American Journal of Epidemiology
Vol. 150, No. 11
Printed in U.S.A.

Variation Over Time of the Effects of Prognostic Factors in a Population-based Study of Colon Cancer: Comparison of Statistical Models

Catherine Quantin,[1] Michal Abrahamowicz,[2,3,4] Thierry Moreau,[3] Gillian Bartlett,[2] Todd MacKenzie,[5] Mohammed Adnane Tazi,[6] Luc Lalonde,[7] and Jean Faivre,[6]

The authors compare the performance of different regression models for censored survival data in modeling the impact of prognostic factors on all-cause mortality in colon cancer. The data were for 1,951 patients, who were diagnosed in 1977–1991, recorded by the Registry of Digestive Tumors of Côte d'Or, France, and followed for up to 15 years. Models include the Cox proportional hazards model and its three generalizations that allow for hazard ratio to change over time: 1) the piecewise model where hazard ratio is a step function; 2) the model with interaction between a predictor and a parametric function of time; and 3) the non-parametric regression spline model. Results illustrate the importance of accounting for non-proportionality of hazards, and some advantages of flexible non-parametric modeling of time-dependent effects. The authors provide empirical evidence for the dependence of the results of piecewise and parametric models on arbitrary a priori choices, regarding the number of time intervals and specific parametric function, which may lead to biased estimates and low statistical power. The authors demonstrate that a single, a priori selected spline model recovers a variety of patterns of changes in hazard ratio and fits better than other models, especially when the changes are non-monotonic, as in the case of cancer stages. *Am J Epidemiol* 1999;150:1188–1200.

colonic neoplasms; Cox regression; goodness-of-fit; models, statistical; multivariate analysis; regression analysis; risk factors; survival analysis

The Cox proportional hazards (PH) model (1) is the basic tool for analyzing censored survival data. The underlying assumption is that the hazard ratio associated with a given predictor is constant over the entire follow-up period. In practice, the Cox PH model is often selected a priori and is the only regression method employed. However, if the impact of a predictor on hazard changes during follow-up, the PH model may produce biased results (2). Several tests have been proposed to check the PH assumption (3) but they are seldom used. A review of survival analyses in cancer journals by Altman et al. (4) showed that only two out of 43 articles that used multivariable Cox model verified the assumptions. One reason is that after having rejected the PH hypothesis it may not be apparent how to remodel the predictor's effect.

The standard method for modeling the effect of a predictor that violates the PH assumption is to include a time-dependent covariate representing an interaction between the predictor and a parametric function of follow-up time $f(t)$, the shape of which represents the changes in hazard ratio (HR) during follow-up (1). If available substantive knowledge makes this choice difficult, two alternative approaches are possible: 1) restricting a priori $f(t)$ to a single arbitrarily selected parametric function; or 2) estimating several different functions and selecting a posteriori the one that fits the best. In the Results section, we present empirical evidence of the difficulties incurred by each approach.

Another method was proposed by Moreau et al. (5), who fit a piecewise PH model. HR becomes a step function that is constant within each a priori determined time interval but varies between intervals. The resulting estimates are un-smooth and the impact of the number of intervals is not obvious.

Received for publication February 20, 1998, and accepted for publication March 1, 1999.

Abbreviations: AIC, Akaike Information Criterion; d.f., degrees of freedom; HR, hazard ratio; LHR, likelihood ratio test; PH model, proportional hazards model.

[1] Department of Biostatistics and Medical Informatics, Teaching Hospital of Dijon, France.
[2] Department of Epidemiology and Biostatistics, McGill University, Division of Clinical Epidemiology, Montreal General Hospital.
[3] Department of Epidemiology and Biostatistics (U472), French Institute for Medical Research (INSERM), Villejuif, France.
[4] International Agency for Research on Cancer, ECP, Lyon, France.
[5] The Children's Hospital Research Institute of Denver and Pediatric Clinical Research Center and Department of Preventive Medicine and Biometrics, University of Colorado, CO.
[6] Registry of Digestive Tumors, Burgundy University, Dijon, France.
[7] Department of Mathematics and Statistics, McGill University, Montréal, Québec, Canada.
Reprint requests to Dr. Michal Abrahamowicz, Division of Clinical Epidemiology Montreal General Hospital, 1650 Cedar Ave., Montréal, Québec H3G 1A4, Canada.

To avoid these difficulties, in the 1990s several authors developed non-parametric methods for modeling time-dependent HR as a smooth flexible function of time (6–11). All of these methods share the ability to select the shape of the HR function directly from the data, without imposing strict a priori assumptions typical of parametric models. Simulations indicate that non-parametric models yield practically unbiased estimates of a broad range of HR functions (6, 7). However, some of the more flexible models do not allow for accurate inference (10). Moreover, increasing flexibility raises concerns about over-fitting bias (12). In this context, a relatively parsimonious regression spline model, recently proposed by Abrahamowicz et al. (6) seems to offer a reasonable compromise between modeling flexibility and the accuracy of inference, as demonstrated in simulations.

The ability of non-parametric time-dependent models to offer new insights into *real* data remains yet to be evaluated, especially when several predictors have to be investigated. In the present study, we use the example of predictors of mortality in colon cancer to assess the performance of different models in the univariate and multivariable analyses of a large cohort followed for several years. Following the Occam's razor principle, we first focus on testing the PH hypothesis, and accept a more complex time-dependent model only if this hypothesis is rejected. Indeed, simulations indicate that if the improvement in fit due to a time-dependent model is statistically nonsignificant, then there is an increased risk that the estimate of pattern of changes is biased (6). Next, we compare the results of different models for non-proportional hazards and explain the observed differences in terms of the underlying assumptions.

MATERIALS AND METHODS

Data source

We analyzed survival data for 1,951 patients, representing all cases of colon cancer diagnosed among the residents of Côte d'Or region in France between 1977 and 1991. Demographic and clinical data were obtained through medical files and recorded in the Registry of Digestive Tumors of Côte d'Or. Information on vital status and date of death was obtained regularly through administrative and medical sources, town hall and medical files, until December 31, 1994 when the follow-up was terminated.

Definition of prognostic variables

Our analyses focused on the predictive ability of several categorical variables evaluated at the time of colon cancer diagnosis. Demographics included patients' age, dichotomized at 65 years, sex and residence (urban vs. rural). Date of diagnosis was categorized into three 5-year periods: 1977–1981, 1982–1986, and 1987–1991. Tumor site was coded according to the 9th revision of the *International Classification of Diseases for Oncology* and merged into two classes: right and left colon. We combined information about the cancer stage at diagnosis (13) and the type of treatment (curative surgery vs. palliative treatment). In patients who had neither surgery nor evidence of distant metastasis, we designated the tumors as "Dukes unclassified." Therefore, the combined stage variable had four categories: 1, Dukes A tumors; 2, Dukes B tumors; 3, Dukes C resected for potential cure, designated as "C curative"; 4, Dukes C with palliative treatment, metastatic, or unclassified tumors (so-called *U*) designated as *C* palliative, *D*, *U*.

