Correcting for sampling variation and measurement error in Cox regression with circulating tumour cells

Cecilia A. Cotton^{*}, Ying Wu[†] and Richard J. Cook

^{*}Correspondence: ccotton@math.uwaterloo.ca

CrossMark

← Click for updates

[†]These authors contributed equally to this work.

Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada.

Abstract

Background: Circulating tumour cell counts are being used increasingly often in oncology for prognosis, stratification and the assessment of response to treatment in phase II trials. Cell counts have typically been based on visual or computer assisted counting of abnormal cells in blood samples drawn at the point of recruitment to a study. The number of cells in a given sample is subject to sampling variation and the recorded number of tumour cells in the sample is subject to measurement error.

Methods: A complete data likelihood was developed for a proportional hazards regression model which recognizes these sources of variation in the observed circulating tumour cell count when it is used as a prognostic covariate.

Results: The performance of the expectation-maximization algorithm was assessed in simulation studies and found to yield estimators with good frequency properties. The asymptotic bias of estimators based on the observed circulating tumour cell count was also examined.

Conclusions: In the situations examined, the proposed estimator was consistent and had considerably smaller empirical bias than estimators arising from analyses based on the observed cell count. The estimated standard error gave confidence intervals with coverage probabilities compatible with the nominal level.

Keywords: Circulating tumour cells, Cox regression, measurement error, sampling variation

Introduction

There has been considerable interest in the prognostic utility of circulating tumour cell (CTC) counts in oncology research in recent years [1-4]. Early research was directed at the use of circulating tumour cell counts as a variable prognostic for disease progression, progression-free survival, or overall survival [5], but these data are being used increasingly often for stratification in clinical trials [5], as a basis for assessing response to treatment in phase II trials [6], and may have the potential to guide treatment selection [7]. Their use in these important settings warrants a careful consideration of the various scientific objectives and issues in statistical analyses.

Miller et al., [8] tabulate data on circulating tumour cell counts in cohorts of normal donors, patients with benign disease, and patients with metastatic breast, colorectal, and prostate cancer. These data reveal that many metastatic cancer patients do not have circulating tumour cells detectable in their blood samples but among those that do, there can be considerable variability in the number detected. Most recent analyses have classified individuals into groups with unfavourable or favourable prognoses depending on whether the observed count was above a specified threshold [9,10]. In metastatic breast and prostate cancer the threshold typically adopted is five so those with a circulating tumour cell count of at least 5 in a 7.5 mL sample are considered to have a poorer prognosis than

those with a count less than 5; in metastatic colorectal cancer the threshold is 3 [5,11,12]. More recently, others have found an association between overall survival and both a baseline CTC count used as a continuous variable and post-treatment change in CTC count [13].

Counts of circulating tumour cells are typically made by visual or computer assisted counting of abnormal cells in blood samples drawn at the point of recruitment to a study [3,14]. The number of cells in a given sample is therefore subject to sampling variation and the recorded number of tumour cells in the sample is subject to measurement error.

Typically when dealing with measurement error, we must first consider the error distribution or the relationship of the mismeasured covariates to the true unobserved covariates. We may classify approaches for measurement error problems as being parametric, semi-parametric or nonparametric by the assumptions of the measurement error distribution. A nonparametric approach was taken by Pepe and Fleming [15] when they empirically estimated the likelihood in the presence of mismeasured covariates with validation data. Huang and Wang [16] also took a nonparametric approach to deal with mismeasured covariates in the Cox model with replicate data available. Tsiatis and Davidan [17] and Kulich and Lin [18] considered semi-parametric approaches of dealing with mismeasured covariates in survival analysis.

There are two fundamentally different interpretations of the unobserved true values of the covariates. In structural modelling the true covariates are regarded as random and a model for their joint distribution is assumed, whereas in functional modelling, the true covariates are considered as a sequence of fixed unknown vectors, or minimal assumptions about the distributional form of the true covariates are made in addition to fixed values. Likelihood methods and Bayesian methods are two general structural modelling methods which require distributional assumptions of the unobserved true covariates. Much literature has concentrated on functional modelling since it involves few or no assumptions regarding the error distribution. For general non-differential error problems, two widely used approaches are regression calibration (RC) and simulation extrapolation (SIMEX). Regression calibration was first proposed by Prentice [19] for the proportional hazard models in survival analysis and required the validation data or replication measurements to estimate the conditional distribution of the unknown true covariates given the observed covariates. Simulation extrapolation was first proposed by Cook and Stefanski [20]. The key idea of this method is that the effect of measurement error can be investigated and therefore adjusted for using simulation techniques. This approach is well suited to additive or multiplicative measurement errors and leads to improve estimates subject to correct model specification [21].

Corrections for the measurement error typically depend on the availability of internal or external data to estimate the error distribution, for example, the replicated measurements or validation dataset. However, in our context regarding CTCs, it is impossible to measure the true concentration; it is also costly to have multiple blood samples to assess the number of CTCs. What's more, we take not only the measurement error but also the sampling variation into consideration in correcting biases. Therefore, in this article, we focus on structural modelling with specified distribution of the true concentration and use likelihood method for correction.

We propose that the implicit goal of prognostic analyses involving circulating tumour cells is to relate the concentration of circulating tumour cells in the total blood volume to the time of interest (e.g., survival time). We develop a complete data likelihood for a proportional hazards regression model based on this concentration which recognizes the sampling variation. The measurement error in the circulating tumour cell count for a specific blood sample is characterized by a simple Bernoulli process for each cell. We consider models using both the concentration in the blood stream and a binary classification of whether this concentration is above a particular threshold. Estimation is carried out based on an expectation-maximization algorithm [22] and we describe how to do this for a parametric proportional hazards model and a semiparametric Cox model [23]. The aim of this article is to demonstrate the potential bias that results when the above sources of error are not accounted for and to describe and

illustrate likelihood based methods for carrying out corrected analyses based on potentially mismeasured or misclassified CTC counts. These methods are assessed through simulation studies and found to yield estimators with good frequency properties.

Methods

Notation

Consider a sample of m independent individuals where $\ddot{C}_i$ denotes the number of circulating tumour cells, $\ddot{V}_i$ denotes the total blood volume, and $\ddot{R}_i = \ddot{C}_i / \ddot{V}_i$ denotes the concentration of circulating tumour cells in the blood stream for individual i , $i = 1, \dots, m$. When a blood sample of $v m \ell$ is drawn we let $\dot{C}_i$ denote the number of tumour cells contained in the sample. Human or automated procedures for counting tumour cells in blood samples are imperfect and it is common for cells present in the sample to be missed. We therefore let C_i denote the number of tumour cells observed in the $v m \ell$ sample and let W_i denote a $p \times 1$ vector of observed auxiliary covariates for individual i , $i = 1, 2, \dots, m$. In the notation above we use "dots" to denote latent variables.

