Research Article

Statistics in Medicine

Received 13 October 2008, Accepted 15 September 2009 Published online in Wiley Interscience

(www.interscience.wiley.com) DOI: 10.1002/sim.3762

Estimating the crude probability of death due to cancer and other causes using relative survival models

P. C. Lambert,[a,b]*† P. W. Dickman,[b] C. P. Nelson[a] and P. Royston[c]

Relative survival is used extensively in population-based cancer studies to measure patient survival correcting for causes of death not related to the disease of interest. An advantage of relative survival is that it provides a measure of mortality associated with a particular disease, without the need for information on cause of death. Relative survival provides a measure of net mortality, i.e. the probability of death due to cancer in the absence of other causes. This is a useful measure, but it is also of interest to measure crude mortality, i.e. the probability of death due to cancer in the presence of other causes. A previous approach to estimate the crude probability of death in population-based cancer studies used life table methods, but we show how the estimates can be obtained after fitting a relative survival model. We adopt flexible parametric models for relative survival, which use restricted cubic splines for the baseline cumulative excess hazard and for any time-dependent effects. We illustrate the approach using an example of men diagnosed with prostate cancer in England and Wales showing the differences in net and crude survival for different ages. Copyright © 2010 John Wiley & Sons, Ltd.

Keywords: relative survival; competing risks; crude mortality

1. Introduction

Data from population-based cancer registries are used to study survival in cancer patients. However, a patient who is diagnosed with a particular type of cancer is also at risk of dying of other causes. For this reason, survival analysis using population-based cancer registry data aims to estimate net survival, a measure of patient survival corrected for the effect of other causes. Net survival aims to estimate the probability of survival in a hypothetical world where the cancer under study is the only possible cause of death.

The net survival can be estimated using cause-specific analysis, where patients who die of causes other than the cancer of interest are treated like a censoring at their death time, or by relative survival [1]. Relative survival is often the preferred method in population-based cancer studies as it provides a measure of mortality associated with a particular disease, without the need for information on the cause of death, which may either not be recorded or considered to be inaccurately recorded [2].

Relative survival is used extensively in cancer survival for both national and international comparisons [3], changes in survival over calendar time [4] and to explore potential risk factors for increased mortality. However, the net probability of death due to cancer (1-relative survival) as a function of time since diagnosis does not provide a measure of the true probability that a patient will die of their cancer as it estimates a probability in the *absence* of other causes, i.e. in the hypothetical world where the cancer under study is the only possible cause of death. However, cancer patients tend to be old and there will be competing causes of death. Therefore, it is also of interest, to both the patient and treating clinician, to estimate the probability of death due to cancer in the *presence* of other causes as a function of time since diagnosis. This is defined as the crude probability of death due to cancer. It is far less common to report measures of crude probabilities in population-based cancer studies. The crude

[a]*Department of Health Sciences, Centre for Biostatistics and Genetic Epidemiology, University of Leicester, 2nd Floor, Adrian Building, University Road, Leicester LE1 7RH, U.K.*

[b]*Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden*

[c]*Hub for Trials Methodology Research, MRC Clinical Trials Unit and UCL, 222 Euston Road, NW1 2DA, London*

**Correspondence to: P. C. Lambert, Department of Health Sciences, Centre for Biostatistics and Genetic Epidemiology, University of Leicester, 2nd Floor, Adrian Building, University Road, Leicester LE1 7RH, U.K.*

†*E-mail: paul.lambert@le.ac.uk*

Contract/grant sponsor: Swedish Cancer Society
Contract/grant sponsor: Swedish Research Council

probability of death due to cancer will be particularly useful when making decisions about treatments with potentially severe side effects or for the planning of future health-care services.

Both the net and the crude probabilities discussed above come from the theory of competing risks [5] and are useful measures, but answer different research questions. For individuals diagnosed with cancer, the net probability of death due to cancer is important when making comparisons over time or between places as mortality due to other causes also varies over time and between places and it is important that the estimates of survival/mortality are not influenced by these mortality changes in other diseases. The crude probability of death due to cancer gives the probability of death due to cancer in the real world where the probability of dying from cancer will vary depending on the risk of dying from the full spectrum of other potential causes of death. It is also possible to calculate the crude probability of death due to other causes in the presence of cancer mortality. The sum of these two probabilities gives the total crude probability of death for individuals diagnosed with cancer. It should be noted that terminology in this area is inconsistent. For example, the crude probability of death is also known as the cumulative incidence [6]. Here, we use the terms crude and net probability of death in a manner consistent with the work of Cronin and Feuer [7] from work in relative survival and Tsiatis [5] from definitions in competing risks methodology.

Cronin and Feuer [7] showed how the crude probability of death due to cancer and other causes can be calculated from life tables for population-based cancer studies. In their approach, subjects were divided into large age groups and the crude probability of death due to cancer and due to other causes were calculated in yearly intervals. A potential problem with this is that mortality may increase dramatically between the lower and the upper boundaries of these age groups. For example, using the expected mortality rates given by Coleman *et al.* [8], the expected probability of death within 10 years is approximately 0.43 for a 70 year old and 0.74 for a 79 year old. In addition, it is not possible to obtain model-based estimates from the life table approach as a separate analysis needs to be performed for each covariate pattern of interest.

This paper shows how the crude probability of death due to cancer and due to other causes can be calculated after fitting a relative survival model to individual patient data. Relative survival models can provide estimates of the net probability of death due to cancer at the individual level. However, the crude probability of death due to cancer and due to other causes can be derived from these models through numerical integration of functions derived from relative survival models. This is illustrated using men diagnosed with prostate cancer in England and Wales between 1986–1988 and followed up for 10 years. We use and further develop the relative survival models proposed by Nelson *et al.* [9] that model on the log cumulative excess hazard scale using restricted cubic splines for the baseline excess hazard and time-dependent effects. Since age at diagnosis is a strong predictor of mortality due to other causes, age is the main factor that will lead to differences between net and crude probability of death. For example, it is possible for relative survival, and hence net mortality, to be identical between age groups, but there will still be differences in crude mortality as the older age groups are more likely to die of other causes. We demonstrate the methods by modeling age at diagnosis as a continuous covariate using restricted cubic splines.

The remainder of the paper is laid out as follows: Section 2 describes relative survival, the flexible parametric models on the log cumulative excess hazard scale and how we use the estimates from these models to estimate the crude probability of death. Section 3 describes the prostate cancer example and implements the methods on this data set. Finally, Section 4 discusses the methods and potential areas of future development.

