Estimating Time-to-Event From Longitudinal Ordinal Data Using Random Effects Markov Models: Application to Multiple Sclerosis Progression

Micha Mandel, The Hebrew University of Jerusalem

and

Rebecca A. Betensky, Harvard School of Public Health

January 23, 2008

Abstract

Longitudinal ordinal data are common in many scientific studies, including those of multiple sclerosis (MS), and are frequently modeled using Markov dependency. Several authors have proposed random effects Markov models to account for heterogeneity in the population. In this paper, we go one step further and study prediction based on random effects Markov models. In particular, we show how to calculate the probabilities of future events and confidence intervals for those probabilities, given observed data on the ordinal outcome and a set of covariates, and how to update them over time. We discuss the usefulness of depicting these probabilities for visualization and interpretation of model results and illustrate our method using data from a phase III clinical trial that evaluated the utility of interferon beta-1a (Avonex) to MS patients of type relapsing-remitting.

KEYWORDS: Markov model, transition model, ordinal response, prediction.

1 Introduction

Multiple sclerosis (MS) is a chronic inflammatory disease of central nervous system myelin. In time, the disease causes disability which is measured on the ordinal expanded disability status scale (EDSS). In MS studies, disability is often evaluated semi-annually with the aim of estimating the probability of progression, as defined on the EDSS scale. The natural assumption of Markov dependency provides a convenient framework for the estimation of probabilities of various time-dependent events that are of biological interest (prediction). Examples of such events are: reaching a certain level, being in a certain level for two consecutive visits, reaching a group of levels and so forth. Mandel et.al. (2007) developed methodology for analyzing these kinds of events using fixed effects transition models and applied it to MS data. However, their first-order models suffered from lack of fit, partially due to the heterogenous nature of the disease.

Transition models for longitudinal data (Diggle et.al., 2002) express the joint distribution of repeated measures as a product of conditional distributions, and are especially useful when the ultimate goal of the analysis is that of prediction. However, when important subject-specific covariates are not measured, the model explains only part of the true heterogeneity in the data, and moreover the Markov assumption may be violated. One remedy is to incorporate random effects into the Markov transition model. The basic assumption is that conditional on an observed set of covariates and an unobserved latent variable, the sequence of the ordinal variables follows a Markov chain.

Two-state Markov transition models with random effects have been studied by several authors (Cook and Ng, 1997; Albert and Waclawiw, 1998; Albert and Follmann, 2003). These papers assume implicitly that the distribution of the first (baseline) state is independent of the random effect given covariates. This strong and somewhat unnatural assumption is relaxed in a series of papers (i.e., Aitkin and Alfó, 1998, 2003; Alfó and Aitkin, 2000) that suggest models for the influence of the first state on the random effect's distribution. A related framework in which subjects make transitions between states according to a continuous time Markov process but are only observed at n time points was studied by Kalbfleisch and Lawless (1985). Random effects were introduced to this model by Cook (1999) and by Cook, et al. (2004).

The current paper focuses on the problem of prediction from random effects Markov models. It is based on estimation results derived in earlier work to make inferences about the future of the process given its past. Specifically, let $Y_0, Y_1, \ldots, Y_n$ be the sequence of the ordinal variables observed for a subject at the equally spanned time points $0, 1, \ldots, n$, and let X be a vector of covariates. We are interested in estimating $P(Y_v = y|Y_0 = y_0, Y_1 = y_1, \ldots, Y_n = y_n, X = x)$ for some $v > n$, or more generally, $P(A_v|Y_0 = y_0, Y_1 = y_1, \ldots, Y_n = y_n, X = x)$, where the event A_v is determined by $(Y_0, Y_1, \ldots, Y_v)$. Only a few papers have dealt with inferences of this kind in the framework of transition models. Albert (1994) and Mandel et.al. (2007) studied parameters such as mean first passage time and time-to-event probabilities, but only for fixed effects models. Albert and Waclawiw (1998) also estimated mean first passage time in random effects models, but only at the population level, and not based on subject-specific history.

In this paper we extend the methodology of Mandel et.al. (2007) by developing methods for estimation of time-to-event probabilities and associated confidence intervals under random effects Markov models. Our estimates take into account subject-specific history and can be updated over time when more data are collected. In the presence of random effects, much more computational effort is required for deriving estimates and confidence intervals and, more importantly, careful interpretation of estimated coefficients and predicted values is necessary. However, proper interpretation of model results with the aid of graphical tools presented below enables important insights to the longitudinal process studied, e.g. to the natural history of MS.

An alternative, more direct, method for estimating $P(Y_v = y|Y_0 = y_0, Y_1 = y_1, \ldots, Y_n = y_n, X = x)$ is through its empirical analog, namely, the proportion of subjects having $Y_v = y$ among those having $(Y_0 = y_0, Y_1 = y_1, \ldots, Y_n = y_n, X = x)$. It is clear, however, that this proportion can be well estimated only with a large number of independent subjects and for data containing only few covariates. A seemingly natural approach is to use survival analysis methods to predict time-to-event probabilities. However, these methods are not adequate because both time and outcome are discrete and because the event of interest is defined at several time points (see Mandel et al., 2007 for a detailed discussion). Markov models address both of these concerns, and thus are attractive. Introducing random effects to Markov models has two advantages. Firstly, it extends to processes that depend on the whole history, but still keeps the model parsimonious (unlike, for example, models that increase the order of the Markov chain). Secondly, it accounts for and provides measures of the heterogeneity of disease courses, which is typical of many diseases such as MS. That is, heterogeneity is described through both variance component estimates and easily interpretable graphical displays, as shown below.

It is important to distinguish between our usage of the term *prediction* to refer to $P(Y_v = y|Y_0 = y_0, Y_1 = y_1, \ldots, Y_n = y_n, X = x)$, and its usage in the linear and generalized linear mixed models framework (Robinson, 1991, Jiang and Lahiri, 2006). The latter usage of prediction is in a subject-specific sense; either the random effect is estimated or a conditional distribution given the random effect is estimated. Thus, letting U denote the latent random effect, the analog of prediction in the generalized linear models framework is estimation of $U|\{Y_0 = y_0, Y_1 = y_1, \ldots, Y_n = y_n, X = x\}$ or $P(Y_t = y|Y_0 = y_0, Y_1 = y_1, \ldots, Y_n = y_n, X = x, U)$ and their associated mean squared errors. Because we do not have many observations on each subject, and subjects are quite heterogeneous, neither of these quantities is suitable in our setting. Instead, we integrate over the random effect to obtain an estimate for future events that is based on observed data alone. This very important distinction is illustrated on the MS data in Section 4.