Interest primarily lies in examining the association between circulating tumour cell counts and times to events such as disease progression or death; we let T_i denote the time of the event of interest. In the context of a prospective clinical trial the plan is to follow individuals over the period $[0, \tau)$ where τ is an administrative censoring time. Individuals may of course withdraw from the study early and we let $\tau_i^\dagger$ denote a random time of withdrawal so that events are ultimately observed over the interval $[0, \tau_i)$ where $\tau_i = \min(\tau_i^\dagger, \tau)$ for

individual i , $i = 1, \dots, m$. Let $X_i = \min(T_i, \tau_i)$ and

$$\delta_i = I(X_i = T_i), \quad i = 1, 2, \dots, m. \text{ Let } Y_i(u) = I(u \leq T_i)$$

indicate that individual i is uncensored and $Y_i^\dagger(u) = I(u \leq T_i)$ indicate that the individual i has not yet failed. Then

$\bar{Y}_i(u) = Y_i(u)Y_i^\dagger(u)$ indicates that individual i is at risk and under observation at time u .

We also let $dN_i(u) = I(u = T_i)$ indicate that the

individual i fails at time u and let $d\bar{N}_i(u) = \bar{Y}_i(u)dN_i(u)$

indicate that failure occurred and was observed at time u .

We make the following assumptions regarding the above random variables.

$$A1. T_i \perp (\dot{C}_i, C_i) | (\ddot{R}_i, W_i)$$

$$A2. C_i \perp (\ddot{R}_i, W_i) | \dot{C}_i$$

$$A3. \tau_i \perp (T_i, \ddot{R}_i) | W_i$$

Assumption A1 states that given the concentration of circulating tumour cells and the auxiliary covariates, the failure time is independent of the number of cells in the sample and the number of observed cells, which in turn implies $\dot{C}_i$ and C_i are surrogates, or in other words, the measurement error model is nondifferential. Assumption A2 states that counting of the circulating tumour cells in the sample is conducted independently of the concentration in the body and the auxiliary covariates given the sample count. This is entirely reasonable since the circulating tumour cells in the sample are counted without knowledge of $(\ddot{R}_i, W_i)$. Assumption A3 is an assumption of conditionally independent random censoring [24] which applies to the censoring time given W_i .

Distributional assumptions and model specification

We adopt a structural modeling approach [25] to deal with the issues of measurement error in this problem. Let $G(\ddot{R}_i | W_i; \gamma)$ denote the distribution function for the concentration of circulating tumour cells given the auxiliary variables relevant for the study population. The circulating tumour cells are assumed to be uniformly distributed throughout the blood stream and so given $\ddot{R}_i$ the number of cells in a $v \text{ ml}$ sample is Poisson distributed with mean $v\ddot{R}_i$ [26]. The probability P that a particular cell in the sample is detected is referred to as the detection efficiency. We assume that the event an individual tumour cell in the sample is detected is independent of whether other cells are detected, and so $C_i | \dot{C}_i$ is binomially distributed with $\dot{C}_i$ trials and probability of success p ; we treat $p = p_0$ as a known feature of the measurement system based on external data published in the literature [4].

The effect of circulating tumour cells on survival is typically assessed based on a proportional hazards model involving either the observed count C_i or an indicator $E_i = I(C_i \geq K)$ of whether the count is elevated above a specified threshold K . In metastatic breast cancer, a threshold of $K = 5$ has been reported and used routinely in the literature, and if $E_i = 1$ the individual would typically be considered high risk. The observed cell count C_i is implicitly interpreted in the context of the $v = 7.5 \text{ ml}$ blood sample. The fact that the samples drawn are all of the same volume implies that the prognostic value of the concentration $\ddot{R}_i$ is of primary interest and C_i/v can be viewed as an estimate of the concentration of circulating tumour cells in the blood stream. The Cox regression model of interest is

$$h(t | \ddot{Z}_i; \theta) = h_0(t; \alpha) \exp(\ddot{Z}_i' \beta) \quad (1)$$

where α indexes the baseline hazard $h_0(t; \alpha)$, $\ddot{Z}_i = (v\ddot{R}_i, W_i)'$, $\beta = (\beta_1, \beta_2)'$ where β_1 is the parameter of interest, and $\theta = (\alpha', \beta')$. We use $v\ddot{R}_i$ since current practise is to use the observed cell count which is thought to represent the concentration per

7.5 ml volume.

Bias ignoring sampling variation and measurement error

Under the current formulation the standard method of fitting regression models to the observed circulating tumour cell count involves a misspecified model since it ignores the sampling variation and measurement error in the observation process. Here we consider the implications of using this approach to analysis by studying the limiting behaviour of the resultant estimator using the theory of misspecified likelihoods [27] for the Cox model [28].

We let $Z_i = (C_i, W_i)'$ denote the covariate vector based on available count and auxiliary covariate and model the effect in $T_i | Z_i$ as

$$d\Lambda(t | Z_i) = d\Lambda_0(t) \exp(Z_i' \phi) = d\Lambda_0(t) \exp(C_i \phi_1 + W_i' \phi_2)$$

where the parameters are denoted as

$$d\Lambda_0 = (d\Lambda_0(s_1), \dots, d\Lambda_0(s_R))', \phi = (\phi_1, \phi_2)'$$

$\eta = (d\Lambda_0' \phi)'$; we use different Greek letters for the parameters in the misspecified model since they represent different quantities than their counterparts in the correctly specified models. The misspecified score equations are then

$$\tilde{U}_{1r}(\eta) = \sum_{i=1}^m \left\{ \frac{d\bar{N}_i(s_r)}{d\Lambda_0(s_r)} - \bar{Y}_i(s_r) \exp(Z_i' \phi) \right\}, \quad r=1, 2, \dots, R \quad (2)$$

$$\tilde{U}_2(\eta) = \sum_{i=1}^m \left\{ \int_0^\infty d\bar{N}_i(s) Z_i - \int_0^\infty \bar{Y}_i(s) d\Lambda_0(s) Z_i \exp(Z_i' \phi) \right\} \quad (3)$$

where $\tilde{U}_1(\eta) = (\tilde{U}_{11}(\eta), \dots, \tilde{U}_{1R}(\eta))'$ and we write

$$\tilde{U}(\eta) = (\tilde{U}_1(\eta)', \tilde{U}_2(\eta)')'$$

The limiting values of the naive estimator $\hat{\eta} = (d\Lambda_0' \phi)'$ from solving (2) and (3) are consistent for the solutions to

$$\bar{U}(\eta; \omega) = E_A[\tilde{U}(\eta); \omega]$$

where the expectation is with respect to $A_i = (A_i', \dots, A_m')'$ where $A_i = (T_i, \tau_i, C_i, W_i)'$, and $\omega = (\psi', \rho)'$, where ρ indexes the random censoring time distribution $f(\tau_i | W_i; \rho)$ [28]. ψ characterizes the distribution of A_i and will be explained in the next section.

