Preface

The material of this volume was inspired by papers presented at BIOSTAT2006, an international conference organized by the University of Cyprus and the European Seminar—"Mathematical Methods in Survival Analysis, Reliability and Quality of Life." The conference was a part of a series of conferences, workshops, and seminars organized or co-organized by the European Seminar over the years. BIOSTAT2006 took place in Limassol, Cyprus between May 29 to 31, 2006 with great success. It attracted over 100 participants from 30 countries. The aim of this event was to bring together scientists from all over the world that work in statistics in general and advance knowledge in fields related to biomedical and technical systems. The publication of this volume comes at a very special time because this year we are celebrating the tenth anniversary of the inauguration of the European Seminar.

The volume consists of selected papers presented at BIOSTAT2006 but it also includes other invited papers. The included papers nicely blend current concerns and research interests in survival analysis and reliability. There is a total of 37 papers which for the convenience of the readers are divided into the following nine parts.

- Cox Models, Analyses, and Extensions
- Reliability Theory - Degradation Models
- Inferential Analysis
- Analysis of Censored Data
- Quality of Life
- Inference for Processes
- Designs
- Measures of Divergence, Model Selection, and Survival Models
- New Statistical Challenges

The editors would like to thank all the authors for contributing their work to this book as well as all the anonymous referees for an excellent job in reviewing the papers and making their presentation the best possible. We would also like to thank Professor Alex Karagrigoriou whose help was invaluable during the organization of the conference as well as the preparation of this volume.

Thanks are also due to Professor N. Balakrishnan for his constant support and guidance, to Mr. Thomas Grasso for his assistance in the production of this volume and to Mrs. Debbie Iscoe for a highly professional job in typesetting this volume in a camera-ready form. Special thanks are due to the Department of Mathematics and Statistics of the University of Cyprus which financially supported the publication of this volume.

Nicosia, Cyprus **F. Vonta**

Bordeaux, France **M. Nikulin**

Compiègne, France **N. Limnios**

Paris, France **C. Huber-Carol**

2

Corrected Score Estimation in the Cox Regression Model with Misclassified Discrete Covariates

David M. Zucker[1] **and Donna Spiegelman**[2]

[1] *Department of Statistics, Hebrew University, Jerusalem, Israel*
[2] *Departments of Epidemiology and Biostatistics, Harvard School of Public Health, Boston, MA, USA*

Summary: We consider Cox proportional hazards regression when the covariate vector includes error-prone discrete covariates along with error-free covariates that may be discrete or continuous. The misclassification in the discrete error-prone covariates is allowed to be of arbitrary form. Building on work of Nakamura and his colleagues, we develop a corrected score method for this setting. The method can handle all three major study designs (internal validation design, external validation design, and replicate measures design), both functional and structural error models, and time-dependent covariates satisfying a certain "localized error" condition. This chapter presents the method, briefly describes its asymptotic properties, and illustrates it on data from a study of the relationship between dietary calcium intake and distal colon cancer. Zucker and Spiegelman (2007, 2008) present further details on the asymptotic theory and a simulation study under Weibull survival with a single binary covariate having known misclassification rates. In these simulations, the method presented here performed similarly to related methods we have examined in previous work. Specifically, our new estimator performed as well as or, in a few cases, better than the full Weibull maximum likelihood estimator. In further simulations for the case where the misclassification probabilities are estimated from an external replicate measures study our method generally performed well. The new estimator has a broader range of applicability than many other estimators proposed in the literature, including those described in our own earlier work, in that it can handle time-dependent covariates with an arbitrary misclassification structure.

Keywords and Phrases: Errors in variables, nonlinear models, proportional hazards

2.1 Introduction

Many regression analyses involve explanatory variables that are measured with error. It is well known that failing to account for covariate error can lead to biased estimates of the regression coefficients. For linear models, theory for handling covariate error has been developed over the past 50 or more years; Fuller (1987) provides an authoritative exposition. For nonlinear models, theory has been developing over the past 25 or so years. Carroll *et al.* (2006) provide a comprehensive summary of the development to date; currently, the covariate error problem for nonlinear models remains an active research area. In particular, beginning with Prentice (1982), a growing literature has developed on the Cox (1972) proportional hazards survival regression model when some covariates are measured with error. In this chapter, we focus on discrete covariates subject to misclassification, which are of interest in many epidemiological studies.

