STATISTICS IN MEDICINE, VOL. 15, 2177-2189 (1996)

REGRESSION MODELS FOR INTERVAL CENSORED SURVIVAL DATA: APPLICATION TO HIV INFECTION IN DANISH HOMOSEXUAL MEN

BENDIX CARSTENSEN

Danish Veterinary Laboratory, Bülowsvej 27, DK-1790 Copenhagen V, Denmark, bxc@svs.dk

SUMMARY

This paper shows how to fit excess and relative risk regression models to interval censored survival data, and how to implement the models in standard statistical software. The methods developed are used for the analysis of HIV infection rates in a cohort of Danish homosexual men.

1. INTRODUCTION

Interval censored data arise when dates of occurrence of events are not known exactly but only by intervals, that is, when it is either known that an event occurred between two dates of testing, or only that no event has occurred by some date. Such data may arise when people are tested for some (typically asymptomatic) condition like HIV-positivity at fixed dates, but where not all have been tested on all occasions, so that for some people we may only know, for example, that HIV-infection occurred during two or three adjacent intervals. Data of this kind are often termed 'panel data'.

A special case is grouped data, where the status of everyone is known at all of a set of fixed time points. Grouped data are easily handled by standard methods because the likelihood factorizes to a set of conditional probabilities. The interval censoring problem becomes non-trivial when the status of some people is unknown at some time points.

Sometimes the event of interest is ill-defined in terms of exact date of occurrence, because a substantial span of time may pass from event to symptoms (for example, onset of cirrhosis of the liver). If ascertainment of the condition is retrospective through hospital records or the like, we may have interval censoring where the intervals in which events are known to have occurred have no or very few common endpoints between individuals. This can also arise with panel data if the time scale of interest is not calendar time, but, for example, age.

Censoring with no or few common endpoints is in principle not different from the first kind, since we can define (a very large) number of fixed intervals by using all endpoints occurring in the data. In the 'non-parametric' setting, as discussed by Peto,[1] Turnbull[2] and Becker and Melbye,[3] this may create some problems with standard errors of estimates, since these methods essentially estimate one parameter per interval.

This paper extends the method of Becker and Melbye to regression models for the intensities of the underlying process. Both additive excess risk models and additive and multiplicative relative risk models will be considered. In this extension it is necessary to assume piecewise constant intensities, but this will not in practice be a serious limitation of the models.

CCC 0277-6715/96/202177-13

Received February 1993
Revised November 1995

Table I. Pattern of diagnosis among 297 Danish homosexuals

Last seen HIV-negative	First seen HIV-positive							Total
	12/81	04/82	03/83	09/84	04/87	05/89	Never	
Never	24	·	2	·	·	·	·	26
Dec 81	·	4	1	10	3	4	61	83
Apr 82	·	·	4	·	1	·	8	13
Mar 83	·	·	·	3	·	2	15	20
Sep 84	·	·	·	·	5	1	22	28
Apr 87	·	·	·	·	·	1	34	35
May 89	·	·	·	·	·	·	92	92
Total	24	4	7	13	9	8	232	297

2. HIV-INFECTION AMONG DANISH HOMOSEXUALS

A cohort of homosexual men from two cities in Denmark has been examined for HIV-antibody positivity on six different dates: December 1981; April 1982; February 1983; September 1984; April 1987, and May 1989. Not everyone appeared for examination at all dates. A total of 297 people have been tested at least once, and 65 diagnosed with infection. Table I shows the time-pattern of diagnoses of HIV-infections in the dataset. A more detailed description of the study may be found in Melbye *et al.*[4]

A question of interest, apart from the description of the intensity of HIV-infection in Denmark by calendar time, is to assess whether sexual habits, as measured by the number of sexual partners per year, influence the infection rates. To this end, year of birth, contact with people from the U.S.A. and number of partners per year were used as covariates.

A preliminary (and indeed frequently used) analysis is based on imputing event dates as the midpoint between the last date seen well and the first date seen ill for each patient, and then calculating a Kaplan–Meier estimator of the cumulative proportion infected. This requires assignment of a date when individuals who are infected at first contact are assumed to have been uninfected; the estimate of the cumulative proportion infected will depend on this choice.

Figure 1 shows Kaplan–Meier estimates of the cumulative proportion infected resulting from choosing different dates when everyone is assumed uninfected. There are considerable differences between these estimates initially, but these disappear around 1988, partly because the infection rate has fallen dramatically by then. The simple approach of just imputing event dates can give misleading results, depending on arbitrary assumptions. A curve showing the estimate based on the method of Becker and Melbye[3] is superimposed; this shows that some 30 per cent are HIV-positive by 1990.

2.1. Regression analysis

A question of interest is whether there are differences in the rate of seroconversion between subgroups in the cohort. This can be addressed by estimating the cumulative infection rate for subgroups of the cohort. However, with covariates such as age and number of partners which have a large number of values, this procedure quickly becomes unfeasible.