The paper is organized as follows. Section 2 defines the model and reviews estimation based on previous papers mentioned above. Section 3 deals with prediction under the mixed effects model. It describes generation of probability estimators and their variances as a function of time conditional on subject-specific covariates and history. Section 4 applies the method to data from a phase III clinical trial of MS patients. It discusses several important points regarding interpretation of the models and provides important insights into the natural history of MS. Section 5 presents results of a simulation study. Section 6 completes the paper with a discussion.

2 Preliminaries

This section gives a brief review of the construction and estimation of the Markov transition mixed model. It also describes an identification problem that affects estimation of model parameters, but not prediction.

2.1 The Model

Consider a discrete time Markov process over the ordinal states $\{1, 2, \ldots, J\}$. Let Y_{iv} be the state of subject i at visit (time) v, let $U_i \sim G$ (with density g) denote a subject-specific latent variable, and let X_i denote a vector of covariates. Our inference will be conditional on the vector of covariates and the baseline state (X_i, Y_{i0}). Using lower case letters for realizations of random variables, the data for subject i having n_i transitions are $(y_{i0}, y_{i1}, \ldots, y_{in_i}, x_i)$ and, omitting the subscript i, its contribution to the (conditional) likelihood, $\mathcal{L}_i$, is

$$\int_{-\infty}^{\infty} \prod_{v=1}^{n} P(Y_v = y_v | Y_{v-1} = y_{v-1}, X = x, U = u) g(u | Y_0 = y_0, X = x) du. \tag{2.1}$$

The two components of the likelihood, the transition probabilities and the distribution of the latent variable, should be modeled. Following Aitkin and Alfó (1998), we assume that $g(u|Y_0 = y_0, X = x) = \sigma^{-1} g_0([u - \beta_{0,y_0}]/\sigma)$, with $\beta_{0,0} = 0$ and for some known g_0, e.g. the standard Normal density. Thus, given the initial state, the random effect is independent of the covariates, and the initial state affects g only through its location. Then, after changing variables, the likelihood contribution reduces to

$$\int_{-\infty}^{\infty} \prod_{v=1}^{n} P(Y_v = y_v | Y_{v-1} = y_{v-1}, X = x, U = \sigma u + \beta_{0,y_0}) g_0(u) du. \tag{2.2}$$

To model the transition probabilities, we follow Mandel et al. (2007) and use the partial proportional odds model for ordinal data:

$$\begin{aligned} p_{k,j}(\beta' x + \gamma u; \alpha) &= P(Y_v = j | Y_{v-1} = k, X = x, U = u) \\ &= \frac{\exp(\alpha_{kj} + \beta' x + \gamma u)}{1 + \exp(\alpha_{kj} + \beta' x + \gamma u)} - \frac{\exp(\alpha_{k(j-1)} + \beta' x + \gamma u)}{1 + \exp(\alpha_{k(j-1)} + \beta' x + \gamma u)}, \end{aligned} \tag{2.3}$$

where $\alpha = (\alpha_{kj})$ is a vector of constants that for each k satisfy the constraints of the proportional odds model (Agresti, 2002). Two important assumptions are embedded in

(2.3). Firstly, the Markov model is time homogeneous. Secondly, the transition probabilities depend on X and U only through their linear combination. The first assumption can be relaxed by considering time-varying models for the parameters α_{kj} or β. The second assumption reflects our view of the random effect as representing unmeasured covariates that, if observed, would be modelled as we do the observed covariates.

Combining (2.2) and (2.3), the likelihood contribution becomes

$$\int_{-\infty}^{\infty} \prod_{v=1}^{n} p_{y_{v-1},y_v}(\beta' x + \gamma[\sigma u + \beta_{0,y_0}]; \alpha) g_0(u) du. \tag{2.4}$$

2.2 Identifiability

It is clear from (2.4) that $(\gamma, \sigma, \beta_0)$ is not identifiable, where $\beta_0 = (\beta_{02}, \ldots, \beta_{0J})$. This, however, does not present a problem for estimation of β or for prediction, hence γ will be set to 1 in the sequel, giving the final working model

$$\mathcal{L}_i = \int_{-\infty}^{\infty} \prod_{v=1}^{n} p_{y_{v-1},y_v}(\beta' x + \sigma u + \beta_{0,y_0}; \alpha) g_0(u) du. \tag{2.5}$$

Suppose that x contains the initial state y_0 as one of its elements. Then, it is clear from (2.5) that the effect of y_0 as a covariate (i.e. the effect of y_0 given $U = u$) is not identifiable. Thus, interpretation of β_0 should be made with care. It seems more reasonable to tie β_0 to the random effect's distribution rather than to the transition probabilities when the initial state is somewhat arbitrary, relating to the sampling time. In that case, interpretation of β_0 is as the center of the random effect's distribution and not in terms of an odds ratio. Moreover, there is nothing special about y_0 in the discussion above, and the same reasoning applies for any other covariate; β is not identifiable if the random effect's density takes the form $g(u|Y_0 = y_0, X = x) = \sigma^{-1} g_0([u - \beta_{0,y_0} - \zeta' x]/\sigma)$. This is quite a reasonable form for $g(u|Y_0 = y_0, X = x)$ when viewing U as an unmeasured covariate and assuming a multivariate Normal distribution for (X, U). Although interpretation of model results is problematic, for prediction purposes this identification issue raises no difficulties since prediction is based on $\beta' x + \gamma[\sigma u + \beta_{0,y_0}]$, which is identifiable (for given x, u and y_0).

2.3 Estimation

The likelihood of N independent subjects is given by

$$\mathcal{L} = \prod_{i=1}^{N} \mathcal{L}_i = \prod_{i=1}^{N} \int_{-\infty}^{\infty} \prod_{v=1}^{n_i} p_{y_{iv-1}, y_{iv}}(\beta' x_i + \sigma u + \beta_{0,y_{0i}}; \alpha) g_0(u) du. \tag{2.6}$$

Estimation of nonlinear random effects model is done via maximization of (2.6). This is a standard, though difficult task. A convenient and flexible routine for maximization of the likelihood is the procedure NLMIXED in SAS, which contains several methods for integration and optimization when g_0 is Normal (Littell et.al., 2006). The EM algorithm is an alternative way of maximizing $\mathcal{L}$ (Aitkin and Alfó, 1998) without requiring specificaion of the distribution of the random effects. However, Agresti found that the random effect distribution has to be extremely non-Normal for the Normal GLMM to suffer from bias or inefficiency (Agresti 2002, pp 547-548). Thus, the Normal model for the random effects distribution seems reasonable in most circumstances in terms of simplicity and interpretability.