Statistical modeling

In all survival analyses, time-to-event was defined as time to death of any cause. Time 0 corresponded to the colon cancer diagnosis. Patients were censored when lost to follow-up or on December 31, 1994.

Cox proportional hazards model

In the Cox PH model (1), the hazard at time t, associated with the specific values $x_1,\ldots,x_k$ of k covariates, $X_1,\ldots,X_k$, is defined as:

$$\lambda(t|x_1,\ldots,x_k) = \lambda_0(t)\exp(\Sigma\beta_j x_j), \qquad (1)$$

where $\lambda_0(t)$ is an unspecified baseline hazard function at time t, i.e., the hazard corresponding to $x_1 = ,\ldots, = x_k = 0$, and β_j is a log HR that represents the effect of covariate X_j on the logarithm of the hazard. The hypothesis of no association of $k \geq 1$ variables with survival is tested by likelihood ratio test (LRT) with k degrees of freedom (d.f.).

Piecewise PH model

The piecewise PH model (5) incorporates non-proportional hazards in the Cox model by representing HR as a step function of time. HR is constant *within* each of r pre-specified time intervals but varies between the intervals. Within the ith interval ($i = 1,..,r$), the hazard is expressed by:

$$\lambda(t|x_1,\ldots,x_k) = \lambda_0(t)\exp[\Sigma(\beta_j + \gamma_{ij})x_j], \qquad (2)$$

where $\gamma_{1j} = 0$. The log HR equals β_j in the first inter-

val and $(\beta_j + \gamma_{ij})$ in the subsequent intervals $i = 2,\ldots,r$. The PH model (model 1) becomes a special case of the piecewise model (model 2), corresponding to $\gamma_2 = \ldots = \gamma_r = 0$. For a set of $k \geq 1$ variables, the PH hypothesis is tested by LRT with $k(r-1)$ d.f.

Our piecewise analyses were limited to univariate models with two and four intervals, delimited, respectively, by the median and the quartiles of the sample distribution of uncensored survival times. In the multivariable analyses, the algorithm for maximum partial likelihood estimation of the piecewise model failed to converge with both two and four intervals. These nonconvergence problems resulted probably from the difficulty in estimating the adjusted hazard ratios for some high-risk subgroups, such as cancer stage 4, in the last time interval, because almost all patients in these subgroups have died in the earlier intervals.

Parametric time-by-covariate interactions.

In his original paper, Cox (1) proposed to represent time-dependent effects by including in the PH model an interaction between the covariate X_j and a pre-specified parametric function $f_j(t)$:

$$\lambda(t|x_1,\ldots,x_k) = \lambda_0(t)\exp[\Sigma\{\beta_j + \gamma_j f_j(t)\}x_j] \quad (3)$$

where, for covariate X_j, log HR equals β_j at time $t = 0$ and equals $\beta_j + \gamma_j f_j(t)$ at $t > 0$. The conventional PH model corresponds to $\gamma_1 = \ldots = \gamma_k = 0$ and, for $k \geq 1$ variables, its validity is tested by LRT with k d.f. Model 3 can be estimated as a PH model with artificial time-dependent covariate(s) $Z_j(t) = X_j f_j(t)$.

Because we had no substantive grounds to select $f(t)$ a priori, in univariate analyses four different functions were estimated for each prognostic factor: linear (t), logarithmic $(\ln(t))$, quadratic (t^2) and inverse $(1/t)$.

In the multivariable analyses, four "homogeneous" versions of model 3 were specified a priori. Each of these models represented time-dependent effects of all prognostic factors by one of the four functions listed above. The fifth model, labeled "optimal," was specified a posteriori and was "heterogeneous" as the effect of each factor was represented by the function that fitted best in the corresponding univariate analysis.

Regression spline model

In the non-parametric regression spline model (6):

$$\lambda(t|x_1,\ldots,x_k) = \lambda_0(t)\exp[\Sigma\beta_j(t)x_j] \quad (4)$$

the log HR becomes a smooth function of time $\beta_j(t)$ and is approximated by a linear combination of piecewise quadratic polynomials known as B-splines (14, 15). The shape of $\beta_j(t)$ is estimated directly from empirical data, eliminating restrictive a priori assumptions and the need to select the best fitting model a posteriori (16, 17).

Flexibility is controlled by d.f., selected depending on sample size and expected complexity of the HR functions to be estimated (6). For large samples, a 5 d.f. model, consisting of three quadratic polynomial pieces, is recommended (6). The 1 d.f. PH model (model 1) is a special case of the spline model (model 4), corresponding to $\beta_j(t) = \beta_j$. Thus, for a set of $k \geq 1$ variables, the PH hypothesis is tested by LRT with $4k$ d.f. The precision of the estimate at time t is assessed by pointwise confidence intervals (6).

Data analytical procedures

In univariate analyses, unadjusted effects of all prognostic factors were estimated mainly to compare properties of different models. However, because of the risks of confounding bias, these univariate results were *not* considered a reliable basis for the decisions regarding multivariable modeling. Thus, in the initial multivariable model, all prognostic factors were included and represented by time-dependent effects. Next, all effects that were not statistically significant at the 0.05 level were removed. In the following, "(non)significant" stands for "statistically (non)significant at the 0.05 significance level." Significance of p dummy variables related to a multi-categorical factor was jointly assessed by the global p d.f. LRT test. Because the null hypothesis (log HR = 0) is a special case of the PH hypothesis (log HR = β), the rejection of the latter indicates that the predictor has a significant effect on survival (6). In the final model, HR associated with each selected variable was represented as either time-dependent or constant, depending on whether the PH hypothesis was rejected. To facilitate comparisons between models, no interactions between covariates were considered.

Akaike Information Criterion (AIC) (18) was employed to compare the goodness-of-fit of different models, while accounting for the differences in d.f. Lower AIC values indicate better fit. If a more complex of the two models yields lower AIC, then the worse fit of the simpler model is likely due to under-fit bias (19). In the opposite case, a visual comparison of the estimates may help in assessing to what extent the more complex model over-fits (19). In general, differences in AIC below 4.0 may be considered as "minor" and differences above 10.0 as "important."

Software

The Cox PH model (model 1), the piecewise model (model 2), and the parametric interactions (model 3)

were implemented using the BMDP program 2L (20). The spline model 4 (model 4) was estimated using a customized software in C (6), available from the authors.

RESULTS

Conventional analyses

There were 1,514 deaths among 1,951 patients followed for up to 15 years, with the median survival time of 22 months. Table 1 shows the distributions of prognostic factors and the results of univariate and multivariable Cox model analyses. Almost all effects are statistically very significant. Given the large number of deaths, the power is of no concern and all nonsignificant predictors are likely to be dismissed as definitely irrelevant. For example, patients diagnosed in the first period (1977–1981) and the second period (1982–1986) seem to have the same risks of mortality ($p = 0.953$ in table 1). However, the validity of this conclusion relies on the PH assumption.