It is therefore desirable to use regression models where the influence of covariates can be assessed on some suitable scale. The usual choice of scale is relative risk, leading to proportional hazards models, which can be formulated either as multiplicative in the covariates (Cox model) or

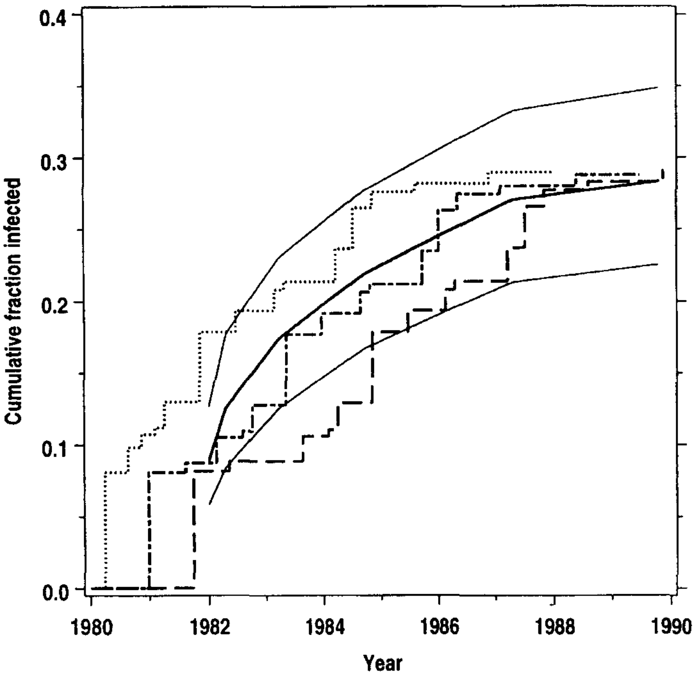

Figure 1. Cumulative fraction of HIV-positive in a cohort of Danish homosexual men, as estimated in the piecewise constant intensity model (with 95 per cent confidence limits based on likelihood profiles) (solid lines), and as estimated by the Kaplan–Meier method using the dates June 78, December 79 and June 81, respectively, as dates of no infection present (broken lines)

additive on the relative risk scale. Another choice is the additive excess risk model, where the covariates act additively on the hazard scale. These models will be considered in detail below.

To introduce regression models I first review and rephrase the ideas of Becker and Melbye.[3]

3. ESTIMATION OF INTENSITIES FROM INTERVAL CENSORED DATA

It is assumed that the condition studied is irreversible, so that once an event has occurred, the individual is removed from the population at risk. Also, it is assumed that entry into the study population is independent of the disease status at entry.

Further, the censoring process, which determines the time when data on individuals are available, should not involve the parameters of interest, that is, the parameters in the conditional likelihood for the event process given the censoring process.[5]

This is for example the case if data arise from panel studies in which subjects are invited for testing at some fixed times. The assumption implies in this case that the occurrence of an event does not alter the probability of appearing for testing. If the event causes a dramatic increase in mortality, then this may not be fulfilled.

3.1. A piecewise constant intensity model

Consider data from a study of P people observed for the occurrence of some event, where for each person, p, we know the date of entry, t_{pe}, a latest date known without the event, t_{pw} (which is the

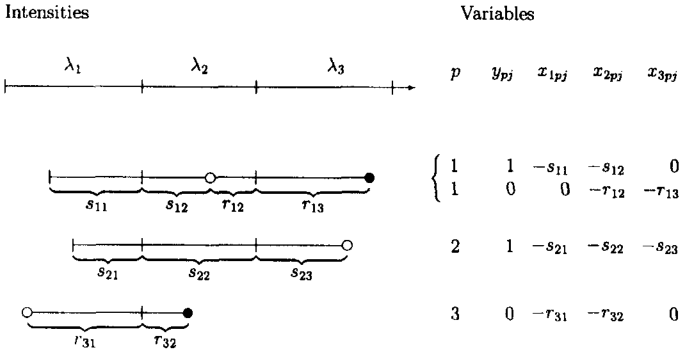
Intensities
Variables
λ_1
λ_2
λ_3
p y_{pj} x_{1pj} x_{2pj} x_{3pj}
s_{11} s_{12} r_{12} r_{13}
1 1 $-s_{11}$ $-s_{12}$ 0
1 0 0 $-r_{12}$ $-r_{13}$
s_{21} s_{22} s_{23}
2 1 $-s_{21}$ $-s_{22}$ $-s_{23}$
r_{31} r_{32}
3 0 $-r_{31}$ $-r_{32}$ 0

Figure 2. Construction of covariates x_{ipj} for the piecewise constant intensity model. Circles represent the last date seen before the 'event', dots the first date seen after

censoring date for those not known to experience the event), and for people for whom the event has occurred, an earliest date known to be after the event, t_{pd}.

Let the time axis, t, be divided into intervals I_1: $t_0 < t \leqslant t_1$, I_2: $t_1 < t \leqslant t_2, \ldots$, and put $\ell_i = t_i - t_{i-1}$. Assume that the intensity of events is constant within each I_i, equal to λ_i, say. The dates t_{pe}, t_{pw}, t_{pd} need not be in the set $\{t_0, t_1, \ldots\}$, or indeed in any other prespecified set of dates. In this setting one may only know that the event occurred between two dates, none of which need be endpoints of the intervals I_i.