2.4 Notation

The following notation will be used in the sequel: $\theta \equiv (\beta', \beta_0')'$ is the vector of fixed effects, $\vartheta \equiv (\alpha', \theta', \sigma)'$ is a vector of length m of all unknowns, $z \equiv (x', I\{y_0 = 2\}, \ldots, I\{y_0 = J\})'$ is the vector of observed predictors, where $I\{\cdot\}$ is the indicator function, and $w \equiv \theta' z + \sigma u$ is the linear predictor. Depending on the context, the transition probabilities will be denoted either by $p_{y_{v-1},y_v}(\theta' z + \sigma u; \alpha)$, $p_{y_{v-1},y_v}(x, u, y_0; \vartheta)$ or $p_{y_{v-1},y_v}(w; \alpha)$. The superscript (s) will be added to denote transitions in s steps: for example, $p^{(s)}_{k,j}(x, u, y_0; \vartheta) = P(Y_{v+s} = j | Y_v = k, X = x, U = u, Y_0 = y_0; \vartheta)$.

3 Prediction

The ultimate goal of our analysis is that of prediction of a future observation given the past observations and covariates:

$$P(Y_v = y_v | \{Y_t = y_t\}_{0 \le t \le n}, X = x). \tag{3.1}$$

Note that here, in contrast to the fixed effects case,

$$P(Y_v = y_v | \{Y_t = y_t\}_{0 \le t \le n}, X = x) \neq P(Y_{v-n} = y_v | Y_0 = y_n, X = x),$$

because the distribution of $(Y_0, Y_1, \dots, Y_v)$ has the Markov property only conditional on U. Thus, the process itself provides information on the latent U which, in turn, is used to predict future events.

Using

$$\begin{aligned} & P(Y_v = y_v | \{Y_t = y_t\}_{0 \le t \le n}, X = x) \\ &= \int_{-\infty}^{\infty} P(Y_v = y_v | \{Y_t = y_t\}_{0 \le t \le n}, X = x, u) g(u | \{Y_t = y_t\}_{0 \le t \le n}, X = x) du \qquad (3.2) \\ &= \int_{-\infty}^{\infty} p^{(v-n)}_{y_n,y_v}(x, u, y_0; \vartheta) \frac{\prod_{t=1}^{n} p_{y_{t-1},y_t}(x, u, y_0; \vartheta)}{\int_{-\infty}^{\infty} \prod_{t=1}^{n} p_{y_{t-1},y_t}(x, u^*, y_0; \vartheta) g_0(u^*) du^*} g_0(u) du, \end{aligned}$$

(3.1) is estimated by

$$\hat{P}(Y_v = y_v | \{Y_t = y_t\}_{0 \le t \le n}, X = x) = \int_{-\infty}^{\infty} p^{(v-n)}_{y_n,y_v}(x, u, y_0; \hat{\vartheta}) \frac{\prod_{t=1}^{n} p_{y_{t-1},y_t}(x, u, y_0; \hat{\vartheta}) g_0(u)}{\int_{-\infty}^{\infty} \prod_{t=1}^{n} p_{y_{t-1},y_t}(x, u^*, y_0; \hat{\vartheta}) g_0(u^*) du^*} du, \tag{3.3}$$

where $\hat{\vartheta}$ is an estimate of ϑ. This last integral can be calculated using numerical methods. A natural simple way is to conduct Monte Carlo integration with respect to the assumed known density g_0. Alternatives are MCMC, which eliminates the burden of approximating the integral in the denominator, and general numerical integration methods.

Of special interest for us is prediction for a subject without any observed transitions. This represents a patient at diagnosis and is analogous to prediction in a model without random effects. For such a case, (3.3) reduces to

$$P_{\hat{\vartheta}}(Y_v = y_v | Y_0 = y_0, X = x) = \int_{-\infty}^{\infty} p_{y_0,y_v}^{(v)}(x, u, y_0; \hat{\vartheta}) g_0(u) du. \tag{3.4}$$

As mentioned in Section 1, it is important to distinguish (3.1) from

$$P(Y_v = y_v | \{Y_t = y_t\}_{0 \le t \le n}, X = x, U = u). \tag{3.5}$$

The problem of estimating quantities similar to (3.5) is referred to as *prediction* in the mixed model literature (Jiang and Lahiri, 2006). In using (3.5), one aims at estimating the *subject-specific* transition probability which is a random variable, even in a frequentist's point of view. A point predictor for (3.5) is $P(Y_v = y_v | \{Y_t = y_t\}_{0 \le t \le n}, X = x, U = \hat{u}_i)$, where $\hat{u}_i$ is the mean or the mode of the distribution of U given data, $U | \{Y_t = y_t\}_{0 \le t \le n}, X = x$, under $\hat{\vartheta}$ (Booth and Hobert, 1998). Jiang (2003) suggests the empirical best predictor $\mathbb{E}\{P(Y_v = y_v | \{Y_t = y_t\}_{0 \le t \le n}, X = x, U)\}$ which is exactly (3.3). Thus, the point estimators of the two prediction problems are the same, but the estimands differ. With small numbers of observations on each subject, the utility of subject-specific parameters, such as those in (3.5), is questionable. This point will be illustrated further in the data analysis in Section 4.

