Vet. Res. (2009) 40:30
DOI: 10.1051/vetres/2009013

www.vetres.org

Original article

Use of posterior predictive assessments to evaluate model fit in multilevel logistic regression

Martin J. GREEN[1,2*], Graham F. MEDLEY[3], William J. BROWNE[4]

[1] School of Veterinary Medicine and Science, University of Nottingham, Sutton Bonington Campus, Sutton Bonington, LE12 5RD, United Kingdom
[2] School of Mathematical Sciences, University of Nottingham, Nottingham, NG7 2RD, United Kingdom
[3] Department of Biological Sciences, University of Warwick, Coventry, CV4 7AL, United Kingdom
[4] Department of Clinical Veterinary Science, University of Bristol, Langford House, Langford, Bristol, BS40 5DT, United Kingdom

(Received 5 November 2008; accepted 24 March 2009)

Abstract – Assessing the fit of a model is an important final step in any statistical analysis, but this is not straightforward when complex discrete response models are used. Cross validation and posterior predictions have been suggested as methods to aid model criticism. In this paper a comparison is made between four methods of model predictive assessment in the context of a three level logistic regression model for clinical mastitis in dairy cattle; cross validation, a prediction using the full posterior predictive distribution and two 'mixed' predictive methods that incorporate higher level random effects simulated from the underlying model distribution. Cross validation is considered a gold standard method but is computationally intensive and thus a comparison is made between posterior predictive assessments and cross validation. The analyses revealed that mixed prediction methods produced results close to cross validation whilst the full posterior predictive assessment gave predictions that were over-optimistic (closer to the observed disease rates) compared with cross validation. A mixed prediction method that simulated random effects from both higher levels was best at identifying the outlying level two (farm-year) units of interest. It is concluded that this mixed prediction method, simulating random effects from both higher levels, is straightforward and may be of value in model criticism of multilevel logistic regression, a technique commonly used for animal health data with a hierarchical structure.

model fit / posterior predictive assessment / mixed predictive assessment / cross validation / Bayesian multilevel model

1. INTRODUCTION

Random effect statistical models are being increasingly used in veterinary sciences within both frequentist and Bayesian frameworks. Models are commonly specified with a binary outcome to represent, for example, 'diseased' or 'non-diseased' states and therefore take the form of multilevel logistic regression [5]. An important element of constructing and finalising a statistical model is to critically assess the fit and performance of the model [8]. However, model checking with discrete data regressions is problematic because usual methods, such as residual plots, have complicated reference distributions that depend on the parameters in the model [7, 4]. Thus, these traditional methods

* Corresponding author: martin.green@nottingham.ac.uk

are considered to be of limited value in discrete outcome, random effects models [2]. It may be because of this that, in the applied literature, particularly when complex discrete response models are specified, attention to model fit is often cursory.

In this research, a recently reported method of mixed predictive model assessment [10] is examined and illustrated in the context of an example from veterinary epidemiology. The concept is extended from the two level Poisson regression originally reported, to a three logistic regression setting with the focus of interest on prediction of bovine clinical mastitis on dairy farms in a specific year [6].

Posterior prediction is a general term used when data are generated under a proposed model, often so that comparisons can be made between specific features of the observed and generated data [3]. The approach provides a useful means for model assessment and cross validatory posterior predictive distributions are generally considered a 'gold standard' [10, 13]. Using cross validation, the data are partitioned '*k*' times into subsets and an analysis is initially performed on the 'training' subset. The other 'testing' subset(s) are retained to validate the initial analysis by making predictions from the data. Data predictions are compared with the observed data. The procedure is repeated *k* times and *k* may equal the total number of data points in the dataset or may represent groups of data within the full set. An important element of cross validation is that predictions made on each subset of testing data are independent of the observed outcome for that subset. The comparisons are used to identify discrepancies between model and data.

