STATISTICS IN MEDICINE
Statist. Med. 2005; **24**:3331–3345
Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/sim.2186

A mixed-effects regression model for three-level ordinal response data

Rema Raman[1,*,†,‡] and Donald Hedeker[2]

[1]*Division of Biostatistics, University of California at San Diego, La Jolla, CA 92093-0949, U.S.A.*
[2]*Division of Epidemiology and Biostatistics, Institute for Health Research and Policy, School of Public Health (MC 923), University of Illinois at Chicago, Chicago, IL 60612-4336, U.S.A.*

SUMMARY

Three-level data occur frequently in behaviour and medical sciences. For example, in a multi-centre trial, subjects within a given site are randomly assigned to treatments and then studied over time. In this example, the repeated observations (level-1) are nested within subjects (level-2) who are nested within sites (level-3). Similarly, in twin studies, repeated measurements (level-1) are taken on each twin (level-2) within each twin pair (level-3). A three-level mixed-effects regression model is described here. Random effects at the second and third level are included in the model. Additionally, both proportional odds and non-proportional odds models are developed. The latter allows the effects of explanatory variables to vary across the cumulative logits of the model. A maximum marginal likelihood (MML) solution is described and Gauss–Hermite numerical quadrature is used to integrate over the distribution of random effects. The random effects are normally distributed in this instance. Features of this model are illustrated using data from a school-based smoking prevention trial and an Alzheimer's disease clinical trial. Copyright © 2005 John Wiley & Sons, Ltd.

KEY WORDS: multilevel data; mixed-effect models; clustered data; longitudinal data; ordinal response

1. INTRODUCTION

In many areas of research, the outcome data are ordinal in nature. For example, in clinical studies, the outcome could be the severity of the disease designed as mild, moderate or severe. In social studies the response could be level of agreement to a particular question (strongly agree, agree, no opinion, disagree, strongly disagree). In addition, data (longitudinal

*Correspondence to: Rema Raman, Department of Family and Preventive Medicine and Neurosciences, Division of Biostatistics, University of California at San Diego, 9500 Gilman Drive, M/C 0949, La Jolla, CA 92093-0949, U.S.A.
†E-mail: rema@ucsd.edu
‡Assistant Professor.

Contract/grant sponsor: National Institutes of Mental Health; contract/grant number: MH56146

Received 17 April 2003
Accepted 1 October 2004

or clustered) are structured in a hierarchical manner. For example, in a multi-centre trial, subjects within a given site are randomly assigned to treatments and prospectively studied over time. Here, the repeated measurements are nested within subjects who are nested within sites to create three levels of data. In prevention studies, various educational programmes designed to minimize risk taking behaviour (e.g. smoking and/or drug prevention) may be compared where randomization to various design conditions is at the level of the school and the intervention is performed at the level of the classroom. In this case each subject is characterized by a single response, but subjects (level 1) are nested within classrooms (level 2) and classrooms are nested within schools (level 3). When analysing these hierarchical data, one would also like to account for all levels of clustering. Ignoring either the variability in level-1 units within the level-2 units (repeated observations of the subject or observations of subjects within classrooms) or between the level-2 units within level-3 units (individuals within the same site, classrooms within the same school) will lead to invalid tests of hypothesis, inconsistent estimates of uncertainty and misleading inferences and conclusions regarding overall significance of the intervention of interest [1, 2].

For both clustered and longitudinal data, mixed-effects models have been extensively developed to model continuous outcome data [3–6]. A review of the literature also reveals that mixed-effects models for binary outcome data have been widely developed, most of the work using logistic or probit response functions [7–10]. Gibbons and Hedeker [2] extended the mixed-effects model to allow for three-level binary data where observations are nested within individuals are nested within study sites. Other researchers that have developed mixed-effects models for correlated binary data include Longford [11], Qaqish and Liang [12], and Breslow and Clayton [13]. Ten Have *et al*. [14] compared the two-level and three-level mixed-effects model for binary responses in the context of three-level toxicity data and found that the three-level model achieves better coverage of confidence intervals for fixed effects log odds ratios than the two-level model that ignores the association at either level of clustering.

For two-level ordinal data, mixed-effects models that account for the correlation within clusters (clustered model) or within individuals (longitudinal model) have been described by several authors. Harville and Mee [15] developed a mixed model procedure to analyse ordinal outcome data using a Taylor series expansion to estimate the random effects. Ezzet and Whitehead [16] presented a mixed-effects model using the Newton–Raphson method that only allowed for a single random effect. Hedeker and Gibbons [1] developed a mixed-effects model for ordinal responses that allow for multiple random effects, and Hedeker and Mermelstein [17] extended this model to allow for non-proportional odds for a set of explanatory variables.

Mixed-effects models for three-level ordinal data have been proposed by Yang [18] and Rabe-Hesketh *et al.* [19]. Yang [18] used the marginal quasi-likelihood (MQL) or predictive quasi-likelihood (PQL) methods to estimate the statistical parameters. These estimation procedures have been shown to have an appreciable bias or higher mean square error, depending on whether the first-order or second-order methods are used [20]. On the other hand, Rabe-Hesketh *et al.* [19] describe a full-likelihood solution that does not suffer from these biases. Both the Yang [18] and Rabe-Hesketh *et al.* [19] methods consider the case when the proportional odds assumption is reasonable.

In this paper, we will extend the three-level binary mixed-effects model of Gibbons and Hedeker [2] to allow for ordinal outcome data. This model will allow random effects at either level of clustering. For the covariates, both proportional and non-proportional odds models will be described. By comparing these two, the proportional odds assumption can be tested

and if rejected, one can estimate partial and non-proportional odds models. A logistic response function is assumed and a full maximum marginal likelihood solution is used to estimate the parameters in the model. The Gauss–Hermite quadrature method is used to approximate the integrals, and the solution to the likelihood equations are obtained by Fisher's method of scoring. Details of the full-likelihood solution are presented in Appendix A. Model features will be illustrated using data from two studies: (a) a school-based smoking prevention trial and (b) an Alzheimer's disease clinical trial.