2. Methods

2.1. Relative survival

In relative survival models, the overall (all-cause) survival function, $S(t)$, is the product of the expected survival function, $S^*(t)$, and the relative survival function, $R(t)$, where t is the time since diagnosis of cancer

$$S(t) = S^*(t)R(t) \tag{1}$$

The net probability of death due to cancer is estimated by $1-R(t)$. Converting equation (1) to the hazard scale implies that the overall hazard function, $h(t)$, is the sum of two components, the expected hazard function, $h^*(t)$ and the excess hazard function, $\lambda(t)$,

$$h(t) = h^*(t) + \lambda(t) \tag{2}$$

Both $S^*(t)$ and $h^*(t)$ are assumed known and are usually obtained from routine data sources by matching on age, sex and year of diagnosis and potentially other variables. The main driver of the expected hazard function is age. An important assumption when interpreting relative survival as a net probability is that the mortality associated with the disease of interest is independent of the mortality associated with other causes conditional on covariates, i.e. there are independent competing risks [10].

It is possible for the excess hazard, $\lambda(t)$, to be negative. In such a situation the relative survival would increase. This could happen if those surviving a certain number of years had a lower mortality rate than that expected in the general population. However, this rarely occurs in practice other when deaths are underascertained.

2.2. Relative survival models

Most models for relative survival are applied on the log excess hazard scale, i.e.

$$h(t) = h^*(t) + \lambda_0(t)\exp(\mathbf{x}\beta)$$

This is a proportional excess hazards model with baseline excess hazard, λ_0. The estimated parameters are log excess hazard ratios. Most relative survival models split the time scale into a number of intervals in order to fit piecewise constant effects for the baseline excess hazard rate [11–13]. These models can be extended to fit non-proportional excess hazard models by fitting interactions between the piecewise time intervals and the covariate(s) of interest. There has been interest in modeling both the baseline excess hazard rate and any time-dependent covariate effects continuously using splines [14–16] or fractional polynomials [17]. These models are an approximation to continuous time as they involve splitting the time scale into a number of small intervals, e.g. monthly intervals, or using numerical integration. This leads to the models being slow to fit on individual level data or require the user to collapse the data over covariate patterns, which precludes continuous covariates.

2.3. Models on the cumulative excess hazard scale

An alternative approach to modeling relative survival is to fit models on the log cumulative excess hazard scale [9]. The advantage of these models over those described above is that it is not necessary to split the time scale or to use numerical integration and that continuous covariates are easy to incorporate. The model is an extension to relative survival of the flexible parametric models for survival analysis developed by Royston and Parmar [18] where restricted cubic splines are used to estimate the baseline log cumulative hazard.

Let $H(t)$ denote the total cumulative hazard, $H^*(t)$ denote the expected cumulative hazard and $\Lambda(t)$ denote the excess cumulative hazard. Integrating equation (2) gives

$$H(t)=H^*(t)+\Lambda(t)$$

Thus we model on the log cumulative excess hazard scale, and introducing covariates, $\mathbf{x}$, write the log cumulative hazard as,

$$\ln(\Lambda(t))=\ln(\Lambda_0(t))+\mathbf{x}\boldsymbol{\beta}$$

This is a proportional excess hazards model as proportional excess hazards also implies proportional cumulative excess hazards. This is important as when proportional cumulative excess hazards are assumed the interpretation of the $\boldsymbol{\beta}$'s is exactly the same as the $\boldsymbol{\beta}$'s from a proportional excess hazards model. The log baseline cumulative excess hazard, $\Lambda_0(t)$, is modeled using restricted cubic splines [19] for $\ln(t)$ with knots located at $\mathbf{k}$, the vector of knot positions. Thus, the log cumulative excess hazard can be written as

$$\ln[\Lambda(t|\mathbf{x})]=\eta=s(\ln(t)|\gamma,\mathbf{k})+\mathbf{x}\boldsymbol{\beta} \tag{3}$$

The linear predictor, η, thus consists of a restricted cubic spline function, $s(\ln(t)|\gamma,\mathbf{k})$, where γ is the vector of parameters associated with the spline variables, and the additive effect of covariates, $\mathbf{x}\boldsymbol{\beta}$. The restricted cubic splines function, $s(\ln(t)|\gamma,\mathbf{k})$, is fitted by including $K-1$ derived variables. For knots, $k_1,\ldots,k_K$, a restricted cubic spline function can be written as

$$s(x)=\gamma_0+\gamma_1 z_1+\gamma_2 z_2+\cdots+\gamma_{K-1}z_{K-1}$$

The derived variables z_j (also known as the basis functions) are calculated as follows:

$$z_1=x$$

$$z_j=(x-k_j)_+^3-\phi_j(x-k_1)_+^3-(1-\phi_j)(x-k_K)_+^3,\quad j=2,\ldots,K-1$$

where

$$\phi_j=\frac{k_K-k_j}{k_K-k_1}$$

and $(u)_+=u$ if $u>0$ and 0 if $u\leqslant 0$. Thus, a model with K knots for the baseline cumulative hazard uses $K-1$ degrees of freedom.

The derived variables can be highly correlated and it can be useful to transform them using Gram–Schmidt orthogonalization [20].

The relative survival function and the excess hazard function can be obtained by the transformation of the model parameters, with the relative survival function obtained using

$$R(t)=\exp(-\exp(\eta)) \tag{4}$$

and the excess hazard function obtained using

$$\lambda(t)=\frac{s(\ln(t)|\gamma,\mathbf{k})}{\mathrm{d}t}\exp(\eta) \tag{5}$$

This involves the derivative of the restricted cubic splines function. The derivative of a restricted cubic spline function, $s(x)$, is calculated using

$$s'(x)=\gamma_1 z_1'+\gamma_2 z_2'+\cdots+\gamma_{K-1}z_{K-1}'$$

where

$$z_1' = 1$$

$$z_j' = 3(x-k_j)_+^2 - 3\phi_j(x-k_1)_+^2 - 3(1-\phi_j)(x-k_K)_+^2$$

Note that there are no constraints for the cumulative excess hazard to be monotonic, i.e. it is possible to estimate a negative excess hazard. However, these situations are very rare in practice and we have not yet encountered a data set where the flexible parametric approach has estimated a cumulative excess hazard with a turning point. However, we see the ability of the approach to the estimated negative excess hazards if they exists as an advantage.

