HHS Public Access

Author manuscript

Lifetime Data Anal. Author manuscript; available in PMC 2017 January 01.

Published in final edited form as:

Lifetime Data Anal. 2016 January ; 22(1): 1–16. doi:10.1007/s10985-014-9315-7.

A semiparametric copula method for Cox models with covariate measurement error

Sehee Kim,

Department of Biostatistics, University of Michigan, Ann Arbor, MI, USA

Yi Li, and

University of Michigan Kidney Epidemiology and Cost Center, Department of Biostatistics, University of Michigan, Ann Arbor, MI, USA

Donna Spiegelman

Departments of Epidemiology, Biostatistics, and Nutrition, Harvard University, Boston, MA, USA

Sehee Kim: seheek@umich.edu; Yi Li: yili@umich.edu; Donna Spiegelman: stdls@hsph.harvard.edu

Abstract

We consider measurement error problem in the Cox model, where the underlying association between the true exposure and its surrogate is unknown, but can be estimated from a validation study. Under this framework, one can accommodate general distributional structures for the error-prone covariates, not restricted to a linear additive measurement error model or Gaussian measurement error. The proposed copula-based approach enables us to fit flexible measurement error models, and to be applicable with an internal or external validation study. Large sample properties are derived and finite sample properties are investigated through extensive simulation studies. The methods are applied to a study of physical activity in relation to breast cancer mortality in the Nurses' Health Study.

Keywords

Bias correction; Copula; Error-prone covariate; Measurement error; Semiparametric; Survival analysis

1 Introduction

In epidemiologic studies, risk factors such as nutrient intake, physical activity or air pollutants are often subject to measurement error. When the error-prone risk factors are included in a Cox (1972) survival model, the estimated effects of these model covariates can be under- or over-estimated, even for covariates that are measured without error. To address the bias caused by the mis-measured exposure, we focus on modeling the association between the true exposure "X" and its surrogate measure "Z" directly, that is $f_{X|Z}$, the conditional density of X given Z. The measurement error model considered here allows the observed exposure distribution to differ from the distribution of the true exposure in unrestricted form and to be estimated from validation data. Our research is motivated by a study of physical activity in relation to breast cancer mortality in the Nurses' Health Study (Holmes et al, 2005). The commonly used linear additive measurement error model fails for

describing the relationship between true and surrogate physical activity measurements because the surrogate has a much heavier density around zero than the true counterpart as shown in Fig. 1, and a non-linear regression model is better fit to the data than the simple linear regression model (Fig. 2).

In the presence of covariate measurement error, the Cox regression model has been studied by many authors. Key issues in measurement error models are the assumption made about the distribution of measurement error and the existence of validation data. Some approaches work under $f_{X|Z}$ to be known up to a parametric form (Hu et al, 1998; Zucker, 2005; Wen, 2010), while others rely less on this. However, the latter requires other assumptions that the former methods do not or has some drawback. Among them, a simple and intuitive way of handling measurement error without any distributional assumption is the regression calibration approach (Wang et al, 1997; Spiegelman et al, 1997; Xie et al, 2001), which replaces the unobserved true exposure X with an "estimate" of X given its surrogate, and then obtains the standard partial likelihood estimator. This method assumes mean function and a constant variance for measurement error, but nothing more. Since these regression calibration estimators are based on a linear approximation to the expectation of X given Z, they may result in an inconsistent regression coefficient estimator. Another approach that does not require any distributional assumption on the measurement error is the "estimated" partial likelihood method proposed by Zhou and Pepe (1995) and Zhou and Wang (2000). Unlike the regression calibration estimators, they deal with the induced relative risk function directly in a nonparametric way, and then estimate the regression coefficients from the resulting estimated partial likelihood function. However, their methods require an internal validation study.

With a single measurement of error-prone covariate, Hu et al (1998) proposed a likelihood-based semiparametric method, assuming $f_{Z|X}(z|x)$ is known, whereas $f_X(x)$, the marginal density of X, is unknown. They showed that their semiparametric estimator was more efficient than a fully nonparametric approach of $f_X(x)$. Another semiparametric approach based on a conditional score estimator (Tsiatis and Davidian, 2001; Song et al, 2002) has been proposed under a known parametric $f_{Z|X}$ and an unknown $f_X(x)$, however, the method requires longitudinal measurements of error-prone covariate. Unlike them, we propose to directly estimate $f_{X|Z}$ by introducing copulas into measurement error framework for the first time to our knowledge. Our approach is semiparametric in that the marginal distributions of X and Z are not specified, but they are linked by a parametric copula. To make our approach more robust to the choice of copula, we take multiple copulas into consideration simultaneously. Detailed description of our estimation procedure based on copula is provided in Section 3.1.

In terms of the existence of validation data, most of work to date requires subsampling within a main cohort study, i.e., an internal validation study (Zhou and Pepe, 1995; Wang et al, 1997; Zhou and Wang, 2000; Chen, 2002) or a reliability study, while some utilizes validation samples from a separate cohort that contains no information on outcome variable, i.e., an external validation study. If such a validation study does not exist, the measurement error model must be fully specified with all its parameters assumed known. Otherwise, the model parameters (β, γ) of equation (1) are not identifiable. Direct use of the surrogate Z as

a substitute for X in the presence of measurement error is known to yield biased estimates. Thus, to correct for measurement errors in covariates, the relationship between X and Z needs to either be fully known or estimated via a validation or reliability study. Depending on the type of validation (or reliability) data available, different approaches were previously proposed as follows. The work by Xie et al (2001) and Huang and Wang (2000) requires replicate measurements on at least a subset of the study population, thereby requiring the classical additive model. In longitudinal study setting, where the error-prone covariates were repeatedly measured, the estimating equation approach has been employed by Tsiatis and Davidian (2001) adopting a conditional score method and by Wang (2006) via a corrected score approach. Recently, Zucker (2005) proposed a method that can be applied with an external validation study, although their approach requires a known parametric form of measurement error. However, in our motivating data example, only external validation samples with no replicates were available and, moreover, X and Z did not follow a known distribution, which precluded the direct use of these existing methods.

Distinct from existing methods, the proposed estimator in this paper does require neither linear additivity nor any parametric distributional assumption on the measurement error model, and yet is applicable to either an internal or external validation study. The key feature that makes this greater flexibility possible is the use of a semiparametric copula-based procedure for estimating the joint distribution of the true exposure X and the surrogate Z, which greatly reduces bias due to exposure measurement error but would still be more efficient than nonparametric approaches.

2 Preliminaries

Let $\tilde{T} = \min(T, C)$ be the observed follow-up time, where T and C are the failure and censoring times, respectively. A natural model that links the outcome T to the covariates is the following proportional hazards model

$$\lambda_c(t|X,W)=\lambda(t)\exp(\beta X+\gamma^T W), \quad (1)$$

where $\lambda(\cdot)$ is an unspecified baseline hazard function, β and γ are unknown regression parameters corresponding to X, an error-prone covariate (e.g., the detailed physical activity diary), and W, a vector of error-free covariates (e.g., age and gender), respectively. Usually, X is the covariate of main interest and is difficult or expensive to measure. In a typical epidemiological study, we observe Z (e.g., the self-administered physical activity measure) in lieu of X, and Z is often termed the surrogate for X.

