Biostatistics (2001), **2**, 4, *pp.* 397–416
Printed in Great Britain

A case study on the choice, interpretation and checking of multilevel models for longitudinal binary outcomes

JOHN B. CARLIN*, RORY WOLFE

Clinical Epidemiology and Biostatistics Unit, Murdoch Children's Research Institute and University of Melbourne Department of Paediatrics, Parkville, VIC 3052, Australia

jbcarlin@unimelb.edu.au

C. HENDRICKS BROWN

Department of Epidemiology & Biostatistics, University of South Florida, Tampa, FL, USA

ANDREW GELMAN

Department of Statistics, Columbia University, New York, USA

SUMMARY

Recent advances in statistical software have led to the rapid diffusion of new methods for modelling longitudinal data. Multilevel (also known as hierarchical or random effects) models for binary outcomes have generally been based on a logistic–normal specification, by analogy with earlier work for normally distributed data. The appropriate application and interpretation of these models remains somewhat unclear, especially when compared with the computationally more straightforward semiparametric or 'marginal' modelling (GEE) approaches. In this paper we pose two interrelated questions. First, what limits should be placed on the interpretation of the coefficients and inferences derived from random-effect models involving binary outcomes? Second, what diagnostic checks are appropriate for evaluating whether such random-effect models provide adequate fits to the data? We address these questions by means of an extended case study using data on adolescent smoking from a large cohort study. Bayesian estimation methods are used to fit a discrete-mixture alternative to the standard logistic–normal model, and posterior predictive checking is used to assess model fit. Surprising parallels in the parameter estimates from the logistic–normal and mixture models are described and used to question the interpretability of the so-called 'subject-specific' regression coefficients from the standard multilevel approach. Posterior predictive checks suggest a serious lack of fit of both multilevel models. The results do not provide final answers to the two questions posed, but we expect that lessons learned from the case study will provide general guidance for further investigation of these important issues.

Keywords: Bayesian methods; Hierarchical models; Interpretation of models; Longitudinal analysis; Model diagnosis; Posterior predictive check; Subject-specific effects.

*To whom correspondence should be addressed

1. INTRODUCTION

Building statistical models to address questions about change over time in binary or ordinal outcomes is substantially more difficult than for continuous outcomes. A large literature has grown up on the use of multilevel or hierarchical models for normally distributed outcomes, including longitudinal data in biostatistics (Laird and Ware, 1982) and numerous applications to clustered data structures in educational and sociological research (Bryk and Raudenbush, 1992; Longford, 1993; Goldstein, 1995). A slightly different approach to an overlapping class of longitudinal models is gaining increasing prominence, under the name of 'latent growth modelling', in the emerging field of prevention science (Muthén, 1993; Muthén and Curran, 1997). With normally distributed outcomes and a linear model, the 'marginal effects', or average differences for subpopulations defined by differing covariate values, are the same as 'subject-specific effects' or expected differences for individual subjects under different covariate values. This is, however, not true for binary outcome measures, for which a nonlinear link function is needed to provide a realistic connection between a linear predictor and the mean of the observable variable (the probability of the outcome) (Neuhaus, 1992; Diggle *et al.*, 1994).

In addition to the greater complexity of parameter interpretation, there are also well known computational problems in fitting hierarchical or multilevel models to a binary outcome, since the (marginal) likelihood function cannot be evaluated in a closed form. These problems are especially acute when there is minimal replication in the data at the first level of variation: that is, occasions within individuals in longitudinal data. In the typical longitudinal epidemiological study there are rarely as many as ten occasions of measurement. In recent years the computational challenges of fitting complex models to binary data have been largely overcome; in particular, software has become available to fit the logistic–normal 'random-effects' model (Hedeker and Gibbons, 1996; Spiegelhalter *et al.*, 1996; SAS, 1999; StataCorp, 1999), and appears to be gaining widespread use. On the other hand, the slightly older estimating-equation ('GEE' and related) methods, which directly model the first-order structure or marginal distribution of the outcome (Liang and Zeger, 1986; Diggle *et al.*, 1994; Carlin *et al.*, 1999), also remain popular.

Despite the appearance of several books and papers reviewing the differences between these modelling approaches (Neuhaus *et al.*, 1991; Neuhaus, 1992; Diggle *et al.*, 1994; Hu *et al.*, 1998), a number of questions about the strengths and weaknesses of the alternative approaches remain unanswered. For example, is it realistic to assume, as is standard in multilevel modelling, that heterogeneity between subjects can be adequately represented by a normally distributed random intercept (and/or slope)? What implications does this assumption have for substantive conclusions that may be drawn from estimated model parameters, and to what extent can the assumption be checked from the data?

In this paper we illustrate and discuss these questions in the context of a specific applied problem, by comparing the results of fitting a standard logistic–normal model with those of fitting an alternative mixture model. The semiparametric marginal modelling approach is also used for comparison. Our new model explicitly assumes that a fraction of the population is 'immune' to exhibiting the outcome of interest. This is in contrast to the logistic–normal specification, which assumes a normal distribution of the unobserved individual-level random intercept, thus allowing the baseline probability to vary smoothly across the whole population.

All the models explored in this paper are ultimately too simple to address the real complexity of questions that might be asked about the underlying epidemiology of smoking uptake in teenagers, the particular application that we consider. In particular, one might well wish to consider more elaborate mixture models, or models where individual-specific slopes (as well as intercepts) are allowed to vary across subjects. It may be argued, however, that a thorough understanding of the relatively simple models used here should be achieved before introducing greater complexity.

In the next section we begin by briefly describing the epidemiological study that motivated this work, and outline some substantive research questions relating to the uptake of regular smoking among the

participants. In Section 3, we introduce a range of models that might be used to draw conclusions about the rate of change over time and the effect of important covariates, in relation to the binary outcome variable representing self-reported regular smoking. In particular, we define the three models that will be fitted to the data: a semiparametric (marginal means) model, a standard multilevel logistic–normal model, and the new discrete mixture model. Methods of parameter estimation for these models are discussed in Section 4, along with approaches to model checking, in particular using the method of posterior predictive distributions. Section 5 presents the results of fitting each of the models and the final section concludes with a detailed discussion of the implications of the results.

2. THE DATA AND RESEARCH QUESTIONS

The Victorian Adolescent Health Cohort Study (VAHCS) was a longitudinal study of teenagers conducted between August 1992 and July 1995 in the state of Victoria, Australia. A sample of participants was identified in an initial cross-sectional survey using a two-stage sampling procedure (Patton *et al.*, 1998a). At this initial survey, students had a mean age of 14.9 years (standard deviation 0.5). At the second wave of data collection, six months later, the sample was augmented by selecting a second intact class from the same age group at each participating school. A total sample of 2032 students was identified from 44 schools. Subsequent waves of data were collected at six-monthly intervals over 3 years, resulting in an intended six time points for the primary cohort and five times for the second sample. The average age at the end of the study was 17.4 years. At each wave, a questionnaire was administered by laptop computer (Hibbert *et al.*, 1996), or where necessary (from wave 4 on) by telephone. The questionnaire was presented as dealing with important health issues for adolescents and included questions on a wide range of health risk behaviours and mental health.

A principal focus of the analysis has been on patterns of cigarette smoking, since this is arguably the single behaviour most likely to impair the long-term health of adolescents. On each occasion, a subject's self-assessed smoking status was determined using a 7-day retrospective diary, and a subject was categorized as a 'daily smoker' if they reported smoking on at least 6 days of the previous week. Demographic and family variables were assessed at entry to the study. For each subject an indicator of parental smoking was based on whether at least one parent was reported to be a daily smoker. Preliminary analyses suggested that the most important explanatory variables for describing differences in daily smoking were age, sex and parental smoking.

