Statistical Methods in Medical Research 2000; **9**: 161–179

Random-effects regression analysis of correlated grouped-time survival data

Donald Hedeker Division of Epidemiology and Biostatistics, Health Policy Research Center, School of Public Health, University of Illinois at Chicago, Chicago, Illinois, USA,
Ohidul Siddiqui Food and Drug Administration, United States Government, USA and
Frank B Hu Department of Nutrition, Harvard School of Public Health, USA

Random-effects regression modelling is proposed for analysis of correlated grouped-time survival data. Two analysis approaches are considered. The first treats survival time as an ordinal outcome, which is either right-censored or not. The second approach treats survival time as a set of dichotomous indicators of whether the event occurred for time periods up to the period of the event or censor. For either approach both proportional hazards and proportional odds versions of the random-effects model are developed, while partial proportional hazards and odds generalizations are described for the latter approach. For estimation, a full-information maximum marginal likelihood solution is implemented using numerical quadrature to integrate over the distribution of multiple random effects. The quadrature solution allows some flexibility in the choice of distributions for the random effects; both normal and rectangular distributions are considered in this article. An analysis of a dataset where students are clustered within schools is used to illustrate features of random-effects analysis of clustered grouped-time survival data.

1 Introduction

Models for grouped-time survival data are useful for analysis of failure-time data when subjects are measured repeatedly at fixed intervals in terms of the occurrence of some event, or when determination of the exact time of the event is only known within grouped intervals of time. For example, in school-based prevention studies students are typically measured annually regarding their smoking, alcohol, and other substance use during the past year. An important question is then to determine the degree to which an intervention prevents or delays substance use initiation. These studies often utilize a cluster randomization scheme so that the schools, rather than the individual students, are randomized to intervention conditions. In analysis of such grouped-time initiation (or survival) data, use of grouped-time regression models that assume independence of observations[1–3] is therefore problematic because of this clustering of students within schools. More generally, this same issue arises for other types of cluster randomization trials in which subjects are observed nested within various types of clusters (e.g. hospitals, firms, clinics, counties), and thus cannot be assumed to be independent. To account for the data clustering, random-effects models (also called multilevel, hierarchical linear, or mixed models) provide a useful approach for simultaneously estimating the parameters of the regression model and the variance components that account for the data clustering.[4–7]

Address for correspondence: D Hedeker, Division of Epidemiology and Biostatistics (M/C 922), School of Public Health, University of Illinois at Chicago, 2121 West Taylor Street, Room 525, Chicago, IL 60612-7260, USA. E-mail: hedeker@uic.edu

 0962-2802(00)SM209RA

For continuous-time survival data that are clustered, several authors[8–14] have developed mixed-effects survival models. These models are often termed frailty models or survival models including heterogeneity, and recent review articles describe many of these models.[15,16] An alternative approach for dealing with correlated data uses the generalized estimating equations (GEE) method described by Liang and Zeger[17] to estimate model parameters. In this regard, Lee *et al*.[18] and Wei *et al*.[19] have developed continuous-time survival models.

Application of these continuous-time models to grouped or discrete-time survival data is generally not recommended because of the large number of ties that result. Instead, models specifically developed for grouped or discrete-time survival data have been proposed. Both Han and Hausman[20] and Scheike and Jensen[21] have described proportional hazards models incorporating a log-gamma distribution specification of heterogeneity. Also, Ten Have[22] developed a discrete-time proportional hazards survival model incorporating a log-gamma random effects distribution, additionally allowing for ordinal survival and failure categories. Ten Have and Uttal[23] used Gibbs sampling to fit continuation ratio logit models with multiple normally distributed random effects. In terms of a GEE approach, Guo and Lin[24] have developed a multivariate model for grouped-time survival data.

Several authors have noted the relationship between ordinal regression models (using complementary log–log and logit link functions) and survival analysis models for grouped and discrete time.[20,25,26] Similarly, others[3,27,28] have described how dichotomous regression models can be used to model grouped and discrete time survival data. The ordinal approach simply treats survival time as an ordered outcome that is either right-censored or not. Alternatively, in the dichotomous approach each survival time is represented as a set of indicators of whether or not an individual failed in each time unit (until a person either experiences the event or is censored). As a result, the dichotomous approach is more useful for inclusion of time-dependent covariates and relaxing of the proportional hazards assumption.

In this paper, we will generalize these fixed-effects regression models for categorical responses by including random effects to account for the data clustering. The resulting models are equivalent to dichotomous and ordinal random-effects regression models,[29] albeit with the extension of the ordinal model to allow for right-censoring of the response. These models allow multiple random effects and a general form for model covariates. Assuming either a proportional or partial proportional hazards or odds model, a maximum marginal likelihood (MML) solution is described using multi-dimensional quadrature to numerically integrate over the distribution of random-effects.

The current article is most closely related to the work of Han and Hausman,[20] Ten Have,[22] and Scheike and Jensen,[21] however, there are important differences. One difference is that these authors used a log-gamma distribution for a single random effect. This specification of the (univariate) random effects distribution leads to a closed-form solution, whereas we use quadrature to numerically integrate the (multivariate) random effect distribution. Although the closed-form solution is mathematically appealing, the quadrature solution does allow us to consider multiple random effects as well as various distributional forms for the random effects, including normally-distributed random effects. Pickles and Crouchley[15] and Preisler[30] also

proposed use of quadrature to estimate survival models with normally distributed random effects, though their models were not as general as the models considered here. Normally-distributed random effects are common in many other types of random-effects models, and the case can be made for normally-distributed random effects as a more natural choice (see discussion and commentary of Lee and Nelder[31]). In particular, as noted by Longford (in the discussion of Lee and Nelder[31]) for models with multiple random effects 'the normal is the only well-established multivariate distribution with a full range of correlation structures'. Also, as noted by Preisler[30] because the random and fixed effects are on the same scale, interpretation of parameters is more straightforward.

Another important distinction is that these authors dealt with relatively few clustered observations (i.e. 2 or 3). In particular, Ten Have[22] noted that estimation was prohibitive in terms of time when the number of clustered observations gets large (i.e. say 10 or more). For cluster randomization trials, this can be a severe limitation. Alternatively, as illustrated by the example, our approach can accommodate many observations per cluster and varying numbers of observations per cluster. Finally, our model can be generalized to accommodate multiple levels of nesting of the random effects (e.g. for repeated observations within subjects within schools). For such nested random effects, Gibbons and Hedeker[32] describe an approach for three-level dichotomous outcomes that can be used to extend the models described here.

The article is organized as follows: Section 2 presents the random-effects grouped-time model, including an extension to allow for nonproportional hazards or odds; Section 3 describes the full-information maximum likelihood estimation procedure; Section 4 illustrates use of the model for clustered data; Section 5 contains some closing remarks.

2 Random-effects grouped-time survival analysis model

Using the terminology of multilevel analysis,[6] let i denote the level-2 units ($i = 1, \ldots, N$) and let j denote the level-1 units ($j = 1, \ldots, n_i$). If subjects are nested within clusters, the subjects and clusters represent the level-1 and level-2 units, respectively. Alternatively, if there are multiple failure times per subject, then the level-2 units are the subjects and the level-1 units are the repeated failure times. Suppose that there is a continuous random variable for the uncensored time of event occurrence (which may not be observed), however assume that time (of assessment) can take on only discrete positive values $t = 1, 2, \ldots, m$. For each level-1 unit, observation continues until time t_{ij} at which point either an event occurs or the observation is censored, where censoring indicates being observed at t_{ij} but not at $t_{ij} + 1$. Define P_{ijt} to be the probability of failure, up to and including time interval t, that is

$$P_{ijt} = \Pr\,[t_{ij} \leq t] \tag{1}$$

and so the probability of survival beyond time interval t is simply $1 - P_{ijt}$.

