Collection of Biostatistics Research Archive
COBRA Preprint Series

Year 2010 *Paper* 67

Recovery of the Baseline Incidence Density in Censored Time-to-Event Analysis

Mikel Aickin*

*University of Arizona, maickin@earthlink.net

http://biostats.bepress.com/cobra/art67

Recovery of the Baseline Incidence Density in Censored Time-to-Event Analysis

Mikel Aickin

Abstract

Abstract Time-to-event analyses are often concerned with the effects of explanatory factors on the underlying incidence density, but since there is no intrinsic interest in the form of the incidence density itself, a proportional hazards model is used. When part of the purpose of the analysis is to use actual cumulative incidence for simulation, or for providing informative visual displays of the results, an estimate of the baseline incidence density is required. The usual method for estimating the baseline hazards in Cox's proportional hazards analysis yields values that are of little use, and furthermore no standard deviations of the estimates (SDEs) are available. In this article we present an alternative approach to recovering an estimate of the baseline incidence density that yields smooth estimates as well as smooth estimates of SDEs. We illustrate the method on a large dataset of inter-visit times for individuals in a diabetes registry, and indicate how it can be used to incorporate different baseline incidence densities in the analysis of different subgroups. Keywords: proportional hazards, exponential regression, survival analysis, diabetes

Recovery of the Baseline Incidence Density in Censored Time-to-Event Analysis

Mikel Aickin

ErgoLogic Consulting & Software
and
Department of Family & Community Medicine
University of Arizona

Abstract

Time-to-event analyses are often concerned with the effects of explanatory factors on the underlying incidence density, but since there is no intrinsic interest in the form of the incidence density itself, a proportional hazards model is used. When part of the purpose of the analysis is to use actual cumulative incidence for simulation, or for providing informative visual displays of the results, an estimate of the baseline incidence density is required. The usual method for estimating the baseline hazards in Cox's proportional hazards analysis yields values that are of little use, and furthermore no standard deviations of the estimates (SDEs) are available. In this article we present an alternative approach to recovering an estimate of the baseline incidence density that yields smooth estimates as well as smooth estimates of SDEs. We illustrate the method on a large dataset of inter-visit times for individuals in a diabetes registry, and indicate how it can be used to incorporate different baseline incidence densities in the analysis of different subgroups.

Keywords: proportional hazards, exponential regression, survival analysis, diabetes

Introduction

The proportional hazards model (Cox 1972) is widely used to assess the effects of factors of interest on the time to a target event, without having to worry about the form of the underlying hazard function. There are, however, some instances in which one requires the fitted hazard function itself. In the application that motivated this research, we wanted to obtain hazard functions for individuals in a large diabetes registry (Brown et al. 1999) in order to construct a model for the progression and sequellae of this disease. To operate such a model, it is necessary to be able to simulate state transitions using actual hazard functions. Specifically, the model takes the general form

$$R(t) = R_0(t)\exp(x'\beta)$$

where $R(t)$ is the cumulative incidence function for an individual with covariate vector x, β is the parameter vector of covariate effects, and R_0 is the baseline cumulative incidence function. (We prefer "incidence" to "hazard", since the target event in this kind of analysis need not be undesirable.) Given that the above person has not experienced the event by time t, the probability of him/her doing so in the next interval to t+dt is

$$P[t<T\leq t+dt|T>t] = 1 - \exp(-(R(t+dt)-R(t)))$$

which can often be approximated by $1 - \exp(-r(t)dt)$ where $r(t)$ is the incidence density. The incidence density is related to the baseline incidence density r_0 by $r(t) = r_0(t)\exp(x'\beta)$.

Kalbfleisch and Prentice (1980) gave a method of estimating $r_0(t)$ for each individual in the sample (at the time t when they experienced the event), evidently as a selection from among several previously proposed procedures. This method has been widely cited, and it is used in the Stata® software that we have employed here (Stata Press 1997). An example of these estimates appears in Figure 1. The data here are the times in days between the first and second post-diagnosis medical encounters among 1000 members of the KPNW Diabetes Registry (Brown et al. 1999). As part of a simulation of the behavior of diabetes patients we wanted to be able to generate the times of medical visits in a realistic way, and the first task was to understand how the times between visits changed in distribution with the passage of time. Because we also wanted to include their effects, the proportional hazards analysis included an indicator of male gender, and the age at diagnosis for these Type 2 diabetics. A graph of the incidence estimates from the same data, by the method to be proposed here, appears in Figure 2, and the underlying cumulative incidence function estimates appear in Figure 3. The hazard estimates in Figure 1 show several features which call for improvement. First, they exhibit a substantial amount of variability, as well as an indication of multiple branches suggesting different hazard functions, which is characteristic of this method. In contrast, the method we propose generally presents the incidence density as a smooth curve, Figure 2 being typical. Moreover, standard deviations of the hazard estimates in Figure 1 are not offered, while in Figure 2 we show smoothed versions of confidence limits. Finally, the hazards in Figure 1 show a marked tendency to increase as the data upon which they are based grows sparser, while our estimates decrease, a pattern that would be indicated by the underlying cumulative incidence function of Figure 3.