There is an important difference between conventional residual analysis and cross validation as a means of assessing outlying data regions in the context of model assessment. In conventional residual analysis, all data points are included in the model fit and thus will have a direct effect on model parameters and fitted values, and hence the difference between observed and fitted values. This is not the case with cross validation when the data points or groups have no influence at all on their cross validatory predicted values, because they are omitted during estimation, and in this respect, classical residual plots are likely to be over-optimistic in the assessment of model fit (i.e. they may not identify all of the true outlying regions) compared with cross validation. Outlying units from cross validation are those for which the other units do not provide sufficient information for the model to fit; outliers from residual analysis are those for which their own influence is insufficient to provide a fit. Therefore, regions of poor fit identified by cross validation will not necessarily be identified by residual analysis indicating the importance of the former method.

A significant disadvantage of cross validation is that it is computationally intensive and thus time consuming. A model has to be re-estimated for each of *k* subsets and this may include hundreds or thousands of data points or regions. If Markov chain Monte Carlo (MCMC) procedures are being used (as has been recommended for random effects logistic regression models [1]), and particularly with large data sets, the timescale required means that cross-validation may often become impractical (depending on the choice of *k*).

Alternative methods to cross-validatory predictions have been suggested that have the advantage of being more straightforward to compute and less computationally intensive. Gelman et al. [3] proposed use of the full model predictive distribution to make predictions on any required aspect of the data. This method may be over-optimistic in the context of model checking (i.e. it may fail to identify true outlying regions) compared to cross-validation because, as for residual analysis, the prediction of any data region tends to be strongly influenced by the equivalent observed data for the region. Marshall and Spiegelhalter [10] proposed a method termed the 'mixed' predictive check which they have illustrated in the context of disease mapping, and which appeared to perform in a similar manner to cross validation. The mixed predictive check incorporates simulated random effects, generated from their underlying distribution which is characterised from fitting the initial model, rather than the random effects estimated directly from the data. Use of the mixed predictive distribution has

also been reported in the context of differential gene expression [9]. In that study, mixed predictive Markov chain P values were used to evaluate hierarchical models [3, 10] but comparisons were not made between different methods of posterior predictions as a means to assess model fit. In this context, Markov chain P values are an indicator of the probability that a predicted data region is numerically higher (or lower) than the observed equivalent. If the probability is high (typically greater than 95% or 97.5%) or low (typically less than 5% or 2.5%) then it suggests that the model is performing poorly in the data region.

The purpose of this paper is to illustrate and compare four methods of model predictive assessment in the context of a multilevel logistic regression model, in which the specific clinical interest was the prediction of disease in a higher level unit (in this example a farm-year). The methods are cross validation, a full posterior predictive assessment and two mixed predictive methods based on the approach proposed by Marshall and Spiegelhalter [10]. An extension to the concept of the mixed prediction is described that is generalisable to three level hierarchical models.

2. MATERIALS AND METHODS

2.1. The data and initial model

The data for this analysis comprises clinical mastitis and farm management information from fifty two commercial dairy herds, located throughout England and Wales, with a mean herd size of approximately 150 cows and has been described in detail previously [6]. Data were collected over a two year period. The aim of the original research was to investigate the influence of cow characteristics, farm facilities and herd management strategies during the dry period, on the rate of clinical mastitis after calving. Interest was focussed on identifying determinants for clinical mastitis occurrence and to assess the extent to which these determinants could be used to predict the occurrence of clinical mastitis in each year on each farm. The response variable was at the cow level; a cow either got a case of clinical mastitis (= 1) or not (= 0) within 30 days of calving and a cow could be at risk in both years of the study. Predictor variables were included at the cow, year and farm levels. The model hierarchical structure was cows within farm-years within farms, and can be summarised as:

$$\begin{aligned} CM_{ijk} &\sim Bernouilli\,(\pi_{ijk}) \\ Logit\,(\pi_{ijk}) &= \beta_0 + \beta_1 X^{(1)}_{ijk} + \beta_2 X^{(2)}_{jk} \\ &+ \beta_3 X^{(3)}_{k} + u_{jk} + v_{0k} + v_{1k} P_{ijk} \\ u_{jk} \sim N(0, \sigma^2_u), v_k &= \begin{pmatrix} v_{0k} \\ v_{1k} \end{pmatrix} \sim MVN(0, \Omega_v) \end{aligned} \quad (1)$$