Suppose we have i.i.d. observations on n individuals in the main cohort study. Using the counting process notation, let $Y_i(t) = I(\tilde{T}_i \geq t)$ be the at-risk process, and $N_i(t) = I(\tilde{T}_i \leq t, T_i \leq C_i)$ be the counting process, where $I(\cdot)$ is the indicator function. The main cohort study consists of $\{\tilde{T}_i, Y_i(t), N_i(t), Z_i, W_i; 0 \leq t \leq \tau\}$ $(i = 1,\ldots, n)$, where τ is the duration of study. We assume that Z_i is independent of outcome T_i given $\{X_i, W_i\}$, and X_i is independent of W_i given Z_i. The former assumption corresponds to the non-differential measurement error assumption, while the latter is assumed for notational ease and can easily be relaxed without loss of generality. Finally, we assume that C_i and T_i are conditionally independent given

observed $\{Z_i, W_i\}$, that is a non-informative censoring condition commonly used in survival analysis.

We specify $S_i^*(t)$ the expected survival function given the observed data only, as discussed by Prentice (1982) and Zucker (2005), i.e.,

$$S_i^*(t)=\Pr[T_i>t|Z_i,W_i]=\int\exp[-\Lambda(t)\exp\{\beta x+\gamma^T W_i\}]f_{X|Z}(x|Z_i)dx,$$

where $f_{X|Z}$ is the conditional density of the true covariate X given the observed surrogate Z, and $\Lambda(t)$ is the cumulative baseline hazard function. The corresponding conditional hazard function is given by

$$d\Lambda_i^*(t)=\lambda(t)\frac{\int\psi_i(x;\beta,\gamma)\exp\{-\Lambda(t)\psi_i(x;\beta,\gamma)\}f_{X|Z}(x|Z_i)dx}{\int\exp\{-\Lambda(t)\psi_i(x;\beta,\gamma)\}f_{X|Z}(x|Z_i)dx}=\lambda(t)E[\psi_i(x;\beta,\gamma)|Z_i,W_i,T_i>t]\equiv\lambda(t)\eta_i(\beta,\gamma,\Lambda(t)), \quad (2)$$

where $\psi_i(x; \beta, \gamma) = \exp\{\beta x + \gamma^T W_i\}$.

3 INFERENCE PROCEDURES

3.1 Estimation of $f_{X|Z}(x|z)$

The treatment of the conditional density $f_{X|Z}(x|z)$ is the key to this development. Indeed, when both X and Z are deemed as random variables, which is always the case in observational studies, it is a matter of mathematical convenience whether the measurement error model is specified conditional on X or on Z. Conditioning on X is more often adopted in classical measurement error settings for ease of interpretation. However, specifying the model in this way makes stronger transportability assumptions than the other way around. Specifically, for an external validation study, the estimated $f_{Z|X}(z|x)$ in the validation study may not be transportable to the main study. This happens because $f_{Z|X}$ may not entirely be identifiable over the support of X as X may distribute differently across the main and external validation samples. On the other hand, working directly with $f_{X|Z}$ may circumvent such a difficulty as Z's are fully observed in both main and external validation samples.

Deviating from the common measurement error literature, we postulate a general framework that directly deals with $f_{X|Z}(x|z)$. Given the availability of a validation study where both X and Z are observed, $f_{X|Z}(x|z)$ can be estimated nonparametrically. However, when the sample size in the validation data set is moderate or small, as in our motivating example, estimation of $f_{X|Z}$ nonparametrically would be less favorable. As a remedy, we propose a semiparametric method that utilizes a copula framework for estimating $f_{X|Z}$. The motivation stems from the fact that $f_{X|Z}$ can be viewed as a functional of F_X and F_Z, the marginal distributions of X and Z respectively, in a copula setting. Specifically, by invoking the Sklar (1959) theorem, it follows that

$$f_{X|Z}(x|z)=\frac{f_{XZ}(x,z)}{f_Z(z)}=C'_\xi(F_X(x),F_Z(z))f_X(x),$$

provided that $f_Z(z) > 0$, where C'_ξ is a copula density function with a dependence parameter ξ in $\mathbb{R}$. Furthermore, if F_X and F_Z are continuous as in our case, then the copula distribution function C_ξ is uniquely determined (Sklar, 1959). Utilizing these nice properties in copulas, we propose to estimate $f_{X|Z}(x|z)$ as follows:

Step 1. Estimate $f_X(x)$ and $f_Z(z)$ separately in the validation study through kernel density estimation (Wand and Jones, 1995). Suppose (X_j, Z_j) is the jth sample in the validation study. The kernel density estimator of f_X at the point x is given by

$$\hat{f}_X(x) = n_\upsilon^{-1} \sum_{j=1}^{n_\upsilon} b^{-1} K\left(\frac{x - X_j}{b}\right),$$

where n_υ is the sample size of the validation study, the kernel K satisfies $\int K(x)\,dx = 1$ and the bandwidth b controls the degree of smoothness of the density function.

Step 2. Compute $\hat{F_X}(x_j)$ and $\hat{F_Z}(z_j)$ at the observed validation samples $\{(x_j, z_j); j = 1,\ldots, n\upsilon\}$.

Step 3. Estimate the dependence parameter ξ by maximizing the likelihood of the validation study, i.e., $\prod_{j=1}^{n_\upsilon} C_\xi(\hat{F}_X(x_j), \hat{F}_Z(z_j))$ for each of the following three forms of C_ξ: Gaussian, Clayton, and Gumbel copulas, for example.

Step 4. Select copula which is best fit for the data at hand by using the Akaike Information Criterion (AIC), and then obtain $\hat{f}_{X|Z}(x|z) = C'_{\hat{\xi}}(\hat{F}_X(x), \hat{F}_Z(z))\hat{f}_X(x)$.

Once $\hat{f_{X|Z}}$ is available, the semiparametric estimator of (β, γ) can be obtained by solving estimating equations proposed in Section 3.2.

In Step 1, the kernel K can be chosen to be a unimodal probability density function symmetric about zero, satisfying the conditions: $\int xK(x)\,dx = 0$ and $\int x^2K(x)\,dx > 0$. Our method can be applied with any choice of K. However, we consider particularly the two most popular choices for K: the Gaussian kernel (Eubank, 1988) $K_G(u) = (2\pi)^{-1/2} \exp(-u^2/2)$ and the Epanechnikov kernel (Eubank, 1988) $K_E(u) = 3/4\,(1 - u^2)I(|u| \le 1)$. In Step 2, a numerical integration algorithm such as Gaussian quadrature can be used. When the Gaussian kernel is specified in Step 1, Gaussian quadrature with the Her-mite polynomials is used for approximating the integral over $(-\infty, \infty)$, while Legendre polynomials are used for approximating the integral over the finite support $[-1, 1]$ when the Epanechnikov kernel is specified.

