Appl. Statist. (1987)
36, *No.* 3, *pp.* 309–317

Regression Analysis of Relative Survival Rates

By. T. HAKULINEN† and L. TENKANEN

Finnish Cancer Registry

[Received April 1986. Revised December 1986]

SUMMARY

Survival from cancer or other chronic diseases is often measured using the relative survival rate. This, in turn, is defined as the ratio of the observed survival rate in the patient group under consideration to the expected survival rate in a group taken from the general population. At the beginning of the follow-up period, apart from the disease under study, factors affecting survival (e.g. age and sex) should be similar in the two groups. This paper outlines how a proportional hazards regression model may be adapted to the relative survival rates using GLIM. The method is illustrated by data on lung cancer patients diagnosed in Finland in 1968–1970.

Keywords: Survival time; Proportional hazards; Relative survival rate; Life table; Competing risks; GLIM; Generalized linear models

1. Introduction

In medical follow-up studies, deaths from causes other (D^c) than the actual disease (D) afflicting the patients under observation unduly reduce the survival rate below what it would be if the patients' disease D were the only possible cause of death. For example, for reasons not necessarily related to D, the observed survival rates of old and young patients are not comparable.

Information on the patients' causes of death may not always be suitable for correcting survival rates (Chiang, 1968) as it may be vague or unavailable (Ederer *et al.* 1963, Hakulinen and Teppo, 1977). Berkson and Gage (1950) estimated the impact of D^c from national life tables. The estimation was based on the assumption that D^c would act independently of D. The resulting quantity, the relative survival rate, was defined as the ratio of the observed survival rate for the group of patients under consideration to the survival rate expected for a group taken from the general population similar to the patients at the beginning of the follow-up period in all possible factors affecting survival except for D. Similarity is usually sought with respect to sex, age and calendar time. Under the independence assumption cited, the relative survival rate should be the survival rate in the event of the patients' disease being the only possible cause of death. Even without the independence assumption, the relative survival rate can be interpreted as the ratio between the observed and the expected proportions of survivors, the latter being based on survival of the general population (Hakulinen, 1977).

Generalized linear models (Baker and Nelder, 1978) have been applied recently to fit additive and multiplicative models to grouped survival data without reference to a general population group (Aranda-Ordaz, 1983; Tibshirani and Ciampi, 1983). These models allow simultaneous analysis of the effect of many prognostic factors on observed survival. Pocock *et al.* (1982) considered a proportional hazards regression model for the relative survival rate. In the

† *Address for correspondence*: Finnish Cancer Registry, LII SANKATU 21 B, 00170, HELSINKI, Finland

0035–9254/87/36309 $2.00

following it is shown how this model, which is a combination of additive and proportional hazards for the total mortality, can be presented within the framework of generalized linear models. The method is illustrated with data on lung cancer patients diagnosed in Finland in 1968–1970.

2. Model

As a basis for reporting the data, the follow-up time is here divided into a number, say g, of subintervals $[x_i, x_{i+1})$, $i = 1, 2, \dots, g$. At large treatment centres, survival is often investigated only at the endpoints of these subintervals. Following Pocock *et al.* (1982), the total force of mortality for the kth patient (group) at a point x during the interval $[x_i, x_{i+1})$ may be expressed as

$$\mu_{ki}(x) = \exp\left\{\alpha_i(x) + \sum_{v=1}^{p} \beta_v z_{vk}\right\} + \mu_{ki}^*(x) \tag{1}$$

in which the $\mu_{ki}^*(x)$ is the force of mortality (due to D^c) in a general population free of the particular disease D. Here $v_{ki}(x) = \mu_{ki}(x) - \mu_{ki}^*(x)$ is the force of mortality due to the disease D and follows a proportional hazards model (Cox, 1972) with $\alpha_i(x)$ and β_v, $v = 1, 2, \dots, p$ as constants. The constant $\alpha_i(x)$ depends on time $x \in [x_i, x_{i+1})$. The values for the independent variables z_{vk} may be different for different intervals $[x_i, x_{i+1})$, $i = 1, 2, \dots, g$.

