Detecting bias arising from delayed recording of time

Bianca L. De Stavola

Centre for Statistical Methodology, London School of Hygiene and Tropical Medicine, London, U.K.

E-mail: bianca.destavola@lshtm.ac.uk

D. R. Cox

Nuffield College, Oxford, U.K.

E-mail: david.cox@nuffield.oxford.ac.uk

Abstract. Sometimes in studies of the dependence of survival time on explanatory variables the natural time origin for defining entry into study cannot be observed and a delayed time origin is used instead. For example, diagnosis of disease may in some patients be made only at death. The effect of such delays is investigated both theoretically and in the context of the England and Wales National Cancer Register.

Keywords: Cancer Registers, cancer survival, exponential distribution, left-censoring, reporting delay, selection bias

1. Introduction

The key elements in defining even the simplest form of survival study are that for each individual involved there should be a clear time origin, that the passage of time should be appropriately measured, and that the outcome of interest should be unambiguous. The measurement of time and identification of outcome have both been extensively discussed in the literature; see, for example, Cox and Oakes (1984). In contrast the identification of the time origin, the time from which individuals are at risk of experiencing the outcome, has not been examined to the same extent.

Common choices of time origin include date of birth and time of first exposure, but there are situations where a different entry point is used instead because the true origin is not observable, for example because of defects in the detection of the start of the 'at risk' period. This could happen because of delays in disease detection (*e.g.* when detection requires extensive investigations) or in registration to a particular scheme (*e.g.* to receive benefits). Using an imprecise entry has consequences in terms of a distorted account of the time scale and, potentially, of distorted associations with the outcome of interest.

We represent this by considering for each individual three time points, an unobserved time origin, an observed delayed entry time and an outcome. There are thus three random variables, V, unobserved, the time between origin and outcome, Z, also unobserved, between origin and entry and also T, which is directly observed, between entry and outcome.

Our aim is to discuss the likely consequences of delayed recording of time for the hazard ratio (HR) of an exposure of interest when that ratio is estimated using the observed entry time and to propose a test of whether such an estimate differs from the HR on the true time scale.

The paper is organized as follows. Section 2 describes a motivating application; Section 3 presents some theoretical results under simple assumptions. Section 4 revisits the application in the light of the theoretical results and Section 5 draws some final remarks.

2. Motivation : cancer survival and deprivation score

Registration of all cancer diagnoses is carried out routinely in England and Wales via the National Cancer Register, with the data then regularly linked to the NHS Central Registration System for assessment of vital status, and recoding of cause and date of death. For a minority of cancer cases inclusion in the Cancer Registration System occurs only because cancer was mentioned in the death certificates. Hence, for these patients, date of diagnosis coincides with their date of death and follow-up time is zero. Such occurrences highlight that detection is then later than the actual onset of disease.

We have access to data on patients registered with a diagnosis of breast cancer (in women only) and lung cancer (both sexes) in the National Cancer Registry of England and Wales in 1995-2007 with follow-up to 31 December 2007. For each patient we know the deprivation score (Carstairs and Morris, 1989) of their area of residence at the time of diagnosis (or at the date of death if date of diagnosis was missing). This index is categorical, with the five groups corresponding to quintiles of the England and Wales distribution of this score.

Overall, breast cancer is relatively more frequent among the least deprived groups, whereas lung cancer is relatively more frequent among the most deprived (Deprivation Gradient for Cancer Incidence, Cancer Research UK, 2016). However the frequency of diagnosis at time of death increases with deprivation score for both cancers (Table 1). This increase is reflected in the increased odds of diagnosis at death certification from the least to the most deprived group, especially for breast cancer patients (Table 2), a factor possibly related to uptake of screening. The same direction of effects is seen in terms of survival hazard rates for the patients whose follow-up time is greater than zero (Table 3). The interpretation of these is, however, not straightforward because of the possible bias

introduced by left-censoring (*i.e.* shortening of the follow-up time) and left-truncation (selection of individual with follow-up time greater than zero).