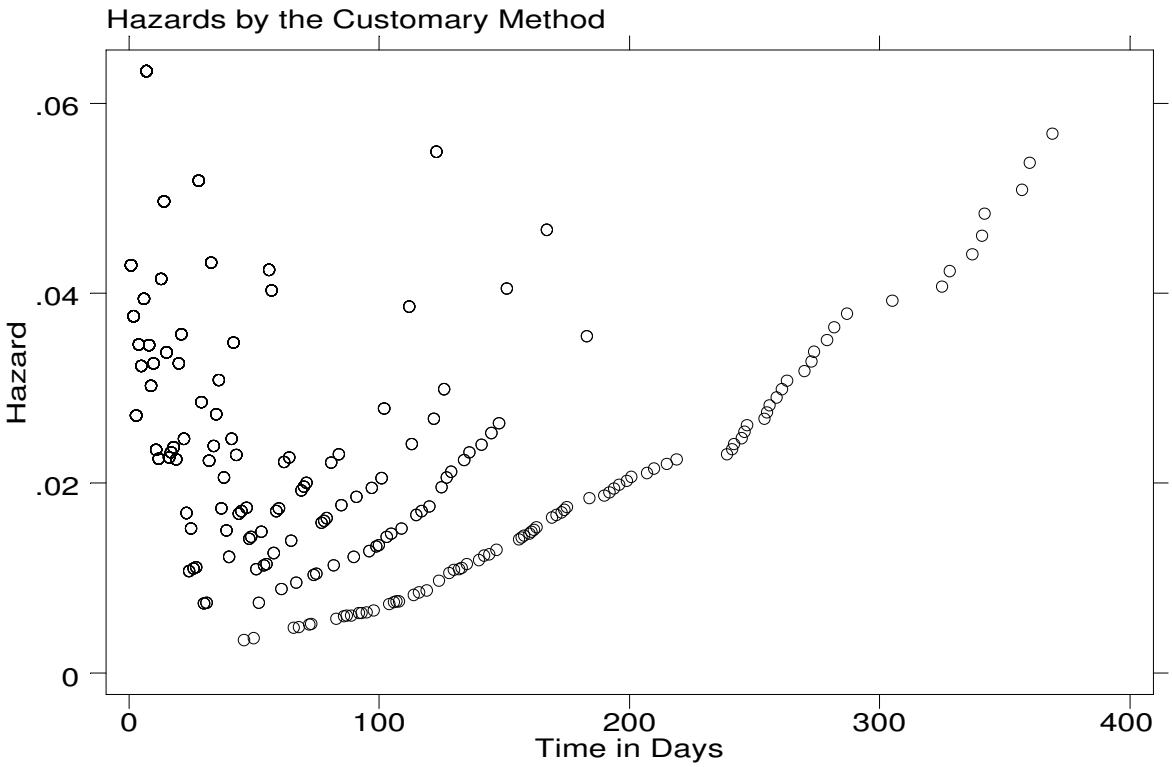


Figure 1. Hazards estimated at each event time, for 1000 times between medical encounters by newly-diagnosed type 2 diabetes patients.