2. THREE-LEVEL MIXED-EFFECTS ORDINAL LOGISTIC MODEL

The ordinal logistic or probit model can be motivated by assuming an underlying latent variable (y) that is related to the ordinal response Y through the threshold concept. When the outcome is ordinal, instead of one threshold value γ (as in the binary case), there are several values (depending on the number of ordinal categories) on the continuum that separate individuals into the various response categories. The response of a given subject is determined by the interval in which their unobserved response falls. In other words, an individual is in category c ($Y=c$) when y exceeds the threshold value γ_{c-1}, but does not exceed the threshold value γ_c. For an ordinal model, a series of thresholds $\gamma_1,\ldots,\gamma_{C-1}$ are assumed, where C is equal to the number of ordered categories and $\gamma_0=-\infty$, and $\gamma_C=\infty$. Assuming a normal or logistic distribution for the underlying latent variable leads to an ordinal probit regression model or an ordinal logistic regression model, respectively.

To describe the model, the three-level data structure is defined as follows: Assume that there are $k=1,\ldots,n_{ij}$ level-1 units that are nested within $j=1,\ldots,n_i$ level-2 units that are in turn nested within $i=1,\ldots,n$ level-3 units. Assume that there are $c=1,\ldots,C$ ordered categories. In terms of the latent response strength for subject j in cluster i on occasion k (y_{ijk}) the three-level mixed-effects model can be written as

$$y_{ijk}=\boldsymbol{x}'_{ijk}\boldsymbol{\beta}+\upsilon_{ij}+\upsilon_i+\varepsilon_{ijk} \tag{1}$$

where $\boldsymbol{x}_{ijk}$ is the covariate vector, $\boldsymbol{\beta}$ are unknown regression parameters, υ_i is the unknown random effect at level-3, υ_{ij} is the unknown random effect at level-2, and ε_{ijk} are the model residuals that follow a logistic distribution (or a normal distribution in the case of the probit model). The distribution of the random effects is assumed to be the following: $\upsilon_{ij}\sim\mathcal{N}(0,\sigma^2_{(2)})$ and $\upsilon_i\sim\mathcal{N}(0,\sigma^2_{(3)})$. Since the level-3 subscript i is present for the both υ_{ij} and υ_i, not all level-2 units are assumed to have the same number of level-1 units nested within, and not all level-3 units are assumed to have the same number of level-2 units nested within. In other words, there is no assumption of equal sample size at any level.

Assuming a logistic response function, for a given level-2 unit j nested within level-3 unit i, the probability that $Y_{ijk}=c$ (that is, a response occurs in category c), conditional on the random effects υ_{ij} and υ_i is given by

$$\Pr(Y_{ijk}=c\mid\upsilon_{ij},\upsilon_i)=\boldsymbol{\Psi}[(\gamma_c-z_{ijk})]-\boldsymbol{\Psi}[(\gamma_{c-1}-z_{ijk})] \tag{2}$$

where $z_{ijk}=\boldsymbol{x}'_{ijk}\boldsymbol{\beta}+\upsilon_{ij}+\upsilon_i$ and $\boldsymbol{\Psi}(\cdot)$ represents the logistic cumulative distribution function (c.d.f.), namely $\boldsymbol{\Psi}(z)=1/[1+\exp(-z)]$. Since the scale of the latent variable y is arbitrary, it is common to assume that the logistic distribution is in its standard form, that is, with mean

0 and variance $\pi^2/3$. Additionally, for identification, one must set either one of the thresholds equal to zero (e.g. $\gamma_1 = 0$) or the intercept to zero (i.e. $\beta_0 = 0$). In what follows, we will specify the latter and therefore estimate $C - 1$ thresholds.

With the cumulative probabilities for the c-ordered categories ($c = 1, \ldots, C$) given by

$$P_{ijkc} = \Pr(Y_{ijk} \leqslant c \mid \upsilon_{ij}, \upsilon_i) = \boldsymbol{\Psi}(\gamma_c - z_{ijk}) \tag{3}$$

Then the mixed-effects logistic regression model can be written in terms of the cumulative logits as

$$\log \frac{P_{ijkc}}{(1 - P_{ijkc})} = \gamma_c - (\boldsymbol{x}'_{ijk}\boldsymbol{\beta} + \upsilon_{ij} + \upsilon_i) \tag{4}$$

By including the random effects υ_i and υ_{ij}, (4) is an extension of the (fixed-effects) proportional odds model described by McCullagh [21]. Since the regression coefficients $\boldsymbol{\beta}$ do not depend on c, the model assumes that the relationship between the explanatory variables and the logits do not depend on c.

To obtain the solution for the likelihood equations, the method used by Gibbons and Bock [9] to orthogonally transform the response model is applied. Letting $\upsilon_i = \theta_i \sigma_{(3)}$ and $\upsilon_{ij} = \theta_{ij}\sigma_{(2)}$, the reparametrized model is then

$$z_{ijk} = \boldsymbol{x}'_{ijk}\boldsymbol{\beta} + \sigma_{(2)}\theta_{ij} + \sigma_{(3)}\theta_i \tag{5}$$

where θ_i and θ_{ij} are the standardized univariate normal random effects at level-3 and level-2. The random-effects variance terms are now explicitly included in the regression model, and so they are on the same scale as the regression coefficients, namely, in terms of the logit of response.

2.1. Non-proportional odds

As noted by Peterson and Harrell [22], violation of the proportional odds assumption is not uncommon. Thus, they described a (fixed-effects) partial proportional odds model in which covariates are allowed to have differential effects on the $C - 1$ cumulative logits. Similarly, Terza [23] developed a similar extension of the (fixed-effects) ordinal probit model. Hedeker and Mermelstein [17] utilize this extension within the context of a two-level ordinal regression model. For this, the model for the $C - 1$ cumulative logits can be written as

$$\log \frac{P_{ijkc}}{(1 - P_{ijkc})} = \gamma_c - (\boldsymbol{x}'_{ijk}\boldsymbol{\beta} + \boldsymbol{u}'_{ijk}\boldsymbol{\alpha}_c + \sigma_{(2)}\theta_{ij} + \sigma_{(3)}\theta_i) \tag{6}$$

where, $\boldsymbol{u}_{ijk}$ is a $h \times 1$ vector containing the values of observation ijk on the set of h covariates for which proportional odds is not assumed.