We define

$$r^{(h)}(u; \omega) = E_A \left[Z_i^{(h)} d\bar{N}_i(u); \omega \right] \\ = E_{W_i} \left(E_{\ddot{R}_i} \left\{ E[Z_i^{(h)} | \ddot{R}_i, W_i] P(\tau_i \geq u | W_i) P(T_i = u | \ddot{R}_i, W_i) \right\} \right)$$

and

$$q^{(h)}(u; \phi, \omega) = E_A[\bar{Y}_i(u) Z_i^{(h)} \exp(Z_i' \phi); \omega]$$

$$= E_{W_i} \left(E_{\tilde{R}_i} \left\{ E[Z_i^{(h)} \exp(Z_i' \phi) | \tilde{R}_i, W_i] P(\tau_i \geq u | W_i) P(T_i \geq u | \tilde{R}_i, W_i) \right\} \right)$$

for: $u \geq 0$, where $Z_i^{(0)} = 1$ and $Z_i^{(1)} = Z_i$. One can then obtain the limiting values of the estimator $d\Lambda_0(s)$ as

$$d\tilde{\Lambda}_0(s_r) \text{ as } d\Lambda^*(s_r) = r^{(0)}(s_r; \omega) / q^{(0)}(s_r; \phi^*, \omega) \text{ where } \phi^*$$

the limiting value of $\tilde{\phi}$ is given by the solution to

$$\int_0^\infty \left\{ r^{(1)}(u; \omega) - \frac{q^{(1)}(u; \phi, \omega)}{q^{(0)}(u; \phi, \omega)} r^{(0)}(u; \omega) \right\} du = 0. \quad (4)$$

Therefore, it remains to evaluate $r^{(h)}(u; \omega)$ and $q^{(h)}(u; \phi, \omega)$, $h=0,1$.

The expectations above can be evaluated numerically by using Monte Carlo integration. In **Figure 1** we display a plot of the limiting value of the estimator of β_1 for the exponential and Cox regression models as a function of (a) the dispersion parameter of the concentration distribution, (b) the mean parameter of the concentration distribution, and (c) the regression parameter itself, for varying degrees of the efficiency of detection. The limiting values are quite different for the parametric and Cox regression models. The plots reveal a clear dependence on the distribution of the concentration and the efficiency of detection; the plotted points represent empirical estimates which are in close agreement with the theoretical results. Among these the most important appears to be the detection efficiency.

Likelihood accommodating sampling variation and measurement error

Using the structural model, if $\Psi = (\theta; \gamma)'$, the complete data likelihood is

$$L_C(\psi) \propto \prod_{i=1}^m P((X_i, \delta_i) | C_i, \dot{C}_i, \ddot{R}_i, W_i; \theta) P(C_i, \dot{C}_i | \ddot{R}_i, W_i) dG(\ddot{R}_i | W_i; \gamma).$$

By Assumption A3,

$$\begin{aligned} P((X_i, \delta_i) | C_i, \dot{C}_i, \ddot{R}_i, W_i; \theta) &= P((X_i, \delta_i) | \ddot{R}_i, W_i; \theta) \\ &= h(X_i | \ddot{Z}_i; \theta)^{\delta_i} \exp(-H(X_i | \ddot{Z}_i; \theta)). \end{aligned}$$

Under Poisson sampling and Assumption A2,

$$P(C_i, \dot{C}_i | \ddot{R}_i, W_i) = P(C_i | \dot{C}_i) P(\dot{C}_i | \ddot{R}_i, W_i),$$

where $P(C_i | \dot{C}_i)$ is binomial with $\dot{C}_i$ trials and known probability of success P_0 , and $P(\dot{C}_i | \ddot{R}_i, W_i)$ is Poisson with

$$E(\dot{C}_i | \ddot{R}_i, W_i) = v\ddot{R}_i. \text{ Since } P(C_i, \dot{C}_i | \ddot{R}_i, W_i) \text{ does not}$$

involve any unknown parameters, we can write the complete data partial log-likelihood as

$$\log L_C(\psi) = \log L_{C1}(\theta) + \log L_{C2}(\gamma), \quad (5)$$

where

$$\log L_{C1}(\theta) = \sum_{i=1}^m \left\{ \int_0^\infty d\bar{N}_i(s) \log dH(s | \ddot{Z}_i; \theta) - \int_0^\infty \bar{Y}_i(s) dH(s | \ddot{Z}_i; \theta) \right\} \quad (6)$$

$$\log L_{C2}(\gamma) = \sum_{i=1}^m \log dG(\ddot{R}_i | W_i; \gamma). \quad (7)$$

Let $D_i = \{(X_i, \delta_i), C_i, W_i\}$ denote the observed data for individual i , $i = 1, 2, \dots, m$ and $D = (D_1, D_2, \dots, D_m)$. Since $\ddot{R}_i$ is unobserved we must take the conditional expectation of (5) using

$$P(\ddot{R}_i | D_i) = \frac{\sum_{\dot{C}_i} h(X_i | \ddot{Z}_i; \theta)^{\delta_i} \exp(-H(X_i | \ddot{Z}_i; \theta)) P(C_i, \dot{C}_i | \ddot{R}_i, W_i) dG(\ddot{R}_i | W_i; \gamma)}{\int_0^\infty \sum_{\dot{C}_i} h(X_i | \ddot{Z}_i; \theta)^{\delta_i} \exp(-H(X_i | \ddot{Z}_i; \theta)) P(C_i, \dot{C}_i | r, W_i) dG(r | W_i; \gamma)} \quad (8)$$

where in the denominator of (8) we integrate with respect to the latent concentration of circulating tumour cells.