As in all cohort studies of this kind, there were problems of sample attrition and missing data. Of the total sample of 2032 students, 1209 students responded at every time period in which they were included in the study. At the other extreme, only 85 failed to respond at all time points, and these were omitted from further consideration. Based on the intended sample, the participation rates were, at each wave respectively: 87, 85, 84, 80, 78 and 75%, indicating some potential for response bias, especially in later waves, despite a generally high rate of follow-up for this type of study. For the present analysis we omitted 187 study participants who did not provide a response on the parental smoking covariate, leaving a dataset of 1760 subjects. There remained substantial amounts of missing data, in the form of missing occasions within subjects, and preliminary analysis showed that the pattern of missing outcomes could not be regarded as completely random.

In a previous methodological review, we have described in detail the primary methods that were used in subject-matter papers to analyse the prevalence and incidence of regular smoking over the course of the study (Carlin *et al.*, 1999). These analyses adopted firstly a marginal modelling perspective, focusing entirely on population-averaged differences in smoking rates with respect to age, sex and parental smoking, and we shall revisit this approach as a point of comparison in the present paper. The second set of previous analyses adopted a transitional modelling approach, investigating factors associated with

the uptake of regular smoking for the first time (Patton *et al.*, 1998a). This approach to longitudinal analysis has considerable practical appeal because of its connections with survival analysis and the natural interpretation of incidence rates and their ratios. But it also has drawbacks. In particular, the method is prone to problems of measurement error, in the sense that great weight is placed on the *first* reported time of smoking at a particular level. It will not be discussed further in this paper.

Methods based on multilevel (random effects or 'mixed') models have recently become popular in longitudinal analysis because they allow the modelling of trends over time at the individual level, by incorporating a subject-specific intercept (and possibly also slope with time) into a regression model for the outcome. In the present context such a modelling approach may offer the possibility of exploring the sources of individual-to-individual variability in the propensity to take up smoking, in a framework that allows for measurement error in the outcome by modelling the level of smoking risk as a latent variable. Previous discussions of multilevel models for binary outcomes (Neuhaus, 1992; Diggle *et al.*, 1994; Hu *et al.*, 1998) have suggested that they offer the scope to answer questions such as 'what is the rate of increase in the (log) odds of regular smoking for an individual, conditional on major covariates and on their latent underlying 'propensity' to exhibit the outcome?' The latent propensity is reflected in the random intercept of the logistic–normal multilevel model. We next describe the specific models that we fitted to the smoking data, and briefly discuss methods of estimation, before returning to consider the models' ability or otherwise to address these questions. For simplicity, we restricted attention to the effect of a small number of covariates, in particular time or wave of study, sex, and parental smoking. The resulting analyses are not therefore intended to capture the full complexity of these data, but to illustrate important issues with respect to the choice of models.

3. MODELS

We consider a binary outcome variable, y_{ij}, which in our example represents self-reported daily smoking (1 = yes, 0 = no), for the ith subject ($i = 1, \ldots, n$) on the jth occasion of measurement ($j = 1, \ldots, J$). In this study there were up to six planned occasions of measurement for each participant, so $J = 6$, although for the majority of subjects at least one value was missing, either by design (in that they were not included for the first wave) or by 'happenstance' in that they failed to complete the survey on one or more occasions.

3.1 *Semiparametric marginal model*

As a point of reference for the subsequent discussion of multilevel models, we first consider a 'marginal model'. This is not in fact a full probability model for the outcome, because it only specifies the form of the mean or expected value of y_{ij}, which we denote π_{ij}. Assumptions about the second-order structure or within-individual correlation in response are typically made when using the GEE approach to estimation (see next section), but these are intended to increase the efficiency of estimation, with less emphasis on making inferences about the higher-level structure. In our example we consider the specification

$$\pi_{ij} = \Pr(y_{ij} = 1) = \text{logit}^{-1}(\beta_0^* + \beta_p^* p_i + \beta_g^* g_i + \beta_{w(m)}^*(1 - g_i)w_{ij} + \beta_{w(f)}^* g_i w_{ij}) \qquad (1)$$

where $\text{logit}(\pi) = \log(\pi/(1-\pi))$ and g_i represents sex (coded 0/1 for males/females), p_i parental smoking (0/1 for no/yes) and w_{ij} wave of study, coded $0, \ldots, 5$, so each 1 unit interval on this variable represents 6 months of age, with the origin corresponding to wave 1. The interaction effect, allowing a different trend with wave for males and females, is coded in a slightly nonstandard manner, so that two wave (or time) coefficients are estimated directly, one for each sex while β_g^* represents the difference in intercepts between the sexes.

3.2 *Logistic–normal model*

To construct a model that fully represents the variation in the data, a method of incorporating correlation between repeated responses on the same individual is clearly required. By analogy with hierarchical normal–normal models, a natural extension of the logistic regression model is to incorporate a subject-specific random intercept, resulting in the following two-level model for the smoking outcome:

$$\begin{aligned} \pi_{ij}|\alpha_i &= \text{logit}^{-1}(\beta_0 + \alpha_i + \beta_p p_i + \beta_g g_i + \beta_{w(m)}(1-g_i)w_{ij} + \beta_{w(f)} g_i w_{ij}) \\ \alpha_i &\sim \text{N}(0, \sigma^2). \end{aligned} \tag{2}$$

This model has been proposed and studied by many authors, often as a specific case of the so-called 'generalized linear mixed model' (GLMM). A great deal of discussion has been given to the fact that the β parameters in (2) are different from the parallel β^* parameters in (1). It has been shown that assuming model (2) holds, the marginal probabilities, $\pi_{ij} = \int (\pi_{ij}|\alpha_i)\, p(\alpha_i)\, \text{d}\alpha_i$ are related to the covariates by an approximate logistic regression of the form (1) where the β^* parameters are attenuated (nearer to zero) in comparison to the corresponding β, by an amount that increases with the variance σ^2 (Zeger *et al.*, 1988; Neuhaus *et al.*, 1991). Although the regression coefficients of the marginal model are clearly interpretable as representing average population differences in the log odds of the outcome (in epidemiological terms, the logs of prevalence odds ratios), the interpretation of the 'subject-specific' parameters β is less clearcut, and we shall return to this issue in Sections 5 and 6. We shall be primarily concerned, however, with the related question of whether model (2) provides a reasonable way of representing subject heterogeneity in data such as ours.

3.3 *Discrete mixture model*

We shall see in Section 5 that only about one-quarter of the study cohort reported regular smoking at any of the six occasions, and clearly there are many teenagers who will never take up regular smoking. This suggests that rather than assuming that the underlying (logit) risk of smoking for each individual follows a normal distribution, as in the logistic–normal model, it might be preferable to suppose that there is a subgroup of the population with extremely low or zero probability of regular smoking. If a substantial proportion of the population of adolescents are 'immune' from becoming smokers, their subject-specific log odds would be negative infinity, and it is impossible for a normal model to capture such a distribution. We therefore propose a mixture model, in which each member of the population may be either 'immune' or 'susceptible'. We use the notation $S_i = 1$ if subject i is in the susceptible group, and $S_i = 0$ if in the immune group. If susceptible, the probability of smoking is modelled by the same logistic–normal specification as previously applied to the whole population. The logistic–normal and discrete mixture models are illustrated in Figure 1.