Three basic design setups are of interest. In all three designs, we have a main survival cohort for which surrogate covariate measurements and survival time data are available on all individuals. The designs are as follows: (1) the internal validation design, where the true covariate values are available on a subset of the main survival cohort; (2) the external validation design, where the measurement error distribution is estimated from data outside the main survival study; and (3) the replicate measurements design, where replicate surrogate covariate measurements are available, either on a subset of the survival study cohort or on individuals outside the main survival study. Also, two types of models for the measurement error are of interest [see Fuller (1987, p. 2) and Carroll *et al.* (2006, Section 1.2)]: structural models, where the true covariates are random variables, and functional models, where the true covariates are fixed values. Structural model methods generally involve estimation of some aspect of the distribution of the true covariate values; in functional model methods, this process is avoided.

The Cox model with covariate error has been examined in various settings. Zucker and Spiegelman (2007, 2008) give a detailed review of the existing work. Much of this work focuses on the independent additive error model, under which the observed covariate value is equal to the true value plus a random error whose distribution is independent of the true value. For discrete covariates subject to misclassification, this model practically never holds, and so the methods built upon it do not apply. Other methods exist, but are subject to various limitations. There is a need for a convenient method for all three study designs that can handle general measurement error structures, both functional and structural models, and time-dependent covariates. The aim of our work is to provide such a method for the case where the error-prone covariates are discrete, with misclassification of arbitrary form. Our method builds on a corrected score

approach developed by Akazawa *et al.* (1998) for generalized linear models. We begin by reviewing their work, and we then present our extension to the Cox model.

2.2 Review of the Corrected Score Technique

We work with a sample of n independent individuals. Associated with each individual i is a response variable T_i and a p-vector of covariates $\mathbf{X}_i$. The conditional density or mass function of T_i given $\mathbf{X}_i$ is denoted by $f(t|\mathbf{X}_i, \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a q-vector of unknown parameters, which includes regression coefficients and auxiliary parameters such as error variances. We have in mind mainly generalized linear models such as linear, logistic, and Poisson regression, but we present the theory in a general way. We denote the true value of $\boldsymbol{\theta}$ by $\boldsymbol{\theta}_0$. Extending Akazawa *et al.* (1998), we partition the vector $\mathbf{X}_i$ into $\mathbf{W}_i$ and $\mathbf{Z}_i$, where $\mathbf{W}_i$ is a p_1-vector of error-prone covariates and $\mathbf{Z}_i$ is a p_2-vector of error-free covariates. We denote the observed value of $\mathbf{W}_i$ by $\tilde{\mathbf{W}}_i$. The vector $\mathbf{W}_i$ is assumed to be discrete, with its possible values (each one a p_1-vector) denoted by $\mathbf{w}_1, \ldots, \mathbf{w}_K$. The range of values of $\tilde{\mathbf{W}}_i$ is assumed to be the same as that for $\mathbf{W}_i$. We denote by $k(i)$ the value of k such that $\tilde{\mathbf{W}}_i = \mathbf{w}_k$. The vector $\mathbf{Z}_i$ of error-free covariates is allowed to be either discrete or continuous. We denote $A_{kl}^{(i)} = \Pr(\tilde{\mathbf{W}}_i = \mathbf{w}_l | \mathbf{W}_i = \mathbf{w}_k, \mathbf{Z}_i, T_i)$, which defines a square matrix $\mathbf{A}^{(i)}$ of classification probabilities. As the notation indicates, we allow the classification probabilities to depend on $\mathbf{Z}_i$ and T_i (e.g., through a suitable model). This feature can be useful in certain applications; in others, it is sensible to assume that the same classification probabilities apply to all individuals. We assume for now that $\mathbf{A}^{(i)}$ is known. We denote by $\mathbf{B}^{(i)}$ the matrix inverse of $\mathbf{A}^{(i)}$. We assume this inverse exists, which will be the case if the misclassification is not too extreme [cf. Zucker and Spiegelman, (2004, Appendix A.1)]. When individual i is a member of an internal validation sample, for the estimation of $\boldsymbol{\theta}$ we set $\tilde{\mathbf{W}}_i = \mathbf{W}_i$ and replace $\mathbf{A}^{(i)}$ by the identity matrix.