Under this model the contribution to the likelihood from person p will be:

$$\Pr(\text{no event from } t_{pe} \text{ to } t_{pw}) \times \{1 - \Pr(\text{no event from } t_{pw} \text{ to } t_{pd} \,|\, \text{no event till } t_{pw})\}$$

$$= \exp\left(-\sum_i \lambda_i s_{pi}\right) \times \left\{1 - \exp\left(-\sum_i \lambda_i r_{pi}\right)\right\} \tag{1}$$

where s_{pi} is the time in I_i during which it is known that person p has not experienced an event (that is $I_i \cap (t_{pe}, t_{pw})$) and r_{pi} is the time in I_i during which an event may have occurred (that is $I_i \cap (t_{pw}, t_{pd})$), see Figure 2. In most cases s_{pi} and r_{pi} are either 0 or ℓ_i, and $s_{pi} + r_{pi} \leqslant \ell_i$.

The contribution to the likelihood from a single individual is thus equal to the likelihood from two independent Bernoulli trials, $y_{p1} = 1$ and $y_{p0} = 0$, with success probabilities $\exp(-\sum_i \lambda_i s_{pi})$ and $\exp(-\sum_i \lambda_i r_{pi})$, respectively. If either $t_{pe} = t_{pw}$ (the patient is only seen well at entry) or $t_{pd} = +\infty$ (the patient is last seen well), one of these contributions will be equal to one, and can thus be left out of consideration.

The full likelihood from the study of P people is the product of P terms of form (1), and equivalent to the likelihood from $2P$ independent Bernoulli trials, $y_{pj} = j$, $j = 0, 1$, with success probabilities (mean):

$$\mu_p = \Pr(y_{pj} = 1) = \exp\left(\sum_i \lambda_i x_{ipj}\right)$$

where

$$x_{ipj} = \begin{cases} -s_{pi} & j = 1 \\ -r_{pi} & j = 0. \end{cases}$$

The intensities, λ_i, can be estimated as parameters in a generalized linear model with logarithmic link function, by pretending that the y_{pj}'s are independent Bernoulli variates. This can be achieved fairly easily in many standard statistical programs.

3.2. Persons first seen diseased

Formally it is assumed that a latest date without the event of intere t is known for everyone, but this may not be the case in all applications. One could then arbitrarily assume that everyone is event-free at t_0, say. This could be age zero if the time scale is age, or some early date if the time scale is calendar time. The intensity estimated for the interval I_1 from t_0 to t_1 will then depend on the choice of t_0 through $\ell_1 = t_1 - t_0$. If t_1 is chosen earlier than the first observed date after an event ($t_1 \leqslant \min_p\{t_{pd}\}$), the estimate of $\lambda_1\ell_1$ and thus of the prevalence at t_1, $1 - \exp(-\lambda_1\ell_1)$, will be independent of the choice of t_0.

3.3. Estimating the survival function

The survival function, $S(t) = \exp\{-\int_0^t \lambda(s)\,\mathrm{d}s\}$, $t_{i-1} < t \leqslant t_i$ may be estimated by $\widehat{S(t)} = \exp\{-\sum_{j=1}^{i-1} \hat{\lambda}_j(t_j - t_{j-1}) - \hat{\lambda}_i(t - t_{i-1})\}$. Confidence intervals for the survival function at these points can be derived from this expression and an estimated variance–covariance matrix of $\hat{\lambda}_1, \ldots, \hat{\lambda}_n$.

Another possibility is to construct confidence intervals by finding the values of $S(t_j), j = 1, 2, \ldots$ for which minus twice the log-likelihood maximized over the remaining parameters exceeds the minimal value of $\chi^2(1)_{1-\alpha}$, for some suitably chosen α. However, this procedure requires that $S(t_i)$, or equivalently, $-\ln\{S(t_i)\} = \Lambda_i = \sum_{j=1}^{i} \lambda_j\ell_j$, is a parameter in the model. The Λ's are linear functions of the λ's; for $n = 3$ (leaving out the (p,j) subscripts on the x's):

$$\lambda_1 x_1 + \lambda_2 x_2 + \lambda_3 x_3 = \Lambda_1(x_1/\ell_1 - x_2/\ell_2) + \Lambda_2(x_2/\ell_2 - x_3/\ell_3) + \Lambda_3(x_3/\ell_3)$$

and in general

$$\sum_{i=1}^{n} \lambda_i x_i = \sum_{i=1}^{n-1} \Lambda_i(x_i/\ell_i - x_{i+1}/\ell_{i+1}) + \Lambda_n(x_n/\ell_n)$$

so the integrated intensity at the end of each interval can be estimated directly by using the covariates $x_i/\ell_i - x_{i+1}/\ell_{i+1}$, $i = 1, \ldots, n-1$ instead of $x_1, \ldots, x_{n-1}$, and x_n/ℓ_n instead of x_n.