The figures of interest are the interval-specific survival rates for those alive at the beginning of the interval in the presence of, cause D only, all causes of death, and causes D^c only, respectively:

$$r_{ki} = \exp\left\{-\int_{x_i}^{x_{i+1}} v_{ki}(x)dx\right\},$$

$$p_{ki} = \exp\left\{-\int_{x_i}^{x_{i+1}} \mu_{ki}(\mathrm{x})\mathrm{dx}\right\}$$

and

$$p_{ki}^* = \exp\left\{-\int_{x_i}^{x_{i+1}} \mu_{ki}^*(x)dx\right\}.$$

The quantities r_{ki}, p_{ki} and p_{ki}^* are often called the relative, observed and expected survival rates, respectively, of the patients during the interval $[x_i, x_{i+1})$ (Ederer *et al.*, 1961). In practical applications the expected survival rates p_{ki}^* are often adopted from life tables for the general population by considering mortality due to D as only a small fraction of the total mortality in a comparable general population(Ederer *et al.*, 1961; Hakulinen, 1982). Moreover, each p_{ki}^* is regarded as a constant because it is based on a section of the large general population. These conventions are followed in this paper, as well.

Equation (1) implies a generalized linear model (Baker and Nelder, 1978) for the p_{ki} as follows:

$$\ln\{-\ln(p_{ki}/p_{ki}^*)\} = \gamma_i + \sum_{v=1}^{p} \beta_v z_{vk}, \tag{2}$$

in which

$$\gamma_i = \ln\left[\int_{x_i}^{x_{i+1}} \exp\{\alpha_i(x)\}dx\right].$$

Using the *GLIM* system, estimation of the parameters in the model is based on an individual *LINK* function, complementary log-log combined with a division by p_{ki}^*, for each observation or stratum of observations during each interval $[x_i, x_{i+1})$, and on a binomial error function.

3. The Macros and Example

Because of the individual *LINK* function for each observation, the standard link and error functions in *GLIM* cannot be applied. Hence, four model macros are employed for the user-defined model:

```
$MAC M1 $CAL EXLP = – %EXP(%LP)

  $CAL %FV = %EXP(EXLP)*LD*PS $ENDMAC

$MAC M2 $CAL %DR = 1/(%FV*EXLP) $ENDMAC

$MAC M3 $CAL %VA = %FV*(1 – %FV/LD) $ENDMAC
```

and

```
$MAC M4 $CAL %DI = 2*(%YV*%LOG(%YV/%FV)
+ (LD – %YV)*%LOG((1 – %YV/LD)/(1 – %FV/LD))) $ENDMAC
```

In the macros, LD = the number of patients in the stratum at the beginning of the follow-up interval and $PS = p_{ki}^*$ for the patient stratum in question. In the example below, a constant of 0.8 has been employed as the initial value for $\%LP$ in the iterations. The directive $\$WARNING$ was applied to avoid *GLIM* warnings when using the macros.

The example involves 5145 male lung cancer patients from Finland whose cancer was diagnosed in 1968–1970 (Table 1). The data were reported to the Finnish Cancer Registry, which is population-based and covers the whole country (Hakulinen *et al.*, 1981). The patients were followed up until the end of 1981. Annual follow-up intervals were employed. For simplicity this example takes into account only the first eight years of follow-up.

Stage of cancer and patient's age at diagnosis were treated as categorical variables. Three classes according to stage and five classes according to age were employed to divide the patients into strata (Table 1). Because there were eight annual follow-up intervals for each stratum the number of potential observation units for *GLIM* was $3 \times 5 \times 8 = 120$.

The annual expected survival rates p_{ki}^* for each stratum of patients were computed as

TABLE 1

The number of new cases of lung cancer (N) and of those alive after an eight-year follow-up (S) in males diagnosed in Finland in 1968–1970, classified according to age and stage. The cumulative eight-year expected survival rates in the comparable general population group (P) (Ederer and Heise, 1959, cf. Hakulinen, 1982) are also shown*

	Stage								
Age	*Localized*			*Non-localized*			*Unknown*		
	N	*S*	*P**	*N*	*S*	*P**	*N*	*S*	*P**
0–49	94	30	0.932	225	7	0.928	28	5	0.934
50–59	429	75	0.834	755	15	0.832	172	14	0.832
60–69	705	61	0.688	1097	13	0.691	394	13	0.686
70–79	335	5	0.466	463	2	0.467	277	4	0.437
80–	54	–	.	62	–	.	55	–	.