3. Theoretical development

3.1. A simple model

As noted in Section 1, there are three random variables, V and Z, which are unobserved, and T, which is directly observed. There is the possibility, assumed in much of the discussion to have small probability, that the outcome has already occurred at the instance of detection, in which case we define T to be zero (Figure 1).

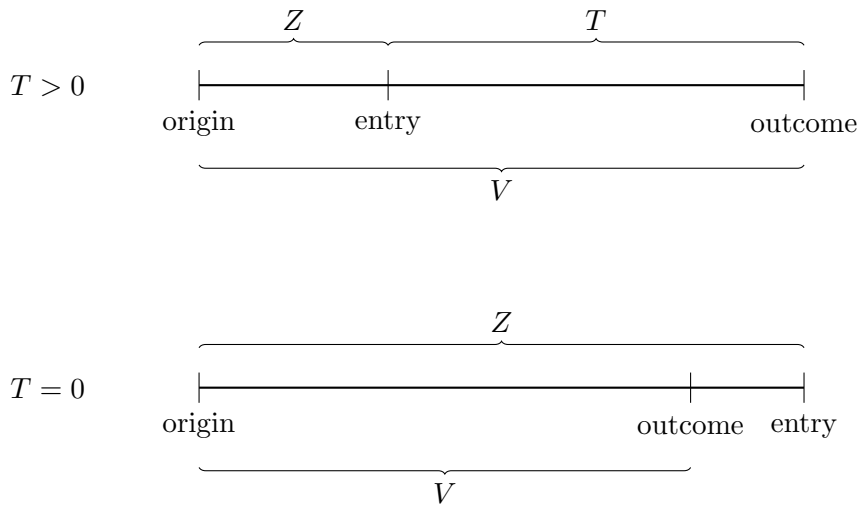

Figure 1. Two typical scenarios where Z represents the unobserved time between origin and entry, V the unobserved time between origin and outcome and $T = (V - X)^+ > 0$ or $= 0$. (In the motivating example, origin is time of true disease onset, entry is time of diagnosis and outcome is death.)

In general we write $T = (V - Z)^+$. The object of study is the dependence of V on a vector $\mathbf{x}$ of explanatory variables. We can observe only the dependence on $\mathbf{x}$ of T, in fact in two parts, namely the dependence of $P(T = 0)$ and the dependence of T conditionally on $T > 0$.

In the simplest special case, V and Z are independently exponentially distributed with rate or hazard parameter, that is the reciprocal of the mean, ρ_V and ρ_Z, respectively, corresponding to events in independent Poisson processes. Then $P(T = 0) = \rho_V/(\rho_V + \rho_Z)$ and the distribution of T^+, that is T conditionally on $T > 0$, is the same as that of V, as are the hazards, $\rho_{T^+} = \rho_V$.

If now we have the representations

$$\rho_V = \rho_{V_0} \exp(\boldsymbol{\beta}_V^T \mathbf{x}), \quad \rho_Z = \rho_{Z_0} \exp(\boldsymbol{\beta}_Z^T \mathbf{x}), \tag{1}$$

then

$$\log\{P(T=0)/P(T>0)\} = (\log \rho_{V_0} - \log \rho_{Z_0}) \; + \; (\boldsymbol{\beta}_V - \boldsymbol{\beta}_Z)^T \mathbf{x}. \tag{2}$$

That is, with exponential V and Z, a direct analysis of T estimates $\boldsymbol{\beta}_V$ and a logistic analysis of the frequency of zero values estimates the log odds ratio $(\boldsymbol{\beta}_V - \boldsymbol{\beta}_Z)$ (for unit changes in $\mathbf{x}$). If it is reasonable to assume that the detection process is independent of $\mathbf{x}$, then $\boldsymbol{\beta}_Z = \mathbf{0}$ and two asymptotically independent estimates of $\boldsymbol{\beta}_V$ are obtained. Subject to their mutual consistency, a mean may be calculated, weighting each contribution inversely by its variance, as estimated from the relevant information matrix.

This simple analysis is based on strong assumptions and we now consider in outline a number of extensions of the analysis.

