STATISTICS IN MEDICINE
Statist. Med. 2007; **26**:2214–2228
Published online 10 August 2006 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.2656

An overall strategy based on regression models to estimate relative survival and model the effects of prognostic factors in cancer survival studies

L. Remontet[1,*,†], N. Bossard[1,‡], A. Belot[1,§], J. Estève[1,¶] and the French network of cancer registries FRANCIM[2,‖]

[1]*Service de Biostatistique, Hospices Civils de Lyon, Lyon, France; Laboratoire de Biostatistique-Santé (UMR 5558), CNRS and Université Claude Bernard Lyon 1, Lyon, France*
[2]*Head office: Réseau FRANCIM, Faculté de médecine, Toulouse, France*

SUMMARY

Relative survival provides a measure of the proportion of patients dying from the disease under study without requiring the knowledge of the cause of death. We propose an overall strategy based on regression models to estimate the relative survival and model the effects of potential prognostic factors. The baseline hazard was modelled until 10 years follow-up using parametric continuous functions. Six models including cubic regression splines were considered and the Akaike Information Criterion was used to select the final model. This approach yielded smooth and reliable estimates of mortality hazard and allowed us to deal with sparse data taking into account all the available information. Splines were also used to model simultaneously non-linear effects of continuous covariates and time-dependent hazard ratios. This led to a graphical representation of the hazard ratio that can be useful for clinical interpretation. Estimates of these models were obtained by likelihood maximization. We showed that these estimates could be also obtained using standard algorithms for Poisson regression. Copyright © 2006 John Wiley & Sons, Ltd.

KEY WORDS: relative survival; regression splines; non-proportional hazards; Akaike Information Criterion; cancer; registry

*Correspondence to: L. Remontet, Service de Biostatistique, Hospices Civils de Lyon, Centre Hospitalier Lyon Sud, 69495 Pierre-Bénite Cedex, France.
†E-mail: laurent.remontet@chu-lyon.fr
‡E-mail: nadine.bossard@chu-lyon.fr
§E-mail: aurelien.belot@chu-lyon.fr
¶E-mail: jacques.esteve@chu-lyon.fr
‖List of registries is given in the Appendix.

Contract/grant sponsor: Ligue Nationale Contre le Cancer

Received 23 December 2005
Accepted 13 June 2006

INTRODUCTION

Relative survival is the main epidemiological indicator in survival studies of data stemming from population-based cancer registries [1]. This concept provides an objective measure of the proportion of patients dying from direct or indirect consequences of cancer without requiring a record of the cause of death. This explains its great popularity in population-based studies where large cohorts of patients are followed for long periods and the causes of death rarely documented or collected. Relative survival can be also useful in clinical studies where, though more clinical information is available, it remains difficult to impute death to the disease under study.

As in other survival studies, relative survival analysis involves estimations of survival at fixed times (generally at 1 and 5 years) and identification of prognostic factors; the latter task being an important goal in many studies. Point estimations may be made using the methods developed by Ederer [2] or Hakulinen [3], where relative survival is the ratio of the survival observed in the cohort of patients under study to the survival expected in this cohort taking into account some factors collected at inclusion (usually sex, age, and year of diagnosis). However, Ederer's methods present some drawbacks, especially in long-term analyses [3–5]. Furthermore, none of the above-cited methods allow to model the effects of prognostic factors.

Regression models have then been proposed to model the excess mortality at hazard level rather than at survival level [5, 6]. With such models, several covariates may be simultaneously adjusted and many tools developed for standard survival models can be applied to relative survival analysis. As the Cox model, the model proposed by Estève is based on the proportional hazard (PH) assumption [5]. However, in relative survival of cancer patients, the prognostic effects of the covariates change frequently with the time since diagnosis [1, 7–9]; thus, time-varying coefficients have been used to model that change [9–11].

In 2001, a population-based survival study was initiated in France; it involved, for the first time, all French cancer registry data. The vital status of 200 000 patients could be obtained using an active and standardized follow-up procedure that have been successful in about 96 per cent of subjects. The goals of this study were: (i) to estimate the relative survival for each cancer site according to age, gender, and year of diagnosis; (ii) to produce relative survivals standardized for age; and (iii) to model the effects of some prognostic factors using appropriate methods. Analyses were carried out using the same regression approach, adapted to each of the three goals. We report here the methodological problems we met and the solutions we adopted.

METHODS

Modelling the log of the baseline hazard with a continuous function

The relative survival was obtained using the excess hazard model [5, 6]. In this model, the mortality hazard λ_0 observed at time t after diagnosis in cancer patients is split up into two components: the mortality hazard due to cancer and the expected mortality hazard due to other causes. This may be written as

$$\lambda_0(t, x) = \lambda_c(t, x) + \lambda_e(a + t, z) \quad (1)$$

DOI: 10.1002/sim

where the parameter of interest λ_c is the cancer-related mortality hazard or *excess mortality hazard*, and x a vector of covariates. 'a' being the age at diagnosis, the expected mortality λ_e at age $a+t$, given the characteristics z, is obtained from published vital statistics. In our study, z were the gender, the year of death, and the Département (i.e. the main territorial and administrative division in France).

In the seminal article by Estéve *et al.* [5] λ_c was written $\lambda_c(t, x) = \exp(\beta x) f(t)$ where f is a step function; that is, the baseline hazard is assumed constant within pre-specified intervals. This model may be implemented with *Stata* software and a specific procedure in case of non-convergence or negative hazard (*strel* command). However, when estimation of the relative survival at specific time points is a priority (as opposed to a study whose objective is the estimation of the covariates effects), it is necessary to use a robust method to estimate $f(t)$ and a step function may not be the optimal choice. This is especially true when the analyses are to be carried on small groups of patients (defined by combinations of cancer site, gender, age, and diagnosis period) with few deaths in each stratum. Within this context of sparse data, the choice of intervals for the step function may be problematic whereas the results may depend on this choice.

