Journal of the
Royal Statistical Society

SERIES A
Statistics
in Society

J. R. Statist. Soc. A (2012)
175, *Part* 3, *pp.* 729–747

Improved probabilistic prediction of healthcare performance indicators using bidirectional smoothing models

Hayley E. Jones

University of Bristol, UK

and David J. Spiegelhalter

Medical Research Council Biostatistics Unit, Cambridge, and University of Cambridge, UK

[Received July 2010. Revised August 2011]

Summary. Smoothing of observed measures of healthcare provider performance is well known to lead to advantages in terms of predictive ability. However, with routinely collected longitudinal data there is the opportunity to smooth either between units, across time or both. Hierarchical generalized linear models with time as a covariate and hierarchical time series models each result in such two-way or 'bidirectional' smoothing. These models are increasingly being suggested in the literature, but their advantages relative to simpler alternatives have not been systematically investigated. With reference to two topical examples of performance data sets in the UK, we compare a range of models on the basis of their short-term predictive ability. Rather than focusing on point predictive accuracy alone, fully probabilistic comparisons are made, using proper scoring rules and tests for uniformity of predictive p-values. Hierarchical generalized linear models with time as a covariate were found to perform poorly for both data sets. In contrast, a hierarchical time series model with a latent AR(1) structure has attractive properties and was found to perform well. Of concern, however, is the large amount of time that is needed to fit this model using the WinBUGS software. We suggest that research into simpler and faster methods to fit models of a similar structure would be of much benefit.

Keywords: Continuous ranked probability score; Hierarchical generalized linear models; Hierarchical time series; Performance monitoring; Predictive performance; Provider profiling

1. Introduction

Routine 'performance' data are increasingly collected on many healthcare providers at regular intervals over time. This has been driven by demands for increased accountability of public services and, from a research perspective, to establish 'what works' (Bird *et al.*, 2005). Performance is monitored both locally and nationally, with strict government targets often calling for annual improvements. For example, in England the Care Quality Commission evaluates the performance of healthcare providers on the basis of a wide range of measures.

Relatively simple methods are required for the routine analysis of these performance data, which often consist of multiple short time series of counts along with covariates representing associated risk factors. In addition to making inferences about the past or present, e.g. to identify potentially poorly performing units, short-term predictions are also of interest. These predictions have various uses, being firstly of interest in their own right, e.g. for the planning of service

Address for correspondence: Hayley E. Jones, School of Social and Community Medicine, University of Bristol, Canynge Hall, 39 Whatley Road, Bristol, BS8 2PS, UK.
E-mail: hayley.jones@bristol.ac.uk

0964–1998/12/175729

provision. In fact, as noted by Leckie and Goldstein (2009), it is the *future* performance of each unit that is of interest for guiding provider choice. Secondly, as soon as the next set of time series data becomes available, short-term predictive distributions can be used to identify units that have experienced recent changes (Jones and Spiegelhalter, 2009). Thirdly, assuming that no sudden unpredictable changes have occurred, assessments of predictive performance can be used to compare models. We focus on this aspect in this paper.

It has long been recognized that observed performance measures should be adjusted for factors such as patient risk on presentation, the control of which cannot reasonably be considered to be in the domain of the healthcare provider. Assuming that such adjustment has taken place, the simplest model for prediction conjectures that each unit's risk-adjusted period T measure will be equal to that observed in the period $T-1$.

However, it is well known that shrinkage estimates lead to improved point predictions over these crude observed rates (James and Stein, 1961; Efron and Morris, 1975). This, in conjunction with other arguments such as increased precision of estimation and a desire to account for observed overdispersion, has led to many researchers suggesting the use of hierarchical or multilevel models for performance monitoring data sets (Morris and Christiansen, 1996; Burgess *et al.*, 2000a; Normand *et al.*, 1997; Goldstein and Spiegelhalter, 1996). Fitting such models need not be computationally demanding: using an empirical Bayes (EB) approach, shrinkage estimates and associated error terms can be computed easily, even using a spreadsheet. Examples of the use of the EB approach in performance monitoring include Greenland and Robins (1991) and Howley and Gibberd (2003).