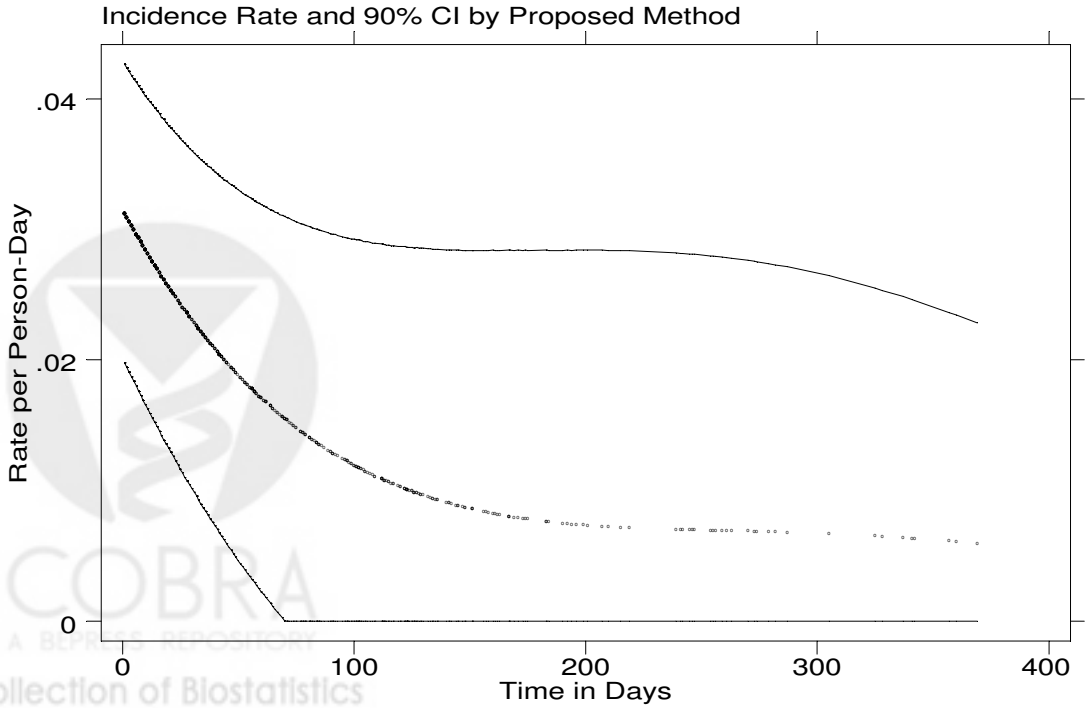


Figure 2. Incidence density estimates and confidence intervals; same data as Figure 1.

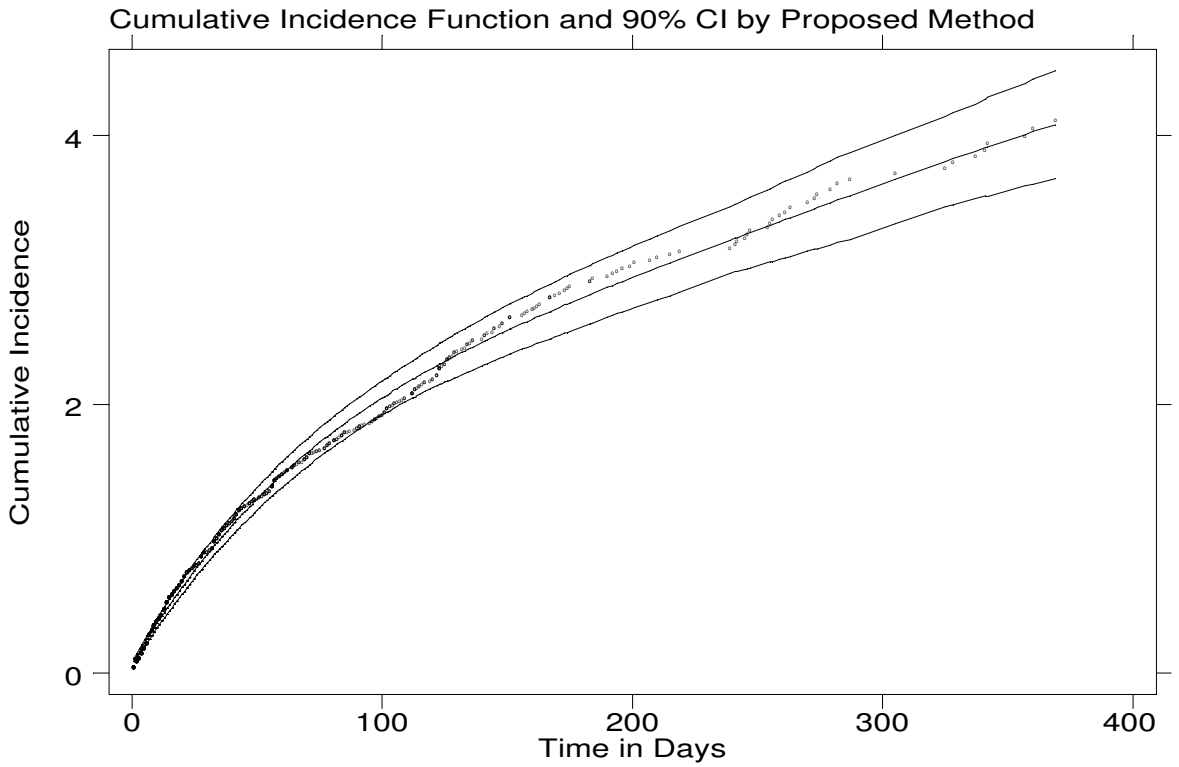


Figure 3. Cumulative incidence estimates (dots), smoothed estimate, and smoothed confidence bound; same data as Figure 1.

The intention of the remainder of this article is to explain the proposed method of estimation, and to illustrate its use in an investigation of the time-between-encounters in the KPNW Diabetes Registry.