As it was already proposed for crude survival [12, 13] or for relative survival [9, 10], we considered regression splines for $\log[\lambda_c(t)]$ (in absence of covariate), written in a truncated power basis, and used the following model:

$$\log[\lambda_c(t)] = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3 + \sum_{j=1}^{k} \theta_j (t - t_j)_+^3 \tag{2}$$

the '+' subscript corresponds to $u_+ = u$ if $u > 0$, $u_+ = 0$ if $u \leqslant 0$. $\mathrm{Log}[\lambda_c(t)]$ is a cubic regression spline; i.e. a piecewise polynomial with the constraint that the function and its first two derivatives should be continuous at the knots $t_1, \ldots, t_k$ (i.e. the join points) [14]. These functions, linear with respect to the parameters, provide a flexible tool: they are able to produce a great variety of curves and may be implemented using any statistical softwares.

Fractional polynomials have been also proposed to model the baseline excess mortality hazard [11].

Estimation based on the full likelihood and on Poisson likelihood approach

Consider the observation t_i, δ_i, and z_i of the ith subject, with $\delta_i = 1$ when t_i corresponds to the time of death and $\delta_i = 0$ when t_i corresponds to a censored observation. Its contribution to the log-likelihood may be written (up to a constant) as

$$l_i(\beta) = -\int_0^{t_i} \lambda_c(u)\, \mathrm{d}u + \delta_i \log[\lambda_c(t_i) + \lambda_e(a_i + t_i, z_i)] \tag{3}$$

The maximum likelihood (ML) estimates of this model, where $\log[\lambda_c(t)]$ is a quadratic spline, have been implemented in S or R software function called *RSurv* using Gauss–Legendre numerical integration [15]. However, as often in epidemiological cohort studies, estimation from equation (3) is greatly facilitated by splitting the data into multiple observations according to the follow-up [1, 16]. More precisely, consider original data (δ_i, t_i) on a subject i split into $n+1$ separate observations $(j = 0, \ldots, n)$ for some arbitrary bands of follow-up $[0; b_1[\ldots [b_j; b_{j+1}[\ldots [b_n; t_i[$ and let $l_{i,j}(\beta)$ be the contribution to the log-likelihood for each subject-band. On these split data, we used Cavalieri–Simpson numerical integration [17] to calculate the contribution to the cumulative

DOI: 10.1002/sim

hazard for each subject-band:

$$l_{i,j}(\beta) = \frac{b_{j+1} - b_j}{6}\left[\lambda_c(b_j) + 4\lambda_c\left(\frac{b_j + b_{j+1}}{2}\right) + \lambda_c(b_{j+1})\right] + \delta_{i,j}\log[\lambda_c(b_{j+1}) + \lambda_e(a_i + t_i, z_i)], \quad j = 0, \ldots, n;\ b_0 = 0;\ b_{n+1} = t_i \tag{4}$$

with $\delta_{i,j} = 0$ for each band but the last where it equals δ_i.

The ML estimates can then be obtained using Newton–Raphson algorithm.

When λ_c is a step function, splitting the data has another advantage: as noted by Dickman, the likelihood (3) can be considered as deriving from a generalized linear model (GLM) with outcome $\delta_{i,j}$, Poisson error structure, and an adequate link function [1]. This approach takes advantage of the GLM framework; in particular, the ML estimates can be obtained by a unique algorithm based on the iterative weighted least squares (IWLS) [18].

When λ_c is a continuous function, this GLM approach on split data can still be used but it requires a specific adaptation. In the Poisson setting, the contribution to the log-likelihood for each subject-band is

$$\begin{aligned} &l_{i,0}(\beta) = -b_1\lambda_c(\text{time}_{i,0}) \\ &\ldots \\ &l_{i,j}(\beta) = -(b_{j+1} - b_j)\lambda_c(\text{time}_{i,j}) \\ &\ldots \\ &l_{i,n}(\beta) = -(t_i - b_n)\lambda_c(\text{time}_{i,n}) + \delta_i \log[\lambda_c(\text{time}_{i,n}) + \lambda_e(a_i + t_i, z)] \end{aligned} \tag{5}$$

Since λ_c is no longer constant within each band, values of time have to be optimally chosen: we considered $\text{time}_{i,j} = (b_{j+1} + b_j)/2$, except for observations with $\delta_i = 1$ for which $\text{time}_{i,n} = t_i$. Doing that, with conveniently chosen bands of follow-up, is equivalent to approximate the integral of equation (3) by the 'point-milieu' method [17]. For observations with $\delta_i = 1$, the value for *time* is chosen so that the right part of equation (5) corresponds to the right part of the likelihood (3). The estimates of the parameters were obtained by implementation of the IWLS algorithm via iterative call of lm(), the S function for linear models.

To determine the point confidence interval of relative survival in each of the three above-described approaches (ML with *Rsurv*, ML with Cavalieri–Simpson integration, Poisson likelihood), we simulated using the Monte-Carlo method 1000 realizations of β by sampling from a multinormal random variable with mean $\hat{\beta}$ and covariance matrix $\text{var}(\hat{\beta})$. Using this sample, 1000 realizations of the cumulative hazard were numerically calculated by the S function *integrate*(). The standard errors of the logarithm of the cumulative hazards were then calculated and used to obtain the confidence intervals. The percentiles of the logarithm of the cumulative hazard could have been used instead; this yielded similar results (not shown).

Model selection

To improve survival estimates, especially at 5 years, we used data from diagnosis to 10 years follow-up. The number and the location of knots (equation (2)) were chosen *a priori*: two knots at

DOI: 10.1002/sim

1 year and 5 years. Furthermore, to obtain parsimonious models with good prediction properties, we fitted six models for each data set: a cubic spline with two knots (described above), a cubic spline with one knot at 1 year, a cubic polynomial, a quadratic polynomial, a linear and a constant model. The final model was selected using the Akaike Information Criterion (AIC) [19]. For very small sample sizes (less than 20 deaths), we restricted the choice to the linear or to the constant model.