Figures 1 and 2 show the Kaplan-Meier (21) survival curves for three periods of diagnosis and four cancer stages, respectively. In epidemiologic studies, the decision to accept the PH hypothesis often depends on whether these curves cross each other. However, very high initial mortality makes it difficult to separate curves in figure 1. Moreover, while the curves for the two first periods seem to cross twice, it is not clear if this departure from the PH assumption is statistically significant.

By contrast, stage-specific curves in figure 2 are well separated, suggesting the PH assumption is met.

Univariate time-dependent analyses

Table 2 presents univariate tests of the PH hypothesis. Rows correspond to different prognostic factors and columns to different models. Table 3 compares goodness-of-fit of different models. For each factor, the best fitting model (minimum AIC) is identified by "–". For other models in the same row, the difference from the minimum AIC is reported, with lower values indicating better fit.

Piecewise models

P values in the first two columns of table 2 show that using both two- and four-intervals models, the PH

TABLE 1. Univariate and multivariable Cox proportional hazards analyses in a population-based study of 1,951 French colon cancer patients, 1977–1991

Predictor	No. of patients	Univariate analysis			Multivariable analysis		
		Hazards ratio	95% CI*	*p* value	Hazards ratio	95% CI*	*p* value
Age (years)							
<65	496	1.00†			1.00		
≥65	1,455	1.60	1.41, 1.81	<0.001	1.65	1.45, 1.87	<0.001
Residence							
Urban	1,221	1.00†			1.00		
Rural	730	1.18	1.07, 1.31	0.002	1.03	0.93, 1.14	0.620
Sex							
Female	968	100†					
Male	983	1.06	0.95, 1.17	0.293	1.17	1.06, 1.30	0.003
Stage							
1	296	1.00†			1.00		
2	644	1.64	1.36, 1.98	<0.001	1.50	1.24, 1.82	<0.001
3	387	2.55	2.10, 3.11	<0.001	2.51	2.06, 3.06	<0.001
4	624	7.69	6.37, 9.30	<0.001	7.50	6.19, 9.08	<0.001
Time period							
1977–1981	578	1.00†			1.00		
1982–1986	653	0.91	0.80, 1.02	0.117	1.00	0.88, 1.13	0.953
1986–1991	720	0.73	0.64, 0.83	<0.001	0.82	0.72, 0.93	0.002
Site							
Left	1,134	1.00†			1.00		
Right	817	1.25	1.13, 1.38	<0.001	1.20	1.09, 1.34	<0.001

* CI, confidence interval.
† Referent category.

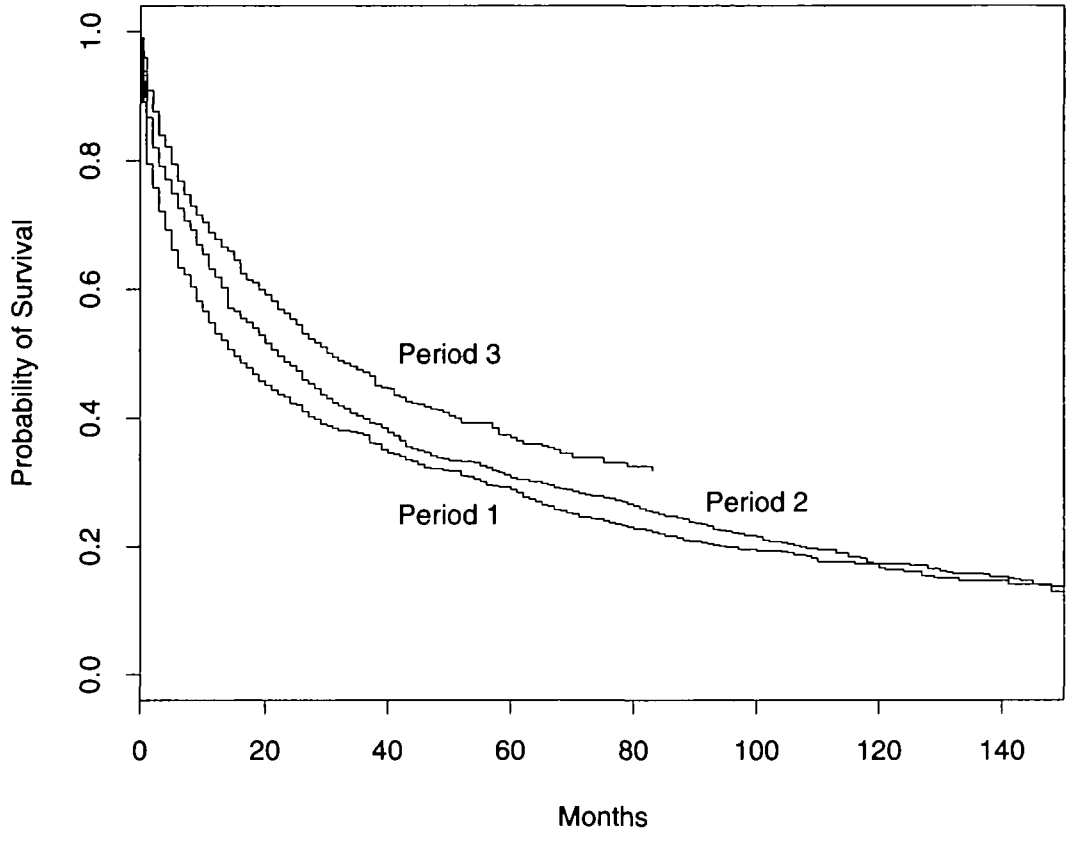

FIGURE 1. Kaplan-Meier survival curves for three periods of colon cancer diagnosis in a population-based study of 1,951 French colon cancer patients: period 1 (1977–1981), period 2 (1982–1986), and period 3 (1987–1991). The estimated probability of survival is plotted against follow-up time, measured in months. Period 3 curve ends at about 7 years (84 months) because the follow-up was terminated in December 1994.

assumption is definitely rejected for both period variables. Moreover, risks for patients enrolled in period 2, compared with period 1, change significantly over time according to all other models in table 2. Thus, the conclusion of no difference in the survival between the first and second periods, implied by the PH model (table 1), is incorrect.

However, table 2 demonstrates also that some conclusions depend on the number of intervals. For age and stage 4, the non-proportionality of hazards is very significant with four intervals ($p < 0.001$), but not with two intervals. Table 3 shows that the 4-interval model fits these variables considerably better, suggesting under-fit bias of the simpler model.

Parametric interactions

Columns 3–6 of tables 2 and 3 refer to interactions between predictors and parametric functions of time. For each prognostic factor, the "T" symbol in table 3 identifies the best fitting of the four parametric models. Each of the four models fits best the effect of at least one predictor but fits quite poorly in at least one other case. For example, while the quadratic model is definitely superior to the inverse model for cancer stage, it is quite inferior for period variables and both differences in AIC are very important (above 40.0 and 20.0, respectively). Thus, restricting parametric analyses of all predictors to any single function would induce considerable bias in some estimates. Moreover, while six variables have significant time-dependent effects in at least one model, this occurs in all four models only for period 2 (table 2).