If $U(\psi) = (U_1'(\theta), U_2'(\gamma))'$, where $U_1(\theta)$ is the complete data score equations for the log partial likelihood (6) and

$$U_2(\gamma) = \sum_{i=1}^m \partial \log dG(\ddot{R}_i | W_i; \gamma) / \partial \gamma, \text{ then}$$

$$\bar{U}(\psi; \psi) = E[U(\psi) | D; \psi] \quad (9)$$

is the observed data score equation. This equation is best solved iteratively using the EM algorithm [22] and we do this by taking the expectation of (9) using an estimate $\hat{\psi}^{(k)}$ at the

$$k \text{ th iteration to compute } \bar{U}(\psi; \hat{\psi}^{(k)}) = E[U(\psi) | D; \hat{\psi}^{(k)}],$$

which is in turn set equal to zero and solved to give $\hat{\psi}^{(k+1)}$.

This procedure is repeated iteratively until convergence.

If (1) is to be modeled using a semi-parametric Cox model, following Martinussen [29], we let $s_1, s_2, \dots, s_R$ be the distinct times at which $d\bar{N}_i(s) = 1$, and

$$dH_{0r} = dH_0(s_r), r = 1, 2, \dots, R, dH_0 = (dH_{01}, dH_{02}, \dots, dH_{0R})'$$

and $\theta = (dH_0, \beta)'$. Then for any estimate $\hat{\psi}$, the expected score vector (9) has components

$$\bar{U}_{11r}(\theta; \hat{\psi}) = \sum_{i=1}^m \{ d\bar{N}_i(s_r) - \bar{Y}_i(s_r) dH_0(s_r) E[\exp(\ddot{Z}_i' \beta) | D_i; \hat{\psi}] \}, \quad r = 1, \dots, R$$

$$\bar{U}_{12}(\theta; \hat{\psi}) = \sum_{i=1}^m \left\{ \int_0^\infty d\bar{N}_i(s) E[\ddot{Z}_i | D_i; \hat{\psi}] - \int_0^\infty \bar{Y}_i(s) dH_0(s) E[\ddot{Z}_i e^{\ddot{Z}_i' \beta} | D_i; \hat{\psi}] \right\}$$

$$\bar{U}_2(\gamma; \hat{\psi}) = \sum_{i=1}^m E[\partial \log dG(\ddot{R}_i | W_i; \gamma) / \partial \gamma | D_i; \hat{\psi}]$$

Figure 1. Asymptotic bias of naive maximum likelihood estimates for β from the exponential and Cox proportional hazards model for $p=0.6, 0.8$ and 1.0 across a range of values of (a) ϕ , (b) μ , and (c) β_1 . Points indicate values of the parameters used in the simulation models of Section 3

conformable with (dH_0, β', γ') , and

$$\bar{U}_{11}(\theta; \hat{\psi}) = (\bar{U}_{111}(\theta; \hat{\psi}), \dots, \bar{U}_{11R}(\theta; \hat{\psi}))',$$

$$\bar{U}_1(\psi; \hat{\psi}) = (\bar{U}_{11}'(\theta; \hat{\psi}), \bar{U}_{12}'(\gamma; \hat{\psi}))'.$$

The equation $\bar{U}_{11r}(\theta; \hat{\psi}) = 0$ can be solved to give

$$dH_0(s; \beta) = \frac{\sum_{i=1}^m d\bar{N}_i(s)}{\sum_{i=1}^m \bar{Y}_i(s) E[\exp(\tilde{Z}_i' \beta) | D_i, \hat{\psi}]}, \quad (10)$$

which is zero except at times of failure, $s_1, s_2, \dots, s_R$.

Letting $s^{(h)}(u; \beta) = \sum_{i=1}^m \bar{Y}_i(u) E[\tilde{Z}_i^h \exp(\tilde{Z}_i' \beta) | D_i, \hat{\psi}]$, $h = 0, 1$, we

substitute (10) into $\bar{U}_{12}(\theta; \hat{\psi})$ to get the estimating equation for β as

$$\sum_{i=1}^m \int_0^\infty \left\{ E[\tilde{Z}_i | D_i; \hat{\psi}] - \frac{s^{(1)}(u; \beta)}{s^{(0)}(u; \beta)} \right\} d\bar{N}_i(u) = 0. \quad (11)$$

To evaluate the estimating equation for γ , we propose specification of the parametric form of $G(\tilde{R}_i | W_i; \gamma)$.

To evaluate these conditional expectations in (10)-(11), we use Gauss-Laguerre quadrature [30] given the conditional probability density function (8). Gauss-Laguerre quadrature involves approximating the integrals on the interval $[0, \infty)$ by:

$$\int_0^{\infty} f(x) dx = \int_0^{\infty} f(x) e^{-x} e^x dx = \int_0^{\infty} g(x) e^{-x} dx = \sum_{s=1}^N w_s g(x_s),$$

where x_s and w_s are the nodes and weights that picked based on weight function e^{-x} . Monte Carlo methods with rejection sampling could alternatively be used to approximate these expectations by simulation. The asymptotic variance of $\hat{\psi}$ can be estimated as suggested by Louis [31] and implemented by Martinussen [29]; see the Appendix for details on the current setting. Further details and examples of using the EM algorithm to fit Cox regression models can be found in Klein [32], Lipsitz and Ibrahim [33], Herring and Ibrahim [34], and Nielsen [35].

Results

Empirical study of finite sample behaviour

Here we present the results of a series of simulation studies conducted to assess the influence of sampling variability and mismeasurement on the behaviour of naive estimators and to examine the performance of the proposed estimator. We generate the data following the model formulation in the previous section. Specifically we generate the total circulating tumour cell count in an individual assuming $\tilde{R}_i$ is gamma distributed with mean μ and variance ϕ ; we consider $(\mu, \phi) = (2, 1)$ and $(2, 4)$ to represent a mean of 2 in the population and a coefficient of ϕ variation of 0.5 and 1.0 respectively. We also consider $(\mu, \phi) = (2/3, 1)$ to correspond to a smaller mean reflecting a higher proportion of zero circulating tumour cell counts but with a coefficient of variation of 1.5. The times of interest were generated from exponential regression models with

$$h(t | \tilde{R}_i) = h_0(t) \exp(v \tilde{R}_i \beta) \quad (12)$$

We considered the case in which the failure times were uncensored and administratively censored at a point where 25% of the events would be unobserved. To reflect the sampling and observation process we generated $\tilde{C}_i | \tilde{R}_i$ as a Poisson random variable with mean $v \tilde{R}_i$ and $C_i | \tilde{C}_i$ as binomially distributed with probability $p_0 = 0.80$.