3.1 Prediction Variance

Letting $\hat{\vartheta}$ be the estimator of ϑ based on N independent subjects, and assuming that $\sqrt{N}(\hat{\vartheta} - \vartheta) \to \mathbb{N}(0, \Sigma)$, we can calculate the asymptotic variance of (3.3) and (3.4) by the delta method. Let $P(x, u, y_0; \vartheta)$ be the transition matrix evaluated at $(x, u, Y_0 = y_0; \vartheta)$, and let $p^{*(v-n)}(y_n, y_v, P(x, u, y_0; \vartheta)) = p^{(v-n)}_{y_n,y_v}(x, u, y_0; \vartheta)$ be the $(v-n)$-step transition probability as a function of the one-step transition matrix. We have that

$$\frac{\partial}{\partial \vartheta} p^{(v-n)}_{y_n,y_v}(x, u, y_0; \vartheta) = \frac{\partial}{\partial \text{vec}(P)} p^{*(v-n)}(y_n, y_v, P(x, u, y_0; \vartheta)) \frac{\partial}{\partial \vartheta} \text{vec}(P(x, u, y_0; \vartheta)), \quad (3.6)$$

where vec(P) is the vector representation of the matrix P. The first term on the right hand side of (3.6) can be calculated by a simple matrix multiplication as shown by Mandel et.al. (2007), and the rows of the second term are

$$\left(\frac{\partial}{\partial \alpha} p(\cdot, \cdot, w; \alpha) \ , \ \frac{\partial}{\partial w} p(\cdot, \cdot, w; \alpha)(z', u) \right)$$

where $p(\cdot, \cdot, w; \alpha)$ is the generic form of the transition probabilities evaluated at w and α. Calculation of these derivatives for the partial proportional odds model (2.3) is given in Appendix C of the supplementary material (http://www.biostatistics.oxfordjournals.org). To calculate the derivative of (3.4), one should average (3.6) with respect to g_0, which can be done again by numerical integration.

Differentiation of (3.3) is more complicated, but can still be carried out analytically as shown in Appendix D of the supplementary material. An alternative approach for estimating the variance is based on a simulation that replaces the analytic differentiation, but makes use of the asymptotic properties of the parameters' estimators:

1. Calculate $\hat{\vartheta}$ and $\hat{\Sigma}$, the estimates of ϑ and Σ.

2. Sample B vectors $\vartheta_1, \ldots, \vartheta_B$ from the Normal distribution with parameters $\hat{\vartheta}$ and $\hat{\Sigma}/N$.

3. Calculate (3.3) with ϑ_b instead of $\hat{\vartheta}$ $(b = 1, \ldots, B)$.

4. Calculate the variance of the estimates in the previous step, or calculate confidence intervals using their distribution.

Note that this algorithm requires numerical integration in step 3 for each of the simulated samples. The calculations of prediction variances for models having fixed effects only are considerably simpler (Mandel et al., 2007).

3.2 Choice of Parameters

Simple manipulations of the estimated transition matrix enable estimation of different parameters of interest. For example, one may be interested in estimating time until the process first visits a certain set of states $\mathcal{S}$ (hitting time), or time until the first two consecutive visits to $\mathcal{S}$. The second parameter is of great interest in MS since EDSS in one visit may indicate a temporary progression that is much less important than sustained progression (see Section 4). As an example, consider time until the first two consecutive visits to $\mathcal{S} = \{k : k > j\}$, with the aim of ultimately estimating the probability of two consecutive visits to states larger than j before time v. To estimate these probabilities, one should replace $P = (p_{ij})$ with the working $(J+1) \times (J+1)$ transition matrix

$$Q_{j+} = \begin{pmatrix} p_{11} & \cdots & p_{1j} & p_{1(j+1)} & \cdots & p_{1J} & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ p_{j1} & \cdots & p_{jj} & p_{j(j+1)} & \cdots & p_{jJ} & 0 \\ p_{(j+1)1} & \cdots & p_{(j+1)j} & 0 & \cdots & 0 & \sum_{k>j} p_{(j+1)k} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ p_{J1} & \cdots & p_{Jj} & 0 & \cdots & 0 & \sum_{k>j} p_{Jk} \\ 0 & \cdots & 0 & 0 & \cdots & 0 & 1 \end{pmatrix}. \tag{3.7}$$

Thus, an additional absorbing state is added to indicate the event of interest (see Mandel et. al. (2007) for other modifications). The $(k, J+1)$'th cell of Q_{j+}^v, the v'th power of Q_{j+}, is the probability that two consecutive visits to $\mathcal{S}$ occurred during the first v transitions when the process started at k.

Prediction and calculation of variances or confidence intervals are based on the modified transition matrix, Q_{j+}. Letting $q_{l,k}^{(v)}(x, u, y_0; \vartheta)$ denote the (l, k) element of Q_{j+}^v under ϑ for $(X = x, U = u, Y_0 = y_0)$, the probability of two consecutive visits in states $> j$ before time v, given the process' history up to time n, is estimated by

$$\int_{-\infty}^{\infty} q_{y_n,J+1}^{(v-n)}(x, u, y_0; \hat{\vartheta}) \frac{\prod_{t=1}^{n} p_{y_{t-1},y_t}(x, u, y_0; \hat{\vartheta})}{\int_{-\infty}^{\infty} \prod_{t=1}^{n} p_{y_{t-1},y_t}(x, u^*, y_0; \hat{\vartheta}) g_0(u^*) du^*} g_0(u) du. \tag{3.8}$$

Equation (3.8) assumes that the event $\{$*two consecutive visits in states* $> j\}$ has not occurred before visit n (otherwise the probability is 1), or alternatively, that we are interested in the probability of that event from the present time to time v. The variance is calculated as described in Section 3.1 (see Appendix D of the supplementary material for some technical details).

3.3 Models Without Random Effects

A model without random effects is obtained as a special case by setting $u \equiv 0$ and considering g_0 as degenerate at zero. The averaging over the random effects is not needed, making the calculation much simpler. Also, (3.3) reduces to (3.4), where the last observed state carries all important information of the history of the process. Interpretation of β_0 is now as the coefficients of the covariate y_0, and may be omitted according to the specification of the model.

4 Progression of Multiple Sclerosis

The data set analyzed here is part of a double-blinded phase III clinical trial that evaluated the utility of interferon beta-1a (Avonex) for MS patients with relapsing-remitting disease (Jacobs et.al., 1996). It includes the subset analyzed by Rudick et al. (1999) of all patients who were accrued early enough to complete two years of follow-up by the end of the study and who had brain MRI scans at baseline and yearly thereafter. As seen in Table 3 of Jacobs et al. (1996), the distribution of the time to sustained progression for this subgroup was the same as that of all study subjects. Visits were scheduled to be every six months, but actual visits deviated slightly from the schedule. We used all visits that had a maximum discrepancy of 30 days from the schedule. This resulted in only 16 missed visits (2.5% of the total scheduled visits). In our analysis, we consider these visits to be missing at random.