Computational Method

Our aim is to estimate $R_0(t)$ at any particular chosen value of t. We do this by constructing the probability model for the data available at time t, employing the assumptions of Cox's proportional hazards model. In fact, we will use part of the output of a proportional hazards analysis to help with the maximum likelihood estimation of $R_0(t)$

The first step of our method consists of a standard Cox proportional hazards analysis, which produces for each individual i the fitted value of $\exp(x'\beta)$. For convenience, we abbreviate this value by ε_i. Let a time t be selected arbitrarily, and consider the estimation of $R_0 = R_0(t)$. Let C(t) be the cumulative incidence function of the censoring mechanism. At time t, our individual is in one of three conditions:

(1) censored before t, probability = 1 - exp(-C(t))

(2) not censored but experienced the event before t, probability =

$\exp(-C(t))(1-\exp(-R_0\varepsilon_i))$

(3) neither of the above, probability = $\exp(-C(t) - R_0\varepsilon_i)$

We interpret the cumulative incidence function R(t) that is commonly estimated in these analyses to be the conditional cumulative incidence of experiencing the event before t given that one was not censored before t. We also take the values ε_i produced by the Cox proportional hazards analysis as given. Consequently, given that individual i was not censored before t, his/her probabilities of experiencing the second and third events above are $1-\exp(-R_0\varepsilon_i)$ and $\exp(-R_0\varepsilon_i)$. It follows that the log likelihood is

$$L = \sum_{(1)} \ln\left(1 - e^{-R_0\varepsilon_i}\right) - \sum_{(2)} R_0\varepsilon_i$$

where the sum(1) is over all event times up to t, and the sum (2) is over all times (event or censoring) after t. The likelihood equation is, therefore,

$$\sum_{(1)} \frac{r_i e^{-R_0\varepsilon_i}}{1 - e^{-R_0\varepsilon_i}} = \sum_{(2)} \varepsilon_i$$

The Newton-Raphson iterative computation procedure becomes

$$R_0 \leftarrow R_0 + \left[\sum_{(1)} \frac{r_i e^{-R_0\varepsilon_i}}{1 - e^{-R_0\varepsilon_i}} - \sum_{(2)} \varepsilon_i\right] \Bigg/ \sum_{(1)} \frac{\varepsilon_i^2 e^{-R_0\varepsilon_i}}{\left(1 - e^{-R_0\varepsilon_i}\right)^2}$$

A starting value for R_0 that we have used successfully is the number of events experienced before t divided by the sum of all ε_i. The term in the denominator on the right is the negative second derivative of the log likelihood, and its inverse is therefore an estimator of the variance of R_0:

$$\mathrm{var}\left[R_0\right] = 1 \Bigg/ \sum_{(1)} \frac{\varepsilon_i^2 e^{-R_0\varepsilon_i}}{\left(1 - e^{-R_0\varepsilon_i}\right)^2}$$

It is worth noting in passing that the likelihood equation can easily be manipulated into the form

$$\sum_{(1)} \frac{\varepsilon_i}{1 - e^{-R_0\varepsilon_i}} = \sum_{(1,2)} \varepsilon_i$$

from which it follows automatically that the estimates of R_0 increase with t.

The values of $R_0(t)$ can be estimated at pre-selected times, only at times of events, or at all times appearing in the dataset. The only reason to select one rather than the other is the number of values desired and the computer time involved. For convenience, we assume $R_0(t)$ estimated at every time in the dataset. This provides values that can be used in subsequent exponential regression analyses. This latter model postulates $R(t) = t \exp(x'\beta)$, but once we have estimates of $R_0(t)$ in hand, we can declare $R_0(t)$ to be the "time" variable in the exponential regression, obtaining a model that is formally equivalent to the original proportional hazards specification, but using the exponential regression method. Note that we would center all explanatory variables (x) at their means for the proportional hazards part of the analysis, so that zero (the mean) would be a sensible value of the covariates at which one might want to estimate the baseline incidence. In the subsequent exponential regression, we might well choose not to do this, so that the implementation in a simulation could use the natural values of the variables, without having to know the means.

For the purposes of using cumulative incidence functions in modeling, it is extremely useful to reduce them to a small number of parameters. Conventional methods of representing functions by polynomial approximations tend to accentuate variations that are not statistically or practically meaningful. The method we employ here is based on a model of the form

$$R_0(t) = \beta_0 + \beta_1 t + \sum_i \alpha_i \ln\left(\left|t - \pi_i\right|\right)$$

The α and β parameters are fitted by ordinary linear regression. The values π_i are generally taken for convenience, since there is little information in typical data to provide precise estimates. Moreover, so long as the π_i values are selected reasonably outside the range of values of t, the fitted functions do not vary much. We have found by practical experience that it is reasonable to place the π_i values as follows:

---------π----------π----------t_{min}----------t_{max}----------π----------π----------

that is, the π values appear Δ below the minimum t and 2Δ below the minimum t, and Δ above and 2Δ above the maximum t, where $\Delta = t_{max}-t_{min}$. Thus, the following four values appear to be adequate in practice:

$$\left|t - \pi_i\right| = \begin{cases} t - t_{min} + 2k\Delta \\ t - t_{min} + k\Delta \\ 2k\Delta + t_{max} - t \\ k\Delta + t_{max} - t \end{cases}$$