In this mixture model, careful consideration needs to be given to the appropriate inclusion of covariate effects. In view of the likely strong impact of subject-level covariates (such as parental smoking) on the likelihood of being susceptible, it is important to construct a model for susceptibility that incorporates these covariates. In general, if we write subject-level covariates as $\mathbf{z}_i$ and occasion-level covariates as $\mathbf{x}_{ij}$, the proposed mixture model may be written as follows:

$$\begin{aligned} \pi_{ij} &= 0 && \text{if } S_i = 0 \\ \pi_{ij}|\gamma_i &= \text{logit}^{-1}(\eta_0 + \gamma_i + \boldsymbol{\eta}^T \mathbf{x}_{ij}) && \text{if } S_i = 1 \\ \text{with } \gamma_i &\sim \text{N}(0, \tau^2) \end{aligned} \tag{3}$$

where, at effectively a third level in the model,

$$\Pr(S_i = 1) = \text{logit}^{-1}(\phi_0 + \boldsymbol{\phi}^T \mathbf{z}_i). \tag{4}$$

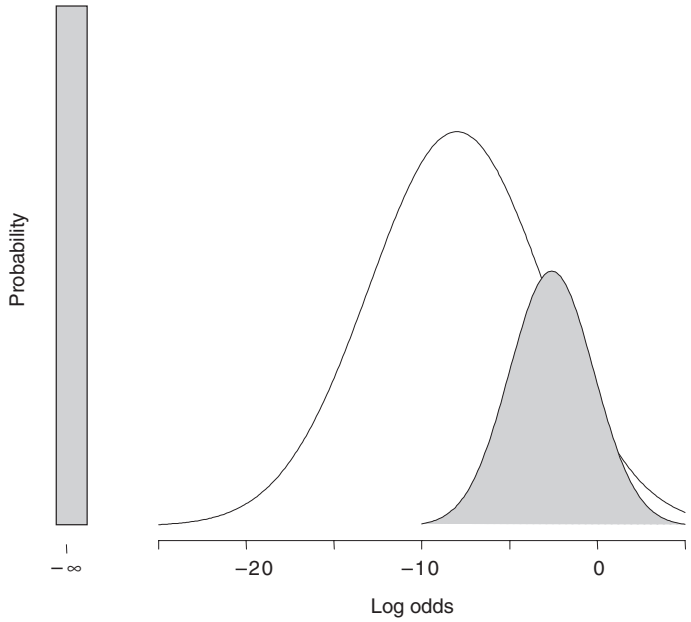

Fig. 1. Illustration of logistic–normal (unshaded normal curve) and discrete mixture (shaded bar and curve) models, as probability distributions for the log odds of smoking.

This model was applied to our data with $\mathbf{z}_i$ consisting of sex (g_i) and parental smoking (p_i), while $\mathbf{x}_{ij}$ included g_i and p_i also, along with the sex-specific time trends as in (2).

4. ESTIMATION AND MODEL CHECKING

Since the semiparametric marginal modelling approach does not specify a full probability model for the data, the traditional methods of statistical inference based on the likelihood function (whether regarded as 'pure likelihood' or Bayesian) are not available. Nevertheless, recent research has shown that a number of so-called quasi-likelihood or estimating equation methods have good repeated-sampling properties for estimating marginal model parameters (Liang and Zeger, 1986; Zeger *et al.*, 1988). Relatively minor differences in estimates may arise depending on how the estimating equations are weighted, in particular within the generalized estimating equation (GEE) framework, on the choice of 'working correlation' structure that is used to capture within-subject correlations. Greater efficiency (smaller standard errors) should be obtained when the working correlation matrix correctly reflects the actual correlation structure in the data. More substantial differences in point estimates may arise if weighting is applied to adjust for missing data (Carlin *et al.*, 1999). For the practitioner, these choices result in a wide array of options both among and within software packages, and in this paper we simply illustrate a range of results for our example, using the packages Stata (StataCorp, 1999) and MLwiN (Goldstein *et al.*, 1998).

Turning to the logistic–normal model, this provides a likelihood function on which to base estimation, but the likelihood is not easy to compute since it involves a high-dimensional integral over the distribution of the n random effects, α_i. Three approaches have been proposed for dealing with this integral. First, it is possible to use the Laplace approximation to derive a so-called penalized quasi-likelihood (PQL), which in certain situations provides a good approximation to the full likelihood (Goldstein, 1991; Breslow

and Clayton, 1993; Wolfinger, 1993). This has been programmed into a SAS macro (GLIMMIX) as well as the package MLwiN. Unfortunately, the approximation works very poorly for data such as that of our example, where there is considerable inter-subject heterogeneity and only a small number of binary responses per subject (Breslow and Clayton, 1993; Rodriguez and Goldman, 1995; Goldstein and Rasbash, 1996). The second approach is to use Gauss–Hermite quadrature to evaluate the likelihood using numerical integration; this has been implemented in the freeware package MIXOR (Hedeker *et al.*, 1994; Hedeker and Gibbons, 1996) and is now also available for models involving one random effect in Stata. Finally, the general Bayesian approach of iterative simulation or Markov chain Monte Carlo (MCMC) is well-suited to this problem, and essentially reproduces a likelihood analysis if diffuse prior distributions are used for the parameters of the random effects distribution (in our example, the variance σ^2). We used the BUGS package (Spiegelhalter *et al.*, 1996) for applying this method.

The discrete mixture model has a complicated likelihood function that not only involves integration over a normal random effects distribution, but also summation over the two possible values of S_i for each individual. Again, however, this model may be handled relatively simply by MCMC, using Gibbs sampling, and we fit it using BUGS. In preliminary estimation runs using very diffuse prior distributions, some instability was observed in the iterative estimation process, where a chain occasionally attempted to switch from the two-component mixture to the original one-component logistic–normal model (corresponding to a singularity in the likelihood function). This led us to incorporate slightly more informative prior distributions (see Section 5). BUGS code used for both models may be obtained from the first author.

4.1 *Target inferences under different models*

There is a long tradition in applied statistics of fitting regression models and reporting coefficient estimates, with appropriate confidence intervals or P-values. With normal-based hierarchical models this tradition appears to be more or less 'safe' since the linearity of structure of the models means that regression coefficients invariably have a fairly direct interpretation in terms of mean differences in outcome measures. Under suitable assumptions, such differences may sometimes be interpreted as causal effects (Rubin, 1974; Holland, 1986). With nonlinear multilevel models, the interpretation of regression coefficients is more difficult.

Consider the case of binary regression with log odds regression parameters (i.e. logistic regression). The attraction of the usual regression specification is that it implies that differences in the transformed outcome parameter—the log odds—associated with differences in the value of a particular covariate, have identical meaning regardless of the value of other covariates (assuming no interaction effects). With a hierarchical model, however, these coefficients are also conditional on the unobservable random effect(s), so any direct interpretation is dependent on the assumption that the effect of interest is constant conditional on an unobservable random component, an assumption that cannot be directly assessed from the data. We shall return to this point below and illustrate in our example the lack of interpretability of the 'subject-specific' coefficients.

One arena in which the semiparametric and multilevel approaches may be directly compared is in the estimation of population mean quantities and their differences. Such quantities are attractive targets of inference since they have a clear meaning in the population—in particular, they could be estimated unambiguously given infinite-sized samples, without the need for any modelling assumptions. One can always derive an implied marginal structure from a multilevel model, since the unobservable components ('random effects') can be integrated out of the multilevel model specification to give marginal probabilities or means at the observable level. Of course, the semiparametric modelling approach estimates marginal probabilities directly, so a natural question to ask is whether improvements in bias or efficiency for estimation of these target quantities is possible by investing in the more complex hierarchical models.