Most MS clinical trials use the ordinal EDSS to define progression. The EDSS ranges from 0 (normal neurologic exam) to 10 (death due to MS) in 0.5 point steps (see *http://www.mult-*

Table 1: Observed transitions of MS patients between EDSS categories. Numbers in parentheses are two-step transitions corresponding to the 16 missed visits.

	Placebo			Avonex		
	1	2	3	1	2	3
1	49 (1)	23 (0)	6 (0)	50 (1)	26 (1)	3 (1)
2	15 (0)	45 (2)	30 (0)	33 (0)	52 (1)	25 (2)
3	4 (0)	21 (0)	124 (5)	1 (0)	21 (0)	79 (2)

sclerosis.org/expandeddisabilitystatusscale.html for a description of the scale). In the current study, the outcome of interest was time to sustained progression, defined as the time to two consecutive visits with EDSS of at least one point greater than baseline (Jacobs et al., 1996).

The data set contains 68 individuals with a total of 290 one-step transitions in the Avonex group and 72 individuals with 317 one-step transitions in the placebo group, with a maximum follow-up of three years (baseline + six visits). Due to the small number of transitions, we collapsed the EDSS values into three categories: category 1 - EDSS$\leq$ 1.5 (no disability), category 2 - EDSS of 2 or 2.5 (mild disability), and category 3 - EDSS$\geq$ 3 (moderate to severe disability). The total number of transitions is summarized in Table 1.

The partial proportional odds model (2.5) was fitted to the data with $J = 3$, and g_0 the standard Normal density. This model does not constrain the baseline transition matrix, but assumes proportional odds of covariates among all transitions. Parameters were estimated by SAS procedure NLMIXED, using the Dual Quasi-Newton algorithm with integrals evaluated by the default adaptive Gaussian quadrature. Convergence problems were solved by first fitting models without random effects, and using the resulting estimates as initial values for the mixed effects models.

We first estimated the model with a treatment indicator as the only component of x. Estimates and their standard errors are listed in Appendix A of the supplementary material (http://www.biostatistics.oxfordjournals.org). The estimated coefficient for treatment is 1.17, with estimated standard error of 0.475. Thus, Avonex significantly decreases the probability of worsening in disability. This finding is consistent with the results of the original study. We then estimated the model for each arm separately, testing for the influence of patient-specific covariates: age, disease duration, sex, brain lesion volume and brain parenchymal fraction, the first two of which were introduced as time-dependent covariates (see Section 6). None of these covariates showed a significant effect. We thus continued our analysis fitting two separate models, one for each arm, without any covariates, i.e. assuming different values for the parameters α, β_0 and σ^2 for the two arms. Results are listed in Appendix A of the supplementary material.

Using the estimated transition matrices, we calculated the probability of two consecutive visits with EDSS higher than the baseline value as a function of time. We used the modification of the transition matrix given in (3.7) for the analysis. In principle, probabilities of progression can be calculated for any future time point, using the appropriate power (n) in (3.8). The utility of such calculations, however, depends on the validity of the model. In the sequel, we present probabilities up to five years to demonstrate the usefulness of the method, and discuss and compare probabilities up to two years, which was the end-point of the original study.

Figure 1 depicts the probability of progression for a subject at the first visit as a function of time (six-month units) stratified by arm and baseline EDSS. After two years, the probability of sustained progression among those who had EDSS of one at baseline is estimated as 0.46 and 0.55 for the Avonex and placebo arms, respectively. For those having EDSS of two at baseline, the difference is more pronounced, being 0.30 for the Avonex and 0.60 for the placebo patients. It appears that Avonex prevents progression for people with mild disability better than for those having no disability. However, this may be related to the nature of the scale; a change from EDSS of one to two is considered a smaller step than a change from two to three.

Jacobs et.al. (1996) estimated time to progression without conditioning on baseline EDSS. To generate a similar estimate, one can weigh the curves according to the probability of baseline EDSS. For example, in the Avonex arm there were 20 and 30 individuals with baseline EDSS of one and two, respectively, and the overall estimate of the probability of progression would use the weights 2/5 and 3/5. Estimating progression of individuals with EDSS of three or higher is impossible, because EDSS values greater than three were combined.

For comparison, models without random effects were fitted to the same data. Figure 2 presents the estimated progression curves and 95% pointwise confidence intervals for patients in the Avonex arm who had baseline EDSS of two. The two curves represent models with (pluses) and without (circles) baseline EDSS as a covariate (see Section 3.3) and their estimated probabilities are quite similar. However, the probabilities are considerably larger than those predicted by the random effects model (dashed line). Under the random effects model, the estimated σ^2 values are 4.96 and 5.83 for the placebo and Avonex arms, respectively. The respective likelihood ratio statistics that contrast the models with and

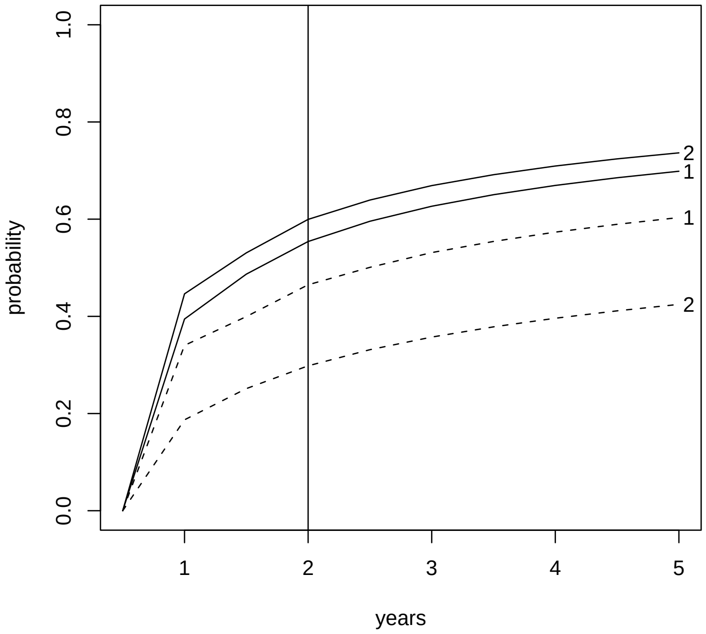

Figure 1: Probability of progression. Solid lines for the Placebo arm and dashed lines for the Avonex arm. The two curves for each arm show different level of baseline EDSS (1 and 2).

without the random effects are 30.7 and 29.2; these are large values for a chi squared distribution with one degree of freedom (in fact, the test statistic under the null hypothesis $\sigma^2 = 0$ is a 50:50 mixture of a χ^2_0 and a χ^2_1 distributions, Self and Liang, 1987, which is stochastically smaller than χ^2_1). These indicate that the heterogeneity in the data is large and support the choice of the random effects model. Various other publications have found empirically that progression is slower than that predicted by the models without random effects (e.g., Jacobs et.al., 1996, Weinshenker et.al., 1989).

