*The International Journal of Biostatistics*

*Volume* 5, *Issue* 1 2009 *Article* 3

# Fitting Smooth-in-Time Prognostic Risk Functions via Logistic Regression

James A. Hanley* Olli S. Miettinen†

*McGill University, james.hanley@mcgill.ca

†McGill University, olli.miettinen@mcgill.ca

# Fitting Smooth-in-Time Prognostic Risk Functions via Logistic Regression*

James A. Hanley and Olli S. Miettinen

## Abstract

When considering treatment options, a physician ideally has access to prognoses for various spans of prospective time, meaning known risks specific for these and also for both treatment and the profile of the patient. Accordingly, investigators ideally would report estimates of such risks from clinical trials and their non-experimental counterparts. To the extent that such risk estimates have been reported at all, they have mainly been based on the semi-parametric regression model of Cox. We focus on a family of fully-parametric hazard models of an attractive, versatile form that readily allows for non-proportionality, yet models that have not been easy to fit with standard statistical software. We elaborate an approach, recently proposed, to fitting such hazard functions via logistic regression. From the fitted hazard function, cumulative incidence and, thus, risk functions of time, treatment and profile can be derived. This approach accommodates any log-linear hazard function of prognostic time, treatment, and the prognostic indicators defining the patient's prognostic profile.

**KEYWORDS:** Cox regression, logistic regression, prognosis, risk function, survival analysis

---

*This work was supported by grants from The Natural Sciences and Engineering Research Council of Canada and Le Fonds Québécois de la recherche sur la nature et les technologies. The authors are also grateful to colleagues who have provided input to this work.

# 1 Introduction

The oldest and best known function for estimating profile-specific risks of illness is the one derived from the Framingham Heart Study data and addressing 10-year risk of coronary heart disease (NHLBI, 2008). The "Gail model" (NCI, 2008) provides estimates of a woman's 5-year risk for (an overt case of) breast cancer. Several recent authors (e.g., Cassidy, 2007; Spitz, 2007; Memorial Sloan-Kettering, 2008) have developed functions for estimating a smoker's risk of lung cancer, depending on whether (s)he discontinues smoking. Probability functions are also increasingly being developed for both diagnosis (e.g., Steyerberg et al., 2001, Bevilacqua et al., 2007; Partin, 2007) and illness-conditional prognosis (e.g., Califf et al., 1997; Kannel et al., 1999; Royston, 2002; Machin and Campbell, 2005; Tisman, 2007). The early – and still more common – functions for profile-specific risk have been for situations in which the risk period is so short that the timing of the outcome within it is irrelevant and the outcome is known for all study subjects, so that the risk function can be derived by logistic regression.

When the time of outcome within the risk period is of intrinsic interest, the object of study is a function that expresses the cumulative incidence ($CI$), or risk ($R$), as a function of time, treatment, and the indicators forming the prognostic profile. As for survival analyses in general, published in the major general medical journals, Julien and Hanley (2008) found that they rarely produce such prognostic functions, even though the requisite software is available in all of the Cox regression packages. They speculated that one of the reasons for this is that the resulting profile-specific $CI$ curves – or their complements, the survival curves – are estimated as steps-in-time functions rather than as smooth-in-time functions. Breslow (1972) did suggest a smoothed estimation of the baseline hazard function, which would lead to a smooth $CI$ function of time. However, the various step function estimators of the complement of $CI$, with as many steps as there are distinct failure times in the dataset, are more easily derived, and so they are the only ones that are available in most statistical packages.

In his review, Hjort (1992) surmised that "the success of Cox regression has perhaps had the unintended side-effect that practitioners too seldomly invest efforts in studying the baseline hazard" and suggested that "a parametric version, ... if found to be adequate, would lead to more precise estimation of survival probabilities." Royston and Parmar (2002), noting that this statement had been "apparently little heeded," were concerned that "in the Cox model, the baseline hazard function is treated as a high-dimensional

nuisance parameter and is highly erratic." Thus, they proposed to estimate it "informatively (that is, smoothly)," by natural cubic splines.

Particularly notable are Cox's own reflections (Reid, 1994) on the uses of his model:

> **Reid**: How do you feel about the cottage industry that's grown up around it [the Cox model]?
> **Cox**: Don't know, really. In the light of some of the further results one knows since, I think I would normally want to tackle problems parametrically, so I would take the underlying hazard to be a Weibull or something. I'm not keen on nonparametric formulations usually.
> **Reid**: So if you had a set of censored survival data today, you might rather fit a parametric model, even though there was a feeling among the medical statisticians that that wasn't quite right.
> **Cox**: That's right, but since then various people have shown that the answers are very insensitive to the parametric formulation of the underlying distribution [see, e.g., Cox and Oakes, Analysis of Survival Data, Chapter 8.5]. And if you want to do things like predict the outcome for a particular patient, it's much more convenient to do that parametrically.

We here address smooth-in-time hazard functions that allow estimation of risks specific to patient profiles, interventions, and various points in prognostic time, and of corresponding risk differences between the interventions. We focus on a family of parametric hazard models of an attractive, versatile form that have not been easy to fit with standard statistical software. Upon preview of the two datasets we use for illustration, we describe the family of parametric hazard models being considered, and how previous authors have managed to fit some of the simpler members of it. We then describe how, using a recently proposed approach (Miettinen, 2004; Miettinen, 2008), all of its members can easily be fitted with standard software. The approach is an adaptation of one that already is familiar to epidemiologists from the study of disease etiology. We illustrate the method, and proceed to show how to use the fitted hazard – incidence density – function to produce a table or nomogram from which estimates of profile- and intervention-specific risks for various spans of prognostic time can be read.

Table 1: For each of the two intervention groups ($I = 1$ for verum, $I = 0$ for placebo) in the Systolic Hypertension in the Elderly Program (SHEP, 1991), distributions of prognostic indicators; also shown are the respective numbers of subjects and strokes.

| I | Age: $Q_{10}$ | Age: $Q_{50}$ | Age: $Q_{90}$ | Gender: % male | Race: % Black | SBP: $Q_{10}$ | SBP: $Q_{50}$ | SBP: $Q_{90}$ | No. of subjects | PT of f-up | No. of strokes |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 64 | 72 | 81 | 43 | 14 | 161 | 168 | 183 | 2351 | 10,392 py | 158 |
| 1 | 64 | 72 | 81 | 44 | 14 | 161 | 168 | 185 | 2350 | 10,502 py | 105 |

$Q_{10}, Q_{50}$ and $Q_{90}$ are 10th, 50th and 90th centiles, respectively. SBP: systolic blood pressure.

# 2 Data used for illustrations

To illustrate the approach, we first use data from the Systolic Hypertension in the Elderly Program (SHEP, 1991). We obtained the data, without subject identifications, under the program NHLBI Datasets Available for Research Use (NHLBI, 2007). That study – a randomized trial – addressed the effectiveness of antihypertensive treatment in reducing the risk of stroke in persons with systolic hypertension. From what we received, we identified 4,701 subjects with complete data on age, gender, race, systolic blood pressure, and intervention (verum/placebo). In the study base of 20,894 person-years of follow-up, 263 first cases of stroke were identified. The data are summarized in Table 1. As the subjects in this study were otherwise healthy, the age-specific stroke hazard (incidence density) was virtually unaffected by time since entry into the study.

In studies with entry as of an acute event such as stroke, the hazard function usually is quite complicated a function of the earliest time since entry into the study. To illustrate the estimation of such a hazard function, we use data, previously analyzed by Efron(1988), from a study contrasting treatment with radiotherapy plus chemotherapy with radiotherapy alone for cancer of the head or neck.

# 3 Parametric models, as fitted

Some fully-parametric functions addressing the specific risks at issue here are possible to fit using packages for survival analysis. However, since the fitting routines are not accessible, the user cannot easily modify the time-component of the hazard function, notably in terms of involving products of the time variate(s) and others.