TABLE 2

Results of fitting proportional hazards regression models to the relative survival rates of male lung cancer patients diagnosed in Finland in 1968–1970 (agecat and agecon: age at diagnosis considered as a categorical and as a continuous variable, respectively)

Model	*Deviance*	*D.f.*
1. General mean only	2296	106
2. Model 1 + follow-up year	772.2	99
3. Model 2 + stage	254.3	97
4. Model 3 + agecat	142.9	93
5. Model 4 + stage.agecat	122.7	85
For comparison:		
6. Model 3 + stage.agecon	132.9	94
7. Model 6 + stage.$(\text{agecon})^2$	125.5	91
8. Model 6 + stage.$(\text{agecon})^2$ (non-loc. only)*	125.5	93

*The last term was allowed for non-localized cases only. The corresponding coefficients for localized and unknown stages were set to zero.

averages of the expected survival probabilities of the individuals alive at the beginning of each follow-up year (Cutler and Axtell, 1963; Hakulinen, 1982). To illustrate, the cumulative eight-year expected survival rates (Ederer and Heise, 1959; Hakulinen, 1982), are also given in Table 1.

Table 2 describes the steps in the fitting of the model. A graphical examination of the fit (Fig. 1) indicated systematic deviation in models 3 and 4 among older patients with localized tumours, when only the main effects of the factors were included in the model. In Fig. 1, the values for relative survival rates exceeding one indicate a better-than-average survival compared with the general population but are as a rule based on no observed deaths among a rather small population under follow-up. The fit was improved by adding an interaction term for age and stage (Table 2). After that, Pearson's chi-squared statistic divided by the corresponding degrees of freedom still tended to indicate over-dispersion (coefficient 1.28 for binomial variance). However, this does not affect the point estimation of the parameters (McGullagh and Nelder, 1983; p.80).

Risk ratios, i.e. the ratios between the estimated hazard rates, were used to measure the effect of the prognostic factors. The parameter estimates (Table 3) quantified the well-known result (e.g. Cancer Registry of Norway, 1975; Axtell *et al.*, 1976; Hakulinen *et al.*, 1981) that the risk of dying from lung cancer increased in lung cancer patients with the spread of the tumour and with increasing age. Correlations between the estimators, especially those related to stage, were also noticed (Table 4). The correlations involving the parameter estimates $\hat{\gamma}_i$ related to follow-up intervals (cf. equation 2) ranged from 0.01 to 0.15. The joint effect of the stage and age factors on the risk of death from lung cancer was less than multiplicative (Fig. 2).

The results could be conveniently summarized by regarding age at diagnosis as a continuous variable. The effect of age on the relative risk was linear within the localized and unknown stage whereas a quadratic term was needed for non-localized cases (Fig. 2). The deviance was virtually unchanged when the second-degree age terms were set to zero for localized and unknown stages but two degrees of freedom were saved (Table 2). An estimated coefficient of 1.20 for binomial variance indicated the presence of some over-dispersion.