Here, $\boldsymbol{\alpha}_c$ is a $h \times 1$ vector of regression coefficients associated with these h covariates. Because $\boldsymbol{\alpha}_c$ carries the c subscript, the effects of these h covariates are allowed to vary across the $C-1$ cumulative logits. These terms are often referred to as threshold interactions. A given model that includes covariates in both $\boldsymbol{x}$ and $\boldsymbol{u}$ would be a partial proportional odds model, whereas one with only $\boldsymbol{u}$ variables would be a non-proportional odds model. Comparison of models (6) and (5), using a likelihood-ratio test, provides a way of assessing the validity of the proportional odds assumption.

Parameters are estimated using a maximum marginal likelihood (MML) estimation method. The details about the estimation are described in Appendix A. Fisher's method of scoring can be used to provide the solution to these likelihood equations. For this, provisional estimates for the vector of parameters $\boldsymbol{\Theta}$, on iteration i are improved by

$$\boldsymbol{\Theta}_{i+1} = \boldsymbol{\Theta}_i - \mathscr{E}\left[\frac{\partial^2 \log L}{\partial \boldsymbol{\Theta}_i \, \partial \boldsymbol{\Theta}_i'}\right]^{-1} \frac{\partial \log L}{\partial \boldsymbol{\Theta}_i} \tag{7}$$

where, following Bock and Lieberman [24], the information matrix, or expectation of the matrix of second derivatives, is given by

$$\mathscr{E}\left[\frac{\partial^2 \log L}{\partial \boldsymbol{\Theta}_i \, \partial \boldsymbol{\Theta}_i'}\right] = -\sum_{i=1}^{N} h^{-2}(\mathbf{Y}_i) \frac{\partial h(\mathbf{Y}_i)}{\partial \boldsymbol{\Theta}_i} \left(\frac{\partial h(\mathbf{Y}_i)}{\partial \boldsymbol{\Theta}_i}\right)'$$

At convergence, the large-sample variance covariance matrix of the parameter estimates is then obtained as the inverse of the information matrix.

2.2. Numerical quadrature

In order to solve the above likelihood equations, numerical integration on the transformed θ space was performed. For this, Gauss–Hermite quadrature was used to approximate the integrals to a practical degree of accuracy. In Gauss–Hermite quadrature, 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 θ space, the summation first goes over Q_2 points for the level-2 effects, then over Q_3 points for the level-3 effects. For the standard normal univariate density, optimal points and weights (which will be denoted B_{q_3}, B_{q_2} and $A(B_{q_3})$, $A(B_{q_2})$, respectively) are given in Reference [25]. Although the random effects were considered to be normally distributed in the described model, other distributional forms can be used. For example, if a rectangular or uniform distribution is assumed, then Q_3 and Q_2 points may be set at equal intervals over an appropriate range (for each dimension) and the quadrature weights are then set equal to $1/Q_3$ and $1/Q_2$. Other distributions are possible: Bock and Atkin [26] discuss the possibility of empirically estimating the random-effect distribution. The two-level mixed-effects models in this article were estimated using the MIXOR program [27], while the three-level models were programmed using FORTRAN code. The starting values for the parameters $\boldsymbol{\beta}$ and γ were obtained using the fixed-effects regression approach, and the random-effects variance terms were arbitrarily set. For the Gauss–Hermite quadrature solution, we used 20 quadrature points and corresponding weights. Convergence was achieved when corrections for all parameters were less than 0.001.

2.3. Intraclass correlation

For random-effects models it is often of interest to express the cluster variance in terms of an intraclass correlation. For a two-level problem, this equals $\sigma^2_{(2)}/(\sigma^2_{(2)} + \pi^2/3)$, where the latter term in the denominator represents the variance of the underlying latent distribution, which is a standard logistic distribution here. For a three-level model, we have several different kinds of intraclass correlation coefficients (ICC) that are of potential interest. For the model described in this paper, $\sigma^2_{(3)}/(\sigma^2_{(2)}+\sigma^2_{(3)}+\pi^2/3)$ represents the ICC at level 3, $(\sigma^2_{(2)}+\sigma^2_{(3)})/(\sigma^2_{(2)}+\sigma^2_{(3)}+\pi^2/3)$

the ICC at level 2 and level 3 and $\sigma^2_{(3)}/(\sigma^2_{(2)} + \sigma^2_{(3)})$ represents the ICC at level 2. Snijders and Bosker [20] describe in greater detail the different interpretations that these ICCs have.

3. ILLUSTRATION

3.1. Study 1: a school-based smoking prevention study

The Television School and Family Smoking Prevention and Cessation Project (TVSFP) study [28], was designed to test independent and combined effects of a school-based social-resistance curriculum and a television-based programme in terms of tobacco use prevention and cessation. For this illustration, a subset of the TVSFP data was used. The data set consisted of data on seventh grade students from 135 classrooms and 28 schools in the Los Angeles area. Students were randomized to four study conditions: (a) a social-resistance classroom curriculum (CC), (b) a media (television) intervention (TV), (c) CC combined with TV, and (d) a no-treatment control group. Randomization for this study was at the school level while the intervention was delivered to the students in the classrooms. One of the primary response variables for this intervention was the tobacco and health knowledge scale (THKS) score. The scale takes values from 0 to 7, with higher scores implying greater knowledge. Only the 1600 children who had both pre- and post-intervention scores were included in the analysis. The resulting data was unbalanced with a range of between 1 and 13 classrooms per school and between 2 and 28 students per classroom. Table I lists the student frequencies for the THKS scores broken down by each of the condition groups. Based on the frequency distribution of the post-intervention THKS scores, ordinal classes of 0–1, 2, 3, and 4–7 correct responses were used for the purpose of analysis.