In addition, even a significant estimate may be considerably biased. For stage 3, the inverse ($1/t$) model yields a significant result (table 2) and the estimate in table 4 suggests that HR *increases* with increasing follow-up time. However, the quadratic model, that shows a significant *decrease* (tables 2 and 4), fits the effects of stage much better (table 3). For cancer site, the results are even more difficult to interpret. Whereas three models are significant (table 2) and the differences in their AIC values in table 3 are very small, the corresponding estimates in table 4 show opposite directions of changes.

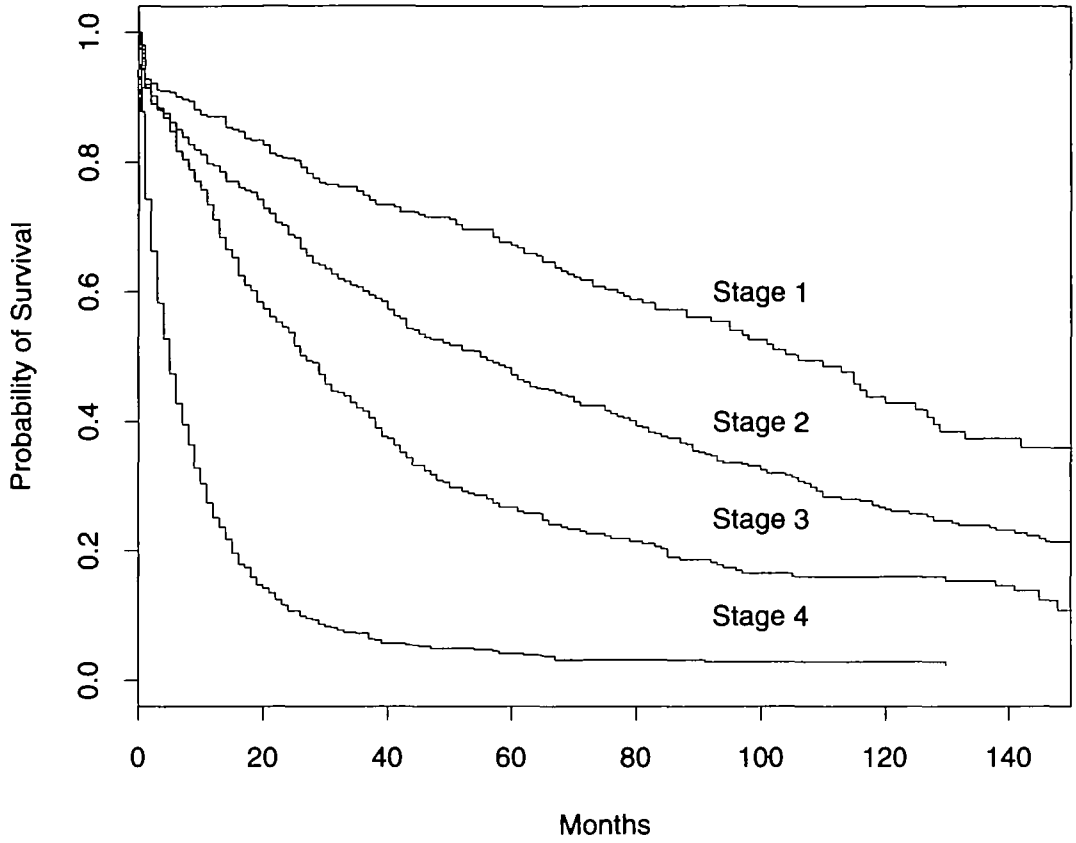

FIGURE 2. Kaplan-Meier survival curves for four cancer stages, determined at the time of colon cancer diagnosis in a population-based study of 1,951 French colon cancer patients. The estimated probability of survival is plotted against follow-up time, measured in months.

TABLE 2. Testing the proportional hazards model (PH) hypothesis for each predictor in a population-based study of 1,951 French colon cancer patients, 1977–1991: *p* values based on univariate models*

Predictor	Piecewise PH model		Parametric interactions				Splines
	r† = 2	*r* = 4	Time	Time²	ln(time)	1/time	
Age	0.133	<0.001	0.027	0.001	0.377	0.004	<0.001
Residence	0.242	0.268	0.127	0.420	0.094	0.394	0.270
Sex	0.148	0.264	0.055	0.124	0.090	0.301	0.390
Stage (global)	<0.001	<0.001	<0.001	<0.001	<0.001	0.007	<0.001
1‡	–	–	–	–	–	–	–
2	0.932	0.855	0.114	0.061	0.624	0.813	0.200
3	0.217	0.127	0.017	0.007	0.888	0.033	<0.001
4	0.061	<0.001	<0.001	<0.001	0.008	0.974	<0.001
Time period (global)	0.002	0.002	0.008	0.048	<0.001	<0.001	<0.001
1977–1981‡	–	–	–	–	–	–	–
1982–1986	0.002	0.014	0.004	0.014	<0.001	<0.001	0.020
1986–1991	0.002	<0.001	0.034	0.494	<0.001	<0.001	0.001
Site	0.005	0.034	0.012	0.039	0.013	0.171	0.026

* *p* values for the test of the H_0 of constant hazard ratio against H_1 of time-dependent hazard ratio obtained with a corresponding model-based likelihood ratio test (see section on statistical modeling).

† *r*, number of intervals used.

‡ Referent category.

TABLE 3. Goodness-of-fit of univariate time-dependent models in a population-based study of 1,951 French colon cancer patients, 1977–1991: differences in Akaike Information Criterion (AIC) values*

Predictor	Piecewise PH†		Parametric interactions				Splines (5 df)
	r‡ = 2 (2 df§)	r = 4 (4 df)	Time (2 df)	Time² (2 df)	ln(time) (2 df)	1/time (2 df)	
Age	20.0	5.0	17.4	11.6¶	21.6	13.8	–
Residence	1.6	3.0	0.6	2.2	–¶	2.2	5.0
Sex	1.6	3.6	–¶	1.2	0.8	2.6	7.0
Stage	69.8	25.4	36.4	32.0¶	72.2	78.6	–
Period	13.2	9.6	16.6	20.2	4.0	–¶	3.4
Site	–	3.0	1.4	3.4	1.4¶	5.8	4.8

* In each row, – indicates the best fitting model (minimum AIC). The numbers in the same row show the difference in AIC values between respective models and the "best" model with smaller numbers indicating better fit.
† PH, proportional hazards model.
‡ r, number of intervals used.
§ df, degrees of freedom.
¶ Indicates the best fitting among four parametric interaction models.

Regression spline modeling

P values in the rightmost column of table 2 indicate that spline-based LRT detected a significant time-dependent effect in all cases in which at least one other test was significant. These results, based on a single a priori selected model, provide unambiguous evidence of significant changes in the impact of stage 3 and age, for which simpler models yielded discrepant results. The PH hypothesis is definitely rejected for cancer stage ($p < 0.001$ for the global test in table 2), even though the four stage-specific curves in figure 2 are well separated. Thus, the acceptance of the PH hypothesis based on the visual assessment of Kaplan-Meier curves may be risky and alternative graphical techniques may be considered. Schemper (22) compared different methods for assessing PH assumption and found that Arjas cumulative hazards plots (23) performed quite well.