For the simultaneous modelling of performance data across multiple time periods, the option exists to smooth observations instead across time periods. For example, simple regressions with time as a covariate (Marshall *et al.*, 1998) will result in the smoothing of observations within but not between providers. Dynamic generalized linear models (GLMs) (West and Harrison, 1997), the associated exponentially weighted moving average type smoothing techniques and related models offer more flexible alternatives.

A natural extension is of course to formulate a model resulting in smoothing both within (over time) and across providers. This has been referred to as 'bidirectional' smoothing by Martz *et al.* (1999). Hierarchical GLMs with time as a covariate (Daniels and Gatsonis, 1999; Bronskill *et al.*, 2002) and hierarchical time series models (West and Aguilar, 1998; Burgess *et al.*, 2000b; Lin *et al.*, 2009) have been suggested, but the properties of these relatively complex models have not been systematically investigated.

With reference to two worked examples, teenage conceptions in English local authorities (LAs) and cases of methicillin-resistant *Staphylococcus aureus* (MRSA) bloodstream infections in National Health Service (NHS) trusts, we compare a range of models on the basis of their ability to forecast rates one and two periods ahead. As well as point forecasts, we take into account the associated level of uncertainty in each predictive distribution. We borrow tools for this purpose from the weather forecasting literature (Gneiting *et al.*, 2007), where the evaluation of full probabilistic forecasts has received increased attention in recent years. The resulting numerical comparisons provide some insight into whether the relatively complex bidirectional smoothing methodology might be worthwhile.

After introducing the two examples and relevant notation in Section 2, the tools used for evaluating predictive accuracy are described in Section 3. We then review the literature on bidirectional smoothing models in Section 4 and describe two such models for counts. The predictive ability of these two models is assessed relative to simpler alternatives in Section 5. We briefly describe how the best performing model based on our results could be extended to incorporate population level trends in Section 6.

The WinBUGS code that was used to fit our preferred model (which has a hierarchical time series structure) can be obtained from

```
http://www.blackwellpublishing.com/rss
```

2. Background

2.1. Notation

We denote the observed count (e.g. of infections) in provider $i=1,\ldots,m$ in time period $t=1,\ldots,T$ as O_{it}. It is assumed throughout that $O_{it}|r_{it} \sim^{\text{indep}} \text{Poisson}(r_{it}E_{it})$. The 'expected' counts (E_{it}s) are estimated in advance according to a simple regression model, based on known risk factors considered to be beyond the influence of the providers. These E_{it}s are then assumed known for the purposes of all further analyses. The quantity r_{it} is the relative risk associated with provider i in time period t, which is considered to represent provider performance. Various models will be considered for the r_{it}s, allowing them to be either

(a) independent,
(b) associated between providers in each time period,
(c) associated within each provider over time or
(d) both.

We now introduce two example data sets, each of which is publicly available on the Internet.

2.2. Teenage conceptions in local authorities

First consider annual numbers of conceptions in 15–17-year-old girls in English LAs. We examine the data from 1997 to 2005 ($T=9$) on each of $m=352$ LAs. The high rates of such pregnancies in the UK relative to many other developed countries are of concern, and LAs are under pressure to obtain reductions. Conception rates for girls under 18 years old are considered a key performance indicator by the government (Department for Children, Schools and Families, 2010).

Among the factors that are recognized as important predictors of teenage conception rates are measures of socio-economic deprivation, educational achievement and population density (Wilkinson *et al.*, 2006; Diamond *et al.*, 1999). On the basis of this knowledge and examination of a range of Poisson regressions, we formulated a risk model which adjusted for the number of girls aged 15–17 years in the LA, the Office of the Deputy Prime Minister index of multiple deprivation 2004 score (Office of the Deputy Prime Minister, 2004), education statistics, a rural–urban classification index and also deprivation–education and deprivation–rurality interactions. Expected numbers of conceptions $\{E_{it}: i=1,\ldots,352; t=1,\ldots,9\}$ were calculated on the basis of this regression analysis.