where the subscripts i, j and k denote the three model levels, π_{ijk} the fitted probability of clinical mastitis (CM) for cow i in year j on farm k, β_0 the regression intercept, $X^{(1)}_{ijk}$ the vector of covariates at cow level, β_1 the coefficients for covariates $X^{(1)}_{ijk}$, $X^{(2)}_{jk}$ the vector of farm-year level covariates, β_2 the coefficients for covariates $X^{(2)}_{jk}$, $X^{(3)}_{k}$ the vector of farm level covariates, β_3 the coefficients for covariates $X^{(3)}_{k}$, P_{ijk} is a covariate (within $X^{(1)}_{ijk}$) that identifies cows of parity one (after first calf), u_{jk} is a random effect to reflect residual variation between years within farms, and v_{0k} and v_{1k} are random effects to reflect residual variation between farms, and for the difference in rates for parity 1 cows between farms respectively.

Model selection was made from a rich dataset of more than 350 covariates. Model building has been described in detail previously [6] but briefly proceeded as follows. Each of the covariates was examined individually, within the specified model framework, to investigate individual associations with clinical mastitis whilst accounting for the data structure. Initial covariate assessment was carried out using penalised quasi-likelihood for parameter estimation (MLwiN, [11]) and final models were selected using MCMC for parameter estimation in WinBUGS [12]. A burn-in of at least 2 000 iterations was used for all MCMC runs during which time model convergence had occurred. Parameter estimates were based on a further 8 000 iterations. The final model included the following predictor variables; cow parity, cow historic infection status, whether the farm maintained a cow standing time of 30 min after administration of treatments at drying off (the end of the previous lactation), whether farms reduced the milk yield of high yielding cows before drying off, whether cow bedding was disinfected during the early dry period, type of cow bedding during the late dry period, the time period between sequential cleaning out of the calving pens, and the time between calving and the cows being first milked after calving.

2.2. Predictive assessments

Of particular clinical interest in the research was the prediction of the incidence rate of clinical mastitis (number of cases per cow at risk) for each of the $j = 1\ldots103$ farm-years and thus the predictions of these rates were used to investigate methods of model assessment. Four methods of predictive assessment were compared; cross validation, a full posterior predictive check and two 'mixed' predictive assessments similar to that suggested by Marshall and Spiegelhalter [10]. After final model selection, each method of prediction was incorporated into the MCMC process. At each iteration after model convergence, a prediction was made for the occurrence of mastitis for each individual cow (y_{ijk}) by drawing from the appropriate conditional probability distribution (see below). Similarly, at each iteration, the number of predicted cases of clinical mastitis were summed over all cows in each farm-year and divided by the total cows at risk in each farm-year, to provide an MCMC estimate of the farm-year incidence rate of clinical mastitis. Predictions were made from 8 000 MCMC iterations after model convergence.

To describe the four methods of predictive assessment, we condense the model terms, such that the disease status for each cow (y_{ijk}) is conditional on a set of model fixed effect parameters β, covariates (at various levels) X_{ijk}, and random effects v_k, and u_{jk}:

$$y_{ijk} \sim p\left(y_{ijk} | \beta, X_{ijk}, V_k, U_{jk}\right)$$

The random effects have parameters represented by σ^2_u and Ω_v.

$$U_{jk} \sim p\left(U_{jk} | \sigma^2_u\right)$$
$$V_k \sim p(V_k | \Omega_V)$$

The four methods of predictive assessment employed were:

A. Cross validation ("xval"). Each of the 103 farm-years was removed from the analysis in turn and the model fitted to a reduced data set excluding the jkth farm-year (denoted $^{(-jk)}$), from which new model parameters were estimated ($\beta^{(-jk)}$, $v^{(-jk)}$, $u^{(-jk)}$, ${\sigma^2}_u{}^{(-jk)}$, $\Omega_v{}^{(-jk)}$). A replicate observation for the omitted data, $y_{ijk}{}^{\text{xval}}$ was simulated from the conditional distribution;

$$\begin{aligned} y_{ijk}{}^{\text{xval}} &\sim p\left(y_{ijk}{}^{\text{xval}} \middle| \beta^{(-jk)}, X_{ijk}, u_{jk}{}^{\text{xval}}, v_k{}^{\text{xval}}\right) \\ u_{jk}{}^{\text{xval}} &\sim p\left(u_{jk}{}^{\text{xval}} \middle| \sigma^2{}_u{}^{(-jk)}\right) \\ v_k{}^{\text{xval}} &\sim p\left(v_k{}^{\text{xval}} | \Omega_v{}^{(-jk)}\right) \end{aligned} \quad (2)$$