Steps 3–4 describe a model selection procedure of copula. We suggest to begin with a visual inspection of the scatter plot. Some dependence structures may already be identified such as stronger dependence in one of the tails. However, in case where one is uncertain which copula functions to consider, the three types of copulas stated in Step 3 are good options because they encompass a variety of dependence structures: the Gaussian copula describes symmetric dependence, while Clayton and Gumbel copulas are suited for stronger negative and positive tail correlations, respectively. There are other parametric copula families

available (see Nelsen, 2006, Chap 4.3), and they can also be accommodated in Step 3. With the candidate copula models, the next step is to estimate copula parameters and AIC for each of the candidates. The model with the smallest AIC is considered the best fitting one. The performance of AIC as a bivariate copula selection criterion has previously been investigated by Manner (2007), Dissmann et al (2013), Grønneberg and Hjort (2014), and Jordanger and Tjøstheim (2014). In our simulation study, we observed that AIC selected the true copula model (approximately 95%) or the model with a similar dependence structure in most cases.

3.2 Semiparametric Copula Estimator for (β, γ)

Let $M_i(t;\beta,\gamma,\Lambda)=N_i(t)-\int_0^t Y_i(s)d\Lambda_i^*(s)=N_i(t)-\int_0^t Y_i(s)\eta_i(\beta,\gamma,\Lambda(s))d\Lambda(s)$. We denote the true value of β by β_0, the true vector of γ by γ_0, and the true $\Lambda(t)$ by $\Lambda_0(t)$. With non-informative censoring, $M_i(t) \equiv M_i(t; \beta_0, \gamma_0, \Lambda_0)$ ($i = 1,\ldots, n$) is a martingale process with respect to the filtration $\sigma\{N_i(s-), Y_i(s), W_i, Z_i, 0 \le s \le t\}$. As such, we propose the following estimating equations:

$$0=\sum_{i=1}^{n}\int_0^\tau \mu(Z_i)dM_i(t;\beta,\gamma,\Lambda), \quad (3)$$

$$0=\sum_{i=1}^{n}\int_0^\tau W_i dM_i(t;\beta,\gamma,\Lambda), \quad (4)$$

$$0=\sum_{i=1}^{n}[dN_i(t)-Y_i(t)\eta_i(\beta,\gamma,\Lambda(t))d\Lambda(t)], \quad (5)$$

where $\mu(z) = \int x\, f_{X|Z}(x|z)\, dx$. Since these estimating equations are functions of $f_{X|Z}$ through $\mu(Z_i)$ and $\eta_i(\beta, \gamma,\Lambda(t))$, which is unknown, we modify the equations (3)–(5) by replacing $f_{X|Z}$ with its estimate $\hat{f}_{X|Z}$ and denote the resulting estimates by $\hat{\mu}(Z_i)$ and $\hat{\eta}_i(\beta, \gamma,\Lambda(t))$. The equation (5) yields a Breslow-type estimator for $\Lambda_0(t)$

$$\hat{\Lambda}(t)=\sum_{i=1}^{n}\int_0^t \frac{dN_i(u)}{\sum_{i=1}^{n}Y_i(u)\hat{\eta}_i(\beta,\gamma,\hat{\Lambda}(u-))}, \text{ or } \hat{\lambda}(u_m)=\frac{d_m}{\sum_{i=1}^{n}Y_i(u_m)\hat{\eta}_i(\beta,\gamma,\hat{\Lambda}(u_{m-1}))},$$

for $u_m \le t < u_{m+1}$, where $u_1, u_2,\ldots, u_M$ are the ordered observed event times, d_m is the number of events at u_m, and $\hat{\Lambda}(0) = 0$. A similar estimate $\hat{\Lambda}(t)$ was obtained by Zucker (2005) in a different context based on a pseudo-partial likelihood function. Define $Q_i=(\hat{\mu}(Z_i), W_i^T)^T$ as a vector of the observed covariates and $\theta_0=(\beta_0,\gamma_0^T)^T$ as the vector of corresponding true regression coefficients. By substituting $\hat{\Lambda}(t)$ and $\hat{f}_{X|Z}$ for $\Lambda(t)$ and $f_{X|Z}$ in $M_i(t; \beta, \gamma, \Lambda)$, the modified estimating equations for θ_0 in (3) and (4) become

$$U(\theta,\hat{\Lambda},\hat{f}_{X|Z})=\sum_{i=1}^{n}\int_0^\tau \left\{Q_i-\frac{S^{(1)}(t;\theta,\hat{\Lambda},\hat{f}_{X|Z})}{S^{(0)}(t;\theta,\hat{\Lambda},\hat{f}_{X|Z})}\right\}dN_i(t), \quad (6)$$

where $S^{(k)}(t;\theta,\Lambda,f_{X|Z})=n^{-1}\sum_{i=1}^{n}Y_i(t)Q_i^{\otimes k}\eta_i(\theta,\Lambda(t))$ (k = 0, 1, 2), $a^{\otimes 0}=1$, $a^{\otimes 1}=a$, and $a^{\otimes 2}=aa^T$. Then, $\hat{\theta}$ is the solution to $U(\theta,\Lambda,\hat{f}_{X|Z})=0$, which can be obtained numerically using the Newton-Raphson algorithm, for example.

To compute the covariance matrix of $\hat{\theta}$, we further define

$$S_\theta^{(k)}(t;\theta,\Lambda,f_{X|Z})=n^{-1}\sum_{i=1}^{n}Y_i(t)Q_i^{\otimes k}\dot{\eta}_{\theta i}(\theta,\Lambda(t)),\ \ S_\Lambda^{(k)}(t;\theta,\Lambda,f_{X|Z})=n^{-1}\sum_{i=1}^{n}Y_i(t)Q_i^{\otimes k}\dot{\eta}_{\Lambda i}(\theta,\Lambda(t)),$$

where

$$\begin{aligned}\dot{\eta}_{\theta i}&=\partial\eta_i(\theta,\Lambda(t))/\partial\theta\\&=\frac{E_{X|Z}\left[\tilde{Q}_i\psi_i(X_i;\theta)e^{-\Lambda(t)\psi_i(X_i;\theta)}-\psi_i^2(X_i;\theta)e^{-\Lambda(t)\psi_i(X_i;\theta)}\{\tilde{Q}_i\Lambda(t)+\partial_\theta\Lambda(t)\}\right]}{E_{X|Z}\left[e^{-\Lambda(t)\psi_i(X_i;\theta)}\right]}\\&+\frac{E_{X|Z}\left[\psi_i(X_i;\theta)e^{-\Lambda(t)\psi_i(X_i;\theta)}\right]E_{X|Z}\left[\psi_i(X_i;\theta)e^{-\Lambda(t)\psi_i(X_i;\theta)}\{\tilde{Q}_i\Lambda(t)+\partial_\theta\Lambda(t)\}\right]}{\{E_{X|Z}\left[e^{-\Lambda(t)\psi_i(X_i;\theta)}\right]\}^2}\end{aligned}$$

and

$$\dot{\eta}_{\Lambda i}=\partial\eta_i(\theta,\Lambda(t))/\partial\Lambda(t)=-\frac{E_{X|Z}\left[\psi_i^2(X_i;\theta)e^{-\Lambda(t)\psi_i(X_i;\theta)}\right]}{E_{X|Z}\left[e^{-\Lambda(t)\psi_i(X_i;\theta)}\right]}+\left\{\frac{E_{X|Z}\left[\psi_i(X_i;\theta)e^{-\Lambda(t)\psi_i(X_i;\theta)}\right]}{E_{X|Z}\left[e^{-\Lambda(t)\psi_i(X_i;\theta)}\right]}\right\}^2$$

at a fixed time t, where $\tilde{Q}_i=(X_i,W_i^T)^T$, $E_{X|Z}$ denotes the conditional expectation with respect to X given Z, and

$$\partial_\theta\Lambda(t)=-\sum_{i=1}^{n}\int_0^t\left\{\sum_i Y_i(u)\dot{\eta}_{\theta i}(\theta,\Lambda(u-))\right\}\left\{\sum_i Y_i(u)\eta_i(\theta,\Lambda(u-))\right\}^{-2}dN_i(u).$$