Before performing any kind of provider profiling, the quality of risk adjustment should of course be considered carefully. In terms of predictive ability, it is likely that this could be improved by use of a more sophisticated risk adjustment model. The focus in this paper is, however, on the more interesting topic of suitable models for the relative risks, which are hoped to be applicable across a wide range of performance monitoring data sets.

Overall, the data exhibit an average annual reduction of 1.3% ($p<0.0001$, based on a generalized estimating equation analysis; Liang and Zeger (1986)). However, since this reduction is rather small, the system might be considered reasonably steady state. Fig. 1(a) shows observed teenage conception rates in the first eight LAs alphabetically, providing some indication of the typical rates and amount of variability.

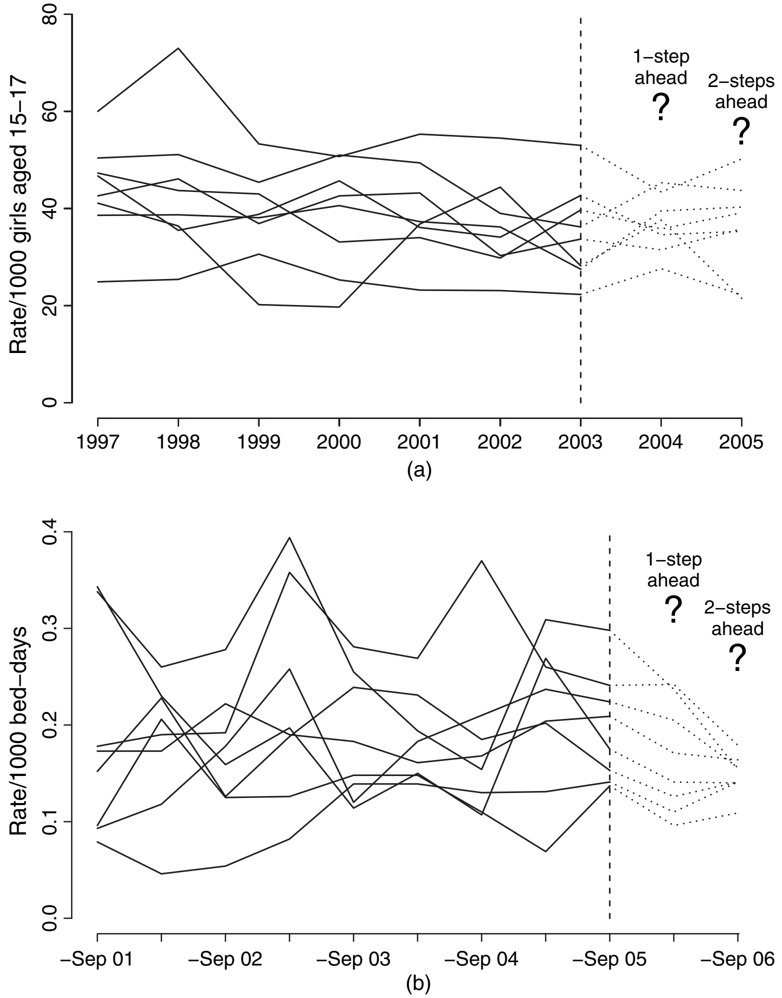

Fig. 1. Observed rates for the first eight LAs or NHS trusts alphabetically (observations to the right of the broken vertical line have been excluded for model fitting; these are instead withheld to assess each model's one-step-ahead and two-step-ahead predictive ability): (a) teenage conceptions; (b) MRSA

2.3. *Methicillin-resistant Staphylococcus aureus rates in National Health Service trusts*

Our second example is numbers of MRSA bloodstream infections in NHS trusts. These infections are a major public health concern in the UK and have received a huge amount of media and public attention. Rates are annually used as an indicator of trust performance by the Care Quality Commission.