Analyses were carried out under an exponential regression model and Cox regression model of the form of (12) using $v \tilde{R}_i$, $\tilde{C}_i$ and C_i . The first of these is not possible in practise since it requires the true concentration of circulating tumour cells. The second is also not possible since it requires perfect recording of the number of circulating tumour cells in the sample, and the third analysis reflects the current practise of using the recorded counts in the samples. The fourth analysis was based on the EM algorithm described above. A second set of

analyses were there carried out using the binary counterparts to the above variables constructed with a threshold of 5. Specifically we let $\tilde{E}_i = I(v \tilde{R}_i \geq 5)$ and $\dot{E}_i = I(\tilde{C}_i \geq 5)$ denote the latent variables based on the unobserved concentration and count, $E_i = I(C_i \geq 5)$ be the binary variable based on the observed count. Based on the parameter configurations $P(\tilde{E}_i = 1) = 0.954, 0.717$ and 0.308 for $(\mu, \phi) = (2, 1), (2, 4)$ and $(2/3, 1)$ respectively. The sensitivities $P(E_i = 1 | \tilde{E}_i = 1)$ and specificities $P(E_i = 0 | \tilde{E}_i = 0)$ of the observed binary variable for the true binary variable are $(0.919, 0.792), (0.899, 0.907)$ and for $(\mu, \phi) = (2, 1), (2, 4)$ and $(2/3, 1)$ respectively. The analyses using the continuous and binary variables are summarized in (Tables 1 and 2) respectively where the empirical biases, the empirical standard errors and the empirical coverage probabilities for nominal 95% confidence intervals are given.

The results for the analysis based on a continuous covariate in (Table 1) reveal negligible empirical bias and excellent empirical coverage probability for the Cox regression model based on the true concentration. Appreciable bias in the estimator is apparent for the analyses based on the sample concentration and there is a considerable impact on the empirical coverage probability. The performance of the estimators based on the observed circulating tumour cell count is also very poor in terms of bias and coverage probability. The proposed estimator has considerably smaller bias and standard errors based on the method of Louis [31] give confidence intervals which are compatible with the 95% level. This performance is roughly the same whether the event times are subject to censoring or not and there is good performance seen as well for the parametric analyses.

Table 2 contains analogous results for the binary covariate. Here the measurement error results in a misclassification of the binary indicator and these misclassification probabilities are reported in Table 2. Again the proposed EM algorithm reduces the bias over the naive analyses.

Discussion

Our proposed methodology is the first we are aware of to rigorously deal with the statistical issues arising from sampling variation and measurement error in the enumeration of CTCs. The utility of circulating tumor cells was originally examined in the context of their prognostic role for progression and survival, but they are now being used increasingly often to form surrogate outcomes in phase II and III clinical trials [36] and are currently receiving consideration as biomarkers to assess treatment response [37] in patients with advanced cancer. The statistical issues of sampling variation and measurement error raised here bear on their use in these contexts as well.

However, our method is not without limitations. We have focussed on the use of circulating tumour cells in the context of parametric and semiparametric proportional hazards models. Other types of regression models are of courses possible including accelerated failure time models, additive models or Cox-Aalen types models [38]. Model diagnostics represent

Table 1. Empirical performance of estimators of the regression coefficient for the continuous covariate under parametric and semi-parametric Cox regression models with and without administrative censoring; sample size $m=500$, number of simulations $nsim=500$.

			True Concentration ($v\tilde{R}_i$)			Sample Count ($\hat{C}_i$)			Observed Count (C_i)			EM Algorithm		
$\exp(\beta_i)$	μ	ϕ	EBIAS [†]	ESE [†]	ECP%	EBIAS [†]	ESE [†]	ECP%	EBIAS [†]	ESE [†]	ECP%	EBIAS [†]	ESE [†]	ECP%
Exponential Regression, No Censoring														
1.1	2/3	1	-0.024	0.610	95.6	1.122	0.639	50.0	-0.689	0.832	80.8	0.161	0.780	95.4
	2	1	0.008	0.588	95.2	2.278	0.594	2.0	0.975	0.748	63.8	-0.030	0.809	94.6
	2	4	-0.037	0.295	95.6	0.922	0.359	18.0	-0.962	0.462	25.6	0.032	0.389	96.4
1.25	2/3	1	-0.057	0.610	96.0	3.500	0.968	0.8	-0.331	1.248	73.6	0.429	1.066	93.0
	2	1	-0.013	0.609	93.8	6.011	0.865	0	3.213	1.114	4.4	-0.090	1.028	95.0
	2	4	-0.028	0.303	95.4	3.073	0.689	0	-0.951	0.929	31.2	0.026	0.592	94.0
Exponential Regression, 25% Administrative Censoring														
1.1	2/3	1	-0.032	0.642	95.4	1.111	0.665	52.8	-0.706	0.858	80.0	0.166	0.800	95.8
	2	1	-0.022	0.670	93.6	2.340	0.669	4.2	1.064	0.832	64.6	0.009	0.956	95.2
	2	4	-0.050	0.319	94.8	0.957	0.382	20.2	-0.908	0.492	38.4	0.045	0.422	95.6
1.25	2/3	1	-0.074	0.633	96.0	3.551	1.017	1.2	-0.256	1.303	76.0	0.469	1.132	92.8
	2	1	-0.018	0.751	94.0	6.793	0.965	0	4.301	1.193	1.4	-0.130	1.250	94.4
	2	4	-0.039	0.331	94.4	3.265	0.752	0	-0.659	1.020	44.2	0.032	0.659	93.6
Cox Regression, No Censoring														
1.1	2/3	1	0.016	0.694	96.8	-1.306	0.689	46.6	0.425	0.877	90.2	0.117	0.877	96.0
	2	1	0.033	0.674	95.2	-2.668	0.631	1.2	-1.511	0.774	42.4	0.004	1.026	95.4
	2	4	0.069	0.461	95.2	-1.451	0.473	10.0	0.198	0.586	87.8	-0.021	0.594	96.0
1.25	2/3	1	0.040	1.078	94.8	-5.388	1.126	0	-2.248	1.395	44.0	0.129	1.698	95.8
	2	1	0.026	1.027	93.6	-9.185	0.872	0	-7.310	1.015	0	-0.426	1.969	93.8
	2	4	0.094	0.929	94.8	-7.334	0.798	0	-4.778	0.983	0.4	-0.584	1.463	92.4
Cox Regression, 25% Administrative Censoring														
1.1	2/3	1	0.018	0.703	96.6	-1.253	0.7	50.6	0.503	0.887	89.0	0.121	0.883	96.6
	2	1	0.035	0.714	95.2	-2.527	0.677	3.0	-1.317	0.832	55.6	0.033	1.070	95.0
	2	4	0.074	0.464	95.2	-1.399	0.475	12.8	0.271	0.588	87.6	-0.016	0.596	96.8
1.25	2/3	1	0.035	1.096	94.0	-5.288	1.148	0.2	-2.104	1.421	48.6	0.124	1.700	94.8
	2	1	0.004	1.071	94.6	-8.907	0.915	0	-6.957	1.066	0	-0.376	1.988	93.8
	2	4	0.097	0.943	94.2	-7.246	0.829	0	-4.661	1.032	0.4	-0.562	1.496	92.8