Note that for panel studies where each individual has been at risk for an entire interval of constant intensity or not at all, the variates x_i/ℓ_i are either 0 or -1, so the reparameterization amounts to the inversion of a lower triangular matrix of -1s.

Using this reparametrization we can directly compute estimates of the survival function as $\widehat{S(t_i)} = \exp(-\hat{\Lambda}_i)$ as well as confidence intervals for the survival function, either as transformed intervals from the integrated intensity scale:

$$[\exp\{-\hat{\Lambda}_i - z_{1-\alpha/2}\widehat{\text{s.e.}}(\hat{\Lambda}_i)\}, \exp\{-\hat{\Lambda}_i + z_{1-\alpha/2}\widehat{\text{s.e.}}(\hat{\Lambda}_i)\}] \tag{2}$$

or by using the δ-method:

$$[\widehat{S(t_i)} - z_{1-\alpha/2}\widehat{\text{s.e.}}(\hat{\Lambda}_i)\widehat{S(t_i)}, \widehat{S(t_i)} + z_{1-\alpha/2}\widehat{\text{s.e.}}(\hat{\Lambda}_i)\widehat{S(t_i)}] \tag{3}$$

where $z_{1-\alpha/2}$ is the $1 - \alpha/2$ quantile of the standard normal distribution.

In drawing the survival function and the corresponding confidence limits, we may use the underlying assumption of constant intensity, and connect the points $(t_i, S(t_i))$ by exponential curves, that is the curve $(t, S(t))$, where:

$$S(t) = S(t_{i-1})\exp\left[\left(\frac{t - t_{i-1}}{t_i - t_{i-1}}\right)\ln\left(\frac{S(t_i)}{S(t_{i-1})}\right)\right], \quad t_{i-1} < t \leqslant t_i. \tag{4}$$

Under the piecewise constant intensity model this will be the maximum likelihood estimate of the survival function at the intermediate points, and for likelihood based confidence intervals it will result in likelihood based confidence intervals at the intermediate points as well. The latter are shown in Figure 2.

4. REGRESSION MODELS

Suppose that, for each individual, a set of covariates, $\mathbf{z} = (z_{p1}, \ldots, z_{pK})$ is recorded and that the influence of these on the intensity of events is of interest. In the framework of models with piecewise constant intensities a number of models will be considered.

First, the (additive) excess risk model, where the covariate effects are described by adding a linear term in the covariates to a baseline hazard:

$$\lambda_i(\mathbf{z}) = \lambda_i(\mathbf{0}) + \sum_k \beta_k z_k.$$

Second, the multiplicative relative risk model (proportional hazards model), where the covariate effects are described by multiplying the baseline hazard by a multiplicative term in the covariates:

$$\lambda_i(\mathbf{z}) = \lambda_i(\mathbf{0}) \times \exp\left(\sum_k \beta_k z_k\right).$$

Finally, the additive relative risk model, where the multiplicative term is replaced by an additive one:

$$\lambda_i(\mathbf{z}) = \lambda_i(\mathbf{0}) \times \left(1 + \sum_k \beta_k z_k\right).$$

In the following j is omitted from the subscripts, and p represents not the individuals, but the units in the generated dataset of 'Bernoulli' variables, i indexes the interval and thus the underlying intensities λ_i, and k indexes the covariates.

4.1. The excess risk model

The likelihood for the data under the (additive) excess risk model is constructed by replacing the terms λ_i in the likelihood for the simple case (1) above by the terms $\lambda_i + \sum_k \beta_k z_k$, which leads to a Bernoulli likelihood for independent observations with success probability (mean):

$$\mu_p(\mathbf{z}) = \exp\left\{\sum_i \lambda_i x_{ip} + \sum_k \beta_k\left(z_{kp}\sum_i x_{ip}\right)\right\}$$

so the inclusion of the covariates $z_1, \ldots, z_K$ in the model for the intensities means that the variables $z_k \sum_i x_i$ should be added to the regression model used for the Bernoulli data. The estimated coefficients associated with these variables in the regression analysis will then be the excess intensity associated with an increase of one in the covariates.

Thus the additive excess risk model for the intensities leads to a likelihood that may be maximized by a program that can perform maximum likelihood estimation in a Bernoulli model with logarithmic link function.

4.2. Relative risk models

4.2.1. *Multiplicative relative risk*

The likelihood for the data under the multiplicative relative risk model (Cox model, proportional hazards model) is constructed by replacing the terms λ_i in the likelihood for the simple case above by the terms $\lambda_i \exp(\sum_k \beta_k z_k)$, which leads to a Bernoulli likelihood for independent observations with success probability (mean):

$$\begin{aligned}\mu_p(\mathbf{z}) &= \exp\left\{\sum_i \lambda_i x_{ip} \exp\left(\sum_k \beta_k z_{kp}\right)\right\}\\ &= \exp\left\{-\exp\left(\ln\left[-\sum_i \lambda_i x_{ip}\right] + \sum_k \beta_k z_{kp}\right)\right\}. \qquad (5)\end{aligned}$$

For fixed β's this is a generalized linear model; the parameters are the λ's and the covariates $x_i \exp(\sum_k \beta_k z_k)$, the error distribution is Bernoulli and the link is logarithmic. For fixed λ's it is also a generalized linear model; the parameters are the β's, the covariates z_k and the error distribution Bernoulli, the link log–log and the offset $\ln(-\sum_i \lambda_i x_i)$.