We focus on the family of hazard functions of the form

$$h(x,t) = \exp[g(x,t)], \tag{1}$$

where $t$ denotes the numerical value (number of units) of a point in prognostic/prospective time and $x$ the realization of the vector $X$ of variates based on the patient's profile and intervention (if any). The only constraint on the function $g(x,t)$ is that it be 'linear' in the parameters, in the 'linear model' sense. This family is quite inclusive in respect to the functional forms.

The simplest member of this family is the parametric hazard model with the longest history, the widest use in demography, and least common presence in statistical software packages: the Gompertz hazard model (Gompertz, 1825). It formulates the age-specific 'force of mortality,' instantaneous incidence density (Miettinen, 1976), or hazard as $h(t) = \exp(\beta_0 + \beta_1 t)$, where $t$ is measured from a given initial value of age. The 'proportional hazards' extension of this to $X = x$ can be written as the $h(x,t)$ in equation (1) above. More complex functions of $t$ in the linear compound $g(x,t)$ can be used to allow more versatile but still smooth-in-time hazard functions. Use of $\log(t)$ yields the Weibull model. The inclusion of product terms involving $t$ and some element of $x$ allows for non-proportional hazards.

If it be possible to fit the hazard (incidence density) function in equation (1), the cumulative incidence function, $CI(x,t)$, could then be derived – by numerical methods if necessary – from the fundamental relationship (Miettinen, 1976)

$$CI(x,t) = 1 - \exp\left[ - \int_0^t h(x,u)du \right]. \tag{2}$$

It is, however, generally impractical to do the fitting by means of the common statistical packages, for the likelihood is quite involved even in the absence of censoring. Whereas it is possible, with some ingenuity, to fit several of the other parametric survival functions iteratively using a generalized linear models package (Aitkin and Clayton, 1980), the programming required to fit log-linear hazard models by such a package is more complex (Clayton, 1983). Despite the claim that the `survreg` routine in `S` can fit Gompertz and Rayleigh functions, even the developer himself has had difficulty doing so (Therneau, 1996a, 1996b).

A number of authors have circumvented these technical problems of fitting by dividing the observed 'survival time' of each subject into a number of time-slices. One of us (Albertsen, Hanley et al., 1998) used this approach to fit hazard functions.

Efron (1977) pointed out that regardless of the form of $h(x,t)$, the likelihood contribution of the $j$th subject, followed for $t_j$ units of time, is equiv-

alent to the likelihood for a sequence of a large number, $n_j = t_j/\Delta$, of Bernoulli trials, with time-dependent probabilities of failure. For a trial that corresponds to the small interval $(t, t+\Delta)$, the failure probability can be well approximated by $p_i = h(x_j, t)\Delta$. The sequence ends with the $n_j^{th}$ trial, at the time of the event of interest or when follow-up was otherwise terminated.

In a subsequent article Efron (1988) focused on discretizations of the $t$-axis and on using logistic regression to fit various smooth-in-$t$ hazard and survival functions in the one-sample situation, where the usual non-parametric alternative is the Kaplan-Meier estimator of survival rate. He modeled the probability, $p_i$, that a patient suffers a first event in the interval $(t_i,\ t_i + \Delta_i)$, as

$$\text{logit}(p_i) = \beta\ x(t) + \log(\Delta_i),$$

where $x(t)$ is a vector of time-variate realizations, and $\log(\Delta_i)$ is an offset.

In that paper he discussed the possible models (including the basic Gompertz model) obtained as limits of his logistic regressions when the time intervals become extremely small. Extensions to the regression situation, including to the model analogous to the basic $h(x, t)$ model above, he mentioned (Remark H, p 423) but did not pursue. His earlier investigation (Efron, 1977) had found that the smooth-in-$t$ form of his extended regression model was not a big improvement on the semi-parametric version of Cox, "at least not for the estimation of $\beta$." However, he considered this form to be attractive when the interest is in "estimating the hazards, rather than just comparing them."

Efron (2002) used discretization ('time-slicing') to fit a fully-parametric hazard function of both follow-up (prognostic) and calendar time, using data on 110 cases in a study base of 2,673.4 patient-years, or 976,447 patient-days ($\Delta = 0.00274$, $n = 620$). In this analysis, even though he visualized the elementary distributions as being of Bernoulli type, Efron took it that "the Poisson model is more tractable." Also, because the size of his dataset – almost one million observations – was impractical for the available statistical software, he aggregated observations from different patients and treated the numbers of events in these aggregates as realizations of Poisson variates. In his 1988 example, 42 of the 51 patients in Arm A, and 31 of the 45 in Arm B, had had a recurrence of their disease. He addressed the events in the 600 and 945 patient-months of follow-up using separate-sample datasets with $N = 47$ and $N = 61$ binomial observations, respectively.

In the next section we show how one can use the case series together with a suitable sample of the base and logistic regression, as an alternative way to fit a fully-parametric log-linear model for $h(x, t)$. The fitted function can then be used to derive estimates of period-specific risks also specific to both

treatment and profile. Although we focus on a particular log-linear hazard formulation, an extension of the Gompertz form, the Weibull model as well as other more versatile ones within the log-linear family can also fitted by this simple technique.

# 4 The proposed approach to fitting

The proposed alternative way to fit parametric hazard models to survival-type data can can be viewed as an extension of the method deployed by Mantel (1973). When describing his motivating example, Mantel made a passing reference[1] to a time dimension but he did not address time in his statistical model. He was faced with 165 instances of $Y = 1$ and a much larger number (about 4,000) of instances of $Y = 0$, an associated regressor vector $x$ for each, a logistic model for $\Pr(Y = 1 \mid X = x)$, and limited capacity of computing. His solution was to form a reduced dataset consisting of all instances of $Y = 1$ and a random sample of the much more numerous instances of $Y = 0$, and to fit the same logistic model to this reduced dataset. He noted, as Anderson (1972) had done, that "such sampling will tend to leave the dependence of the log odds on the variables unaffected except for an additive constant."

Figure 1 addresses our first example here. The follow-up of these patients in the aggregate, represented by the shaded area, constitutes a *study base* of 20,894 person-years – or just over $10^{10}$ person-minutes, consisting of an infinite number of *person-moments*, where a 'moment' is a point in time. Any given person-moment in the study base is characterized by the point in time, $t$, and the person's value, $x$, for the variates based on prognostic indicators (at prognostic $T_0$) and intervention. The value of the hazard function at this person-moment is $h(x, t)$.

We take our dataset to consist of the $(y, x, t)$ information for the $c = 263$ person-moments at which stroke occurred – the *case series* – and for a *representative* sample – the *base series* – of size $b$, of the infinite number of person-moments that constitute the $B = 20,894$ person-years in the study base. For such a dataset for $(c + b)$ person-moments, $c$ with $Y = 1$ and $b$ with $Y = 0$,

$$\frac{\Pr(Y = 1|x, t)}{\Pr(Y = 0|x, t)} = \frac{h(x, t) \times B(x, t)}{b \times [B(x, t)/B]} = h(x, t) \times (B/b), \tag{3}$$

[1] "The actual number of individuals was substantially less than 4,000. An initial cohort of about 1,350 men was studied to evaluate the short-term prognostic value of various factors in coronary heart disease. Men remaining free of disease for two years could be re-entered into the analysis for the next two years using their new X values."