[†]EBIAS is the empirical bias $\times 100$, ESE is the empirical standard error $\times 100$.

an important step in survival analysis to assess the plausibility of the model assumptions. Key assumptions to check include the way the covariate effects have been expressed and the proportional hazards assumption. These can be assessed through the analysis of Schoenfeld or martingale residuals [39]. In settings where empirical covariate distributions might involve extreme values, influence statistics can play an important role. Further research is warranted on the development of methods for carrying out these diagnostic tests when covariates are measured with error.

An additional limitation is due to the fact that the concentration of circulating tumour cells is not observed in any patients it is difficult to assess the validity of any parametric assumptions made for the distribution of this concentration. In such settings model expansion offers one feasible approach to check for the adequacy of the specification of $G(\cdot)$; score tests are particularly appealing in this setting since they do not require fitting the expanded model [40]. Power of these tests will typically be rather low since they are based on the latent

true concentration. See Waagepetersen [41] and Alonso et al., [42]. Additionally, depending on the disease severity of the subjects used in a CTC study there may be a large proportion of subjects with a true concentration of zero or subjects with very high concentrations. Other approaches for dealing with the difficulty in specifying the distribution of the concentration in the sample include nonparametric estimation but this is really only feasible in the absence of auxiliary covariates or if W_i is discrete. Spline-based approaches of kernel density estimation offer alternatives [43].

Many statistical methods for correcting for measurement error rely on data from validation studies. If the correct underlying value of the covariate of interest is available in a validation sub-sample then one can estimate features of the measurement process. Validation studies will not be possible in this setting, but auxiliary data in which repeated blood samples are taken and circulating tumor cells are counted could yield more precise estimates of the true concentration and enhance the precision of the corrected estimators we described [44].

Table 2. Empirical performance of estimators of the regression coefficient for the binary covariate under parametric and semi-parametric Cox regression models with and without administrative censoring; sample size $m=500$, number of simulations $nsim=500$.

exp(β_1)	μ	ϕ	True Concentration ($\tilde{E}_i$)			Sample Count ($\tilde{E}_i$)			Observed Count (E_i)			EM Algorithm		
			EBIAS [†]	ESE	ECP%	EBIAS [†]	ESE	ECP%	EBIAS [†]	ESE	ECP%	EBIAS [†]	ESE	ECP%
Exponential Regression, No Censoring														
1.5	2/3	1	0.368	0.098	93.6	9.259	0.100	83.2	7.831	0.107	84.6	-0.479	0.118	93.8
	2	1	1.782	0.232	93.6	23.635	0.193	75.4	28.068	0.152	48.2	-9.314	0.387	94.0
	2	4	0.429	0.095	94.6	7.897	0.096	87.6	10.019	0.092	83.0	-0.369	0.111	95.6
2	2/3	1	-0.378	0.089	96.0	16.652	0.099	58.2	13.829	0.103	71.2	0.233	0.109	96.0
	2	1	3.526	0.210	95.4	35.715	0.207	51.4	43.804	0.150	14.2	-8.262	0.318	95.0
	2	4	0.059	0.094	97.0	13.361	0.095	73.4	15.706	0.092	62.2	-0.419	0.107	96.4
Exponential Regression, 25% Administrative Censoring														
1.5	2/3	1	0.860	0.105	95.4	8.994	0.109	86.0	8.005	0.113	90.2	-0.744	0.129	95.8
	2	1	-1.927	0.301	94.8	24.461	0.216	80.2	29.121	0.173	56.8	-1.338	0.570	95.8
	2	4	-0.007	0.116	94.6	7.933	0.112	91.4	10.840	0.105	86.4	0.123	0.137	95.8
2	2/3	1	-0.298	0.102	95.6	14.478	0.108	71.0	12.947	0.111	76.2	0.375	0.126	95.8
	2	1	-0.176	0.312	94.2	40.487	0.235	53.4	48.858	0.175	21.0	4.514	0.747	94.8
	2	4	-0.734	0.122	95.8	13.723	0.120	80.0	17.789	0.110	66.6	0.299	0.144	97.0
Cox Regression, No Censoring														
1.5	2/3	1	-0.211	0.098	94.2	-9.342	0.100	82.8	-8.001	0.106	85.0	-0.232	0.122	95.4
	2	1	-1.543	0.237	93.6	-23.878	0.192	74.0	-28.380	0.152	48.0	-10.127	0.442	94.0
	2	4	-0.039	0.100	94.8	-7.951	0.098	87.8	-10.149	0.095	82.4	0.245	0.121	94.4
2	2/3	1	0.524	0.094	96.0	-17.621	0.102	56.0	-14.948	0.106	68.0	0.484	0.122	95.6
	2	1	-2.889	0.217	95.0	-36.702	0.202	51.0	-44.949	0.146	12.6	-7.136	0.338	96.2
	2	4	0.213	0.100	97.4	-15.107	0.094	70.2	-17.508	0.091	58.2	-0.051	0.118	97.0
Cox Regression, 25% Administrative Censoring														
1.5	2/3	1	-0.834	0.106	95.4	-9.046	0.109	86.0	-8.071	0.113	90.2	-0.725	0.130	95.2
	2	1	1.889	0.301	94.6	-24.511	0.215	80.0	-29.169	0.172	56.6	-1.915	0.585	95.8
	2	4	0.003	0.116	94.6	-7.967	0.112	90.8	-10.865	0.106	86.4	0.174	0.139	96.0
2	2/3	1	0.298	0.103	96.2	-14.852	0.108	71.0	-13.351	0.111	76.0	0.404	0.129	95.4
	2	1	0.204	0.312	94.2	-40.506	0.235	53.6	-48.860	0.175	21.2	1.285	0.637	93.6
	2	4	0.688	0.123	95.8	-13.997	0.120	80.0	-18.047	0.110	66.4	0.204	0.147	96.4

[†]EBIAS is the empirical bias $\times 100$.