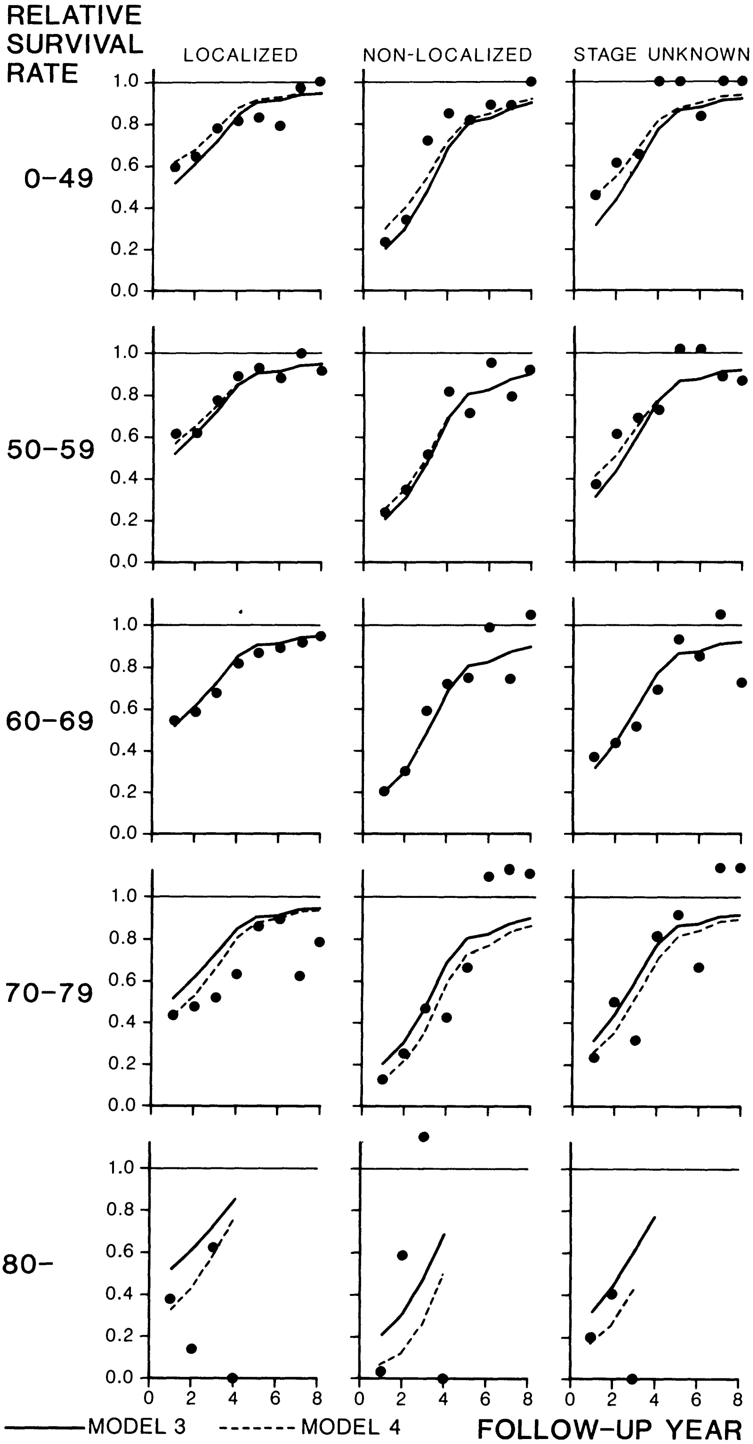

Fig. 1. Graphical illustration of the fit of models 3 (solid lines) and 4 (dashed lines) to the relative survival rates (small circles) of male lung cancer patients diagnosed in Finland in 1968–1970, according to stage, age and follow-up year (see Table 2 for models).

TABLE 3

Relative risks of dying from lung cancer (RR) and their approximate 95% confidence intervals (CI) in male lung cancer patients diagnosed in Finland in 1968–1970, according to stage and age. The relative risks in the groups with the localized stage and aged 60–69 years, respectively, have been defined as unity

Factor and level	*Model 3**		*Model 4**	
	RR	*CI*	*RR*	*CI*
Stage				
-localized	1.00		1.00	
-non-localized	2.39	(2.11, 2.71)	2.44	(2.21, 2.68)
-unknown	1.67	(1.42, 1.96)	1.57	(1.39, 1.78)
Age (years)				
- 0–49	.		0.77	(0.60, 0.90)
-50–59	.		0.88	(0.80, 0.98)
-60–69	.		1.00	
-70–79	.		1.34	(1.20, 1.50)
-80-	.		1.76	(1.36, 2.28)

* Cf. Table 2

4. Discussion

The advantage of the present methodology is that it brings the analysis of relative survival rates within the framework of the generalized linear models. Unlike the software prepared for the analysis of these rates specifically (Hakulinen and Abeywickrama, 1985), the *GLIM* package is widely available throughout the world. Moreover, estimation and testing can be treated uniformly in a regression analysis environment. The estimation is a regression extension of the method devised by Ederer and Heise (1959)–cf. also Rothman and Boice (1979), Hakulinen (1982). The tests are an extension of those suggested by Hakulinen *et al* (1987). The regression technique provides a parsimonious way to condense larger sets of survival data and to analyze prognostic factors having an effect on the relative survival rates. In addition to relative survival rates, risk ratios related to the prognostic factors are also obtained.