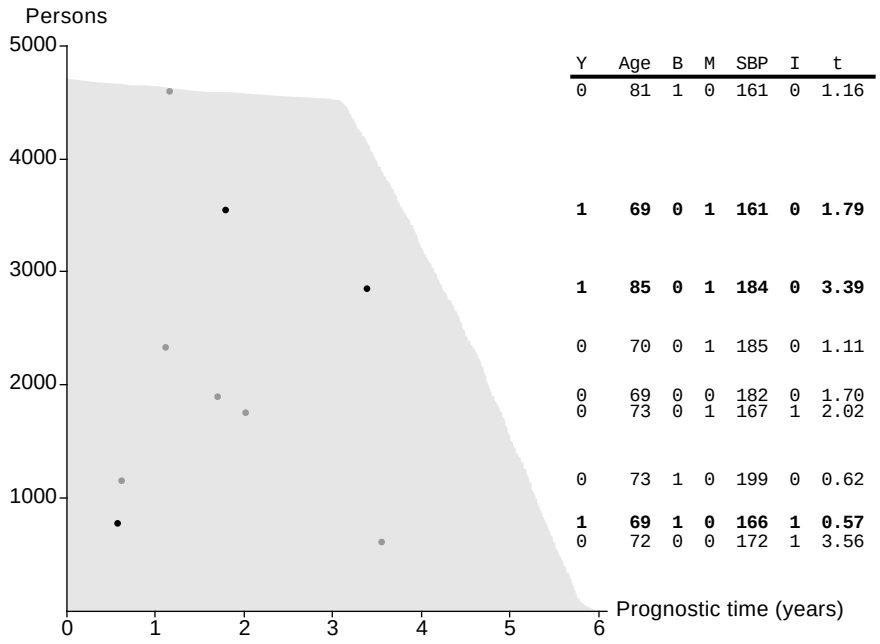


Figure 1: Formation of the dataset for the logistic regression approach (schematic). Shown at the left is the study base: 20,894 person-years in the SHEP (1991). It can be thought of as 4,701 rows of infinitely small rectangles (person-moments). Also shown in the figure are select person-moments from the base and case series: the person-moments of 3 case events (darker dots) and a random sample of 6 person-moments (lighter dots). Shown on the right are the corresponding data for these selected person-moments. Logistic regression, with the appropriate offset, fitted to data like these produces the parameter estimates for the empirical log-incidence-density function.

where $B(x,t)$ is the population-time element in the study base with $(X,T) = (x,t)$. Thus the fitted hazard function is

$$\widehat{h(x,t)} = (b/B)\exp(\hat{L}), \tag{4}$$

where $\hat{L}$ is the linear compound from the fitting of the *logistic* regression model. In practice, one can obtain $\log[\widehat{h(x,t)}]$ directly by fitting a logistic regression model to the $(c+b)$ vectors $[y, x, t, \log(B/b)]$ and specifying $\log(B/b)$

as an 'offset,' that is, a term whose regression coefficient is forced to be 1 (McCullagh and Nelder, 1989, p 206). The empirical log incidence-density function, $\widehat{h(x,t)}$, in equation (4) translates to the corresponding cumulative incidence or risk function on the basis of equation (2). An interval estimate for the risk, or risk difference, can be derived by making use also of the variance-covariance matrix of the fitted regression coefficients.

The formation of the dataset for such logistic regression is illustrated schematically in Figure 1, as in the actual dataset used to estimate the parameters in Table 2, there were $c = 263$ black dots, and $b = 26,300$ grey ones – an average of 5.6 moments per patient – too many to show in the graph. Drawing a representative sample of $b$ person-moments from the total population-time $B$ of follow-up of $n$ individuals can be done in a number of ways. In completely random selection one first generates a realization $(b_1, \ldots, b_n)$ from a multinomial distribution with $b$ trials and probabilities $(\pi_1, \ldots, \pi_n)$, where $\pi_j = t_j/B$, and $t_j$ is the duration of follow-up[2] for subject $j$. The $b_j$ moments are then selected independently from the $U(0, t_j)$ distribution. A systematic sample of size $b$ of $B$ can be formed upon concatenating the $n$ subject-specific intervals into a single interval from 0 to $B$. A third way, one that yields fully-reproducible estimates, is used in section 5.2. However, given the large sample size we worked with (see next paragraph), the way the representative sample is chosen made little difference.

How large should $b$ be on relation to $c$? We quote Mantel (1973), only using our notation: "By the reasoning that $cb/(c + b)$ $[= (1/c + 1/b)^{-1}]$ measures the relative information in a comparison of two averages based on sample sizes of $c$ and $b$ respectively, we might expect by analogy, which would of course not be exact in the present case, that this approach would result in only a moderate loss of information. (The practicing statistician is generally aware of this kind of thing. There is little to be gained by letting the size of one series, $b$, become arbitrarily large if the size of the other series, $c$, must remain fixed)." With computer capacity no longer a concern, we can easily use a $b/c$ ratio as high as 100. Such a ratio assures variances and covariances for the estimated regression coefficients that are only 1 percent larger than those obtained using all of the information in the study base, as they are proportional to $1/c + 1/100c$ rather than $1/c + 1/\infty$. Thus, virtually all of the information about the parameters in the hazard model is retained in a dataset involving a base-series of size $b = 100c$, given a case-series of size $c$.

[2] Ideally, separate letters should be used to distinguish a (generic) point on the time axis from the observed duration of follow-up for subject $j$. We, as textbooks do, use $t$ (without a subscript) for the former, and $t_j$ for the latter. Which meaning is intended will also be obvious from the context.