We applied a very simple risk model to calculate the E_{it}s, which adjusted only for trust volume, as measured by n_{it}, which is the 'number of patient bed-days', and trust type, of which there are five. The existence of systematic differences between rates in different types of trust is well established, with the highest rates tending to be found in acute teaching trusts. These are generally large and urban, and receive referrals from other trusts for specialist services. The Health Protection Agency has noted that these differences are probably due to patient casemix and the higher risk that is posed by more invasive procedures.

The original data set consisted of the recorded number of cases in each of $T=11$ 6-month periods (from April 2001 to September 2006) in each of 172 NHS trusts. Two trusts were removed owing to missing data or no observed cases over the entire period of surveillance, leaving $m = 170$ for our analyses. A generalized estimating equation analysis applied to the entire data set indicated no evidence for population level change over this time period.

Observed rates in the first eight trusts alphabetically are shown in Fig. 1(b). It can be seen that these MRSA rates are somewhat more volatile than teenage conceptions. Unit-specific predictions are of interest in both cases, however.

2.4. Formulation of predictive distributions

For each data set, various models for the r_{it}s were fitted. Data from the final two periods $(t = T - 1, T)$ were excluded in each case, allowing one- and two-step-ahead predictive performance to be evaluated.

We focus on comparisons of the full probabilistic predictive ability of the various models, which requires the formulation of predictive distributions rather than just point predictions. For some simple models, closed form predictive distributions are readily available. In particular, a Poisson–gamma EB model fitted to period 1 data will provide closed form negative binomial predictive distributions for period 2 (see for example Jones and Spiegelhalter (2009)). Likewise, if we assume independence of the r_{it}s both within and across units and a Jeffreys prior, $p(r_i) \propto r_i^{-1/2}$, for each.

Consider for example Fig. 2, relating to the number of MRSA *bacteraemia* in an arbitrarily selected NHS trust. This plot is a smoothed histogram of the trust's predictive density for one period based on the data from only the period before, under the independent Poisson and Poisson–gamma EB models. The plot shows that fewer cases were observed than predicted by either of the models. It can be seen that the EB point prediction was the more accurate of the two in this particular instance. The plot also demonstrates that the EB predictive interval is narrower than the interval arising from the independence model. Using the terminology of Gneiting *et al.* (2007), this forecast is said to be the *sharper* of the two. A useful measure that is associated with both the point predictive accuracy and the sharpness is the predictive p-value: the area under each curve to the left of the vertical line. Related measures of predictive accuracy are discussed later.

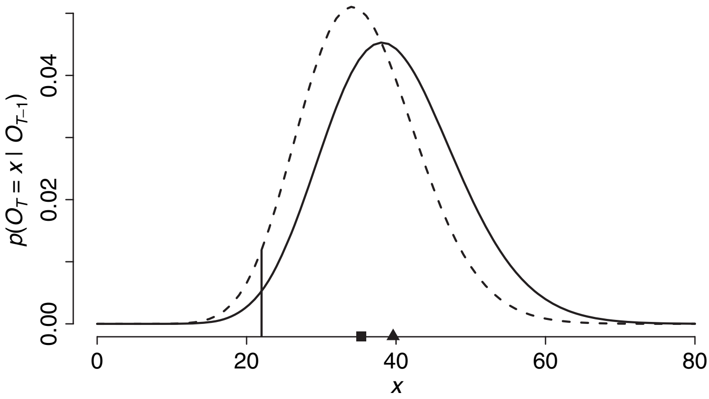

Fig. 2. Predictive density function for the number of MRSA bloodstream infections in a 6-month period in an arbitrarily selected NHS trust, under each of two models: ———, independent Poisson; – – –, Poisson–gamma EB; |, actual observed count in the final period, O_T; ▲, point prediction $\hat{O}_T$ under the independence model; ■, 'shrunken' point prediction under the EB model

For more complex models, we used the WinBUGS software to sample from the resulting predictive distributions by using Markov chain Monte Carlo techniques. Predictive p-values are then estimated by comparing each sampled value from the predictive distribution with the realized observation.