The above-described strategy allows dealing with sparse data that frequently stem from computing relative survivals standardized for age because estimates must be obtained for each data set defined by its cancer site (forty seven categories), age group (five categories), gender, and period of diagnosis (three categories).

Modelling covariates effects using time-dependent hazard ratios

Age at diagnosis is known to be a strong prognostic factor in cancer relative survival, though the pattern of its effect is still unknown or has not been conveniently modelled yet. However, it is agreed that the effect of age is neither linear nor constant over the follow-up period [7, 9]. To estimate simultaneously non-linear and time-dependent effects, we fitted the following non-proportional hazard model to data up to 5 years of follow-up:

$$\log[\lambda_c(t, \text{age})] = f(t) + g(\text{age}) + h(t) \times \text{age} \tag{6}$$

where f and h are cubic splines with a knot at 1 year and g is a cubic spline with a knot at the mean age. According to equation (6), the log hazard ratio function (LHRF) for two arbitrary values of age a_0 and a_1 is

$$\log[\lambda_c(t, \text{age} = a_1)/\lambda_c(t, \text{age} = a_0)] = g(a_1) - g(a_0) + h(t) \times (a_1 - a_0) \tag{7}$$

This LHRF is not linear according to age and depends on t, two characteristics often observed when analysing prognostic factors in cancer patients. We tested the proportional hazard assumption for age by testing the significance of the $h(t) \times$ age term in model (6): this was done using the likelihood ratio test (LRT) with 4 degrees of freedom. Similar non-proportional hazard models may be easily built for categorical factors.

APPLICATION

Estimation based on the full likelihood and on a Poisson likelihood approach

To compare various procedures of estimation, including that of the free function *RSurv*, which is also an implementation of the full likelihood method, we considered the model $\log[\lambda_c(t)] = f(t) + h(t) \times$ gender, where f and h are two quadratic splines, both with two knots at 1 year and 5 years, and where *gender* is 0 for male, 1 for female. This model was chosen because, in *RSurv*, modelling baseline hazard with continuous function requires at least one covariate in the model. Furthermore, *RSurv* handles only quadratic splines, and, finally, the number of knots must be two. This model was applied to patients with bladder cancer aged 15 and over. Table I shows the results of ML with *Rsurv*, ML with Cavalieri–Simpson integration, and Poisson likelihood according to two subject-band sizes: (i) 'large intervals': 0.5 year-bands from 0 to 10 years of follow-up; (ii) 'short intervals' 0.05 year-bands from 0 to 1 year then 0.1 year-bands up to 10 years of follow-up.

DOI: 10.1002/sim

Table I. Log-likelihood (model M_1: $\log[\lambda_c(t)] = f(t) + h(t) \times$ gender), likelihood ratio test (M_0: $\log[\lambda_c(t)] = f(t) +$ gender *versus* M_1) and survival estimates in men and women from M_1 with confidence intervals (per cent) according to the approach and to the size of the subject-bands (bladder cancer).

Size of the		Log-likelihood	Likelihood	Survival (men)			Survival (women)		
subject-bands	Approach	model M_1	ratio test	1 year	3 years	5 years	1 year	3 years	5 years
Large interval	Poisson likelihood	−10587.33	22.1	80.6 [79.5–81.6]	65.7 [64.3–67.0]	59.1 [57.5–60.6]	70.6 [68.0–73.1]	54.9 [52.0–57.7]	49.7 [46.6–52.8]
Large interval	Maximum likelihood (Cavalieri–Simpson)	−10613.91	23.1	81.3 [80.2–82.4]	65.8 [64.4–67.2]	59.7 [58.1–61.2]	71.6 [69.1–74.0]	55.2 [52.2–58.1]	50.5 [47.4–53.6]
Short interval	Poisson likelihood	−10613.50	23.1	81.3 [80.3–82.3]	65.8 [64.5–67.1]	59.7 [58.2–61.2]	71.6 [69.1–73.9]	55.2 [52.2–58.2]	50.5 [47.2–53.7]
Short interval	Poisson likelihood without the specific adaptation	−10591.28	22.4	81.0 [80.0–82.0]	65.8 [64.5–67.1]	59.7 [58.2–61.2]	71.3 [68.8–73.7]	55.3 [52.2–58.2]	50.6 [47.3–53.8]
Short interval	Maximum likelihood (Cavalieri–Simpson)	−10613.87	23.1	81.3 [80.2–82.4]	65.8 [64.4–67.2]	59.7 [58.1–61.2]	71.6 [69.1–73.9]	55.2 [52.2–58.1]	50.5 [47.3–53.6]
Not concerned	Maximum likelihood (*RSurv*)	−10613.87	23.1	81.3 [80.2–82.4]	65.8 [64.4–67.2]	59.7 [58.1–61.2]	71.6 [69.1–74.0]	55.2 [52.2–58.1]	50.5 [47.4–53.6]

DOI: 10.1002/sim

We also present in this table, for short intervals, the results of the Poisson likelihood approach with $\text{time}_{i,j} = (b_{j+1} + b_j)/2$ for all observations (i.e. without specific adaptation).

When the follow-up period was divided into sufficiently small bands, there were no differences between the results of these approaches, except for the Poisson likelihood without specific adaptation: its estimates may differ from the ML estimates even when small intervals were used (see likelihood and likelihood ratio test). However, in terms of relative survival probabilities, the results of all approaches were very close.

'Short intervals' seemed sufficient to obtain reliable estimates with the Poisson likelihood approach (with specific adaptation) even if the likelihood of model M_1 was not exactly the same as that of the other approaches. These short intervals have been subsequently used in the analysis.

ML estimates obtained with *Rsurv* or based on Cavalieri–Simpson approximation with split data were identical even with 'large intervals'.

Figure 1 illustrates these results in terms of hazard estimates in men (Poisson likelihood approach without specific adaptation is not shown).