B. Posterior predictive assessment from the full data ("full"). The predictive distribution was conditional on all fixed effect and random effect parameters estimated in the final model and a replicate observation $y_{ijk}{}^{\text{full}}$ generated from the conditional distribution;

$$y_{ijk}{}^{\text{full}} \sim p\left(y_{ijk}{}^{\text{full}} \middle| \beta, X_{ijk}, v_k, u_{jk}\right) \quad (3)$$

C. Mixed prediction 1 ("mix1"). This predictive distribution was conditional on the fixed effect parameters and the random effect distributions from which new random effects, $u_{jk}{}^{\text{mix1}}$ and $v_k{}^{\text{mix1}}$, were simulated to make the prediction. Thus a replicate observation $y_{ijk}{}^{\text{mix1}}$ was generated from the conditional distribution;

$$\begin{aligned} y_{jk}{}^{\text{mix1}} &\sim p\left(y_{jk}{}^{\text{mix1}} \middle| \beta, X_{ijk}, u_{jk}{}^{\text{mix1}}, v_k{}^{\text{mix1}}\right) \\ u_j{}^{\text{mix1}} &\sim p\left(u_{jk}{}^{\text{mix1}} \middle| \sigma^2{}_u\right) \\ v_k{}^{\text{mix1}} &\sim p\left(v_k{}^{\text{mix1}} | \Omega_v\right) \end{aligned} \quad (4)$$

D. Mixed prediction 2 ("mix2"). This predictive distribution was conditional on the fixed effect parameters, the random effects distribution at level 2, (from which new random effects, $u_{jk}{}^{\text{mix2}}$ were simulated), and the level 3 random effects from the model, v_k. Thus a replicate observation $y_{ijk}{}^{\text{mix2}}$ was simulated from the conditional distribution;

$$\begin{aligned} y_{ijk}{}^{\text{mix2}} &\sim p\left(y_{ijk}{}^{\text{mix2}} \middle| \beta, X_{ijk}, u_{jk}{}^{\text{mix2}}, v_k\right) \\ u_{jk}{}^{\text{mix2}} &\sim p\left(u_{jk}{}^{\text{mix2}} \middle| \sigma^2_u\right) \end{aligned} \quad (5)$$

2.3. Comparisons between methods of predictive assessments

In each case, predictions of farm-year incidence rates of clinical mastitis were compared with observed rates. Predictions from cross validation (taken as a gold standard) were also compared to the other methods of prediction to assess which best mimicked this procedure. To assess the degree of discrepancy between observed and predicted farm-year incidence rate of mastitis, the predicted distributions, $y_{jk}{}^{\text{pred}}$ were compared to the observed values using Monte Carlo predictive P values. At each iteration of the MCMC procedure, an indicator variable was set to 1 when $y_{jk}{}^{\text{pred}} > y_{jk}$, to 0.5 if $y_{jk}{}^{\text{pred}} = y_{jk}$ and to 0 if $y_{jk}{}^{\text{pred}} < y_{jk}$; the Monte Carlo P value was estimated as the mean of this indicator variable.