Because $1 - P_{ijt}$ represents the survivor function, McCullagh[25] proposed the following grouped-time version of the continuous-time proportional hazards model:

$$\log\left[-\log\left(1-P_{ijt}\right)\right]=\alpha_{0t}+\mathbf{x}_{ij}'\boldsymbol{\beta} \tag{2}$$

This is the so-called complementary log–log function, which can be re-expressed in terms of the cumulative failure probability, $P_{ijt}=1-\exp\left(-\exp\left(\alpha_{0t}+\mathbf{x}_{ij}'\boldsymbol{\beta}\right)\right)$. In this model, $\mathbf{x}_{ij}$ is a $p\times 1$ vector including covariates that vary either at level 1 or 2, however they do not vary with time (i.e. they do not vary across the ordered response categories). They may, however, represent the average of a variable across time or the value of the covariate at the time of the event.

Since the integrated hazard function equals $-\log\left(1-P_{ijt}\right)$, this model represents the covariate effects ($\boldsymbol{\beta}$) on the log of the integrated hazard function. The covariate effects are identical to those in the grouped-time version of the proportional hazards model described by Prentice and Gloeckler.[2] As such, the $\boldsymbol{\beta}$ coefficients are also identical to the coefficients in the underlying continuous-time proportional hazards model. Furthermore, as noted by Allison,[3] the regression coefficients of the model are invariant to interval length. Augmenting the coefficients $\boldsymbol{\beta}$, the intercept terms α_{0t} are a set of m constants that represent the logarithm of the integrated baseline hazard (i.e. when $\mathbf{x}=\mathbf{0}$). As such, these terms represent cutpoints on the integrated baseline hazard function; these parameters are often referred to as threshold parameters in descriptions of ordinal regression models. While the above model is the same as that described in McCullagh,[25] it is written so that the covariate effects are of the same sign as the Cox proportional hazards model. A positive coefficient for a regressor then reflects increasing hazard with greater values of the regressor.

Adding random effects to this model, we get

$$\log\left[-\log\left(1-P_{ijt}\right)\right]=\alpha_{0t}+\mathbf{x}_{ij}'\boldsymbol{\beta}\ +\mathbf{w}_{ij}'\boldsymbol{\upsilon}_i \tag{3}$$

or

$$P_{ijt}=1-\exp\left(-\exp\left(\alpha_{0t}+\mathbf{x}_{ij}'\boldsymbol{\beta}\ +\mathbf{w}_{ij}'\boldsymbol{\upsilon}_i\right)\right)=1-\exp\left(-\exp z_{ijt}\right) \tag{4}$$

where $\boldsymbol{\upsilon}_i$ is the $r\times 1$ vector of unknown random effects for the level-2 unit i, and $\mathbf{w}_{ij}$ is the design vector for the r random effects. The distribution of the r random effects $\boldsymbol{\upsilon}_i$ is assumed to be multivariate with mean vector $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}_\upsilon$. An important special case is when the distribution is assumed to be multivariate normal. For convenience, the random effects are often expressed in standardized form. Specifically, let $\boldsymbol{\upsilon}=\mathbf{S}\boldsymbol{\theta}$, where $\mathbf{SS}'=\boldsymbol{\Sigma}_\upsilon$ is the Cholesky decomposition of $\boldsymbol{\Sigma}_\upsilon$. The model for z_{ijt} then is written as:

$$z_{ijt}=\alpha_{0t}+\mathbf{x}_{ij}'\boldsymbol{\beta}\ +\mathbf{w}_{ij}'\mathbf{S}\boldsymbol{\theta}_i \tag{5}$$

As a result of the transformation, the Cholesky factor $\mathbf{S}$ is estimated instead of the covariance matrix $\boldsymbol{\Sigma}_\upsilon$. As the Cholesky factor is essentially the square-root of the covariance matrix, this allows more stable estimation of near-zero variance terms.

The model given in (5) can accommodate separate random-effect variance terms for groups of either level-1 or level-2 units. For example, suppose that treatment group is a level-2 variable (i.e. at the cluster level) and there is interest in allowing varying random-effect variance terms by groups. If there are two treatment groups, $\mathbf{w}_{ij}$ is then

specified as a 2×1 vector of dummy codes indicating membership in groups 1 and 2, respectively, and $\mathbf{S}$ is a 2×1 vector of independent random-effect standard deviations for these two groups. In this case, θ_i is a scalar that is pre-multiplied by the vector $\mathbf{S}$. Similarly, in a longitudinal context, educational testing models (see Bock[33]) allow separate random-effect variance terms by items, which represent groupings of level-1 units (i.e. item responses nested within subjects). In either case, the model is specified with $\mathbf{S}$ as a $r \times 1$ vector that is pre-multiplied by the transpose of a $r \times 1$ vector of indicator variables $\mathbf{w}_{ij}$, and so $\mathbf{S}$ pre-multiplies a scalar random effect θ_i instead of a $r \times 1$ vector of random effects $\boldsymbol{\theta}_i$.

2.1 Proportional odds model

As applied to survival data, the proportional odds model is described by Bennett.[34] For grouped-time, the random-effects proportional odds model is written in terms of the logit link function as

$$\log [P_{ijt}/(1 - P_{ijt})] = z_{ijt} \tag{6}$$

or alternatively as $P_{ijt} = 1/[1 + \exp(-z_{ijt})]$. The choice of which link function to use is not always clear-cut. Bennett[34] noted that the proportional odds model is useful for survival data when the hazards of groups of subjects are thought to converge with time. This contrasts to the proportional hazards model where the hazard rates for separate groups of subjects are assumed proportional at all timepoints. However, this type of nonproportional hazards effect can often be accommodated in the complementary log-log link model by including interactions of covariates with the baseline hazard cutpoints.[35] Also as Doksum and Gasko[36] note, large amounts of high quality data are often necessary for link function selection to be relevant. Since these two link functions often provide similar fits, Ten Have[22] suggests that the choice of which to use depends on whether inference should be in terms of odds ratios or discrete hazard ratios. Similarly, McCullagh[25] notes that link function choice should be based primarily on ease of interpretation.

2.2 Pooling of repeated observations and nonproportional hazards

Thus far, survival time has been represented as an ordered outcome t_{ij} that is designated as censored or not. An alternative approach for grouped-time survival data, described by Allison[3], D'Agostino *et al.*,[27] Singer and Willett[28] and others, treats each individual's survival time as a set of dichotomous observations indicating whether or not an individual failed in each time unit until a person either experiences the event or is censored. Thus, each survival time is represented as a $t_{ij} \times 1$ vector of zeros for censored individuals, while for individuals experiencing the event the last element of this $t_{ij} \times 1$ vector of zeros is changed to a one. These multiple person-time indicators are then treated as distinct observations in a dichotomous regression model. In the case of clustered data, a random-effects dichotomous regression model is used. This method has been called the pooling of repeated observations method by Cupples.[37] It is particularly useful for handling time-dependent covariates and fitting nonproportional hazards models because the covariate values can change across each individuals' t_{ij} timepoints.

