9-19-2006

# COX MODELS WITH NONLINEAR EFFECT OF COVARIATES MEASURED WITH ERROR: A CASE STUDY OF CHRONIC KIDNEY DISEASE INCIDENCE

Ciprian M. Crainiceanu
*Johns Hokins Bloomberg School of Public Health, Department of Biostatistics*, ccrainic@jhsph.edu

David Ruppert
*School of Operational Research and Industrial Engineering, Cornell University*

Josef Coresh
*Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health*

# Cox models with nonlinear effect of covariates measured with error: A case study of chronic kidney disease incidence

Ciprian M. Crainiceanu* David Ruppert† Josef Coresh‡

September 19, 2006

**Abstract**

We propose, develop and implement the simulation extrapolation (SIMEX) methodology for Cox regression models when the log hazard function is linear in the model parameters but nonlinear in the variables measured with error (LPNE). The class of LPNE functions contains but is not limited to strata indicators, splines, quadratic and interaction terms. The first order bias correction method proposed here has the advantage that it remains computationally feasible even when the number of observations is very large and multiple models need to be explored. Theoretical and simulation results show that the SIMEX method outperforms the naive method even with small amounts of measurement error. Our methodology was motivated by and applied to the study of time to chronic kidney disease (CKD) progression as a function of baseline kidney function and applied to the Atherosclerosis Risk in Communities (ARIC), a large epidemiological cohort study.

**Keywords: survival analysis, nonlinear log hazard, measurement error**

*Assistant Professor, Department of Biostatistics, Johns Hopkins University, 615 N. Wolfe St. Baltimore, MD 21205 USA. E-mail: ccrainic@jhsph.edu

†Andrew Schultz Jr. Professor of Engineering and Professor of Statistical Science, School of Operational Research and Industrial Engineering, Cornell University, Rhodes Hall, NY 14853, USA. E-mail: dr24@cornell.edu.

‡Professor, Department of Epidemiology, Johns Hopkins University, 2024 E. Monument Street, Suite 2-600, Baltimore, MD 21205 USA. E-mail: coresh@jhu.edu

# 1 Introduction

Survival analysis has developed from the analysis of life tables in actuarial sciences and has enjoyed remarkable success with modern applications in medicine, epidemiology, and the social sciences. The popularity of survival analysis models, such as the Cox proportional hazards model, is probably surpassed only by the popularity of standard linear regression models.

Survival data are the product of a continuous death process coupled with a censoring mechanism. Typically, the death rate depends on a number of factors and time to death is only partially observed for those subjects with censored observations. Standard analyses of survival data assume that all covariates affecting survival rates are observed without error. However, in many applications some of the covariates are subject to measurement error or are available without error only for a subsample of the population.

Survival analysis when covariates are measured with error has witnessed an explosion in the last decade. However, the case when the log hazard rate is nonlinear in the covariates measured with error has been overlooked, probably because of its inherent inferential complexity, not because of its lack of relevance. Indeed, once measurement error has been acknowledged it is only natural to ask how it may impact the fitting of nonlinear functions, such as strata indicators, splines, quadratic or interaction terms. One common feature of such functions is that they are linear in the model parameters but nonlinear in the covariates measured with error (LPNE). We propose a new inferential method based on simulation extrapolation (SIMEX) for Cox regression models when some of the risk factors are observed with error and the log–hazard function is LPNE.

Substantial methodological and applied research has been dedicated in recent years to survival analysis with covariates subject to measurement error, starting with the seminal paper by Prentice (1982) [17]. The regression calibration approach was expanded and refined by Pepe, Self & Prentice (1989) [16] and Wang, Hsu, Feng & Prentice (1997) [24]. Clayton (1991) [5] used regression calibration within risk sets thus avoiding the rare disease assumption. For data containing a validation sample, Zhou & Pepe (1995) [25] and Zhou & Wang (2000) [26] proposed nonparametric estimators of the induced hazard function. For data with at least two replicates Huang & Wang (2000) [11] have proposed a consistent nonparametric estimator based on a modification of the partial likelihood score equation. Augustin (2004) [2] showed that Nakamura's (1992) [15] methodology of adjusting the likelihood can be applied to the Breslow likelihood to provide an exact corrected likelihood. This result circumvented the impossibility result derived by Stefanski (1989) [19] for the partial likelihood. Hu, Tsiatis & Davidian (1998) [10] have proposed likelihood maximization algorithms for parametric and nonparametric specifications of the distribution of the unobserved variables. Greene & Cai (2004) [8] established the asymptotic properties of the

SIMEX estimators for models with measurement error and multivariate failure time data. Hu & Lin (2004) [9] introduced a modified score equation and established the asymptotic properties of the estimators for multivariate failure time data. Li & Lin (2003) [13] used the EM algorithm and SIMEX respectively to provide maximum likelihood estimators for frailty models with variables observed with error. Song & Huang (2005) [18] compare the conditional score estimation of Tsiatis & Davidian (2001) [23] with Nakamura's (1992) [15] parametric adjustment.

Current statistical approaches for Cox regression with risk factors measured with sizeable error have one or more of the following theoretical and practical limitations: 1) Biased and misspecified variability of risk factor effect estimators leading to invalid tests for exposure effect; 2) Sensitivity of results to assumptions such as rare disease and linearity; 3) Need for intensive computing that drastically limits the size of data sets; 4) Focus on linear log-hazard function. SIMEX has emerged as a natural, computationally usable methodology that accounts for nonlinearity, measurement error structure and circumvents the rare disease assumption.

Our proposed methodology was motivated by the analysis of time to event data from the Atherosclerosis Risk In Communities (ARIC) study. ARIC is a large multipurpose epidemiological study conducted in four US communities (Forsyth County, NC; suburban Minneapolis, MN; Washington County, MD; and Jackson, MS). A detailed description of the ARIC study is provided by the ARIC investigators (1989) [12]. In short, from 1987 through 1989, $15,792$ male and female volunteers aged 45 through 64 were recruited from these communities for a baseline and three subsequent visits.

Time to event data is observed continuously for multiple end points, but for reasons of brevity we focus here on the event incidence of CKD, the least severe phase of kidney disease. For the purpose of this study all primary CKD events up to January 1, 2003 were included and the time to event data was obtained from annual participant interviews and review of local hospital discharge lists and county death certificates.

The relationship between various risk factors, such as race, age or sex, and progression time to incidence of CKD may be confounded by a series of baseline confounders. An important confounder is the baseline kidney function as measured by the glomerular filtration rate (GFR). Because GFR can only be obtained through a long and awkward procedure, the estimated GFR (eGFR) is used in practice. eGFR is obtained from a prediction equation based on creatinine, sex, gender and age and is subject to regression and biological measurement error. Incidence CKD is defined as either achievement of follow–up eGFR $< 60$ ml/min/1.73m$^2$ or a post–baseline hospitalization or death with CKD (Marsh–Manzi et al., 2005 [14]). As is the case in many biological applications, subjects with lower baseline GFR (higher risk) are expected to progress faster towards primary CKD (clinical endpoint), but the effect of GFR on the risk of CKD is likely to be nonlinear. Given the measurement

error in eGFR, estimating the dose–response curve is a difficult inferential problem, while its result is of interest to nephrologists, internists and researchers (Stevens et al. 2006, [21]).

## 2 Notations and assumptions

Assume that $n$ subjects are observed over time and their failure times $T_1, \ldots, T_n$ are subject to right censoring and $C_1, \ldots, C_n$ are the corresponding censoring times. Let $\delta_i = I(T_i < C_i)$ be the failure indicator and $Y_i = \min(T_i, C_i)$ be the time to failure or censoring for subject $i$. Denote by $R_i = \{j : Y_j \geq Y_i\}$ the risk set when the event corresponding to subject $i$ occurs. $R_i$ is the index set for those subjects who have not failed and are uncensored at the time the $i$th subject fails or is censored. The at-risk indicator process for the $i$th subject is defined as $Y_i(t) = I(Y_i \geq t)$.