We investigate this issue in our case study by comparing estimated marginal effects from the multilevel models with those estimated directly using the semiparametric approach. In particular, we calculate estimates of marginal differences in probabilities under the three modelling approaches. These marginal differences may be interpreted as differences in probability or 'risk' that are 'attributable' to changes in a single covariate (Rothman and Greenland, 1998). But the nonlinear structure of the model means that they must also depend on the values of the other covariates included in the regression model. For simplicity, we chose to fix the other covariates at certain values of substantive interest; in particular, we estimated the difference in the probability of regular smoking:

- between a subject whose parents smoked and whose parents did not smoke, for males and females separately, at wave 5 (age $\approx$17), and
- between a subject at the inception of the cohort (wave 1) and at wave 5 (a 2 year interval), again for males and females separately, and holding parental smoking at 0 (no smoking).

Calculation of the estimated marginal probabilities (predictive means) under the multilevel models was performed within the MCMC iterations, by simulating 1000 values from the normal random-effect distribution at each drawn value of the variance parameter (σ^2 for the logistic–normal and τ^2 for the mixture). The drawn random effects were used in conjunction with the fixed covariate values to generate 1000 values of the probability of interest, the mean of which represents (at convergence) a draw from the posterior distribution of the marginal probability. This approach automatically produces an accompanying estimate of uncertainty, appropriately accounting for the uncertainty of estimation of the variance parameters.

4.2 *Model checking using posterior predictive distributions*

As models become more complex, it becomes more important to assess whether there are substantively relevant ways in which they fail to fit the data (Gelman *et al.*, 1995). A simple and effective tool for model checking, based on comparing the observed data with the type of data that would be 'typical' under the assumed model, is the method of posterior predictive check distributions (Gelman *et al.*, 1996). Briefly, the method requires one to choose one or more test statistics, T, which are (usually scalar) quantities that reflect features of the data and possibly the unknown parameters that one would expect the model to represent faithfully. Then we compare the posterior predictive distribution of this statistic with its posterior distribution based on the observed y value. If the test statistic does not depend on the unknown parameters, the concept simplifies to a comparison between the observed value of a statistic, $T(y)$, and the posterior predictive distribution of the same quantity, which may be regarded as the posterior distribution of $T(y^{\text{rep}})$, where y^{rep} denotes a 'replicated' version of the data y. A simple summary of this comparison is provided by the posterior predictive P-value: $\Pr(T(y^{\text{rep}}) > T(y) \mid y)$. The computation of posterior predictive distributions and corresponding P-values requires integration over the posterior distribution of the unknown parameters in the model, but this is simple to perform numerically using posterior simulation draws within the iterations of a MCMC estimation algorithm. The method is essentially classical goodness-of-fit testing with the Bayesian machinery used to average over uncertainty in the parameter estimates—which can be important in the case of random effects, since these are individually not well estimated from the data.

To apply this method to our example, it remains to specify suitable test statistics. We consider that a sensible requirement of an adequate model is that it should reproduce some of the simplest global features of the data, such as the proportions of subjects who exhibit particular patterns of change over time. Also, there is no need to test aspects of the data that are fitted directly by the model—the choice of check statistics should encompass important substantive features of the data that the model is not designed to represent directly. To this end, we defined the following three test statistics:

(1) T_n = proportion of subjects who never report regular smoking,
(2) T_a = proportion of subjects who report regular smoking on all occasions, and
(3) T_i = proportion of subjects who are incident, non-remitting smokers; in other words who report one or more occasions of not smoking, followed by one or more occasions of regular smoking, without alternating more than once between the '0' and '1' states.

These definitions were applied to all subjects regardless of their pattern of missing data (occasions where outcome was missing were ignored); so, for example, an incident smoker in the sense underlying T_i could have just two occasions of measurement, or could have all six waves.

A further set of posterior predictive checks was based on the marginal odds ratios between successive time points:

$$T_{OR}^{jk} = \frac{\Pr(y_{ij} = 1, y_{ik} = 1)\Pr(y_{ij} = 0, y_{ik} = 0)}{\Pr(y_{ij} = 1, y_{ik} = 0)\Pr(y_{ij} = 0, y_{ik} = 1)} \tag{5}$$

(for $j \neq k$ in $1 \ldots J$), where these are simple empirical measures of pairwise association between the binary outcome at different times, obtained by substituting crude proportions (over available pairs of nonmissing data values) for the joint probabilities. We expected that most models for dependent binary outcomes should be able to reproduce these second-order pairwise associations reasonably well, but it might be of interest to examine patterns across the lag structure: for example, would odds ratios between outcomes separated by four or five time points be as well fitted as odds ratios corresponding to shorter lags?

5. Results

Marginal smoking frequencies are summarized in Figure 2, which shows that females had a higher linear trend in the prevalence of regular smoking than males (using available data at each wave). In Table 1, descriptive information about missing data patterns is provided, along with a breakdown of subjects according to their overall pattern of response, in parallel with the three posterior predictive check statistics defined in the last section. In particular, we see that 77% of individuals never reported smoking at the daily level at any wave of the study. Table 2 displays the pairwise odds ratios between waves, and indicates that the strength of association tends to decrease with increasing separation between time points. One might expect this feature would be difficult to reproduce with the relatively simple models examined here, since the random intercept model implies approximately equal association between all pairs of time points.

Parameter estimates from three different semiparametric approaches to marginal model estimation, all available in well-established statistical packages, are shown in Table 3. The different methods, which correspond to the use of different estimating equations (Carlin *et al.*, 1999) all give generally similar results, and the GEE method might be slightly preferred because of its explicit allowance for intrasubject correlation.

In Table 4 we display parameter estimates under three different approaches to the logistic–normal model. The first method used the so-called 'PQL-1' approximation (Breslow and Clayton, 1993; Goldstein and Rasbash, 1996) available in the MLwiN package and the GLIMMIX macro for SAS (Wolfinger and O'Connell, 1993), and its estimates were very poor, giving values much closer to the marginal models than to the full-likelihood estimates. These problems with PQL when applied to binary data have been discussed by others (Rodriguez and Goldman, 1995); its results with data such as these are meaningless. It should also be noted that the second-order approximation suggested by Goldstein and Rasbash (1996) and also implemented in MLwiN ('PQL-2') did not converge in this problem.

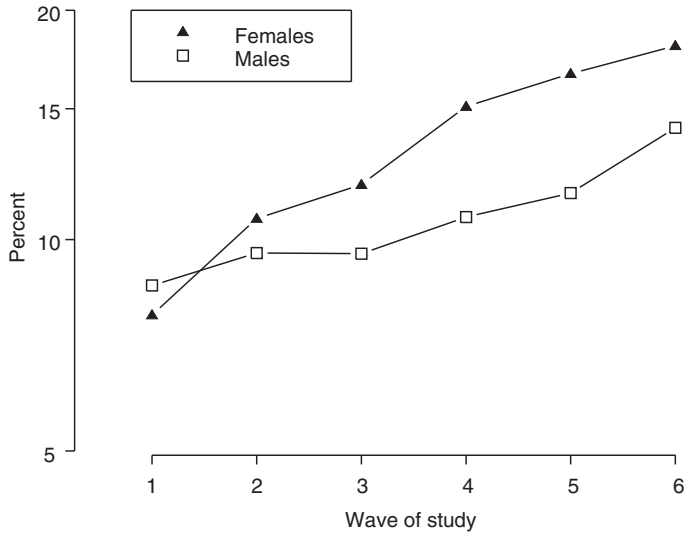

Fig. 2. Prevalence of regular (daily) smoking among participants responding at each wave of the study, for males and females, in logit scale.