Conclusions

The present study demonstrates that using observed CTC count as a prognostic covariate may lead to biased results when sources of sampling variation and measurement error are not accounted for. This could lead to false conclusions regarding the association between CTC count and patient outcomes. A likelihood based methods for carrying out corrected analyses is described and illustrated. These methods are assessed through simulation studies and found to yield estimators with good frequency properties.

Competing interests

The author declares that they have competing interests.

Authors' contributions

Authors' contributions	CAC	YW	RJC
Research concept and design	✓	✓	✓
Collection and/or assembly of data	--	✓	--
Data analysis and interpretation	✓	✓	✓
Writing the article	✓	✓	✓
Critical revision of the article	✓	✓	✓
Final approval of article	✓	✓	✓
Statistical analysis	--	✓	--

Acknowledgement and funding

The authors wish to thank Dr Alison Allan (London Regional Cancer Program, London Health Sciences Centre; Departments of Oncology and Anatomy & Cell Biology, Schulich School of

Medicine and Dentistry, University of Western Ontario, London, Ontario) for helpful discussions and comments. This research was supported by grants from the Natural Sciences and Engineering Research Council of Canada (RGPIN 402474 for CAC and RGPIN 155849 for RJC) and the Canadian Institutes for Health Research (FRN 13887). Ying Wu was supported by a grant from the Division of High Impact Clinical Trials of the Ontario Institute for Cancer Research. Richard Cook is a Canada Research Chair in Statistical Methods for Health Research.

Publication history

EIC: Jimmy Efird, East Carolina University, USA.