Here, k is the multiplier of the range $t_{max}-t_{min}$, with larger values spreading the π-values further above and below the actual time values. For $k \leq 1$ this procedure fits curvature within the data range quite well, while larger values of k provide increasingly stiffer fitted functions. The selection of k is not made with any optimality criterion in mind, but rather with the practical aim of obtaining a stiff estimate of cumulative incidence (that is, one

that is insensitive to isolated bumps). Experience suggests that k in the range from 1 to 2 work well. With this approach, the values of the incidence density $r_0(t)$ can be computed at any time by

$$r_0(t) = \beta_1 + \sum_i \frac{\alpha_i}{t - \pi_i}$$

An estimate of the variance of $r_0(t)$ can be provided at any t for which the corresponding R_0 has been estimated. The key observation is that if we let R_0 stand for $R_0(t_{j-1})$, and $\delta = R_0(t_j) - R_o(t_{j-1})$, then the log likelihood can be written

$$L = \sum_{(1)} \ln\left(1 - e^{-R_0 \varepsilon_i}\right) - \sum_{(2,3)} R_0 \varepsilon_i + \sum_{(2)} \ln\left(1 - e^{-\delta \varepsilon_i}\right) - \sum_{(3)} \delta \varepsilon_i$$

where (1) denotes event times up to t_{j-1}, (2) denotes event times (and censoring times in its first appearance above) between t_{j-1} and t_j, and finally (3) denotes events and censorings after t_j. From this expression it is obvious that the estimates of R_0 and δ are asymptotically independent. This in turn implies

$$\text{var}[\delta] = \text{var}[R_0(t_j)] - \text{var}[R_0(t_{j-1})]$$

and so finally we have approximately

$$\text{var}[r_0(t_j)] = \text{var}[\delta]/(t_j - t_{j-1})^2$$

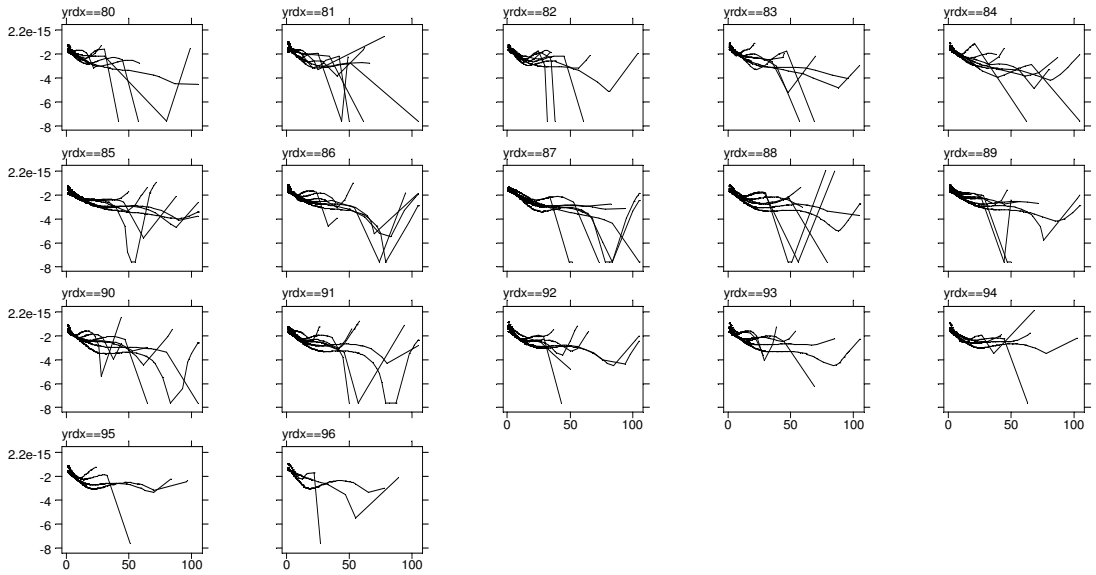
This value tends to infinity as the time interval in the denominator shrinks to zero. In fact, the variance expression on the right pertains to the time average of $r_0(t)$ over the interval from t_{j-1} to t_j.

Results

The main purpose of the analysis is to produce cumulative incidence functions that are specific to gender and to age at diagnosis. A complicating feature is that the patterns of medical encounters may have changed over time, as measured by year of diagnosis, and may also shift as one considers later encounters. In order to provide a realistic simulation based on proportional hazards modeling, it is advisable to perform some check to see that these latter two factors do not affect the proportionality assumption unduly. A secondary purpose of the analysis is to assess the effects of gender and age at diagnosis on inter-encounter times, removing potential secular and encounter number effects.

Figure 4 shows the incidence density plots by year of diagnosis (1980-1996) and within each plot, by encounter number (1,10,20,30,40,50,60; the later encounter numbers

are not included in the most recent years). The vertical axis is ln-scaled, in order to visually spread out the early portion of the graphs. The overall impression is one of homogeneity, with no obvious secular shift. This is re-inforced by Figure 5, which shows all of the plots of Figure 4 superimposed. The sparse, straight line segments beyond about 24 weeks largely represent the tail ends of the incidence density curves, where there are few events, and where the rate is particularly poorly estimated. The relatively dense clustering of lines suggests both proportionality as well as reasonable homogeneity over the diagnosis years and encounter numbers.