We assume that the survival probability for each subject depends on covariates that are subject to measurement error, $\boldsymbol{X}_i$, as well as on covariates that are not, $\boldsymbol{Z}_i$. The covariate $\boldsymbol{X}_i$ is measured through the usual classical measurement error model

$$\boldsymbol{W}_i = \boldsymbol{X}_i + \boldsymbol{U}_i \tag{1}$$

where $\boldsymbol{U}_i$ is the additive measurement error. For simplicity of presentation we will assume that $\boldsymbol{U}_i \sim N(0, \boldsymbol{\Omega})$, which is a reasonable assumption in most applications after appropriate data transformations. The covariance matrix $\boldsymbol{\Omega}$ is assumed to be known or estimable. Note, however, that the methodology proposed in this paper can be easily extended to other measurement error distributions as well as non-additive measurement error.

We also assume that $(T_i, \boldsymbol{X}_i^t, C_i, \boldsymbol{U}_i^t)^t$ are independent random vectors, $C_i$ is independent of $(T_i, \boldsymbol{X}_i^t)^t$ and $\boldsymbol{U}_i$ is independent of $(T_i, \boldsymbol{X}_i^t, C_i)^t$. Note that our methodology will allow $\boldsymbol{X}_i$ to depend on other covariates, $\boldsymbol{Z}_i$, even though our application to CKD progression does not contain this additional complexity. The observed data are the vectors $(Y_i, \delta_i, \boldsymbol{W}_i^t, \boldsymbol{Z}_i^t)^t$ where $(Y_i, \delta_i)^t$ is a proxy observation for $(T_i, C_i)^t$ and $\boldsymbol{W}_i$ is a proxy observation for $\boldsymbol{X}_i$.

The distribution of the failure time of subject $i$, $T_i$, is completely described by the hazard rate, $\lambda_i(t|\boldsymbol{X}_i, \boldsymbol{Z}_i)$. The proportional hazards model introduced by Cox (1972) [7] is the most commonly used model for the hazard rate and assumes that

$$\lambda_i(t|\boldsymbol{X}_i, \boldsymbol{Z}_i) = \lambda_0(t) \exp(\boldsymbol{\alpha}_x^t \boldsymbol{X}_i + \boldsymbol{\alpha}_z^t \boldsymbol{Z}_i), \tag{2}$$

where $\lambda_0(\cdot)$ is an unspecified baseline hazard function that does not depend on the covariate values. With these notations, the log hazard rate, $\boldsymbol{\beta}_x^t \boldsymbol{X}_i + \boldsymbol{\beta}_z^t \boldsymbol{Z}_i$, is linear both in the parameters and in the variables observed with error, $X_i$. In order to accommodate such

simple functions as strata indicators, splines, quadratic and interaction terms we consider a hazard ratio of the type

$$\lambda_i(t|\boldsymbol{X}_i, \boldsymbol{Z}_i) = \lambda_0(t) \exp\{\boldsymbol{\beta}^t h(\boldsymbol{X}_i, \boldsymbol{Z}_i)\}. \tag{3}$$

In the standard regression case when $\boldsymbol{X}_i$ are observed model (3) would be indistinguishable from model (2) by simply relabelling the components of the $h(\cdot, \cdot)$ function. However, when $\boldsymbol{X}_i$ is observed with error and $h(\cdot, \cdot)$ is non-linear at least in one of the components of $\boldsymbol{X_i}$, $\boldsymbol{\beta}^t h(\boldsymbol{X}_i, \boldsymbol{Z}_i)$ is linear in the model parameters (LP) but nonlinear in the measurement error (NE). We label this class of functions LPNE for the remainder of the paper.

In the standard regression case, Cox (1972) [7] suggested that inference on $\boldsymbol{\beta}$ be based on the log partial likelihood function

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \delta_i \left( \boldsymbol{\beta}^t h(\boldsymbol{X}_i, \boldsymbol{Z}_i) - \log \left[ \sum_{j \in R_i} \exp\{\boldsymbol{\beta}^t h(\boldsymbol{X}_i, \boldsymbol{Z}_i)\} \right] \right) \tag{4}$$

which does not depend on $\lambda_0(\cdot)$. An alternative strategy is to use the log of the full likelihood of the model (2)

$$L(\boldsymbol{\beta}) = \sum_{i=1}^{n} \delta_i \left[ \boldsymbol{\beta}^t h(\boldsymbol{X}_i, \boldsymbol{Z}_i) + \log\{\lambda_0(Y_i)\} \right] - e^{\boldsymbol{\beta}^t h(\boldsymbol{X}_i, \boldsymbol{Z}_i)} \int_0^{Y_i} \lambda_0(s) ds. \tag{5}$$

In measurement error models the variables $\boldsymbol{X}_i$ are not available and standard Cox regression cannot be used. It is well documented, e.g. in Carroll et al., 2006 [4], that naively replacing $\boldsymbol{X}_i$ by its missmeasured version $\boldsymbol{W}_i$ typically leads to biased estimates and misspecified variability of exposure effects. In principle, regression calibration techniques could be used to provide a first order bias corrected estimator of $\boldsymbol{\beta}$ by simply replacing $h(\boldsymbol{X}_i, \boldsymbol{Z}_i)$ by its conditional expectation $E\{h(\boldsymbol{X}_i, \boldsymbol{Z}_i)|\boldsymbol{W}, \boldsymbol{Z}\}$, where $\boldsymbol{W} = \{\boldsymbol{W}_1, \ldots, \boldsymbol{W}_n\}$ and $\boldsymbol{Z} = \{\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n\}$. However, when $h(\cdot, \cdot)$ is not linear the conditional expectation is hard or impossible to obtain explicitly requiring additional approximations. Moreover, as described by Prentice (1982) [17], regression calibration requires the rare disease assumption, which is violated in many practical applications, including our study of progression to primary CKD. In contrast, the SIMEX method avoids these problems and remains computationally feasible.

# 3 SIMEX for LPNE Functions

SIMEX was proposed by Cook and Stefanski, 1995 [6] and further developed by Carroll, Küechenhoff, Lombard and Stefanski (1996) [3] and Stefanski and Cook (1995) [20]. For the case of multivariate failure time data Greene and Cai (2004) [8] have established the consistency and asymptotic normality of the SIMEX estimator for a linear log hazard function. Li and Lin (2003) [13] have used SIMEX coupled with the EM algorithm to provide inference for clustered survival data when some of the covariates are subject to measurement error.

The SIMEX idea is to simulate new data by adding increasing amounts of noise to the measured values $\boldsymbol{W}_i$ of the error prone covariate $\boldsymbol{X}_i$, compute the estimator on each simulated data set, model the expectation of the estimator as a function of the measurement error variance, and extrapolate back to the case of no measurement error. More precisely, if $\boldsymbol{\Omega}$ is a known positive definite $Q \times Q$ measurement error covariance matrix and $\boldsymbol{\Omega}^{1/2}$ is its positive square root then remeasured data is generated as

$$\boldsymbol{W}_{b,i}(\upsilon) = \boldsymbol{W}_i + \sqrt{\upsilon}\,\boldsymbol{\Omega}^{1/2}\,\boldsymbol{U}_{b,i}, \quad b = 1, \ldots, B$$

where $\boldsymbol{U}_{b,i}$ are independent Normal$(0, I_Q)$ vectors, $\upsilon$ is a positive scalar, and $B$ is the number of simulations for each value of $\upsilon$. The measurement error covariance of the contaminated observations $\boldsymbol{W}_{b,i}(\upsilon)$ is

$$\mathrm{Cov}\big\{\boldsymbol{W}_{b,i}(\upsilon)\big\} = (1+\upsilon)\,\Omega,$$

which converges to the zero matrix as $\upsilon \to -1$. By replacing $\boldsymbol{X}_i$ with $\boldsymbol{W}_{b,i}(\upsilon)$ in the hazard function (3) we obtain

$$\lambda_i\big\{t|\boldsymbol{W}_{b,i}(\upsilon), \boldsymbol{Z}_i\big\} = \lambda_0(t)\exp\big[\boldsymbol{\beta}^t h\{\boldsymbol{W}_{b,i}(\upsilon), \boldsymbol{Z}_i\}\big] \tag{6}$$

and either the partial likelihood (4) or the full likelihood (5) could be used to produce estimators $\widehat{\boldsymbol{\beta}}^b(\upsilon)$. Here $\boldsymbol{\beta}$ is a $P \times 1$ dimensional vector characterizing the proportional hazard function. The linearity assumption of the log hazard function in $\boldsymbol{\beta}$ plays an important because when both $\boldsymbol{W}_{b,i}$ and $\boldsymbol{Z}_i$ are known, model (6) is a standard Cox regression model. Thus, fitting model (6) can be done using existent software designed for proportional hazards models, such as R or S–plus (`coxph()` and `survreg()` functions) or SAS (`PHREG` procedure).