Table 1. *Summary statistics on regular smoking outcome in Victorian Adolescent Health Cohort Study—crosstabulation of smoking response pattern by completeness of subject data (with percentage by column)*

Pattern of response	Number of waves completed				Total	
	1–4		5–6			
Never positive	278	(69.3)	1083	(79.7)	1361	(77.3)
Always positive	46	(11.5)	44	(3.2)	90	(5.1)
'Incident'	39	(9.7)	108	(7.9)	147	(8.4)
Other	38	(9.5)	124	(9.1)	162	(9.2)
Total	401	(100.0)	1359	(100.0)	1760	(100.0)

Table 2. *Crude odds ratios* (5) *for association between regular smoking outcome at each pair of study waves*

Wave j	Wave k				
	2	3	4	5	6
1	61	35	43	33	35
2		52	52	41	37
3			65	40	32
4				65	46
5					89

Table 3. *Parameter estimates (with SEs) for semiparametric (marginal) logistic regression model* (1) *for regular smoking*

Parameter		Estimation method		
		ML/RSE[a]	GEE[b]	MQL[c]
Intercept[d]	β_0^*	−2.8	−2.8	−2.8
		(0.14)	(0.13)	(0.13)
Parental smoking	β_p^*	0.93	0.93	0.92
		(0.12)	(0.11)	(0.12)
Sex	β_g^*	0.07	0.09	0.01
		(0.17)	(0.16)	(0.16)
Wave (M)	$\beta_{w(m)}^*$	0.11	0.14	0.13
		(0.03)	(0.03)	(0.02)
Wave (F)	$\beta_{w(f)}^*$	0.18	0.17	0.18
		(0.02)	(0.02)	(0.02)

[a]Maximizing likelihood based on marginal model only, with information sandwich 'robust' SEs, using Stata.
[b]Generalized estimating equations based on unstructured working correlation, using Stata.
[c]Marginal quasi-likelihood (Goldstein, 1991; Breslow and Clayton, 1993), using MLwiN.
[d]At wave 1 (≈15 years).

Table 4. *Parameter estimates (with SEs) for logistic–normal regression model* (2) *for regular smoking*

Parameter		Estimation method		
		PQL[a]	ML[b]	Bayes[c]
Intercept[d]	β_0	−3.7	−7.9	−8.2
		(0.17)	(0.42)	(0.46)
Subject SD	σ	2.0	4.5	5.0
				(0.28)
Parental smoking	β_p	1.11	2.7	2.6
		(0.14)	(0.29)	(0.35)
Sex	β_g	−0.15	−0.27	−0.28
		(0.21)	(0.33)	(0.41)
Wave (M)	$\beta_{w(m)}$	0.20	0.35	0.36
		(0.04)	(0.05)	(0.05)
Wave (F)	$\beta_{w(f)}$	0.30	0.45	0.53
		(0.03)	(0.05)	(0.05)

[a]Penalized quasi-likelihood (Goldstein, 1991; Breslow and Clayton, 1993; Wolfinger and O'Connell, 1993), using MLwiN.
[b]Exact maximum likelihood by Gauss–Hermite (20 point) quadrature, using Stata or MIXOR.
[c]Posterior means and SEs combining five parallel MCMC chains of length 40 000, after burn-in of 20 000, using BUGS.
[d]At wave 1 (≈15 years).

Full-likelihood maximization using Gauss–Hermite numerical quadrature was performed using 12, 16, 20 and 24 point quadrature with the MIXOR program and the 'xtlogit' command in Stata 6.0. Results from the two programs agreed closely, but there was considerable sensitivity of estimates to the number of quadrature points used (which suggests that the component of the integrand from the binomial likelihood is not well behaved). In particular, the estimated coefficient for parental smoking increased by a factor of more than two when the number of quadrature points was increased from 12 to 20. If only one approximation (choice of quadrature points) had been used, this instability would have been difficult to detect, because the approximate SE at each approximation was much smaller than the difference between estimates.

Finally, MCMC was used, as implemented in BUGS, and assuming diffuse prior distributions for all parameters. Convergence was achieved after a burn-in of 10 000 iterations (based on five parallel chains, showing potential scale reduction factors of 1.01 or less (Gelman and Rubin, 1992)) despite high autocorrelation within each chain for some of the parameters, and a high correlation between the intercept parameter and the random-effect variance. We also experimented briefly with the MCMC technology now available in the MLwiN package, and the results were consistent with those from BUGS. Parameter estimates were very similar to those found by the Gauss–Hermite algorithms when using a reasonably large number of quadrature points (20), which provided reassurance that both approaches were converging to the same region of the parameter space.

Examination of the parameter estimates for the logistic–normal model illustrates why the assumption of a normal distribution for the random intercepts may be questionable. In particular, the fitted model for the random effects exhibited a huge variance and a highly negative mean (see Figure 1). The reason is the relative rarity of regular smoking, with 77% of participants never reporting smoking at this level. The wide dispersion of the random intercepts is the model's only way of capturing this feature of the data. Finally, the large between-subject variance is also the reason why the estimated β are so much larger than the corresponding β^* (Neuhaus *et al.*, 1991).

Parameter estimates for the discrete mixture model are displayed in Table 5. These were obtained under the following prior distributions:

$$\phi_0 \sim \mathrm{N}(-1, 1)$$
$$\phi_g \sim \mathrm{N}(0, 1)$$
$$\phi_p \sim \mathrm{N}(0, 1)$$

with diffuse specifications used for the subject-level coefficients, $\boldsymbol{\eta}$, as for the logistic–normal. The rationale for using a standard deviation of 1 in the prior distributions for the parameters of the susceptibility model was that odds ratios greater than $e^2 = 7.4$ are very unlikely. The location of the intercept parameter was set at -1, since $e^{-1}/(1 + e^{-1}) = 0.27$ seemed a reasonable prior estimate for the proportion of susceptibles among males ($g_i = 0$) with non-smoking parents ($p_i = 0$). (These specifications would have the effect of slightly damping any estimates of (positive) association with sex or parental smoking.) Using five chains from dispersed starting points, convergence was apparent after 20 000 iterations (PSRs < 1.02), and the same number of iterations was used as for the logistic–normal model. The estimation was repeated with different prior distributions (for example, changing the intercept specification to $\phi_0 \sim \mathrm{N}(0, 4)$), with only minor changes in the results.

The parameter estimates show firstly that the proportion estimated to be susceptible, among males with nonsmoking parents, was about $e^{-0.78}/(1 + e^{-0.78}) = 0.31$, which was somewhat greater than (and so consistent with) the frequency of 'ever-smokers' in the data. The evidence for a sex effect on susceptibility was weak, but the parental smoking effect appeared to operate at both levels of the model, in predicting susceptibility and in predicting smoking uptake within susceptible individuals.