Define $\mathbf{u}(t, \mathbf{w}, \mathbf{z}, \boldsymbol{\theta}) = [\partial/\partial\boldsymbol{\theta}] \log f(t|\mathbf{w}, \mathbf{z}, \boldsymbol{\theta})$ and $\mathbf{u}_i(\boldsymbol{\theta}) = \mathbf{u}(T_i, \mathbf{W}_i, \mathbf{Z}_i, \boldsymbol{\theta})$. The classical normalized likelihood score function when there is no covariate error is then given by $\mathbf{U}(\boldsymbol{\theta}) = n^{-1} \sum_i \mathbf{u}_i(\boldsymbol{\theta})$, and the maximum likelihood estimate (MLE) is obtained by solving the equation $\mathbf{U}(\boldsymbol{\theta}) = \mathbf{0}$. Under classical conditions, $E_{\boldsymbol{\theta}_0}[\mathbf{U}(\boldsymbol{\theta}_0)] = \mathbf{0}$ and the MLE is consistent and asymptotically normal. The idea of the corrected score approach is to find a function $\mathbf{u}^*(t, \tilde{\mathbf{w}}, \mathbf{z}, \boldsymbol{\theta})$ such that

$$E[\mathbf{u}^*(T_i, \tilde{\mathbf{W}}_i, \mathbf{Z}_i, \boldsymbol{\theta}) | \mathbf{W}_i, \mathbf{Z}_i, T_i] = \mathbf{u}(T_i, \mathbf{W}_i, \mathbf{Z}_i, \boldsymbol{\theta}). \tag{2.1}$$

Then, with $\mathbf{u}_i^*(\boldsymbol{\theta}) = \mathbf{u}^*(T_i, \tilde{\mathbf{W}}_i, \mathbf{Z}_i, \boldsymbol{\theta})$, we use the modified likelihood score function $\mathbf{U}^*(\boldsymbol{\theta}) = n^{-1} \sum_i \mathbf{u}_i^*(\boldsymbol{\theta})$ in place of $\mathbf{U}(\boldsymbol{\theta})$ as the basis for estimation.

The estimation equation thus becomes $\mathbf{U}^*(\boldsymbol{\theta}) = \mathbf{0}$. In the case of discrete error-prone covariates, as shown by Akazawa *et al.* (1998), a function $\mathbf{u}^*$ satisfying (2.1) is given by a simple formula:

$$\mathbf{u}_i^*(\boldsymbol{\theta}) = \sum_{l=1}^{K} B_{k(i)l}^{(i)} \mathbf{u}(T_i, \mathbf{w}_l, \mathbf{Z}_i, \boldsymbol{\theta}). \tag{2.2}$$

Let $\mathbf{J}_i(\boldsymbol{\theta})$ be the matrix with elements $J_{i,rs}(\boldsymbol{\theta}) = (\partial/\partial\theta_s)u_{i,r}(\theta)$ and let $\mathbf{J}_i^*(\boldsymbol{\theta})$ be defined correspondingly with $\mathbf{u}_i^*$ in place of $\mathbf{u}_i$.

Under the typical conditions assumed in generalized estimation equations (GEE) theory, the estimator $\hat{\boldsymbol{\theta}}$ will be consistent and asymptotically normal. The limiting covariance matrix $\mathbf{V}$ of $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ can be estimated using the sandwich estimator $\hat{\mathbf{V}} = \mathbf{D}(\hat{\boldsymbol{\theta}})^{-1}\mathbf{H}(\hat{\boldsymbol{\theta}})\mathbf{D}(\hat{\boldsymbol{\theta}})^{-1}$, where $\mathbf{H}(\boldsymbol{\theta}) = n^{-1}\sum_i \mathbf{u}_i^*(\boldsymbol{\theta})$ $\mathbf{u}_i^*(\boldsymbol{\theta})^T$ and $\mathbf{D}(\boldsymbol{\theta}) = -n^{-1}\sum_i \mathbf{J}_i^*(\boldsymbol{\theta})$.

