AN APPROACH TO ESTIMATION IN RELATIVE SURVIVAL REGRESSION

Maja Pohar*
Department of Biomedical Informatics
University of Ljubljana,
Vrazov trg 2, SI-1000 Ljubljana; Slovenia
e-mail: maja.pohar@mf.uni-lj.si

Robin Henderson
Mathematics & Statistics
University of Newcastle, UK

Janez Stare
Department of Biomedical Informatics
University of Ljubljana, Slovenia

*Corresponding author

Abstract

The goal of relative survival methodology is to compare the survival experience of a cohort with that of the background population. Most often an additive excess hazard model is employed, which assumes that each person's hazard is a sum of two components - the population hazard obtained from life tables and an excess hazard attributable to the specific condition. Usually covariate effects on the excess hazard are assumed to have a proportional hazards structure with parametrically modelled baseline. In this paper we introduce a new fitting procedure using the EM algorithm, treating the cause of death as missing data. The method requires no assumptions about the baseline excess hazard thus reducing the risk of bias through misspecification. It accommodates the possibility of knowledge of cause of death for some patients, and as a side effect the method yields an estimate of the ratio between the excess and population hazard for each subject. More importantly, it estimates the baseline excess hazard flexibly with no additional degrees of freedom spent. Finally it is a generalization of the Cox model, meaning that all the wealth of options in existing software for the Cox model can be used in relative survival.

The method is applied to a data set on survival after myocardial infarction, where it shows how a particular form of the hazard function could be missed using the existing methods.
Keywords: Relative survival; EM algorithm; additive model

1 Introduction

The goal of relative survival methodology is to compare the survival experience of a cohort with that of the background population. The observed cohort is defined by a certain condition, such as poverty, wealth, heart attack, diabetes, high blood

pressure etc, and the interest of the study lies in identifying the possible increase (or decrease) of the cohort mortality compared to the population. When an increase in mortality is expected, for example in cancer patients, an additive relative survival model is often used to model the effect of the covariates on the survival. The focus of this paper is on improving parameter estimation for this model.

Letting $S_O(t)$ and $S_P(t)$ denote the observed and population survival functions respectively, the cumulative relative survival function is defined by Ederer et al. (1961) as

$$r(t) = S_O(t)/S_P(t).$$

If $r(t)$ is a decreasing function we can write

$$S_O(t) \quad = \quad S_P(t) * r(t) = \exp[-\int_0^t \lambda_P(u)du] \exp[\int_0^t \lambda_E(u)du],$$

where λ_P is the population hazard and λ_E is the excess hazard experienced by the cohort. This implies an additive hazard relationship, $\lambda_O(t) = \lambda_P(t) + \lambda_E(t)$. The effect of a $p-$dimensional vector of covariates Z on relative survival can be incorporated via a regression model (Hakulinen and Tenkanen, 1987; Dickman et al., 2004)

$$\lambda_O(t, Z) = \lambda_P(t) + \lambda_0(t)e^{\beta Z}, \tag{1}$$

where $\lambda_0(t)$ denotes the baseline excess hazard, which is, for estimation purposes, usually taken to be piecewise constant over a partition of the follow-up interval $[0, \tau]$. Hence we can write

$$\lambda_O(t) = \lambda_P(t) + \exp[\sum_k \tau_k I_k(t)] \exp[\beta Z], \tag{2}$$

where $I_k(t)$ is an indicator function for the k-th time interval. The population

hazard $\lambda(t)$ is in practice a piecewise constant function as well, usually available in yearly intervals.

Although multiplicative models have also been suggested for relative survival (Andersen et al., 1985), the additive model (2) has had considerable attention and success (Hakulinen and Tenkanen, 1987; Dickman et al., 2004; Estève et al., 1990) and in the relative survival literature it seems to be almost exclusively the first-choice model. However, there is some variety in the choice of estimation procedures and a number have been proposed, some based on generalized linear models (Hakulinen and Tenkanen, 1987; Dickman et al., 2004) and extensions (Cheuvart and Ryan, 1991), and others on full maximum likelihood estimation (Estève et al., 1990).