To estimate prediction curves for patients for whom EDSS history is available, realizations of curves can be generated using the posterior distribution of the random effect (given history and covariates), as seen in (3.2). Such realizations represent the hypothetical popu-

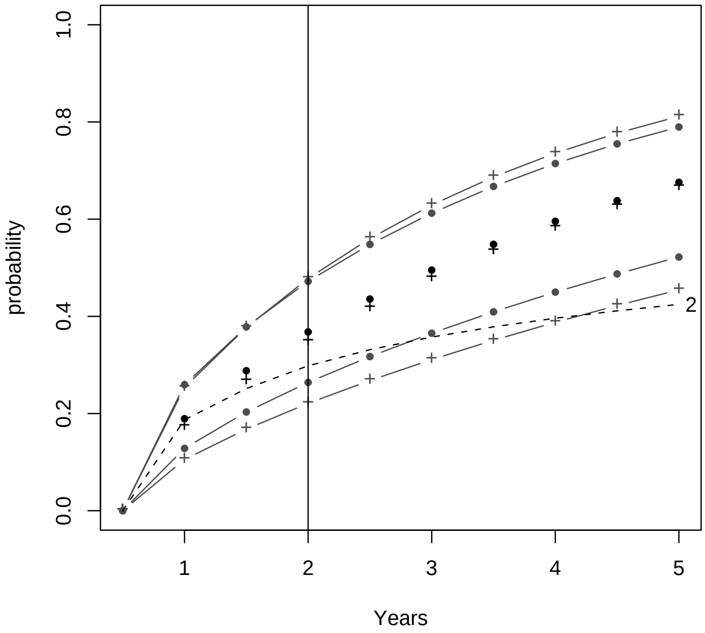

Figure 2: Probability of progression starting in state 2 of the Avonex arm under the fixed models. Pluses and circles denote estimates with and without baseline EDSS as a covariate. Lines are the corresponding estimates of the 95% pointwise confidence intervals. The dashed line is the estimate under the mixed model (lowest curve of Figure 1).

lation of curves that the patient specific curve comes from, and their mean is the probability of progression given the data, i.e., unconditional on the random effect. Depicting curves from the posterior distribution is a useful descriptive tool that helps with interpretation. To illustrate this, Figure 3 depicts 100 curves for two hypothetical subjects in the Avonex group. The left panel represents a subject with baseline EDSS of two and without follow-up data (i.e., at the first visit). The curve on the right represents a hypothetical subject after ten visits with EDSS history (2,2,3,2,3,2,1,2,2,2), and can be considered as a five year update of the progression curve for the subject on the left. Because of a short follow-up, the validity of the model cannot be assessed for five years, hence the curves may not reflect the

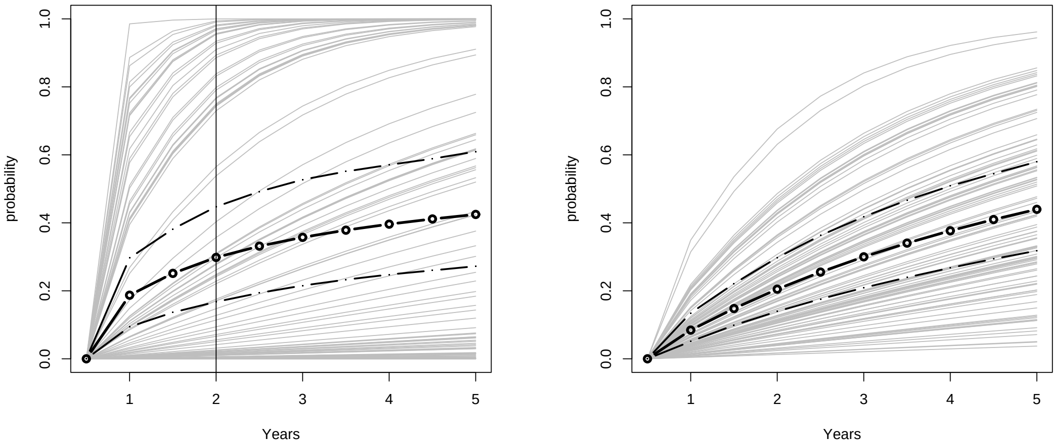

Figure 3: Distribution of curves of the probability of progression. One hundred realizations of the estimated model for the Avonex arm (gray lines) with the estimated mean curve and 95% confidence intervals based on the simulation method. Left - baseline EDSS of 2 no follow-up data; right - ten visits with observed data (2,2,3,2,3,2,1,2,2,2).

real probabilities of progression for such a history. The graph is depicted to demonstrate the utility of the method for valid models. The mean curves and 95% pointwise confidence intervals are depicted too. The variability of the curves on the right is much smaller than that on the left as a result of the additional information. The confidence intervals are not much smaller since they contain the sampling variability of the coefficient estimators. With increasing follow-up data on the same individual, the variability of the gray curves disappears and the graph shows the predicted (or estimated) probability of progression of a specific subject.

Figure 3 illustrates the heterogeneity in the course of the disease and indicates that subject-specific prediction is very difficult. It provides a nice platform for understanding the distinction between (3.1) and (3.5). The quantity (3.5) is essentially one of the gray curves appearing in the figure, while (3.1) is the average of the curves. Thus, (3.1) is a functional of the distribution of (3.5), and can be estimated consistently from the data.

To generate realizations of progression curves for Figure 3, a sampling algorithm from the posterior $g(u|\{Y_t = y_t\}_{0\leq t\leq n}, X = x)$ is required. A natural choice is the MCMC algorithm. However, for the current purpose, a rather small number of independent realizations is needed and a simple alternative is to use a direct sampling. Let $c > \sup_u P(\{Y_t = y_t\}_{0\leq t\leq n}|X = x, u)$. In practice, it is enough to approximate c by calculating $P(\{Y_t = y_t\}_{0\leq t\leq n}|X = x, u)$ on a fixed grid. The rejection/acceptance algorithm (e.g. Evans and Swartz, 2000) for generating one realization is then

1. Generate independently u from g_0 and v from $\mathrm{U}(0, 1)$.

2. If $P(\{Y_t = y_t\}_{0\leq t\leq n}|X = x, u) > cv$ stop and return u. Otherwise go back to step 1.

Independent applications of this algorithm produce independent realizations from the posterior distribution, and these are used in the graph.