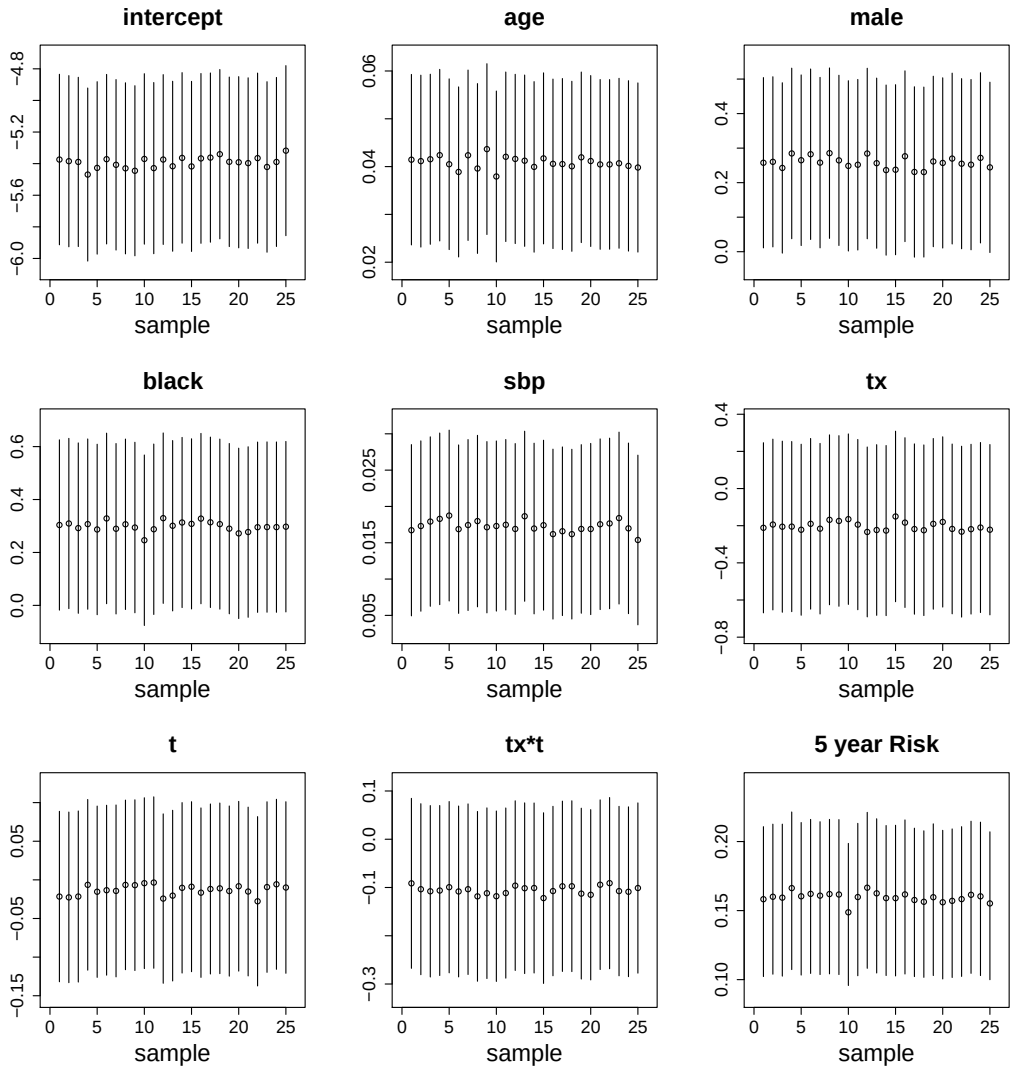


Figure 2: Stability of point and (95% confidence) interval estimates of 8 parameters of the hazard function, and of 5-year risk for a specific (untreated) high-risk profile (bottom right panel). Fits are based on 25 different random samples of 26,300 from the infinite number of person-moments in the study base. Estimates were derived by fitting a logistic regression model to a case-series of $c = 263$ and a base-series of $b = 26,300$. Data and model are those described in the illustration in section 5.1.

Figure 2 illustrates the virtually 100% efficiency, and negligible Monte Carlo error, achievable by such sampling of the study base. Whereas the point

estimate is subject to the additional variability induced by the sampling, the overall statistical uncertainty, reflected in the width of the confidence interval, is determined almost entirely by the amount of information in the study base, and only trivially by the fact that only a finite sample of the infinite number of person-moments is used. This phenomenon has some parallels with the two-part variance used following multiple imputation for missing data. No matter how many 'copies' are used, and their fitted coefficients averaged, the main component of variance is determined by the size of the study and remains irreducible. In our approach, the additional contribution to variance that arises from the sampling can be made arbitrarily small.

# 5 Illustrations

## 5.1 The SHEP

For the study base of $B = 20,894$ person-years of follow-up, in which $c = 263$ events of stroke were observed, we adopted for the hazard the model

$$h(x,t) = \exp(\Sigma \beta_k x_k),$$

where $x_k$ is based on

$X_0 \equiv 1$
$X_1 =$ Age (in yrs) - 60
$X_2 =$ Indicator of male gender
$X_3 =$ Indicator of Black race
$X_4 =$ Systolic BP (in mmHg) - 140
$X_5 =$ Indicator of verum treatment
$X_6 =$ Prognostic time (in yrs)
$X_7 = X_5 \times X_6$.

To estimate the eight parameters, we formed a person-moments dataset pertaining to the case series of size $c = 263$ (with $Y = 1$) and a randomly-selected base series of size $b = 26,300$ (with $Y = 0$). Each of the 26,563 rows was constituted by the realizations of $X_0, \ldots, X_7$ and $Y$ at the person-moment and the offset, $\log(B/b) = \log(20,894/26,300)$.

The logistic model involving these variates was fitted to the data in the union of the two series. The resulting fitted values for the parameters in this fully-parametric non-proportional hazards model, together with their SEs, are given in column (1a) of Table 2. The fully-parametric proportional hazards model – the one without the $X_7$ term – was also fitted to the data, with the corresponding statistics shown in the second column (1b) of Table 2, and the corresponding results under the Gompertz model are shown in column (2), those under the semi-parametric proportional hazards model in column (3).

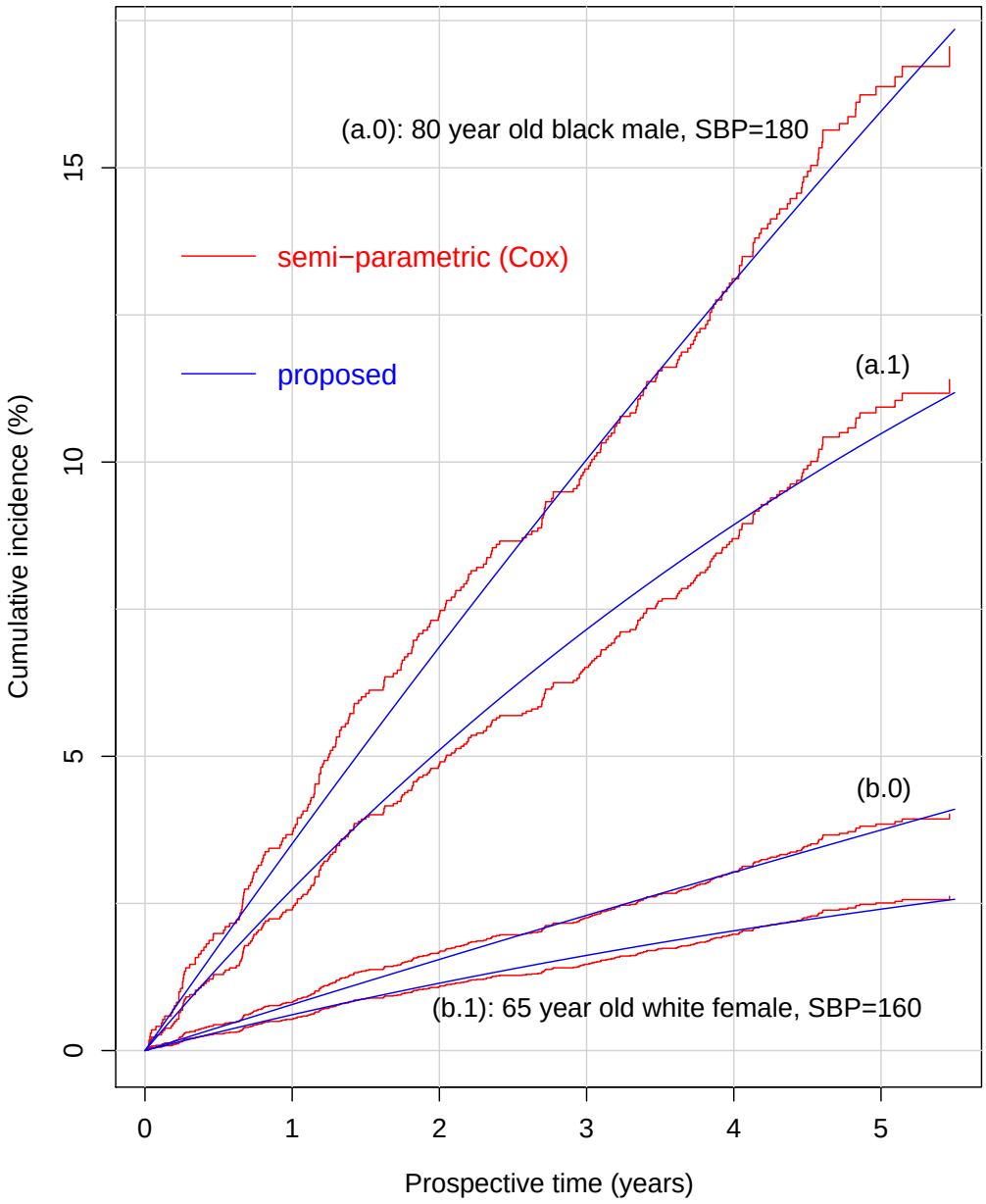


Figure 3: Estimated cumulative incidence (risk) of stroke for patients with higher-risk (a.0 if untreated, a.1 if treated) and lower-risk (b.0 and b.1) profiles, fitted by the proposed fully-parametric approach, and by the semi-parametric Cox regression. Data are from the SHEP (1991).