3.2. Some developments

The nature of the detection process may make the assumption of exponentially distributed Z reasonable and, moreover, it is likely that for most purposes, so long as Z is small compared with V, the precise form of the distribution of Z may not be critical. We therefore continue to assume that Z is exponentially distributed but allow an arbitrary distribution for V. Then provided ρ_Z is relatively large, so that Z is small, and with the probability density of V denoted by $f_V(v)$, we have that

$$\begin{aligned} P(T=0) = P(Z>V) &= \int_0^\infty f_V(v) e^{-\rho_Z v} dv & (3) \\ &= f_V(0)/\rho_Z - f_V'(0)/\rho_Z^2 + \ldots, & (4) \end{aligned}$$

so that if the density of V varies only slowly near the true origin, essentially the previous result is recovered, with

$$\log\{P(T=0)/P(T>0)\} \approx \alpha_0 + (\boldsymbol{\beta}_V - \boldsymbol{\beta}_Z)^T \mathbf{x}. \tag{5}$$

If, however, as may happen in some applications, there is a relatively particularly high risk of failure at very small times, *e.g.* if the distribution of V is Weibull with index less than one, then $f_V'(0)$ will be large and negative and $P(Z>V)$ increases. If that happens then the true log odds ratio will be larger than $(\boldsymbol{\beta}_V - \boldsymbol{\beta}_Z)$.

In the region $V > Z$ in which T is therefore positive, the improper density of T is

$$f_{T^+}(t) = \int_0^\infty f_Z(z) f_V(t+z) dz \tag{6}$$

and if the values of Z are all small this can be written as

$$f_{T^+}(t) = f_V(\tilde{t})\{1 + \frac{1}{2}\sigma_Z^2 f_V''(\tilde{t})/f_V(\tilde{t})\}, \tag{7}$$

where $\tilde{t} = t + \mu_Z$. If Z is exponentially distributed then $\sigma_Z^2 = \mu_Z^2$ and to the first order the consequence of observing T rather than V is to displace the argument of the density, and in fact also the hazard, by μ_Z. If also V is exponentially distributed then there is no change in the hazard, as is clear on general grounds. Then displacement might be of little concern unless Z depends strongly on the explanatory variables $\mathbf{x}$.

A further possibility, usually not assessable directly, is that Z and V are dependent given the explanatory variables $\mathbf{x}$. As an approximation for small levels of dependence we write the joint density of (Z, V) in the form

$$f_Z(z)f_v(v)\{1 + \eta\sigma_Z^{-1}\sigma_V^{-1}(z - \mu_Z)(v - \mu_V)\}, \tag{8}$$

where μ and σ denote mean and standard deviation and $\eta = \text{corr}(Z, V)$. This could be regarded formally as the leading term of an expansion in terms of orthogonal polynomials. Note that here η is assumed sufficiently small that contributions from formally negative values of the density may be ignored. Assuming that Z is marginally exponentially distributed so that $\sigma_Z = \mu_Z$, local dependence can be represented approximately by writing for small η the joint density as

$$\rho_Z e^{-\rho_Z z} f_V(v)\{1 + \eta(\rho_Z z - 1)(\rho_V v - 1)\}, \tag{9}$$

where $\rho_V = 1/\mu_V$. Then

$$P(Z > V) = \int_0^\infty e^{-\rho_Z v}\{1 + \eta v(v - \mu_V)\}f_V(v)dv \tag{10}$$

and this may be evaluated in terms of the moment generating function of V. The integral can be evaluated explicitly if V has a gamma distribution or may be approximated for large ρ_Z/ρ_V. We outline here the case where Z is exponentially distributed and V has a gamma distribution with index δ, that is $\sigma_V/\mu_V = 1/\sqrt{\delta}$. Then, with $\rho_V = 1/\mu_V$ we have that

$$P(T = 0) = P(Z > V) = \delta^\delta(\rho_V/\rho_Z)^\delta(1 + \delta\rho_V/\rho_Z)^{-\delta}\left\{1 - \eta\frac{1 - \rho_V/\rho_Z}{(1 + \delta\rho_V/\rho_Z)^2}\right\}. \tag{11}$$

The leading term shows that for given small values of ρ_V/ρ_Z the value of $P(T = 0)$ decreases with δ. That is, if the distribution of V is relatively more dispersed than the exponential distribution , then $P(T = 0)$ decreases. If this happens then the approximation outlined in (5) would lead to an overestimate of the true log odds ratio and hence an underestimate of $(\boldsymbol{\beta}_V - \boldsymbol{\beta}_Z)$.