Regression splines until 10 years of follow-up to estimate 5-year relative survival

Modelling the hazard until 10 years of follow-up produce more reliable patterns of the hazard function over the first 5 years. This point was illustrated by the analysis of four subgroups that present contrasting and not trivial patterns of mortality: Figure 2 shows the hazard function obtained after model selection by AIC with data up to 5 and 10 years of follow-up (cubic splines with two

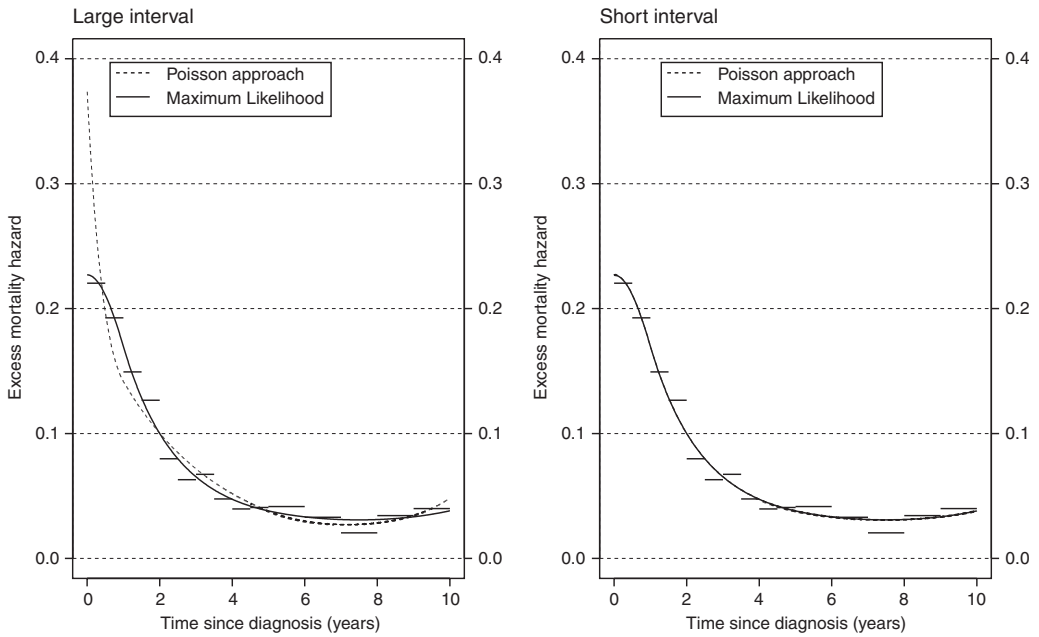

Figure 1. Excess mortality hazard estimates with a quadratic spline for bladder cancer in men aged 15 and over. Comparison between Poisson approach and maximum likelihood according to the size of the subject-bands. Horizontal bars depict the corresponding step function estimates.

DOI: 10.1002/sim

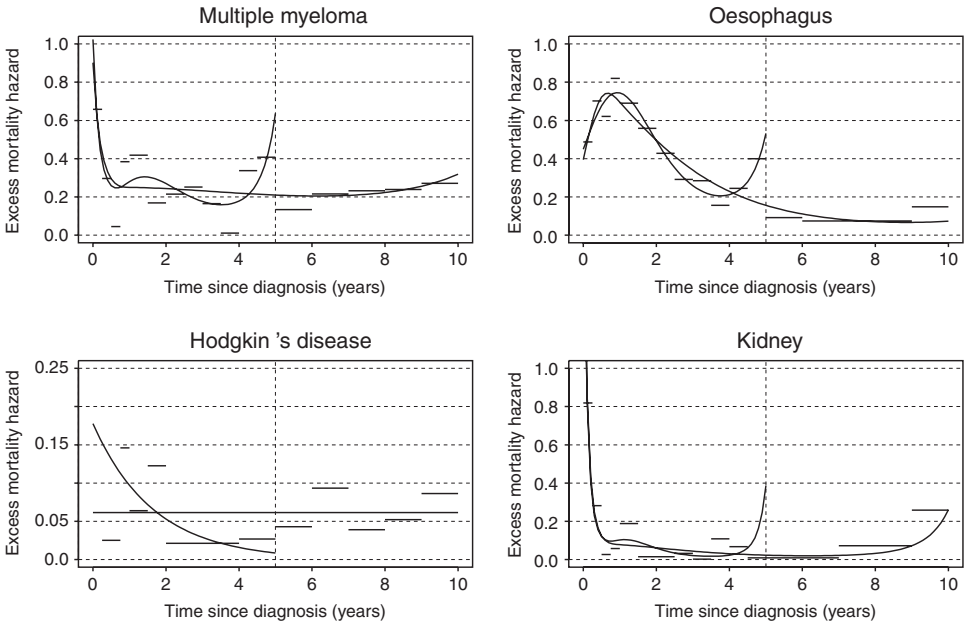

Figure 2. Continuous hazard function selected by AIC, based on data up to 5 years or up to 10 years follow-up, and estimates from models assuming a step function (horizontal bars), for four particular subgroup analyses.
Data are: multiple myeloma in women aged 75 and over diagnosed between 1989 and 1991, oesophagus cancer in men aged 45–55 in three French Départements (preliminary analysis), Hodgkin's disease in men aged 55–65, kidney cancer in men aged 75 and over diagnosed between 1992 and 1994.

knots were not considered when the fit concerned the five-year data). This figure shows also, as horizontal bars, the estimates from a model assuming a step function.

The advantages of modelling up to ten years are particularly important in the neighbourhood of 4–5 years because the hazards observed within this period are compared with the hazards observed in the subsequent period. In multiple myeloma, for example, the increase observed between 4 and 5 years was judged highly improbable because this increase did not extend beyond the 5th year. In contrast, the fit with 0–5 years data followed this increase and seemed less plausible. This consideration was reinforced by data related to the same cancer and the same age group in another period: the outlying elevated hazard between 4 and 5 years was no longer observed. The same comment applied to the oesophagus cancer. Note that Figure 2 shows the great flexibility of cubic splines, which are able to describe very different patterns of mortality hazard.