For this approach, define p_{ijt} to be the probability of failure in time interval t, conditional on survival prior to t:

$$p_{ijt} = \Pr\ [t_{ij} = t \mid t_{ij} \geq t] \tag{7}$$

Similarly, $1 - p_{ijt}$ is the probability of survival beyond time interval t, conditional on survival prior to t. The proportional hazards model is then written as

$$\log\ [-\log\ (1 - p_{ijt})] = \alpha_{0t} + \mathbf{x}'_{ijt}\boldsymbol{\beta} + \mathbf{w}'_{ij}\mathbf{S}\boldsymbol{\theta}_i \tag{8}$$

and the corresponding proportional odds model is

$$\log\ [p_{ijt}/(1 - p_{ijt})] = \alpha_{0t} + \mathbf{x}'_{ijt}\boldsymbol{\beta} + \mathbf{w}'_{ij}\mathbf{S}\boldsymbol{\theta}_i \tag{9}$$

where now the covariates $\mathbf{x}$ can vary across time and so are denoted as $\mathbf{x}_{ijt}$. Augmenting the model intercept α_{01}, the remaining intercept terms α_{0t} $(t = 2, \ldots m)$ are obtained by including as regressors $m - 1$ dummy codes representing deviations from the first timepoint. Because the covariate vector $\mathbf{x}$ now varies with t, this approach automatically allows for time-dependent covariates, and relaxing the proportional hazards assumption only involves including interactions of covariates with the $m - 1$ timepoint dummy codes.

Under the complementary log–log link function, the two approaches characterized by (3) and (8) yield identical results for the parameters that do not depend on t, namely the regression coefficients of time-independent covariates and the Cholesky factor.[38,39] For the logit link, similar, but not identical, results are obtained for these parameters. Comparing these two approaches, notice that for the ordinal approach each observation consists of only two pieces of data: the (ordinal) time of the event and whether it was censored or not. Alternatively, in the dichotomous approach each survival time is represented as a vector of dichotomous indicators, where the size of the vector depends on the timing of the event or censoring. Thus, the ordinal approach can be easier to implement and offers savings in terms of the dataset size, especially as the number of timepoints gets large, while the dichotomous approach is superior in its treatment of time-dependent covariates and relaxing of the proportional hazards or odds assumption.

Relaxing the proportional hazards or odds assumption in the ordinal model is possible; for fixed-effects models this has been discussed by Terza,[40] Peterson and Harrell,[41] and Cox.[42] Similarly, for clustered ordinal data, Hedeker and Mermelstein[43] have developed and described a random-effects partial proportional odds model. For this, the model can be rewritten as:

$$z_{ijt} = \alpha_{0t} + (\mathbf{u}^*_{ij})'\boldsymbol{\alpha}^*_t + \mathbf{x}'_{ij}\boldsymbol{\beta} + \mathbf{w}'_{ij}\mathbf{S}\boldsymbol{\theta}_i \tag{10}$$

or absorbing α_{0t} and $\boldsymbol{\alpha}^*_t$ into $\boldsymbol{\alpha}_t$

$$z_{ijt} = \mathbf{u}'_{ij}\boldsymbol{\alpha}_t + \mathbf{x}'_{ij}\boldsymbol{\beta} + \mathbf{w}'_{ij}\mathbf{S}\boldsymbol{\theta}_i \tag{11}$$

where, $\mathbf{u}_{ij}$ is a $(l + 1) \times 1$ vector containing (in addition to a 1 for α_{0t}) the values of

observation ij on the set of l covariates for which interactions with the cutpoints of the integrated baseline hazard are desired. Here, $\boldsymbol{\alpha}_t$ is a $(l+1) \times 1$ vector of regression coefficients associated with the l variables (and the intercept) in $\mathbf{u}_{ij}$.

Note that because the dichotomous and ordinal approaches only yield identical results under the proportional hazards model (i.e. the complementary log–log link and covariates with effects that do not vary across time), differences emerge for covariates allowed to have varying effects across time under these two approaches. For covariates of this type, the dichotomous approach is generally preferred because it models the covariate's influence in terms of the conditional probability of failure given prior survival (i.e. the hazard function), rather than the cumulative probability of failure (i.e. the integrated or cumulative hazard function) as in the ordinal model.

3 Maximum marginal likelihood estimation

For the dichotomous approach, the maximum likelihood solution as described in Hedeker and Gibbons[29] and implemented in the MIXOR software program[44] can be used without modification. For the ordinal treatment of survival times, the solution must be extended to accommodate right-censoring of the ordinal outcome. For this, let $\delta_{ij} = 0$ if level-1 unit ij is a censored observation and equal to 1 if the event occurs (fails). Thus, t_{ij} denotes the value of time ($t = 1, \ldots, m$) when either the ijth unit failed or was censored. It is assumed that the censoring and failure mechanisms are independent. With the above mixed-effects regression model, the probability of a failure at time t for a given level-2 unit i, conditional on $\boldsymbol{\theta}$ (and given $\boldsymbol{\alpha}_t, \boldsymbol{\beta}$, and $\mathbf{S}$) is:

$$Pr(t_j = t \cap \delta_j = 1 \mid \boldsymbol{\theta}; \boldsymbol{\alpha}_t, \boldsymbol{\beta}, \mathbf{S}) = P_{jt} - P_{j,t-1} \tag{12}$$

where $P_{j0} = 0$ and $P_{j,m+1} = 1$. The corresponding probability of being right censored at time t equals the cumulative probability of not failing at that time, $1 - P_{jt}$.

Let $\mathbf{t}_i$ denote the vector pattern of failure times from level-2 unit i for the n_i level-1 units nested within. Similarly, let $\boldsymbol{\delta}_i$ denote the vector pattern of event indicators. The joint probability of patterns $\mathbf{t}_i$ and $\boldsymbol{\delta}_i$, given $\boldsymbol{\theta}$, assuming independent censoring is equal to the product of the probabilities of the level-1 responses:

$$\ell(\mathbf{t}_i, \boldsymbol{\delta}_i \mid \boldsymbol{\theta}; \boldsymbol{\alpha}_t, \boldsymbol{\beta}, \mathbf{S}) = \prod_{j=1}^{n_i} \prod_{t=1}^{m} \left[(P_{ijt} - P_{ij,t-1})^{\delta_{ij}} (1 - P_{ijt})^{1-\delta_{ij}} \right]^{d_{ijt}} \tag{13}$$

where $d_{ijt} = 1$ if $t_{ij} = t$ (and $= 0$ if $t_{ij} \neq t$).

The marginal density of $\mathbf{t}_i$ and $\boldsymbol{\delta}_i$ in the population is expressed as the following integral of the conditional likelihood, $\ell(\cdot)$, weighted by the prior density $g(\cdot)$:

$$h(\mathbf{t}_i, \boldsymbol{\delta}_i) = \int_{\boldsymbol{\theta}} \ell(\mathbf{t}_i, \boldsymbol{\delta}_i \mid \boldsymbol{\theta}; \boldsymbol{\alpha}_t, \boldsymbol{\beta}, \mathbf{S})\, g(\boldsymbol{\theta})\, d\boldsymbol{\theta} \tag{14}$$

where $g(\boldsymbol{\theta})$ represents the multivariate distribution of the standardized random effects vector $\boldsymbol{\theta}$ in the population. The marginal log-likelihood for the patterns from the N level-2 units is then written as $\log L = \sum_i^N \log h(\mathbf{t}_i, \boldsymbol{\delta}_i)$. Maximizing this likelihood

then provides MML estimates. The derivation for all model parameters is provided in the appendix, as is a discussion of computer implementation of the estimation procedure.

3.1 Numerical quadrature

Many authors have assumed a log-gamma distribution for the random effects to obtain a closed-form solution. Alternatively, as mentioned, the case can be made for normally-distributed random effects. Here, we use numerical integration to integrate over the distribution of the random effects $\boldsymbol{\theta}$ to allow estimation of a model with normally-distributed random effects. The integration is approximated by a summation on a specified number of quadrature points Q for each dimension of the integration; thus, for the transformed $\boldsymbol{\theta}$ space, the summation goes over Q^r points.