To establish asymptotic properties of a semiparametric estimator $\hat{\theta}$, we will first prove rigorously asymptotic normality of $\tilde{\theta}$ in Theorem 1, where $\tilde{\theta}$ is estimated under the assumption that $f_{X|Z}(x|z;\xi)$ belongs to a known parametric family indexed by a vector of unknown parameters ξ. Then, we will extend to the asymptotic normality of $\hat{\theta}$ under the semiparametric estimation of $f_{X|Z}(x|z)$ in Theorem 2. The regularity conditions and proofs of Theorems 1 – 2 (appear below) are given in the Web Appendices.

Theorem 1—*Under regularity conditions (C1)–(C7),* $\tilde{\theta}$ *is a consistent estimator of* θ_0 *and* $n^{1/2}(\tilde{\theta}-\theta_0)$ *is asymptotically normally distributed with mean 0 and variance-covariance matrix* $D(\theta_0)^{-1}+D(\theta_0)^{-1}[H+V(\xi_0)\Omega V(\xi_0)^T]D(\theta_0)^{-1}$, *where* ξ_0 *is the true value of* ξ, Ω *is the variance of the maximum likelihood estimator* $\tilde{\xi}$ *and* $V(\xi_0)$ *is the limit of* $n^{-1}\partial U(\theta_0,\Lambda_0,\xi_0)/\partial\xi$.

The exact forms of $D(\theta_0)$, H, and $V(\xi_0)$ are provided in Web Appendices, which can be consistently estimated by

$$\tilde{D}=n^{-1}\sum_{i=1}^{n}\int_0^\tau \left\{S_\theta^{(1)}/S^{(0)}(t;\tilde{\theta},\tilde{\Lambda},\tilde{\xi}) - S^{(1)}S_\theta^{(0)}/(S^{(0)})^2(t;\tilde{\theta},\tilde{\Lambda},\tilde{\xi})\right\} dN_i(t),$$

$$\tilde{H}=n^{-1}\sum_{i=1}^{n}\int_0^\tau \frac{G(t;\tilde{\theta},\tilde{\Lambda},\tilde{\xi})^{\otimes 2}\tilde{R}(t-)^2}{\left\{\sum_{i=1}^n Y_i(t)\eta_i(\tilde{\theta},\tilde{\Lambda}(t))\right\}^2} dN_i(t),$$

$$\tilde{R}(t)=\prod_{u\le t}\left\{1+\sum_{i=1}^{n} nS_\Lambda^{(0)}/(S^{(0)})^2(u;\tilde{\theta},\tilde{\Lambda},\tilde{\xi})dN_i(u)\right\},$$

$$G(t;\tilde{\theta},\tilde{\Lambda},\tilde{\xi})=\sum_{i=1}^{n}\int_t^\tau \{S^{(1)}S_\Lambda^{(0)}/(S^{(0)})^2(u;\tilde{\theta},\tilde{\Lambda},\tilde{\xi})-S_\Lambda^{(1)}/S^{(0)}(u;\tilde{\theta},\tilde{\Lambda},\tilde{\xi})\}/\tilde{R}(u)dN_i(u),$$

and

$$\tilde{V}=n^{-1}\sum_i\int_0^\tau \left[\dot{Q}_{\xi i}(\tilde{\xi})+\frac{\sum_i Y_i(t)\eta_i(\tilde{\theta},\tilde{\Lambda})Q_i(\tilde{\xi})\sum_i Y_i(t)\dot{\eta}_{\xi i}(\tilde{\theta},\tilde{\Lambda})^T}{\{\sum_i Y_i(t)\eta_i(\tilde{\theta},\tilde{\Lambda})\}^2}\right] dN_i(t)$$

$$-n^{-1}\sum_{i=1}^{n}\int_0^\tau \left[\frac{\sum_i Y_i(t)Q_i(\tilde{\xi})\dot{\eta}_{\xi i}(\tilde{\theta},\tilde{\Lambda})^T\sum_i Y_i(t)\eta_i(\tilde{\theta},\tilde{\Lambda},)\dot{Q}_{\xi i}(\tilde{\xi})}{\sum_i Y_i(t)\eta_i(\tilde{\theta},\Lambda_0,)}\right] dN_i(t),$$

where $\dot{Q}_{\xi i}$ and $\dot{\eta}_{\xi i}$ are the partial derivatives of $Q_i(\xi)$ and $\eta_i(\theta, \Lambda)$ with respect to ξ, and $\tilde{\theta}$ and $\tilde{\Lambda}$ are estimates based on $f_{X|Z}(x|z; \tilde{\xi})$ in the replacement of $\hat{f}_{X|Z}$ in (5) and (6).

The next theorem summarizes the asymptotic properties of the resulting estimator $\hat{\theta}$ based on $\hat{f}_{X|Z}$ in the estimating equation (6).

Theorem 2—*Under regularity conditions (C1) – (C11),* $\hat{\theta}$ is a consistent estimator of θ_0, *and* $n^{1/2}(\hat{\theta} - \theta_0)$ *weakly converges to a distribution with mean 0 and asymptotic covariance* $D(\theta_0)^{-1} + D(\theta_0)^{-1}[H + V_f]D(\theta_0)^{-1}$, *where* V_f *is the variance of* $U_f'(M)$, U_f' *is the Hadamard derivative of* $U(\theta, \Lambda, f_{X|Z})$ at $f_{X|Z}$, *and M is a random variable following a mean-zero normal distribution with covariance matrix* $\{C'_{\xi_0}(F_X, F_Z)\}^2 f_X \int K^2(u)du$.

However, the variance estimator based on the formula given in Theorem 2 requires knowledge of the exact forms of the functional and the kernels, which is not realistic in practice. A more practical alternative approach to variance estimation is a nonparametric bootstrap, which we found to perform well as demonstrated in Section 4.