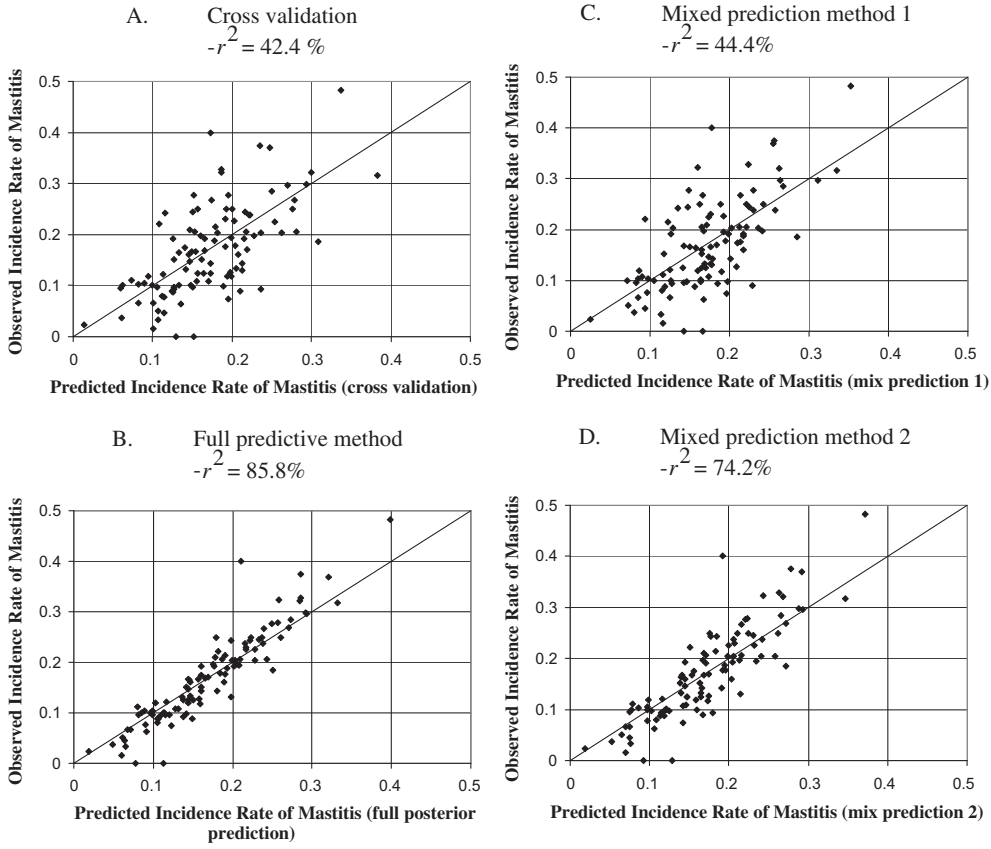

Figure 1. Plots of observed against predicted farm-year incidence rates of clinical mastitis (cases per cow at risk per year).

Therefore predictive P values > 0.975 or < 0.025 indicated that the probability of the observed incidence rate of clinical mastitis being within the predicted distribution was less than 5% and represented a relatively extreme result.

3. RESULTS

Figure 1 (A–D) illustrates the mean predicted incidence rate of clinical mastitis for each method of posterior prediction, plotted against the observed incidence of clinical mastitis. The graphs illustrate that the full posterior predictive method most closely resembled the observed data and cross validation and the "mix1" method displayed considerably more variability. The "mix2" method provided an intermediate result. Figure 2 illustrates the comparison between mixed and full predictive methods and cross validation. Both mixed predictive methods yielded better estimates of the cross validatory prediction than the full posterior predictive method, and the "mix2" method produced estimates most similar to cross validation.

The median error for each predictive method was calculated as the median of the unsigned differences between predicted and cross validatory farm-year incidence rates of clinical mastitis, as