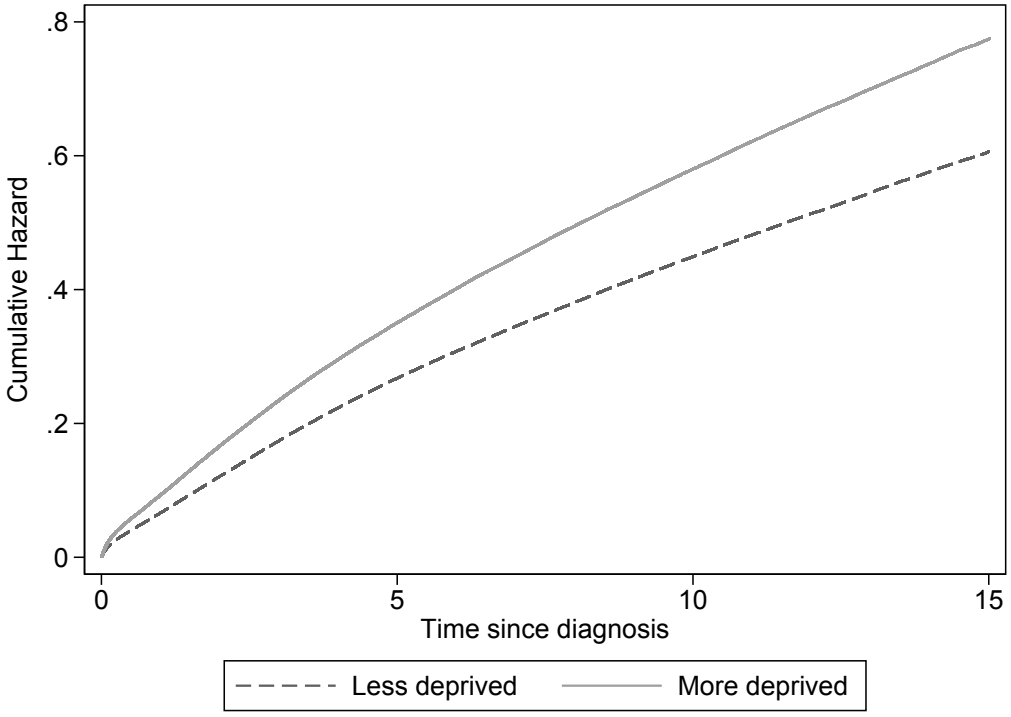

Figure 2. Estimated cumulative hazard function for breast cancer patients.