3. Evaluation criteria

Accurate probabilistic predictions of future events should be a key objective of biostatisticians (Spiegelhalter, 1986). Given the importance of predictions in the performance monitoring context as discussed in Section 1, model comparison based on predictive ability seems particularly appealing.

We assess one- and two-step-ahead predictive accuracy using three approaches, drawing heavily on recommendations by Gneiting *et al.* (2007). First, point predictions are compared with the set of realized observations by using summary measures such as the empirical root-mean-squared error (RMSE). We then assess the 'calibration' of the predictive distributions by testing for uniformity of the predictive p-values. Finally, two proper scoring rules are applied, models offering the lowest mean scores across providers being preferred.

3.1. Accuracy of point predictions

Denote the set of point predictions for period t by $\hat{O}_{1t}, \ldots, \hat{O}_{mt}$. To assess the accuracy of these, we use

$$\mathrm{RMSE} = \sqrt{\left\{ \frac{1}{m} \sum_{i=1}^{m} (O_{it} - \hat{O}_{it})^2 \right\}},$$

and also

$$\chi^2 = \sum_{i=1}^{m} \frac{(O_{it} - \hat{O}_{it})^2}{\hat{O}_{it}},$$

a statistic which appropriately scales down contributions from large providers.

3.2. Calibration of predictive distributions

The concept of sharpness was introduced briefly in Section 2.4. Sharpness is a desirable feature in a forecast, but only if the distribution is also correctly calibrated, i.e. events declared to have probability q occur a proportion q of the time on average (Gneiting *et al.*, 2005). We shall assess calibration by using the sets of predictive p-values.

Very wide predictive intervals will tend to lead to many p-values close to 0.5, whereas very narrow ('sharp') intervals might result in many extreme p-values if the forecasting system is incorrectly calibrated. Ideal forecasts would result in p-values which are approximately independent draws from a uniform$(0, 1)$ distribution. The uniformity of the predictive p-values can be assessed visually or by using standard test statistics.

This approach has been used commonly for the evaluation of single long series of continuous observations, particularly in an economic setting. Denoting the tth sequential observation Y_t and its predictive distribution function based on data up to time $t-1$ F_t, the 'probability integral transform' is defined as $U_t = F_t(Y_t)$. Standard tests of uniformity can be applied to the series $U_1, \ldots, U_T$. For applications, refer to Diebold *et al.* (1998), Gneiting *et al.* (2005, 2007) and Elder *et al.* (2005).

Uniformity of the probability integral transform values holds exactly for continuous measures. However, for discrete predictive densities $f_i(O_i)$, the measure $U_i = F_i(O_i)$ will not be exactly uniform. The corrected measure $U_i^* \equiv F_i(O_i) - u\, f_i(O_i)$ where $u \sim \text{uniform}(0,1)$ has an exact uniform distribution for perfectly calibrated forecasts (Denuit and Lambert, 2005). We approximate this by replacing the u-values by their expectation of 0.5, therefore applying the tests of uniformity to the 'mid-' predictive p-values defined as $p_i \equiv F_i(O_i) - 0.5\, f_i(O_i)$.

We examine the uniformity of each set $\{p_1, \ldots, p_m\}$ by using two empirical statistics, which measure discrepancies between a theoretical and an empirical distribution function: the Kolmogorov–Smirnov D and Cramér–von Mises's W^2. Computational formulae are given in Stephens (1974).

Approximate critical values for D and W^2 were obtained by taking a random sample of size m from a uniform$(0,1)$ distribution 10000 times and calculating the resulting test statistics in each case. The 9000th and 9500th largest values were then taken as approximate $\alpha = 0.10$ and $\alpha = 0.05$ critical values.

3.3. Proper scoring rules

As discussed by Gneiting *et al.* (2007), uniformity of the probability integral transform values is a necessary but not sufficient condition for a forecasting system to be 'ideal'. Alternative approaches to comparing predictive distributions are therefore required, such as the use of suitable scoring rules. For each model, we shall evaluate both the mean logarithmic score and the mean continuous ranked probability score (CRPS) across providers. As with the measures of point predictive accuracy in Section 3.1, models leading to low scores are preferred. Both of these scoring rules are 'proper', meaning that the expectation of the score is minimized by reporting one's beliefs honestly.