In practice, however, the step function assumption for the baseline excess hazard is unrealistic and estimates can only be interpreted as averages over the specified intervals. While analysts often concentrate on coefficients in the model (relative hazards), the knowledge of the baseline excess hazard function behaviour in the additive model is crucial for understanding the whole picture and plays an important role in the interpretation of the model results and prognostics. Also, the baseline excess hazard is estimated simultaneously with the coefficients of the model and misspecification can lead to biased estimation of these coefficients. Diagnostics might reveal misspecification of the model, but it is invariably impossible to say where the misspecification comes from. We describe this in more detail in Stare et al. (2005b). It is therefore essential to have a flexible method to estimate the baseline excess hazard. Most of the work in this area (Giorgi et al., 2003; Lambert et al., 2005) has focussed on fully parametric approaches, an exception being Sasieni (1996).

In this paper, we propose a new approach to fitting the model (1) that makes no assumptions about the form of the baseline excess hazard and is based on an EM

algorithm with the cause of death treated as missing data. We introduce the idea and investigate basic properties in Section 2. Standard error estimation is described in Section 3. In Section 4 we describe how residuals and model extensions that have been developed for other fitting methods can be used with our method. We also provide some further extensions specific to the EM approach. Section 5 describes the properties of the EM approach estimates. Section 6 applies the EM-based fitting approach to a data set on survival after myocardial infarction. Some closing remarks in Section 7 complete the paper.

2 EM algorithm

We define the study cohort by the presence a particular condition $\mathcal{C}$ and assume interest is in mortality attributable to this condition, whether direct or indirect. We assume model (1) holds. Our proposal is simply stated and easily implemented: we treat cause of death as a potentially missing variable and adopt EM estimation.

Let δ_i be a death (1) or censoring (0) indicator for patient i and let δ_{Ei} be the indicator of a death attributable to condition $\mathcal{C}$. Analogously, δ_{Pi} is an other-cause death indicator. In some cases δ_{Ei} and δ_{Pi} will be explicitly interpreted, for example when we are interested in death from, say, myocardial infarction, but we lack information on primary cause of death. So δ_{Ei} would indicate myocardial infarction and δ_{Pi} would indicate any other cause. In other situations these terms may be less tangible. For instance, suppose we are interested in excess bladder cancer mortality of dyestuff workers, due to prolonged exposure to particular chemicals. A case of bladder cancer may be caused by the exposure ($\delta_{Ei} = 1$) or may have arisen anyway ($\delta_{Pi} = 1$). We have no way, and no need in this work, to identify these at the individual level.

We therefore assume that in $\delta_i = \delta_{Pi} + \delta_{Ei}$, we always observe δ_i, but may not know δ_{Pi} and δ_{Ei}. As usual we assume cause $\mathcal{C}$ contributes a sufficiently small proportion of all population deaths for all-cause mortality tables to be effectively the same as other-cause mortality if $\mathcal{C}$ is removed.

If the cause of death were known for all patients we might consider a cause-specific Cox model treating δ_{Ei} as the death/censoring indicator. The standard partial likelihood obtained by profiling out the baseline hazard (Andersen et al., 1993, p.482) would be

$$L(\beta|X) = \prod_{i=1}^{n} \left(\frac{\exp(\beta Z_i)}{\sum_{j \in R_i} \exp(\beta Z_j)} \right)^{\delta_{Ei}}, \tag{3}$$

where n is the total number of patients, R_i the risk set at the time of the ith patient event, and $X = \{Z, T, \delta_E\}$ the set containing all the data. The baseline hazard could be estimated nonparametrically by the Breslow estimator as usual.

The idea of our approach is to base an EM algorithm on the partial likelihood (3) even though we have not assumed proportional hazards but rather the proportional excess hazards model (1). In the Supplementary Material (*http://www.biostatistics.oxfordjournals.org*) we show the method is valid since the full data likelihood for our model profiled over nonparametric maximum likelihood estimates of the baseline excess $\lambda_0(t)$ provides the same score equations as those obtained from (3). Thus, given δ_{Ei} $(i = 1, 2, \ldots, n)$ we can estimate first β and then $\lambda_0(t)$ very easily.

