2000/m$ out of 100 units selected within each selected cluster at the second stage, using simple random sampling at both stages.

TABLE 3

Simulations under a 2-stage complex design^a with varying censoring rates

	Initial Health Status $\ln(y_0)$		Drift Rate μ	
	Intercept	X	Intercept	X
<u>Censoring Rate = 8%</u>				
<u>MLE</u>				
Relative bias ($\times 100$)	36.189	-1.335	-79.368	7.287
emp.var ($\times 100$)	1.978	0.072	7.718	0.312
taylor.var/emp.var	1.095	1.110	1.054	1.057
<u>TR</u>				
Relative bias ($\times 100$)	5.238	-0.242	9.370	-0.525
emp.var ($\times 100$)	1.927	0.072	8.461	0.340
taylor.var/emp.var	1.085	1.098	1.037	1.043
<u>Censoring Rate = 15%</u>				
<u>MLE</u>				
Relative bias ($\times 100$)	21.184	-0.321	-91.874	8.762
emp.var ($\times 100$)	2.264	0.085	9.914	0.404
taylor.var/emp.var	1.125	1.120	1.145	1.135
<u>TR</u>				
Relative bias ($\times 100$)	-1.664	0.269	-4.060	1.003
emp.var ($\times 100$)	2.157	0.083	10.220	0.419
taylor.var/emp.var	1.119	1.111	1.139	1.126
<u>Censoring Rate = 25%</u>				
<u>MLE</u>				
Relative bias ($\times 100$)	17.857	-0.084	-80.699	8.038
emp.var ($\times 100$)	2.879	0.108	15.896	0.638
taylor.var/emp.var	1.072	1.092	1.082	1.098
<u>TR</u>				
Relative bias ($\times 100$)	1.776	0.051	7.142	-0.147
emp.var ($\times 100$)	2.642	0.101	15.393	0.622
taylor.var/emp.var	1.076	1.095	1.076	1.090

^a A sample of 2000 observations is randomly selected by 2-stage sample designs: 400 out of 4000 clusters are selected by selection probabilities proportional to the size (PPS), where the size is defined by x and t , at the first stage. Within each of the 400 clusters sampled, 5 observations are selected by simple random sampling.

TABLE 4

Selected characteristics of the NHANES III study cohort by blood lead level

Characteristic	Blood Lead ($\mu\text{g/dL}$)	
	<5	5
Sample size	2415	800
Mortality status; no. (%) [*]		
No	1384(67)	336(50)
Yes	1031(33)	464(50)
Smoking status at baseline; no. (%) [*]		
Never	1304(50)	255(25)
Former	724(32)	294(40)
Current	387(17)	251(35)
Race/ethnicity; no. (%) [*]		
Non-Hispanic white	1293(84)	317(75)
Non-Hispanic black	444(6)	270(15)
Mexican-American OR Other	678(10)	213(10)
Education; no (%) [*]		
<12 years	974(22)	427(38)
Completed high school	725(33)	227(37)
>High school	703(45)	138(25)
Sex; no (%) [*]		
MALE	871(42)	542(67)
FEMALE	1544(58)	258(33)
ALAD; no. (%) [*]		
GG	2072(85)	703(86)
GC AND CC	272(15)	70(14)
Urban/rural; no. (%) [*]		
Urban	1034 (45)	409 (47)
Rural	1381 (55)	391 (53)
Year of birth; no. (%) [*]		
[1888,1918)	407(11)	159(15)
[1918,1930)	657(21)	258(30)
[1930,1942)	619(29)	213(30)
[1942,1954]	732(39)	170(25)
Age at baseline (years); mean	57	61
Follow-up time (years); mean	15.8	14

^{*} The percentages are based on weighted data.

TABLE 5

A comparison of Cox and threshold regression analysis using the NHANES III Linked Mortality File

	Cox Regression		TR Regression			
			$\ln(y_0)$		μ	
	Coeff.	SE (P-value)	Coeff.	SE (P-value)	Coeff.	SE (P-value)
Age at interview	.104	.005 (<.001)	-.007	.004 (.08)	-.018	.002 (<.001)
Smoking status						
Non smoker	0	0	0	0	0	0
Former smoker	.265	.121 (.039)	-.243	.141 (.085)	.030	.048 (.529)
Current smoker	.850	.163 (<.001)	.019	.154 (.904)	-.165	.067 (.014)
Race and Hispanic						
Non-Hispanic white	0	0	0	0	0	0
Non-Hispanic black	.205	.103 (.059)	-.131	.185 (.48)	-.022	.066 (.742)
Mexican American or other	-.123	.116 (.301)	-.060	.262 (.818)	.014	.091 (.88)
Education						
<12 years	0	0	0	0	0	0
12 years	-.057	.053 (.294)	-.209	.139 (.132)	.007	.043 (.057)
>12 years	-.327	.109 (.007)	.234	.098 (.017)	-.098	.052 (.899)
Sex						
Male (=1)	0	0	0	0	0	0
Female (=2)	-.293	.076 (.001)	.527	.126 (<.001)	-.076	.05 (.048)
Blood lead level ($\mu\text{g/dL}$)						
[0,5) (=1)	0	0	0	0	0	0
[5, ∞) (=2)	.090	.105 (.401)	.228	.084 (.007)	-.087	.035 (.031)
<i>ALAD rs1800435</i>						
GG (=1)	0	0	0	0	0	0
CC or GC (=2)	-.070	.075 (.355)	.306	.117 (.009)	.064	.042 (.04)
Urban/rural						
Urban (=1)	0	0	0	0	0	0
Rural (=2)	.038	.07 (.594)	-.211	.118 (.074)	-.165	.039 (.103)
Constant			.532	.521 (.307)	1.372	0.217 (<.001)

TABLE 6

Comparison between the MLE estimator and the TR estimator with pseudo likelihood under 2-stage designs^a with the number of clusters selected at the first stage $m = 25$

	Initial Health Status $\ln(y_0)$		Drift Rate μ	
	Intercept	X	Intercept	X
<u>Intraclass Correlation $\rho_t \equiv 0.05$</u>				
Relative bias ($\times 100$)	3.737	-0.122	3.035	0.216
emp.var ($\times 100$)	0.848	0.032	5.237	0.188
mle.var/emp.var	0.920	0.938	0.741	0.859
taylor.var/emp.var	0.967	0.947	0.945	0.919
<u>Intraclass Correlation $\rho_t \equiv 0.1$</u>				
Relative bias ($\times 100$)	8.055	-0.307	6.137	0.127
emp.var ($\times 100$)	0.953	0.033	6.770	0.207
mle.var/emp.var	0.819	0.908	0.575	0.786
taylor.var/emp.var	0.962	0.955	0.921	0.927
<u>Intraclass Correlation $\rho_t \equiv 0.2$</u>				
Relative bias ($\times 100$)	14.474	-0.753	12.154	-0.309
emp.var ($\times 100$)	1.345	0.039	10.154	0.269
mle.var/emp.var	0.582	0.759	0.387	0.608
taylor.var/emp.var	0.900	0.881	0.841	0.838

^a A sample of 2,000 observations is randomly selected by two-stage sample designs with 25 out of 4000 clusters selected at the first stage and 80 out of 100 units selected within each selected cluster at the second stage, using simple random sampling at both stages.