A three-level ordinal logistic model was fit to these data. This data set was previously analysed using a two-level ordinal probit regression model [1] and a three-level binary probit regression model [2]. For the present illustration of the model, we will consider a three-level ordinal logistic regression model. Additionally, for each analysis we will compare the

Table I. THKS post-intervention results subgroup frequencies (and percentages).

Subgroup		THKS score				
CC	TV	0–1	2	3	4–7	Total
No	No	117 (27.8)	129 (30.6)	89 (21.1)	86 (20.4)	421
No	Yes	110 (26.4)	105 (25.2)	91 (21.9)	110 (26.4)	416
Yes	No	62 (16.3)	78 (20.5)	106 (27.9)	134 (35.3)	380
Yes	Yes	66 (17.2)	86 (22.5)	114 (29.8)	117 (30.5)	383
Total		355 (22.2)	398 (24.9)	400 (25.0)	447 (27.9)	1600

Table II. THKS post-intervention (ordinal) scores: comparison of two- and three-level mixed-effect logistic model estimates (standard errors).

Term	Fixed-effect model	Two-level model school	Three-level model
Threshold 1	−0.040	−0.089	−0.099
	(0.123)	(0.313)	(0.336)
Threshold 2	1.185	1.153	1.175
	(0.124)	(0.282)	(0.300)
Threshold 3	2.345	2.332	2.385
	(0.133)	(0.293)	(0.310)
Baseline THKS	0.422***	0.403***	0.408***
	(0.038)	(0.043)	(0.044)
Social resistance (CC)	0.863***	0.924**	0.885**
	(0.132)	(0.371)	(0.388)
Television (TV)	0.253**	0.275	0.232
	(0.125)	(0.315)	(0.342)
Interaction (CC × TV)	−0.367**	−0.466	−0.368
	(0.183)	(0.406)	(0.419)
Class SD			0.385
			(0.115)
School SD		0.271	0.215
		(0.091)	(0.122)
Log L	−2125.10	−2119.74	−2114.52

***$p<0.01$; **$p<0.05$.

proportional odds model to a non-proportional odds model in order to assess the assumption of proportional odds for the explanatory variables.

In the clustered situation, subjects (level 1) were nested within classrooms (level 2) and schools (level 3). The post-intervention THKS scores was modelled in terms of the baseline THKS score, and effects of CC and TV (as main effects) and the CC × TV interaction. Table II lists results from the analysis.

The results from the fixed-effects model indicates a significant CC × TV interaction. However, results from the mixed-effects logistic regression model allowing for the nesting of students within schools (column 2 in Table II) are slightly different from those obtained by the fixed-effects model in that neither the interaction term nor the main effect due to television are significant. The standard errors (s.e.) for the two-level model are larger than those from the fixed-effects model indicate ignoring the clustering leads to underestimated standard errors. The three-level model (students within classrooms within schools) leads to similar conclusions as the two-level mixed-effects model and indicates a positive effect of the social resistance curriculum. However, the standard errors for the three-level model are the largest, as expected and the effect of the social resistance intervention is not as strong as in the two level case. Both the two and three-level model provide a significantly better fit than the fixed-effects model ($\chi^2_1 = 10.72$, $p<0.01$ and $\chi^2_2 = 21.18$, $p<0.01$, respectively).[§]

[§]Likelihood-ratio tests of variance parameters tend to be overly conservative [20], so if anything, the significance of the random-effects variance parameters is understated.

In addition, the three-level model also provided a significantly better fit than the two-level model ($\chi^2_1 = 10.46$, $p < 0.01$), supporting the inclusion of both random classroom and school term in the model. The intraclass correlation was 0.043 and the intra-school correlation was 0.013. Hence, approximately 4.3 per cent of the variability was due to classrooms and 1.3 per cent was due to schools. It should be noted that there is concern in using Wald statistics for hypothesis testing of random-effects variance terms [29]. Hence, we do not list these for the random-effects parameters in the tables. Finally, fitting a three-level non-proportional odds model (not shown), which allowed the four explanatory variables to have varying effects across the three cumulative logits, resulted in a log-likelihood value of -2110.39. Thus, the assumption of proportional odds for these variables is not rejected ($\chi^2_8 = 8.26$, $p = 0.41$), and so the reported results for the three-level proportional odds model are reasonable.

The results from this analysis generally agree with the results reported earlier [1] when these data were analysed using a two-level ordinal model (students within classrooms), though as one would expect the standard errors are larger for the treatment effects in the presented three-level model. This is especially true since they considered students nested within classrooms, rather than students within schools (as we present in Table II). Additionally, the current analysis which accounts for all levels of clustering in the data is more appropriate than models that ignore either level of nesting, and the likelihood ratio test clearly supports the significance of including a random effect to the model at all levels of nesting.

3.2. Study 2: an Alzheimer's disease clinical trial

The haloperidol, trazodone, and behaviour management techniques in Alzhemier's disease patients with disruptive agitated behaviours (AP) trial [30] was a randomized, placebo-controlled multi-centre clinical trial designed to compare haloperidol, trazadone and BMT with placebo in the treatment of agitation in AD patients. The study enrolled a total of 149 patients randomized to one of the three treatment arms or the placebo. Assessment was conducted at the mid-point (week 9), after 16 weeks of treatment (posttreatment), and at 3 months follow-up. Of the 149 patients, 129 (87 per cent) had data at week 9, 114 (77 per cent) had data at posttreatment and 84 (56 per cent) had data at 3 months follow-up. The primary outcome measure for this study was the ADCS clinical global impression of change (ADCS-CGIC), completed by a trained clinician blind to treatment assignment. The ADCS-CGIC is a 7-point ordinal scale (that range from marked improvement to marked deterioration) used to assess change from baseline. The data set had a three-level clustered data structure with a range of 1–3 ADAS-CGIC repeated observations per participant and between 1 and 17 participants per clinical centre. Hence, these data are highly unbalanced with a varying number of observations at both levels of clustering.