For each level of added noise $\upsilon$ and each scalar component $\beta_p$ of $\boldsymbol{\beta}$, $p = 1, \ldots, P$, one obtains

$$\widehat{\beta}_p(\upsilon) = \frac{1}{B}\sum_{b=1}^{B}\widehat{\beta}_p^b(\upsilon).$$

A quadratic or rational extrapolant can then be used to obtain the estimated values corre-

sponding to $\upsilon = -1$. More precisely, if $\widehat{m}_p(\upsilon)$ is an extrapolant fitted to $\widehat{\beta}_p(\upsilon)$ then

$$\widehat{\beta}_{p,\text{SIMEX}} = \lim_{\upsilon \to -1} \widehat{m}_p(\upsilon) \tag{7}$$

The variance estimation of the SIMEX estimator is obtained using a similar, but slightly more involved procedure. Indeed, there are two components of the variability at each level of added noise that need to be extrapolated separately: 1) the average of the variances of parameter estimators; 2) the variance of the parameter estimates around their average. More precisely, the average of the variances at noise level $(1 + \upsilon)$ is

$$\overline{\text{Var}}\{\widehat{\beta}_p(\upsilon)\} = \frac{1}{B}\sum_{b=1}^{B} \widehat{\text{Var}}\{\widehat{\beta}_p^b(\upsilon)\},$$

where $\widehat{\text{Var}}\{\widehat{\beta}_p^b(\upsilon)\}$ is the estimated variance of the $\beta_p$ parameter based on the $b$th simulation. The variance of the parameter estimates around their average is

$$\widehat{\text{Var}}\{\widehat{\beta}_p(\upsilon)\} = \frac{1}{B}\sum_{b=1}^{B} \{\widehat{\beta}_p^b(\upsilon) - \widehat{\beta}_p(\upsilon)\}^2.$$

Extrapolating $\overline{\text{Var}}\{\widehat{\beta}_p(\upsilon)\}$ can be done using, for example, a quadratic extrapolant. However, extrapolating $\widehat{\text{Var}}\{\widehat{\beta}_p(\upsilon)\}$ is more challenging because $\widehat{\text{Var}}\{\widehat{\beta}_p(\upsilon = 0)\} = 0$ and all extrapolants at $\upsilon = -1$ will be negative. To overcome this problem, Stefanski and Cook (1995) [20] suggested the following SIMEX estimator of the variance

$$\widehat{\sigma}^2_{p,\text{SIMEX}} = \lim_{\upsilon \to -1} \{\overline{\sigma}^2_p(\upsilon) - \widehat{\sigma}^2_p(\upsilon)\}, \tag{8}$$

where $\overline{\sigma}^2_p(\upsilon)$ and $\widehat{\sigma}^2_p(\upsilon)$ are the extrapolant functions corresponding to $\overline{\text{Var}}\{\widehat{\beta}_p(\upsilon)\}$ and $\widehat{\text{Var}}\{\widehat{\beta}_p(\upsilon)\}$, respectively.

We end this section by noting that SIMEX works well on any linear transformation of the model parameters using exactly the same extrapolation techniques. This observation will be especially useful when we discuss changes in the relationship between GFR and log hazard of CKD.

# 4 Theoretical results

To establish the theoretical properties of the SIMEX estimator a few notations need to be introduced. Suppose that when $\boldsymbol{X}$'s are known the unknown parameter $\boldsymbol{\beta}_0$ could be estimated solving the estimating equation

$$\sum_{i=1}^{n} \psi(Y_i, \delta_i, \boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\beta}) = 0 \tag{9}$$

where

$$\psi(Y_i, \delta_i, \boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\beta}) = \delta_i \left[ h(\boldsymbol{X}_i, \boldsymbol{Z}_i) - \frac{\exp\{\boldsymbol{\beta}^t h(\boldsymbol{X}_i, \boldsymbol{Z}_i)\}}{\sum_{j \in R_i} \exp\{\boldsymbol{\beta}^t h(\boldsymbol{X}_i, \boldsymbol{Z}_i)\}} \right].$$

Note that $\psi(\cdot)$ is a $P \times 1$ vector and equation (9) is a set of $P$ equations with $P$ unknowns. Following Carroll et al. (1996) [3] define $\widehat{\boldsymbol{\beta}}^b(\upsilon)$ as the solution to

$$\sum_{i=1}^{n} \psi(Y_i, \delta_i, \boldsymbol{W}_i + \upsilon^{1/2} \boldsymbol{\Omega}^{1/2} \boldsymbol{\epsilon}_{ib}, \boldsymbol{Z}_i; \boldsymbol{\beta}) = 0$$

where $\boldsymbol{\epsilon}_{ib}$ are i.i.d. random variables $N(0, \boldsymbol{I}_Q)$, where $Q = \dim(\Omega)$ is the number of covariates measured with error. Let $\widehat{\boldsymbol{\beta}}(\upsilon) = \sum_{b=1}^{B} \widehat{\boldsymbol{\beta}}^b(\upsilon)/B$. Under regularity conditions described originally by Tsiatis (1981) [22] , $\widehat{\boldsymbol{\beta}}(\upsilon)$ converges in probability to $\boldsymbol{\beta}_0(\upsilon)$, the solution of

$$E\{\psi(Y, \delta, \boldsymbol{W} + \upsilon^{1/2} \boldsymbol{\Omega}^{1/2} \boldsymbol{\epsilon}, \boldsymbol{Z}; \boldsymbol{\beta})\} = 0.$$

The SIMEX procedure assumes that a true extrapolant is available such that $\boldsymbol{\beta}(\upsilon) = \Gamma(\boldsymbol{\Theta}, \upsilon)$, where $\boldsymbol{\Theta}$ is a vector of parameters. Typical extrapolants are linear, quadratic or fractional, all with at most 3 parameters for each component of $\boldsymbol{\beta}(\upsilon)$. For simplicity of presentation we focus on the popular quadratic extrapolant, but results hold more generally. Thus, one fits the model

$$\widehat{\beta}_p(\upsilon_m) = \theta_{p1} + \theta_{p2}\upsilon_m + \theta_{p3}\upsilon_m^2 + \eta_{pm}, \quad p = 1, \ldots, P; m = 1, \ldots, M,$$