The case where there are replicate measurements $\tilde{\mathbf{W}}_{ij}$ of $\tilde{\mathbf{W}}$ on the individuals in the main study can be handled in various ways. A simple approach is to redefine the quantity $\mathbf{u}_i^*(\boldsymbol{\theta})$ given in (2.2) by replacing $B_{k(i)l}^{(i)}$ with the mean of $B_{k(i,j)l}^{(i)}$ over the replicates for individual i, with $k(i,j)$ defined as the value of k such that $\tilde{\mathbf{W}}_{ij} = \mathbf{w}_k$. The development then proceeds as before.

2.3 Application to the Cox Survival Model

2.3.1 Setup

We now show how to apply the foregoing corrected score approach to the Cox model. Denote the survival time by T_i° and the censoring time by C_i. The observed survival data then consist of the observed follow-up time $T_i = \min(T_i^\circ, C_i)$ and the event indicator $\delta_i = I(T_i^\circ \leq C_i)$. We let $Y_i(t) = I(T_i \geq t)$ denote the at-risk indicator. We assume the failure process and the censoring process are conditionally independent given the covariate process in the sense described by Kalbfleisch and Prentice (2002, Sections 6.2 and 6.3).

The covariate structure is as described in the preceding section, except that the covariates are allowed to be time-dependent, so that we write $k(i,t)$ and $\mathbf{Z}_i(t)$. We assume that the measurement error process is "localized" in the sense that it depends only on the current true covariate value. More precisely, the assumption is that, conditional on the value of $\mathbf{X}_i(t)$, the value of $\tilde{\mathbf{W}}_i(t)$ is independent of the survival and censoring processes and of the values of $\mathbf{X}_i(s)$ for $s \neq t$. This assumption is plausible in many settings, for example, when the main source of error is technical or laboratory error, or reading/coding error, as with diagnostic X-rays and dietary intake assessments. With no change in

the theory, the classification probabilities $A_{kl}^{(i)}$ can be allowed to depend upon t. This extension permits accounting for improvements in measurement techniques over time. In addition, if internal validation data are available, this extension allows us to dispense with the localized error assumption.

In the proportional hazards model, the hazard function is taken to be of the form $\lambda(t|\mathbf{X}(t)) = \lambda_0(t)\psi(\mathbf{X}(t);\boldsymbol{\beta})$, with $\lambda_0(t)$ being a baseline hazard function of unspecified form. The function $\psi(\mathbf{x};\boldsymbol{\beta})$, which involves a p-vector $\boldsymbol{\beta}$ of unknown regression parameters which are to be estimated, represents the relative risk for an individual with covariate vector $\mathbf{x}$. The classical Cox model assumes $\psi(\mathbf{x};\boldsymbol{\beta}) = e^{\boldsymbol{\beta}^T\mathbf{x}}$. In line with Thomas (1981) and Breslow and Day (1993, Section 5.1(c)), we allow a general relative risk function $\psi(\mathbf{x};\boldsymbol{\beta})$ which is assumed to be positive in a neighborhood of the true $\boldsymbol{\beta}$ for all $\mathbf{x}$ and to be twice differentiable with respect to the components of $\boldsymbol{\beta}$. We assume further that $\psi(\mathbf{x};\mathbf{0}) = 1$, which simply means that $\boldsymbol{\beta} = \mathbf{0}$ corresponds to no covariate effect. In many applications, it will be desirable to take $\psi(\mathbf{x};\boldsymbol{\beta})$ to be a function that is monotone in each component of $\mathbf{x}$ for all $\boldsymbol{\beta}$. We let $\boldsymbol{\beta}_0$ denote the true value of $\boldsymbol{\beta}$.