Based on the same subsamples data as Figure 2, Table II shows the results, in terms of survival estimates, of the three above-cited strategies: step function, modelling up to 5 years, and modelling up to 10 years.

The Hodgkin's disease example showed that the differences between the estimates can be substantial (see survivals at 1 and 3 years); this is also the case for the kidney example (see survival at 5 years).

DOI: 10.1002/sim

Table II. Survival estimates with confidence intervals (per cent) for particular subgroup analyses according to three strategies: (1) a step function is assumed and data are grouped into broader time intervals in case of non-convergence or negative hazard (*strel* command from *Stata* software), (2) modelling the hazard until 5 years of follow-up with five regression splines; final model selected by AIC, (3) modelling the hazard until 10 years of follow-up with six regression splines; final model selected by AIC.
Data are: multiple myeloma in women aged 75 and over diagnosed between 1989 and 1991, oesophagus cancer in men aged 45–55 in three French Départements (preliminary analysis), Hodgkin's disease in men aged 55–65, kidney cancer in men aged 75 and over diagnosed between 1992 and 1994.

	Survival at 1 year			Survival at 3 years			Survival at 5 years		
Cancer type	Step function	Modelling up to 5 years	Modelling up to 10 years	Step function	Modelling up to 5 years	Modelling up to 10 years	Step function	Modelling up to 5 years	Modelling up to 10 years
Multiple myeloma	70.4 [60.7–78.2]	69.8 [60.8–77.2]	69.0 [60.1–76.3]	41.2 [31.2–51.0]	41.8 [32.1–51.1]	42.4 [33.1–51.4]	25.7 [16.7–35.6]	25.9 [16.8–35.9]	27.3 [18.8–36.5]
Oesophagus	51.8 [45.1–58.0]	52.3 [46.5–57.7]	51.8 [45.8–57.5]	19.8 [14.8–25.2]	19.2 [14.7–24.2]	18.9 [14.5–23.8]	11.7 [7.8–16.4]	11.3 [7.5–16.0]	11.8 [8.2–16.2]
Hodgkin's disease	88.3 [75.6–94.6]	87.5 [76.4–93.6]	94.0 [90.4–96.3]	75.9 [60.7–85.9]	78.2 [63.8–87.4]	83.1 [73.8–89.4]	73.0 [57.3–83.7]	75.6 [57.9–86.7]	73.5 [60.3–82.9]
Kidney	74.3 [64.8–81.6]	73.3 [64.2–80.4]	73.3 [64.2–80.5]	64.7 [52.8–74.2]	64.7 [52.7–74.3]	64.8 [53.5–74.1]	54.2 [40.9–65.7]	57.4 [34.7–74.8]	60.9 [43.8–74.2]