.
.

Figure 4. Ln incidence densities plotted by encounter number within diagnosis year (yrdx).

The ln incidence density was regressed on a fifth-degree polynomial in time, and diagnosis year and encounter number, and the resulting plot of fitted values is shown in Figure 6. This figure incorporates the proportionality assumption, and cannot therefore be used to test it, but the important point is that the visual comparison of Figure 6 with Figure 5 does not suggest that the former is a gross misrepresentation of the latter.

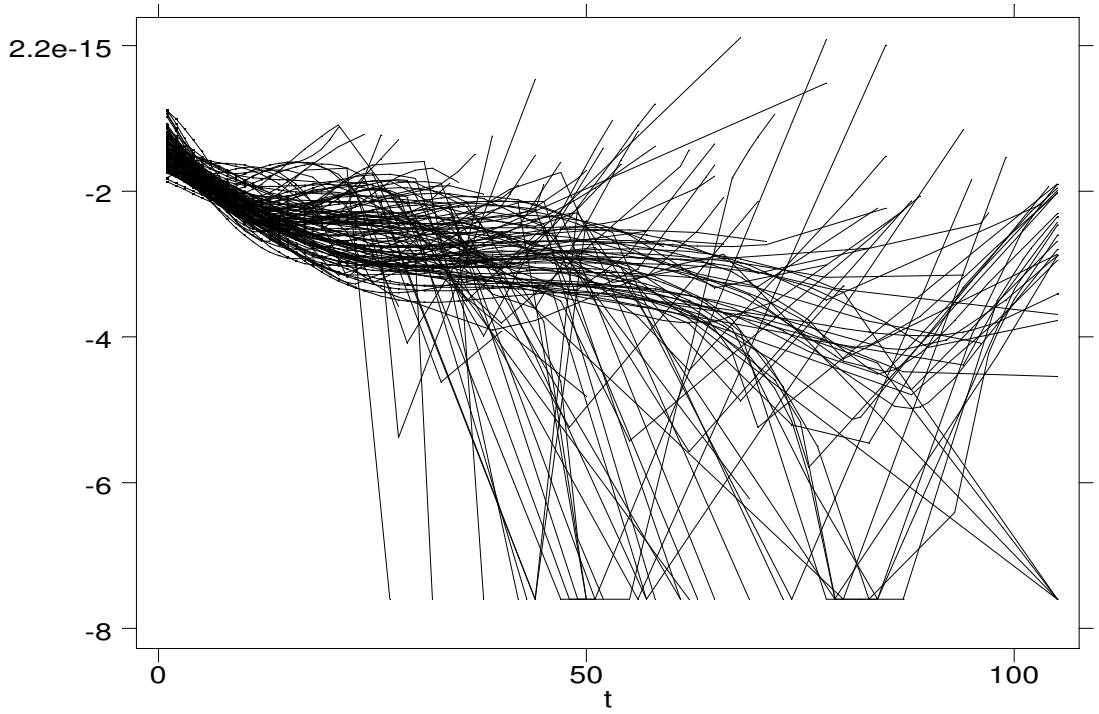


Figure 5. The curves of Figure 4 superimposed (t=Time in Weeks)