Table 5. *Parameter estimates for discrete mixture model* (3) *for regular smoking. Values are posterior means (and standard deviations) from combining five parallel MCMC chains of length* 40 000*, after burn-in of* 20 000*, using BUGS*

'Susceptibility' model		
Intercept	ϕ_0	−0.78
		(0.33)
Parental smoking	ϕ_p	0.72
		(0.20)
Sex	ϕ_g	−0.30
		(0.27)
Logistic–normal model given susceptible		
Intercept	η_0	−2.6
		(0.74)
Subject SD	τ	2.4
		(0.39)
Parental smoking	η_p	0.89
		(0.42)
Sex	η_g	0.43
		(0.50)
Wave (M)	$\eta_{w(m)}$	0.33
		(0.05)
Wave (F)	$\eta_{w(f)}$	0.54
		(0.05)

Of primary interest, however, is that the estimated wave coefficients, for males and females, were almost identical to those obtained in the logistic–normal model. In the mixture model, these coefficients refer to change within individuals regarded as susceptible, while in the former model they ostensibly apply to the whole population. It seems counter-intuitive that there is estimated to be the same 'subject-specific rate of change' for susceptible individuals (mixture model) as for all individuals (logistic–normal model), and this raises questions regarding the interpretability of these coefficients, a point to which we return in Section 6.

5.1 *Estimated marginal differences*

In Table 6 we display the marginal risk differences estimated as described in Section 3. The differences between estimates from the three modelling approaches were not large from a practical point of view. The estimates of the parental smoking effect were slightly higher under the semiparametric method, while those for the wave/time effects were slightly lower, with the largest differences under the discrete mixture model. It might be expected that the inter-wave differences would be more affected by the modelling assumptions than the parental smoking effect, since estimation of the former could be more affected by the separation of within and between-subject information that is provided by multilevel modelling.

Table 6. *Estimates of marginal risk differences under three modelling approaches*

	Marginal	Logistic–normal	Discrete mixture
Parental smoking (yes versus no)[a]			
Males	0.115	0.113	0.108
Females	0.136	0.123	0.126
Wave 5 versus wave 1 (2 years)[b]			
Males	0.038	0.040	0.043
Females	0.056	0.061	0.066

[a] At wave 5; standard errors for all estimates were between 0.016 and 0.018.
[b] For nonsmoking parents; standard errors for all estimates were between 0.007 and 0.008.

Table 7. *Summary of posterior predictive distribution for three check statistics, as described in text. For each model, table shows* 2.5, 50 *and* 97.5%*-ile of posterior distribution, and posterior predictive* P*-value:* $P = \Pr(T(y^{\text{rep}}) > T(y) \mid y)$

		Logistic–normal				Discrete mixture			
	Observed	2.5%	50%	97.5%	P	2.5%	50%	97.5%	P
T_n	0.773	0.755	0.769	0.782	0.27	0.748	0.774	0.799	0.53
T_a	0.051	0.050	0.057	0.065	0.95	0.038	0.050	0.063	0.44
T_i	0.084	0.053	0.065	0.079	0.005	0.049	0.063	0.078	0.004

5.2 *Model checking*

Posterior predictive distributions for the three diagnostic check statistics are shown in Table 7, alongside the observed values reproduced from Table 1. For T_n (proportion never smoking) and T_a (proportion always smoking), the only suggestion of a lack of fit was associated with T_a for the logistic–normal model ($P = 0.95$), which indicates that the estimated model typically produces a higher proportion of 'always-smokers' than found in the data. This is apparently because the logistic–normal model is unable to reproduce the right-hand end of the distribution of π_{ij} very well. In this regard, the mixture model fared better. For both models, however, the P-values associated with T_i were extremely small, indicating that both models produce too few 'pure incident' smokers. Another way of expressing this result is that the models produce more within-subject variability in response than was observed in the data.

Posterior predictive P-values for the pairwise crude odds ratios, T_{OR}^{jk}, are shown in Figure 3. None of these P-values is particularly extreme, but plotted against time lag they show that both models predict relatively lower odds ratios at small lags and higher odds ratios at large lags, than observed in the data. In other words, as might be expected, the tendency seen in the raw data to decreasing association with increasing lag between time points is not well represented by these models.

6. DISCUSSION

Although the logistic–normal (or 'random-effects') model has been widely described as providing estimates of regression coefficients that represent effects *specific to a subject*, or, more precisely, conditional on the value of the random-effect (Neuhaus, 1992; Diggle *et al.*, 1994; Hu *et al.*, 1998), we question the value of these interpretations with binary outcomes. With data such as ours, where

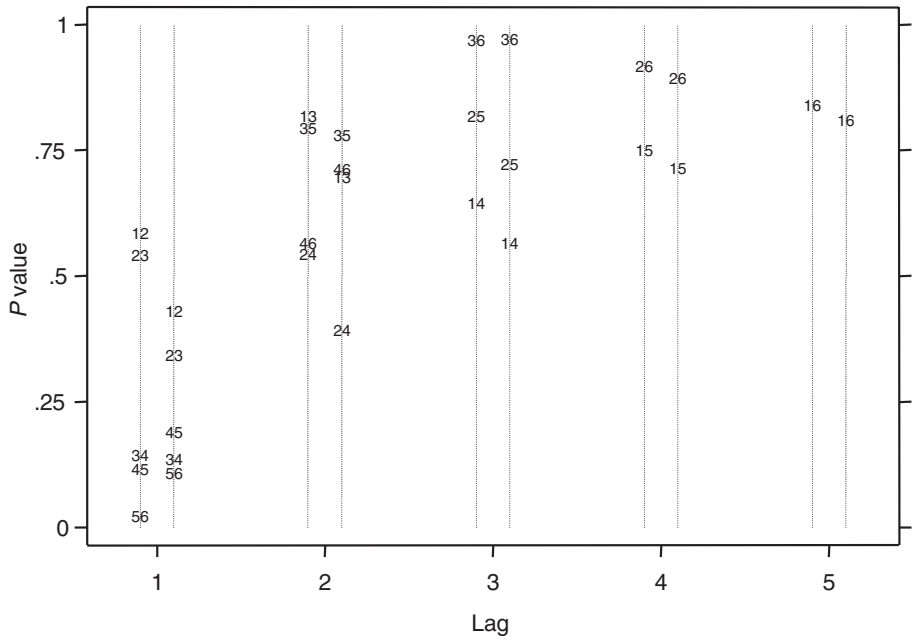

Fig. 3. Posterior predictive P-values for the pairwise crude odds ratios, T_{OR}^{jk}, defined by (5), plotted against the lag, $k - j$. The left-hand line at each lag corresponds to the logistic–normal model, the right-hand line to the discrete mixture. Points are labelled by the wave numbers jk.

the outcome is always negative in a substantial proportion of subjects, the estimated 'subject-specific time effects' ($\beta_{w(m)}$ and $\beta_{w(f)}$ in our example) are determined largely by the strength of persistence of the positive outcome response *within subjects who exhibit any positive responses*. This is shown in our example by the fact that the estimated wave coefficients are very similar in the mixture and logistic–normal models, despite the fact that in the mixture model these effects explicitly represent parameters specified only for the 'susceptible' subpopulation. It therefore seems problematic to give these parameter estimates in the logistic–normal model the usual regression-type interpretation concerning the relative odds of the outcome at one time point compared to a previous one, an interpretation that is implied in a number of reviews of methods for analysing repeated binary outcomes. For example, in the context of a similar example concerning adolescent smoking, Hu *et al.* (1998) describe logistic–normal models as 'estimating the changes in individuals' smoking behaviour across time'. A similar interpretation is implied by Heagerty (1999), who presents methods for obtaining estimates of these conditional regression coefficients from marginally specified models.