4 Numerical Results

4.1 Simulation Studies

Extensive simulations were conducted to evaluate the finite sample properties of the proposed estimator in various settings representative of what may occur in practice. We considered the proportional hazards model with two covariates, the error-prone covariate $X \sim N(0, 1)$ and the error-free covariate $W \sim N(0, 1)$. The surrogate Z was generated from the two different settings:

Model A: $Z = X + e$ (i.e., the classical measurement error model),

Model B: $Z = XI(X < -1) - I(X \leq -1) + e$,

where e followed a mean-zero normal distribution with variance $\sigma^2 = 0.5$. The latter simulated a nonlinear relationship between the true exposure and surrogate, mimicking the setting of our motivating data example. The former, a linear measurement error model, was considered to compare the proposed estimator to the regression calibration-based methods that are valid under the classical measurement error model (Xie et al, 2001) or require a good linear approximation for the $E(X|Z, W)$ at least (Spiegelman et al, 1997; Wang et al, 1997) to achieve consistency.

To generate right-censored failure time data, we used an exponential baseline hazard with mean ν and fixed the censoring time to be 1. The constant hazard ν was varying according to the desired censoring rates of 40% and 90%, reflecting the relatively high censoring rates typically found in epidemiologic applications. We considered sample sizes $n = 500$ and $n_\upsilon = 100$ for the common disease setting, and $n = 3000$ and $n_\upsilon = 200$ for the rare disease setting as in our motivating example.

To estimate $f_X(x)$, we considered the Gaussian kernel and Epanechnikov kernel functions with a bandwidth $b=0.9\min(\hat{\sigma}_X, \mathrm{IQR}(X)/1.34)n_\upsilon^{-1/5}$, where $\hat{\sigma_X}$ is the sample standard deviation of X and IQR is the interquartile range (Silverman, 1986, Chap 3.4), satisfying the bandwidth conditions in Theorem 2. The nonparametric bootstrap was used to estimate the variance of $\hat{\beta}_{SCG}$ and $\hat{\beta}_{SCE}$ with 100 replacement resamples.

Table 1 summarizes the main results of estimating the regression coefficient $\beta_0 = 1$, based on 1000 replicates. The naive Cox estimator, $\hat{\beta}_{NC}$, was always biased toward the null. The ordinary regression calibration estimator, $\hat{\beta}_{ORC}$, yielded smaller bias, but had larger variance compared to $\hat{\beta}_{NC}$. The bias in $\hat{\beta}_{ORC}$ became larger when the truncated surrogate model (Model B) was considered (see Table 1). In addition, in the common disease case, $\hat{\beta}_{ORC}$ did not perform as well as in the rare disease case. On the other hand, the proposed semiparametric-copula estimators, $\hat{\beta}_{SCG}$ and $\hat{\beta}_{SCE}$, performed consistently well in all scenarios studied. With both linear and non-linear measurement error, the proposed estimators were virtually unbiased. Although the variances of $\hat{\beta}_{SCG}$ and $\hat{\beta}_{SCE}$ were larger than those for $\hat{\beta}_{ORC}$, by the mean-squared error (MSE) criterion, the semiparametric-copula estimators were among the best in all the scenarios considered. Finally, we note that the coverage probabilities for the semiparametric-copula estimators under both the common and rare diseases scenarios lay in a reasonable range around the nominal 0.95.

4.2 Motivating Example

We applied our method to a study of the effect of physical activity on survival after breast cancer diagnosis in the Nurses' Health Study (NHS) (Holmes et al, 2005). The key features of this study are very similar to the simulation study described in Section 4.1, especially for Model B and the rare disease case shown in Table 1. The NHS is a prospective observational study, following 121,700 female registered nurses since 1976 who were 30–55 years of age at the start of follow-up. Our analysis focused on the 2987 women who were diagnosed with breast cancer in stages I, II or III between 1984 and 1998. These women were followed until death or June 2002, with a median follow-up time of 8 years, and 280 women (9.4%) died from breast cancer during the follow-up period. Physical activity, the primary exposure, was assessed as metabolic equivalent task (MET) hours per week at least 2 years after cancer diagnosis (a median of 3.2 years) to avoid bias due to declining physical activity immediately prior to and after cancer diagnosis.

An external validation study was conducted in the Nurses' Health Study II cohort, where the validity of the self-administered physical activity measure based on questionnaires (i.e., surrogate) was assessed using a detailed activity diary (i.e., the gold standard). The validation data from these 149 women were available to build the measurement error model (Wolf et al, 1994). Because preliminary analyses revealed that there was larger variation in the surrogate than the true measure, we used a functional measurement error model as in Section 4.1. The challenge is three-fold: the assessment of physical activity based on self-reported questionnaires was subject to measurement error, and as shown in Fig. 1 and Fig. 2, neither the true physical activity nor its surrogate measure were normally distributed, and their relationship could not be described by a simple linear additive model. In Fig. 2, restricted cubic splines (Durrleman and Simon, 1989) were used to assess nonlinearity between true and surrogate physical activity measures, and there was a significant evidence of nonlinearity (P=0.042). Hence the classical error model would not be appropriate to correct the measurement error.

Following the procedure described in Section 3.1, we first estimated f_X and f_Z using the Gaussian kernel and Epanechnikov kernel, and then used the AIC for copula model selection. However, a model selection procedure based on AIC alone does not provide any information on the adequacy of the copula chosen. Thus, we further suggest conducting a goodness-of-fit test for the model chosen as the last step of the model selection procedure. For goodness-of-fit tests of copula, Genest et al (2009) systematically reviewed seven test statistics based on the Cramèr-von Mises criterion and the Kolmogorov-Smirnov distance, and compared them in terms of the size and power under numerous scenarios, varying by the form of true copula and the level of dependence. Their extensive simulation study verified that goodness-of-fit tests based on a parametric bootstrap procedure approximated well the null distribution of the various test statistics, and the Cramèr-von Mises statistics yielded more powerful tests than the Kolmogorov-Smirnov statistics. In particular, a test statistic based on the Rosenblatt's transformation, $S_n^{(B)}$ (see equation (9) of Genest et al, 2009) was recommended since it was more powerful across the range of copulas considered. In our NHS II validation data, a good fit of the Clayton copula, the model with the smallest AIC, was obtained for $S_n^{(B)}=0.311$ (P=0.237). This result was consistent with the graphical

inspection of the scatter plot and contour lines in Fig. 3 where a stronger lower tail correlation was observed.

Table 2 shows the estimated effect of physical activity on breast cancer survival using the new semiparametric-copula estimators, compared to the naive Cox approach. The univariate analysis results were from a model with exposure only, and the multivariate analysis results were adjusted for possible other confounders, including age at diagnosis (1 unit change representing 10 year), body mass index (BMI) and cancer stage. In both analyses, increasing average daily physical activity had a significant protective effect on breast cancer survival, and the magnitude of the effect was attenuated (i.e., hazard ratio closer to 1) when the surrogate measure was used without adjusting for the measurement error (see results for $\hat{\beta}_{NC}$). In contrast, the proposed semiparametric-copula approach substantially corrected for the attenuation effect (see results for $\hat{\beta}_{SCG}$ and $\hat{\beta}_{SCE}$). Measurement error in physical activity had minimal impact on the estimated effect of age and BMI because they were not confounders. In some cases where the variables measured with error are more highly correlated with other covariates, however, measurement error will induce bias on the other estimated model coefficients too.