Received: 26-Sep-2013 Revised: 11-Oct-2013

Accepted: 18-Oct-2013 Published: 23-Jan-2014

References

- Johnson PW, Burchill SA and Selby PJ. **The molecular detection of circulating tumour cells.** *Br J Cancer.* 1995; **72**:268-76. | [Article](#) | [PubMed Abstract](#) | [PubMed Full Text](#)
- Nagrath S, Sequist LV, Maheswaran S, Bell DW, Irimia D, Utkus L, Smith MR, Kwak EL, Digumarthy S, Muzikansky A, Ryan P, Balis UJ, Tompkins RG, Haber DA and Toner M. **Isolation of rare circulating tumour cells in cancer patients by microchip technology.** *Nature.* 2007; **450**:1235-9. | [Article](#) | [PubMed Abstract](#) | [PubMed Full Text](#)
- Paterlini-Brechot P and Benali NL. **Circulating tumor cells (CTC) detection: clinical impact and future directions.** *Cancer Lett.* 2007; **253**:180-204. | [Article](#) | [PubMed](#)
- Tibbe AG, Miller MC and Terstappen LW. **Statistical considerations for enumeration of circulating tumor cells.** *Cytometry A.* 2007; **71**:154-62. | [Article](#) | [PubMed](#)
- Cristofanilli M, Budd GT, Ellis MJ, Stopeck A, Matera J, Miller MC, Reuben JM, Doyle GV, Allard WJ, Terstappen LW and Hayes DF. **Circulating tumor cells, disease progression, and survival in metastatic breast cancer.** *N Engl J Med.* 2004; **351**:781-91. | [Article](#) | [PubMed](#)
- Pierga JY, Bidard FC, Mathiot C, Brain E, Delalogue S, Giachetti S, de Cremoux P, Salmon R, Vincent-Salomon A and Marty M. **Circulating tumor cell detection predicts early metastatic relapse after neoadjuvant chemotherapy in large operable and locally advanced breast cancer in a phase II randomized trial.** *Clin Cancer Res.* 2008; **14**:7004-10. | [Article](#) | [PubMed](#)
- Dotan E, Cohen SJ, Alpaugh KR and Meropol NJ. **Circulating tumor cells: evolving evidence and future challenges.** *Oncologist.* 2009; **14**:1070-82. | [Article](#) | [PubMed](#)
- Miller MC, Doyle GV and Terstappen LW. **Significance of Circulating Tumor Cells Detected by the CellSearch System in Patients with Metastatic Breast Colorectal and Prostate Cancer.** *J Oncol.* 2010; **2010**:1-8. | [Article](#) | [PubMed Abstract](#) | [PubMed Full Text](#)
- Hayes DF, Cristofanilli M, Budd GT, Ellis MJ, Stopeck A, Miller MC, Matera J, Allard WJ, Doyle GV and Terstappen LW. **Circulating tumor cells at each follow-up time point during therapy of metastatic breast cancer patients predict progression-free and overall survival.** *Clin Cancer Res.* 2006; **12**:4218-24. | [Article](#) | [PubMed](#)
- Lucci A, Hall CS, Lodhi AK, Bhattacharyya A, Anderson AE, Xiao L, Bedrosian I, Kuerer HM and Krishnamurthy S. **Circulating tumour cells in non-metastatic breast cancer: a prospective study.** *Lancet Oncol.* 2012; **13**:688-95. | [Article](#) | [PubMed](#)
- Cohen SJ, Punt CJA, Iannotti N, Saidman BH, Sabbath KD, Gabrail NY, Picus J, Morse M, Mitchell E, Miller MC, Doyle GV, Tissing H, Terstappen LWM and Meropol NJ. **Relationship of circulating tumor cells to tumor response, progression-free survival, and overall survival in patients with metastatic colorectal cancer.** *J Clin Oncol.* 2008; **26**:3213-21. | [Pdf](#)
- de Bono JS, Scher HI, Montgomery RB, Parker C, Miller MC, Tissing H, Doyle GV, Terstappen LW, Pienta KJ and Raghavan D. **Circulating tumor cells predict survival benefit from treatment in metastatic castration-resistant prostate cancer.** *Clin Cancer Res.* 2008; **14**:6302-9. | [Article](#) | [PubMed](#)
- Scher HI, Jia X, de Bono JS, Fleisher M, Pienta KJ, Raghavan D and Heller G. **Circulating tumour cells as prognostic markers in progressive, castration-resistant prostate cancer: a reanalysis of IMMC38 trial data.** *Lancet Oncol.* 2009; **10**:233-9. | [Article](#) | [PubMed Abstract](#) | [PubMed Full Text](#)
- Allan AL and Keeney M. **Circulating tumor cell analysis: technical and statistical considerations for application to the clinic.** *J Oncol.* 2010; **2010**:1-10. | [Article](#) | [PubMed Abstract](#) | [PubMed Full Text](#)
- Pepe MS and Fleming TR. **A nonparametric method for dealing with mismeasured covariate data.** *J Am Stat Assoc.* 1991; **86**:108-13. | [Article](#)
- Huang Y and Wang CY. **Cox regression with accurate covariates unascertainable: a nonparametric- correction approach.** *J Am Stat Assoc.* 2000; **95**:1209-19. | [Article](#)
- Tsiatis AA and Davidian M. **A semiparametric estimator for the proportional hazards model with longitudinal covariates measured with error.** *Biometrika.* 2001; **88**:447-58. | [Article](#)
- Kulich M and Lin DY. **Additive hazards regression with covariate measurement error.** *J Am Stat Assoc.* 2000; **95**:238-48. | [Article](#)
- Prentice RL. **Covariate measurement errors and parameter estimation in a failure time regression model.** *Biometrika.* 1982; **69**:331-42. | [Article](#)
- Cook J and Stefanski LA. **A simulation extrapolation method for parametric measurement error models.** *J Am Stat Assoc.* 1994; **89**:464-7. | [Article](#)
- Yi GY and Cook RJ. **Errors in the measurement of covariates.** *Encyclopedia of Biostatistics.* 2nd Edition 2005; **3**:1741-8. | [Book](#)
- Dempster AP, Laird NM and Rubin DB. **Maximum likelihood from incomplete data via the EM algorithm.** *J R Stat Soc Series B Stat Methodol.* 1977; **39**:1-38. | [Article](#)
- Cox DR. **Regression models and life tables.** *J R Stat Soc Series B Stat Methodol.* 1972; **34**:187-220. | [Pdf](#)
- Lawless JF. **Statistical Models and Methods for Lifetime Data.** 2nd edition. John Wiley and Sons: Hoboken, New Jersey, 2003. | [Book](#)
- Carroll RJ, Ruppert D, Stefanski LA and Crainiceanu CM. **Measurement Error in Nonlinear Models.** Chapman and Hall: London, 2006. | [Book](#)
- Allard WJ, Matera J, Miller MC, Repollet M, Connelly MC, Rao C, Tibbe AG, Uhr JW and Terstappen LW. **Tumor cells circulate in the peripheral blood of all major carcinomas but not in healthy subjects or patients with nonmalignant diseases.** *Clin Cancer Res.* 2004; **10**:6897-904. | [Article](#) | [PubMed](#)
- White H. **Maximum likelihood estimation of misspecified models.** *Econometrica.* 1982; **50**:1-26. | [Article](#)
- Struthers CA and Kalbfleisch JD. **Misspecified proportional hazard models.** *Biometrika.* 1986; **73**:363-9. | [Article](#)
- Martinussen T. **Cox regression with incomplete covariate measurements using the EM-algorithm.** *Scand J Statist.* 1999; **26**:479-91. | [Article](#)
- Golub GH and Welsch JH. **Calculation of Gaussian quadrature rules.** *Math Comput.* 1969; **23**:221-30. | [Article](#)
- Louis TA. **Finding the observed information matrix when using the EM algorithm.** *J R Stat Soc Series B Stat Methodol.* 1982; **44**:226-233. | [Article](#)
- Klein JP. **Semiparametric estimation of random effects using the Cox model based on the EM algorithm.** *Biometrics.* 1992; **48**:795-806. | [Article](#) | [PubMed](#)
- Lipsitz SR and Ibrahim JG. **Estimating equations with incomplete categorical covariates in the Cox model.** *Biometrics.* 1998; **54**:1002-13. | [Article](#) | [PubMed](#)
- Herring AH and Ibrahim JG. **Likelihood-based methods for missing covariates in the Cox proportional hazards model.** *JASA* 2001; **96**:292-302. | [Article](#)
- Nielsen SF. **Survival analysis with coarsely observed covariates.** *SORT.* 2003; **27**:41-64. | [Pdf](#)
- Nakamura S, Yagata H, Ohno S, Yamaguchi H, Iwata H, Tsunoda N, Ito Y, Tokudome N, Toi M, Kuroi K and Suzuki E. **Multi-center study evaluating**

- circulating tumor cells as a surrogate for response to treatment and overall survival in metastatic breast cancer.** *Breast Cancer*. 2010; **17**:199-204. | [Article](#) | [PubMed](#)
37. Tewes M, Aktas B, Welt A, Mueller S, Hauch S, Kimmig R and Kasimir-Bauer S. **Molecular profiling and predictive value of circulating tumor cells in patients with metastatic breast cancer: an option for monitoring response to breast cancer related therapies.** *Breast Cancer Res Treat*. 2009; **115**:581-90. | [Article](#) | [PubMed](#)
38. Martinussen T and Scheike TH. **Dynamic Regression Models for Survival Data.** Springer: New York, 2006. | [Book](#)
39. Therneau TM and Grambsch PM. **Modeling Survival Data: Extending the Cox Model.** Springer: New York, 2000. | [Book](#)
40. Cox DR and Hinkley DV. **Theoretical Statistics.** Chapman and Hall: London, 1974. | [Book](#)
41. Waagepetersen T. **A simulation-based goodness-of-fit test for random effects in generalized linear mixed models.** *Scand J Statist*. 2006; **33**:721-31. | [Article](#)
42. Alonso A, Litière S and Molenbergh G. **A family of tests to detect misspecifications in the random-effects structure of generalized linear mixed models.** *Comput Stat Data An*. 2008; **52**:4474-86. | [Article](#)
43. Izenman AJ. **Modern multivariate statistical techniques: regression, classification, and manifold learning.** Springer: New York, 2008. | [Book](#)
44. Schafer DW and Purdy KG. **Likelihood analysis for errors-in-variables regression with replicate measurements.** *Biometrika*. 1996; **93**:813-24. | [Article](#)

Citation:

Cotton CA, Wu Y and Cook RJ. **Correcting for sampling variation and measurement error in Cox regression with circulating tumour cells.** *J Med Stat Inform*. 2014; **2**:1. <http://dx.doi.org/10.7243/2053-7662-2-1>