In assessing the compatibility of a predictive density function f_i and an observation y_i, the logarithmic score is defined simply as $-\log\{f(y_i)\}$. This is a well-known and commonly used scoring rule, which has been used in a variety of applications including the assessment of football game predictions and weather forecasting evaluation (Winkler, 1971; Bröcker and Smith, 2008). In practice we truncate, taking the average across providers of $-\log[\max\{f(y_i), 0.0001\}]$. However, the mean logarithmic score has been found to be highly sensitive to individual extreme cases, the CRPS being recommended as a more robust alternative (Gneiting and Raftery, 2007). This has been used for example by Gneiting *et al.* (2005) to compare forecasting systems for sea level pressure and surface temperature in the Pacific.

Given an observation y_i, the CRPS contribution for unit i is defined as

$$\begin{aligned}\text{CRPS}(F_i, y_i) &= \int_{-\infty}^{\infty} \{F_i(x) - I(y_i \leqslant x)\}^2 \,\mathrm{d}x \\ &= \int_{-\infty}^{y_i} F_i(x)^2 \,\mathrm{d}x + \int_{y_i}^{\infty} \{1 - F_i(x)\}^2 \,\mathrm{d}x. \qquad (1)\end{aligned}$$

These individual contributions are averaged over units to give a mean CRPS for each model. Alternatively, this measure can be thought of as the integral of the well-known Brier score over all possible thresholds (Hersbach, 2000). Unlike the logarithmic score, which depends only on the probability mass that is assigned to the value that materializes, the CRPS is also influenced by the sharpness.

Gneiting and Raftery (2007) have proved the following result:

$$\text{CRPS}(F_i, y_i) = E|Y_i^{\text{pred}} - y_i| - \tfrac{1}{2} E|Y_i^{\text{pred}} - Y_i^{\text{pred}'}|, \qquad (2)$$

where Y_i^{pred} and $Y_i^{\text{pred}'}$ are independent copies of a random variable drawn from the predictive distribution. For deterministic forecasts, these two copies are identical so the second term vanishes and equation (2) reduces to $|\hat{y}_i - y_i|$. As such, the mean CRPS reduces to the mean absolute difference and can be considered a generalization of this for probabilistic forecasts (Gneiting *et al.*, 2007). The result is also extremely useful in that it allows the CRPS to be evaluated easily using software such as WinBUGS.

In the next section we review models resulting in the smoothing of observed rates both within and between providers. The predictive performance of these relatively new models is then compared with that of more standard, simpler models by using the tools discussed in this section.

4. Bidirectional smoothing methods

The literature on 'bidirectional smoothing' models in performance monitoring can be divided roughly into three categories. The first entails a two-stage smoothing approach which is not fully model based. The second category involves models in which time is explicitly fitted as a covariate. The regression coefficients are then assumed to be drawn from a common distribution, thus resulting in shrinkage between providers as well as within. The models in the third category assume a more flexible dynamic process within each unit. Again, some assumptions of exchangeability between providers are made, leading to shrinkage in both directions.

4.1. Two-stage approach

Martz *et al.* (1999) smoothed unplanned 'scrams' in nuclear power plants, which are an indicator of plant performance and reliability. A Poisson–gamma model was assumed for the counts in each year independently, resulting in rates being smoothed across plants. Smoothing within plants over time was then achieved on a more *ad hoc* basis by applying an exponentially weighted moving average to these shrinkage estimates, rather than using a fully model-based approach. A similar method was mentioned by Jensen *et al.* (2009), who described a system of predicting the home run rates of baseball players. A two-stage approach is applied in the opposite direction here: each player's rate is first smoothed across 3 years by using a weighted average. These weighted averages are then shrunk towards the overall league mean in each year.