5 Conclusion

We have proposed a semiparametric copula approach for consistent estimation of the effect of an error-prone covariate in the Cox model through the derivation of simple unbiased estimating equations. The semiparametric density estimation procedure for $f_{X|Z}$ that we propose makes use of copula models, allowing for covariate errors of arbitrary structures not restricted to a linear additive measurement error model or Gaussian measurement error as in many previous methods. We provide methods for selection of the form of the copula to be used, thereby decreasing sensitivity of results to a priori arbitrary choice. In addition, we found in our simulation study that when the likelihood values obtained from different copulas were similar, the choice of copula did not have much impact on estimation of regression parameters. Other attractive features of the method include that its applicability to both external and internal validation studies, for which very few existing methods are available, since our method does not require the data $\{T_i, X_i\}$ to be obtained from the same subjects. We note that it may not be the most efficient way of utilizing internal validation data, but the asymptotic properties proven here are still valid. Moreover, compared to the alternative methods for measurement error correction such as regression calibration, this new semiparametric approach performed well even in the common disease setting.

While the method has been restricted to a univariate error-prone covariate, extending the methodology to allow multiple mis-measured exposures is straightforward when the multi-dimensional distribution of exposures is known or can be estimated from validation or reliability data. Future efforts will be devoted to estimating the multi-dimensional conditional distribution of the exposures given the surrogates using approaches for conditional copulas, for example. Based on our investigation, results were robust to the choice of kernel function and its bandwidth, in terms of the mean squared errors of $\hat{\beta}$ For example, regardless of whether Gaussian, Epanechnikov or biweight kernel functions were applied with different bandwidths $b{=}c \times \min(\hat{\sigma}_X, \mathrm{IQR}(X)/1.34)n_v^{-1/5}$, where c = 0.9, 2.34,

or 2.78, the MSEs of $\hat{\beta}$ changed by no more than 0.004. However, it will be worthwhile to investigate whether efforts to reduce the asymptotic mean integrated squared error of the measurement error distribution function itself via bandwidth selection tools such as cross-validation would appreciably improve the overall performance of β estimation.

Finally, although we mainly work on $f_{X|Z}(x|z)$ in the paper, with the identity $f_{Z|X}(z|x) = f_{X|Z}(x|z)f_Z(z)/f_X(x)$, our copula-based approach can also handle $f_{Z|X}(z|x)$ when both X and Z are continuous variables. This way, our formulation would encompass almost all the major measurement error models.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

The authors gratefully acknowledge the support of NIH/NCI grant R01 CA050597 and NIH/NIEHS grant R01 ES009411.

References

Chen YH. Cox regression in cohort studies with validation sampling. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2002; 64(1):51–62.

Dissmann J, Brechmann EC, Czado C, Kurowicka D. Selecting and estimating regular vine copulae and application to financial returns. Computational Statistics & Data Analysis. 2013; 59:52–69.

Durrleman S, Simon R. Flexible regression models with cubic splines. Statistics in Medicine. 1989; 8(5):551–561. [PubMed: 2657958]

Eubank, RL. Spline smoothing and nonparametric regression. New York: Dekker; 1988.

Genest C, Rémillard B, Beaudoin D. Goodness-of-fit tests for copulas: A review and a power study. Insurance: Mathematics and economics. 2009; 44(2):199–213.

Grønneberg S, Hjort NL. The copula information criterion. Scandinavian Journal of Statistics. 2014; 41:436–459.

Holmes MD, Chen WY, Feskanich D, Kroenke CH, Colditz GA. Physical activity and survival after breast cancer diagnosis. Journal of the American Medical Association. 2005; 293(20):2479–2486. [PubMed: 15914748]

Hu P, Tsiatis A, Davidian M. Estimating the parameters in the cox model when covariate variables are measured with error. Biometrics. 1998; 54:1407–1419. [PubMed: 9883541]

Huang Y, Wang C. Cox regression with accurate covariates unascertainable: A nonparametric-correction approach. Journal of the American Statistical Association. 2000; 95(452):1209–1219.

Jordanger LA, Tjøstheim D. Model selection of copulas: Aic versus a cross validation copula information criterion. Statistics & Probability Letters. 2014; 92:249–255.

Manner, H. METEOR Research Memorandum 056. Maastricht University; 2007. Estimation and model selection of copulas with an application to exchange rates.

Nelsen, R. An introduction to copulas. New York: Springer Verlag; 2006.

Prentice RL. Covariate measurement errors and parameter estimation in a failure time regression model. Biometrika. 1982; 69:331–342.

Silverman, BW. Density estimation for statistics and data analysis. London: Chapman & Hall; 1986.

Sklar A. Fonctions de répartition à n dimensions et leurs marges. Publ Inst Statist Univ Paris. 1959; 8:229–231.

Song X, Davidian M, Tsiatis AA. An estimator for the proportional hazards model with multiple longitudinal covariates measured with error. Biostatistics. 2002; 3(4):511–528. [PubMed: 12933595]

Spiegelman D, McDermott A, Rosner B. The regression calibration method for correcting measurement error bias in nutritional epidemiology. American Journal of Clinical Nutrition. 1997; 65:1179S–1186S. [PubMed: 9094918]

Tsiatis AA, Davidian M. A semiparametric estimator for the proportional hazards model with longitudinal covariates measured with error. Biometrika. 2001; 88(2):447–458.

Wand, MP.; Jones, MC. Kernel smoothing. London: Chapman & Hall; 1995.

Wang C. Corrected score estimator for joint modeling of longitudinal and failure time data. Statistica Sinica. 2006; 16(1):235–253.

Wang C, Hsu L, Feng Z, Prentice RL. Regression calibration in failure time regression. Biometrics. 1997; 53(1):131–145. [PubMed: 9147589]

Wen CC. Semiparametric maximum likelihood estimation in cox proportional hazards model with covariate measurement errors. Metrika. 2010; 72(2):199–217.

Wolf AM, Hunter DJ, Colditz GA, Manson JE, Stampfer MJ, Corsano KA, Rosner B, Kriska A, Willett WC. Reproducibility and validity of a self-administered physical activity questionnaire. International Journal of Epidemiology. 1994; 23(5):991–999. [PubMed: 7860180]

Xie SX, Wang C, Prentice RL. A risk set calibration method for failure time regression by using a covariate reliability sample. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2001; 63(4):855–870.

Zhou H, Pepe MS. Auxiliary covariate data in failure time regression. Biometrika. 1995; 82(1):139–149.

Zhou H, Wang CY. Failure time regression with continuous covariates measured with error. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2000; 62(4):657–665.

Zucker D. A pseudo-partial likelihood method for semiparametric survival regression with covariate errors. Journal of the American Statistical Association. 2005; 100(472):1264–1277.