For the normal density, optimal points and weights for the Gauss–Hermite quadrature are given in Stroud and Sechrest.[45] If another distribution is assumed, other points and density weights may be used. For example, if a rectangular or uniform distribution is assumed, then Q points are set at equal intervals over an appropriate range (for each dimension) and the quadrature weights equal to $1/Q$. Other distributions are possible: Bock and Aitkin[33] discuss the possibility of empirically estimating the random-effect distribution. In the example below, results are compared assuming a normal and a uniform distribution, thus providing some information about the sensitivity of the results to the assumed normal distribution.

3.2 Estimation of random effects and marginal probabilities

In some cases, it may be of interest to estimate values of the random effects $\boldsymbol{\theta}_i$ within the sample. A reasonable choice for this is the expected *a posteriori* (EAP) or empirical Bayes estimator $\bar{\boldsymbol{\theta}}_i$ (see Bock and Aitkin[33]). For the univariate case, this estimator $\bar{\theta}_i$, given the vector of survival times $\mathbf{t}_i$ censor indicators $\boldsymbol{\delta}_i$, and covariate matrices $\mathbf{X}_i$ and $\mathbf{U}_i$, is given by:

$$\bar{\theta}_i = E(\theta_i \mid \mathbf{t}_i, \boldsymbol{\delta}_i, \mathbf{X}_i, \mathbf{U}_i) = \frac{1}{h(\mathbf{t}_i, \boldsymbol{\delta}_i)} \int_\theta \theta_i \, \ell(\mathbf{t}_i, \boldsymbol{\delta}_i \mid \theta; \boldsymbol{\alpha}_t, \boldsymbol{\beta}, \sigma) \, g(\theta) \, d\theta \tag{15}$$

The variance of this estimator is obtained similarly as:

$$V(\bar{\theta}_i \mid \mathbf{t}_i, \boldsymbol{\delta}_i, \mathbf{X}_i, \mathbf{U}_i) = \frac{1}{h(\mathbf{t}_i, \boldsymbol{\delta}_i)} \int_\theta (\theta_i - \bar{\theta}_i)^2 \, \ell(\mathbf{t}_i, \boldsymbol{\delta}_i \mid \theta; \boldsymbol{\alpha}_t, \boldsymbol{\beta}, \sigma) \, g(\theta) \, d\theta \tag{16}$$

At convergence, these quantities can be obtained using an additional round of quadrature. They might then be used, for example, to evaluate the failure probabilities for particular level-2 units. Also, Ten Have[22] suggests how these empirical Bayes estimates can be used in performing residual diagnostics.

To obtain estimated marginal probabilities, an additional step is required. First, so-called 'cluster-specific' probabilities[46,47] are estimated for specific values of covariates and random effects $\boldsymbol{\theta}_i$ using $\hat{z}_{ijt} = \mathbf{u}'_{ij}\hat{\boldsymbol{\alpha}}_t + \mathbf{x}'_{ij}\hat{\boldsymbol{\beta}} + \mathbf{w}'_{ij}\hat{\mathbf{S}}\boldsymbol{\theta}_i$. These are referred to as cluster-specific probabilities because they indicate response probabilities conditional on the random cluster effects $\boldsymbol{\theta}_i$. Denoting these cluster-specific probabilities as $\hat{P}_{cs}$, marginal probabilities $\hat{P}_m$ are then obtained by integrating over the random-effect

distribution, namely $\hat{P}_m = \int_\theta \hat{P}_{cs}\, g(\theta)\, d\theta$. Again, numerical quadrature can be used for this. These estimated marginal probabilities can then be compared to the observed marginal proportions to examine model fit, either for the whole sample or stratified by covariates.

Analogous to the situation for the probabilities, there is a distinction between cluster-specific and marginal, or population-averaged, model parameters that is important to note. As the models presented in this article are mixed-effects models, the parameters $\boldsymbol{\alpha}_t$ and $\boldsymbol{\beta}$ (and their estimates) are cluster-specific parameters that indicate the covariate effects adjusted or conditional on the random cluster effects $\boldsymbol{\theta}_i$. Alternatively, marginal parameter estimates, like those obtained from GEE models, indicate the (averaged) effect for the population of clusters. As noted by Neuhaus, Kalbfleisch, and Hauck,[46] the values of the subject-specific parameters will exceed the marginal parameters, in absolute value, as the variance attributable to the random effects increases. Also, as noted by these authors, interpretation is more satisfactory for cluster-specific estimates of within-cluster covariates, and for marginal estimates of cluster-level covariates. In many cluster-randomization trials, the degree of data clustering is not large and so the two types of estimates will not differ greatly in scale. Alternatively, this issue is more critical in the analysis of longitudinal data where observations are nested within individuals and the degree of data clustering is usually large.

4 Example: smoking prevention project

4.1 The data set

The Television School and Family Smoking Prevention and Cessation Project (TVSFP) study[48] was designed to test independent and combined effects of a school-based social-resistance curriculum and a television-based program in terms of tobacco use prevention and cessation. The initial study sample consisted of seventh-grade students who were pretested in January, 1986. Students who took the pretest (Wave A) completed an immediate post-intervention questionnaire in April, 1986 (Wave B), a one-year follow-up questionnaire (April, 1987; Wave C), and a second year follow-up (April, 1988; Wave D). The study involved students of schools from Los Angeles and San Diego. A cluster randomization design was used to assign schools to the design conditions, while the primary outcome variables were at the student level. In the analyses below, a subset of the TVSFP data was used. We concentrated on students from Los Angeles schools, where the schools were randomized to one of four study conditions: (a) a social-resistance classroom curriculum (SR); (b) a media (television) intervention (TV), (c) a combination of SR and TV conditions; and (d) a no-treatment control group. These conditions form a 2 x 2 factorial design of SR (yes or no) by TV (yes or no).

One outcome of interest from the study is the onset of cigarette experimentation. At each of the four timepoints, students answered the question: 'have you ever smoked a cigarette?' In analysing the data below, because the intervention was implemented following the pretest, we focused on the three post-intervention timepoints and included only those students who had not answered yes to this question at pretest. Thus, our analysis examines the degree to which the intervention prevented or delayed

Table 1 Onset of cigarette experimentation across three waves: frequencies (and percentages) for gender and condition subgroups

	Wave B			Wave C			Wave D		
	with event	censored	total	with event	censored	total	with event	censored	total
Males	156 (21.0)	83 (11.2)	742	89 (17.7)	134 (26.6)	503	63 (22.5)	217 (77.5)	280
Females	130 (16.0)	105 (12.9)	814	117 (20.2)	154 (26.6)	579	79 (25.6)	229 (74.4)	308
Control	66 (16.5)	60 (15.0)	401	53 (19.3)	69 (25.1)	275	34 (22.2)	119 (77.8)	153
SR only	75 (19.1)	27 (6.9)	392	53 (18.3)	61 (21.0)	290	49 (27.8)	127 (72.2)	176
TV only	71 (17.3)	54 (13.2)	410	60 (21.1)	79 (27.7)	285	38 (26.0)	108 (74.0)	146
SR and TV	74 (21.0)	47 (13.3)	353	40 (17.2)	79 (34.1)	232	21 (18.6)	92 (81.4)	113

students from initiating smoking experimentation. Because the intervention was also aimed at smoking cessation for individuals who had initiated smoking, here we are examining only a part of the intervention aims.