These two-stage approaches are very easy to apply, but the resulting one-step-ahead predictive distributions might not be straightforward to derive. Since we are more interested for the moment in assessing whether bidirectional smoothing is potentially worthwhile, we do not investigate this further in this paper.

4.2. Hierarchical generalized linear models with time as a covariate

Hierarchical linear models with time as a covariate have been fitted to educational performance data by Willms and Raudenbush (1989) and Gray *et al.* (1995, 1996), and in the modelling of healthcare performance by van Houwelingen *et al.* (2000). These models assume a simple regression structure, with provider-specific intercepts and gradients that are assumed to be random effects. Longitudinal binomial–beta models with a similar structure have also been applied in performance monitoring by Daniels and Gatsonis (1999) and Bronskill *et al.* (2002).

A model of this type can similarly be formulated for count data, and fitted using the WinBUGS software. Let us assume that $\log(r_{it}) = \beta_{i0} + \beta_{i1}(t - t^*)$, where t^* is the mid-value of t (or some other value chosen for ease of interpretability). It is then further assumed that the two unit-specific parameters are drawn from a common bivariate normal distribution:

$$\begin{pmatrix} \beta_{i0} \\ \beta_{i1} \end{pmatrix} \sim \mathrm{MVN}\left\{ \begin{pmatrix} \mu_0 \\ \mu_1 \end{pmatrix}, \begin{pmatrix} \tau_0^2 & \rho\tau_0\tau_1 \\ \rho\tau_0\tau_1 & \tau_1^2 \end{pmatrix} \right\}$$

where ρ is the correlation between β_{i0} and β_{i1} in the population.

In fitting this model, we assumed vague priors for the five hyperparameters ($\rho \sim \text{uniform}(-1, 1)$, $\mu_k \sim N(0, 100)$ and $\tau_k \sim \text{uniform}(0, 5)$, $k = 0, 1$).

4.3. Hierarchical time series models

Several researchers have discussed 'hierarchical time series' models, in which more flexible time trends are allowed in each unit. For example, Martz *et al.* (1999) noted the possibility of a fully integrated version of their two-stage approach, but they did not fit such a model. Hierarchical auto-regressive models have, however, been fitted by van Houwelingen *et al.* (2000), West and Aguilar (1998) and Lin *et al.* (2009). These models involve a marginal hierarchical model for each provider's performance at each time point, combined with a common auto-regressive structure over time.

West and Aguilar (1998) formulated a model for rates and applied this to psychiatric care performance data. It was assumed that each provider's underlying rate on the logit scale was marginally drawn from a common normal distribution in each period, with time-specific marginal mean and constant variance. The difference between this transformed rate and the population mean was then assumed to follow an auto-regressive process with lag 1 (AR(1)). This reflects the view that individual deviations from the population mean are expected to be relatively stable over time. A similar model has been applied to patient level data by Burgess *et al.* (2000b) and has also been extended to analyse K related performance indicators simultaneously, by assuming an auto-regressive process for each K-dimensional vector of deviations (West and Aguilar, 1998).

More recently, a Poisson model of a similar structure has been applied by Lin *et al.* (2009) in the modelling of mortality data from over 3000 renal dialysis centres. Like West and Aguilar (1998), Lin *et al.* (2009) used Markov chain Monte Carlo techniques to fit their proposed models. A similar but somewhat simpler model has, however, been fitted by van Houwelingen *et al.* (2000) by using EB methodology and the expectation–maximization algorithm. A transformation to approximate normality was employed to make estimation simpler. Also of note within the category of hierarchical time series models are the multivariate models that were applied by Daniels and Normand (2006) and McClellan and Staiger (1999).

Let us consider the model of Lin *et al.* (2009) in more detail. We note that the model itself received relatively little attention in the original publication, the main focus of which was on optimal ranking systems. In addition, some unintuitive prior distributions were used for population level parameters: we suggest what we believe to be more sensible alternatives below. We also extend the model to allow short-term predictions to be made automatically.