The models can be extended to time-dependent effects by introducing a new set of knots, $\mathbf{k}_m$ for the mth time-dependent effect. If there are D time-dependent effects then equation (3) can be extended as follows:

$$\ln[\Lambda(t|\mathbf{x})] = s(\ln(t)|\boldsymbol{\gamma}, \mathbf{k}_0) + \sum_{m=1}^{D} s(\ln(t)|\boldsymbol{\delta}_m, \mathbf{k}_m)x_m + \mathbf{x}\boldsymbol{\beta}$$

The number of spline variables for a particular time-dependent effect will depend on the number of knots, $\mathbf{k}_m$. For any time-dependent effect, there is an interaction between the covariate and the spline variables. This is a further development of the relative survival models proposed by Nelson *et al.* [9] and the survival models of Royston and Parmar [18] as these previous models forced any time-dependent effects to have the same number of knots at the same locations as the baseline effect. This tended to lead to over parameterization of time-dependent effects and the approach outlined above will usually lead to more parsimonious models being fitted.

The time dependence is modelled on the log cumulative excess hazard scale, but it is usually of more interest to report on the (log) excess hazard scale. Log excess hazard ratios are a non-linear function of the model parameters and standard errors are obtained using the delta method.

Non-linear effects of continuous variables can be incorporated into a proportional excess hazards model using restricted cubic splines. In addition, time dependence of non-linear effects can be incorporated by forming interactions between the spline variables for the continuous covariate and the spline variables for time-dependent effects. This has the effect of making the coefficients for the spline variables for the continuous covariate vary as a function of time [21].

2.4. Estimation

The contribution to the log-likelihood for the ith individual for a relative survival model can be written as

$$\ln L_i = d_i \ln[h^*(t_i) + \lambda(t_i)] + \ln[S^*(t_i)] + \ln[R(t_i)] \tag{6}$$

where d_i is the event indicator. The term $\ln[S^*(t_i)]$ does not depend on the model parameters and can be excluded from the likelihood. This means that to fit these models the expected mortality rate, $h^*(t_i)$, at time of death, t_i is incorporated into the data for each subject. The expected (or background) mortality rate for each individual is treated as known.

Thus, by substituting equations (4) and (5) into (6), the contribution to the log-likelihood for the ith individual is

$$\ln L_i = d_i(h^*(t_i) + \ln[s'(\ln(t_i)|\boldsymbol{\gamma}, \mathbf{k}_0)] + \eta_i) - \exp(\eta_i)$$

This likelihood can be maximized using Stata's optimizer, `ml`. The maximization is implemented by defining an additional equation for the derivatives of the spline function and constraining the parameters to be equal to the equivalent spline functions in the main linear predictor. The models are implemented in a Stata command, `stpm2` [22].

2.5. Estimating the crude probability of death

The crude probability of death due to cancer, $\mathrm{Cr_c}$, and due to other causes, $\mathrm{Cr_o}$, are calculated after fitting the relative survival model using standard competing risks definitions [5].

The crude probability of death due to cancer is defined as

$$\mathrm{Cr_c}(t) = \int_0^t S^*(u)R(u)\lambda(u)\,\mathrm{d}u \tag{7}$$

This gives the probability of dying from the cancer of interest by time t in the presence of the competing risk of death due to other causes.

Similarly, the crude probability of death due to other causes is defined as

$$\mathrm{Cr_o}(t) = \int_0^t S^*(u)R(u)h^*(u)\,\mathrm{d}u \tag{8}$$

This gives the probability of dying of causes other than the cancer of interest by time t in the presence of cancer mortality.

$S^*(t)$ and $h^*(t)$ are considered to be known (for a particular covariate pattern). As we are interested in individual level predictions, $S^*(t)$, $h^*(t)$, $R(t)$ and $\lambda(t)$ will vary depending on the covariate pattern. The total crude probability of death by time t is the sum of $Cr_c(t)$ and $Cr_o(t)$.

The integrals generally need to be obtained numerically. The integrands in equations (7) and (8) are non-linear functions of the model parameters. The integration is performed using a similar method to that proposed by Carstensen [23] in the following steps:

1. The time scale is split into a large number, n, of small intervals. For example, 1000 intervals.
2. For a particular covariate vector, $\mathbf{x}_0$, the predicted value of the integrand, (7) or (8) at each of the j time intervals, t_j, is calculated, $\hat{f}(t_j|\mathbf{x}_0)$.
3. The variance–covariance matrix of $\hat{f}(t_j|\mathbf{x}_0)$ is obtained using the delta method. In order to do this, the observation-specific derivatives for each parameter in the model need to be calculated. These are calculated numerically using the Stata command `predictnl` [24]. Let $\mathbf{G}$ be the $n \times p$ matrix of observation-specific derivatives. The variance–covariance matrix of $\hat{f}(t_j|\mathbf{x}_0)$ is

$$\mathrm{Var}(\hat{f}(t_i)) = \mathbf{G}\hat{\mathbf{V}}\mathbf{G}'$$

   where $\hat{\mathbf{V}}$ is the estimated variance matrix for the model parameters.
4. The crude probability of death is calculated by summing the values of the integrand for the n time intervals. This is done by creating a triangular matrix L. For example, for five intervals this looks like

$$\mathrm{Cr}(t) = \ell \times \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} \hat{f}(t_1) \\ \hat{f}(t_2) \\ \hat{f}(t_3) \\ \hat{f}(t_4) \\ \hat{f}(t_5) \end{bmatrix} = \mathbf{L} \begin{bmatrix} \hat{f}(t_1) \\ \hat{f}(t_2) \\ \hat{f}(t_3) \\ \hat{f}(t_4) \\ \hat{f}(t_5) \end{bmatrix}$$

   where ℓ is the interval length.
5. The variance–covariance matrix for the crude probability is then calculated using

$$\mathrm{Var}(\hat{\mathrm{Cr}}) = \mathbf{LG}\hat{\mathbf{V}}\mathbf{G}'\mathbf{L}'$$

We compare the confidence intervals obtained using the above method, with those obtained using bootstrapping [25].

3. Application to cancer of the prostate in England and Wales

3.1. Description of the data

Data were obtained from the public-use data set of all England and Wales cancer registrations between 1 January 1971 and 31 December 1990 with follow-up to 31 December 1995 [8]. We present results for 28 943 males with cancer of the prostate aged between 40 and 90 and diagnosed between 1986 and 1988 inclusive with follow-up to the end of 1995. There were a total of 24 432 deaths. Background mortality rates were obtained from England and Wales national mortality statistics by age, geographical region, period of diagnosis and deprivation group [26].