On the other hand, if we knew the baseline excess hazard $\lambda_0(t)$ and the coefficients β, the conditional probability of patient i dying due to condition $\mathcal{C}$, given observed exit time t_i and death indicator δ_i, is shown in the Supplementary Material to be

$$
\begin{aligned}
P(\delta_{Ei}=1|\delta_i,t_i) &= \delta_i \frac{\lambda_{Ei}(t_i)}{\lambda_{Pi}(t_i)+\lambda_{Ei}(t_i)} \\
&= \delta_i \frac{\lambda_0(t_i)\exp(\beta Z_i)}{\lambda_{Pi}(t_i)+\lambda_0(t_i)\exp(\beta Z_i)}. \qquad (4)
\end{aligned}
$$

Hence iterating the partial likelihood maximization and updating the value of δ_{Ei} forms the EM algorithm. More specifically, the algorithm consists of the following steps (Dempster et al., 1977):

1. Specify initial values of unknown parameters $\theta : \theta^{(0)} = (\beta^{(0)}, \lambda_0^{(0)})$.

2. E-step: Obtain the log-likelihood for the working Cox model as

$$
\log L\left(\theta|X\right) = \sum_{i=1}^{n} (\beta Z_i - \log \sum_{j\in R_i} e^{\beta Z_j})\delta_{Ei},
$$

   and its expected value with respect to (4) as

$$
\begin{aligned}
Q(\theta,\theta^{(0)}) &= E\left\{\sum_{i=1}^{n}\left(\beta Z_i - \log\sum_{j\in R_i} e^{\beta Z_j}\right)\delta_{Ei}|\delta_i,t_i\right\} \qquad (5) \\
&= \sum_{i=1}^{n}\left(\beta Z_i - \log\sum_{j\in R_i} e^{\beta Z_j}\right) E(\delta_{Ei}|\delta_i,t_i) \\
&= \sum_{i=1}^{n}\left(\beta Z_i - \log\sum_{j\in R_i} e^{\beta Z_j}\right)\left(\frac{\lambda_0^{(0)}(t_i)e^{\beta^{(0)}Z_i}}{\lambda_{Pi}(t_i)+\lambda_0^{(0)}(t_i)e^{\beta^{(0)}Z_i}}\right)\delta_i.
\end{aligned}
$$

3. M-step: Maximize the Q function with respect to β to get new values of

parameters $\beta^{(1)}$. Estimate $\lambda_0^{(1)}$ using the Breslow estimator

$$\begin{aligned}\lambda_0^{(1)}(t_i) &= \frac{E(\delta_{Ei})}{\sum_{j\in R_i}\exp(\beta^{(1)}Z_j)} \\ &= \frac{\lambda_0^{(0)}(t_i)e^{\beta^{(0)}Z_i}}{\lambda_{Pi}(t_i)+\lambda_0^{(0)}(t_i)e^{\beta^{(0)}Z_i}}\left\{\sum_{j\in R_i}\exp(\beta^{(1)}Z_j)\right\}^{-1} \qquad (6)\end{aligned}$$

4. Back to step 2.

With respect to practical implementation of the algorithm, note the following:

- For each individual, only one number must be obtained from the population tables, i.e. the population hazard at the time of their death. No population data are used for censored patients.
- The Q function (5) is the log-likelihood of a weighted Cox model, enabling us to use any existing software that can deal with weighted Cox models.
- The extension to time-varying covariates can be handled in the usual way with the Cox model and Breslow estimator.
- Ties can also be handled in the usual way.