TABLE 4

Correlations between estimators of the stage and age parameters in model 4 (cf. Table 2)

	Stage		*Age*			
	Non-localized	*Unknown*	0–49	50–59	70–79	80–
Stage						
Non-localized	1.00					
Unknown	0.48	1.00				
Age						
0–49	−0.05	0.03	1.00			
50–59	−0.03	0.04	0.24	1.00		
70–79	0.05	−0.05	0.21	0.33	1.00	
80–	0.05	−0.04	0.08	0.14	0.14	1.00

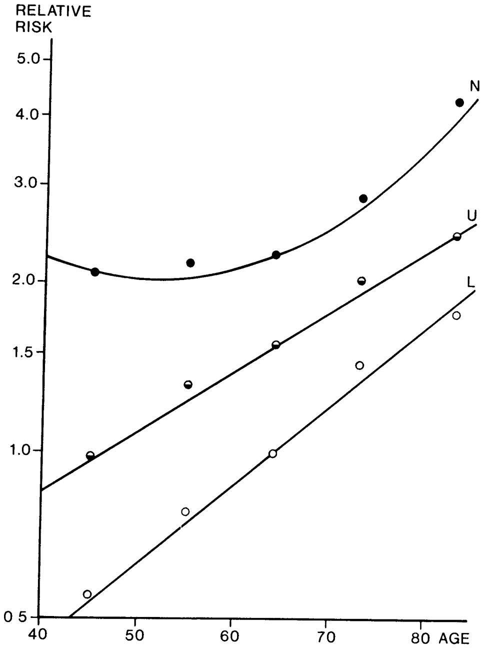

Fig. 2. Age-specific relative risks of dying from lung cancer in male lung cancer patients diagnosed in Finland in 1968–1970 according to stage (L = localized, N = non-localized, U = unknown). The points indicate estimates under model 5, the lines those under model 8 (cf. Table 2). In both models the relative risk for the group with localized stage aged 60–69 years has been defined as unity.

A drawback is that with *GLIM* items in larger data sets need to be combined. As pointed out by Clayton and Cuzick (1985), when the proportional hazards model is analysed with *GLIM*, the number of "observation units" is the number of person-trials (person-intervals) rather than the number of persons. The 5145 patients included in the example may be treated as individual patients using the program by Hakulinen and Abeywickrama (1985); with *GLIM*, strata must be formed in order to prevent the material from becoming too large. If the number of follow-up intervals is increased, say by shortening them to a length of 2 months (6 intervals per year), the number of prognostic factors or their levels may have to be decreased due to data size limitations on *GLIM*. On the other hand, stratification is not a problem with small clinical materials in which each patient can be his own stratum with $LD = 1$ and the number of deaths either zero or one in a follow-up interval (cf. McCullagh and Nelder, 1983; p. 79).

Due to an efficient system of population registry in Finland and to a sufficiently long potential follow-up for every patient, there were no withdrawals from the follow-up in this example (Hakulinen, 1982). Persons due to withdraw may be treated in the analysis by removing them from the material at the beginning of the interval during which they are due to withdraw. Another possibility, which saves on material, is to use the "effective" numbers of patients at risk (Cutler and Ederer, 1958).

The method provides an alternative to the least squares procedure used by Pocock *et al.* (1982). The binomial errors for the observed survival rates applied in the present paper may be a more reasonable choice than the Poisson errors assumed by Pocock *et al.*, at least, if fatality among patients is high. In fact, Pocock *et al.* applied the method to those surviving after a five-year follow-up, when the fatality of breast cancer patients considered there was lower.

As illustrated by the example, *GLIM* also facilitates analysis of the combined effects of prognostic factors other than the multiplicative ones specified by the proportional hazards model. An interaction may be introduced not only between the prognostic factors but also between a factor and the follow-up year. This implies the assumption of different baseline hazards for different patient strata (Thall and Lachin, 1986). Major interactions would indicate that a proportional hazards model may not be appropriate for mortality due to D.

Some authors (Moon, 1979; Breslow *et al.*, 1983; Andersen, 1984) have considered a model in which the hazard rate due to D is proportional instead of additive to the hazard rate in a general population. This means, for example, that the D-specific mortality in young patients is low and is proportional to the overall mortality in a general young population, whereas the D-specific mortality in old patients is high and is proportional to the overall mortality in a general old population. This model may provide a good alternative when the additivity assumption does not hold. On the other hand, a model in which the mortality due to D is proportional to the overall mortality of a general population can be easily fitted with *GLIM* using the *$OFFSET* technique, and may consequently be an attractive initial choice. If this model does not fit, the model suggested in this paper in which the D-specific mortality is additive to the overall mortality, is a good alternative worth considering.

References

Andersen, P. K. (1984) A Cox regression model for the excess mortality in long-term follow-up studies. In *XIIth International Biometric Conference Proceedings*. Invited Papers. Tokyo, pp. 157–162.

Aranda-Ordaz, F. J. (1983) An extension of the proportional-hazards model for grouped data. *Biometrics*, **39**, 109–117.