DOI: 10.1002/sim

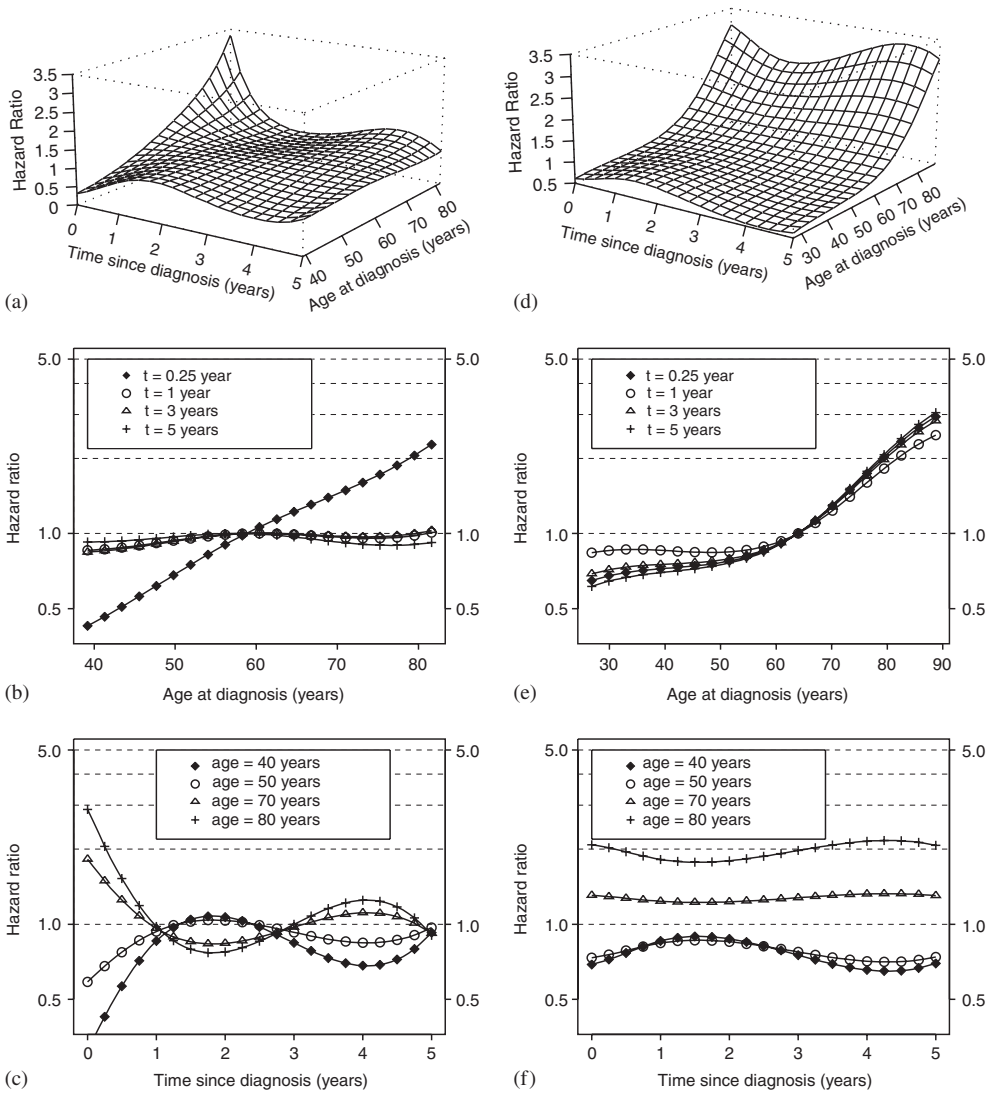

Figure 3. Hazard ratio (mean age as reference) according to age and time since diagnosis (a) and (d); hazard ratio according to age at specified time t since diagnosis (b) and (e); and hazard ratio according to time t since diagnosis at specified age (c) and (f).
Data are: Oropharynx cancer (a–c) and non-Hodgkin's lymphoma (d–f) in subjects aged 15 and over.

Non-proportional hazards model

An example of non-proportional hazard model was used above for bladder cancer (Table I) where the effect of sex changed with time since diagnosis. Such a model is an effective method to analyse

DOI: 10.1002/sim

the impact of gender on survival: a proportional hazard model would give a biased picture while estimating one curve per gender is more cumbersome and does not provide a clear picture of the non-proportionality. Furthermore, comparison of the likelihoods of models $M0$ and $M1$ permits to assess quantitatively the level of non-proportionality.

The advantage of such modelling is even clearer with a continuous covariate such as age. Figure 3 illustrates different patterns of the time-dependent effect of age: (i) 3(a) and (d) show the hazard ratio function (HRF, equation (7)) according to time and age; (ii) 3(b) and (e) show the HRF plotted at fixed values of t (0.25, 1, 3 and 5 years) on a logarithmic scale; (iii) 3(c) and (f) show the HRF plotted at fixed values of age (40, 50, 70, and 80 years). In all parts, the reference age a_0 is the mean age at diagnosis.

As to the oropharynx cancer, the PH assumption was rejected ($\mathrm{LRT} = 60.1$, $p < 0.001$): during early follow-up (first six months), old age is an unfavourable factor whereas this is not true at 1 year or at 5 years, times at which the hazard ratios depended no more on age. This limitation of the effect of age to the first months after diagnosis has been quite often observed in our analysis. It may correspond to cancers for which a large proportion of deaths occur early during follow-up because of post-surgical complications that affect especially the oldest patients [7]. More advanced cancer stages in oldest patients may be another explanation.

The PH assumption is acceptable for non-Hodgkin's lymphoma ($\mathrm{LRT} = 9.2$, $p = 0.06$); the hazard increased with age starting from 60-year old, whatever the time since diagnosis (Figure 3(e)). In other words, the hazard at 80 years was twice the hazard at the reference age, whatever the time since diagnosis (Figure 3(f)).

DISCUSSION

Our approach contrasts with other publications of population survival data that generally use 'life table methods' to calculate point relative survival probabilities in subgroups and restrict the regression model to the analysis of covariates effects. We did not see practical or theoretical advantages to such a strategy. We rather thought it is more sensible to base the estimation of relative survival on an ML estimate of the excess mortality hazard in various subgroups rather than on the calculation of one of the available versions of expected survival [2, 3]. Two obvious benefits of our approach are that the excess hazard method provides a genuine survival curve and that it is consistent with the results obtained when covariates are taken into account. For example, in agreement with our results (Table II), the 5-year survival probability of Hodgkin disease is about 75 per cent with the Hakulinen method. However, this method provides a 4-year survival (77 per cent) that is greater than the 3-year survival (76 per cent). The other life table methods produce the same anomaly. This cannot happen with an excess hazard model. Furthermore, our strategy to estimate relative survival at 1, 3 and 5 years is based on a model that fitted the data up to 10 years and on a continuous function for the baseline hazard. Using almost all the follow-up period stabilizes the estimates by avoiding random fluctuations, in particular around 5 years. In contrast to piecewise step functions, continuous functions produce clinically convincing patterns of hazard (which are interesting *per se*) and reduce considerably the number of parameters. Finally, it avoids the question of partitioning the follow-up period in intervals, which can be a real problem with sparse data.

The maximization of the likelihood of the regression models is facilitated by splitting the follow-up period of each subject into small intervals. Furthermore, taking advantage of the similarity

DOI: 10.1002/sim

between the likelihood of our model and a well-chosen Poisson likelihood, the estimates could be also obtained using standard algorithms for the GLM.

To estimate relative survival from piecewise constant excess hazard models, the GLM approach is the one recommended by Dickman. It gives the same results as other methods (e.g. the full likelihood approach) and is easier to implement [1]. Furthermore, when subject-band observations are 'collapsed' to give one observation for each covariate pattern ('grouped data') and when only categorical covariates are investigated, Dickman argues that the GLM theory may be used to check the assumptions and to assess the goodness-of-fit.

When the baseline hazard is modelled using a continuous function, the GLM approach has been also used on grouped data [11]. However, we have shown that, using individual data and continuous covariates, the Poisson approach is equivalent to the approximation of the cumulative hazard by the 'point-milieu' method. For an adequate approximation of this integral, small intervals are required, especially during the first year when the hazard pattern is the more complex. Furthermore, for subject-bands ending by death, we took the time of death for $\text{time}_{i,n}$ to favour the correspondence with the right part of the likelihood. This is not strictly equivalent to the method of 'point-milieu' regarding the cumulative hazard part but this approximation is perfectly adequate when small subject-bands are used. This specific adaptation, with the retained subject-bands (short intervals in Table I and Figure 1) is necessary to obtain estimates close to ML ones.