This suggests the following fitting algorithm:

1. Fit a model as given in Section 2, to obtain initial estimates of the λ's.
2. Fix the λ's, and fit the model with covariates z_k, log–log-link and offset $\ln(-\sum_i \lambda_i x_i)$ to obtain estimates of the β's.
3. Fix the β's, form the covariates $x_i \exp(\sum_k \beta_k z_k)$, and fit a model with these covariates and log-link.
4. Repeat 2 and 3 until convergence.

If this algorithm converges to a point, it will be a stationary point of the likelihood function. An algorithm such as this, where one set of parameters is estimated while another set is held fixed, will yield variances of the parameter estimates in one of the sets (the β's or the λ's) that are estimates of the *conditional* variances given the value of the maximum likelihood estimates of the parameters in the other set, and as such underestimate the marginal variances, since conditional variances are never greater than marginal variances. Confidence intervals for parameters in models estimated this way should accordingly not be constructed using the approximate standard errors from the (conditional) log-likelihood, but, for example, by calculating the profile likelihood for the parameters and using the χ^2-approximation to minus twice the log-likelihood ratio.

4.2.2. *Additive relative risk*

The likelihood under this model is constructed by replacing the term $\exp(\sum_k \beta_k z_k)$ by $(1 + \sum_k \beta_k z_k)$ in (5) leading to a Bernoulli likelihood for independent observations with mean (success probability):

$$\begin{aligned}\mu_p(\mathbf{z}) &= \exp\left\{\sum_i \lambda_i x_{ip}\left(1 + \sum_k \beta_k z_{kp}\right)\right\}\\ &= \exp\left\{\sum_i \lambda_i x_{ip} + \sum_k \beta_k\left(z_{kp} \sum_i \lambda_i x_{ip}\right)\right\}. \qquad (6)\end{aligned}$$

For fixed β's this is a generalized linear model; the parameters are the λ's and the covariates $(1 + \sum_k \beta_k z_k)x_i$, the error distribution is Bernoulli and the link is logarithmic. For fixed λ's it is also a generalized linear model; the parameters are the β's, the covariates $z_k \sum_i \lambda_i x_i$, the error distribution Bernoulli, the link logarithmic and the offset $\sum_i \lambda_i x_i$.

The fitting algorithm above thus applies again, with obvious modifications and with the same reservations about variances of parameter estimates.

5. PRACTICALITIES

5.1. Parameter constraints

The parameters in the model must satisfy $\lambda_i > 0$ (at least in the model without covariates and in the regression models if $\mathbf{z} = 0$ is a covariate value in the data). For the additive excess risk models one must further assume that $\lambda_i + \sum_k \beta_k z_k > 0$, that is $\sum_k \beta_k z_k > -\min\{\lambda_i\}$ and for the additive relative risk models that $\sum_k \beta_k z_k > -1$ for all values of z in the data.

These constraints are fairly easily checked in actually fitted models; if they are not fulfilled, one must inspect the data to see if the assumptions made are unrealistic, for example, if the chosen number of intervals (number of λ's) is too big. In such cases the number of intervals must be reduced, for example, by combining adjacent intervals.

5.2. Data layout

For Bernoulli trials (with log-link function), an observation of $y_p = 1$ with covariate vector $(x_1, \ldots, x_a)$, say, in which all x_i are non-zero, has the same likelihood as a independent observations $y_p^{(1)}, \ldots, y_p^{(a)}$, all equal to 1, with covariate vectors $(x_1, 0, \ldots, 0)$, $(0, x_2, \ldots, 0)$, $\ldots$, $(0, \ldots, 0, x_a)$, respectively. Choosing this latter form of data layout avoids problems with negative estimates for the intensities if such covariate patterns can be generated corresponding to all λ's; a negative estimate yields a fitted value exceeding 1 for such units, and this will never be the case at a maximum for the likelihood.

This procedure does not, however, guarantee that the usual fitting algorithm converges, but it has been my experience that this coding of data actually removes some difficulties with convergence.

For the regression models these remarks only apply to the x's, not to the covariates (the z's); these should be coded in the same way for all the 'independent' Bernoulli variables generated.

Since we have generated a dataset of Bernoulli variables we may condense the dataset to one of binomial trials by collecting all units with the same covariate pattern, and thus simplify the analysis. This is most likely to give substantial reductions in the number of units for analysis if the number of intervals, and thereby the number of variables (x's), is fairly small.

5.3. Standard errors of parameter estimates

The estimation algorithms outlined above are based on the fact that the likelihood function for observation of P people is equal to that of a number of independent Bernoulli trials larger than P. However, data do *not* come from independent Bernoulli trials, so in software, where standard errors are based on the *expectation* of the second derivative of the log-likelihood function (the expected information), as for many statistical programs, notably GLIM, we get estimated standard errors calculated under a model different from the one we are actually analysing (albeit a model with the same likelihood function).