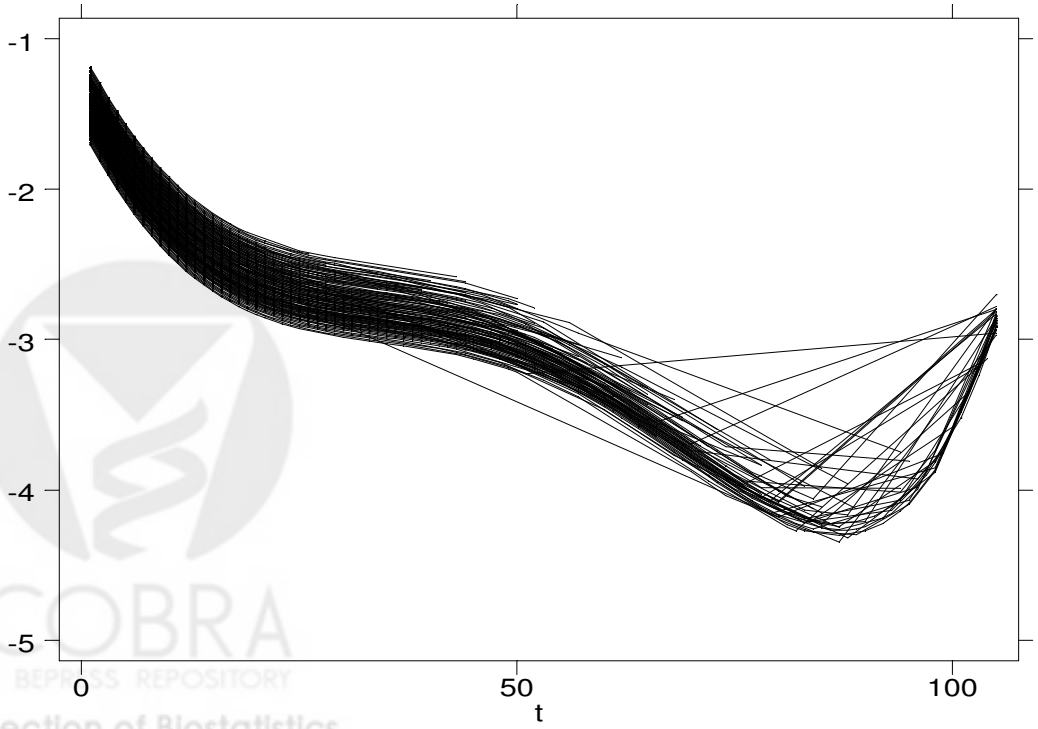


Figure 6. Fitted values of the ln incidence density for all diagnosis years and encounter numbers (t=Time in Weeks)

Finally, the ln incidence density regression was used to obtain fitted values of the incidence density itself (by applying exp to the predicted values of the ln), and then these were plotted against the actual values, in Figure 7 (for times before 24 weeks). The correlation between the two (both for all values and for values with times ≤ 24 weeks) is 0.925. Taken together, there appears to be reasonable visual and analytic evidence that the proportionality assumption would be acceptable for simulating encounters.

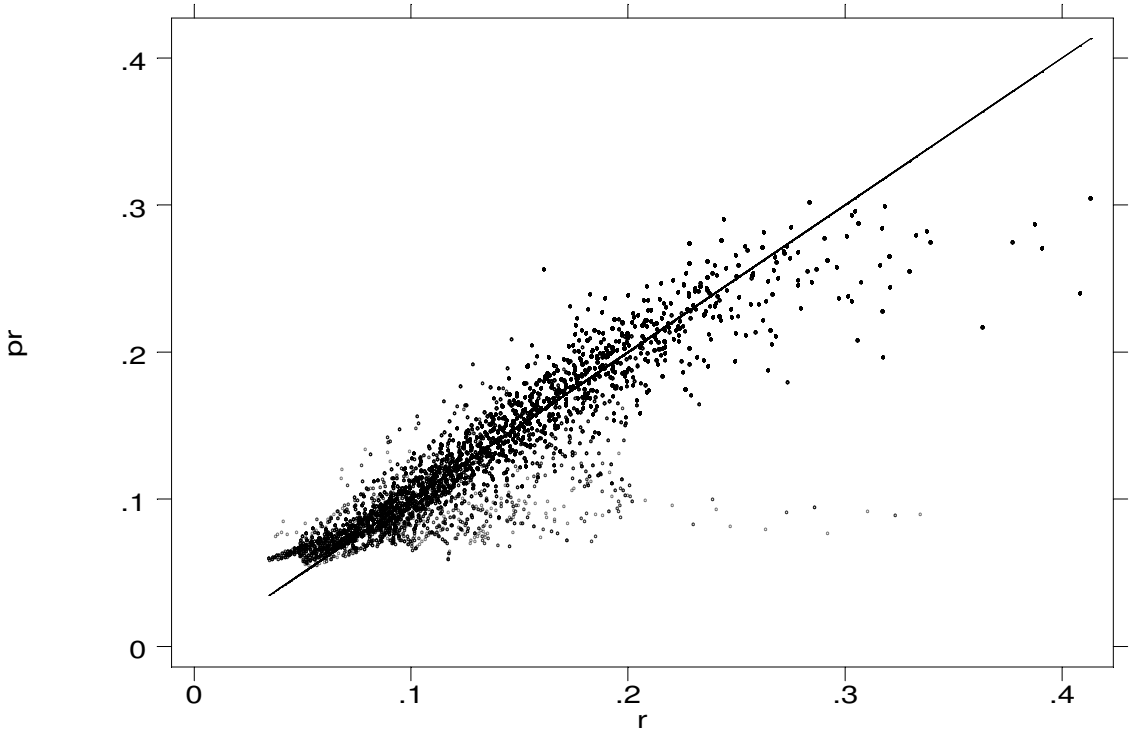


Figure 7. Agreement between the fitted (pr) and observed (r) values of the incidence density.

Whether proportionality is adequate for analytic purposes is a slightly different question, and one that can be addressed by fitting two kinds of models. The first is exponential regression, with the cumulative incidence function (specific to the diagnosis year and encounter number) playing the role of "time," and secondly, a Cox proportional hazards analysis, which implicitly estimates a single underlying cumulative incidence function for the entire dataset. We include terms yrdx = diagnosis year - 1980 and vn=serial number of encounter. (Each individual appears multiple times in this dataset, but the correlations between the encounter times are exceedingly small.)