For the analysis, the ADCS-CGIC was recorded into a 3-point scale ($1 =$ any improvement; $2 =$ no change; $3 =$ any deterioration) due to the fact that there were very few participants in several of the cells and to facilitate the clinical interpretation of the results. The test groups were also combined since analysis showed a similar result for all three treatments. Three ordinal logistic models were fit to these data: A two-level proportional odds model (ignoring the clustering due to site), a three-level proportional odds model, and a three-level non-proportional odds model. For all models, two dummy-codes were included for the time effect using mid-point (week 9) as the reference cell. Similarly, a dummy-code was included for treatment group (active treatment vs placebo). The Cohen–Mansfield agitation inventory

Table III. CGIC response over time. Ordinal logistic model estimates and standard errors.

	Two-level model	Three-level model		
	Proportional odds model 1	Proportional odds model 2	Partial-proportional odds model 3	
Term	Estimate (s.e.)	Estimate (s.e.)	No improvement[†] Estimate (s.e.)	Deterioration[‡] Estimate (s.e.)
Threshold 1	−0.40	−0.35	−0.53	
	(0.46)	(0.55)	(0.55)	
Threshold 2	1.27	0.92		1.07
	(0.13)	(0.55)		(0.55)
Baseline CMAI (High)	0.54	0.81*	0.35	1.15**
	(0.37)	(0.40)	(0.43)	(0.39)
Time 1 (Week 17)	−0.09	−0.12	−0.11	
	(0.28)	(0.28)	(0.28)	
Time 2 (Month 3 FU)	0.22	0.21	0.21	
	(0.31)	(0.43)	(0.42)	
Treatment (Active)	0.28	0.14	0.13	
	(0.43)	(0.43)	(0.42)	
Site SD (Level 3)		1.03	1.02	
		(0.41)	(0.39)	
Patient SD (Level 2)	1.59	1.29	1.22	
	(0.26)	(0.31)	(0.30)	
$-2\log L$	688.14	678.32	669.11	

*$0.025 < p < 0.05$; **$p < 0.01$.
[†]Logit comparing deterioration and no change *versus* improvement.
[‡]Logit comparing deterioration *versus* no change and improvement.

(CMAI) was dichotomized into participants with high and low scores at baseline based on a median split. For the non-proportional odds model, the effects of CMAI, time and treatment were allowed to vary across the two cumulative logits. The likelihood ratio test supported the non-proportional odds assumption but the only variable where the effect was different across the two logits was CMAI. Hence, the final model treats the effect of time and treatment to be constant across the cumulative logits and allows the CMAI effect to vary (i.e. a partial-proportional odds model).

Table III presents the results of these analyses.

The first column gives the results from the two-level regression model that ignores clustering of the sites. The analysis shows no effect of baseline CMAI score, time or treatment on the ADAS-CGIC scores during the study. The three-level proportional odds model, on the other hand, indicates a significant effect due to baseline CMAI score ($p = 0.0452$). In other words, patients with a poor baseline performance in CMAI (i.e. higher CMAI scores) tend to do worse consistently across time.

Like the two-level results, there is no significant treatment or time effect in the three-level model. The likelihood ratio χ^2_1 statistic comparing models 1 and 2 is 9.82 ($p < 0.001$), indicating that the three-level model provides a better fit compared to the two-level model. Turning to column three, comparison of the two three-level models indicates that the

Table IV. Effect of CMAI scores (aggregated across all timepoints) across the cumulative logits.

	CGIC score		
CMAI group	Improved	No change	Worsened
Lower baseline score	66	52	67
Cumulative odds	66/121	118/67	
logit	−0.5	0.6	
Higher baseline score	56	21	77
Cumulative odds	56/98	77/77	
logit	−0.5	0	

assumption of proportional odds is not resonable for CMAI (likelihood-ratio $\chi^2_1 = 9.21$; $p < 0.001$). In other words, the baseline CMAI effect is not identical across the two cumulative logits. From Table III, we see that the baseline CMAI effect is not significant in terms of the no improvement logit (that is, comparing deterioration and no change categories combined *versus* improvement), while there is a significant effect of baseline CMAI performance in terms of the deterioration logit (that is, comparing deterioration *versus* no change and improvement categories combined). Thus, patients with higher baseline CMAI scores are more likely to show deterioration during the course of the study, relative to those with lower CMAI baseline scores. The effects of time and treatment are not significant and are similar across the two three-level models.

To better describe the non-proportional odds results, the breakdown of CGIC scores and their cumulative logits (aggregated across all time points) by CMAI groups is presented in Table IV. From the table, we see the varying group difference on the two cumulative logits (group difference of 0 for the first logit and 0.6 for the second logit). This difference in CMAI effect across the cumulative logits is consistent with the results obtained by the three-level partial proportional odds model. Namely, whereas the groups are very similar in terms of the improved category, the higher CMAI group has relatively more responses in the worsened category (or, the lower CMAI group has relatively more responses in the no change category).

In terms of the random effects variance terms, there is considerable variability associated with both patients and sites, as expected. The estimated intraclass correlation for patients is 0.36 and the estimated intraclass correlation for sites is 0.20. Hence, the data are highly correlated with 36 per cent of the variance attributable to the patient and 20 per cent of the variance attributable to the site. Clearly, ignoring either of these levels, with their moderately high degree of intraclass correlation, would not be reasonable.

4. DISCUSSION

A mixed-effects regression model that allows for both an ordinal response and a three-level data structure is described in this article. This model allows random effects to be included at the second and third level of clustering. This model can be used both in the case of clustered and longitudinal data and is an extension of the three-level mixed-effects model for binary data proposed by Gibbons and Hedeker [2]. Covariate effects are allowed to be the same or vary across the cumulative logits of the model. Thus, the proportional odds assumption

can be tested and, when violated, non-proportional or partial-proportional models can be fit. Parameter estimation is done using maximum marginal likelihood estimation, with Gauss–Hermite quadrature used to numerically integrate over the distribution of random effects. The solution to the non-linear equations is obtained iteratively using Fisher's method of scoring.