Table II. Deviance differences relative to model without regression terms for models for infection intensities among 297 Danish homosexual men: AER*, additive excess risk model with separate regression parameters for the period prior to the first examination; AER, additive excess risk model; ARR, additive relative risk model; MRR, Multiplicative relative risk model. The variables in the models are: BTH, year of birth; BT, $1\{BTH > 1950\}$; PTN, annual number of partners; PT, $1\{PTN > 11\}$; LPT, $\log_2(PTN)$; US, contact with person from the U.S.A.

Regression terms	Model type			
	AER*	AER	ARR	MRR
None	0·00	0·00	0·00	0·00
BT	0·23	0·05	0·63	0·63
BTH	0·70	0·00	1·44	1·48
PT	6·04	5·95	8·72	8·72
PTN	9·89	9·30	8·79	10·40
LPT	4·69	3·70	9·12	5·66
US	3·88	3·30	7·31	7·31
BT + PT	8·73	6·83	8·81	8·78
BT + PTN	12·67	9·86	9·46	10·72
BT + LPT	7·19	4·28	9·15	5·79
BT + US	4·14	3·39	7·64	6·56
BTH + PT	7·35	5·97	9·75	9·58
BTH + PTN	10.94	9.35	11.29	12.00
BTH + LPT	5·28	3·83	10·51	7·14
BTH + US	4·24	3·30	8·73	8·51
PT + US	9·25	7·06	11·74	12·76
PTN + US	**12·70**	**10·29**	**12·67**	**14·76**
LPT + US	7·62	5·79	11·78	11·50
BT + PT + US	11·77	7·99	11·78	12·76
BT + PTN + US	15·32	10·94	13·08	14·91
BT + LPT + US	9·71	6·47	11·83	11·52
BTH + PT + US	10·43	7·05	12·83	13·70
BTH + PTN + US	13·55	10·41	14·72	16·23
BTH + LPT + US	7·98	5·98	13·15	12·69

Standard errors of parameter estimates should therefore be based on the *observed* values of the second derivative of the log-likelihood function (Appendix I).

5.4. Starting values

With a link function (like the logarithm) which imposes constraints on the value of the linear predictor, one may encounter problems in defining a proper set of starting values for the linear predictor. For binomial data with reasonably large denominators one can often use the link function of the observed proportions, but in the case of Bernoulli data, the observed proportions will be either 0 or 1.

In practice this problem can often be overcome by fitting an ordinary logistic regression model with covariates $x_1, \ldots, x_n$, and then use the logarithm of the fitted values from this model as starting values for the linear predictor.

6. REGRESSION ANALYSIS OF HIV-INFECTION AMONG DANISH HOMOSEXUAL MEN

I have fitted a number of models to the dataset of HIV-infection in the cohort of Danish homosexuals for combinations of the variables year of birth, contact with people from the U.S.A. and number of partners per year.

Minus twice the log-likelihood ratio between the regression models and the model without covariate appears in Table II, where it is seen that the best fitting model is that with PTN (number of partners per year) and US (contacts with people from U.S.A.) for the relative risk models. The additive excess risk model renders US non-significant, apparently because the influence of US simply is not additive on the intensity scale (compare the fit of the models with US as the only covariate). Neither of the models are significantly improved by adding the year of birth, in either form. The estimates of the regression parameters from the models with PTN and US are:

(i) Additive excess risk model:

$$\mathrm{ER} = 0{\cdot}000718 \times \mathrm{PTN} + 0{\cdot}0123 \times \mathrm{US}\ (\mathrm{years}^{-1}).$$

Thus, the infection rate increases by 0·7 per cent per year for every ten partners. People with contacts in the U.S.A. have rates which are 1.2 per cent per year higher.

(ii) Multiplicative relative risk:

$$\mathrm{RR} = \exp(0{\cdot}0106\,\mathrm{PTN} + 0{\cdot}507\,\mathrm{US}).$$

In this model an increase of 10 partners give an increase in relative risk of $100\,(e^{0{\cdot}106} - 1)$ per cent = 11 per cent, whereas those with contacts in the U.S.A. have a relative risk of $e^{0{\cdot}507} = 1{\cdot}66$ compared with those who have none.

(iii) Additive relative risk:

$$\mathrm{RR} = 1 + 0{\cdot}0335\,\mathrm{PTN} + 0{\cdot}995\,\mathrm{US}.$$

In this model an increase of 10 partners increases the relative risk by 0·34, whereas people with contacts in the U.S.A. have a relative risk which is 1·00 higher than those without.

Figure 3 shows the estimated cumulative probability of being infected for men with 10 annual partners, with and without contacts in the U.S.A. respectively, for each of the three models. The additive excess risk model shows a smaller difference between those with and those without U.S.A. contacts, which may be attributed to this effect not being additive on the intensity scale.

The addition of the dichotomous variable US to the additive excess risk model gives a deviance reduction of only 3·30, a much poorer fit than the corresponding relative risk model that produce a deviance reduction of 7·3 (Table II). Note that the two relative risk models are identical if only a single categorical covariate is included.