Table 2: Fitted values $\hat{\beta}$ (with SE in parentheses) for the regression coefficients of log-linear hazard model from the proposed logistic regression approach – with its representative sampling of the study population-time and the corresponding offset – and those of Gompertz and Cox regression models. Data are from the SHEP (1991). $I$ denotes "indicator of." The logistic function from the proposed approach provides estimates of the parameters in a smooth-in-time hazard function. This function in turn yields smooth-in-time estimates of cumulative hazard and, thus, profile- and treatment-specific risk.

| | Proposed logistic regression | | Gompertz regression | Cox regression |
|---|---|---|---|---|
| Term | (1a) | (1b) | (2) | (3) |
| $Age - 60$ | 0.041 *(0.009)* | 0.041 *(0.009)* | 0.041 *(0.009)* | 0.041 *(0.009)* |
| $I$(male) | 0.257 *(0.126)* | 0.258 *(0.126)* | 0.259 *(0.125)* | 0.259 *(0.125)* |
| $I$(black) | 0.302 *(0.164)* | 0.301 *(0.164)* | 0.304 *(0.163)* | 0.303 *(0.163)* |
| $SBP - 140$ | 0.017 *(0.006)* | 0.017 *(0.006)* | 0.017 *(0.006)* | 0.017 *(0.006)* |
| $I$(verum) | -0.200 *(0.234)* | -0.435 *(0.127)* | -0.435 *(0.126)* | -0.435 *(0.126)* |
| *Intercept* ($\beta_0$) | *-5.390 (0.274)* | *-5.295 (0.261)* | *-5.295 (0.260)* | |
| $t$ | -0.014 *(0.056)* | -0.057 *(0.044)* | -0.055 *(0.044)* | |
| $t \times I$(verum) | -0.107 *(0.090)* | | | |

(1a) Although the coefficient for the product term is not statistically significant, the results are shown to illustrate the ease with which non-proportional hazards models can be fitted.

(1b) a smooth-in-$t$ proportional hazards model, for comparison with its fully-parametric and semi-parametric counterparts in (2) and (3).

(2) Fitted using **streg** in Stata, where the coefficient for $t$ is called *gamma*.

Based on the full (8-term) logistic model, one might address the estimated 5-year risk of stroke for a 65 year old white female with a SBP of 160 mmHg. The estimated hazard is $h(t) = \exp(-4.86t)$ if $I(\text{verum}) = 0$, and $\exp(-5.06-$

Table 3: Risk estimate (%) for stroke in the next $1, \ldots, 5$ years, if the SBP will not be treated ($I = 0$) and if it will be treated ($I = 1$), as a function of the four prognostic indicators incorporated in the Total Score, **Cox model**, column (3) of Table 2. Data are from the SHEP (1991).

| | **Total Score** | I | Year 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|---|
| | 200 | 0 | 3.4 | 7.0 | 9.3 | 12.3 | 15.4 |
| | | 1 | 2.2 | 4.6 | 6.1 | 8.2 | 10.3 |
| (No. Years beyond 60) × 4 _____ | | | | | | | |
| | 150 | 0 | 2.1 | 4.3 | 5.7 | 7.7 | 9.7 |
| | | 1 | 1.4 | 2.8 | 3.8 | 5.0 | 6.4 |
| Black ... 30 _____ | | | | | | | |
| | 100 | 0 | 1.3 | 2.6 | 3.5 | 4.7 | 6.0 |
| | | 1 | 0.8 | 1.7 | 2.3 | 3.1 | 3.9 |
| Male ... 26 _____ | | | | | | | |
| | 50 | 0 | 0.8 | 1.6 | 2.2 | 2.9 | 3.7 |
| | | 1 | 0.5 | 1.0 | 1.4 | 1.9 | 2.4 |
| (Every 10 mm SBP above 140) × 17 _____ | | | | | | | |
| | 0 | 0 | 0.5 | 1.0 | 1.3 | 1.8 | 2.2 |
| | | 1 | 0.3 | 0.6 | 0.8 | 1.1 | 1.5 |
| **Total Score** _____ | | | | | | | |

$0.124t$) if $I(\text{verum}) = 1$. The 5-year integrals of these are 0.037 and 0.024, respectively, so that the $CI$ estimates are $1 - \exp(-0.037) = 0.036$ and $1 - \exp(-0.024) = 0.024$ respectively. While these imply the risk reduction estimate of 1.2 percent, the corresponding estimate for an 80 year old black male with a SBP of 180 mmHg is $(0.16 - 0.10 =)$ 6 percent, both appreciably different from the overall estimate of (0.076 - 0.049 =) 2.7 percent. The fitted cumulative incidence functions for these two profiles, along with those from the model of Cox, are shown in Figure 3. The fully-parametric fit is very good.

Even though they involve integrals, the risk estimates corresponding to a given profile can be obtained from the corresponding risk scores together with either a table (e.g. Table 3 or Table 4) or a nomogram (Figure 4) that converts these into risk estimates. Figure 4 was formed using the nomogram function in the "Design" package (Harrell, 2001, Harrell 2007) in `R` (details are available on request). As described more fully in our earlier article advocating greater use of profile-specific risk estimates (Julien and Hanley, 2008), points – proportional to fitted regression coefficients – for the four factors

Table 4: Risk estimate (%) for stroke in next $1, \ldots, 5$ years, if the SBP will not be treated ($I = 0$) and if it will be treated ($I = 1$), as a function of the four prognostic indicators incorporated in the Total Score, **proposed approach**, model in column (1a) of Table 2. Data are from the SHEP (1991).

| | **Total Score** | I | Year 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|---|
| | 200 | 0 | 3.3 | 6.4 | 9.4 | 12.3 | 15.0 |
| | | 1 | 2.6 | 4.8 | 6.7 | 8.4 | 9.8 |
| (No. Years beyond 60) × 4 _____ | | | | | | | |
| | 150 | 0 | 2.0 | 4.0 | 5.8 | 7.6 | 9.4 |
| | | 1 | 1.6 | 2.9 | 4.1 | 5.2 | 6.1 |
| Black ... 30 _____ | | | | | | | |
| | 100 | 0 | 1.2 | 2.4 | 3.6 | 4.7 | 5.8 |
| | | 1 | 1.0 | 1.8 | 2.5 | 3.2 | 3.7 |
| Male ... 26 _____ | | | | | | | |
| | 50 | 0 | 0.7 | 1.5 | 2.2 | 2.9 | 3.6 |
| | | 1 | 0.6 | 1.1 | 1.5 | 1.9 | 2.3 |
| (Every 10 mm SBP above 140) × 17 _____ | | | | | | | |
| | 0 | 0 | 0.5 | 0.9 | 1.3 | 1.8 | 2.2 |
| | | 1 | 0.4 | 0.7 | 0.9 | 1.2 | 1.4 |
| **Total Score** _____ | | | | | | | |

are summed and transferred to "Total Points" scale. The corresponding risk estimates are read from the bottom two scales. In that article, we created a modification of Harrel's nomogram to address risk *differences*. Although the content of the nomogram in that article was based on results from Cox regression, the format of that nomogram could also be used to display the results of the regression models described here.

To calculate the point estimate of the 5-year risk for the (untreated) high-risk profile shown in the bottom right panel of Figure 2, we substituted the value of the integral (obtained via the `integrate` function in `R`) into equation (2). To obtain a confidence interval to accompany this integral, we calculated the variance of a 10-point sum, using the variance-covariance matrix of the estimated coefficients, and the delta method (as in Efron, 1988).