In all, there were 1556 students included in the analysis of smoking initiation. Of these students, approximately 40% ($n = 634$) answered yes to the smoking question at one of the three post-intervention timepoints, while the other 60% ($n = 922$) either answered no at the last timepoint or were censored prior to the last timepoint. The breakdown of cigarette onset for gender and condition subgroups is presented in Table 1. In terms of the clustering, these 1556 students were from 28 schools with between 13 and 151 students per school ($\bar{n} = 56$, SD = 38) Thus, the data are highly unbalanced with large variation in the number of clustered observations.

4.2 A comparison of models

Several proportional hazards models utilizing the complementary log–log link function were fit to these data. Results from these analyses for cigarette onset are given in Table 2. For all models, gender is included as a dummy variable expressing the male versus female difference. For the condition terms, because the SR by TV interaction was observed to be nonsignificant in all analyses, only a main effects model is presented. The first three columns of Table 2 list results ignoring the clustering of students in schools, while the last three columns list results for random-effects modelling incorporating a random school effect. Within each set, comparisons are made between the dichotomous and ordinal data analysis approaches. Additionally, for the sake of comparison, the first column lists results from an ordinary Cox regression analysis carried out using SAS PROC PHREG (with the TIES=EXACT option).

Table 2 Grouped-time onset of cigarette experimentation: 1556 students clustered within 28 schools proportional hazards model estimates (standard errors)

Parameter	Without clustering			With clustering		
	SAS PHREG	Dichot	Ordinal	Dichot normal RE	Ordinal normal RE	Ordinal uniform RE
Intercept α_1		–1.652 (0.094)	–1.652 (0.094)	–1.656 (0.107)	–1.656 (0.107)	–1.656 (0.107)
Intercept α_2		–1.613 (0.096)	–0.939 (0.084)	–1.616 (0.126)	–0.943 (0.107)	–0.943 (0.107)
Intercept α_3		–1.344 (0.106)	–0.428 (0.081)	–1.346 (0.130)	–0.431 (0.096)	–0.431 (0.096)
Male β_1	0.056 (0.080)	0.056 (0.080)	0.056 (0.080)	0.057 (0.124)	0.057 (0.124)	0.057 (0.124)
SR β_2	0.041 (0.080)	0.041 (0.080)	0.041 (0.080)	0.045 (0.104)	0.045 (0.104)	0.045 (0.104)
TV β_3	0.023 (0.080)	0.023 (0.080)	0.023 (0.080)	0.021 (0.094)	0.021 (0.094)	0.021 (0.093)
School SD σ_v				0.051 (0.161)	0.051 (0.161)	0.012 (0.036)
$-2 \log L$						
Full model	3166.7	3187.5	3187.5	3187.4	3187.4	3187.4
With $\beta_2 = \beta_3 = 0$	3167.0	3187.8	3187.8	3187.7	3187.7	3187.7

Dichot = dichotomous complementary log–log regression.
Ordinal = ordinal complementary log–log regression.
RE = random-effects distribution.

Finally, the final two columns list results allowing for the nesting of students within schools assuming a normal and uniform distribution, respectively, for the random-effects distribution.

For the fixed-effects models, it is apparent that the Cox regression results are replicated exactly by either dichotomous or ordinal regression approaches. While the likelihood values are not identical, the differences in deviances ($-2 \log L$) are. The results from the random-effects analyses are similar to those obtained from ordinary analysis at the student-level, and again the results for the dichotomous and ordinal approaches do not differ in terms of the time-invariant effects. Assuming either a normal or uniform distribution for the random effects yields near-identical estimates and standard errors for the model terms and for the value of the deviance. Of course, the estimated standard deviation of the random effect distribution changes depending on the assumed distributional form. Generally, the intervention was not effective in influencing students onset of cigarette experimentation. The effects for both SR and TV are close to zero and nonsignificant. Likelihood-ratio tests for the joint influence of the SR and TV effects (based on the deviances given at the bottom of Table 2) are

clearly nonsignificant. While the effects of these school-level covariates are similar in the models with and without the random school effects, the standard errors are appreciably larger in the random-effects models, relative to the models ignoring the clustering of students. In terms of the gender effect, there is no evidence of a significant effect in any of the models, though the positive estimate is consistent with increased hazard for males, relative to females.

The variability attributable to schools that is estimated in the random-effects models is small and, based on a likelihood-ratio test, the addition of the random-effect variance term is not significant. It should be noted, however, that this test usually has very little power for detecting small, but important, values of the random-effect variance. Thus, for the purpose of assessing the importance of the variation attributable to data clustering, this test should not be relied upon in a strict sense. Furthermore, from a design perspective, because schools were the treatment assignment level, the random school effect should remain in the model regardless of its significance. For models assuming normally distributed random-effects, the estimated school variance can be expressed as an approximate intraclass correlation, $\sigma_v^2/(\sigma_v^2 + \sigma^2)$, where σ^2 represents the variance of the latent continuous event time variable. For the complementary log–log link the standard variance $\sigma^2 = \pi^2/6$, while for the logit link $\sigma^2 = \pi^2/3$ (see Agresti[49]). Applying this formula, for these data the estimated intraclass correlation equals 0.002 under the proportional hazards model. This value is consistent with results from a previous paper by our group,[50] in which, based on the same study dataset, intraclass correlations were evaluated across variable type, time, race, and gender. In that paper, the range of intraclass correlations equalled 0.001 to 0.14 for a smoking behaviour variable. Also, while in the present example conclusions do not differ between the fixed and random-effects models, a previous report[7] shows that conclusions can change even at relatively low levels of intraclass correlation (e.g. 0.02).

In order to test the proportional hazards assumption, interactions with the timepoint dummy-codes were introduced into the dichotomous random-effects model. For the intervention terms, this resulted in a likelihood-ratio $\chi^2 = 4.1$ for the four parameters (two intervention terms by two dummy-codes), indicating that the proportional hazards assumption is acceptable. Alternatively, for gender, a likelihood-ratio $\chi^2 = 8.0$, on two degrees of freedom, rejects the proportional hazards assumption ($p < 0.02$). Allowing the gender effect to vary across time yields estimates of .306 (SE = 0.142), –0.146 (SE = 0.221), and –0.151 (SE = 0.279) for the gender difference at waves B, C, and D, respectively. Thus, as can be seen from Table 1, males have significantly increased hazard for smoking onset at Wave B, but not at Waves C and D.

To examine model fit, estimates from the mixed-effects nonproportional hazards model were compared to Kaplan–Meier estimates of the survival function by gender. These are plotted in Figure 1. The marginal estimates are obtained as described in Section 3.2 using quadrature to integrate over the random school effects. As can be seen, these model-based marginal estimates agree very well with the Kaplan–Meier estimates. Consistent with the analysis, the plot shows an increased hazard for males initially that diminishes across time.