Finally, we comment on model selection. Choosing the correct model is always a difficult task in parametric analysis. As discussed above, we contrasted the models with and without random effects, and the analysis strongly supported the inclusion of random effects. We also embedded the random effects model into larger models, such as models with baseline and time-dependent covariates, and models with time-varying vectors of constants, α. Using likelihood ratio tests, all of these extended models were rejected when compared to our final model. As pointed out by an associate editor, probabilities such as (3.1) can be read directly from the data when the sample size is very large. Unfortunately, we do not have a large sample size and thus do need to impose strong parametric assumptions. Even though the sample sizes are very small, we calculated Kaplan-Meier curves for the probability of progression in order to compare informally the data to the models' results. We imputed

Table 2: Kaplan Meier (KM) and model based estimates of the probability of sustained progression during the first two years. CI - confidence interval

Arm	baseline EDSS	KM (95% CI)	model
Avonex	1	.489 (.117,.577)	.465
	2	.338 (.143,.488)	.298
Placebo	1	.520 (.278,.681)	.554
	2	.550 (.269,.723)	.600

values for missing visits by the last observation carried forward approach. Table 2 lists the results and shows good agreement between the data and model.

Since the event of interest (sustained increase in one point) is non-standard in survival analysis and necessitated imputation of missing values, we further compared the model's estimates of two-year transition probabilities to empirical data. We again found good agreement between the two sets of estimates (Appendix A of the supplementary material, http://www.biostatistics.oxfordjournals.org). These findings, and the sensitivity analysis described in the next section, support the validity of the chosen random effects model and the adequacy of the resulted progression curves.

5 Simulation

We conducted a small simulation study to examine the performance of the confidence intervals. We considered two settings, both of which result in 300 transitions. The first was of 100 subjects each with 3 transitions and the second was of 20 subjects each with 15 transitions. We compared the confidence intervals based on the analytical delta method to those obtained by the simulation method described at the end of Section 3.1. We also calculated intervals for models without random effects to illustrate the consequences of model misspecification. Detailed description of the models used and tables of results appear in Appendix B.1 of the supplementary material (http://www.biostatistics.oxfordjournals.org).

We found that confidence intervals with 100 subjects perform well, whereas those for 20 subjects were anti-conservative. This was probably a result of poor Normal approximation for the distribution of the fixed effect estimators in data sets with few subjects. Another feature of the confidence intervals was their uneven distribution on the left and right of the true value, where most intervals that did not include the true parameter assigned values that were too small. This was less pronounced in the percentile method, and suggests that it was mostly related to the linear approximation of the delta method. The similarity between the analytical delta method and the simulation-based approach for the 100 subjects setting was remarkable. In summary, the simulation approach demands more computer time, but saves the burden of calculating and programming complicated derivatives. It also has the additional merit of eliminating the linear approximation of the delta method.

Confidence intervals based on models without random effects perform poorly. In the setting of 100 subjects, there is a tendency towards overestimation which is consistent with the finding of the data analysis in Section 4 (see Figure 2).

We also tested the sensitivity of the model to time-varying transition probabilities. For that, we generated data using log(time) as a covariate, but fitted time-homogeneous models. The results, given in Appendix B.2 of the supplementary material (http://www.biostatistics.oxfordjournals.org), show moderate sensitivity which, as expected, increases with the effect of log(time). For example, when the coefficient of log(time) is -0.5, the bias at time 4 (corresponding to two years) is 0.012, and the bias at time 10 (five years) is -.039. The true probabilities are 0.593 and 0.897, so the bias is relatively small.

6 Discussion

We have presented prediction and confidence interval estimation in the mixed effects Markov model framework, and have provided several graphical tools that help to interpret model results. These graphs, and especially Figure 3, indicate that MS is indeed a heterogeneous disease, and provide useful tools for better understanding the course of the disease. Although we have introduced our models for time-independent covariates, they can be extended to time-varying covariates. Letting x_v be the values of the covariates as measured at visit v, (2.5) is replaced by

$$\mathcal{L}_i = \int_{-\infty}^{\infty} \prod_{v=1}^{n} p_{y_{v-1},y_v}(\beta' x_{v-1} + \sigma u + \beta_{0,y_0}; \alpha) g_0(u) du, \tag{6.1}$$

and estimation of ϑ follows, similar to this extension in models without random effects (Mandel, et al., 2007). If the covariate process is external to the Markov process and its future is known at time v, then prediction can be done as in (3.3) by

$$\hat{P}(Y_v = y_v | \{Y_t = y_t\}_{0 \le t \le n}, \{x_t\}_{t=0}^{v-1})$$
$$= \int_{-\infty}^{\infty} p_{y_n,y_v}^{(v-n)}(\{x_t\}_{t=n}^{v-1}, u, y_0; \hat{\vartheta}) \frac{\prod_{t=1}^{n} p_{y_{t-1},y_t}(x_{t-1}, u, y_0; \hat{\vartheta}) g_0(u)}{\int_{u^*} \prod_{t=1}^{n} p_{y_{t-1},y_t}(x_{t-1}, u^*, y_0; \hat{\vartheta}) g_0(u^*) du^*} du,$$

where $p_{y_n,y_v}^{(v-n)}(\{x_t\}_{t=n}^{v-1}, u, y_0; \hat{\vartheta})$ is the (y_n, y_v) element of the transition matrix $P(x_n, u, y_0; \vartheta) \times P(x_{n+1}, u, y_0; \vartheta) \times \cdots \times P(x_{v-1}, u, y_0; \vartheta)$. The variance can be estimated by the simulation algorithm discussed in Section 3.1.

Our study involves ordinal responses and assumes the partial proportional odds model (2.3), but extensions to other models and to nominal responses are quite straightforward. A general treatment of modelling and estimating random effects for categorical data is given by Hartzel et al. (2001). One can specify such models for each row of the transition matrix, or define one model for all the rows, similar to the approach taken in this work. Prediction is then conducted by integrating over the random effects, exactly as is done here.