3.2. Proportional excess hazards model

A proportional hazards model was fitted using 6 knots, which uses 5 df. The knots were placed at the 0th, 20th, 40th, 60th, 80th and 100th centiles of the uncensored log event times. Age was included as a continuous covariate using restricted cubic splines with 4 knots (3 df) at the 0th, 33rd, 67th and 100th centiles of the age distribution. Figure 1(a) shows the estimated excess hazard rates, for selected ages, namely 45, 55, 65, 75 and 85. The predicted excess hazard rate for an 85 year old is notably higher than the other ages. Figure 1(b) shows the estimated excess hazard ratio as a function of age at diagnosis with age 65 as the reference. This shows that there is some evidence that the excess hazard ratio is increased for those diagnosed at a younger age, but the confidence interval is wide due to the small numbers. From about 70 years there is a large increase in the excess hazard ratio with those aged 90 having a 2.25 times higher excess mortality rate than those aged 65.

3.3. Non-proportional excess hazards model

The model was extended to include non-proportional effects by incorporating new restricted cubic spline variables of $\ln(t)$ with 4 knots at the 0th, 33rd, 67th and 100th centiles of the uncensored survival times and forming interactions between these derived variables and the derived spline variables for age. This is an interaction between the terms representing the non-linear effects of

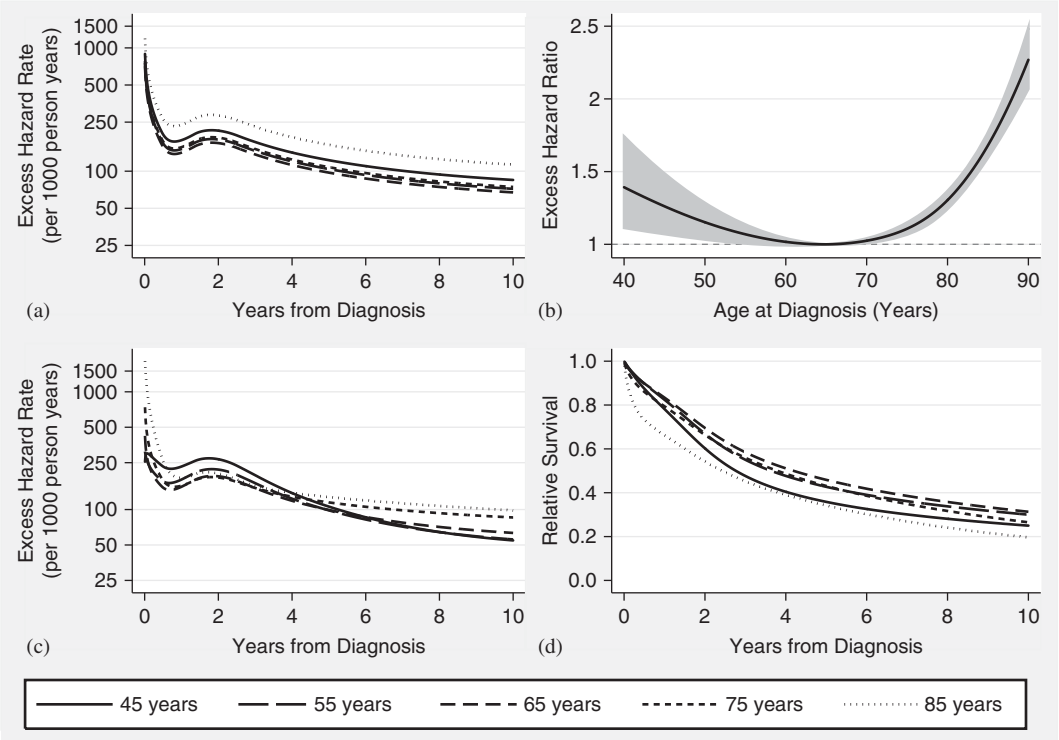

Figure 1. Cancer of the prostate example: (a) predicted excess hazard rate for 5 selected ages from a proportional excess hazards model; (b) predicted excess hazard ratio (with 95 per cent confidence interval) for age with age 65 as the reference from a proportional hazards model; (c) predicted excess hazard rate for 5 selected ages from a non-proportional excess hazards model; and (d) predicted relative survival from a non-proportional excess hazards model.

two continuous variables, with three parameters for the non-linear effect of age and three parameters for the time dependence of age. This makes a total of $3\times3=9$ additional parameters to model the time dependence of age. The likelihood ratio test gives $\chi^2_9=355.19$, $P<0.001$. Figure 1(c) shows the estimated excess hazard rates from the non-proportional hazards model. These are noticeably different from Figure 1(a) where the hazard rates were forced to be proportional. There is a greater difference between the selected ages early on in the time scale. The most notable difference is for the older subjects. Figure 1(d) shows the estimates of relative survival. This indicates that there are broadly similar profiles for ages 55, 65 and 75, but a different pattern for ages 45 and 85.

Figure 2 shows the estimated time-dependent excess hazard ratios for 45, 55, 75 and 85 year old compared with a 65 year old. The excess hazard ratio for age 45 has a wide confidence interval due to the small number of subjects, although there is some evidence of an increased risk in the first couple of years. A 55 year old has a similar excess hazard rate to 65 year old as the excess hazard ratio is close to 1. For ages 75 and 85 there is an increase in risk in the first year after diagnosis when compared with age 65, this is particularly noticeable in the 85 year old.

3.4. Crude probability of death

Figure 3(a) shows the estimated net probability of death due to cancer, i.e. $1-R(t)$ and Figure 3(b) shows the estimated crude probability of death due to cancer obtained using equation (7) for men aged 45, 55, 65, 75 and 85. This shows that as age increases there is a greater difference between the estimated net and the crude probabilities. This is because the competing risks of other causes will clearly increase with age. Thus, an 85 year old has the highest net probability of death due to cancer, but the lowest crude probability of death due to cancer.

Figure 4 shows the estimated crude probabilities of death due to cancer and due to other causes as well as the total crude probability of death for each of the selected ages. The total crude probability of death is the sum of the two crude probabilities. These graphs clearly show how the crude probability of death due to other causes increases with age and the effect this has on the crude probability of death due to cancer. Confidence intervals for the crude probability of death due to cancer are also shown and illustrate the greater amount of uncertainty in the estimated probabilities for younger ages, which is due to prostate cancer being a rare disease in younger men.