2.3.2 The method

We now describe the method. Let $\psi_r'(\mathbf{x};\boldsymbol{\beta})$ denote the partial derivative of $\psi(\mathbf{x};\boldsymbol{\beta})$ with respect to β_r and define $\xi_r(\mathbf{x};\boldsymbol{\beta}) = \psi_r'(\mathbf{x};\boldsymbol{\beta})/\psi(\mathbf{x};\boldsymbol{\beta})$. Then the classical Cox partial likelihood score function in the case with no measurement error is given by

$$U_r(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\delta_i\left(\xi_r(\mathbf{X}_i(T_i);\boldsymbol{\beta}) - \frac{e_{1r}(T_i)}{e_0(T_i)}\right), \tag{2.3}$$

where

$$e_0(t) = \frac{1}{n}\sum_{j=1}^{n}Y_j(t)\psi(\mathbf{X}_j(t);\boldsymbol{\beta}),\ \ e_{1r}(t) = \frac{1}{n}\sum_{j=1}^{n}Y_j(t)\psi_r'(\mathbf{X}_j(t);\boldsymbol{\beta}).$$

Now define

$$\psi_i^*(t,\boldsymbol{\beta}) = \sum_{l=1}^{K}B_{k(i,t)l}^{(i)}\psi(\mathbf{w}_l,\mathbf{Z}_i(t);\boldsymbol{\beta}),\ \ \eta_{ir}(t,\boldsymbol{\beta}) = \sum_{l=1}^{K}B_{k(i,t)l}^{(i)}\psi_r'(\mathbf{w}_l,\mathbf{Z}_i(t);\boldsymbol{\beta}),$$

$$\xi_{ir}^*(t,\boldsymbol{\beta}) = \sum_{l=1}^{K}B_{k(i,t)l}^{(i)}\xi_r(\mathbf{w}_l,\mathbf{Z}_i(t);\boldsymbol{\beta}),\ \ e_0^*(t) = \frac{1}{n}\sum_{j=1}^{n}Y_j(t)\psi_j^*(t,\boldsymbol{\beta}),$$

$$e_{1r}^*(t) = \frac{1}{n}\sum_{j=1}^{n}Y_j(t)\eta_{jr}(t,\boldsymbol{\beta}).$$

Then our proposed corrected score function is the following obvious analogue of (2.3):

$$U_r^*(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} \delta_i \left(\xi_{ir}^*(T_i, \boldsymbol{\beta}) - \frac{e_{1r}^*(T_i)}{e_0^*(T_i)} \right). \tag{2.4}$$

As before, the proposed corrected score estimator is the solution to $\mathbf{U}^*(\boldsymbol{\beta}) = \mathbf{0}$, where $\mathbf{U}^*$ denotes the vector whose components are U_r^*.

Using an iterated expectation argument, under the localized error assumption, we can show that

$$E[Y_i(t)\psi_i^*(t,\boldsymbol{\beta})|\mathbf{X}_i(t)] = E[Y_i(t)\psi(\mathbf{X}_i(t);\boldsymbol{\beta})|\mathbf{X}_i(t)], \tag{2.5}$$

$$E[Y_i(t)\eta_{ir}^*(t,\boldsymbol{\beta})|\mathbf{X}_i(t)] = E[Y_i(t)\psi_r'(\mathbf{X}_i(t),\boldsymbol{\beta})|\mathbf{X}_i(t)], \tag{2.6}$$

$$E[Y_i(t)\xi_{ir}^*(t,\boldsymbol{\beta})|\mathbf{X}_i(t)] = E[Y_i(t)\xi_r(\mathbf{X}_i(t),\boldsymbol{\beta})|\mathbf{X}_i(t)]. \tag{2.7}$$

Thus, referring to the quantity in parentheses in (2.4), the first term and the numerator and denominator of the second term all have the correct expectation. It follows that $\mathbf{U}^*(\boldsymbol{\beta})$ is an asymptotically unbiased score function.

Accordingly, under standard conditions such as those of Andersen and Gill (1982) and of Prentice and Self (1983), our corrected score estimator will be consistent and asymptotically normal. The asymptotic covariance matrix of $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ may be estimated by the sandwich formula $\hat{\mathbf{V}} = \mathbf{D}(\hat{\boldsymbol{\beta}})^{-1}\mathbf{H}(\hat{\boldsymbol{\beta}})\mathbf{D}(\hat{\boldsymbol{\beta}})^{-1}$. Here $\mathbf{D}(\boldsymbol{\beta})$ is -1 times the matrix of derivatives of $\mathbf{U}^*(\boldsymbol{\beta})$ with respect to the components of $\boldsymbol{\beta}$ and $\mathbf{H}(\boldsymbol{\beta})$ is an empirical estimate of the covariance matrix of $\sqrt{n}\,\mathbf{U}^*(\boldsymbol{\beta})$.