The analyses appear in Table 1. In the exponential regressions, it makes virtually no difference to the effects of interest whether yrdx and vn effects are included or not, and even the linear effects terms in these variables are nonsignificant, which is of some note in this dataset of over 72,000 encounter times. The proportional hazards model (with a single underlying cumulative incidence function) gives more extreme estimates of

both coefficients, the difference amounting to about 1.8 SDEs for the male effect and about 1 SDE for the agedx effect. The proportional hazards model with linear yrdx and vn effects shows that they are both significant, and the result of including them is to move the estimates of the coefficients of interest closer to the values obtained by exponential regression. We can conclude that the proportional hazards model (with yrdx and vn) or the exponential regression (without them) are adequate for assessing gender and diagnosis age effects, and since the latter is simpler to present, and has adjusted for any possible confounding by time trend or encounter number, it may be the analysis of choice.

Exponential Regression				Proportional Hazards			
Cr	Coef.	Std. Err.	P>|z|	t	Coef.	Std. Err.	P>|z|
male	−.0564492	.0080196	0.000	male	−.0705685	.0080292	0.000
agedx	.0278315	.0030949	0.000	agedx	.0308443	.0030903	0.000
_cons	−.1477217	.0195335	0.000				
Cr	Coef.	Std. Err.	P>|z|	t	Coef.	Std. Err.	P>|z|
male	−.0563333	.0080402	0.000	male	−.061260	.008048	0.000
agedx	.0276774	.0031017	0.000	agedx	.0276288	.0031109	0.000
yrdx	.0006249	.0010574	0.555	yrdx	.0109975	.0010789	0.000
vn	.0001575	.0002237	0.481	vn	.0066517	.0002249	0.000
_cons	−.2057665	.0959636	0.032				

Table 1. Statistical analysis by exponential regression (left) and proportional hazards (right). Male is an indicator, age at diagnosis (agedx) is in decades, year of diagnosis (yrdx) is in years, and vn is the numeric encounter number.

Conclusions

Interest in recovery of the baseline cumulative incidence goes back at least as far as Breslow's (1974) estimate. A number of strategies have been proposed, including splines (Angelos et al. 1991, Gray 1994, Herndon and Harrell 1995) and kernel density estimates (Gray 1990). These methods have considerable theoretical appeal, and extend to the case of non-proportional hazards models, thus providing formal tests of the proportional hazards assumption. These methods are, however, computationally intensive those, and understanding their distributional properties requires advanced methods.

In contrast, the method proposed here requires, in addition to widely available software routines, only the solution of a simple, one-parameter likelihood problem. Moreover, the asymptotic likelihood-based analysis is transparent, yielding asymptotic SDEs for both the cumulative incidence function and its derivative, the incidence density (averaged over small time windows). One of the continuing problems in incidence density recovery is the tendency of estimation methods to give the appearance of "bumps" or other shapes, which scientists would like to interpret as meaningful, when in fact they reflect either random variability or artifacts of the estimation method. Penalized likelihood approaches (Gray 1994) are then used, but this leads to the problem of how

severe the penalty should be, increasing the complexity of the analysis. The proposed approach produces maximum likelihood estimates of the cumulative incidence function at certain points, and then uses reasonably stiff modeling functions to smooth the cumulative incidence function (by regression, or weighted regression), and then to estimate the incidence density by taking the derivative. This provides enough flexibility to fit a reasonable range of different forms for the cumulative incidence function, and also does a reasonable job at suppressing artificial bumps in the incidence density.

As illustrated here, the approach can be used routinely to generate incidence densities within subgroups. These can be used for judgments about the adequacy of the proportional hazards model across the subgroups. Perhaps more importantly, if one is concerned about confounding due to failure of proportional hazards, then the smoothed cumulative incidence functions in subgroups can be substituted for the "time" variable in exponential regression, and effect estimates of interest can be computed by what is essentially proportional hazards, with the hazard function varying by subgroup.

References

Angelos, J., Lee, C.M-S., Singh K.P. (1991) B-spline approximation for the baseline hazard function. Environmetrics 2:323-339

Breslow, N. (1974) Covariance analysis of censored survival data. Biometrics 30:89-99

Brown, JB, K.L. Pedula, and A.W. Bakst (1999) The progressive cost of complications in type 2 diabetes. Archives of Internal Medicine (in press)

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

Gray, R.J. (1990) Some diagnostic methods for Cox regression models through hazard smoothing. Biometrics 46:93-102

Gray, R.J. (1994) Spline-based tests in survival analysis. Biometrics 50:640-652

Kalbfleisch, J.D., and R.L Prentice (1980) The Statistical Analysis of Failure Time Data. John Wiley & Sons, New York, p. 84 ff.

Herndon II, J.E., and F.E. Harrell, Jr. (1995) The restricted cubic spline as baseline hazard in the proportional hazards model with step function time-dependent covariates. Statistics in Medicine 14:2119-2129.

Stata Press (1997). Stata Reference Manual, Release 5. College Station TX.