Figure 5 shows the estimated crude probability of death due to cancer for those aged 45 and 55 with 95 per cent confidence intervals calculated using the delta method and also by using bootstrapping with 1000 replications. The bias-corrected method was used to calculate the bootstrapped confidence intervals [25]. The upper and lower bounds of the confidence intervals are very similar for the two methods. The equivalent plots for the other ages showed similar agreement.

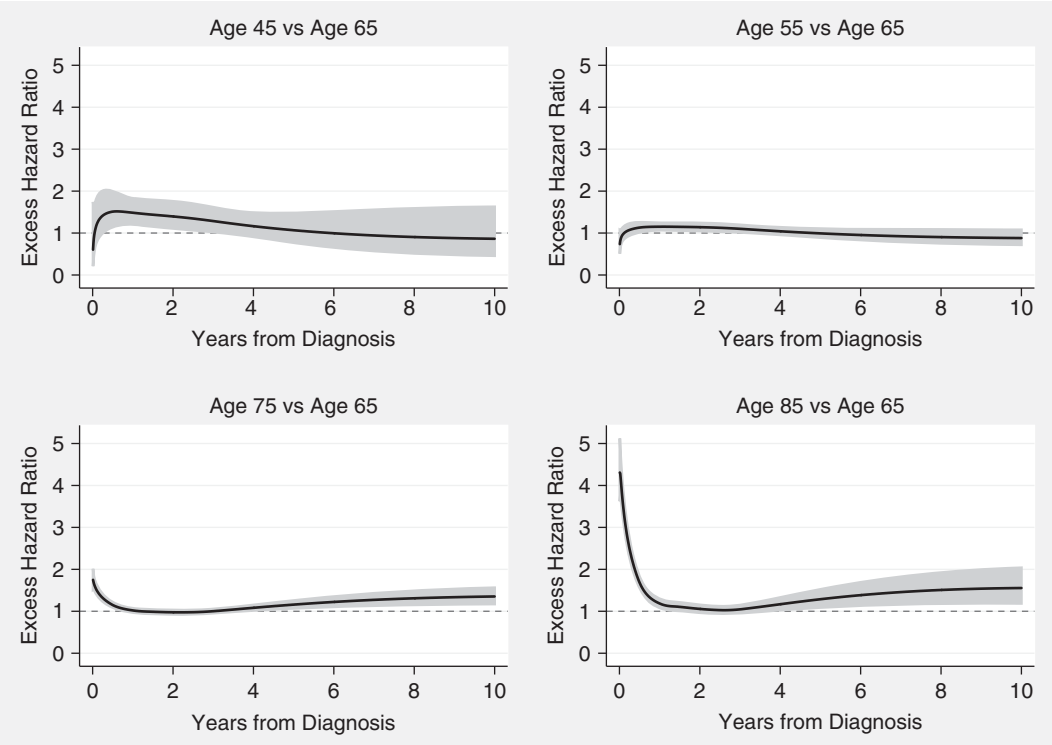

Figure 2. Time-dependent excess hazard ratios for ages 45, 55, 75, 85 compared with those age 65 from a non-proportional hazards model with 5 df for the log cumulative excess hazard and 12 df for the time-dependent effects for age. 95 per cent confidence intervals are shown by the shaded area. See Figure 1 for the predicted excess hazard rate for those aged 65.

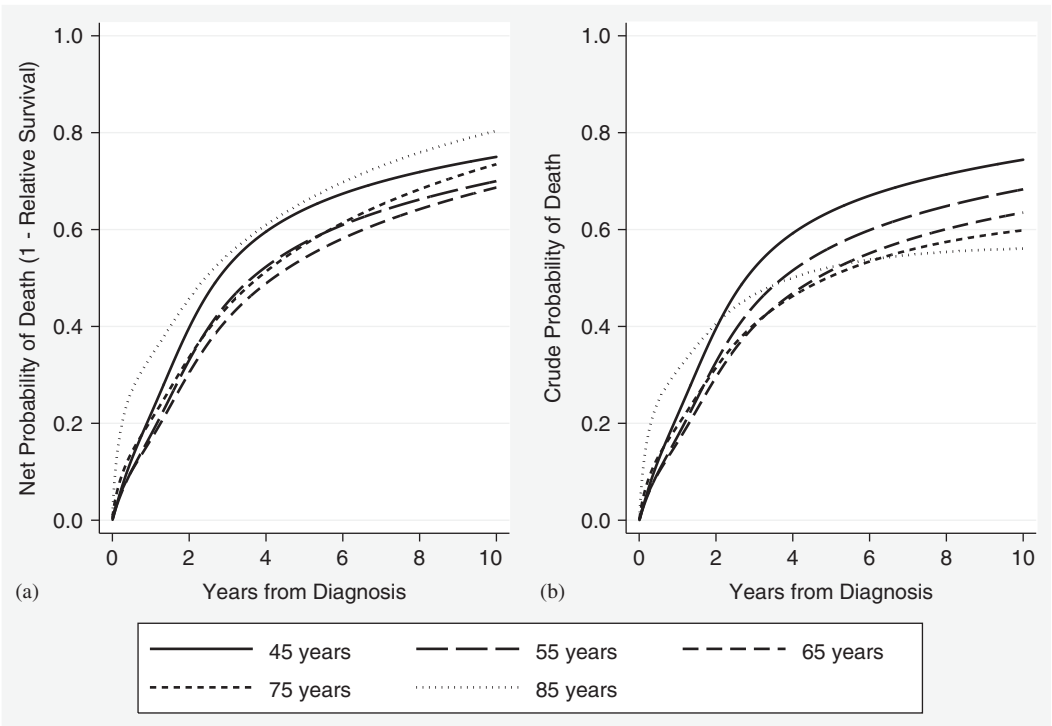

Figure 3. Net probability of death, $1-R(t)$, and crude probability of death due to cancer for ages 45, 55, 65, 75 and 85 from a non-proportional hazards model with 5 df for the log cumulative excess hazard and 12 df for the time-dependent effects for age.