The fact that the time effect estimates are very similar in the mixture and logistic–normal models shows that a similar estimate of the rate of change parameter arises whether or not the large group of never-smokers is included in the analysis. Upon reflection, this is not surprising because it is intuitively clear that the never-smokers contribute little if any information about subject-specific rate of change. The phenomenon illustrated here is closely analogous to that arising in the standard conditional analysis of matched pair binary outcome data, where concordant pairs do not contribute to estimation of the odds ratio. This feature of estimation in paired binary data has been said to cause discomfort to many data analysts (Rothman and Greenland, 1998). Several authors have examined the phenomenon in the matched pair context, from the point of view of efficiency of estimation *given the common odds ratio model* (Liang

and Zeger, 1988; Neuhaus and Lesperance, 1996). The present analysis suggests a different point of view on this issue: that in some applications it may be appropriate to question the underlying model itself.

Is it really sensible to estimate the 'within-subject effect' of a covariate such as wave of study? What do we mean by this? It presumably would mean the difference between the probability (or logit) of the outcome in a given individual at one wave compared to another. In a different example, one might be concerned with the effect of a treatment (or pseudo-treatment) variable. For example, Neuhaus (1992) describes a longitudinal study of AIDS behaviour in which one of the aims was to estimate the effect of learning the result of an HIV test on the risk of engaging in unsafe sex. The aim was thus to estimate the difference for a given subject between their probability of unsafe sex with and without the knowledge of the test result, an aim that addresses a *causal* question under the counterfactual definition of causality proposed by Rubin (1974, see also Holland (1986)). Unfortunately, in either context ('causal' or not), the question posed seems to make sense only under the strong and untestable model that there is a constant effect (usually assumed in the logit scale) across all individuals. This is because, although each individual is observed under both covariate conditions (so in principle a direct estimate of causal effect is available for all), the fact that the outcome is binary means that no probability estimate can be obtained except by grouping individuals together in some way, with the usual approach being the constant effect assumption. Notice, moreover, that if our discrete-mixture model is adopted, the within-subject question is no longer sensible when applied to the whole population.

By the very nature of a binary outcome, if a subject never exhibits the outcome (or indeed, always exhibits it), it is not possible to answer the question of whether or not a within-subject effect (the time trend in our example) is the same for them as for others, who do exhibit change. This is very different from the situation with continuous outcome data, where even two or three occasions of measurement provide some information about each individual's slope parameter, no matter what their baseline or final values are. Thus with continuous outcomes there is always some information in the data with which to assess an assumption of common effect across subjects. (With binary data in which many subjects do exhibit change, it is still impossible to use the data to check the assumption of common slope within subjects, but perhaps it is a reasonable first assumption in such data, since the concept of change at least seems meaningful for all subjects.)

Although the regression parameters of the logistic–normal or similar complex models may be difficult to interpret, it does not follow that such models have no role in efforts to understand longitudinal data. In determining an appropriate role, it is critically important to clarify the ultimate aim of the statistical modelling. In this paper we limited attention to the rather simple aim of estimating average differences between subgroups defined by different covariate values, which could at best be interpreted as estimating *average* causal effects. One reason for this is the questionable meaning of the 'within-subject effect', as discussed, while another is the desire to explore simple questions thoroughly before moving onto more complex ones, where the answers may be more model dependent.

The mean difference in probabilities between individuals with distinct covariate values ('treated' and 'untreated') is of course estimated directly in the semiparametric modelling approach. Building a full multilevel model may, however, result in reduced bias and greater efficiency if a well-fitting model can be found. The two approaches estimate a similar quantity, in principle, because integrating over the unobservable effects distribution(s) in a multilevel specification is linear in the probability scale (so the integrated risk difference is the same as the difference of the integrated risks). Others have shown that if the data are truly generated by a logistic–normal specification, the marginal logistic model provides a good approximation to the true marginal model (Zeger *et al.*, 1988; Neuhaus *et al.*, 1991). However, biases may arise from fitting the marginal model to unbalanced data where missing data are not missing completely at random, essentially due to the combining together of between-subject and within-subject information in the estimation (Liang and Zeger, 1986; Carlin *et al.*, 1999). In this case study, there was little evidence

of major differences between the semiparametric and logistic–normal methods, or the alternative mixture model, for these particular inferential targets.

Comparing the marginal probability differences suggested that inferences for these quantities were not unduly influenced by the alternative model assumptions, but such immediate comparisons are not available when models are used for making conclusions about more complex or more abstract aspects of behavioural development. Multilevel models are being used more commonly as tools for building better understanding of developmental processes. While the background theory of relevant subject matter is important in judging model assumptions, it is also critical to assess the fit of the model to key aspects of the observed data. Approaches to model checking have recently been discussed for hierarchical models involving continuous outcomes (Hodges, 1998; Langford and Lewis, 1998) but less has been written on model fit for multilevel models with binary outcomes. We approached model diagnosis in our example by using the method of posterior predictive checks, comparing observed features of the data with the distribution of such features in replicated datasets under the fitted model. In fact, in our example, the posterior predictive check method suggested that there was at least one major aspect on which neither the logistic–normal nor the mixture model provided a good fit to the data (reproduction of the proportion of pure incident cases). The logistic–normal fit was also questionable with respect to one of the other check statistics.

Although the discrete mixture model has considerable appeal as an alternative to the logistic–normal in this example, it also has its own difficulties. For example, the realism of assuming a precise zero probability for a large subpopulation might be questioned. This represents a strong prior assumption, but it appears to be more reasonable than the alternative assumption of a normal distribution across the population. The resulting model achieved slightly better results in our diagnostic model checking although, as mentioned, still failed to fit on the check statistic of proportion of incident cases. The mixture model may be interpreted as incorporating an interaction effect between an unobservable latent variable and the other covariate effects, and we believe such interaction effects may need to be more widely used to capture complex behavioural phenomena realistically (Muthén and Curran, 1997). The mixture model is not straightforward to fit, although general-purpose Bayesian estimation software (BUGS) handled this particular problem adequately.

Further motivation for exploring multilevel models for binary outcomes in more depth may come from defining more complex substantive research questions. For example, it may be that a full probability model of the 'natural history' of a behavioural development such as the commencement of tobacco use can be used to inform the development of preventive interventions and the study of the effects of such interventions, which may not produce constant effects across the entire population. On the other hand, caution seems warranted by the limited information in binary outcomes (in this regard, see also Longford, 1993) and it may be more fruitful to develop more refined outcome scales for analysis.

In conclusion, the ease with which a fully specified random effects Bayesian model, as opposed to incompletely specified marginal models, can be used to obtain inferences regarding a wide range of population quantities, is both valuable and dangerous. The fact that neither the standard logistic–normal model nor the mixture model in our example reproduced the proportion of pure incident cases very well needs to be taken seriously when deriving population-based conclusions from the model. For example, these models do not seem to be specified sufficiently well to reproduce the population frequencies of identifiable developmental trajectories. Application of this model in the analysis of a preventive intervention trial, for example, might lead to misleading conclusions about the population impact of the intervention in reducing the incidence of smoking.

Furthermore, although many authors have suggested that the subject-specific logit difference parameter may be interpreted in its own right as estimating a 'subject-specific' effect, our example suggests that the usefulness of this interpretation is at best questionable. The smoking example demonstrates that the proper role of increasingly popular logistic–normal models is not yet clear and requires further research.

In particular, to what extent do the conclusions reached after fitting such models depend on aspects of the model that demonstrably fail to fit aspects of the data, or on untestable and perhaps unreasonable assumptions? As part of this, further thought needs to be given to clarifying the types of conclusions that are sought from the fitting of such models, moving from the use of regression coefficients to risk differences and other parameters that have clearer substantive interpretations.