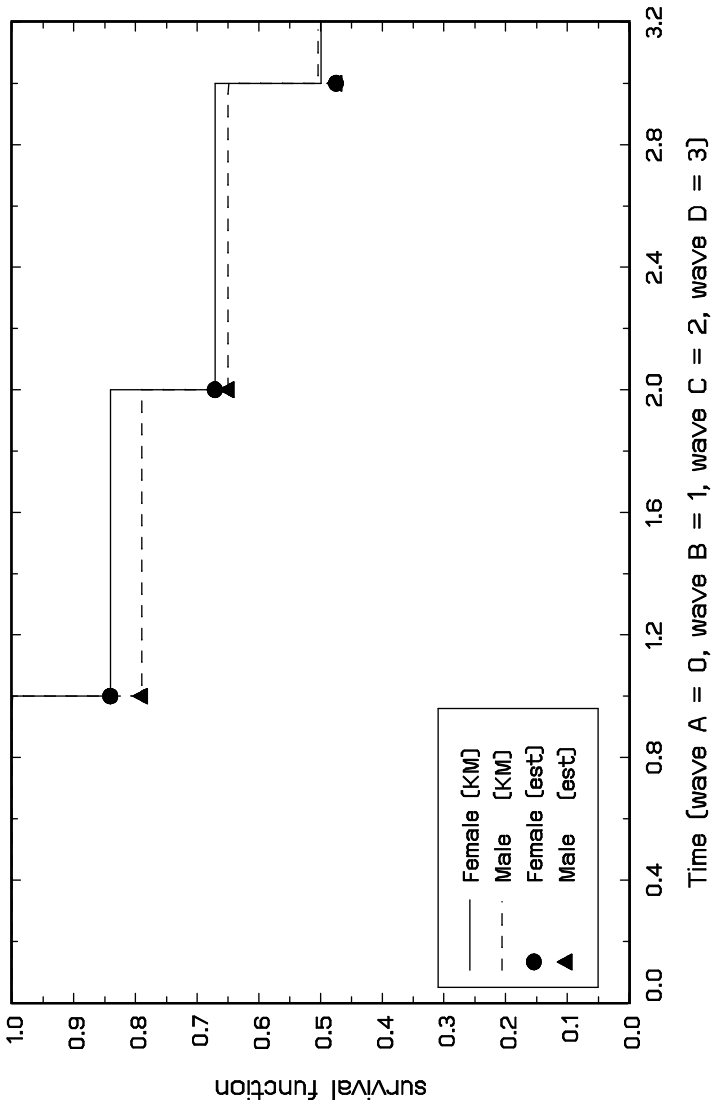

Figure 1 Smoking onset by gender: Kaplan–Meier and marginalized model estimates

5 Discussion

Random-effects categorical regression models are proposed for analysis of clustered grouped-time survival data, using either a proportional or partial proportional hazards or odds assumption. Maximum marginal likelihood methods are used to estimate the model parameters. For this solution, quadrature is utilized to numerically integrate over the distribution of random effects. For models without time-dependent covariates, and assuming proportional hazards or odds, the data are analysed utilizing an ordinal random-effects regression model. In this approach, survival times are represented as ordinal outcomes that are right-censored or not. Alternatively, to relax the proportional hazards assumption and/or to include time-dependent covariates, survival times are represented as sets of binary indicators of survival and analysed using a dichotomous random-effects regression model.

The solution via quadrature can involve summation over a large number of points when the number of random-effects is increased. An issue, then, is the number of necessary quadrature points to insure accurate estimation of the model parameters. As Jansen[52] noted in the unidimensional quadrature solution for a random-effects ordinal model, the estimation is affected very little when the number of points is 5 or greater. Also, as suggested by Bock, Gibbons and Muraki[53] in the context of a dichotomous factor analysis model, the number of points in each dimension can be reduced as the dimensionality is increased. These authors noted that as few as three points per dimension were sufficient for a five-dimensional solution. In the present example, we used between five and 20 quadrature points and observed little change beyond five points, though this might be due to the small degree of data clustering that was evident in the example. In general, to completely resolve this issue for a particular data set a sensitivity analysis varying the number of quadrature points may be advisable.

The use of Gibbs sampling and related methods[54] provides an alternative way of handling the integration over the random effect distribution. As mentioned, Ten Have and Uttal[23] used Gibbs sampling for a discrete-time survival model with multiple random effects. While the quadrature solution is relatively fast and computationally tractable for models with few random effects, Gibbs sampling may be more advantageous for models with many random effects. For example, if there is only one random effect, the quadrature solution requires only one additional summation over Q points relative to the fixed effects solution. For models with $r > 1$ random effects, however, the quadrature is performed over Q^r points, and so becomes computationally burdensome for $r > 5$ or so. Recently, however, Bock and Schilling[55] described a method of adaptive quadrature that uses a fewer number of points per dimension (e.g. three or so) that are adapted to the location and dispersion of the distribution to be integrated. They examined dichotomous factor analysis models with five and eight factors (i.e. random effects) and found similar results as compared to a Gibbs sampling approach.

While only three waves of data were considered in the dataset presented, the model can readily accommodate grouped-time data from many more timepoints. For instance, Han and Hausman[20] used the ordinal logistic model to analyse unemployment duration data from 40 weekly intervals, while Teachman *et al.*[26] used the same

model to analyse unemployment data from 14 timepoints (12 months for the first year and 1 month each for the next 2 years). Since the cutpoints for each time interval are estimated, the number of timepoints considered may depend on the sparseness of the data. In some cases, data may need to be recoded into fewer time intervals. Note, though, that the timepoints are only assumed to be ordinally related, and so, equally spaced timepoints are not necessary.

The methods and analyses described in this article are valid under the standard assumption of noninformative censoring. For the TVSFP dataset this may not be plausible to the degree that students who were censored at time t (observed 'surviving' at t but not observed at $t+1$) were unrepresentative of all students (with the same covariate values) who 'survived' at t. For example, it is possible that censored students at t were not observed at $t+1$ because they were more engaged in delinquent behaviour (i.e. missing school, dropping out of school, etc.), relative to students who survived at t (and were measured at $t+1$), and so could have a higher probability of initiating smoking or alcohol at $t+1$. One way of assessing whether this type of informative censoring influences the results, suggested by Allison,[56] is to perform a sensitivity analysis that treats censoring at t as being equal to experiencing the event at $t+1$. When this was done, parameter estimates did change somewhat, however the general conclusions remained consistent. Notice that with only three timepoints, it is the treatment of censoring at the first two timepoints (Waves B and C) that is of issue. As can be seen from Table 1, for gender the percentages of censored observations are very similar at these two timepoints. For the intervention groups the percentages differ somewhat, though not greatly.

Our example illustrated the utility of the random-effects approach for clustered grouped-time survival data. In particular, random-effects models are useful in accounting for variability attributable to data clustering, while concurrently estimating effects of model covariates. The degree of data clustering is reflected by the estimated random-effects variance term, which can be expressed as an intraclass correlation estimate. Although, in the present article the conclusions did not change with the inclusion of the random effects, such changes can occur even with relatively small intra-cluster correlation.[7] As methods and software are increasingly available for many types of outcome variables (e.g. normal, dichotomous, ordinal, nominal, counts), analysis of data from cluster randomization trials can now appropriately include the design components under which such data were obtained.

Acknowledgements

The authors are grateful to Dr Richard Campbell for many helpful discussions and for suggesting the development of the random-effects ordinal model within the grouped-time survival data context; Dr Brian Flay for use of the data and for constructive comments on earlier drafts; Drs Thomas Ten Have and Jay Teachman for several helpful e-mail notes; Drs Allan Donner and Neil Klar for organizing this special issue of *Statistical Methods in Medical Research*, and to an anonymous reviewer for many helpful and constructive comments. This work was supported by National Institute of Mental Health Grant MH56146 and National Institute of Drug Abuse Grant DA06307.

References

1 Thompson WA. On the treatment of grouped observations in life studies. *Biometrics* 1977; **33**: 463–70.

2 Prentice RL, Gloeckler LA. Regression analysis of grouped survival data with application to breast cancer data. *Biometrics* 1978; **34**: 57–67.

3 Allison PD. Discrete-time methods for the analysis of event histories. In Leinhardt S ed. *Sociological methodology*. San Francisco, CA: Jossey-Bass, 1982: 61–98.

4 Aitkin M, Longford N. Statistical modelling issues in school effectiveness studies (with discussion). *Journal of the Royal Statistical Society, Series A* 1986; **149**: 1–43.