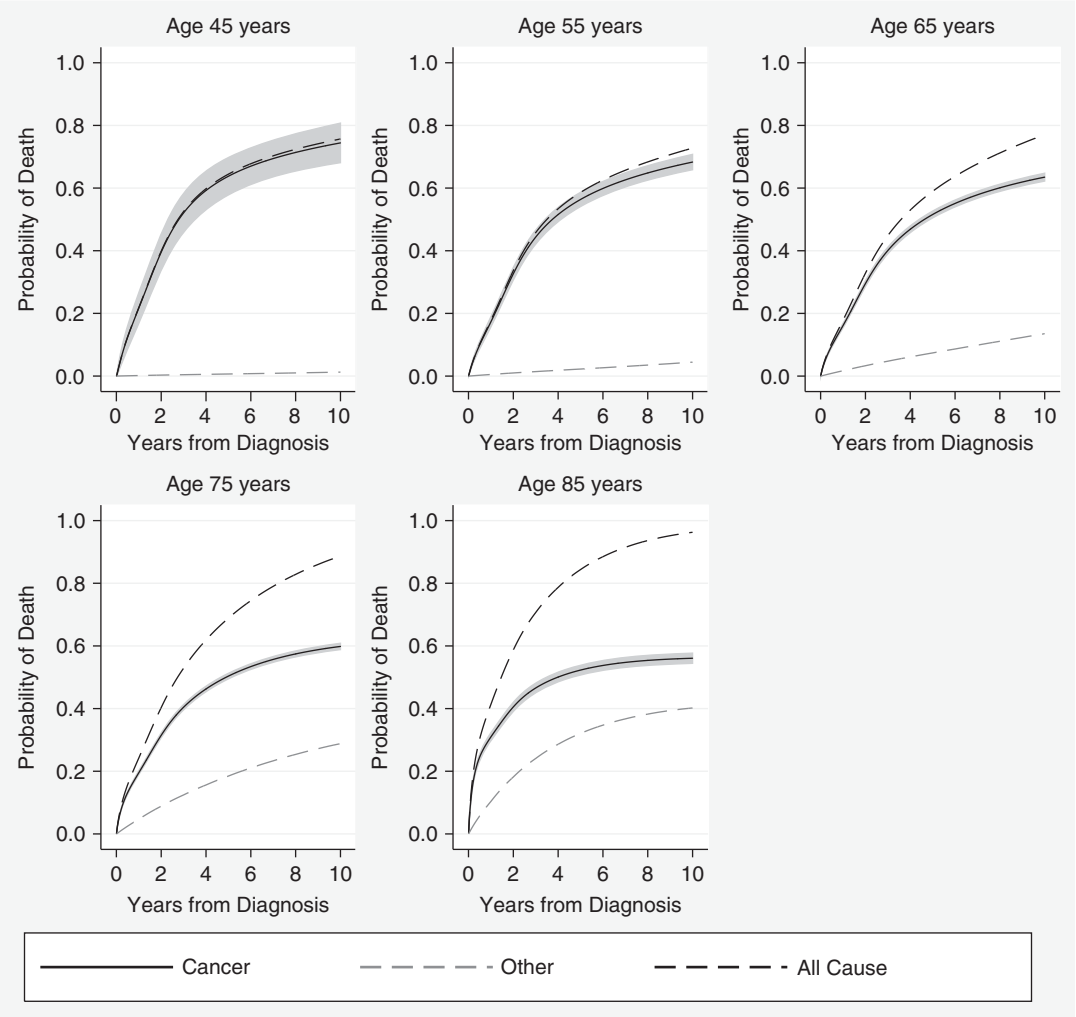

Figure 4. Crude probability of death due to cancer and due to other causes and total crude probability of death for ages 45, 55, 65, 75 and 85 from a non-proportional excess hazards model with 5 df for the log cumulative excess hazard and 12 df for the time-dependent effects for age.

3.5. Sensitivity analysis

As a sensitivity analysis, four further models were fitted that compared the number and locations of the knots for the baseline log cumulative excess hazard, the main effect of age at diagnosis and the time-dependent effect of age at diagnosis. Model (a) is the time-dependent model that produced the estimated crude probabilities in Figure 4. Table I shows four further models with varying degrees of freedom for the baseline log cumulative excess hazard, df_b, the main effect of age at diagnosis, df_a and the time-dependent effect of age, df_t (age $\times$ time interaction). The number of parameters for the baseline hazard is equal to the df_b+1 (including the constant), there are a further df_a+df_t parameters to model the time-dependent effect of age.

In terms of the AIC and BIC, the best-fitting model from this limited set of models is 8 df for the baseline log cumulative excess hazard, 5 df for the main effect of age at diagnosis and 5 df for the time-dependent effect of age at diagnosis. However, Figure 6 shows that the estimated crude probability of death due to cancer for the 5 selected ages is very similar for the 5 different models. It is only for the youngest age that a noticeable difference appears, which given the smaller numbers and the size of the confidence interval in Figure 4 is perhaps not surprising.

4. Discussion

We have shown how the crude probability of death due to cancer and due to other causes can be calculated in population-based cancer studies after fitting a relative survival model. We see the reporting of these crude probabilities as a complement to the more usual net probability of death (or more usually net survival). The two measures help to answer different questions, but the crude probability of death due to cancer measures will generally be of more interest to individual patients and for future planning of services.