where $P$ is the number of Cox model parameters, $M$ is the number of grid points used in the simulation step, and $\eta_{pm} \sim \text{Normal}(0, \sigma_\eta^2)$ are mutually independent. Denoting by $\widehat{\boldsymbol{\beta}}(\upsilon_m) = \text{vec}\{\widehat{\beta}_p(\upsilon_m) : p = 1, \ldots, P\}$, $\widehat{\boldsymbol{\beta}}(\Upsilon) = \text{vec}\{\widehat{\boldsymbol{\beta}}(\upsilon_m) : m = 1, \ldots, M\}$, $\boldsymbol{\theta}_p = (\theta_{p1}, \theta_{p2}, \theta_{p3})^t$, $\boldsymbol{\theta} = \text{vec}\{\boldsymbol{\theta}_p : p = 1, \ldots, P\}$, $\boldsymbol{\Upsilon}_m = \text{diag}\{(1, \upsilon_m, \upsilon_m^2)\}$, and $\boldsymbol{\Upsilon}$ the $MP \times 3P$ dimensional

matrix obtained by stacking the $\boldsymbol{\Upsilon}_m$ matrices, $m = 1, \ldots, M$, then

$$\widehat{\boldsymbol{\beta}}(\boldsymbol{\Upsilon}) = \boldsymbol{\Upsilon}\boldsymbol{\theta} + \boldsymbol{\eta}$$

where $\boldsymbol{\eta} = \text{vec}\{\eta_{pm} : p = 1, \ldots, P; m = 1, \ldots, M\}$. The true parameters in the case of no measurement error are $\beta_p(-1) = \theta_{p1} - \theta_{p2} + \theta_{p3}$ or, in matrix format $\boldsymbol{\beta}(-1) = \boldsymbol{\Upsilon}(-1)\boldsymbol{\theta}$, where $\boldsymbol{\Upsilon}(-1) = \text{diag}\{(1 - 1\ 1)\}$ is a $P \times 3P$ dimensional matrix. It follows that $\widehat{\boldsymbol{\beta}}(-1) = \boldsymbol{\Upsilon}(-1)(\boldsymbol{\Upsilon}^t\boldsymbol{\Upsilon})^{-1}\boldsymbol{\Upsilon}^t\widehat{\boldsymbol{\beta}}(\boldsymbol{\Upsilon})$. We are now in the position to provide the main theoretical result of our paper.

**Theorem 1** *Assume that the covariance matrix of the measurement error, $\boldsymbol{\Omega}$, is known and an exact extrapolant is available. Under conditions similar to those in Andersen and Gill (1982) [1] the SIMEX estimator satisfies*

$$n^{1/2}(\widehat{\boldsymbol{\beta}}_{\text{SIMEX}} - \boldsymbol{\beta}_0) \Rightarrow \text{Normal}(0, \boldsymbol{\Sigma}).$$

*If the extrapolant is quadratic then $\boldsymbol{\Sigma} = \boldsymbol{\Upsilon}(-1)(\boldsymbol{\Upsilon}^t\boldsymbol{\Upsilon})^{-1}\boldsymbol{\Upsilon}^t\boldsymbol{\Xi}\boldsymbol{\Upsilon}(\boldsymbol{\Upsilon}^t\boldsymbol{\Upsilon})^{-1}\boldsymbol{\Upsilon}^t(-1)$, where $\boldsymbol{\Xi} = \boldsymbol{A}^{-1}\boldsymbol{C}\boldsymbol{A}^{-t}$ and $\boldsymbol{A}$ and $\boldsymbol{C}$ are given by*

*i. $\boldsymbol{A} = \text{diag}[\mathcal{A}\{\upsilon_m, \boldsymbol{\beta}(\upsilon_m)\} : m = 1, \ldots, M]$ which is an $MP \times MP$ matrix with diagonal elements given by the $P \times P$ matrices*

$$\mathcal{A}\{\upsilon_m, \boldsymbol{\beta}(\upsilon_m)\} = -\lim_{n\to\infty}\frac{1}{n}\frac{\partial}{\partial\beta^t}\sum_{i=1}^{n}\psi\{Y_i, \delta_i, \boldsymbol{W}_i + \upsilon^{1/2}\boldsymbol{\Omega}^{1/2}\boldsymbol{\epsilon}_{ib}, \boldsymbol{Z}_i; \boldsymbol{\beta}^*(\upsilon_m)\}$$

*for any random $\boldsymbol{\beta}^*(\upsilon_m)$ such that $\boldsymbol{\beta}^*(\upsilon_m) \to \boldsymbol{\beta}(\upsilon_m)$.*

*ii. $\boldsymbol{C} = \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\text{Var}\{\boldsymbol{D}_i(\upsilon_1)^t, \ldots, \boldsymbol{D}_i(\upsilon_M)^t\}^t$ where*

$$\boldsymbol{D}_i(\upsilon) = \frac{1}{B}\sum_{b=1}^{B}\psi\{Y_i, \delta_i, \boldsymbol{W}_i + \upsilon^{1/2}\boldsymbol{\Omega}^{1/2}\epsilon_{ib}, \boldsymbol{Z}_i; \boldsymbol{\beta}(\upsilon)\}.$$

To prove this result we follow essentially the steps described in the proof of Carroll et al. (1996) [3] in the general case of unbiased estimating equations. The major difference between our proof and the proof of Carroll et al. (1996) [3] is that the the Taylor expansion of the estimating equation around the true parameter cannot be obtained using *standard asymptotic theory results*. Instead, one needs to use martingale representation theory to show that the Taylor series expansion of the score equation is asymptotically equivalent to a sum of i.i.d. random vectors. The same strategy was used by Greene and Cai (2004) [8] who then went on to prove the asymptotic normality following step by step the proof of

Carroll et al. (1996) [3]. Because our model is linear in the parameters $\boldsymbol{\beta}$ the proof of the asymptotic normality results also follows the same steps, even though the log-hazard function is non-linear in the measurement error.

The result in Theorem 1 was derived for the case of a known measurement error covariance matrix $\boldsymbol{\Omega}$. The estimating equation approach accommodates the case when the variance of the measurement error is estimated. For simplicity we consider the case when one variable is observed with error and $\sigma^2$, the variance of the measurement error, is estimated based on an estimating equation. The strategy for proving asymptotic normality of the SIMEX estimator is based on stacking the Cox model estimating equation, $\psi_B(\boldsymbol{\beta}(\boldsymbol{\Upsilon}))$, with the measurement error variance estimating equation, $\psi_{\text{var}}(\sigma^2)$. The proofs use methods similar to those in Caroll et al. (1996) [3] and Carroll et al. (2006) [4]. Let $\widehat{\boldsymbol{\beta}}(\Upsilon)$ and $\widehat{\sigma}^2$ be the solution to the system of estimating equations

$$\left\{ \begin{array}{rcl} \psi_B(\boldsymbol{\beta}(\boldsymbol{\Upsilon})) & = & 0 \\ \psi_{\text{var}}(\sigma^2) & = & 0 \end{array} \right. \tag{10}$$

Under standard regularity assumptions for estimating equations one can show that

$$n^{1/2} \left[ \left\{ \begin{array}{c} \widehat{\boldsymbol{\beta}}(\Upsilon) \\ \widehat{\sigma}^2 \end{array} \right\} - \left\{ \begin{array}{c} \boldsymbol{\beta}_0(\Upsilon) \\ \sigma^2 \end{array} \right\} \right] \Rightarrow \text{Normal}(0, \boldsymbol{\Xi}^*) \tag{11}$$

where $\boldsymbol{\Xi}^* = (\boldsymbol{A}^*)^{-1}\boldsymbol{C}^*(\boldsymbol{A}^*)^{-t}$,

$$C^* = \text{var}\left[ n^{-1/2} \left\{ \begin{array}{c} \psi_B(\boldsymbol{\beta}_0(\Upsilon)) \\ \psi_{\text{var}}(\sigma_0^2) \end{array} \right\} \right], \quad \boldsymbol{A}^* = \left( \begin{array}{cc} \boldsymbol{A}_{11}^* & \boldsymbol{A}_{12}^* \\ \boldsymbol{0}_{1\times PM} & \boldsymbol{A}_{22}^* \end{array} \right),$$

$\boldsymbol{A}_{11}^* = \boldsymbol{A}$, $\boldsymbol{A}_{12}^* = \lim_{n\to\infty} \frac{1}{n}\left[ -\frac{\partial}{\partial \sigma^2}\psi_B\{\boldsymbol{\beta}(\Upsilon)\} \right]|_{(\boldsymbol{\beta},\sigma^2)^*}$ for any random $(\boldsymbol{\beta},\sigma^2)^* \to (\boldsymbol{\beta}_0(\Upsilon), \sigma_0^2)$, and $\boldsymbol{A}_{22}^* = \lim_{n\to\infty} \frac{1}{n}\left[ -\frac{\partial}{\partial \sigma^2}\psi_{\text{var}}\{\sigma^2\}|_{(\sigma^2)^*} \right]$ for any random $(\sigma^2)^* \to \sigma_0^2$.