ACKNOWLEDGEMENTS

This research was supported by a grant from Australia's National Health and Medical Research Council, and by the US National Institute of Mental Health, Grants No. MH40859, 'Designs and Analyses for Mental Health Preventive Trials', No. MH01259, 'Methodologic Advances in Mental Illness Prevention', and No. MH38725, 'Epidemiologic Center for Early Risk Behaviors'. We thank George Patton for providing data from the VAHCS, and for generous comments and encouragement. We also acknowledge our colleagues in the Prevention Science and Methodology Group, for their discussion of this work.

REFERENCES

BRESLOW, N. E. AND CLAYTON, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of American Statistical Association* **88**, 9–25.

BRYK, A. S. AND RAUDENBUSH, S. W. (1992). *Hierarchical Linear Models: Applications and Data Analysis Methods*. Newbury Park, CA: Sage.

CARLIN, J. B., WOLFE, R., COFFEY, C. AND PATTON, G. C. (1999). Analysis of binary outcomes in longitudinal studies using weighted estimating equations and discrete-time survival methods: prevalence and incidence of smoking in an adolescent cohort (Tutorial in Biostatistics). *Statistics in Medicine* **18**, 2655–2679.

DIGGLE, P. J., LIANG, K.-Y. AND ZEGER, S. L. (1994). *Analysis of Longitudinal Data*. Oxford: Oxford University Press.

GELMAN, A., CARLIN, J. B., STERN, H. S. AND RUBIN, D. B. (1995). *Bayesian Data Analysis*. London: Chapman and Hall.

GELMAN, A., MENG, X.-L. AND STERN, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica* **6**, 733–807.

GELMAN, A. AND RUBIN, D. B. (1992). Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science* **7**, 457–511.

GOLDSTEIN, H. (1991). Nonlinear multilevel models, with an application to discrete response data. *Biometrika* **78**, 45–51.

GOLDSTEIN, H. (1995). *Multilevel Statistical Models*. London: Arnold.

GOLDSTEIN, H. AND RASBASH, J. (1996). Improved approximations for multilevel models with binary responses. *Journal of the Royal Statistical Society* A **159**, 505–513.

GOLDSTEIN, H., RASBASH, J., PLEWIS, I., DRAPER, D., BROWNE, W., YANG, M., WOODHOUSE, G. AND HEALY, M. (1998). *A User's Guide to MLwiN*. London: Institute of Education.

HEAGERTY, P. J. (1999). Marginally specified logistic–normal models for longitudinal binary data. *Biometrics* **55**, 688–698.

HEAGERTY, P. J. AND ZEGER, S. L. (1998). Lorelogram: a regression approach to exploring dependence in longitudinal categorical responses. *Journal of the American Statistical Association* **93**, 150–162.

HEDEKER, D. AND GIBBONS, R. D. (1996). MIXOR: a computer program for mixed-effects ordinal regression analysis. *Computer Methods and Programs in Biomedicine* **49**, 157–176.

HEDEKER, D., GIBBONS, R. D. AND FLAY, B. R. (1994). Random-effects regression models for clustered data with an example from smoking prevention research. *Journal of Consulting and Clinical Psychology* **62**, 757–765.

HIBBERT, M., HAMILL, M., ROSIER, M., CAUST, J., PATTON, G. AND BOWES, G. (1996). Computer administration of a school-based adolescent health survey. *Journal of Paediatrics and Child Health* **32**, 372–377.

HODGES, J. S. (1998). Some algebra and geometry for hierarchical models, applied to diagnostics (with discussion). *Journal of the Royal Statistical Society* B **60**, 497–536.

HOLLAND,, P. W. (1986). Statistics and causal inference (with discussion). *Journal of the American Statistical Association* **81**, 945–970.

HU, F. B., GOLDBERG, J., HEDEKER, D., FLAY, B. R. AND PENTZ, M. A. (1998). Comparison of population-averaged and subject-specific approaches for analyzing repeated binary outcomes. *American Journal of Epidemiology* **147**, 694–703.

LAIRD, N. AND WARE, J. H. (1982). Random-effects models for longitudinal data. *Biometrics* **38**, 963–974.

LANGFORD, I. H. AND LEWIS, T. (1998). Outliers in multilevel data. *Journal of the Royal Statistical Society* A **161**, 121–160.

LIANG, K.-Y. AND ZEGER, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13–22.

LIANG, K.-Y. AND ZEGER, S. L. (1988). On the use of concordant pairs in matched case-control studies. *Biometrics* **44**, 1145–1156.

LONGFORD, N. (1993). *Random Coefficient Models*. Oxford: Clarendon.

MUTHÉN, B. (1993). Latent variable modelling of growth with missing data and multilevel data. In Cuadras, C. M. and Rao, C. R. (eds), *Multivariate Analysis: Future Directions 2*, Amsterdam: Elsevier, pp. 199–210.

MUTHÉN, B. O. AND CURRAN, P. J. (1997). General longitudinal modelling of individual differences in experimental designs: a latent variable framework for analysis and power estimation. *Psychological Methods* **2**, 371–402.

NEUHAUS, J. M. (1992). Statistical methods for longitudinal and clustered designs with binary responses. *Statistical Methods in Medical Research* **1**, 249–273.

NEUHAUS, J. M., KALBFLEISCH, J. D. AND HAUCK, W. W. (1991). A comparison of cluster-specific and population-averaged approaches for analyzing correlated binary data. *International Statistical Review* **59**, 25–35.

NEUHAUS, J. M. AND LESPERANCE, M. L. (1996). Estimation efficiency in a binary mixed-effects model setting. *Biometrika* **83**, 441–446.

PATTON, G. C., CARLIN, J. B., COFFEY, C., WOLFE, R. AND BOWES, G. (1998). Depression, anxiety and smoking initiation: a prospective study over three years. *American Journal of Public Health* **88**, 1518–1522.

PATTON, G. C., CARLIN, J. B., COFFEY, C., WOLFE, R., HIBBERT, M. E. AND BOWES, G. (1998). The course of early smoking: a population based cohort study over three years. *Addiction* **93**, 1251–1260.

RODRIGUEZ, G. AND GOLDMAN, N. (1995). An assessment of estimation procedures for multilevel models with binary responses. *Journal of Royal Statistical Society* A **158**, 73–90.

ROTHMAN, K. G. AND GREENLAND, S. (1998). *Modern Epidemiology*. Philadelphia, PA: Lippincott-Raven.

RUBIN, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* **66**, 688–701.

SAS (1999). *What's New in SAS Software for Version 8*. Cary, NC, USA: SAS Institute, Inc.

SPIEGELHALTER, D. J., THOMAS, A., BEST, N. G. AND GILKS, W. R. (1996). *BUGS: Bayesian inference Using Gibbs Sampling, Version 0.5 (version ii)*. Cambridge, UK: MRC Biostatistics Unit.

STATACORP (1999). *Stata: Release 6.0*. College Station, TX: Stata Corporation.

WOLFINGER, R. (1993). Laplace's approximation for nonlinear mixed models. *Biometrika* **80**, 791–795.

WOLFINGER, R. AND O'CONNELL, D. (1993). Generalized linear mixed models: a pseudo-likelihood approach. *Journal of Statistical Computation and Simulation* **48**, 233–243.

ZEGER, S. L., LIANG, K.-Y. AND ALBERT, P. S. (1988). Models for longitudinal data: a generalized estimating equation approach. *Biometrics* **44**, 1049–1060.

[*Received 17 January, 2000; revised 27 October, 2000; accepted for publication 1 November, 2000*]