The AIC values in table 3 show that for age and stage the spline model fits the data considerably better than any of the six simpler models. Thus, the flexibil-

TABLE 4. Unadjusted hazard ratio estimates for cancer stage 3 and site at selected times in a population-based study of 1,951 French colon cancer patients, 1977–1991

Predictor and model	Time					
	1 month	6 months	1 year	2 years	5 years	10 years
Stage 3 vs. stage 1						
Cox PH*	2.55	2.55	2.55	2.55	2.55	2.55
2 intervals	2.15	2.15	2.86	2.86	2.86	2.86
4 intervals	1.34	2.86	2.61	2.61	2.24	2.24
Time	3.38	3.26	3.14	2.87	2.23	1.47
Time²	3.08	3.07	3.04	2.91	2.15	0.73
ln(time)†	–	–	–	–	–	–
1/time	1.67	2.51	2.71	2.83	2.90	2.93
Splines	1.07	3.47	4.11	4.23	2.47	1.18
Site (right vs. left)						
Cox PH*	1.25	1.25	1.25	1.25	1.25	1.25
2 intervals	1.44	1.44	1.08	1.08	1.08	1.08
4 intervals	1.35	1.54	1.08	1.08	1.08	1.08
Time	1.37	1.35	1.31	1.25	1.08	0.85
Time²	1.30	1.30	1.30	1.31	1.35	1.51
ln(time)†	1.47	1.31	1.23	1.16	1.07	1.00
1/time	–	–	–	–	–	–
Splines	1.33	1.57	1.29	1.07	1.11	0.97

* PH, proportional hazards model.
† As the PH hypothesis was not rejected in this model (table 2), the time-dependent estimate should not be interpreted.

ity of the non-parametric model was necessary to account for the complex changes in their effects. For stage 3, the spline estimate in table 4 initially increases and then decreases. Simple parametric models are unable to represent such non-monotone changes and each is *locally biased*, which explains why their estimates diverge (table 4).

A similar pattern is seen in figure 3, which explores the discrepancies between different estimates for stage 4. The constant PH model-based estimate considerably underestimates early risks and overestimates later risks, compared with most time-dependent estimates, that all suggest very significant changes over time (table 2). The bias is important as the 95 percent confidence intervals (constant dotted lines in figure 3d) do not overlap with the spline-based pointwise confidence intervals (vertical bars), both around the early peak in HR and near the end of follow-up. Interestingly, at 25 months, when the two point estimates are almost identical, the two confidence intervals are equally narrow, suggesting similar precision. The spline model (solid curve in figure 3d) offers a *smooth* representation of a complex pattern of changes. By contrast, figure 3c shows that parametric interaction models are unable to recover the initial rise-and-fall phase of changes in HR. Finally, the 2-interval piecewise model (figure 3a) fits poorly because major changes occur *within* each of the two intervals. Increasing the number of intervals to four (figure 3b) reduces the bias, but yields a clinically implausible estimate, with big "jumps."

For other variables, the differences in AIC between the spline and the best-fitting among the simpler models reflect mostly the penalty for the additional d.f.'s (table 3). Thus, for example, a (slight) non-monotonicity of the non-parametric estimate for cancer site (table 4) may be due to overfitting.

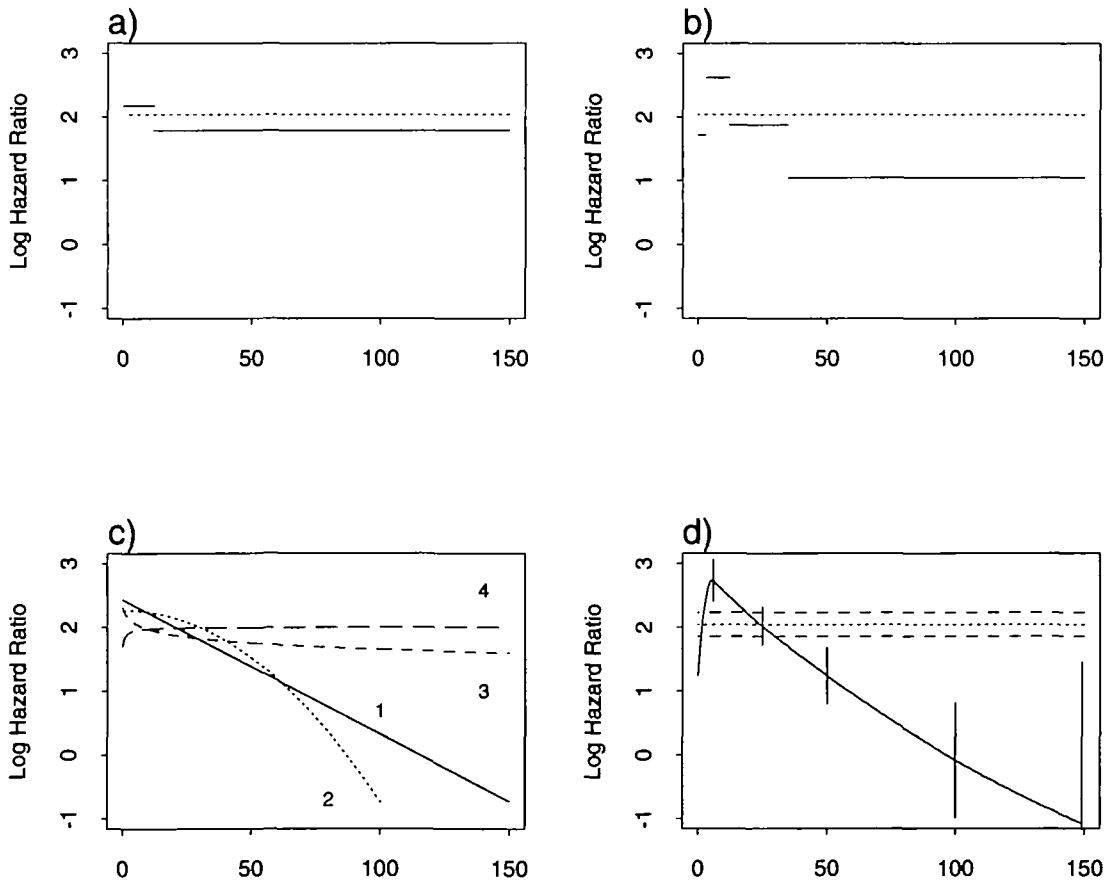

FIGURE 3. Comparison of unadjusted hazard ratio (HR) estimates for stage 4 (vs. stage 1) of colon cancer, obtained with different univariate models. HR for all-causes mortality, associated with stage 4, is plotted against time since colon cancer diagnosis (in months). In panels a) and b) dotted lines represent the constant HR estimate yielded by the conventional PH model and the solid lines represent piecewise PH estimates with two and four time intervals, respectively. Panel c) shows estimates obtained with four parametric models of time-by-predictor interactions: 1, linear; 2, quadratic; 3, logarithmic; 4, inverse. Panel d) compares constant PH estimate (dotted line) and its 95% confidence limits (dashed lines) with the 5 d.f. spline estimate (solid curve), for which pointwise 95% confidence intervals (vertical bars) are shown at selected points in time.