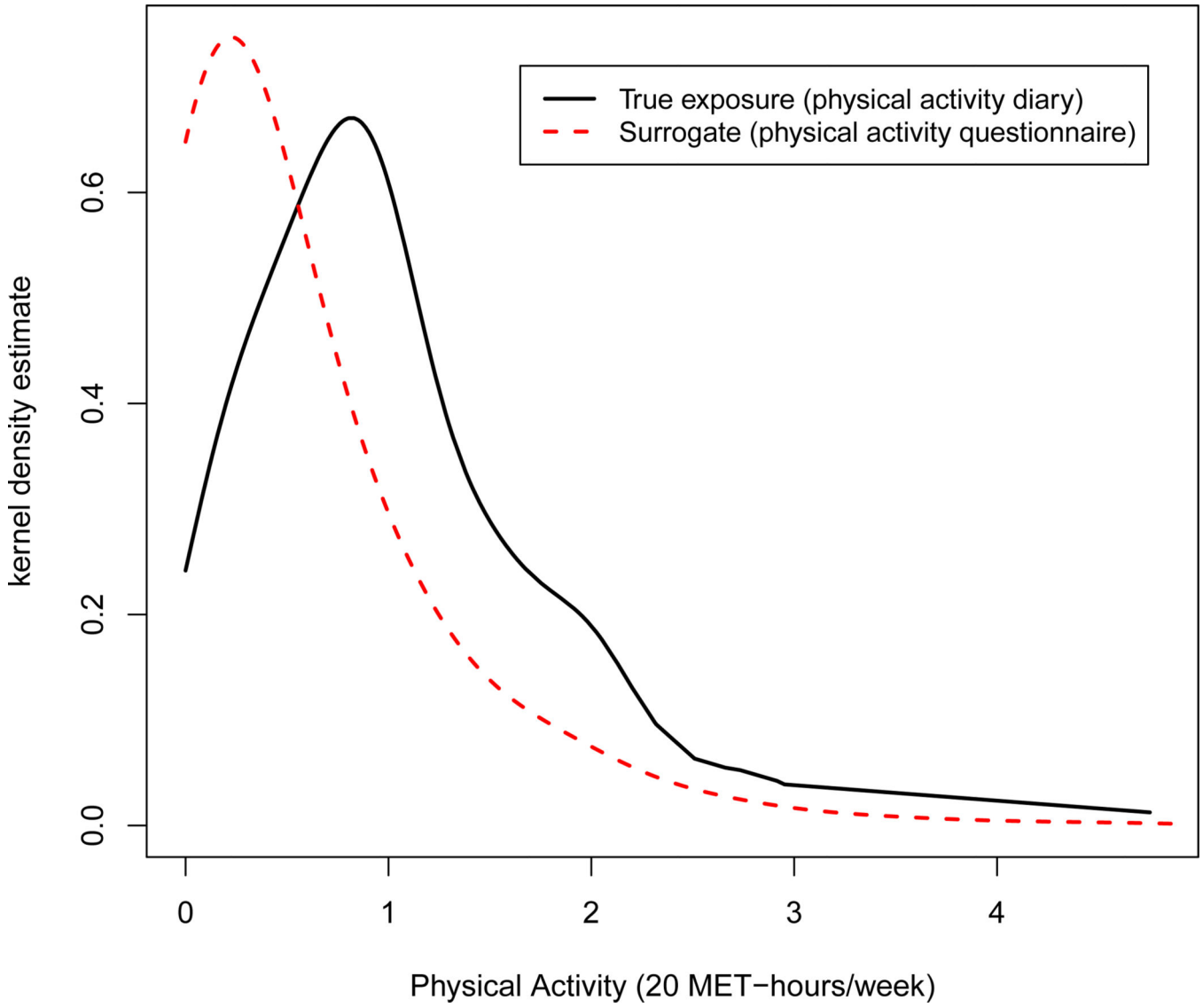

Fig. 1.

Estimated marginal kernel densities of true and surrogate physical activity measurements. Physical activity was the error-prone covariate.

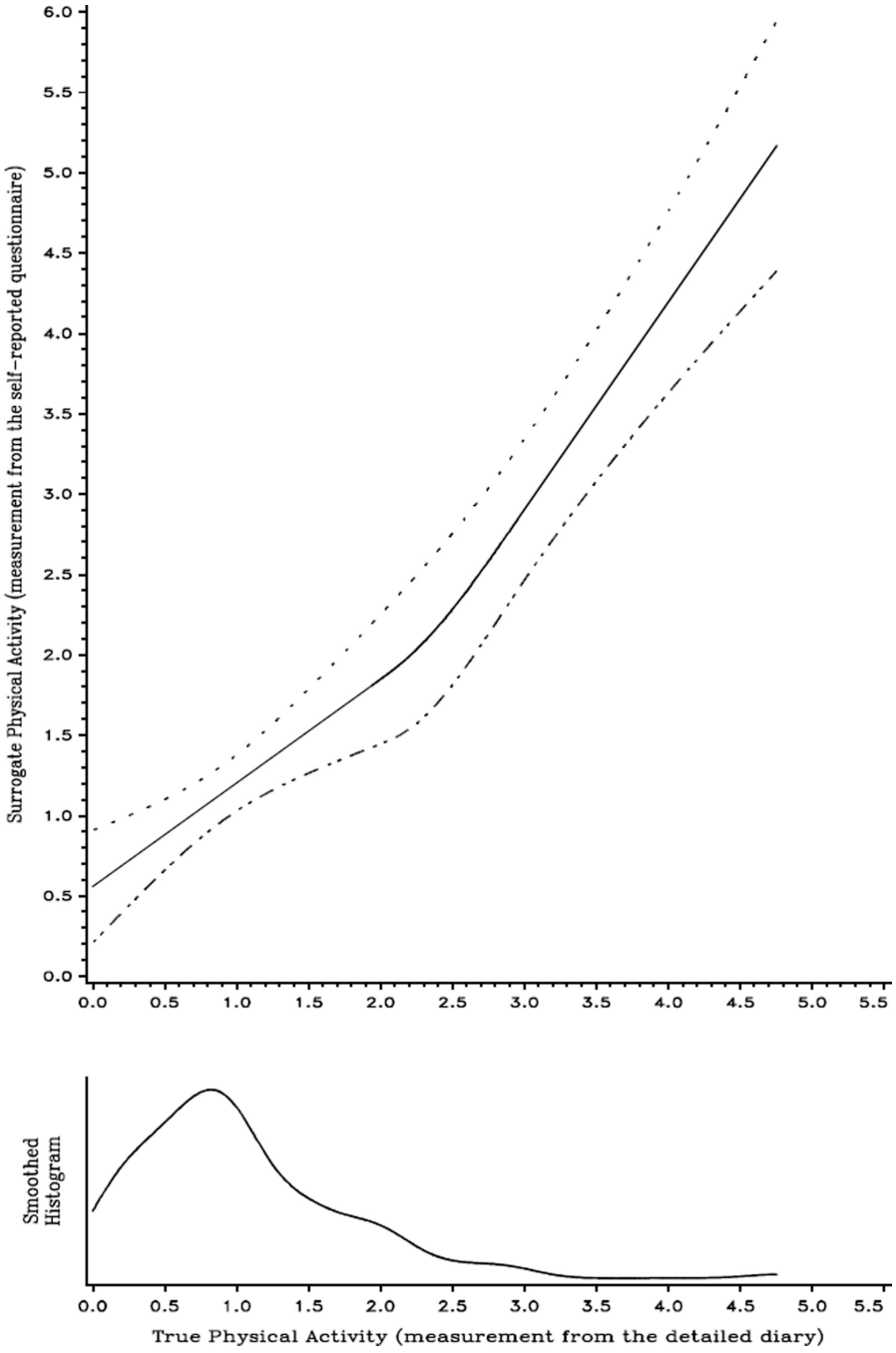

Fig. 2.