5 Bryk AS, Raudenbush SW. *Hierarchical linear models: applications and data analysis methods*. London: Sage, 1992.

6 Goldstein H. *Multilevel statistical models*, 2nd edn. London: Edward Arnold, 1995.

7 Hedeker D, Gibbons RD, Flay BR. Random-effects regression models for clustered data: with an example from smoking prevention research. *Journal of Consulting and Clinical Psychology* 1994; **62**: 757–65.

8 Vaupel JW, Manton KG, Stallard E. The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography* 1979; **16**: 439–54.

9 Lancaster T. Econometric methods for the duration of unemployment. *Econometrica* 1979; **47**: 939–56.

10 Clayton D, Cuzick J. Multivariate generalizations of the proportional hazards model (with discussion). *Journal of the Royal Statistical Society, Series A* 1985; **148**: 82–117.

11 Self SG, Prentice RL. Incorporating random effects into multivariate relative risk regression models. In Moolgavkar SH, Prentice RL eds. *Modern statistical methods in chronic disease epidemiology*. New York: John Wiley, 1986: 167–78.

12 Guo G, Rodriquez G. Estimating a multivariate proportional hazards model for clustered data using the EM algorithm. With an application to child survival in Guatemala. *Journal of the American Statistical Association* 1992; **87**: 969–76.

13 Paik MC, Tsai W-Y, Ottman R. Multivariate survival analysis using piecewise gamma frailty. *Biometrics* 1994; **50**: 975–88.

14 Shih JH, Louis TA. Assessing gamma frailty models for clustered failure time data. *Lifetime Data Analysis* 1995; **1**: 205–20.

15 Pickles A, Crouchley R. A comparison of frailty models for multivariate survival data. *Statistics in Medicine* 1995; **14**: 1447–61.

16 Hougaard P. Frailty models for survival data. *Lifetime data analysis* 1995; **1**: 255–73.

17 Liang K-Y, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika* 1986; **73**: 13–22.

18 Lee EW, Wei LJ, Amato DA. Cox-type regression analysis for large numbers of small groups of correlated failure time observations. In Klein JP, Goel PK eds. *Survival analysis: state of the art*. Dordrecht: Kluwer, 1992: 237–47.

19 Wei LJ, Lin DY, Weissfeld L. Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *Journal of the American Statistical Association* 1989; **84**: 1065–73.

20 Han A, Hausman JA. Flexible parametric estimation of duration and competing risk models. *Journal of Applied Econometrics* 1990; **5**: 1–28.

21 Scheike TH, Jensen TK. A discrete survival model with random effects: an application to time to pregnancy. *Biometrics* 1997; **53**: 318–29.

22 Ten Have TR. A mixed effects model for multivariate ordinal response data including correlated discrete failure times with ordinal responses. *Biometrics* 1996; **52**: 473–91.

23 Ten Have TR, Uttal DH. Subject-specific and population-averaged continuation ratio logit models for multiple discrete time survival profiles. *Applied Statistics* 1994; **43**: 371–84.

24 Guo SW, Lin DY. Regression analysis of multivariate grouped survival data. *Biometrics* 1994; **50**: 632–39.

25 McCullagh P. Regression models for ordinal data (with discussion). *Journal of the Royal Statistical Society, Series B* 1980; **42**: 109–42.

26 Teachman JD, Call VRA, Carver KP. Marital status and the duration of joblessness among white men. *Journal of Marriage and the Family* 1994; **56**: 415–28.

27 D'Agostino RB, Lee M-L, Belanger AJ, Cupples LA, Anderson K, Kannel WB. Relation of pooled logistic regression to time dependent Cox regression analysis: the Framingham Heart Study. *Statistics in Medicine* 1990; **9**: 1501–15.

28 Singer JD, Willett JB. It's about time: using discrete-time survival analysis to study duration and the timing of events. *Journal of*

Educational and Behavioral Statistics 1993; **18**: 155–95.
29 Hedeker D, Gibbons RD. A random-effects ordinal regression model for multilevel analysis. *Biometrics* 1994; **50**: 933–44.
30 Preisler HK. Analysis of a toxicological experiment using a generalized linear model with nested random effects. *International Statistical Review* 1989; **57**: 145–59.
31 Lee Y, Nelder JA. Hierarchical generalized linear models (with discussion). *Journal of the Royal Statistical Society, Series B* 1996; **58**: 619–78.
32 Gibbons RD, Hedeker D. Random-effects probit and logistic regression models for three-level data. *Biometrics* 1997; **53**: 1527–37.
33 Bock RD, Aitkin M. Marginal maximum likelihood estimation of item parameters: an application of the EM algorithm. *Psychometrika* 1981; **46**: 443–59.
34 Bennett S. Analysis of survival data by the proportional odds model. *Statistics in Medicine* 1983; **2**: 273–77.
35 Collett D. *Modelling survival data in medical research*. New York: Chapman & Hall, 1994.
36 Doksum KA, Gasko M. On a correspondence between models in binary regression analysis and in survival analysis. *International Statistical Review* 1990; **58**: 243–52.
37 Cupples LA, D'Agostino RB, Anderson K, Kannel WB. Comparison of baseline and repeated measure covariate techniques in the Framingham Heart Study. *Statistics in Medicine* 1985; **7**: 205–18.
38 Engel J. On the analysis of grouped extreme-value data with GLIM. *Applied Statistics* 1993; **42**: 633–40.
39 Läärä E, Matthews JNS. The equivalence of two models for ordinal data. *Biometrika* 1985; **72**: 206–207.
40 Terza JV. Ordinal probit: a generalization. *Communications in Statistical Theory and Methods* 1985; **14**: 1–11.
41 Peterson B, Harrell FE. Partial proportional odds models for ordinal response variables. *Applied Statistics* 1990; **39**: 205–17.
42 Cox C. Location-scale cumulative odds models for ordinal data: a generalized nonlinear model approach. *Statistics in Medicine* 1995; **14**: 1191–203.
43 Hedeker D, Mermelstein RJ. A multilevel thresholds of change model for analysis of stages of change data. *Multivariate Behavioral Research* 1998; **33**: 427–55.
44 Hedeker D, Gibbons RD. MIXOR: a computer program for mixed-effects ordinal probit and logistic regression analysis. *Computer Methods and Programs in Biomedicine* 1996; **49**: 157–76.
45 Stroud AH, Sechrest D. *Gaussian quadrature formulas*. Englewood Cliffs, NJ: Prentice Hall, 1966.
46 Neuhaus JM, Kalbfleisch JD, Hauck WW. A comparison of cluster-specific and population-averaged approaches for analyzing correlated binary data. *International Statistical Review* 1991; **59**: 25–35.
47 Zeger SL, Liang K-Y, Albert PS. Models for longitudinal data: a generalized estimating equation approach. *Biometrics* 1988; **44**: 1049–60.
48 Flay B, Brannon B, Johnson C *et al*. The Television School and Family Smoking Prevention and Cessation Project: I. Theoretical Basis and Program Development. *Preventive Medicine* 1988; **17**: 585–607.
49 Agresti A. *Categorical data analysis*. New York: John Wiley, 1990.
50 Siddiqui O, Hedeker D, Flay BR, Hu FB. Intraclass correlation estimates in a school-based smoking prevention study: outcome and mediating variables, by sex and ethnicity. *Americal Journal of Epidemiology* 1996; **144**: 425–33.
51 Snijders TAB, Boskers RJ. *Multilevel analysis*. London: Sage, 1999.
52 Jansen J. On the statistical analysis of ordinal data when extravariation is present. *Applied Statistics* 1990; **39**: 75–84.
53 Bock RD, Gibbons RD, Muraki E. Full-information item factor analysis. *Applied Psychological Measurement* 1988; **12**: 261–80.
54 Tanner MA. *Tools for statistical inference: methods for the exploration of posterior distributions and likelihood functions*, 3rd edn. New York: Springer, 1996.
55 Bock RD, Shilling S. High-dimensional full-information item factor analysis. In Berkane M ed. *Latent variable modeling and applications to causality*. New York: Springer, 1997: 163–76.
56 Allison PD. *Survival analysis using the SAS system: a practical guide*. Cary, NC: SAS Institute, 1995.
57 Magnus JR. *Linear structures*. London: Charles Griffin, 1988.
58 Bock RD. Measurement of human variation: a two-stage model. In Bock RD ed. *Multilevel analysis of educational data*. San Diego, CA: Academic Press, 1989: 319–42.