A different estimating method pursuing the same goal of estimating the coefficients in the additive relative survival model without specifying the form of the baseline excess hazard was introduced by Sasieni (1996).
To compare methods we will turn briefly to counting process notation, letting $N_i(t)$ be the observed number of events to time t for subject i. Then the score equation arising from (5) can be written

$$U(\beta)=\sum_i\int_0^\tau\left(Z_i-\frac{\sum_j Y_j(u)Z_je^{\beta Z_j}}{\sum_j Y_j(u)e^{\beta Z_j}}\right)w_i(u,\beta^{(0)})dN_i(u),$$

where τ is the maximum follow-up time, $Y_j(.)$ is an at-risk indicator, and

$$w_i(u,\beta) = \frac{\lambda_0(u)e^{\beta Z_i}}{\lambda_{Pi}(u) + \lambda_0(u)e^{\beta Z_i}}.$$

Sasieni's method by contrast is based on the discounted counting process

$$\tilde{N}_i(t) = N_i(t) - \int_0^t Y_i(u)\lambda_{Pi}(u)du$$

and involves the score

$$U_S(\beta) = \sum_i \int_0^\tau \left(Z_i - \frac{\sum_j w_j(u,\beta)Y_j(u)Z_j e^{\beta Z_j}}{\sum_j w_j(u,\beta)Y_j(u)e^{\beta Z_j}} \right) w_i(u,\beta)d\tilde{N}_i(u).$$

This requires $d\tilde{N}_i(u)$ at all times, not just event times. The integration is feasible when $\lambda_{Pi}(u)$ is piecewise constant, as it always is in practice, but nonetheless implementation is much easier under our approach, which can use standard software.

A problem common to all additive regression models occurs when there is little or no genuine excess hazard due to condition $\mathcal{C}$ over parts of the time scale. Since the estimates of $\lambda_0(t)$ are forced to be non-negative there can be finite-sample positive bias in $\hat{\lambda}_0(t)$ for at least some t, leading to overestimation of the cumulative baseline excess hazard. Some local smoothing of the baseline excess hazard function in the E-step of the algorithm is therefore needed. The details about the procedure used can be found in the supplementary material. An R function for fitting the additive model with the EM algorithm is available from CRAN (R Development Core Team, 2005) as a part of the `relsurv` package (Pohar and Stare, 2006).
The Sasieni method encounters similar problems. It also needs several reiterations and some baseline hazard smoothing to achieve the same goal.

In terms of practical use, an important advantage of our method is that it iterates

between two very standard routines: the Cox model fitting and a ratio calculation. Either part of the iteration is easy to implement using standard methods and therefore any extensions, such as for example splines or frailties, can be directly used. This does not apply to the Sasieni method, which perhaps explains why it seems to have been little used in practical relative survival applications.

3 Standard error estimation

Standard errors are estimated via the Fisher information matrix, which can be expressed as the complete information minus the missing information (see Supplementary Material). To evaluate the complete information we use the Hessian matrix obtained by fitting the Cox model at the final M step. To estimate the missing information contribution, $\mathcal{I}_m$ say, we use methods described by Louis (1982) to get

$$\begin{aligned}
\mathcal{I}_m &= \mathrm{Var}\left\{\frac{\partial \log L(\theta|Y,\delta_E)}{\partial\theta}\right\} \\
&= Var\left\{\sum_{i=1}^{n}\left(Z_i - \frac{\sum_{j\in R_i} Z_j e^{\beta Z_j}}{\sum_{j\in R_i} e^{\beta Z_j}}\right)\delta_{Ei}\right\} \\
&= \sum_{i=1}^{n}\left(Z_i - \frac{\sum_{j\in R_i} Z_j e^{\beta Z_j}}{\sum_{j\in R_i} e^{\beta Z_j}}\right)^{\otimes 2} Var(\delta_{Ei}). \qquad (7)
\end{aligned}$$

We thus estimate the observed information as

$$\begin{aligned}
\hat{\mathcal{I}}_O &= \sum_{i=1}^{n}\frac{\hat\lambda_{Ei}}{\hat\lambda_{Oi}}\left\{\frac{\sum_{j\in R_i} Z_i Z_i' e^{\hat\beta Z_i}}{\sum_{j\in R_i} e^{\hat\beta Z_j}} - \left(\frac{\sum_{j\in R_i} Z_i e^{\hat\beta Z_i}}{\sum_{j\in R_i} e^{\hat\beta Z_j}}\right)^{\otimes 2}\right\} \\
&\quad - \sum_{i=1}^{n}\left(Z_i - \frac{\sum_{j\in R_i} Z_i e^{\hat\beta Z_i}}{\sum_{j\in R_i} e^{\hat\beta Z_j}}\right)^{\otimes 2}\frac{\hat\lambda_{Ei}}{\hat\lambda_{Oi}}\left(1-\frac{\hat\lambda_{Ei}}{\hat\lambda_{Oi}}\right). \qquad (8)
\end{aligned}$$

4 Residuals and extensions

The most important advantage of our approach is the fact that the EM fitting method is based on the Cox model and therefore any further extensions used in the Cox model framework (splines, frailties, etc) for classical survival can be straightforwardly incorporated into the study of relative survival. This makes the method very flexible.