Axtell, L. M., Asire, A. J. and Myers, M. H. (eds) (1976) *Cancer Patient Survival, Report No. 5.* DHEW Publication No. (NIH) 77–992. Bethesda: U. S. Department of Health, Education and Welfare.

Baker, R. J. and Nelder, J. A. (1978) *The GLIM System, Release 3. Generalized Linear Interactive Modelling.* Oxford: Numerical Algorithms Group.

Berkson, J. and Gage, R. P. (1950) Calculation of survival rates for cancer. *Proc. Staff Meet. Mayo Clin.*, **25**, 270–286.

Breslow, N. E., Lubin, J. H., Marek, P. and Langholz, B. (1983) Multiplicative models and cohort analysis. *J. Amer. Statist. Ass.*, **78**, 1–12.

Cancer Registry of Norway (1975) *Survival of Cancer Patients. Cases Diagnosed in Norway 1953–1967.* Oslo: Norwegian Cancer Society.

Chiang, C. L. (1968) *Introduction to Stochastic Processes in Biostatistics.* New York: Wiley.

Clayton, D. and Cuzick, J. (1985) The *EM* algorithm for Cox's regression model using GLIM. *Appl. Statist.*, **34**, 148–156.

Cox, D. R. (1972) Regression models and life-tables (with Discussion). *J. R. Statist. Soc.* B, **34**, 187–220.

Cutler, S. J. and Axtell, L. M. (1963) Partitioning of a patient population with respect to different mortality risks. *J. Amer. Statist. Ass.*, **58**, 701–712.

Cutler, S. J. and Ederer, F. (1958) Maximum utilization of the life table method in analyzing survival. *J. Chron. Dis.*, **8**, 699–712.

Ederer, F., Axtell, L. M. and Cutler, S. J. (1961) The relative survival rate: a statistical methodology. *Natl. Cancer Inst. Monogr.*, **6**, 101–121.

Ederer, F., Cutler, S. J., Goldenberg, I. S. and Eisenberg, H. (1963) Causes of death among long-term survivors from breast cancer in Connecticut. *J. Natl. Cancer Inst.*, **30**, 933–947.

Ederer, F. and Heise, H. (1959) Instructions to IBM 650 programmers in processing survival computations. Methodological Note No. 10. Bethesda: End Results Evaluation Section, National Cancer Institute.

Hakulinen, T. (1977) On long-term relative survival rates. *J. Chron. Dis.*, **30**, 431–443.

Hakulinen, T. (1982) Cancer survival corrected for heterogeneity in patient withdrawal. *Biometrics*, **38**, 933–942.

Hakulinen, T. and Abeywickrama, K. H. (1985) A computer program package for relative survival analysis. *Comp. Progr. Biomed.*, **19**, 197–207.

Hakulinen, T., Pukkala, E., Hakama, M., Lehtonen, M., Saxén, E. and Teppo, L. (1981) Survival of cancer patients in Finland 1953–1974. *Ann. Clin. Res.*, **13**, Suppl. 33.

Hakulinen, T., Tenkanen, L., Abeywickrama, K. H. and Päivärinta, L. (1987) Testing equality of relative survival patterns based on aggregated data. *Biometrics*, **43**. In press.

Hakulinen, T. and Teppo, L. (1977) Causes of death among female patients with cancer of the breast and intestines. *Ann. Clin. Res.*, **9**, 15–24.

McCullagh, P. and Nelder, J. A. (1983) *Generalized Linear Models.* London: Chapman and Hall.

Moon, S. Y., Woolson, R. F. and Bean, J. A. (1979) A computer program for survival comparisons to a standard population. *Comp. Progr. Biomed.*, **10**, 91–104.

Pocock, S. J., Gore, S. M. and Kerr, G. R. (1982) Long term survival analysis: the curability of breast cancer. *Statist. Med.*, **1**, 93–104.

Rothman, K. J. and Boice, J. D. (1979) *Epidemiologic Analysis with a Programmable Calculator*. NIH Publication No. 79–1649. Bethesda: U.S. Department of Health, Education and Welfare.

Thall, P. F. and Lachin, J. M. (1986) Assessment of stratum-covariate interactions in Cox's proportional hazards regression model. *Statist. Med.*, **5**, 73–83.

Tibshirani, R. J. and Ciampi, A. (1983) A family of proportional- and additive- hazards models for survival data. *Biometrics*, **39**, 141–147.