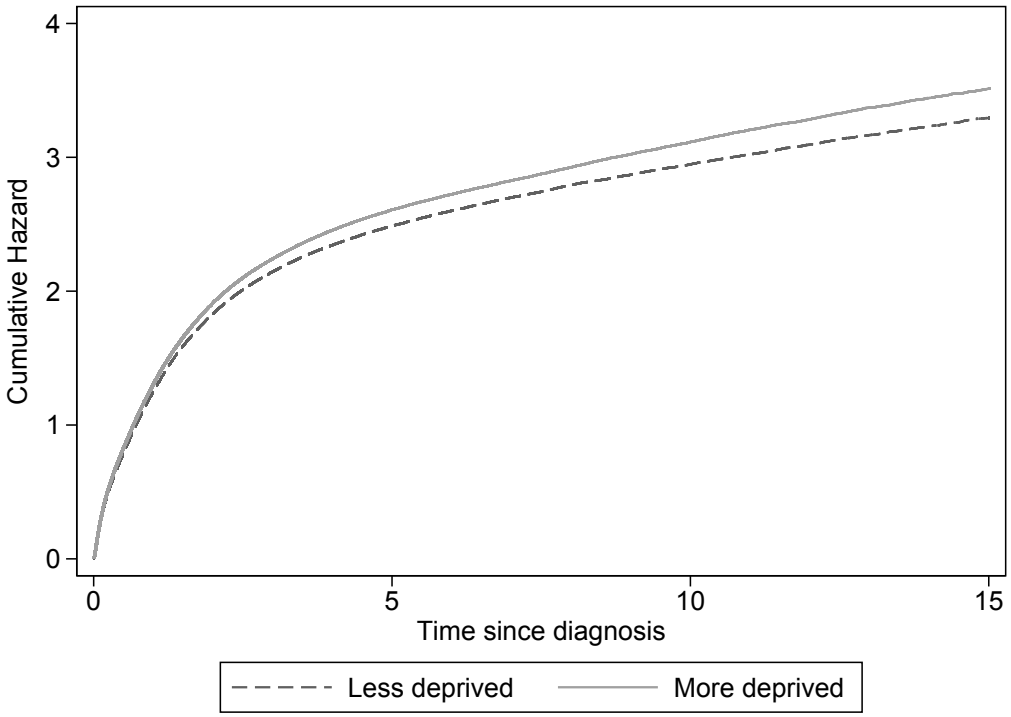

Figure 3. Estimated cumulative hazard function for lung cancer patients.

4. The cancer data revisited

In the cancer data we have shown that both the odds of $T = 0$ and the hazard of death measured on the T^+ scale increase with deprivation score. In the following for simplicity we consider a dichotomy of the deprivation score, corresponding to the top two fifths of the distribution in the general population.

Assuming initially both that Z is exponentially distributed and that V either follows an exponential distribution or its density varies slowly near the time origin, then log(OR) of a death certificate only diagnosis of exposed (higher deprivation score) versus unexposed (lower deprivation score) should give an (approximate) estimate of $(\beta_V - \beta_Z)$. For the breast cancer patients this is 0.1369 (SE=0.0160) and for lung cancer patients it is 0.0259 (SE=0.0087). The corresponding values of log(HR) estimated on the T^+ scale are 0.2598 (0.0043) and 0.0473 (0.0028), respectively.

Formally, comparing these two sets of independent estimates we find for breast cancer that 0.2598 and 0.1369 are statistically significantly different ($z = 7.418, p < 0.001$). For lung cancer the two estimates, 0.0473 and 0.0259, are also statistically significantly different ($z = 2.338, p = 0.01$). Under the assumption that Z and V are both exponentially distributed, these results imply that Z is positively associated with deprivation score, with $\hat{\beta}_Z = 0.1229$ (SE=0.0166) for breast cancer and $\hat{\beta}_Z = 0.0214$ (SE=0.0092) for lung cancer. In other words, time to diagnosis to either cancer is shorter on average when suffering deprivation, especially so for breast cancer cases.

The form of the cumulative hazard functions on the T scale suggests a deceleration of the hazards (Figures 2 and 3) and hence, extrapolating this pattern from T to V, since Z is assumed to be small relative to V, a deceleration of the hazards on the V scale. This implies that $f'_V(v)$ is negative, leading to the true log odds ratio of exposure (for $T = 0$) being larger than $(\beta_V - \beta_Z)$. In this case the calculations above give, in absolute terms, overestimates of β_Z, when $\hat{\beta}_T$ is a good approximation for β_V. For lung cancer, since the log odds ratio above is small, the implication is that $\hat{\beta}_T$ is a good approximation for β_V (see the interpretation of equation (7)). For breast cancer however, an overestimation of β_Z has more substantial consequences for this approximation, with the displacement of the hazard functions on the T^+ scale in exposed and unexposed possibly being less serious than anticipated. Since the size of the bias affecting $\hat{\beta}_T$ when used as an approximation for β_V cannot be deduced directly, we consider a range of values for β_Z in sensitivity analyses using observed T to capture the shape of $f_V(v)$. This is shown in Figure 4 (details in the Web Appendix). There we report mean $\hat{\beta}_V$ obtained in simulations where V is generated as the sum of observed T and a random draw from a variable generated with hazard $\rho_Z = \rho_{Z_0} \exp(\beta_Z x)$, where x is the deprivation indicator, ρ_{Z_0} is set to be either 0.5