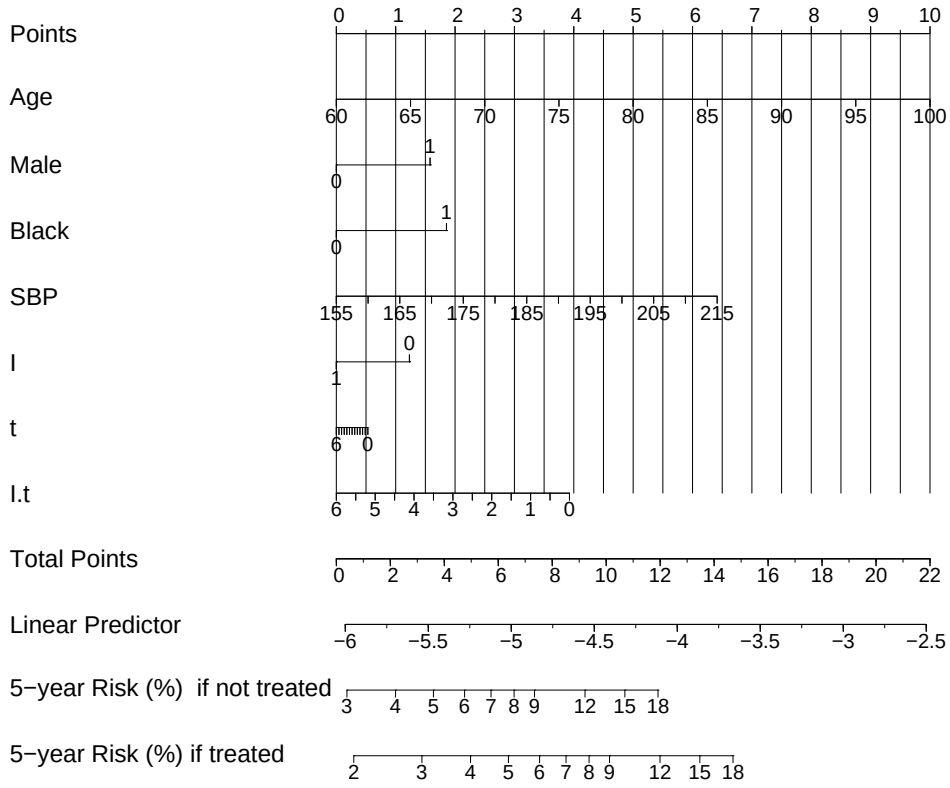


Figure 4: Nomogram to calculate estimated 5-year risk of stroke if untreated, or if treated. Points – proportional to fitted logistic regression coefficients – for the 4 indicators (Age to SBP) are summed and transferred to the "Total Points" scale. The corresponding estimates of risk are read from the bottom two scales. The scores for "I", "t" and "I.t", already included in these bottom two scales, are shown merely for completeness. Data are from the SHEP.

## 5.2 Head-and-neck cancer study

Efron (1988) found that a hazard function modelled as a cubic-linear spline, with the join point at $t = 11$ months, fitted the time-to-recurrence data better than a linear-in-time or cubic-in-time function. Figure 5 shows this hazard function, fitted by his method to the discretized data for arm A of the study,

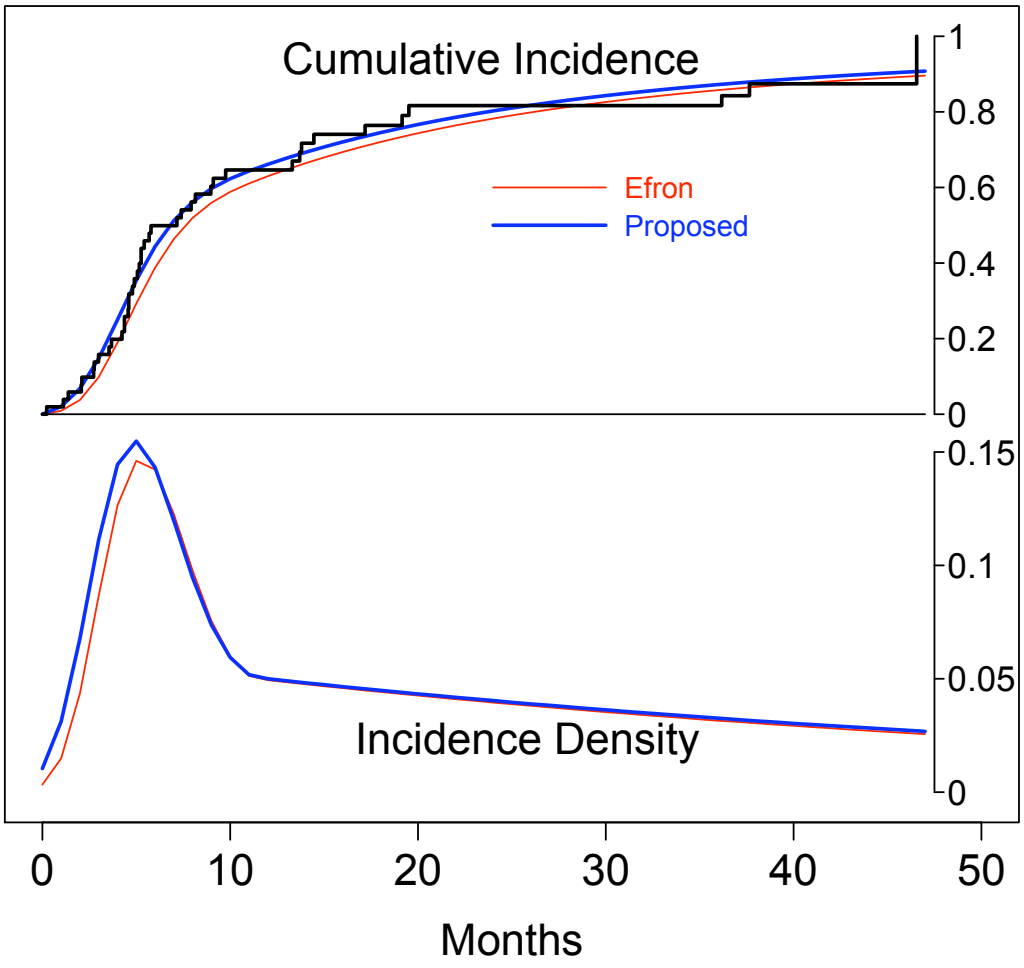


Figure 5: Cubic-linear incidence density (i.e., hazard) function, with join at 11 months, and cumulative incidence function derived from it, fitted to data from arm A of head-and-neck cancer study (Efron, 1988). Thicker, blue, lines: curves fitted by proposed method, based on person-moments; thinner, red, lines: curves fitted by Efron's 'parametric analysis' method, based on discretized data; step function, black: Kaplan Meier estimated cumulative incidence curve.

together with a hazard function of the same form fitted by the proposed method, based on person-moments (i.e., a non-discretized version of the same data). The slight difference between them probably reflects the discretization versus non-discretization. For this example, the representative sample of