Multivariable time-dependent analyses

The second row of table 5 shows that, for each time-dependent model, the global test rejected the PH assumption, indicating a very significant improvement in fit. Thus, the PH model fails to correctly represent the effects of prognostic factors on survival of colon cancer patients. The top row of table 5 compares the fit to data, in terms of AIC values, of multivariable time-dependent models. Each other row shows the p values for testing time-dependence of a particular variable, adjusted for time-dependent effects of all other variables.

Multivariable parametric models

Table 5 shows that while the quadratic and linear models provide much better *overall* fit to data than two other "homogeneous" parametric models, none of the four models fits well the effects of *every* prognostic factor. Moreover, as in univariate analyses, the four models disagree with respect to the significance of time-dependent effects of particular variables. As expected, the "optimal" model (fifth column) detects all significant time-dependent effects and fits the data considerably better. However, specific time-dependent functions in the optimal model (e.g., quadratic, linear, logarithmic, and inverse functions for, respectively, age, sex, residence, and period) were selected a posteriori, in order to optimize the fit to our data. As in the classic multiple testing context, this induces optimistic bias (6, 24). The p values are optimistically low, the goodness-of-fit and the predictor effects are overestimated, and the magnitude of this bias increases with the number of alternative models considered (6, 25). As our "optimal" model was specified by selecting, independently for each of six predictors, the best-fitting among four parametric functions, the resulting bias is probably big enough to undermine the reliability of conclusions.

Interpretation of spline estimates

AIC values in table 5 show that the spline model overall fits much better than even the "optimal" parametric model. Moreover, a priori selected spline model detected significant time-dependent effects more frequently than any of the four homogeneous parametric

TABLE 5. Multivariable modeling and testing of time-dependent effects in a population-based study of 1,951 French colon cancer patients, 1977–1991: p values for tests of the proportional hazards model (PH) hypothesis

	Parametric model					Splines
	Time	Time²	ln(time)	1/time	"Optimal"	
Model's AIC*	33.4	28.8	70.9	75.4	12.4	
Overall test of PH†	79.0	83.6	41.5	37.0	100.0	166.4
	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001
Age	0.005	<0.001	0.733	0.013	<0.001	<0.0001
Residence	0.204	0.581	0.201	0.518	0.127	0.487
Sex	0.008	0.016	0.053	0.385	0.005	0.089
Stage (global)	<0.001	<0.001	0.003	0.018	<0.001	<0.0001
1‡	–	–	–	–	–	–
2	0.582	0.433	0.885	0.562	0.310	0.535
3	0.152	0.038	0.463	0.019	0.029	0.002
4	<0.001	<0.001	0.079	0.508	<0.001	<0.0001
Time period (global)	0.034	0.073	0.002	<0.001	<0.001	0.003
1977–1981‡	–	–	–	–	–	–
1982–1986	0.011	0.023	0.001	0.002	0.002	0.002
1986–1991	0.195	0.736	0.006	<0.001	<0.001	0.006
Site	0.010	0.078	0.017	0.285	0.015	0.031

* Spline model had the lowest (best) Akaike Information Criterion (AIC) value. For other models, the difference in AIC from the spline model is shown with lower values indicating better fit.

† Likelihood ratio test (LRT) based on the difference in log likelihood between a respective time-dependent model and the PH model (9 degrees of freedom (d.f.) for parametric models, 36 d.f. for spline model).

‡ Referent category.

models. These significant results indicate that the corresponding time-dependent estimates fit better than the constant HR estimates and, thus, reflect systematic trends in the data rather than over-fitting (6). Figure 4 shows the spline estimates of the adjusted effects of selected prognostic variables, and pointwise 95 percent confidence intervals. The variation of the shapes presented in figure 4 and figure 3d (stage 4) is rather remarkable, given that all these shapes were estimated using a single model. However, while statistical significance of a given time-dependent effect allows for a meaningful interpretation of the general pattern of changes, confidence intervals have to be taken into account when assessing finer aspects of the estimate (6). For example, the confidence intervals in figure 4 consistently become wider during late follow-up, indicating that pattern of changes cannot be reliably estimated beyond 10 years after diagnosis.

The spline estimate in figure 3d shows that the effect of stage 4 is rather modest at the very beginning but immediately increases sharply, reaching the peak in the second year, and then gradually declines with increasing follow-up. To interpret this pattern, we take into account that, whereas we don't know the actual causes of death, very early mortality in colon cancer is caused mostly by post-surgical complications. Thus, initial cancer stage appears to have a relatively minor impact on (early) post-surgical fatality but becomes very important after few months, when mortality reflects mainly disease severity. However, with increasing follow-up duration, the initial stage becomes a less adequate proxy for the *current* unknown cancer severity, because patients presumably progress at different rates. This decline in the predictive ability of the initial cancer stage suggests that current stage should be updated after a few years. A simi-