To fit the models, we developed our own tool instead of using the S function *RSurv* [15]. This function accepts only quadratic splines and the number of knots is fixed to two, while we needed other models such as a simple linear or more complex cubic splines. Furthermore, our 'non-user-friendly' function (with short intervals and Poisson likelihood approach) is about seven times faster than the user-friendly *RSurv* (about 85 s *versus* 10 min for a data set on 7000 patients) and higher-order splines may be easily implemented. In particular, we did not meet the computing difficulties encountered by Bolard that forced him to limit spline complexity [7, 10], which confirms the easier implementation of approaches based on split data. To control each step of the algorithm, we implemented IWLS algorithm using an iterative call of the S function lm(), rather than the glm() function with user-defined link function.

Regression splines have been already used in survival analyses to model baseline hazards [9, 10, 12, 13], to model covariates effects [20–22] or to model time-dependent hazard ratios [8–10, 21, 23, 24]. The main advantage of regression splines (as opposed to smoothing splines) is that they offer a great flexibility to fit the data while being linear in the parameter [22, 24–26]. In particular, cubic splines are a good compromise between flexibility and smoothness [20]. They can be equally based on a truncated power basis or on a B-spline basis, which is a more numerically stable way to form the design matrix [9, 20]. However, the truncated power basis is more intuitive and allows extrapolating beyond the knots. Furthermore, when modern methods for matrix inversion are used, the truncated power basis seldom presents computing problems, especially when the spline order and the number of knots are small, as in our study [27].

To avoid potential instability in the tails, cubic splines can be constrained to be linear beyond the extreme knots [28]. These natural cubic splines have been already used in survival studies [10, 22, 24], but they made inference difficult: splines of different orders cannot be compared and nested knot sequences do not produce nested models [20]. Furthermore, high instability in the tails is seldom reported by authors using unconstrained splines [9, 13, 20, 21, 23]. In our study, we have chosen to search for a parsimonious model using AIC. In order to detect high instability in the tail, we explored a great number of data sets including very sparse ones and found no evidence of instability in the finally retained model.

DOI: 10.1002/sim

Ideally, for correct inference using standard methods, the location of spline knots must be chosen *a priori*, independently of the response variable. In general, there is no prior information; thus, the percentiles of time to death are used to locate the knots. However, modelling the hazard baseline in cancer survival, we considered that there is a prior experience on which pre-specification of knot location should be based. High mortality is often observed during the first year; we then decided to locate one knot at 1 year, when a change in curvature is expected and a second knot at 5 years, to cover the remaining time interval. This *a priori* choice was strengthened by *a posteriori* analysis comparing the likelihood of the model with knot at 1 year with the likelihood of models with knots at 1.5, 2, and 3 years. For all 18 types of cancer analysed for this purpose, locating the knot at 1 year was the best choice among all other positions (we did not consider locations before 1 year because extreme positions of knots are not recommended). When splines are used to model the effect of prognostic factors such as age, prior information on the exact pattern of the effect is not usually available, so we located the knot at the mean age.

The number of knots depends on the expected complexity of the phenomenon under consideration and parsimony is desirable because the variance of the estimators and the risk of overfitting increase with the number of knots. For age, a cubic spline with one knot (five parameters) seemed sensible because we did not expect more than three inflection points.

Splines are natural choices to model covariates effects because detecting and testing non-linearity are thus easier: the linear model (as the 'no age effect' model) is nested within any spline model and the corresponding likelihood ratio statistics—rather than Wald statistics—may be favourably used [20]. This non-linearity should be taken into account because, for example, the 40- *versus* 50-year old hazard ratio is often different from the 70- *versus* 80-year old hazard ratio.

Furthermore, analysing relative survival of cancer patients, time-dependent hazard ratios should be taken into account, especially for age [1, 7, 9]. Time-varying coefficients, including the $h(t) \times x$ term, have been proposed for crude and relative survival, x being mainly a categorical factor, as in the examples given by several authors [9, 10, 23, 24] or, exceptionally, a continuous factor [21]. We proposed to model simultaneously non-linearity and time-dependent hazard ratio by equation (6), using continuous values of age. We used only data up to 5 years of follow-up to overcome the potential change of the effect of age after 5 years. We did not follow the strategy of Abrahamowicz *et al*. [23] to determine the terms that compose $h(t)$. Instead, we took systematically a cubic spline with one knot at 1 year for $h(t)$ and tested it against the PH assumption with LRT. This strategy based on a 4-df test is certainly not the most powerful one but we used it because we wanted to underscore only important violations of the PH assumption. This simultaneous modelling of non-linearity and time-dependent effect led to a 3D graphical representation of the hazard ratio function according to time and age (Figure 3(a) and (d)). This representation might be a useful tool for clinical interpretation because the effect of age is not averaged over the whole follow-up period. Note that this kind of modelling is possible because $h(t)$ is a parsimonious function while adjusting time-varying coefficients would be more problematic with a step function (due to the involvement of a high number of additional parameters). Equation (6) is a special case of bivariate response surface, relatively easy to implement due to the additivity and the linearity of the parameters. It would be interesting to develop and fit more general multivariate splines. However, estimation problems may occur if, for example, older patients die early and the data to estimate the remaining surface become insufficient [23]. In this case, it is necessary to assume a particular relationship between hazard, time, and age as in equation (6). However, this is not always sufficient and for some cancers, estimations from equation (6) were not possible.

DOI: 10.1002/sim

To model the baseline, we have restricted the choice to classical polynomial functions and cubic spline functions. These functions were compared on AIC; that is, on a reasonable trade-off between model parsimony and goodness-of-fit. This approach was already proposed by Abrahamowicz *et al.* [23] for $h(t)$. As expected with the AIC, this strategy tends to select models that have better predictive value in comparison with the strategy of the *a priori* fixed model—fixed order and fixed knots. However, as with any *a posteriori* selection procedure, traditional inference methods are not perfectly accurate. In particular, the coverage rates of the confidence intervals tend to be below the nominal value of 95 per cent (from 69 to 95 per cent in Abrahamowicz simulation). In practice, survival estimates are used as prognostic tools and inference is not of primary interest; thus, choosing the appropriate strategy, we favoured the predictive ability.