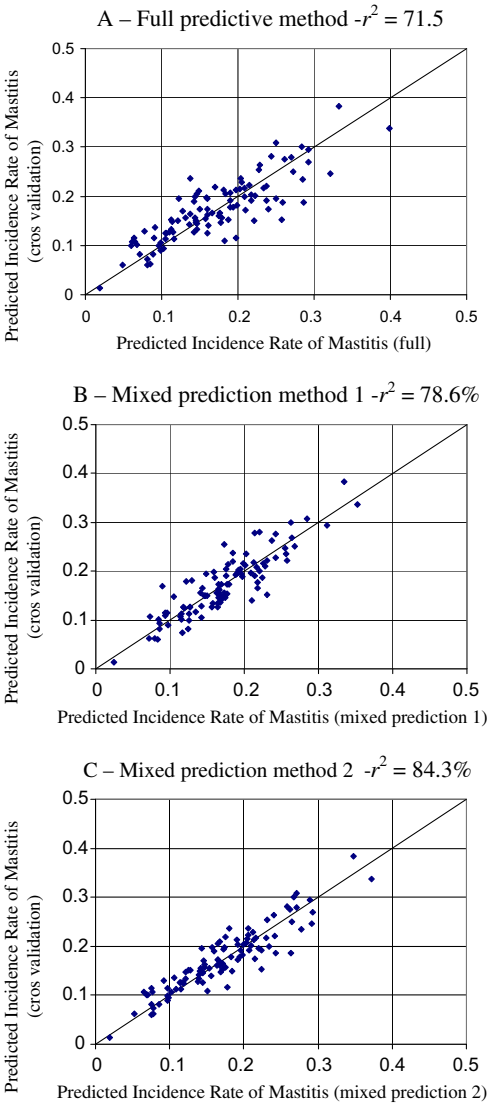

Figure 2. Plots of cross validatory predictions of farm-year clinical mastitis incidence against full and mixed predictive methods of farm-year clinical mastitis incidence (cases per cow at risk per year).

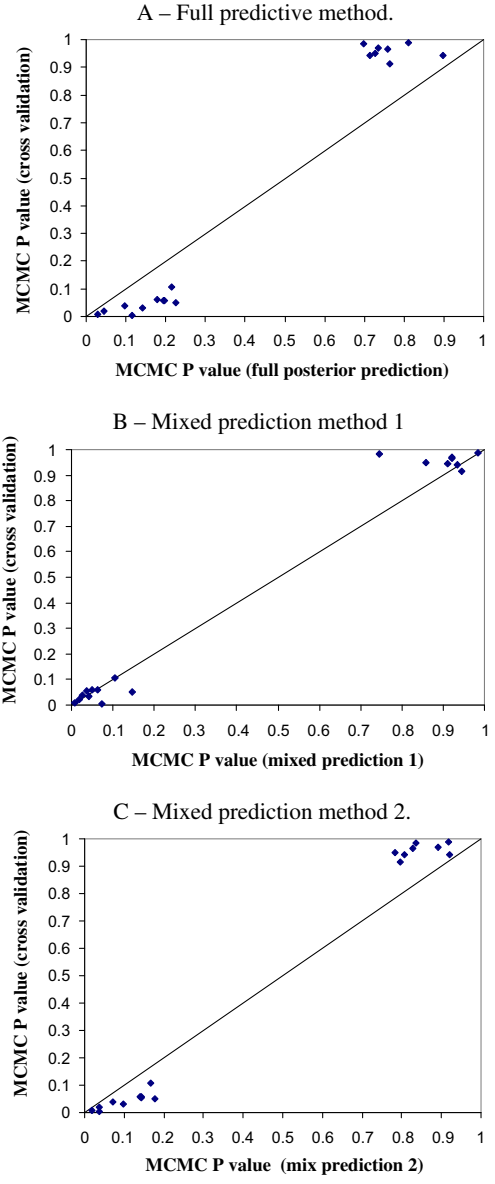

Figure 3. Comparison of MCMC *P* values from cross validation (for values > 0.80 and < 0.20) and from different methods of predictive assessment for farm-year incidence of clinical mastitis.

a percentage of the cross validatory farm-year incidence rate of clinical mastitis. The median errors were 13.7%, 11.5% and 9.4% for the full posterior prediction, the mixed prediction 1, and for mixed prediction 2 respectively.

Figure 3 illustrates the MCMC *P* values obtained from the different predictive methods to compare with the most extreme *P* values

Table I. Sensitivity and specificity of MCMC *P* values for each prediction method (full = full posterior predictive method, mix 1 and mix 2 = mixed predictive methods 1 and 2 respectively) compared to MCMC *P* values for cross validation, at different *P* value thresholds (as specified).