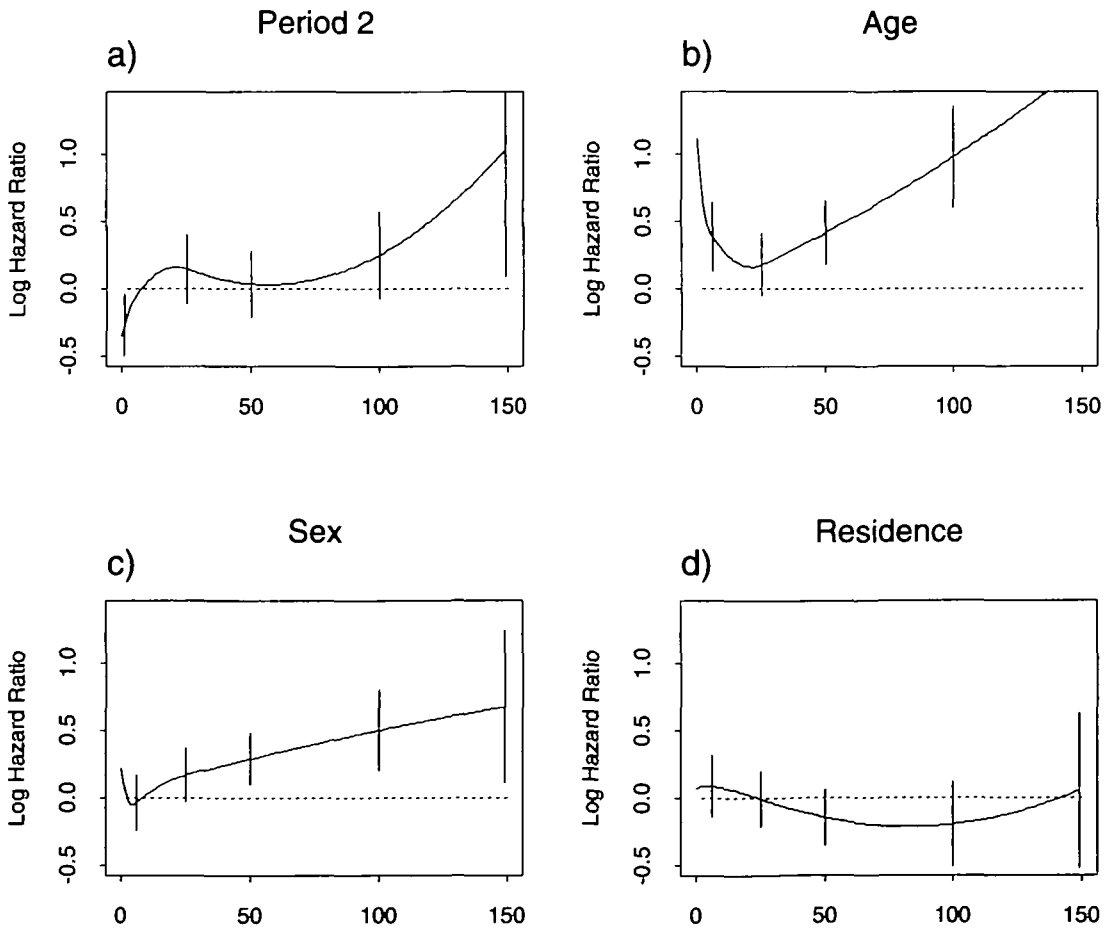

FIGURE 4. Non-parametric 5 d.f. estimates of adjusted time-dependent effects of selected prognostic factors for all-causes mortality in French colon cancer patients. Each panel shows a hazard ratio estimate (solid curve) obtained from the multivariable spline model and the corresponding pointwise 95% confidence intervals (vertical bars) at selected times: a) period 2 (diagnosis in 1982–1986) vs. period 1 (1977–1981); b) age at diagnosis (>65 years vs. <65 years); c) sex (male vs. female); d) residence (rural vs. urban). Horizontal lines correspond to the relative risks of 1.0.

lar pattern was observed for stage 3 (table 4), while for stage 2 the changes were too small to reach statistical significance (table 5).

Figure 4a shows the estimated HR for patients diagnosed in period 2 (1982–1987). The estimate suggests that post-surgical mortality was reduced compared with period 1 (1977–1981), but there was no improvement in (later) cancer-related mortality. The fact that the estimated effect of period 2 inverts, from an early "protective" effect (HR below 1.0) to a slightly detrimental effect later on, explains why this effect was completely nonsignificant in the multivariable PH model ($p > 0.9$ in table 1). The PH estimate, constrained to be constant, represents the average-over-time of the actual HR, which in the case of risk inversion is close to 1.0 (6). Whereas the general shape of the estimate for period 3 (not shown) was similar to period 2, the HR did not exceed 1.0 over the entire follow-up period, explaining its significance in the PH model (table 1).

Figure 4b shows that the pattern of changes in the effect of patient's age is opposite to that observed for higher cancer stages. While elderly patients have much higher risks of post-surgical mortality, the impact of age is practically null in the second and third year and increases only later. Figure 4c shows a similar increasing pattern for the effect of sex (male vs. female). Both estimates may partly reflect the effect of patient's demographics on mortality from causes other than colon cancer, the share of which likely increases with the aging of the cohort. The estimate for sex is almost perfectly linear so that the test based on a complex 5 d.f. regression spline model becomes inefficient, yielding a marginally nonsignificant result (table 5). Finally, figure 4d shows convincingly that the patient's residence (rural vs. urban) has no any effect on mortality.

After having removed all nonsignificant effects, in the final multivariable spline model age, cancer stages 3 and 4, cancer site and both period variables are represented by time-dependent spline HR estimates, whereas stage 2 and sex are represented by the constant HR.

DISCUSSION

Our results demonstrate the importance of accounting for non-proportionality of hazards and advantages of flexible modeling. Restricting analyses to the conventional PH model resulted in incorrect conclusions regarding the nonsignificance of period 2 and substantially biased estimates for higher cancer stages. The ability to reject the incorrect PH assumption depends, however, on the method. The example of cancer stages illustrates the risks of relying on visual assessment of Kaplan-Meier curves. Piecewise models (5) yield "jumpy," clinically implausible, estimates (figures 3a, b) and are difficult to implement in multivariable modeling. Moreover, the results depend on the arbitrary number of time intervals.

Parametric modeling of time-by-predictor interactions in the Cox model (1) implies choosing between two sub-optimal alternatives. Restricting the analyses to a single a priori selected parametric function may result in biased estimates and/or loss of statistical power, if the actual pattern of changes is not consistent with the shape of the selected function. An alternative strategy requires estimating different parametric functions and selecting, a posteriori, the best fitting model. However, in the case, for example, of cancer site, very different parametric estimates fit the data equally well (tables 3 and 4). A similar phenomenon was illustrated in a recent study of various parametric models of exposure intensity (27). Moreover, a posteriori model selection invalidates conventional statistical inference and induces overestimation bias, especially in multivariable analyses which entail data-dependent decisions regarding each of several predictors (6, 24, 25). Whereas testing a posteriori selected model in an independent study might corroborate the findings, it may be difficult to find a comparable data set of similar size and quality. Moreover, simulations suggest that the results of a posteriori identification of multivariable Cox models are quite unstable (28), making such a corroboration rather unlikely. Finally, as illustrated in figure 3, it is possible that even if several parametric functions are considered, none is able to represent the complex pattern of actual changes.

By contrast, the regression spline method (6) accurately represented *both* simple and complex patterns of changes with a single a priori selected model, that produced smooth, clinically plausible estimates. The spline model fitted well the effects of *all* predictors and offered the best fit in multivariable analyses. The variety of shapes revealed by non-parametric estimates yielded potentially important insights into the role of some prognostic factors. Finally, the model-based test of the PH hypothesis detected non-monotone changes and avoided problems related to a posteriori model selection.

Some limitations of our study should be acknowledged. First, we could not separate deaths due specifically to colon cancer. Thus, our estimates may partly reflect an increase in the proportion of deaths from other causes, due to aging of the cohort. This could explain why, with increasing follow-up, the impact of cancer-specific factors (cancer stage) decreases while the predictive ability of "generic" predictors of mortality, such as patient's age and sex, increases. Ideally, relative survival approach could be used to establish if the estimated effects of patient's age and sex simply

reflect their impact on overall mortality in the general population, rather than their effects specific to colon cancer (29, 30). Unfortunately, we are not aware of any method or software that combines estimation of time-dependent HR with relative survival modeling. Even so, we expect the methodologically relevant conclusions of our study to be rather robust. The implications of parametric assumptions underlying conventional models and the advantages of the splines over simpler time-dependent models were most evident in estimating the effects of higher cancer stages, especially during the early follow-up, and natural mortality probably has only a very minor impact on these estimates.