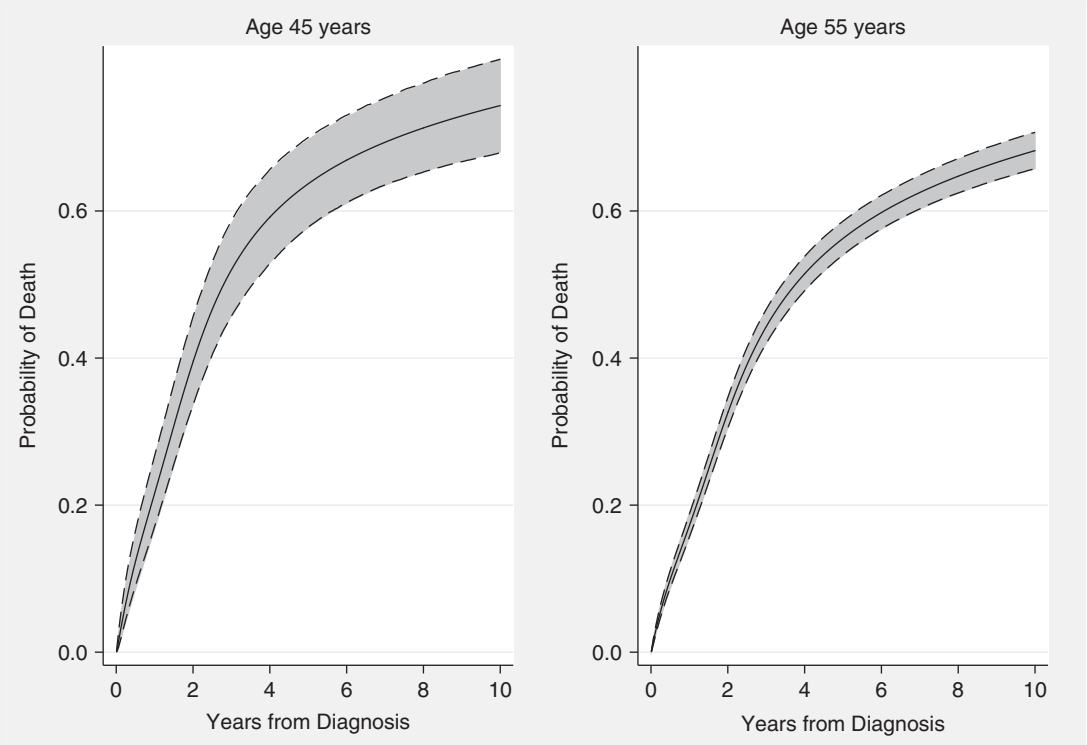

Figure 5. Comparison of 95 per cent confidence intervals for the crude probability of death due to cancer using the delta method (shaded area) and bootstrapping (dashed lines).

Table I. Models with varying degrees of freedom for the baseline log cumulative excess hazard, df_b, the main effect of age at diagnosis, df_a and the time-dependent effect of age.

Model	Baseline df_b	Time-dependent df_t	Age df_a	No. of parameters	AIC	BIC
Model (a)	5	3	3	18	97250.11	97399.02
Model (b)	8	5	5	39	97059.30	97381.95
Model (c)	5	5	3	24	97235.68	97434.23
Model (d)	3	3	3	16	97447.35	97579.72
Model (e)	8	8	8	81	97105.8	97775.92

For 3 df knots are placed at centiles (0, 33, 67, 100), for 5 df at centiles (0, 20, 40, 60, 80, 100) and for 8 df at centiles (0, 12.5, 25, 37.5, 50, 62.5, 75, 87.5, 100). For the baseline log cumulative excess hazard and time-dependent effects, these are placed on the distribution of uncensored event times.

The main advantage of our method over the life table method of Cronin and Feuer [7] is that it provides predictions at the individual level. This is particularly important for older patients as the probability of death due to other causes increases dramatically and thus predictions can vary noticeably over the range of a potential age grouping used for a life table analysis. The prediction at the individual level is important, for example the use of the crude probability of death will be more useful than the net probability of death in the choice of whether to take a treatment with potential severe side effects. However, for predictions at the individual level to be clinically useful, it will be important to record and model as many relevant covariates as possible. Of particular interest will be to include stage of disease as this a strong predictor of survival.

We have demonstrated the calculations of crude probabilities of death after fitting an extension of the flexible parametric models of Nelson *et al.* [9]. These models have the advantage of not having to split the time scale or to use numerical integration methods and are a natural way to model continuous covariates. The models are parametric and hence it is straightforward to obtain smooth predictions of the quantities needed for the numerical integration when calculating the crude probabilities of death. The models can be criticized as the number and location of the knots are subjective, but the sensitivity analysis in Section 3.5 indicates that the overall conclusions do not vary according to the knot location. This has also been reported elsewhere [18, 22, 9]. We recommend that such a sensitivity analysis should be performed when fitting these models. However, given that these models are likely to be applied to the large data sets and are of a descriptive nature, we do not see the

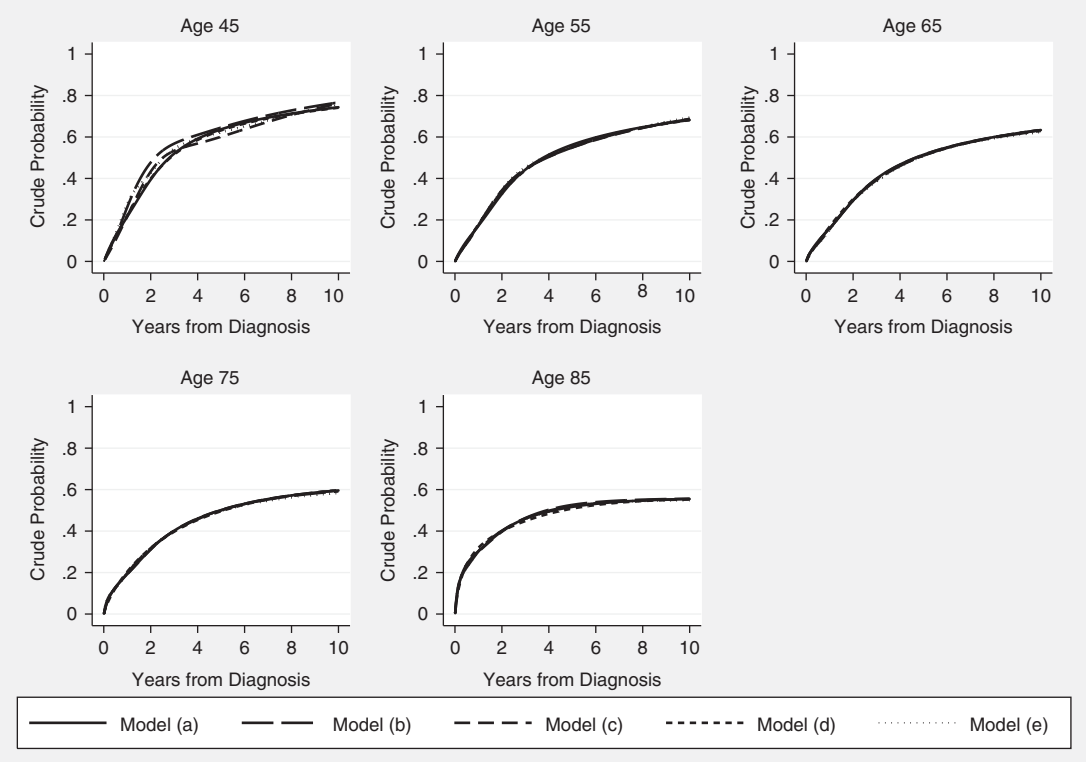

Figure 6. Comparison of estimated crude probability of death due to cancer for models with varying number of knots.

knot selection as the major issue that it may be in smaller data sets [27]. Further work is needed in the area of selecting the number and location of knots, for example via a simulation study [21, 27]. An alternative to using restricted cubic splines would be to use fractional polynomials [28]. This approach is likely to require fewer parameters than the restricted cubic splines. Equations (7) and (8) could be applied after fitting other relative survival models. Of particular interest is the use of these equations after fitting cure models [29].