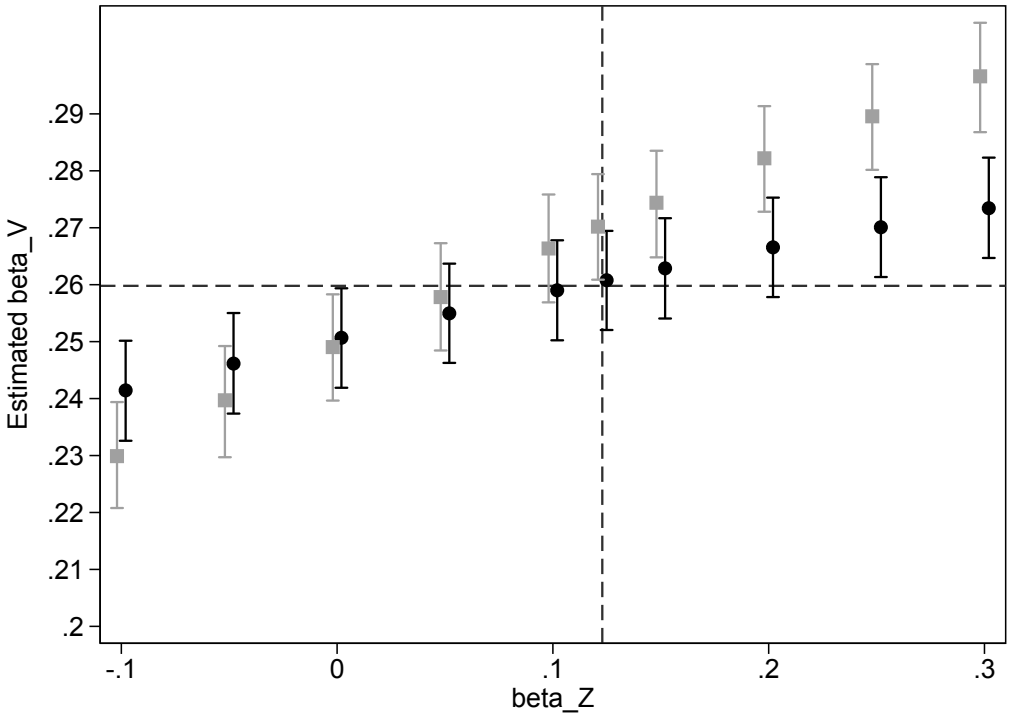

Figure 4. Sensitivity analysis of the effect of deprivation on β_V on breast cancer when β_Z takes different values and $\rho_{Z_0} = 1$ (black) or $= 2$ (grey) (N=660,025; 100 simulations per combination of parameters). The horizontal dotted line depicts $\hat{\beta}_T$ as obtained from the original data and the vertical dotted line depicts $\hat{\beta}_Z$ obtained under the assumption of exponential Z and V.

or 1 (*i.e.* mean time to diagnosis in unexposed two or one year since cancer onset, as is realistic for this example), and β_Z varies from -0.10 to 0.30. We also show the minimum lower bound and maximum upper bound of the 95% confidence intervals for each of these groups of estimates.

When β_Z is set to be 0 there is no differential displacement between exposed and unexposed individuals and $\hat{\beta}_T = 0.2598$ overestimates β_V because of the departure from the exponential distribution in V. When $\beta_Z \gg 0$, $\hat{\beta}_T = 0.2598$ underestimates β_V, more substantially when ρ_{Z_0} is smaller, but not critically if $\beta_Z \sim 0.12$ as suggested by the earlier analyses. In summary, with a positive β_Z, $\hat{\beta}_T$ can be taken as a lower bound for the log hazards ratio of survival by deprivation, β_V.

The analysis has assumed a proportional hazard dependence on the V scale. This can be checked to some extent by studying the dependence of T by censoring the follow-up times at 3,5 and 10 years and examining the effect on the estimates of $\hat{\beta}_T$. For lung cancer the resulting estimates change by less than 2 per cent suggesting that the proportionality assumption is reasonably satisfactory. For breast cancer the changes are systematic from 0.2964 for the data censored at 3 years to 0.2740 and 0.2643 for the other censored data

to 0.2598 for the original data. That is, there is reasonable evidence that the effect of the explanatory variable is relatively greater at short times than at longer times.