In conclusion, our approach yielded smooth and reliable estimates of the mortality hazard taking into account all the available information. It also provided a clear representation of the covariates effects on hazard.

APPENDIX

The cancer registries of the network FRANCIM are: Bas-Rhin General Cancer Registry (M. Velten), Calvados Digestive Cancer Registry (G. Launoy), Calvados General Cancer Registry (M. H. Amar), Côte d'Or Digestive Cancer Registry (J. Faivre), Côte d'Or Malignant Haemopathies Cancer Registry (P. M. Carli), Doubs General Cancer Registry (A. Danzon), Hérault General Cancer Registry (B. Trétarre), Haut-Rhin General Cancer Registry (A. Buemi), Isère General Cancer Registry (M. Colonna), Loire-Atlantique Breast & Colon Cancer Registry (M. J. Leroux), Manche General Cancer Registry (N. Maarouf), Marne & Ardennes Thyroid Cancer Registry (C. Schvartz), Saône et Loire Digestive Cancer Registry (J. Faivre), Somme General Cancer Registry (A. Dubreuil, N. Raverdy), Tarn General Cancer Registry (P. Grosclaude, M. Sauvage).

ACKNOWLEDGEMENTS

This work was funded by a grant from the 'Ligue Nationale Contre le Cancer'. We thank Dr J. Iwaz, PhD, scientific advisor for helpful comments on the manuscript. We also thank the two referees for their interesting suggestions.

REFERENCES

1. Dickman PW, Sloggett A, Hills M, Hakulinen T. Regression models for relative survival. *Statistics in Medicine* 2004; **23**(1):51–64.
2. Ederer F, Axtell LM, Cutler SJ. The relative survival rate: a statistical methodology. *National Cancer Institute Monograph* 1961; **6**:101–121.
3. Hakulinen T. Cancer survival corrected for heterogeneity in patient withdrawal. *Biometrics* 1982; **38**:933–942.
4. Hakulinen T. On long-term relative survival rates. *Journal of Chronological Disease* 1977; **30**:431–443.
5. Estéve 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.
6. Hakulinen T, Tenkanen L. Regression analysis of relative survival rates. *Applied Statistics* 1987; **36**:309–317.
7. 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.
8. Quantin C, Abrahamowicz M, Moreau T, Bartlett G, MacKenzie T, Tazi MA, Lalonde L, Faivre J. Variation over time of the effects of prognostic factors in a population-based study of colon cancer: comparison of statistical models. *American Journal of Epidemiology* 1999; **150**(11):1188–1200.

DOI: 10.1002/sim

9. 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.
10. Bolard P, Quantin C, Abrahamowicz M, Esteve J, Giorgi R, Chadha-Boreham H, Binquet C, Faivre J. Assessing time-by-covariate interactions in relative survival models using restrictive cubic spline functions. *Journal of Cancer Epidemiology and Prevention* 2002; **7**(3):113–122.
11. 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**(24):3871–3885.
12. Etezadi-Amoli J, Ciampi A. Extended hazard regression for censored survival data with covariates: a spline approximation for the baseline hazard function. *Biometrics* 1987; **43**:181–192.
13. Kooperberg C, Stone CJ, Truong YK. Hazard regression. *Journal of the American Statistical Association* 1995; **90**(429):78–94.
14. Smith PL. Splines: as a useful and convenient statistical tool. *The American Statistician* 1979; **33**(2):57–62.
15. Giorgi R, Payan J, Gouvernet J. RSURV: a function to perform relative survival analysis with S-PLUS or R. *Computer Methods and Programs in Biomedicine* 2005; **78**(2):175–178.
16. Breslow NE, Day NE. *The Design and Analysis of Cohort Studies*. International Agency for Research on Cancer Press: Lyon, 1987.
17. Quarteroni A. *Méthodes Numériques Pour Le Calcul Scientifique*. Springer-Verlag: Paris, 2000; 279–315.
18. McCullagh P, Nelder JA. *Generalized Linear Models*. Chapman & Hall: London, New York, 1989; 40–43.
19. Akaike H. Information theory and an extension of the maximum likelihood principle. In *2nd International Symposium on Information Theory*, Petrov BN, Csaki F (eds). Budapest, 1973; 268–281.
20. Sleeper LA, Harrington DP. Regression splines in the Cox model with application to covariate effects in liver disease. *Journal of the American Statistical Association* 1990; **85**(412):941–949.
21. Gray RJ. Flexible methods for analysing survival data using splines, with applications to breast cancer prognosis. *Journal of the American Statistical Association* 1992; **87**(420):942–951.
22. Durrleman S, Simon R. Flexible regression models with cubic splines. *Statistics in Medicine* 1989; **8**(4):551–561.
23. Abrahamowicz M, MacKenzie T, Esdaile JM. Time-dependent hazard ratio: modeling and hypothesis testing with application in Lupus Nephritis. *Journal of the American Association* 1996; **91**(436):1432–1439.
24. Hess KR. Assessing time-by-covariate interactions in proportional hazards regression models using cubic spline functions. *Statistics in Medicine* 1994; **13**(10):1045–1062.
25. Harrell FEJ, Lee KL, Pollock BG. Regression models in clinical studies: determining relationships between predictors and responses. *Journal of the National Cancer Institute* 1988; **80**(15):1198–1202.
26. Green PJ, Silverman BW. *Nonparametric Regression and Generalized Linear Models*: *A Roughness Penalty Approach*. Chapman & Hall: London, 1994.
27. Harrell FEJ. *Regression Modeling Strategies With Applications to Linear Models*, *Logistic Regression*, *and Survival Analysis*. Springer-Verlag: New York, 2002; 19–23.
28. Herndon JE, Harrell FEJ. The restricted cubic spline as baseline hazard in the proportional hazards model with step function time-dependent covariables. *Statistics in Medicine* 1995; **14**(19):2119–2129.

DOI: 10.1002/sim