Finally, relative survival assumes that the competing risks of death due to cancer and due to other causes are independent. Further work should address under what situations and for which particular cancer sites the assumption may not be reasonable and the impact this has on the estimate of the crude probability of death due to cancer.

Acknowledgements

The part of this work was carried out when the first author visited Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden a visit funded by the Swedish Cancer Society (Cancerfonden) and the Swedish Research Council.

References

1. Ederer F, Axtell LM, Cutler SJ. The relative survival rate: a statistical methodology. *National Cancer Institute Monograph* 1961; **6**:101–121.
2. Begg CB, Schrag D. Attribution of deaths following cancer treatment. *Journal of the National Cancer Institute* 2002; **94**:1044–1045.
3. Coleman MP, Quaresma M, Berrino F, Lutz J-M, De Angelis R, Capocaccia R, Baili P, Rachet B, Gatta G, Hakulinen T, Micheli A, Sant M, Weir HK, Elwood JM, Tsukuma H, Koifman S, Silva GAE, Francisci S, Santaquilani M, Verdecchia A, Storm HH, Young JL and The CONCORD Working Group. Cancer survival in five continents: a worldwide population-based study (CONCORD). *Lancet Oncology* 2008; **9**:730–756.
4. Dickman PW, Adami HO. Interpreting trends in cancer patient survival. *Journal of Internal Medicine* 2006; **260**(2):103–117.
5. Tsiatis AA. Competing risks. *Encyclopedia of Biostatistics* (2nd edn). Wiley: New York, 2005; 824–834.
6. Klein JP, Moeschberger ML. *Survival Analysis*: *Techniques for Censored and Truncated Data* (2nd edn). Springer: Berlin, 2003.
7. Cronin KA, Feuer EJ. Cumulative cause-specific mortality for cancer patients in the presence of other causes: a crude analogue of relative survival. *Statistics in Medicine* 2000; **19**(13):1729–1740.
8. Coleman MP, Babb P, Damiecki P, Grosclaude P, Honjo S, Jones J, Knerer G, Pitard AJQ, Sloggett A, De Stavola B. *Cancer Survival Trends in England and Wales, 1971–1995*: *Deprivation and NHS Region*. Office for National Statistics: London, 1999.
9. Nelson CP, Lambert PC, Squire IB, Jones DR. Flexible parametric models for relative survival, with application in coronary heart disease. *Statistics in Medicine* 2007; **26**(30):5486–5498.
10. Gamel JW, Vogel RL. Non-parametric comparison of relative versus cause-specific survival in surveillance, epidemiology and end results (SEER) programme breast cancer patients. *Statistical Methods in Medical Research* 2001; **10**(5):339–352.

11. Dickman PW, Sloggett A, Hills M, Hakulinen T. Regression models for relative survival. *Statistics in Medicine* 2004; **23**:41–64.
12. Esteve J, Benhamou E, Croasdale M, Raymond L. Relative survival and the estimation of net survival: elements for further discussion. *Statistics in Medicine* 1990; **9**(5):529–538.
13. Hakulinen T, Tenkanen L. Regression analysis of relative survival rates. *Applied Statistics* 1987; **36**(3):309–317.
14. Remontet L, Bossard N, Belot A, Esteve J. An overall strategy based on regression models to estimate relative survival and model the effects of prognostic factors in cancer survival studies. *Statistics in Medicine* 2007; **26**(10):2214.
15. Bolard P, Quantin C, Esteve J, Faivre J, Abrahamowicz M. Modelling time-dependent hazard ratios in relative survival: application to colon cancer. *Journal of Clinical Epidemiology* 2001; **54**(10):986–996.
16. Giorgi R, Abrahamowicz M, Quantin C, Bolard P, Esteve J, Gouvernet J, Faivre J. A relative survival regression model using b-spline functions to model non-proportional hazards. *Statistics in Medicine* 2003; **22**(17):2767–2784.
17. Lambert PC, Smith LK, Jones DR, Botha JL. Additive and multiplicative covariate regression models for relative survival incorporating fractional polynomials for time-dependent effects. *Statistics in Medicine* 2005; **24**:3871–3885.
18. Royston P, Parmar MKB. Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Statistics in Medicine* 2002; **21**(15):2175–2197.
19. Durrleman S, Simon R. Flexible regression models with cubic splines. *Statistics in Medicine* 1989; **8**(5):551–561.
20. Golub GH, van Loan CF. *Matrix Computations* (2nd edn). The John Hopkins University Press: Baltimore, MD, 1990.
21. Abrahamowicz M, MacKenzie TA. Joint estimation of time-dependent and non-linear effects of continuous covariates on survival. *Statistics in Medicine* 2007; **26**(2):392–408.
22. Lambert PC, Royston P. Further development of flexible parametric models for survival analysis. *The Stata Journal* 2009; **9**:265–290.
23. Carstensen B. Who needs the Cox model anyway? *Technical Report*, Steno Diabetes Center, Denmark, 2004. Available from: http://www.staff.pubhealth.ku.dk/ bxc/Talks/WntCma–xrp.pdf.
24. *Stata User's Guide Release 10.0*. Stata Press: College Station, 2007.
25. Efron B, Tibshirani RJ. *An Introduction to the Bootstrap*. Chapman & Hall: New York, 1993.
26. Coleman MP, Babb P, Mayer DJQM, Sloggett A. *Cancer Survival Trends in England and Wales, 1971–1995*: *Deprivation and NHS Region* (*CDROM*). Office for National Statistics: London, 1999.
27. Abrahamowicz M, MacKenzie TA. Time-dependent hazard ratio: modeling and hypothesis testing with application in lupus nephritis. *Journal of the American Statistical Association* 1996; **91**:1432–1439.
28. Royston P. The use of fractional polynomials to model continuous risk variables in epidemiology. *International Journal of Epidemiology* 1999; **28**(5):964–974.
29. Lambert PC, Thompson JR, Weston CL, Dickman PW. Estimating and modelling the cure fraction in population-based cancer survival analysis. *Biostatistics* 2007; **8**:576–594.