Adding PTN to the relative risk model with US produces substantial decreases in deviance; 5·4 and 7·5 for the additive and multiplicative relative risk models, respectively. Thus the two relative risk models provide much the same fit to the data. The rather small difference between the deviance reductions suggests that it is difficult to discriminate between the two in this dataset. From Figure 3 it is also apparent that the fit of the two models is quite similar.

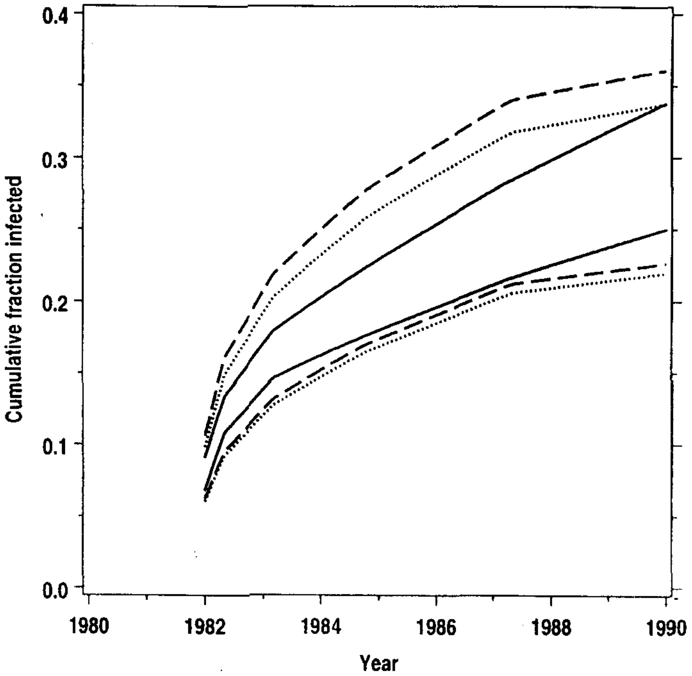

Figure 3. Estimated cumulative fraction infected estimated by three different regression models (additive ER —; additive RR – – –; multiplicative RR · · ·) for 10 partners per year (PTN = 10) and for contact with people from the U.S.A.: none (US = 0, lower curves) and some (US = 1, upper curves)

7. DISCUSSION

The models presented here represent a slightly more limited approach than Becker and Melbye by assuming a model of piecewise constant intensities, where the intervals of constant intensities need not coincide with any intervals formed by the observations. Becker and Melbye do not assume any parametric form for the hazard, they simply estimate the survival probabilities (or equivalently, the integrated hazard) at time points where there are observations. Thus, they assume that the number of time points is small compared with the number of events. If the number of time points is large the approach proposed by Becker and Melbye would not be useful in the construction of standard errors of parameter estimates because of the large number of parameters estimated.

The assumption of piecewise constant intensities allows one to keep the number of estimated parameters in reasonable balance with the available information from the data, without sacrificing precision, and still enables use of standard statistical software.

Finkelstein[6] developed a proportional hazards model for interval censored data, and proposed to fit it by Newton–Raphson iteration. She used the $\ln(-\ln[S(t_i)])$, that is the log of the integrated hazard as parameters. Finkelstein's approach is fully parametric too, and also has a potentially large number of parameters, except for panel data. She proposed to resolve the problems that this may cause for standard errors of parameter estimates by a suitable grouping of data prior to analysis. This is in reality the same as assuming piecewise constant intensities, because the parameters used will be the integrated intensity at the grouping points. Thus the number of intervals (number of parameters) that we can entertain depends crucially on the available amount of data. If a lot of events are recorded throughout the time scale, it is of course possible to use a large number of intervals each with a parameter to describe the hazard, whereas a scarcity of data could ultimately force us to assume a constant hazard. Empirical studies suggest that this has surprisingly little influence on the estimates of the regression parameters and their estimated standard errors.[7]

In a piecewise constant intensity model, Finkelstein's parameterization corresponds to using the log of the hazards as parameters. As hazards necessarily are positive, this parameterization seems more reasonable to use for calculation of approximate confidence intervals. In Appendix I are shown the second derivatives of the log-likelihood with respect to the hazards and the regression parameters. If the log-hazard were used as a parameter instead, one would get the second derivatives with respect to this parameter simply by multiplying the expressions for the derivatives with respect to $\lambda_a \lambda_b$ by $\lambda_a \lambda_b$ and the expressions for the derivatives with respect to $\lambda_a \beta_b$ by λ_a.

The advantage of the approach developed here is that standard software may be used for estimation, and that a variety of different models may be accommodated in the framework. The price to pay is the assumption of piecewise constant intensities, but this is not a serous limitation if there are sufficient data available. On the other hand, if the data are scarce then likelihood based methods may not be the best choice anyway. The drawback is that the standard errors that are produced by standard programs are not reliable, partly because they may be based on the expected information (expectation taken under a wrong model), and partly because for the relative risk models they are conditional standard errors. Thus it is necessary to resort to explicit evaluation of the observed information (Appendix I), or to construction of confidence intervals by profile likelihood. The latter is recommendable in any case for the additive excess risk models where the parameter estimates cannot be expected to be symmetrically distributed.