Several statistical software programs are now available to estimate mixed-effects regression models for ordinal data, but they vary as to which specific models can be fit. The packages can be broadly divided into three groups based on the models that they fit: two-level non-proportional or proportional odds models (SAS NLMIXED, MIXOR, HLM), three-level models without using full-likelihood estimation methods (MLwiN) and three-level proportional odds models using full-likelihood estimation procedures (GLLAMM). The two level models in this manuscript were fit using MIXOR while a FORTRAN program was written to fit the proposed three-level proportional odds and partial proportional odds models.

One of the issues when using the quadrature method is the determination of the number of points used to approximate the integrals. In the analysis of the data described in this paper, we used 20 quadrature points per dimension for the solutions. The use of Gibbs sampling and related methods [31] provides an alternative way of handling the integration over the random-effects distribution. While the quadrature solution is relatively fast and computationally tractable for models with few random effects, Gibbs sampling is 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. In the present case with one random effect per level, but two levels of clustering, the quadrature solution requires summation over Q^2 points, and so the quadrature solution was viable. For models with many more random effects, the adaptive quadrature method might also be more beneficial. This method uses fewer points per dimension, since the quadrature is adapted to the location and dispersion for each subject [32]. More work is underway to explore this issue further.

The examples demonstrate the use of the random-effects approach for correlated multi-level data. There are alternative approaches, such as the generalized estimating equations (GEE) approach [4] that can be applied to correlated data. This approach has the advantage that it is easier to implement, the parameter estimates are on the marginal scale and the model is robust to misspecification of the variance–covariance structure. In contrast, since the random effects are specifically included in the model, parameter estimates from mixed-effects models depend on the specified distribution of the random effects and the statistical tests are valid only under the assumed correlation structure. However, the quadrature solution used in this article allows the use of varying distributions for the random effects. The random-effects approach proposed in this article, has several advantages of its own. Often, the random effect themselves are of interest (for example, in behavioural research, twin studies). The random-effects model allows the estimates of these random effects to be computed. Another issue that is important to consider between the two approaches is the assumptions that the method makes about missing data: the random-effects approach assumes that the missing data are missing at random (MAR) as compared to the GEE approach that requires a more restrictive missing completely at random (MCAR) assumption. It should be noted that this comment relates to the original representation of GEE, now often referred to as GEE1 [33], and that generalizations of the GEE approach to allow MAR missing data have been proposed (e.g. [34]). More details on the differences between GEE and random effects models are discussed by several authors [35–37].

Data from two studies were used to demonstrate this model. The first study was a behavioural intervention study where the outcome assessment was the THKS score, an ordinal outcome with four response categories. The analyses demonstrated the necessity to account for the multiple levels of nesting, and indicated that the proportional odds assumption was reasonable. The second study was a Alzheimer's disease therapy trial where the outcome was the ADAS-CGIC score, an ordinal variable with three categories. In this analysis, the proportional odds assumption was violated for one of the model covariates. Both these examples illustrate the usefulness of the model in accounting for the variability due to clustering at both levels of these data when estimating the effects of model covariates. The second example illustrates the usefulness of the model in testing the assumption of proportional odds and, if not met, providing an appropriate alternative to this restrictive assumption.

APPENDIX A: MAXIMUM MARGINAL LIKELIHOOD ESTIMATION

Let $\mathbf{Y}_i$ denote the vector pattern of ordinal responses from level-3 unit i for the n_i level-2 units nested within the level-3 units, and the n_{ij} level-1 units nested within the level-2 units. Assuming that the responses are independent conditional on the random effects, the probability of any pattern $\mathbf{Y}_i$, given θ_i and θ_{ij}, is equal to the product of the probabilities of the level-1 responses:

$$\ell(\mathbf{Y}_i \mid \theta_{ij}, \theta_i) = \prod_{j=1}^{n_i} \prod_{k=1}^{n_{ij}} \prod_{c=1}^{C} [\boldsymbol{\Psi}(\gamma_c - z_{ijkc}) - \boldsymbol{\Psi}(\gamma_c - z_{ijk,c-1})]^{d_{ijkc}} \tag{A1}$$

where $d_{ijkc} = 1$ if $Y_{ijk} = c$ and 0 otherwise (i.e. for each ijkth observation, $d_{ijkc} = 1$ for only one of the C categories), and $z_{ijkc} = \boldsymbol{x}'_{ijk}\boldsymbol{\beta} + \boldsymbol{u}'_{ijk}\boldsymbol{\alpha}_c + \sigma_{(2)}\theta_{ij} + \sigma_{(3)}\theta_i$.

The marginal density of $\mathbf{Y}_i$ is given by the following integrand of the likelihood, $\ell(\cdot)$, weighted by the prior density $g(\cdot)$:

$$h(\mathbf{Y}_i) = \int_{\theta_i} \int_{\theta_{ij}} \ell(\mathbf{Y}_i \mid \theta_{ij}, \theta_i) g(\theta_{ij})\, \mathrm{d}\theta_{ij}\, g(\theta_i)\, \mathrm{d}\theta_i \tag{A2}$$

where $g(\theta_i)$ and $g(\theta_{ij})$ are the univariate standard normal densities. From this point on, for convenience of notation, θ_i will be replaced by $\theta_{(3)}$ and θ_{ij} will be replaced by $\theta_{(2)}$ when it is used within the integral.