Similarly, goodness of fit of the model (1) with a non-parametric λ_0 can be addressed in the same way as in the case (2), described in Stare et al. (2005b). We define the partial residuals as

$$U_i(\beta) := Z_i(t_i) - \frac{\sum\limits_{j \in R_i} Z_j(t_i) \left\{ \lambda_{Pj}(t_i) + \lambda_0(t_i) e^{\beta Z_j} \right\}}{\sum\limits_{j \in R_i} \left\{ \lambda_{Pj}(t_i) + \lambda_0(t_i) e^{\beta Z_j} \right\}}. \tag{9}$$

Assuming that the integral of the baseline excess hazard function is bounded on the time interval of interest, proofs follow in the same way as for the stepwise case. We can therefore use the residuals U_i for a graphical examination of the proportional excess hazards assumption as well as for formal testing with Brownian bridge statistics. An example of goodness of fit checking is presented in Section 6.

An extension that is unique to the EM fitting method is to include the possibility of cause of death being reliably known for some but not all patients (Cheuvart and Ryan, 1991). As long as the availability of cause-of-death information is missing at random and ignorable for likelihood purposes, this additional piece of information can be straightforwardly incorporated. All that needs to be done is to fix the values of δ_{Ei} in the E step of the algorithm for the individuals whose cause of death is known. This partial information will lead to more precise estimation. An

example is given in Section B of the Supplementary Material.

Estimates of $E(\delta_{Ei}|\delta_i, t_i)$ are sometimes useful, i.e. the post-hoc probabilities that death is attributable to condition $\mathcal{C}$. These could of course be calculated after fitting with any method, but they are automatically generated by the EM procedure and hence immediately available. The sum of these values is the expected number of deaths due to the condition, which can give us an idea of the importance of using relative survival methods in our analysis. An example is given in Section 6.

5 Properties of the EM based approach

We performed a simulation study in order to evaluate the properties of the EM approach and compare it with the standard fully parametric approaches (for details see Supplementary Material). The simulations were designed so that the parametric model assumptions hold.
The estimated coefficients using either the EM approach or the parametric model are close and in both cases some bias is present when the percentage of condition-attributable deaths is low. The EM approach also seems to provide good variance estimates (8), on the one hand these are close to the actual observed variance of the coefficient, on the other hand the variance of the semi-parametric model is not much larger than that of the parametric one, even though the parametric model has an advantage of additional fulfilled assumptions.

The simulations also illustrate two properties of both both the semi-parametric and the parametric additive model. Firstly, the additive model may not be the best choice in situations with less than 30% deaths due to the excess risk, this coincides well with the recommendations of Sasieni (1996). Instead, one should

rather turn to using the multiplicative model (Andersen et al., 1985) or the transformation approach (Stare et al., 2005a).
Secondly, the variance of the estimates strongly depends on the proportion of deaths due to the excess risk. If the survival of the observed cohort is very similar to the survival of the population, the variance of the estimated coefficients for excess hazard can become very large, regardless of the fitting procedure.

6 An application to long-term survival after myocardial infarction

We have applied our method to data from a study of survival of patients after acute myocardial infarction. The study included all patients who were admitted to and later discharged from the Centre for Cardiovascular Diseases in Ljubljana, with the diagnosis of acute myocardial infarction, between May 1st 1982 and January 1st 1987. The aim of the study was to investigate the impact of different risk factors on mortality for patients who survive a myocardial infarction. A patient was considered a survivor of the infarction if she/he was discharged from the hospital, and survival times were measured from this point. Data on 1040 patients were collected, the follow-up was up to 14 years, during which time 53% of the subjects died. This is an appropriate scenario for relative survival methodology since, on the one hand, a substantial number of deaths would be expected even in a healthy population given the long follow-up, and on the other hand, Figure 1 implies that the excess risk is substantial.