Finally, if there were a small positive correlation between Z and V, beyond that due to their common dependence on the deprivation score, and if V was relatively more dispersed than the exponential distribution (*e.g.* followed a gamma distribution), our estimates of the true log odds ratio of deprivation (for a death certificate only diagnosis) would in expectation be smaller than $(\beta_V - \beta_Z)$. Then the underestimation discussed above would be compensated.

5. Concluding remarks

Our aims were to discuss the likely consequences of the bias affecting the hazards ratio of an exposure of interest estimated on the observed time scale T, as opposed to the true time scale V. We give a simple procedure for exploring such a bias. The simplest assumption, that exponential distributions are involved for both Z and V, leads to direct and easily interpreted answers. When the assumed exponential distribution for V is inappropriate we have given alternative more realistic possibilities that focus on the expression for the odds of $T = 0$. More elaborate results would be required if $\mathrm{P}(T = 0)$ were large, as would happen when Z is not small relative to V.

References

Cancer Research UK, Deprivation gradient for cancer incidence (2016) `http://www.cancerresearchuk.org/health-professional/cancer-statistics/incidence/deprivation-gradient#heading-One`, accessed on 27/09/2016.

Carstairs, V. and Morris, R. (1989) Deprivation, mortality and resource allocation. *Community Medicine*, 11 (4), 364–372.

Cox, D.R. and Oakes, D. (1984) *Analysis of Survival Data*. London: Chapman and Hall.

Acknowledgements

We are extremely grateful to Dr Bernard Rachet for alerting us to the problem and granting us access to the cancer data and to the Associate Editor and Reviewers whose comments have led to substantial improvements of this work.

The LSHTM Centre for Statistical Methodology is supported by the Wellcome Trust Institutional Strategic Support Fund, 097834/Z/11/B.

Table 1. Number of breast and lung cancer diagnosis by deprivation index categories (fifths) and type of registration

	Breast cancer			Lung cancer		
		Death			Death	
Deprivation index	All	cert. only		All	cert. only	
category	N	N	Row %	N	N	Row %
1=Least	146,078	3,016	2.06	81,071	7,525	9.28
2	144,204	3,727	2.58	101,578	9,662	9.51
3	139,029	3,851	2.77	119,318	11,971	10.03
4	129,734	3,874	2.99	145,945	14,595	10.00
5=Most	100,980	2,729	2.70	162,017	16,294	10.06
All	660,025	17,197	2.61	609,929	60,047	9.84

Table 2. Odds ratios (ORs)* of being a death certificate only registration by deprivation index categories (in fifths) and cancer type

Deprivation index	Breast cancer		Lung cancer	
category	OR	95%CI	OR	95%CI
1=Least	1	-	1	-
2	1.26	1.20, 1.32	1.02	0.99, 1.06
3	1.35	1.29, 1.42	1.08	1.05, 1.11
4	1.45	1.38, 1.52	1.07	1.04, 1.10
5=Most	1.29	1.22, 1.36	1.07	1.04, 1.10
Linear trend (p-value)	< 0.001		< 0.001	

* ORs estimated by logistic regression adjusted for year of diagnosis and gender (the latter only for lung cancer).

Table 3. Hazard ratios (HRs)* of survival by deprivation index categories (in fifths) and cancer type

Deprivation index	Breast cancer			Lung cancer		
category	N	HR	95%CI	N	HR	95%CI
1=Least	143,062	1	-	73,546	1	-
2	140,477	1.17	1.15, 1.18	91,916	1.05	1.03, 1,06
3	135,178	1.30	1.28, 1.32	107,347	1.07	1.06, 1.09
4	125,860	1.43	1.41, 1.45	131,350	1.09	1.08, 1.10
5=Most	98,251	1.56	1.54, 1.59	145,723	1.10	1.09, 1.11
Linear trend (p-value)		< 0.001			< 0.001	

* HRs estimated by semi-parametric proportional hazards regression stratified by year of diagnosis and gender (the latter only for lung cancer).