Conditional on the level-3 effect θ_i, the responses from the n_i units at level i are independent; hence, the marginal probability can be written as

$$h(\mathbf{Y}_i) = \int_{\theta_{(3)}} \left\{ \prod_{j=1}^{n_i} \int_{\theta_{(2)}} \left(\prod_{k=1}^{n_{ij}} \prod_{c=1}^{C} [\boldsymbol{\Psi}(\gamma_c - z_{ijkc}) - \boldsymbol{\Psi}(\gamma_c - z_{ijk,c-1})]^{d_{ijkc}} \right) g(\theta_{(2)})\, \mathrm{d}\theta_{(2)} \right\} g(\theta_{(3)})\, \mathrm{d}\theta_{(3)} \tag{A3}$$

This marginal probability can be decomposed as

$$h(\mathbf{Y}_i) = \int_{\theta_{(3)}} \ell_i(\theta_{(3)}) g(\theta_{(3)})\, \mathrm{d}(\theta_{(3)}) \tag{A4}$$

where

$$\ell_i(\theta_{(3)}) = \prod_{j=1}^{n_i} h(\mathbf{Y}_{ij}) \tag{A5}$$

$$h(\mathbf{Y}_{ij}) = \int_{\theta_{(2)}} \ell_{ij}(\theta) g(\theta_{(2)})\, \mathrm{d}(\theta_{(2)}) \tag{A6}$$

and

$$\ell_{ij}(\theta) = \prod_{k=1}^{n_{ij}} \prod_{c=1}^{C} [\boldsymbol{\Psi}(\gamma_c - z_{ijkc}) - \boldsymbol{\Psi}(\gamma_{c-1} - z_{ijkc})]^{d_{ijkc}} \tag{A7}$$

As a result, differentiating the marginal log-likelihood from the n level-3 units, $\log L = \sum_i^n \log h(\mathbf{Y}_i)$, with respect to an arbitrary parameter vector $\boldsymbol{\eta}$ yields

$$\begin{aligned} \frac{\partial \log L}{\partial \boldsymbol{\eta}} &= \sum_{i=1}^{N} h^{-1}(\mathbf{Y}_i) \frac{\partial h(\mathbf{Y}_i)}{\partial \boldsymbol{\eta}} \\ &= \sum_{i=1}^{N} h^{-1}(\mathbf{Y}_i) \int_{\theta_{(3)}} \ell_i(\theta_{(3)}) \frac{\partial \log \ell_i(\theta_{(3)})}{\partial \boldsymbol{\eta}} g(\theta_{(3)})\, \mathrm{d}(\theta_{(3)}) \end{aligned} \tag{A8}$$

where

$$\frac{\partial \log \ell_i(\theta_{(3)})}{\partial \boldsymbol{\eta}} = \sum_{j=1}^{n_i} h^{-1}(\mathbf{Y}_{ij}) \int_{\theta_{(2)}} \ell_{ij}(\theta) \frac{\partial \log \ell_{ij}(\theta)}{\partial \boldsymbol{\eta}} g(\theta_{(2)})\, \mathrm{d}(\theta_{(2)}) \tag{A9}$$

Thus, only $\partial \log \ell_{ij}(\theta)/\partial \boldsymbol{\eta}$ needs to be separately obtained for each parameter or parameter vector. For the parameters that do not vary with c, namely, $\sigma_{(2)}$, $\sigma_{(3)}$, and $\boldsymbol{\beta}$, we get:

$$\frac{\partial \log \ell_{ij}(\theta)}{\partial \boldsymbol{\eta}} = \sum_{k=1}^{n_i} \sum_{c=1}^{C} d_{ijkc} \frac{\psi(\gamma_{c-1} - z_{ijkc}) - \psi(\gamma_c - z_{ijkc})}{\boldsymbol{\Psi}(\gamma_c - z_{ijkc}) - \boldsymbol{\Psi}(\gamma_{c-1} - z_{ijkc})} \frac{\partial z_{ijkc}}{\partial \boldsymbol{\eta}} \tag{A10}$$

with,

$$\frac{\partial z_{ijkc}}{\partial \boldsymbol{\beta}} = \boldsymbol{x}_{ijk}, \quad \frac{\partial z_{ijkc}}{\partial \sigma_{(2)}} = \theta_{(2)} \quad \text{and} \quad \frac{\partial z_{ijkc}}{\partial \sigma_{(3)}} = \theta_{(3)}$$

For the threshold parameters, we get for a particular $\gamma_{c'}$,

$$\frac{\partial \log \ell_{ij}(\theta)}{\partial \gamma_{c'}} = \sum_{k=1}^{n_i} \sum_{c=1}^{C} d_{ijkc} \frac{\psi(\gamma_c - z_{ijkc})\delta_{c,c'} - \psi(\gamma_{c-1} - z_{ijkc})\delta_{c-1,c'}}{\boldsymbol{\Psi}(\gamma_c - z_{ijkc}) - \boldsymbol{\Psi}(\gamma_{c-1} - z_{ijkc})} \tag{A11}$$

with

$$\delta_{c,c'} = \begin{cases} 1 & \text{if } c = c' \\ 0 & \text{if } c \neq c' \end{cases}$$

Finally, for the threshold interaction parameters $\boldsymbol{\alpha}_c$, the form of the derivative is the same as in (A11), albeit with $-\partial z_{ijkc}/\partial \boldsymbol{\alpha}_c = -\boldsymbol{u}_{ijk}$ additionally as a product term in the summation.

Note that for the logistic formulation, $\psi(\cdot) = \mathbf{\Psi}(\cdot)(1 - \mathbf{\Psi}(\cdot))$. For a probit regression formulation, the logistic function $\mathbf{\Psi}(\cdot)$ is replaced by the normal response function $\Phi(\cdot)$, and $\psi(\cdot)$ is replaced by the standard normal density function $\phi(\cdot)$ in the above equations.

ACKNOWLEDGEMENTS

The authors thank Dr Brian Flay for the use of the TVSFP data set and Dr Linda Teri, Dr Leon Thal and Dr Ronald Thomas for the use of the Alzheimer's disease data from the Agitation Study, a clinical trial conducted by the Alzheimer's Disease Cooperative Study. The second author was supported in this work by National Institutes of Mental Health Grant MH56146.