**Theorem 2** *Assume that $\sigma^2$ is estimated from the estimating equations (10) and an exact extrapolant is available. Under conditions similar to those in Andersen and Gill (1982) [1] the SIMEX estimator satisfies*

$$n^{1/2}(\widehat{\boldsymbol{\beta}}_{\text{SIMEX}} - \boldsymbol{\beta}_0) \Rightarrow \text{Normal}(0, \boldsymbol{\Sigma}^*).$$

*If the extrapolant is quadratic then $\boldsymbol{\Sigma}^* = \boldsymbol{\Upsilon}(-1)(\boldsymbol{\Upsilon}^t\boldsymbol{\Upsilon})^{-1}\boldsymbol{\Upsilon}^t\boldsymbol{\Xi}^*\boldsymbol{\Upsilon}(\boldsymbol{\Upsilon}^t\boldsymbol{\Upsilon})^{-1}\boldsymbol{\Upsilon}^t(-1)$, where $\boldsymbol{\Xi}_{11}^*$*

*is the upper left $PM \times PM$ submatrix of $\mathbf{\Omega}^*$.*

In the following sections we turn our attention to the application of SIMEX to the study of progression to primary CKD, while a simulation study will show the practical relevance of our methodology.

# 5 SIMEX for CKD Progression

## 5.1 Simple example

To illustrate our proposed methodology we consider the following simple Cox proportional hazard model for time to primary CKD

$$\lambda_i(t) = \lambda_0(t)\, \exp\{\beta_1 \mathrm{AA}_i + \beta_2 \mathrm{Age}_i + \beta_3 \mathrm{Sex}_i + f(\mathrm{eGFR}_i)\}, \tag{12}$$

where $f(\cdot)$ is a function of the eGFR, and AA denotes the African-American race. We used a linear regression spline with four knots at 90, 105, 125, and 140. More precisely,

$$f(x) = \beta_4 x + \sum_{j=1}^{4} \beta_{j+4}(x - \kappa_j)_+, \tag{13}$$

where $\kappa_j,\ j = 1, \ldots, 4$ are the knots of the spline and $a_+$ is equal to $a$ if $a > 0$ and 0 otherwise. In this parameterization the $\beta_j$ parameter represents the change in the slope of the log hazard ratio at knot $\kappa_j$. The proportional hazard model (12) using the linear spline (13) with fixed knots to describe the effect of eGFR is linear in the $\boldsymbol{\beta}$ parameters but is nonlinear in the variable measured with error. The sample size was $15,080$ with $1,605$ cases of incidence CKD after a median follow-up time of $5,089$ days (roughly 14 years) per subject.

Following the SIMEX methodology we simulated data sets from $\mathrm{eGFR}_i^{\upsilon,b} \sim N(\mathrm{eGFR}_i, \upsilon\sigma_u^2)$ $b = 1, \ldots, B$ where $\sigma_u^2 = 77.56$ was estimated from a different replication study. We used 10 values for $\upsilon$ on an equally spaced grid between 0.2 and 2 and $B = 50$ simulated data sets for each value of $\upsilon$. The entire program was implemented in R and ran in approximately 5 minutes on a PC (3.6GHz CPU, 3.6Gb RAM), with more than 99% of the computation time being dedicated to fitting the 500 Cox models, each with $15,080$ observations.

The parameter estimates $\widehat{\beta}_j^{\upsilon,b},\ j = 1, \ldots, 8$, were obtained by replacing $\mathrm{eGFR}_i^{\upsilon,b}$ for $\mathrm{GFR}_i$ in model (12), and the SIMEX estimates $\widehat{\beta}_j^{\upsilon}$ were obtained by averaging $\widehat{\beta}_j^{\upsilon,b}$ over $b$. Figure 1 (web supplement) displays $\widehat{\beta}_j^{\upsilon},\ j = 1, 2, 3$ in the left column as filled black circles. The parameter estimates are obtained using a quadratic extrapolant evaluated at $\upsilon = -1$, which corresponds to zero measurement error variance and are shown as empty circles. The

| | AA | Age | Sex |
|---|---|---|---|
| Naive | 0.50 | 0.070 | 0.011 |
| SE | 0.059 | 0.0047 | 0.051 |
| SIMEX | 0.63 | 0.054 | 0.061 |
| SE | 0.062 | 0.0049 | 0.052 |

Table 1: *Estimates and standard errors (SE) of risk factors using all subjects with* eGFR $> 60$ *(*$15,080$ *subjects) using events up to 2002. Naive is the regression using the observed eGFR; Here "AA" is African-American race, sex* $= 1$ *indicates males.*

variance of the parameter estimates was obtained using the procedure described in Section 3 and corresponding estimates are presented in the right column of Figure 1 (web supplement). Table 1 provides a comparison between the naive and SIMEX estimates showing that taking measurement error into account increased the log relative hazard for African-American race by 22% The effect of age was decreased by 23%. The effect of sex was not statistically significant either under the naive or the SIMEX procedure.

To obtain the SIMEX estimator of the GFR effect we estimated the function $f(\cdot)$ on an equally spaced grid of points $x_g$, $g = 1, \ldots, G = 100$, between the minimum and maximum observed eGFR. For each level of added noise, $\upsilon\sigma_u^2$, the SIMEX estimator at each grid point, $x_g$, is $\widehat{f}^{\upsilon}(x_g)$, the average over the estimated functions at $x_g$, $\widehat{f}^{\upsilon,b}(x_g)$. For every grid point we then used a quadratic extrapolant to obtain the SIMEX estimator $\widehat{f}^{\{\upsilon=-1\}}(x_g)$. The solid lines in Figure 2 (web supplement) represent the estimated function $f(\cdot)$, $\widehat{f}^{\upsilon}(x_g)$, for $\upsilon = 0, 0.4, 0.8, 1.2, 1.6, 2$, with higher values of noise corresponding to higher intercepts and less shape definition. The bottom dashed line is the SIMEX estimated curve. All functions correspond to 60 year old non-African American males.

The dose/response model implied by the SIMEX estimator displays a number of scientifically interesting details: 1) steeper relative risk for GFR between $[60, 90]$; 2) close to no association with risk between [90,140]; 3) higher relative risk for GFR above 140. Testing whether these features are actually statistically relevant is the focus of the next section.

## 5.2 Hypotheses testing

Hypothesis testing starts with the observation that many of the hypothesis about the dose/effect function can be expressed in terms of linear combinations of the model parameters. Indeed, the log relative hazard rate for GFR between knots $j$ and $j+1$ is $\theta_{j+1} = \beta_4 + \sum_{l=1}^{j} \beta_{4+l}$, for $j = 0, \ldots, 4$. To simplify notation, knots 0 and 5 are de-

fined as the minimum and maximum observed eGFR, respectively. If $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_5)^t$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_8)^t$ then $\boldsymbol{\theta} = \boldsymbol{L}\boldsymbol{\beta}$, where $\boldsymbol{L} = [\mathbf{0}_{5\times 3}|\boldsymbol{T}]$, $\mathbf{0}_{5\times 3}$ is a $5 \times 3$ matrix with zero entries and $\boldsymbol{T}$ is a $5 \times 5$ matrix with all entries equal to zero above the main diagonal and equal to one below and on the main diagonal.