Our data and our analyses were limited to fixed covariates, measured at time of diagnosis. A more refined analysis could use time-dependent covariates to incorporate information on changes that occurred during follow-up. For example, whereas treatment was usually determined very shortly after the diagnosis, time-dependent covariates could reflect later changes in treatment of some patients. Recently Heinzl et al. (31) proposed a model that incorporates time-dependent effects of binary time-dependent covariates, and the SAS macros developed by those authors may facilitate analyses of non-proportional hazards for both fixed and time-dependent covariates (32).

Finally, our arbitrary choice of specific numbers of intervals in the piecewise model and of parametric time-by-predictor interactions should not affect our findings. For example, although increasing further the number of parametric models could increase the probability that *one* of these models would fit the effect of a given predictor as well as the spline model, it would further increase problems related to inference based on a posteriori selected model and to discrepancies between competing parametric estimates.

We demonstrated the advantages of flexible modeling of non-proportional hazards in a large multivariable data set. With more than 1,500 deaths, we had no numerical problems estimating a spline model with 45 d.f. (5 d.f. for each of nine variables). However, in smaller samples, it may be difficult to estimate models of that complexity. Moreover, simulations in Abrahamowicz et al. (6) show that at least 10 observed deaths for each model's d.f. are required to ensure accurate inference. Thus, multivariable analyses of smaller data sets will require some a priori decisions regarding such issues as 1) which predictor effects will be considered as possibly time-dependent; and/or 2) how many d.f. should be used to model specific effects. Further research is necessary to develop sound methodological strategies to deal with such situations. One efficient solution may be to rely on relevant previous analyses of larger data sets when specifying plausible parsimonious models for smaller studies. Analyses similar to ours may provide some insights about the clinical importance of changes over time in particular effects and about possible patterns of these changes. For instance, our estimates suggest that in future analyses the time-dependent effect of sex may be represented by a simple linear function whereas considerable flexibility is necessary for higher cancer stages or age. Finally, the time-dependent effect of cancer site, while statistically significant, is of limited clinical importance so that representing this effect by a constant hazard ratio would induce only minor bias. Thus, the results of complex modeling may ultimately suggest some model simplifications. However, such empirically valuable findings would not be possible if the model were a priori constrained to yield simple estimates.

ACKNOWLEDGMENTS

This research was supported by the INSERM/FRSQ scholarship and NSERC grant awarded to Dr. M. Abrahamowicz and by grants from the Burgundy Regional Council, the French Ligue Bourguignonne contre le Cancer, and the French Association for Research against Cancer (ACR). Dr. Abrahamowicz is a Health Scientist of the Medical Research Council of Canada.

The authors thank Roxane du Berger for valuable comments and assistance and Dr. Jack Siemiatycki for his many helpful suggestions in regard to the study.

REFERENCES

1. Cox DR. Regression models and life tables (with discussion). J R Stat Soc B 1972;4:187–200.
2. O'Quigley J, Pessione F. The problem of a covariate-time qualitative interaction in a survival study. Biometrics 1991;47:101–15.
3. Lin DY, Wei LJ, Zing Z. Checking the Cox model with cumulative sums of martingale-based residuals. Biometrika 1993;80:557–72.
4. Altman DG, De Stavola BL, Love SB, et al. Review of survival analyses published in cancer journals. Br J Cancer 1995;72:511–18.
5. Moreau T, O'Quigley J, Mesbah M. A global goodness-of-fit statistic for the proportional hazards model. Appl Stat 1985;34:212–18.
6. Abrahamowicz M, MacKenzie T, Esdaile JM. Time-dependent hazard ratio: modeling and hypothesis testing with application in lupus nephritis. J Am Stat Assoc 1996;91:1432–9.
7. Gray RJ. Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis. J Am Stat Assoc 1992;87:942–51.
8. Hastie TJ, Tibshirani RJ. Varying-coefficient models (with discussion). J R Stat Soc (B) 1993;55:757–96.
9. Hess KR. Assessing time-by-covariate interactions in proportional hazards regression models using cubic spline functions. Stat Med 1994;13:1045–62.
10. Kooperberg C, Stone CJ, Truong YK. Hazard regression. J Am

Stat Assoc 1995;90:78–94.
11. Zucker DM, Karr AF. Nonparametric survival analysis with time-dependent covariate effects: a penalized partial likelihood approach. Ann Stat 1990;18:329–53.
12. Abrahamowicz M, du Berger R, Grover SA. Flexible modeling of the effects of serum cholesterol on coronary heart disease mortality. Am J Epidemiol 1997;145:714–29.
13. Dukes CE. The classification of cancer of rectum. J Pathol Bacteriol 1932;35:323–32.
14. de Boor C. A practical guide to splines. New York: Springer, 1978.
15. MacKenzie T, Abrahamowicz M. B-Splines without divided differences. Student 1996;1(4):223–30.
16. Ramsay JO. Monotone regression splines in action, (with discussion). Stat Sci 1988;3:425–61.
17. Wegman EJ, Wright JW. Splines in statistics. J Am Stat Assoc 1983;78:351–66.
18. Akaike H. A new look at statistical model identification. IEEE Transactions on Automatic Control. AC 1974;19:716–23.
19. Abrahamowicz M, Ciampi A. Information theoretic criteria in nonparametric density estimation: bias and variance in the infinite dimensional case. Computat Stat Data Anal 1991;21: 239–47.
20. Dixon WJ, Brown MB, eds. BMDP statistical software. Berkeley, CA: University of California Press, 1990.
21. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. J Am Stat Assoc 1958;53:457–81.
22. Schemper M. Cox analysis of survival data with non-proportional hazard functions. Statistician 1992;41:455–65.
23. Arjas E. A graphical method for assessing goodness of fit in Cox's proportional hazards model. J Am Stat Assoc 1988;83: 204–12.
24. Hurvich CM, Tsai CL. The impact of model selection on inference in linear regression. Am Statist 1990;44:214–17.
25. Lausen B, Schumacher M. Maximally selected rank statistics. Biometrics 1992;48:73–85.
26. Hastie TJ, Tibshirani RJ. Generalized additive models. New York: Chapman & Hall, 1990.
27. Vacek PM. Assessing the effects of intensity when exposure varies over time. Stat Med 1997;16:505–13.
28. Altman DG, Anderson PK. Bootstrap investigation of the stability of a cox regression model. Stat Med 1989;8:771–83.
29. Esteve J, Benhamou E, Croasdale M, et al. The relative survival and the estimation of the net survival: elements for further discussion. Stat Med 1990;4:529–38.
30. Monnet E, Boutron MC, Arveux P, et al. Different multiple regression models for estimating survival: use in a population-based series of colorectal cancers. J Clin Epidemiol 1992;45: 267–73.
31. Heinzl H, Kaider A, Zlabinger G. Assessing interactions of binary time-dependent covariates with time in cox proportional hazards regression models using cubic spline functions. Stat Med 1996;15:2589–601.
32. Heinzl H, Kaider A. Gaining more flexibility in cox proportional hazards regression models with cubic spline functions. Comput Methods Prog Biomed 1997;54:201–8.