Estimates of the regression coefficients for the variables age, sex and year of diagnosis are presented in Table 1. The table includes estimates using the EM procedure with nonparametric baseline and the maximum likelihood estimates

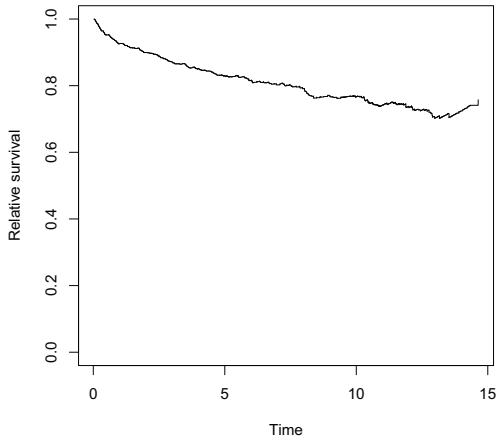

Figure 1: *The relative survival curve $r(t)$ for the myocardial infarction data.*

method	variable	coeff	se	z
EM	sex	0.6900	0.1730	3.997
	age	0.0307	0.0078	3.948
	year	−0.0005	0.0002	−3.043
step	sex	0.6885	0.1857	3.708
	age	0.0299	0.0083	3.616
	year	−0.0006	0.0002	−3.071
splines	sex	0.6914	0.1855	3.728
	age	0.0295	0.0083	3.558
	year	−0.0006	0.0002	−3.082

Table 1: *Results of the EM fitting method and two parametric options - piecewise constant (step) function and splines for the baseline excess hazard*

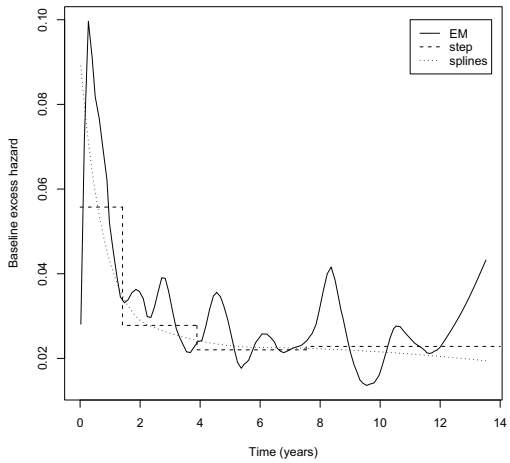

Figure 2: *Comparison of the baseline excess hazard estimated under three different models. The baseline hazard estimated by the EM method was further smoothed to enable a better comparison.*

using either splines or a piecewise constant baseline. In the case of splines we use a quadratic B-spline function with two interior knots as proposed in Giorgi et al. (2003) with the defaults as programmed in the `RSURV` function (Giorgi et al., 2005) (two knots set at the respective quantiles of event times). In the case of the piecewise constant baseline four intervals were chosen, with boundaries at the quartiles of the overall survival function. All three methods lead to similar estimated values.

Figure 2 shows the baseline hazard estimated by the three methods. Here, the two parametric models are used only for comparison and therefore we have not attempted to improve the original model setting (interval and knot position), which might lead to estimates closer to those given by the EM method. Without knowledge of the EM results of course we may have no reason to doubt the estimates obtained by the other methods.

The baseline excess hazard estimate obtained by the EM algorithm is of course

very ragged as the smoothing introduced was only intended to make the cumulative baseline excess hazard monotonically increasing. For an easier comparison, we further smoothed the estimate (using the R `loess` routine) to obtain the curve shown in Figure 2. The curve clearly shows that the baseline excess hazard starts low, increases to its highest point in the next few months, then steeply decreases for about a year and evens out thereafter. While the estimated coefficients with the three fitting options are very similar, the results differ materially with respect to the goodness-of-fit of the model. Test statistics based on cumulative sums of Schoenfeld-type residuals were given by Stare et al. (2005b). These behave in large samples like functions of Brownian bridges. One option is based on the maximum of a weighted bridge, with more weight at the beginning of the follow-up interval where risk sets are large. We used this for the myocardial infarction data and found differences between the approaches with respect to the effect of age. While the proportionality assumption seems to be violated when using splines (p=0.016), no problems are evident by the EM method (p=0.719). The reason for this disagreement lies in the fact that the baseline excess hazard is needed for the residual calculations (9) and the spline method gives much higher values for this over the important early months.