		Cross validation		Total	Sens (%)	Spec (%)
		0	1			
P value > 0.90 or < 0.10						
full	0	86	14	100	17.6	100.0
	1	0	3	3		
	Total	86	17	103		
mix 1	0	84	3	87	82.4	97.7
	1	2	14	16		
	Total	86	17	103		
mix 2	0	86	10	96	41.2	100.0
	1	0	7	7		
	Total	86	17	103		
P value > 0.95 or < 0.05						
full	0	93	8	101	20.0	100.0
	1	0	2	2		
	Total	93	10	103		
mix 1	0	90	5	95	50.0	96.8
	1	3	5	8		
	Total	93	10	103		
mix 2	0	93	7	100	30.0	100.0
	1	0	3	3		
	Total	93	10	103		
P value > 0.975 or < 0. 025						
full	0	98	5	103	0.0	100.0
	1	0	0	0		
	Total	98	5	103		
mix 1	0	98	2	100	60.0	100.0
	1	0	3	3		
	Total	98	5	100		
mix 2	0	98	4	102	20.0	100.0
	1	0	1	1		
	Total	98	5	103		

identified with cross validation, these being the most divergent regions eligible for identification and further investigation. At large and small *P* values ($P < 0.20$ or > 0.80) the mixed predictive methods performed more similarly to cross validation than the full posterior prediction with the "mix1" method most closely representing cross validatory MCMC *P* values. This is confirmed in Table I that provide the sensitivity and specificity for each predictive method, taking cross validation MCMC *P* values as the "gold standard", and different *P* value thresholds. The "mix1" method had the highest sensitivity indicating that this method identified the largest proportion of "true" extreme values as determined by cross validation. The "mix1" method identified 82.4% (14 out of 17) of extreme values when a threshold of < 0.10 or > 0.90 was used and 60% (3 out of 5) of extreme values with a threshold set at < 0.025 or > 0.975.

The computing times to complete 10 000 iterations (using an Intel Centrino 2.0 GHz Processor, 1.5GB RAM) for 103 cross validatory predictions and the "mix1" method were 334

h and 3.6 h respectively. This did not include the time required to format the data and set up each model and this took approximately the same time per model. Thus it took approximately 103 times longer for the cross validatory predictions than the "mix1" method.

4. DISCUSSION

Identifying divergent data regions in statistical modelling is important for two reasons. Firstly, numerous divergent regions could indicate that underlying statistical assumptions are incorrect, for example the model does not capture the true data structure. Secondly, individual divergent units could represent those that are fundamentally different from other units in the dataset after accounting for predictor variables, and the possible absence of unknown but important explanatory covariates. In either case, further investigations would be warranted. Cross validation provides a useful method of accurately identifying divergent units in complex statistical models, but faster methods would be of practical value in model assessment and it was for this reason that the alternative strategies were investigated in this research.

The predictions of clinical mastitis incidence rates obtained from the different methods show clear differences in results obtained, as shown in Figure 1. The full predictive method provided predicted incidence rates of clinical mastitis that most closely resembled the observed incidence rates, but these appeared to be over-optimistic in terms of model performance in comparison to cross validatory predictions. This is not surprising since the random effects from the initial model are directly incorporated into the prediction steps but it does highlight the difference between this method and cross validation.

For the three level logistic regression models in this example, the mixed predictive methods provided a better approximation to cross-validation than the full posterior predictive assessment. This is concordant with the first study that used a mixed prediction for approximating cross validation in a two level Poisson model for disease mapping [10]. In the current study using a three level logistic regression model, the "mix2" method provided the closest overall approximation to cross validatory predictions of farm-year incidence of clinical mastitis. However, the "mix1" method performed best for the more extreme outlying values identified by cross validation and thus this method was more useful for identifying the most divergent higher level units in these data. The mixed predictive methods look promising as a means of practical model assessment for the relatively common statistical approach of multilevel logistic regression and as such, warrant further investigations.

Importantly, the mixed predictive methods take considerably less time to implement (in this example approximately one hundredth of the time of cross validation) and therefore provide a clear advantage in terms of practical use. The "mix2" method is essentially a compromise between the "mix1" method and a full posterior prediction. The method simulates a new random effect at level 2 but uses the estimated random effects from the model at level 3. In the current example there were only two level 2 units for each level 3 unit and it may be that if more level two units existed for each level 3 units, mixed prediction method 2 would tend to become similar to mixed method 1 (the higher level unit having less influence on the predicted data). Similarly, the relative performance of the two mixed predictive methods may depend on the relative sizes of the higher level variances and more research into the importance of the relative size of higher level variances when using mixed predictive methods would be beneficial. In this example the variance at level two (farm-year) was 0.06 and at level three (farm) was 0.10 (for cows greater than parity one) and 0.64 (for cows of parity one). If the level three variances had been very small in comparison to the level 2 variance, it is possible that both mixed predictive methods used in this study would have yielded similar results. Further investigations of mixed predictive methods using different types of models, numbers of levels, units per level and relative sizes of higher unit variances would be worthwhile.