Appendix: maximum marginal likelihood estimation

Estimation follows the procedure outlined for the mixed-effects ordinal regression model described in Hedeker and Gibbons[29] and implemented in MIXOR[44] with a few additions. First, the conditional likelihood $\ell(\mathbf{t}_i, \boldsymbol{\delta}_i \mid \boldsymbol{\theta}; \boldsymbol{\alpha}_t, \boldsymbol{\beta}, \mathbf{S})$ takes into account right censoring and, as a result, the event indicator vector $\boldsymbol{\delta}_i$. Further, the model presented in this article has been extended to permit nonproportional hazards or odds. Finally, the random-effect variance terms are allowed to vary by groups of i or j units.

To simplify the notation in the derivation that follows, the conditional likelihood is denoted as ℓ_i and the marginal density as h_i. Differentiating first with respect to the parameters that vary with t, we get for a particular vector $\boldsymbol{\alpha}_k$ $(k = 1, \ldots, m)$

$$\frac{\partial \log L}{\partial \boldsymbol{\alpha}_k} = \sum_{i=1}^{N} h_i^{-1} \frac{\partial h_i}{\partial \boldsymbol{\alpha}_k}$$

where

$$\frac{\partial h_i}{\partial \boldsymbol{\alpha}_k} = \int_{\boldsymbol{\theta}} \sum_{j=1}^{n_i} \sum_{t=1}^{m} d_{ijt} \left[\delta_{ij} \frac{(\partial P_{ijt}) a_{tk} - (\partial P_{ij,t-1}) a_{t-1,k}}{P_{ijt} - P_{ij,t-1}} - (1 - \delta_{ij}) \frac{(\partial P_{ijt}) a_{tk}}{1 - P_{ijt}} \right] \ell_i \, g(\boldsymbol{\theta}) \, \mathbf{u}_{ij} \, d\boldsymbol{\theta} \tag{17}$$

and $a_{tk} = 1$ if $t = k$ (and $= 0$ if $t \neq k$). For the logit formulation $\partial P_{ijt} = P_{ijt}(1 - P_{ijt})$, while for the complementary log–log formulation $\partial P_{ijt} = (\exp z_{ijt})(1 - P_{ijt})$. Let $\boldsymbol{\eta}$ represent an arbitrary parameter vector; then for $\boldsymbol{\beta}$ and the vector $\mathrm{v}(\mathbf{S})$ which contains the unique elements of the Cholesky factor $\mathbf{S}$, we get:

$$\frac{\partial \log L}{\partial \boldsymbol{\eta}} = \sum_{i=1}^{N} h_i^{-1} \int_{\boldsymbol{\theta}} \sum_{j=1}^{n_i} \sum_{t=1}^{m} d_{ijt} \left[\delta_{ij} \frac{\partial P_{ijt} - \partial P_{ij,t-1}}{P_{ijt} - P_{ij,t-1}} - (1 - \delta_{ij}) \frac{\partial P_{ijt}}{1 - P_{ijt}} \right] \ell_i \, g(\boldsymbol{\theta}) \, \frac{\partial z_{ijt}}{\partial \boldsymbol{\eta}} \, d\boldsymbol{\theta} \tag{18}$$

where

$$\frac{\partial z_{ijt}}{\partial \boldsymbol{\beta}} = \mathbf{x}_{ij} \qquad \frac{\partial z_{ijt}}{\partial (\mathrm{v}(\mathbf{S}))} = \mathbf{J}_r(\boldsymbol{\theta} \otimes \mathbf{w}_{ij})$$

and $\mathbf{J}_r$ is the transformation matrix of Magnus[57] which eliminates the elements above the main diagonal. If $\mathbf{S}$ is a $r \times 1$ vector of independent random effect variance terms (i.e. if $\mathbf{w}_{ij}$ is a $r \times 1$ vector of level-1 or level-2 grouping variables), then $\partial z_{ijt} / \partial \mathbf{S} = \mathbf{w}_{ij} \theta$ in the equation above.

Fisher's method of scoring can be used to provide the solution to these likelihood equations as described in Hedeker and Gibbons.[29] In general, the scoring solution converges must faster than the EM algorithm when applied to random-effects models.[58] Additionally, the Fisher scoring solution provides standard errors for all model parameters.

Computer implementation

The procedure described in this article has been implemented for use in an extended-version of the original MIXOR program.† The program starts by reading in for each level-2 unit the $n_i \times 1$ time and status vectors $\mathbf{t}_i$ and $\boldsymbol{\delta}_i$, the $n_i \times r$ random-effect design matrix $\mathbf{W}_i$, and the $n_i \times p$ matrix of covariates $\mathbf{X}_i$. Provisional starting values for the model parameters must be specified prior to the start of the iterative procedure. These are estimated by the program using an approximate fixed-effects ordinal regression solution for coefficient vector $\boldsymbol{\beta}$ and intercepts α_{0t} $(t = 1, \ldots, m)$. Starting values for the Cholesky factor $\mathbf{S}$ of the random-effects covariance matrix are specified arbitrarily as a diagonal matrix, with each diagonal element set equal to some fraction of the assumed residual variance value. At each iteration and for each level-2 unit, the solution goes over the Q^r quadrature points, with summation replacing the integration over the random-effect distribution. The conditional probabilities $\ell(\mathbf{t}_i, \boldsymbol{\delta}_i \mid \boldsymbol{\theta}; \alpha_t, \boldsymbol{\beta}, \mathbf{S})$ are obtained substituting the random-effect vector $\boldsymbol{\theta}$ by the current r-dimensional vector of quadrature points $\mathbf{B}_q$. The marginal density for each level-2 unit is then approximated as

$$h(\mathbf{t}_i, \boldsymbol{\delta}_i) \approx \sum_{q}^{Q^r} \ell(\mathbf{t}_i, \boldsymbol{\delta}_i \mid \mathbf{B}_q; \boldsymbol{\alpha}_t, \boldsymbol{\beta}, \mathbf{S})\, A(\mathbf{B}_q)$$

At each iteration, computation of the first derivatives and information matrix then proceeds summing over level-2 units and quadrature points. In the summation over the Q^r quadrature points, substitutions are made in the equations for the first derivatives and information matrix as follows: the $\boldsymbol{\theta}$ random-effect vector is substituted by the current vector of quadrature points $\mathbf{B}_q$, and the evaluation of the multivariate standard density $g(\boldsymbol{\theta})$ is substituted by the current quadrature weight $A(\mathbf{B}_q)$. Following the summation over level-2 units and quadrature points, parameters are corrected according to the scoring solution, and the entire procedure is repeated until convergence. With 20 quadrature points for the one dimensional examples described in this article, convergence (corrections of less than 0.0001 for all parameters) was typically obtained within 40 iterations.

†This program can be obtained from http://www.uic.edu/~hedeker/mix.html