REFERENCES

1. Hedeker D, Gibbons RD. A random-effects ordinal regression model for multilevel analysis. *Biometrics* 1994; **50**:933–944.
2. Gibbons RD, Hedeker D. Random-effects probit and logistic regression models for three-level data. *Biometrics* 1997; **53**:1527–1537.
3. Laird NM, Ware JH. Random effects models for longitudinal data. *Biometrics* 1982; **38**:963–974.
4. Liang K-Y, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika* 1986; **73**:13–22.
5. Goldstein H. *Multilevel Models in Educational and Social Research*. Oxford University Press: New York, 1987.
6. Bock RD. The discrete Bayesian. In *Modern Advances in Psychometric Research*, Wainer H, Messick S (eds). Hillsdale: Earlbaum, 1983; 103–115.
7. Stiratelli MH, Laird NM, Ware JH. Random-effects models for serial observations with binary response. *Biometrics* 1984; **40**:961–971.
8. Wong GY, Mason WM. The hierarchical logistic regression model for multilevel analysis. *Journal of the American Statistical Association* 1985; **80**:513–524.
9. Gibbons RD, Bock RD. Trends in correlated proportions. *Psychometrika* 1987; **52**:113–124.
10. Goldstein H. Nonlinear multilevel models, with an application to discrete response data. *Biometrika* 1991; **78**:45–51.
11. Longford NT. Logistic regression with random coefficients. *Computational Statistics Data Analysis* 1994; **17**:1–15.
12. Qaqish BF, Liang K-Y. Marginal models for correlated binary responses with multiple classes and multiple levels of nesting. *Biometrics* 1992; **48**:939–950.
13. Breslow NE, Clayton DG. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* 1993; **88**:9–35.
14. Ten Have TR, Kunselman AR, Tran L. A comparison of mixed effects logistic regression models for binary response data with two nested levels of clustering. *Statistics in Medicine* 1999; **18**:947–960.
15. Harville DA, Mee RW. A mixed-model procedure for analyzing ordered categorical data. *Biometrics* 1984; **40**:393–408.
16. Ezzet F, Whitehead J. A random effects model for ordinal responses from a crossover trial. *Statistics in Medicine* 1991; **10**:901–907.
17. Hedeker D, Mermelstein RJ. A multilevel thresholds of change model for analysis of stages of change data. *Multivariate Behavioral Research* 1998; **33**(4):427–455.
18. Yang M. Multinomial regression. In *Multilevel Modelling of Health Sciences*, Leyland AH, Goldstein H (eds). Wiley: New York, 2001; 107–125
19. Rabe-Hesketh S, Pickles A, Skrondal A. *GLAAM Manual. Technical Report 2000/01*. Department of Biostatistics and Computing Institute of Psychiatry, King's College, University of London, 2001.
20. Snijders T, Bosker R. *Multilevel Analysis*. Sage: Thousand Oaks, CA, 1999.
21. McCullagh P. Regression models for ordinal data (with discussion). *Journal of the Royal Statistical Society, Series B* 1980; **42**:109–142.
22. Peterson B, Harrell FE. Partial proportional odds models for ordinal response variables. *Applied Statistics* 1990; **39**:205–217.
23. Terza JV. Ordinal probit: a generalization. *Communications in Statistical Theory and Methods* 1985; **14**:1–11.
24. Bock RD, Lieberman M. Fitting a response model for *n* dichotomously scored items. *Psychometrika* 1970; **35**:443–459.
25. Stroud AH, Sechrest D. *Gaussian Quadrature Formulas*. Prentice Hall: Englewood Cliffs, NJ, 1966.

26. Bock RD, Atkin M. Marginal maximum likelihood estimation of item parameters: an application of the EM algorithm. *Psychometrika* 1981; **46**:443–459.
27. Hedeker D, Gibbons RD. MIXOR: a computer program for mixed-effects ordinal regression analysis. *Computer Methods and Programs in Biomedicine* 1996; **49**:157–176.
28. Flay B, Brannon B, Johnson C, Hansen W, Ulene A, Whitney-Saltiel D, Gleason L, Sussman S, Gavin M, Kimarie G, Sobol D, Spiegel D. The television school and family smoking and cessation project: I. Theoretical basis and program development. *Preventive Medicine* 1988; **17**:585–607.
29. Bryk A, Raudenbush SW. *Hierarchical Linear Models*: *Applications and Data Analysis Methods*. Sage: Newbury Park.
30. Teri L, Logsdon RG, Peskind E, Raskind M, Weiner MF, Tractenberg RE, Foster NL, Schneider LS, Sano M, Whitehouse P, Tariot P, Mellow AM, Auchus AP, Grundman M, Thomas RG, Schafer K, Thal LJ. Treatment of agitation in AD: a randomized, placebo-controlled clinical trial. *Neurology* 2000; 1271–1278.
31. Tanner MA. *Tools for Statistical Inference*: *Methods for the Exploration of Posterior Distributions and Likelihood Functions* (3rd edn). Springer: New York, 1996.
32. Rabe-Hesketh S, Skrondal A, Pickles A. Reliable estimation of generalized linear mixed models using adaptive quadrature. *The Stata Journal* 2002; 1–21.
33. Hardin JW, Hilbe JM. *Generalized Estimating Equations*. Chapman & Hall: Boca Raton, FL, 2003.
34. Heagerty PJ, Zeger SL. Marginalized multilevel models and likelihood inference. *Statistical Science* 2000; **15**: 1–26.
35. Kenward MG, Lasaffre E, Molenberghs G. An application of maximum likelihood and generalized estimating equations to the analysis of ordinal data from a longitudinal study with cases missing at random. *Biometrics* 1994; **50**:945–953.
36. Zeger SL, Liang KY, Albert PA. Models for longitudinal data: a generalized estimating equation approach. *Biometrics* 1988; **44**:1049–1060.
37. Neuhaus JM, Klabfleisch JD, Hauck WW. A comparison of cluster-specific and population-averaged approaches for analyzing correlated binary data. *International Statistical Review* 1991; **59**:25–35.