We note again that, for the internal validation design, the available true $\mathbf{W}$ values can be used in the estimation of $\boldsymbol{\beta}$ by replacing $\tilde{\mathbf{W}}_i$ with $\mathbf{W}_i$ and $\mathbf{A}^{(i)}$ by the identity matrix when individual i is in the internal validation sample. Alternatively, the hybrid scheme of Zucker and Spiegelman (2004, Section 5) can be used. Also, the case where there are replicate measurements $\tilde{\mathbf{W}}_{ij}$ of $\tilde{\mathbf{W}}$ on the individuals in the main study can be handled as described at the end of the preceding section.

In Zucker and Spiegelman (2007, 2008) we give an outline of the asymptotic argument, explicit expressions for the matrices $\mathbf{H}$ and $\mathbf{D}$, an estimator of the cumulative hazard function, and an extension of the theory to the case where the classification matrix $\mathbf{A}^{(i)}$ is estimated. We also give results of a finite-sample simulation study under Weibull survival with a single binary covariate having known misclassification rates. The performance of the method described here was similar to that of related methods we have examined in previous work [Zucker and Spielgelman (2004) and Zucker (2005)]. Specifically, our new estimator performed as well as or, in a few cases, better than the full Weibull maximum likelihood estimator. We also present simulation results for our method for the case where the misclassification probabilities are estimated from an external replicate measures study. Our method generally performed well in these simulations.

2.4 Example

We illustrate our method on data from the Nurses Health Study concerning the relationship between dietary calcium (Ca) intake and incidence of distal colon cancer [Wu *et al.* (2002, Table 4)]. The data consist of observations on female nurses whose calcium intake was assessed through a food frequency questionnaire (FFQ) in 1984 and were followed up to May 31, 1996 for distal colon cancer occurrence. Our analysis includes data on 60,575 nurses who reported in 1984 that they had never taken calcium supplements, and focuses on the effect of baseline calcium intake after adjustment for baseline body mass index (BMI) and baseline aspirin use. In line with Wu *et al.*'s analysis, we use the classical Cox relative risk function $\psi(\boldsymbol{\beta}; \mathbf{x}) = e^{\boldsymbol{\beta}^T \mathbf{x}}$, and, as in Wu *et al.*'s Table 4, we work with a binary "high Ca" risk factor defined as 1 if the calcium intake was greater than 700 mg/day and 0 otherwise. Note that one glass of milk contains approximately 300 mg of calcium. BMI is expressed in terms of the following categories: <22 kg/m^2, 22 to <25 kg/m^2, 25 to <30 kg/m^2, and 30 kg/m^2 or greater. Aspirin use is coded as yes (1) or no (0). Thus, our model has five explanatory variables, one for the binary risk factor (W), three dummy variables for BMI (Z_1, Z_2, Z_3), and one for aspirin use (Z_4). BMI and aspirin use status are assumed to be measured without error.

It is well known that the FFQ measures dietary intake with some degree of error and more reliable information can be obtained from a diet record (DR) [Willett (1998, Chapter 6)]. We thus take W to be the Ca risk factor indicator based on the DR and $\tilde{W}$ to be the Ca risk factor indicator based on the FFQ. The classification probabilities are estimated using data from the Nurse's Health Study validation study [Willett (1998, pp. 122–126)]. The estimates obtained were $\Pr(\tilde{W} = 0|W = 0) = 0.78$ and $\Pr(\tilde{W} = 1|W = 1) = 0.72$, with corresponding estimated standard errors of 0.042 and 0.046.

Table 2.1 presents the results of the following analyses: (1) a naive classical Cox regression analysis ignoring measurement error, corresponding to an assumption that there is no measurement error; (2) our method with $\mathbf{A}$ assumed known and set according to the foregoing estimated classification probabilities, ignoring the estimation error in these probabilities; and (3) our method with $\mathbf{A}$ estimated as above with the estimation error in the probabilities taken into account (main study/external validation study design).