Surrogate physical activity in relation to the true measure. One unit equals 20 MET-hours/week.

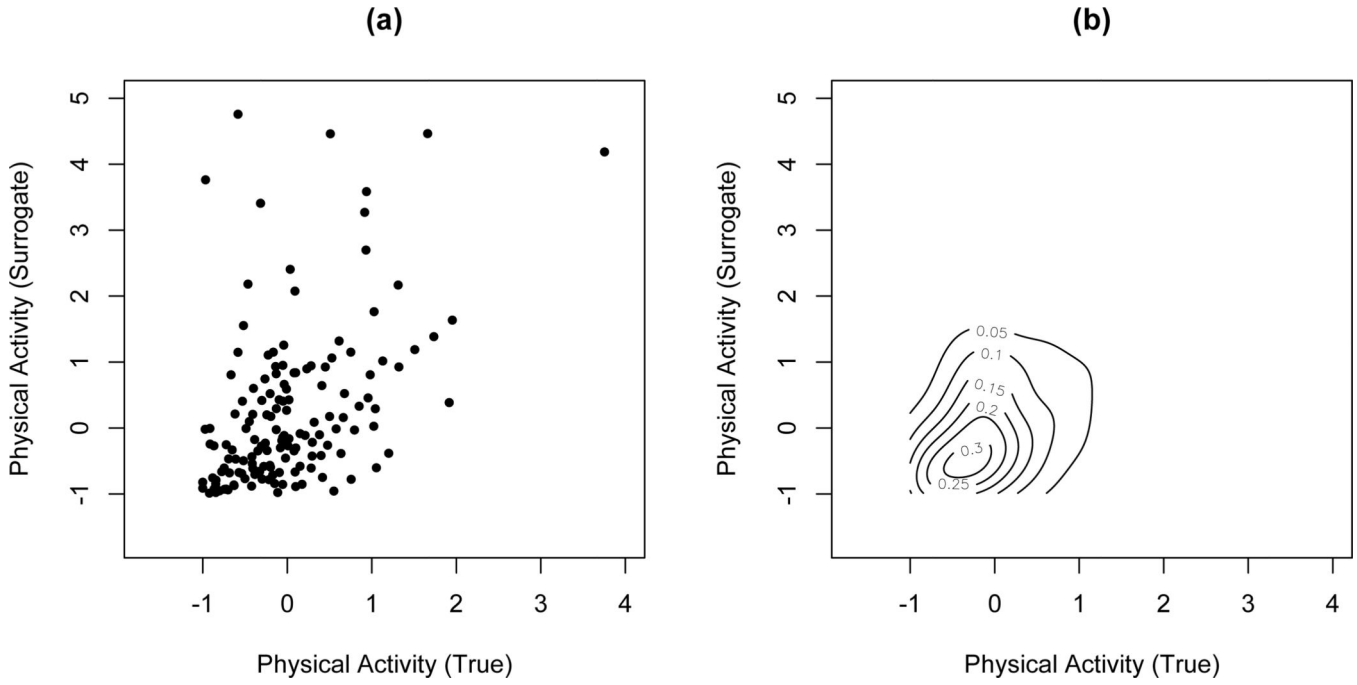

Fig. 3.
(a) Scatter plot and (b) Contour lines of true and surrogate physical activity measurements.

Table 1

Simulation results (β = 1). Common disease setting is with n = 500, n_{υ} = 100, and 60% event rate, while rare disease setting is with n = 3000, n_{υ} = 200, and 10% event rate. $\hat{\beta}_{NC}$ is the naive Cox estimator; $\hat{\beta}_{ORC}$ is the ordinary regression calibration estimator; $\hat{\beta}_{SCG}$ and $\hat{\beta}_{SCE}$ are the proposed semiparametric-copula estimators, using the Gaussian and Epanechnikov kernel smoothings, respectively; SSD is the sample standard deviation; SEE is the standard error estimate; CP is the coverage probability of the 95% confidence interval; MSE is the mean squared error.

		Model A		Model B	
		Common	Rare	Common	Rare
$\hat{\beta}_{NC}$	Bias	−0.412	−0.357	−0.351	−0.318
	SSD	0.054	0.048	0.059	0.048
	SEE	0.054	0.047	0.059	0.047
	CP (%)	0.0	0.0	0.1	0.0
	MSE×10^2	17.3	13.0	12.7	10.4
$\hat{\beta}_{ORC}$	Bias	−0.114	−0.032	−0.158	−0.115
	SSD	0.104	0.085	0.105	0.080
	SEE	0.105	0.086	0.105	0.081
	CP (%)	75.7	91.5	60.9	67.0
	MSE×10^2	2.4	0.8	3.6	2.0
$\hat{\beta}_{SCG}$	Bias	−0.024	−0.024	0.001	−0.019
	SSD	0.152	0.102	0.159	0.106
	SEE	0.160	0.109	0.158	0.101
	CP (%)	93.2	93.0	93.4	92.0
	MSE×10^2	2.4	1.1	2.5	1.2
$\hat{\beta}_{SCE}$	Bias	−0.026	−0.016	0.020	−0.003
	SSD	0.173	0.112	0.170	0.116
	SEE	0.201	0.131	0.166	0.104
	CP (%)	94.7	95.2	93.3	92.1
	MSE×10^2	3.0	1.3	2.9	1.3

Table 2

Analysis results for the study of physical activity in relation to breast cancer mortality in the Nurses' Health Study. Physical activity is an error-prone covariate with 1 unit change representing 20 MET-hours/week, while the other covariates are error-free. HR = Hazard Ratio; P = p-value.

	Univariate Analysis		Multivariate Analysis	
Effect	**HR (95% CI)**	**P**	**HR (95% CI)**	**P**
Naive Cox Estimator ($\hat{\beta}_{NC}$)				
Physical activity	0.788 (0.645, 0.963)	.019	0.794 (0.651, 0.970)	.024
Age at diagnosis			1.107 (0.946 1.295)	.210
Overweight (BMI≥25)			1.043 (0.819, 1.327)	.740
Cancer stage II or III			3.815 (2.934, 4.961)	<.001
Semiparametric-copula estimator using the Gaussian kernel ($\hat{\beta}_{SCG}$)				
Physical activity	0.445 (0.201, 0.986)	.023	0.453 (0.201, 1.020)	.028
Age at diagnosis			1.099 (0.933, 1.295)	.129
Overweight (BMI≥25)			1.023 (0.799, 1.310)	.429
Cancer stage II or III			3.846 (2.949, 5.015)	<.001
Semiparametric-copula estimator using the Epanechnikov kernel ($\hat{\beta}_{SCE}$)				
Physical activity	0.433 (0.229, 0.818)	.005	0.448 (0.236, 0.849)	.007
Age at diagnosis			1.102 (0.934, 1.301)	.124
Overweight (BMI≥25)			1.023 (0.799, 1.312)	.427
Cancer stage II or III			3.844 (2.953. 5.003)	<.001