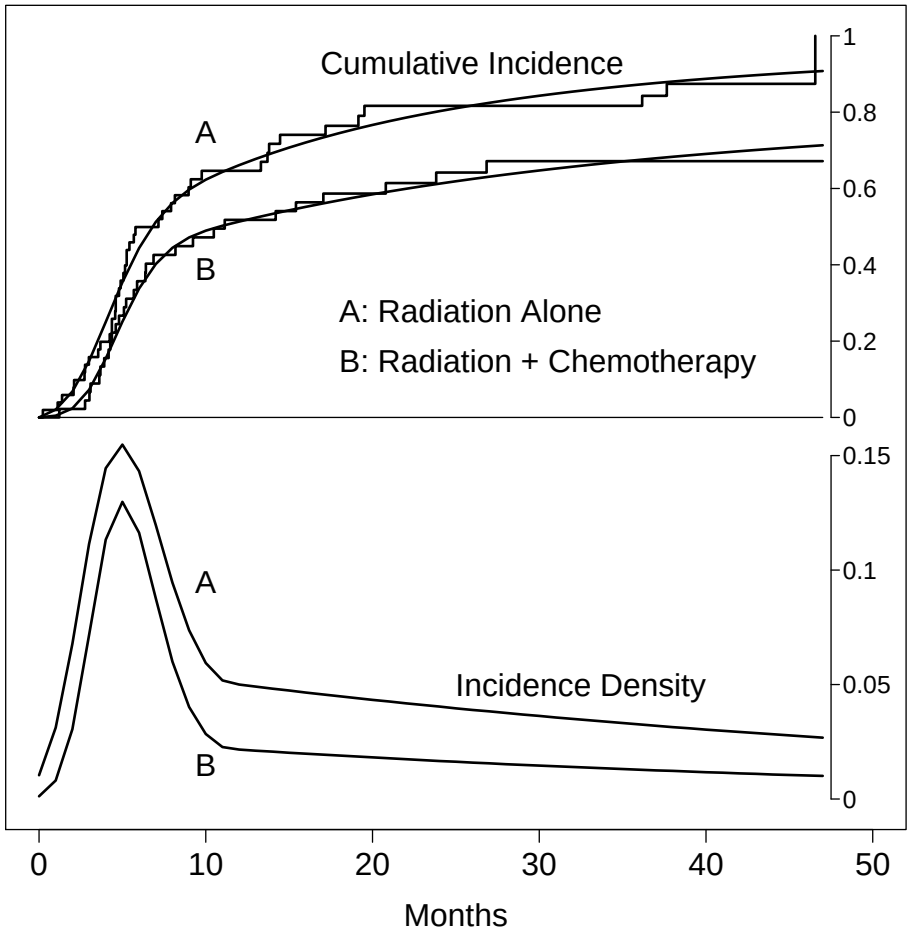


Figure 6: Cubic-linear incidence density (i.e., hazard) functions, with join at 11 months, and cumulative incidence functions derived from them, fitted – separately – to data from arms A and B of head-and-neck cancer study (Efron, 1988). Smooth curves: fitted by proposed method, based on person-moments; step functions: Kaplan Meier estimated cumulative incidence curves.

$b = 100 \times c = 100 \times 42 = 4,200$ person-moments from the 599.58 patient-months in the base was determined by allocating the number of moments for each of the $n = 51$ patients in proportion to the lengths of their follow-up times, 4,200/599.58 = 7.00 per month. Thus, for example, for a patient with

follow-up time of 1.38 months, allocated were $1.38 \times 7.00 = 10$ moments, at $t = 1.38 \times [1/11, 2/11, \ldots, 10/11]$ months.

Figure 6 shows the separately fitted hazard functions for the two treatment arms, along with the cumulative incidence functions derived from them, and their Kaplan-Meier counterparts. Just as in Efron's analysis, the two separately fitted hazard functions for A and B (4 parameters each) appear to be non-proportional: the ratio of the fitted *ID*s is 1.2 at 6 months but 2.3 at 12 months. In the 5-parameter proportional cubic-in-time hazard model the constant-in-time $ID$ ratio is 1.69 (95% CI 1.05 to 2.71). However, an 8-parameter hazard function that uses a cubic function, a treatment indicator, and 3 products – which duplicated the 2 separate cubic fits – did not fit appreciably better than the 5-parameter proportional hazards model. Some of this "complicated early structure" (Efron, 1988, p. 416) that required the cubic form undoubtedly stems from our ignorance of important disease-extent descriptors for each patient.

# 6 Discussion

Our focus has been on 'individualized' – specific for prospective time, profile and treatment – risk functions derived from smooth-in-time hazard functions. Despite their wide availability, the profile-specific estimates of risk available from Cox's semi-parametric approach are seldom used, and we suspect that end-users are averse to the fitted risk function's 'steps-in-time' – and thus 'raw' and 'unsophisticated-looking' – form.

The parametric, smooth-in-time, log-linear form we adopted for the hazard function—as the basis for the risk function derived from it— is not new. It is a natural extension of the form proposed by Gompertz (1825). There are closed-form expressions for the ML estimates of the parameters of the Gompertz model fitted to a homogeneous sample (Carriere, 1994); and a proportional-hazards form of this model can be fitted using `streg` in Stata. However, with more complicated terms in $t$, such as a treatment and $t$ product term or any other time-varying covariate, computing difficulties appear to have prevented this very natural form from being used in a broader regression framework. To avoid maximizing a log-likelihood that involves logs of integrals of this hazard function, one can use a time-slicing approach to fit the smooth model; but this too involves some compromises.

The method described here is a very different way of fitting this hazard function. It is based on bringing to the context of survival analysis the general structure of an etiologic study in epidemiology (Mantel 1973; Miettinen 1976; Miettinen, 2004; Miettinen, 2008): a *case series* (of the event at issue) coupled with a *base series* (sample of the study population-time), together

with logistic regression analysis of the data on these. While this approach generally allows estimation of only functions for ratios of the hazard – incidence density – the proposed approach provides for estimation of the hazard function *per se*, and thereby estimation of cumulative incidence and risk. Essential to this end is *representative sampling* of the study population-time for the base series, so very different from the risk set sampling that is in the essence of Cox regression.

The log-linear modelling for incidence density, which underlies this logistic-regression approach, opens up the possibility of fitting – and using standard methods to assess the fit of – a wide range of functional forms for the time-dimension of the hazard function, and of effortlessly handling censored data. It handles not just the one-sample situations that were the main focus of Efron (1988), but also the regression analog he proposed in that article. In addition, the proposed model allows flexibility in explicitly modeling non-proportionality over $t$. Statistical research in this area has heretofore focused on using splines to model a hazard *ratio* that changes over time. The proposed approach allows splines to be used within a standard logistic regression framework to smooth the hazard function itself, as an alternative to smoothing the cumulative hazard function (Royston and Parmar, 2002).

By replacing $t$ by $\log(t)$ in the linear predictor, and using case and base series, one can also use standard logistic regression to fit Weibull models. For example, this approach reproduces the parameter estimates for the (exponential and) Weibull models reported in Table 1 of Aitkin and Clayton (1980).

Since the choice of model form will depend on the context, there can not be general principles for this choice. The purely Gompertzian hazard model, with log rates having a straight-line relation to age and/or follow-up time, is suited to studies, such as the SHEP (1991), involving persons who have not yet developed the illness of interest, whereas more complex time functions tend to be needed for the prognosis in the context of a newly-diagnosed condition. Efron (1988) used a cubic-linear spline to model the "more complicated early structure" of the hazard functions in each of the two arms of a head-and-neck cancer study. Some of the complexity of those hazard functions might be removable by including patient-level covariates. However, features such as early surgical mortality and the (partially latent) mixture of curable and incurable patients can still be expected to complicate the early structure and to create non-proportionality of the hazard function. Quite complex hazard models can, however, be fitted by standard logistic regression in the approach addressed here, and this means that the checking of the model fit can be carried out using standard and familiar regression techniques. As with Efron's approach, no additional complexities are created

by the presence of censored observations. Thus, other than pointing to the assessments of fit he used, no additional specifics of model fit and selection are discussed here.

Depending on the currently available software used, the fitting of semi- and fully-parametric hazard models with time-varying covariates, and product terms involving $t$, requires different levels of sophistication on the part of the end-user. In the time-slicing approach used by Stata and the `survival` package in `R`, the user must split each record into several, or use the built-in facilities for doing so. The `phreg` procedure in SAS allows the user to accommodate time-varying covariates 'on-the-fly' but is subject to being misused by less sophisticated users. The approach described here, with the opportunity to incorporate the covariate history as of each person-moment in the two series makes the modelling more transparent.

The information about the hazard function, descriptive of the study base, is constrained by the size ($c$) of the case series; and this information is captured practically in its entirety when the base series, though only a finite sample of $b$ of the infinite number of person-moments constituting the study base, is suitably large in proportion to the case series. It is quite feasible to use, as we have done, a $b/c$ ratio that is sufficiently large so that the imprecision of the estimates results, in all essence, from the paucity of cases, that is, so that the information, proportional to $(1/c + 1/b)^{-1}$, is effectively proportional to $c$.