MS is known to be a heterogenous disease. Some patients experience no progression for ten or more years (benign MS) whereas others experience a fast and continuous progression from onset (primary progressive MS). The patients analyzed in this paper are relatively homogeneous, since enrollment was subjected to the strict criteria of a phase III clinical trial. Nonetheless, our results indicate that heterogeneity is present and is significant. Prediction of the random effect U would be too ambitious with the typical short follow-up data on each subject, as is well illustrated by Figure 3. In other contexts, predicting U and calculating the mean squared error of prediction, using the approach of Booth and Hobert (1998), is of both theoretical and applied interest and is a topic of current research. When predicting U, it is of interest to calculate also "plug-in" prediction intervals (see Lawless and Fredette, 2005 and references therein). These are calculated by the pointwise percentiles of the realizations of the curves at each time point, and can be included in Figure 3.

Funding

National Institutes of Health (CA075971 to MM and RAB); Harvard Center for Neurodegeneration and Repair (to MM and RAB); Partners Multiple Sclerosis Center (to MM).

Acknowledgments

We thank Dr. Richard Rudick, Director of the Mellen Center for Multiple Sclerosis at the Cleveland Clinic Foundation, for providing us the Avonex trial data on behalf of the Multiple Sclerosis Collaborative Research Group (Principal Investigator, Lawrence Jacobs, MD, deceased). We are also grateful to the editor, an associate editor and a referee for helpful comments and suggestions.

References

[1] AGRESTI, A. (2002). *Categorical Data Analysis* (second edition), Wiley & Sons (New Jersey).

[2] AITKIN, M., and ALFÓ, M. (1998). Regression models for binary longitudinal responses, *Statistics and Computing*, **8**, 289-307.

[3] AITKIN, M., and ALFÓ, M. (2003). Longitudinal analysis of repeated binary data using autoregressive and random effect modelling, *Statistical Modelling*, **3**, 291-303.

[4] ALBERT, P. S. (1994). A Markov model for sequences of ordinal data from a relapsing-remitting disease, *Biometrics*, **50**, 51-60.

[5] ALBERT, P. S., and FOLLMANN, D. A. (2003). A random effects transition model for longitudinal binary data with informative missingness. *Statistica Neerlandica*, **57**, 100-111.

[6] ALBERT, P. S., and WACLAWIW, M. A. (1998). A two-state markov chain for heterogeneous transitional data: a quasi-likelihood approach, *Statistics in Medicine*, **17**, 1481-1493.

[7] ALFÓ, M., and AITKIN, M. (2000). Random coefficient models for binary longitudinal responses with attrition, *Statistics and Computing*, **10**, 279-287.

[8] BOOTH, J. G., and HOBERT, J. P. (1998). Standard errors of prediction in generalized linear mixed models, *Journal of the American Statistical Association*, **93**, 262-272.

[9] COOK, R. J. (1999). A mixed model for two-state Markov processes under panel observation, *Biometrics*, **55**, 915-920.

[10] COOK, R. J., and NG, E. T. M. (1997). A logistic-bivariate normal model for overdispersed two-state Markov processes, *Biometrics*, **53**, 358-364.

[11] COOK, R. J., YI, G. Y., LEE, K. A., and GLADMAN, D. D. (2004). A conditional Markov model for clustered progressive multistate processes under incomplete observation, *Biometrics*, **60**, 436-443.

[12] DIGGLE, P., HEAGERTY, P., LIANG, K. Y., and ZEGER, S. L. (2002). *Analysis of longitudinal data* (second edition), Oxford University Press, (Oxford).

[13] EVANS, M., and SWARTZ, T. (2000). *Approximating Integrals Via Monte Carlo and Deterministic Methods*, Oxford University Press, (Oxford).

[14] HARTZEL, J., AGRESTI, A., and CAFFO, B. (2001) Multinomial logit random effects models. *Statistical Modelling*, **1**, 81-102.

[15] JACOBS, L. D., COOKFAIR, D. L., RUDICK, R. A., HERNDON, R. M., RICHERT, J. R., SALAZAR, A. M., FISCHER, J. S., GOODKIN, D. E., GRANGER, C. V., SIMON J. H., and others (1996). Intramuscular interferon beta-1 alpha for disease progression in relapsing multiple sclerosis, *Annals of Neurology*, **39**, 285-294.

[16] JIANG, J. M. (2003). Empirical best prediction for small-area inference based on generalized linear mixed models, *Journal of Statistical Planning and Inference*, **111**, 117-127.

[17] JIANG, J., and LAHIRI, P. (2006). Mixed model prediction and small area estimation (with discussion), *Test*, **15**, 1-96.

[18] KALBFLEISCH, J. D., and LAWLESS, J. F. (1985). The analysis of panel data under a Markov assumption, *Journal of the American Statistical Association*, **80**, 863-871.

[19] KALBFLEISCH, J. D., and PRENTICE, R. L. (2002). *The Statistical Analysis of Failure Time Data* (second edition), Wiley & Sons (New Jersey).

[20] LAWLESS, J. F., and FREDETTE, M. (2005). Frequentist prediction intervals and predictive distributions, *Biometrika*, **92**, 529-542.

[21] LITTELL, R. C., MILLIKEN, G. A., STROUP, W. W., WOLFINGER, R. D., and SCHABENBERBER, O. (2006). *SAS for Mixed Models* (second edition), SAS Publishing.

[22] MANDEL, M., GAUTHIER, S. A., GUTTMANN. C. R. G., WEINER, H. L., and BETENSKY, R. A. (2007). Estimating time to event from longitudinal categorical

data: an analysis of multiple sclerosis progression, *Journal of the American Statistical Association*, **102**, 1254-1266.

[23] ROBINSON, G. K. (1991). That BLUP is a good thing: the estimation of random efffects (with discussion), *Statistical Science*, **6**, 15-32.

[24] RUDICK, R. A., FISHER, E., LEE, J.-C., SIMON, J. AND JACOBS, L. (1999). Use of the brain parenchymal fraction to measure whole brain atrophy in relapsing-remitting MS, *Neurology*, **53**, 1698-1704.

[25] WEINSHENKER, B. G., BASS, B., RICE, G. P. A., NOSEWORTHY, J., CARRIERE, W., BASKERVILLE, J., and EBERS, G. C. (1989). The natural history of multiple sclerosis: a geographically based study 2. Predictive value of the early clinical course, *Brain*, **112**, 1419-1928.