To explore this behaviour, we performed a simulation study based on the myocardial infarction data set. Death times were simulated using the estimates of the coefficients and the baseline excess hazard provided by the EM-based model (Table 1), with censoring mimicking that in the original data. The model was fitted using both the EM and spline approaches for each simulated data set. Figure 3 shows the two empirical cumulative distribution functions (cdf's) of the goodness-of-fit test statistics for age and gives for reference the appropriate theoretical asymptotic value for a correctly specified model. In the EM case there is no misspecification and the empirical cdf is very close to the theoretical one, so the test statistic will yield reliable results. On the other hand, in the spline case,

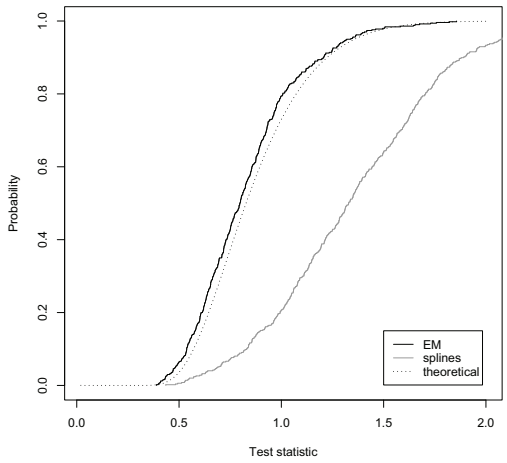

Figure 3: *Empirical cumulative distribution functions for the goodness of fit test statistics compared to the theoretical cdf.*

the parametric baseline excess hazard assumptions do not hold exactly and the test statistic too often leads to rejection of the null hypothesis, even though the hazard for age is proportional. Misspecification of the baseline can evidently lead to erroneous conclusions with respect to the proportionality effects of covariates.

As mentioned in Section 4, the EM method automatically generates estimated probabilities of dying due to the disease for each individual. By summing them up, we see that for this application the expected number of deaths due to the infarction is 256.9, some 47% of the deaths recorded in this study. Figure 4 presents the individual probabilities of dying due to the infarction plotted against follow-up time. As the average age at diagnosis is relatively low, at 63, the population hazard starts low and most deaths at the beginning of the follow-up period can be attributed to the infarction. As time progresses the population hazard increases rapidly, with the excess hazard simultaneously decreasing. The chance of disease-specific death therefore decreases over time, as can be seen in Figure 4.

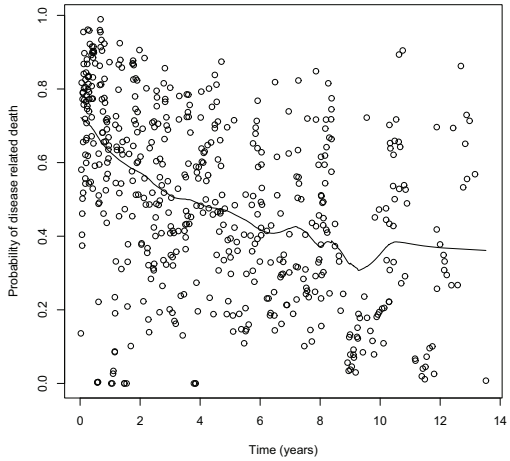

Figure 4: *Individual probabilities of dying due to the infarction plotted against the follow-up time, with smooth trend.*