Some readers may wonder why, since our first illustrative example deals with the experience of patients 65 and older, we did not deal with competing risks of death. One reason was to not detract from the central topic: parametric modeling of hazard functions, leading to smooth-in-time profile-specific cumulative incidence functions. Another was to accommodate the outlook that even in the face of competing risks, prognosis should be conditional on otherwise surviving. Others, such as Albertsen et al. (1998), take the opposite view: to produce a display of cumulative proportions of prostate cancer and all-other-cause mortality graphs, they used time-slicing and Poisson regression to fit separate parametric log-linear hazard functions for these two rates of mortality. The approach proposed here, based on Bernoulli rather than Poisson variates, is a simpler and more natural way to fit such functions.

# 7 Software

The website `http://www.biostat.mcgill.ca/hanley/software` contains an `R` function that accepts as input a dataset containing the prognostic indicators, the duration until follow-up was terminated, and whether termination

was by the event of interest, or by censoring. The function returns a person-moment file suitable for the proposed logistic regression approach.

# 8 References


Aitkin, M. and Clayton, D. (1980) The fitting of exponential, Weibull and extreme value distributions to complex censored survival data using GLIM. *Applied Statistics* **29**, 156-163.

Albertsen, P.C., Hanley J.A., Gleason D.F., Barry M.J. (1998). Competing risk analysis of men aged 55 to 74 years at diagnosis managed conservatively for clinically localized prostate cancer. *Journal of American Medical Association* **280**, 975-980.

Anderson, J. A. (1972). Separate sample logistic discrimination. *Biometrika* **59**, 19-35.

Bevilacqua, J.L., Kattan, M.W., Fey, J.V., Cody, H.S.,, Borgen, P.I., Van Zee, K.J. (2007). Doctor, what are my chances of having a positive sentinel node? A validated nomogram for risk estimation. *Journal of Clinical Oncology* **25**, 3670-3679. Epub 2007 Jul 30.

Breslow, N.E. (1972). Contribution to the discussion of paper by D.R. Cox. *Journal of the Royal Statistical Society, Series B* **34**, 216-217.

Califf, R.M., Woodlief, L.H., Harrell, F.E. Jr, Lee, K.L., White, H.D., Guerci, A., Barbash, G.I., Simes, R.J., Weaver, W.D., Simoons, M.L., Topol, E.J. (1997). Selection of thrombolytic therapy for individual patients: development of a clinical model. *American Heart Journal* **133**, 630-639.

Carriere, J.F. (1994). An investigation of the Gompertz law of mortality. *Actuarial Research Clearing House* **2**, 161-177. `http://www.soa.org`

Cassidy, A., Myles, J.P., van Tongeren, M., Page, R.D., Liloglou, T., Duffy, S.W., Field, J.K. (2008). The LLP risk model: an individual risk prediction model for lung cancer. *British Journal of Cancer* **98**, 270-276.

Clayton, D.G. (1983). Fitting a General Family of Failure-Time Distributions using GLIM. *Applied Statistics* **32**, 102-109.

Cox, D.R. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society, Series B* **34**, 187-220.

Efron B. (1977). The Efficiency of Cox's Likelihood Function for Censored Data. *Journal of the American Statistical Association* **72**, 557-565.

Efron B. (1988). Logistic Regression, Survival Analysis, and the Kaplan-Meier Curve. *Journal of the American Statistical Association* **83**, 414-425

Efron, B. (2002). The two-way proportional hazards model. *Journal of the*

*Royal Statistical Society, Series B* **64**, 899-909.

Gompertz, B., (1825). On the Nature of the Function Expressive of the Law of Human Mortality, and on a New Mode of Determining the Value of Life Contingencies. *Philosophical Transactions of the Royal Society of London* **115**, 513-585.

Hanley, J.A. (2008). The Breslow Estimator of the Nonparametric Baseline Survivor Function in Cox's Regression Model: Some Heuristics. *Epidemiology* **19**, 101-102.

Harrell, F.E. (2007). The 'Design' package for R and S-Plus. `http://cran.r-project.org/src/contrib/Descriptions/Design.html`. (accessed 01 Dec 2007).

Harrell F.E. (2001). *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis*. New York: Springer.

Hjort N.L. (1992) On inference in parametric survival data models. *International Statistical Review* **60**, 355-387.

Julien, M, and Hanley, J.A. (2008) Profile-specific survival estimates: Making reports of clinical trials more patient-relevant. *Clinical Trials* **5**, 107-115.

Kannel, W.B., D'Agostino, R.B., Silbershatz, H., Belanger, A.J., Wilson, P.W., Levy, D. (1999). Profile for estimating risk of heart failure. *Archives of Internal Medicine* **159**, 1197-1204.

Machin, D., Campbell, M.J. (2005). Prognostic Factor Studies. Chapter 11 in *Design of Studies for Medical Research* Published Online: 29 Nov 2005 .John Wiley & Sons, Ltd.

Mantel, N. (1973). Synthetic retrospective studies and related topics. *Biometrics* **29**, 479-486.

McCullagh, P., Nelder, J.A. (1989). *Generalized Linear Models, 2nd Edition* London: Chapman and Hall.

Memorial Sloan-Kettering Cancer Center. Lung Cancer Risk Assessment `http://www.mskcc.org/mskcc/html/12463.cfm` Last accessed, March 29, 2008.

Miettinen OS (1976). Estimability and estimation in case-referent studies. *American Journal of Epidemiology* **103**, 226-235.

Miettinen, O.S. (2004). Epidemiology: quo vadis? *European Journal of Epidemiology* **19**, 713-718.

Miettinen, O.S. (2008). Important concepts in epidemiology. in: *Teaching Epidemiology: A Guide for Teachers in Epidemiology, Public Health and Clinical Medicine*, 3rd edn, Olsen, Saracci, Trichopoulos (eds.). In press.

NHLBI (National Heart Lung and Blood Institute, USA) (2008). Estimating coronary heart disease (CHD) risk using Framingham Heart Study prediction score sheets.

http://www.nhlbi.nih.gov/about/framingham/riskabs.htm
Last accessed, March 29, 2008.

NCI (National Cancer Institute, USA) (2008). The breast cancer risk assessment tool. http://www.cancer.gov/bcrisktool/ Last accessed, March 29, 2008.

NHLBI (National Heart Lung and Blood Institute, USA) (2007). The NHLBI Limited Access Dataset Program.
http://www.nhlbi.nih.gov/resources/deca/directry.htm.
Last accessed, March 06, 2007.

Partin tables for the probability of pathological stage of prostate cancer. (2007).
http://urology.jhu.edu/prostate/partintables.php
Last accessed, March 06, 2007.

Reid, N. (1994). A Conversation with Sir David Cox. *Statistical Science* **9**, 439-455.

Royston, P., Parmar, M.K. (2002). Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Statistics in Medicine* **21**, 2175-2197.

SHEP Cooperative Research Group (1991). Prevention of stroke by antihypertensive drug treatment in older persons with isolated systolic hypertension. Final results of the Systolic Hypertension in the Elderly Program (SHEP). *Journal of American Medical Association* **265**, 3255-3264.

Spitz, M.R., Hong, W.K., Amos, C.I., Wu, X., Schabath, M.B., Dong, Q., Shete, S., Etzel, C.J. (2007). A risk model for prediction of lung cancer. *Journal of the National Cancer Institute* **99**, 715-726.

Steyerberg, E.W., Vergouwe, Y., Keizer, H.J., Habbema, J.D.F. (2001). Residual mass histology in testicular cancer: development and validation of a clinical prediction rule. *Statistics in Medicine* **20**, 3847-3859.

Therneau T.M. (1996a) A Package for Survival Analysis in S.
http://www.asu.edu/splus/survival.pdf.

Therneau T.M. (1996b) Fitting a Gompertz distribution.
E-mail to S-news@utstat.toronto.edu Wed, 10 Jul 1996 11:16:43 -0500.

Tisman, G. (2007). Using nomograms to predict prostate cancer treatment outcomes. http://www.prostate-cancer.org/resource/pdf/Is8-4.pdf
Last accessed, December 01, 2007.