The same SIMEX methodology can be applied for estimation of $\boldsymbol{\theta}$ by noting that at each level of added noise, $\upsilon$, $\widehat{\boldsymbol{\theta}}(\upsilon) = \boldsymbol{L}\widehat{\boldsymbol{\beta}}(\upsilon)$ while $\overline{\mathrm{Var}}\{\widehat{\theta}(\upsilon)\} = \boldsymbol{L}\overline{\mathrm{Var}}\{\widehat{\beta}(\upsilon)\}\boldsymbol{L}^t$. Moreover, $\widehat{\mathrm{Var}}\{\widehat{\theta}(\upsilon)\}$ can be obtained as the variance of $\widehat{\boldsymbol{\theta}}^b(\upsilon) = \boldsymbol{L}\widehat{\boldsymbol{\beta}}^b(\upsilon)$ calculated over all simulations $b = 1, \ldots, B$. Thus, all quantities needed for inference about $\boldsymbol{\theta}$ can be obtained by simple matrix manipulations of standard Cox regressions output software. Point and variance estimators of $\boldsymbol{\theta}$ are obtained using techniques identical to those described for $\boldsymbol{\beta}$. Table 2 displays the naïve and SIMEX estimators of the $\theta$ parameters, their variances and the t-statistic based p-values for testing that $H_0 : \theta_j = 0$, $j = 1, \ldots, 5$.

Results in Table 2 coupled with inspection of Figure 2 (web supplement) shows how our method can uncover and quantify interesting biological features of the dose/response relationship.

Indeed, the nonmonotonic shape of all curves is clear in Figure 2 (web supplement), with unexpected estimated increase in CKD hazard for very large values of GFR. This increase is statistically significant under the naïve approach (p-value=0.009) but is not significant under the SIMEX approach (p-value=0.067). Such results should be interpreted cautiously for two reasons. First, there are only 30 CKD cases with baseline eGFR $> 140$. The total number of CKD cases in our data set was 1605. Second, eGFR is obtained from a prediction equation based on creatinine, which is a muscle product that is filtered out by the kidney. Thus, very low values of creatinine may occur either because the kidney filtration is high or because the subject has lower muscular mass. The latter mechanism may actually be the one that is providing the increasing pattern corresponding to eGFR $> 140$, irrespective of its statistical significance. While SIMEX cannot identify such subjects, accounting for measurement error reduces (not enhances!) the spurious signal.

Interestingly, the GFR effect is strongly significant in the $[60, 90]$ interval both under the naïve and SIMEX approaches, indicating high progression hazards above the current standard, $GFR < 60$. Correcting for measurement error de-attenuated the already strong signal in the interval $[60, 90]$, left signals largely unchanged between $[90, 140]$, and attenuated the spurious signal above 140. The latter effect may just be the result of a lucky combination of factors, since we do not model the likely cause of the measurement error above 140.

## 5.3 Multiple models

Because the SIMEX methodology is computationally fast, multiple models can be fit in reasonable time. For illustration we fitted a series of models similar to (12), called here Model

| | [60, 90] | [90, 105] | [105, 125] | [125, 140] | [140, 200] |
|---|---|---|---|---|---|
| Naive | −.084 | −.014 | −.007 | .000 | .022 |
| SE | .003 | .009 | .010 | .018 | .008 |
| p–value | < .001 | .108 | .527 | .980 | .009 |
| | | | | | |
| SIMEX | −.143 | .023 | −.012 | .014 | .019 |
| SE | .005 | .013 | .017 | .028 | .011 |
| p–value | < .001 | .088 | .486 | .633 | .067 |

Table 2: *Results for the Cox proportional hazard model of time to primary CKD using a* 4 *knot linear spline to model GFR and AA, age, and sex as risk factors. t–tests for statistical significance of log relative risk. The naive testing is based on the naive point estimates and standard errors obtained by using the observed eGFR value. The SIMEX testing is based on the SIMEX point estimates and their standard errors that take into account the measurement error in the observed eGFR.*

1, where we added scientifically relevant confounders while GFR was modeled as a linear spline with 4 knots. More precisely, Model 2 added education (3 ordered levels), income (3 ordered levels), and insurance status (YES/NO); Model 3 added smoking (YES/NO), drinking (YES/NO) and physical activity (continuous); Model 4 added glucose (continuous), body mass index (continuous) and triglycerides (continuous); Model 5 added diabetes (YES/NO), myocardial infarction (YES/NO) and hypertension (YES/NO).

Table 3 displays results for all 5 models using the SIMEX and the naïve approach. For lack of space, we only provide the point estimates where "*" denotes significance at the 95% level using a t-test. Interestingly, the relative hazards for African American race and age are attenuated but remain statistically significant even under increasing covariate adjustments. Education remains statistically significant even in Model 5, while income becomes insignificant after adjustment for behavioral risk factors (smoking, drinking, activity). Insurance and BMI are statistically insignificant under all models, while activity level becomes insignificant after adjusting for biological risk factors (glucose, etc.). Smoking, glucose, triglycerides, diabetes, myocardial infarction, and hypertension are significant in all models that include them. Baseline drinking status is significant in all models except Model 5.

## 5.4 Model selection

The 4 knots used for our linear spline in model (12) were chosen based on scientific input. After addressing the GFR measurement error issues, considerable scientific interest has centered on the sensitivity of estimators and dose/response curve to the number and locations

| | Model 1 | | Model 2 | | Model 3 | | Model 4 | | Model 5 | |
|---|---|---|---|---|---|---|---|---|---|---|
| | N | S | N | S | N | S | N | S | N | S |
| AA | .50* | .63* | .36* | .50* | .27* | .41* | .27* | .40* | .21* | .35* |
| age | .070* | .054* | .065* | .049* | .065* | .050* | .061* | .046* | .056* | .042* |
| sex | .011 | .061 | .046 | .078 | .056 | .082 | .013 | .051 | .005 | .031 |
| $\text{educ.}_2$ | | | −.182* | −.207* | −.142* | −.167* | −.117 | −.150* | −.110 | −.137* |
| $\text{educ.}_3$ | | | −.163* | −.193* | −.091 | −.124 | −.046 | −.093 | −.014 | −.057 |
| $\text{inco.}_1$ | | | −.139 | −.064 | −.116 | −.037 | −.092 | −.011 | −.053 | .029 |
| $\text{inco.}_2$ | | | −.237* | −.189* | −.178* | −.126 | −.123 | −.072 | −.074 | −.031 |
| insur. | | | −.003 | .019 | −.006 | .009 | .034 | .021 | −.010 | −.011 |
| smok. | | | | | .141* | .172* | .168* | .185* | .181* | .201* |
| drink. | | | | | −.193* | −.170* | −.160* | −.144* | −.131* | −.107 |
| activ. | | | | | −.134* | −.114* | −.092 | −.072 | −.090 | −.063 |
| gluc. | | | | | | | .297* | .303* | .190* | .192* |
| bmi | | | | | | | .008 | .005 | .001 | −.000 |
| tri | | | | | | | .284* | .243* | .217* | .190* |
| diab. | | | | | | | | | .444* | .456* |
| MI | | | | | | | | | .376* | .312* |
| HTN | | | | | | | | | .358* | .278* |

Table 3: *Cox proportional hazard model of time to primary CKD using a 4 knot linear spline to model GFR and increasing adjustment. Comparing point estimates using the naive procedure (N) with SIMEX estimates (S). In all these models the GFR is modeled as a linear spline with 4 knots. Here "*" in the exponent denotes significance at the 95% level using a t-test using the corresponding model.*

of knots. We address this issue by considering 8 models starting with a model without knots (linear log-hazard) and ending with a model with 8 knots placed at 65, 75, 80, 85, 90, 105, 125, and 140. More knots were added in the $[60, 90]$ interval because current scientific controversy has focused on the shape of risk just above 60 ml/min/1.73m$^2$.