APPENDIX I

This appendix contains the expressions for the second derivatives of the log-likelihood functions for the three models. These are necessary for evaluation of the observed information needed to give the approximate covariance matrix of the parameter estimates.

The additive excess risk model has second derivatives of the log-likelihood function which are

$$\frac{\partial^2 l}{\partial \lambda_a \partial \lambda_b} = \sum x_a x_b \frac{\mu}{(1-\mu)^2}(y-1)$$

$$\frac{\partial^2 l}{\partial \lambda_a \partial \beta_b} = \sum x_a \left(z_b \sum_i x_i \right) \frac{\mu}{(1-\mu)^2}(y-1)$$

$$\frac{\partial^2 l}{\partial \beta_a \partial \beta_b} = \sum \left(z_a \sum_i x_i \right) \left(z_b \sum_i x_i \right) \frac{\mu}{(1-\mu)^2}(y-1)$$

where $\mu = \exp\{\sum_i \lambda_i x_i + \sum_k \beta_k (z_k \sum_i x_i)\}$. In the simple model without regression parameters the first of these expressions with $\mu = \exp(\sum_i \lambda_i x_i)$ is the second derivative of the log-likelihood function.

The multiplicative relative risk model has second derivatives of the log-likelihood function which are

$$\frac{\partial^2 l}{\partial\lambda_a\partial\lambda_b} = \sum x_a x_b \exp\left(2\sum_k \beta_k z_k\right)\frac{\mu}{(1-\mu)^2}(y-1)$$

$$\frac{\partial^2 l}{\partial\lambda_a\partial\beta_b} = \sum x_a z_b \exp\left(\sum_k \beta_k z_k\right)\ln(\mu)\,\frac{\mu}{(1-\mu)^2}(y-1)$$

$$\frac{\partial^2 l}{\partial\beta_a\partial\beta_b} = \sum z_a z_b \ln(\mu)^2\,\frac{\mu}{(1-\mu)^2}\left\{1+\frac{1-\mu}{\mu\ln(\mu)}\right\}\times\left\{y-\frac{\mu(1-\mu)+\mu\ln(\mu)}{1-\mu+\mu\ln(\mu)}\right\}$$

where $\mu = \exp\{\sum_i \lambda_i x_i \exp(\sum_k \beta_k z_k)\}$.

The additive relative risk model has second derivatives of the log-likelihood function which are

$$\frac{\partial^2 l}{\partial\lambda_a\partial\lambda_b} = \sum x_a x_b\left(\sum_k \beta_k z_k\right)^2\frac{\mu}{(1-\mu)^2}(y-1)$$

$$\frac{\partial^2 l}{\partial\lambda_a\partial\beta_b} = \sum x_a z_b\left(\sum_i \lambda_i x_i\right)\left(\sum_k \beta_k z_k\right)\frac{\mu}{(1-\mu)^2}(y-1)$$

$$\frac{\partial^2 l}{\partial\beta_a\partial\beta_b} = \sum z_a z_b\left(\sum_k \beta_k z_k\right)^2\frac{\mu}{(1-\mu)^2}(y-1)$$

where $\mu = \exp\{\sum_i \lambda_i x_i (1 + \sum_k \beta_k z_k)\}$.

ACKNOWLEDGEMENTS

I am grateful to Niels Becker for constructive comments on an earlier version of this paper, and to Adam Gottschau for pointing out to me that the expected information is not usable as a basis for variance calculations. Finally, I thank Mads Melbye for supplying data from the Danish cohort of homosexual men.

REFERENCES

1. Peto, R. 'Experimental survival curves for interval-censored data', *Applied Statistics*, **22**, 86–91 (1973).
2. Turnbull, B. W. 'The empirical distribution function with arbitrarily grouped censored and truncated data', *Journal of the Royal Statistical Society, Series B*, **38**, 290–295 (1976).
3. Becker, N. G. and Melbye, M. 'Use of a log-linear model to compute the empirical survival curve from interval-censored data, with application to data on tests for HIV-positivity', *Australian Journal of Statistics*, **33**, 125–133 (1990).
4. Melbye, M., Biggar, R. J., Ebbesen, P., Sarngadharan, M. G., Weiss, S. H., Gallo, R. C. and Blattner, W. A. 'Seroepidemilogy of HTLV-III antibody in Danish homosexual men: prevalence, transmission and disease outcome', *British Medical Journal*, **289**, 573–575 (1984).
5. Grüger, J., Kay, R. and Schuhmacher, M. 'The validity of inferences based on incomplete observations in disease state models', *Biometrics*, **47**, 595–605 (1991).
6. Finkelstein, D. M. 'A proportional hazards model for interval censored failure time data', *Biometrics*, **42**, 845–854 (1986).
7. Selmer, R. 'A comparison of Poisson regression models fitted to multiway summary tables and Cox's survival model using data from a blood pressure screening in the city of Bergen, Norway', *Statistics in Medicine*, **9**, 1157–1166 (1990).