The results followed the expected pattern. Adjusting for the misclassification in calcium intake had a marked effect on the estimated relative risk for high calcium intake. Accounting for the error in estimating the classification probabilities increased (modestly) the standard error of the log relative risk estimate. The relative risk estimates for high calcium intake and corresponding 95% confidence intervals obtained in the three analyses were as follows.

Table 2.1. Estimated coefficients and standard errors for the Nurses Health Study of the relationship between dietary calcium intake and distal colon cancer incidence

	High Calcium		BMI of 22 to <25		BMI of 25 to <30		BMI of 30+		Aspirin Use	
Method	Estimate	Std Err	Estimate	Std Err	Estimate	Std Err	Estimate	Std Err	Estimate	Std Err
Cox	−0.3448	0.1694	0.6837	0.2240	0.5352	0.2395	0.5729	0.2876	−0.4941	0.1954
CS0	−0.7121	0.3690	0.7124	0.2247	0.5776	0.2419	0.6157	0.2892	−0.4994	0.1955
CS1	−0.7121	0.3832	0.7124	0.2249	0.5776	0.2423	0.6157	0.2896	−0.4994	0.1955

Cox = Classical Cox regression analysis.
CS0 = Corrected score method, observed classification matrix taken as known.
CS1 = Corrected score method, accounting for uncertainty in the classification matrix.

Method	Estimate	95% CI
Naive Cox	0.71	[0.51,0.99]
A known	0.49	[0.24,1.01]
A estimated	0.49	[0.23,1.04]

The misclassification adjustment had a small effect on the estimated regression coefficients for the BMI dummy variables and essentially no effect on the estimated regression coefficient for aspirin use.

References

1. Akazawa, K., Kinukawa, N., and Nakamura, T. (1998). A note on the corrected score function corrected for misclassification, *Journal of the Japan Statistical Society,* **28**, 115–123.

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

3. Breslow, N. and Day, N. E. (1993). *Statistical Methods in Cancer Research,* Volume 2: *The Design and Analysis of Cohort Studies,* Oxford University Press, Oxford.

4. Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective*, 2nd ed. Chapman and Hall/CRC, Boca Raton.

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

6. Fuller, W. A. (1987). *Measurement Error Models*, John Wiley & Sons, New York.

7. Kalbfleisch, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data,* 2nd ed. John Wiley & Sons, New York.

8. Prentice, R. (1982). Covariate measurement errors and parameter estimation in a failure time regression model, *Biometrika,* **69**, 331–342.

9. Prentice, R. L. and Self, S. G. (1983). Asymptotic distribution theory for Cox-type regression models with general relative risk form, *The Annals of Statistics,* **11**, 804–812.

10. Thomas, D. C. (1981). General relative-risk models for survival time and matched case-control analysis, *Biometrics,* **37**, 673–686.

11. Willett, W. C. (1998). *Nutritional Epidemiology*, 2nd ed., Oxford University Press, New York.

12. Wu, K., Willett, W. C., Fuchs, C. S., Colditz, G. A., and Giovannucci, E. L. (2002). Calcium intake and risk of colon cancer in women and men, *Journal of the National Cancer Institute,* **94**, 437–446.

13. Zucker, D. M. (2005). A pseudo partial likelihood method for semiparametric survival regression with covariate errors, *Journal of the American Statistical Association,* **100**, 1264–1277.

14. Zucker, D. M. and Spiegelman, D. (2004). Inference for the proportional hazards model with misclassified discrete-valued covariates, *Biometrics,* **60**, 324–334.

15. Zucker, D. M. and Spiegelman, D. (2007). Corrected score estimation in the proportional hazards model with misclassified discrete covariates. Technical Report, Hebrew University. Available online at http://pluto.mscc.huji.ac.il/∼mszucker.

16. Zucker, D. M. and Spiegelman, D. (2008). Corrected score estimation in the proportional hazards model with misclassified discrete covariates, *Statistics in Medicine*, in press.