To illustrate our methodology we focus on the model having African American race, Age and Sex effects in addition to GFR. Figure 3 (web supplement) displays superimposed SIMEX inferences for these three parameters in three models: 1) linear log-hazard (small empty circle); 2) four knot linear spline (larger empty circle); 3) eight knot linear spline. As was probably expected, point estimators become closer to those of the full model, while variances increase, as the number of knots increases.

To compare the effect of increasing the number of knots on estimation of the linear effects we used the mean square error (MSE). The MSE for a model parameter is estimated by

$$\widehat{\mathrm{MSE}}^2(\widehat{\beta}_{j,k}) = \widehat{\mathrm{bias}}^2(\widehat{\beta}_{j,k}) + \widehat{\mathrm{Var}}(\widehat{\beta}_{j,k}),$$

where $\widehat{\beta}_{j,k}$ is the SIMEX estimator of $\beta_j$ based on the model with $k$ knots. The squared bias $\widehat{\mathrm{bias}}^2(\widehat{\beta}_{j,k})$ is estimated by assuming that the the model with 8 knots is the saturated model, that is

$$\widehat{\mathrm{bias}}^2(\widehat{\beta}_{j,k}) = (\widehat{\beta}_{j,k} - \widehat{\beta}_{j,8})^2.$$

The variance $\widehat{\mathrm{Var}}(\widehat{\beta}_{j,k})$ is the estimated SIMEX variance of $\widehat{\beta}_{j,k}$. Figure 4 (web supplement) displays the squared bias (empty circles), variance (empty squares) and MSE (filled circles) for the parameter estimates of African American race, Age and Sex across models. An interesting feature of the plot is that the square bias and not the variance plays an important role in choosing a reasonable number of knots. The MSE risk seems roughly similar across the range of models for the African American race. However, the MSE for age drops dramatically from the model with a linear log-hazard function to the model with two knots, and remains roughly constant thereafter. An interesting, and reassuring, conclusion is that if one is interested in parameter estimation the MSE does not change dramatically with the adjustment for baseline GFR, as long as the adjustment captures the shape of the dose/response function reasonably well.

We now turn our attention to estimating the dose response relationship. One way of comparing the likelihood of various models would be to simulate and extrapolate the log-likelihood of various models following the same recipe as the one used for parameters. While, in theory this sounds reasonable, we discovered a practical problem that has prevented the implementation of this approach. Indeed, the log-likelihood based on observed data does not change much when knots are added. Moreover the slope of the estimated log-likelihood with increased noise is, in some cases, shallower for more complex models which leads to smaller

extrapolated log-likelihood for more complex models. This lead us to compare models based on their log-likelihood calculated on the observed data.

Using observed data, twice the log-likelihood ratio between the models with two knots (at 90 and 125) and without knots (linear log-hazard) was 285.06 (p-value $\approx 0$) indicating strong evidence against the linearity of the log-hazard as a function of the eGFR. The log-likelihood improved only marginally when adding knots in the following sequence 105, 140, 75, 65, 80, 85. The two cases worth mentioning were adding the knots 75 and 85 with log-LRT=3.68 (p-value=0.055) and log-LRT=2.92 (p-value=0.087) respectively. To better understand the effect of adding knots on the shape of the dose/response model Figure 5 (web supplement) shows the same type of results as Figure 2 (web supplement) but for 8 knots instead of 4. Even though two knots were added above 90 the general shape of the dose/response remains very similar to the shape obtained with 4 knots above 90. A small detail is revealed below 90 by the addition of knots at 80 and 85. These knots were not statistically significant in the observed data indicating that there is no evidence in the data for the small bump in the dose response curve between knots 80 and 85. Even if the bump were statistically significant it would not be scientifically relevant. Indeed, its existence would merely indicate insufficient covariate adjustment and not the existence of a "magic" very narrow range of kidney function corresponding to small hazard nestled between intervals with high hazard for progression to primary CKD. Interestingly, SIMEX exacerbates such spurious features of the data leading us to conclude that likelihood ratio testing on the *observed data* performs well in this particular application.

# 6 Simulations

Even the most appealing theories have to pass the minimal performance tests under various relevant simulation scenarios. Given the complexity of the problem described in this paper, especially when the methodology is applied to large cohort studies, we designed our simulation study to be relevant for the CKD application. We were especially interested in the effects of measurement error both at and below the estimated level of variability.

More precisely we used the Cox model (12) and the covariates (race, age, sex, GFR) from the original ARIC data with $\beta_1 = 0.63$ (race), $\beta_2 = 0.054$ (age), $\beta_3 = 0.06$ (sex), $\beta_4 = 0.14$, $\beta_5 = -0.14$, $\beta_6 = \beta_7 = \beta_8 = 0$ (GFR spline). The observed eGFR in the ARIC data was treated as the known GFR (this values will be denoted by GFR in this section and will be treated as true values). The follow-up times for each subject were the follow-up times from the original ARIC study.

Simulations were conducted according to the following algorithm

1. Simulate survival times by

$$Y \sim \text{Exponential}(\boldsymbol{X}\boldsymbol{\beta} + \alpha)$$

where $\boldsymbol{X}$ is the matrix with columns corresponding to race, age, sex and the linear regression spline with 4 knots at 90, 105, 125, 140, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_8)^t$, and $\alpha = 6$ was chosen to ensure roughly 50% censoring.

3. Obtain eGFR by injecting independent normal noise into each observed GFR value by simulating from

$$\text{eGFR}_i^{(s)} \sim \text{Normal}(\text{GFR}_i, \sigma_u^2)$$

where $\sigma_u^2$ is known at fixed at one of the levels 9, 16, or 77.56. This is a way to simulate noisy measurements and is not part of the SIMEX fitting algorithm.

4. Obtain the naive estimates by fitting the Cox model (12) using the $\text{eGFR}^{(s)}$ values instead of the true GFR.

5. Obtain the SIMEX corrected estimates for the Cox model (12) using $\sigma_u^2$ as the known measurement error variance.

For every level of noise, $\sigma_u^2$, we have simulated 100 data points and calculated the naive and SIMEX estimates. For each parameter the log of the squared bias and MSE was estimated from simulations as

$$\widehat{\log(\text{Bias}^2)}(\widehat{\beta}_j) = \log\{\frac{1}{S}\sum_{s=1}^{S}(\widehat{\beta}_j - \beta_j)^2\}, j = 1, \ldots, 8,$$

and

$$\widehat{\log(\text{MSE})}(\widehat{\beta}_j) = \log\{\frac{1}{S}\sum_{s=1}^{S}(\widehat{\beta}_j - \beta_j)^2 + \frac{1}{S}\sum_{s=1}^{S}\widehat{\text{Var}}(\widehat{\beta}_j)\},$$

respectively, where $S = 100$ is the number of simulations, $\widehat{\beta}_j$ is the estimated $\beta_j$ parameter, and $\widehat{\text{Var}}\widehat{\beta}_j\}$ is the estimated variance for a given method.

Table 4 displays the log squared bias and MSE for all the parameters of model (12) calculated based on $S = 100$ simulations using naive and SIMEX estimation. As expected, SIMEX consistently outperforms the naive method both in terms in squared bias and MSE, most of the improvement being in reducing the squared bias with a small price paid in terms of variance. Another consistent feature is that both squared bias and MSE seem to be getting smaller when the variance of the measurement error process increases. The only exceptions are the estimators of the $\beta_6 = \beta_7 = \beta_8 = 0$ parameters for which the bias and

MSE decrease when measurement error variance increases and SIMEX is outperformed by the naive analysis when $\sigma_u^2 = 77.56$. Interestingly the bias of these estimators is at least one order of magnitude smaller than the bias for parameters that are not zero. Moreover the biases of the naive and SIMEX estimators are both scientifically negligible. For example, the mean of the naive and SIMEX estimators for $\beta_7$ are 0.001 and $-0.01$ respectively. Both these values correspond to changes of the log hazard that are not scientifically relevant.