7 Discussion

The main purpose of the newly introduced method is to fit the additive relative survival model with a non-parametric baseline excess hazard function. In practice, in cancer studies as in many other diseases it is reasonable to expect the excess hazard to be decreasing, at least during the first year after the diagnosis, and the assumption of a constant hazard will usually be invalid. While we can try making a better piece-wise constant approximation by forming more intervals, we quickly run into estimation problems. The more flexible methods often work well but can sometimes fail to capture subtle but important changes in baseline shape. In particular there are no methods for checking the assumptions about the baseline excess hazard, and therefore no way to detect problems with the spline model if they arise. Coefficient estimates can be, and goodness of fit statistics are, affected by misspecification of baseline hazards, and therefore misleading conclusions as to covariate effects may be obtained. When the assumptions about the baseline excess hazard hold, the performance of both parametric and non-

parametric approaches will be very similar, but when the assumptions are violated, the EM approach is more reliable. Therefore, if the results of the EM approach and a parametric model match, then either model could be used. However, if the results differ, we suggest the more flexible EM approach is preferable.

Apart from its performance and simplicity, an attraction of the proposed method lies in providing information about the form of baseline excess hazard as well as in automatically yielding the individual post-hoc disease related death probabilities, given death with unknown cause. In this way, the method helps in understanding the results of fitting an additive relative survival model.

Last but not least, the fitting method is nothing but an iteration between Cox model fitting and a simple ratio calculation. This means that the additive model can be seen as a generalization of the Cox model and the wealth of extensions available for the Cox model can be straightforwardly incorporated into relative survival.

Further work should include a study of the asymptotic properties of the newly proposed method as the usual properties of the EM algorithm can be affected by the smoothing in the E-step.

List of Figures

List of Tables

References

Andersen, P. K., Borch-Johnsen, K., Deckert, T., Green, A., Hougaard, P., Keiding, N., and Kreiner, S. (1985). A Cox regression model for relative mortality and its application to diabetes mellitus survival data. *Biometrics*, 41:921–932.

Andersen, P. K., Borgan, O., Gill, R. D., and Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer-Verlag, New York.

Cheuvart, B. and Ryan, L. (1991). Adjusting for age-related competing mortality in long-term cancer clinical trials. *Statistics in Medicine*, 10:65–77.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society*, 39:1–38.

Dickman, P. W., Sloggett, A., Hills, M., and Hakulinen, T. (2004). Regression models for relative survival. *Statistics in Medicine*, 23:51–64.

Ederer, F., Axtell, L. M., and Cutler, S. J. (1961). *The relative survival rate: a statistical methodology*, volume 6, pages 101–121. National Cancer Institute Monograph.

Estève, J., Benhamou, E., Croasdale, M., and Raymond, M. (1990). Relative survival and the estimation of net survival: elements for further discussion. *Statistics in Medicine*, 9:529–538.

Giorgi, R., Abrahamowicz, M., Quantin, C., Bolard, P., Estève, J., Gouvernet, J., and Faivre, J. (2003). A relative survival regression model using b-spline functions to model non-proportional hazards. *Statistics in Medicine*, 22:2767–84.

Giorgi, R., Payan, J., and Gouvernet, J. (2005). RSURV: A function to perform relative survival analysis with S-PLUS or R. 78:175–178.

Hakulinen, T. and Tenkanen, L. (1987). Regression analysis of relative survival rates. *Journal of the Royal Statistical Society — Series C*, 36:309–317.

Lambert, P. C., Smith, L. K., Jones, R. J., and Botha, J. L. (2005). Additive and multiplicative covariate regression models for relative survival incorporating fractional polynomials for time-dependent effects. *Statistics in Medicine*, 24:3871–3885.

Louis, T. A. (1982). Finding the observed information matrix when using the em algorithm. *Journal of the Royal Statistical Society — Series B*, 44:226–233.

Pohar, M. and Stare, J. (2006). Relative survival analysis in R. *Computer Methods and Programs in Biomedicine*, 81:272–278.

R Development Core Team (2005). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Sasieni, P. D. (1996). Proportional excess hazards. *Biometrika*, 83:127–141.

Stare, J., Henderson, R., and Pohar, M. (2005a). An individual measure of relative survival. *Journal of the Royal Statistical Society — Series C*, 54:115–126.

Stare, J., Pohar, M., and Henderson, R. (2005b). Goodness of fit of relative survival models. *Statistics in Medicine*, 24.