From our results, it would appear that, out of the methods examined, the "mix1" method is

likely to provide the closest representation of cross validation for potentially divergent data regions in multilevel logistic regression. However, it is important to note that these results apply only to one dataset and whilst in agreement with a previous study [10], need to be viewed with this perspective. It may be possible to generalise this approach to logistic regression and other multilevel models, but more research in this area is required.

Our results indicate that whilst mixed predictions provide a reasonable approximation to cross validation, they do not provide precise replication of the results. Therefore, a pragmatic approach for implementation of mixed predictive assessments may be for an initial highlighting of possible divergent data regions on which to undertake further model checking using cross validation. Thus, instead of undertaking cross validation on all possible regions an intermediate step could be to first use a mixed prediction approach and then to use cross validation for data regions that are potentially divergent based on the mixed prediction. A reduced mixed prediction MCMC *P* value threshold could be used to improve the likelihood that all 'true' outliers are identified, possibly the central 80 percentile region and cross validation then carried out on regions that fall outside this interval. This would increase the sensitivity of identifying "true" divergent regions using the mixed methods but would reduce the computing time required compared to using cross validation for all regions.

Assessment of model performance is important and problematic particularly when large datasets and complex model structures are used. Posterior predictions are recognised as a useful method to investigate model fit and more research on mixed posterior predictions may be useful to facilitate straightforward, fast assessments for these types of model.

Acknowledgements. Martin Green is funded by a Wellcome Trust Intermediate Clinical Fellowship.

REFERENCES

[1] Browne W.J., Draper D., A comparison of Bayesian and likelihood-based methods for fitting multilevel models, Bayesian Analysis (2006) 1: 473–514.

[2] Dohoo I.R., Martin W., Stryhn H., Veterinary epidemiologic research, Atlantic Veterinary College Inc., Prince Edward Island, Canada, 2003.

[3] Gelman A., Meng X., Stern H., Posterior predictive assessment of model fitness via realized discrepancies, Statistica Sinica (1996) 6:733–807.

[4] Gelman A., Goegebeur Y., Tuerlinckx F., van Mechelen I., Diagnostic checks for discreet data regression models using posterior predictive simulations, Appl. Stat. (2000) 49:247–268.

[5] Goldstein H., Multilevel Statistical Models, London, Edward Arnold, 1995.

[6] Green M.J., Bradley A.J., Medley G.F., Browne W.J., Cow, farm and management factors during the dry period that determine the rate of clinical mastitis after calving, J. Dairy Sci. (2007) 90:3764–3776.

[7] Landwehr J.M., Pregibon D., Shoemaker A.C., Graphical methods for assessing logistic regression models (with discussion), J. Am. Stat. Assoc. (1984) 79:61–83.

[8] Langford I., Lewis T., Outliers in multilevel data, J. R. Stat. Soc. Ser. A (1998) 161:121–160.

[9] Lewin A., Richardson S., Marshall C., Glazier A., Aitman T., Bayesian modelling of differential gene expression, Biometrics (2006) 62:1–9.

[10] Marshall E.C., Spiegelhalter D.J., Approximate cross-validatory predictive checks in disease mapping, Stat. Med. (2003) 22:1649–1660.

[11] Rasbash J., Browne W.J., Healy M., Cameron B., Charlton C., MLwiN Version 2.02, Multilevel Models Project, Centre for Multilevel Modelling, Bristol, UK, 2005.

[12] Spiegelhalter D.J., Thomas A., Best N., WinBUGS Version 1.4.1., Imperial College and MRC, UK, 2004.

[13] Stern H.H., Cressie N., Posterior predictive model checks for disease mapping models, Stat. Med. (2000) 19:2377–2397.