# 7 Comments

In this paper we proposed a simple and computationally usable extension of the SIMEX methodology for first order bias correction in Cox models with a log hazard function that is linear in parameters but non-linear in variables measured with error (LPNE). Our solution addresses a real need for new and feasible inferential methodology in apparently simple cases when the log-hazard contains strata indicators, splines, quadratic and interaction terms of variables observed with error. While SIMEX is consistent only if a correct extrapolant is available, our results indicate that SIMEX can substantially improve estimation even when this is not the case. An important characteristic of our methodology is that it can be used with realistic data sets, such as the ARIC study.

# Acknowledgements

Ciprian Crainiceanu and Joe Coresh's research was supported by NIH/NHLBI grant R01HL62985-01A1. David Ruppert's research was supported by NSF Grant DMS 04-538 and NIH Grant CA57030.

Table 4: Log squared bias and MSE estimated from $S = 100$ simulated data sets from model (12) using naive and SIMEX estimation (smaller is better)

| | | $\sigma_u^2 = 9$ | | $\sigma_u^2 = 77.56$ | |
|---|---|---|---|---|---|
| | | Naive | SIMEX | Naive | SIMEX |
| $\beta_1$ | $\log(\text{Bias}^2)$ | -4.94 | -6.02 | -2.93 | -3.98 |
| | $\log(\text{MSE})$ | -4.84 | -5.72 | -2.92 | -3.93 |
| $\beta_2$ | $\log(\text{Bias}^2)$ | -10.59 | -10.94 | -11.16 | -12.04 |
| | $\log(\text{MSE})$ | -10.42 | -10.73 | -10.86 | -11.50 |
| $\beta_3$ | $\log(\text{Bias}^2)$ | -6.80 | -7.05 | -5.69 | -5.97 |
| | $\log(\text{MSE})$ | -6.41 | -6.53 | -5.55 | -5.75 |
| $\beta_4$ | $\log(\text{Bias}^2)$ | -6.91 | -9.25 | -5.10 | -6.08 |
| | $\log(\text{MSE})$ | -6.91 | -9.20 | -5.10 | -6.08 |
| $\beta_5$ | $\log(\text{Bias}^2)$ | -5.39 | -7.54 | -4.46 | -5.16 |
| | $\log(\text{MSE})$ | -5.38 | -7.49 | -4.46 | -5.15 |
| $\beta_6$ | $\log(\text{Bias}^2)$ | -5.71 | -7.67 | -7.00 | -6.53 |
| | $\log(\text{MSE})$ | -5.70 | -7.56 | -6.96 | -6.47 |
| $\beta_7$ | $\log(\text{Bias}^2)$ | -6.74 | -8.14 | -9.15 | -7.71 |
| | $\log(\text{MSE})$ | -6.67 | -7.81 | -8.47 | -7.31 |
| $\beta_8$ | $\log(\text{Bias}^2)$ | -8.19 | -8.68 | -9.09 | -8.10 |
| | $\log(\text{MSE})$ | -7.86 | -8.16 | -8.41 | -7.59 |

# References


[1] P.K. Andersen and R.D. Gill. Cox's regression model for counting processes: A large sample study. *The Annals of Statistics*, 10:1100–1120, 1982.

[2] T. Augustin. An exact corrected log-likelihood function for cox's proportional hazards model under measurement error and some extensions. *Scandinavian Journal of Statistics*, 31:43–50, 2004.

[3] R.J. Carroll, H. Küechenhoff, F. Lombard, and L.A. Stefanski. Asymptotics of the simex estimator in structural measurement error models. *Journal of the American Statistical Association*, 91:242–250, 1996.

[4] R.J. Carroll, D. Ruppert, L.A. Stefanski, and C.M. Crainiceanu. *Measurement Error in Nonlinear Models, A modern Perspective, Second Edition.* Chapman & Hall CRC, New York, 2006.

[5] D.G. Clayton. *Models for the analysis of cohort and case-control studies with inaccurately measured exposures. In Statistical Models for Longitudinal Studies of Health, eds.: J.H. Dwyer, et al.* Oxford University Press, New York, USA, 1991.

[6] J. Cook and L.A. Stefanski. A simulation extrapolation method for parametric measurement error models. *Journal of the American Statistical Association*, 89:1314–1328, 1995.

[7] D.R. Cox. Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, Series B*, 34:187–220, 1972.

[8] W.F. Greene and J. Cai. Measurement error in covariates in the marginal hazards model for multivariate failure time data. *Biometrics*, 60:987–996, 2004.

[9] C. Hu and D.Y. Lin. Semiparametric failure time regression with replicates of mismeasured covariates. *Journal of the American Statistical Association*, 99:105–118, 2004.

[10] P. Hu, A.A. Tsiatis, and M. Davidian. Estimating the parameters of the cox model when covariate variables are measured with errors. *Biometrics*, 54:1407–1419, 1998.

[11] Y. Huang and C.Y. Wang. Cox regression with accurate covariates unascertainable: a nonparametric correction approach. *Journal of the American Statistical Association*, 96:1469–1482, 2000.

[12] The ARIC INVESTIGATORS. The atherosclerosis risk in communities (aric) study: design and objectives. *American Journal of Epidemiology*, 129:687–702, 1989.

[13] Y. Li and X. Lin. Functional inference in frailty models for clustered survival data using the simex approach. *Journal of the American Statistical Association*, 98:191–203, 2003.

[14] J. Marsh-Manzi, C.M. Crainiceanu, B.C. Astor, N.R. Powe, M.J. Klag, H.A. Taylot, and J. Coresh. Increased risk of ckd progression and esrd in african americans: The atherosclerosis risk in communities (aric) study. *under review*.

[15] T. Nakamura. Proportional hazards models with covariates subject to measurement error. *Biometrics*, 48:829–838, 1992.

[16] M.S. Pepe, S.G. Self, and R.L. Prentice. Further results in covariate measurement errors in cohort studies with time to response data. *Statistics in Medicine*, 8:1167–1178, 1989.

[17] R.L. Prentice. Covariate measurement errors and parameter estimation in a failure time regression model. *Biometrika*, 69:331–342, 1982.

[18] X. Song and Y. Huang. On corrected score approach for proportional hazards model with covariate measurement error. *Biometrics*, 61:702–714, 2005.

[19] L.A. Stefanski. Unbiased estimation of a nonlinear function of a normal mean with application to measurement error models. *Communications in Statistics, Series A*, 18:4335–4358, 1989.

[20] L.A. Stefanski and J. Cook. Unbiased estimation of a nonlinear function of a normal mean with application to measurement error models. *Simulation extrapolation: the measurement error jackknife*, 90:1247–1256, 1995.

[21] L.A. Stevens, J. Coresh, T. Greene, and A.S. Levey. Assessing kidney function–measured and estimated glomerular filtration rate. *New England Journal of Medicine*, 354(23):2473–2483, 2006.

[22] A.A. Tsiatis. A large sample study of cox's regression model. *The Annals of Statistics*, 9:93–108, 1981.

[23] A.A. Tsiatis and M. Davidian. A semiparametric estimator for the proportional hazards model with longitudinal covariates measured with error. *Biometrika*, 88:447–458, 2001.

[24] C.Y. Wang, L. Hsu, Z.D. Feng, and R.L. Prentice. Regression calibration in failure time regression. *Biometrics*, 53:131–145, 1997.

[25] H. Zhou and M.S. Pepe. Auxiliary covariate data in failure time regression. *Biometrika*, 82:139–149, 1995.

[26] H. Zhou and C.Y. Wang. Failure time regression with continuous covariates measured with error. *Journal of the Royal Statistical Society, Series B*, 62:657–665, 2000.