The model, which we apply below, assumes that a simple hierarchical structure holds marginally at each time point, with $\theta_{it} \equiv \log(r_{it}) \sim N(\mu_t, \tau_t^2)$. Dynamic structure is then imposed by the assumption that the standardized difference of each provider's log-relative-risk from the population mean follows an AR(1) process:

$$\frac{\theta_{it} - \mu_t}{\tau_t} = \phi\left(\frac{\theta_{i,t-1} - \mu_{t-1}}{\tau_{t-1}}\right) + \eta_{it}, \qquad t = 2, \ldots, T, \tag{3}$$

where the auto-regressive parameter $\phi \in [0, 1]$ and $\eta_{it} \sim^{\text{IID}} N(0, \nu^2)$.

As a result of the assumed marginal distributions, by taking the variance of expression (3) it is seen that the constraint $\phi^2 + \nu^2 = 1$ must be obeyed. The implied conditional distribution of θ_{it} given its previous value is then

$$\theta_{it} | \theta_{i,t-1} \sim N\left\{\mu_t + \phi \frac{\tau_t}{\tau_{t-1}}(\theta_{i,t-1} - \mu_{t-1}), (1 - \phi^2)\tau_t^2\right\}. \tag{4}$$

In the special case of constant marginal mean and variance ($\mu_t = \mu$ and $\tau_t = \tau \; \forall t$), the conditional mean from distribution (4) simplifies to $E(\theta_{it}|\theta_{i,t-1}) = \phi\theta_{i,t-1} + (1-\phi)\mu$, which is a weighted average of the previous risk and the population mean.

The original model as presented by Lin *et al.* (2009) involves estimating the population mean log-relative-risk (μ_t) and the amount of variation around this mean (τ_t) independently in each time period. This allows full flexibility in the population process over time but has the drawback of not allowing predictions to be made automatically for the next period. We therefore formulated simple models for these parameters. To allow some flexibility in any overall trend rather than a straight line, we assume a random walk: $\mu_t \sim N(\mu_{t-1}, \sigma_\mu^2)$, $t = 2, \ldots, T$. The same structure was assumed for the logarithm of the population standard deviation.

Lin *et al.* (2009) assumed the priors $\frac{1}{2}\log\{(1+\phi)/(1-\phi)\} \sim N(0, 0.2)$, $\tau_t^{-2} \sim^{\text{IID}}$ gamma(0.05,0.2) and $\mu_t \sim^{\text{IID}} N(0, 10)$, $t = 1, \ldots, T$. Although the last is a standard vague prior, we do not consider these priors for either ϕ or τ_t to be appropriate. Simulations showed that these priors correspond to a 95% credible interval of $(3.0, 3 \times 10^{16})$ for each τ_t, the lower limit of which seems to us to be larger than we would expect to find in practice, and a 95% credible interval of $(-0.7, 0.7)$ for ϕ. Initial fits of the model showed clear conflict between the data and each of these prior distributions. In practice we instead assumed a uniform(0,1) prior for ϕ and vague uniform priors for μ_1, τ_1, σ_μ and σ_τ.

5. Results

Table 1 shows summary statistics of one- and two-step-ahead predictive performance for a range of models fitted to the teenage conceptions data. As a baseline for comparison, the first row of Table 1 shows the results from fitting independent Poisson models to each period $T-2$ observation, the resulting predictions for periods $T-1$ and T being smoothed neither between nor within LAs. This model essentially issues the crude prediction that both $O_{i,T-1}/E_{i,T-1}$ and O_{iT}/E_{iT} will be equal to the last observed relative risk $O_{i,T-2}/E_{i,T-2}$ in each LA.

For both one- and two-step-ahead predictions, the RMSE, χ^2, CRPS and log-score all demonstrate an improvement made by smoothing these observations between units, on the basis of a simple cross-sectional hierarchical model (as shown in the second and third rows). The results and others not shown also demonstrated that, in terms of overall predictive ability, it made very little difference whether a gamma or log-normal distribution was assumed for the random effects, and whether an EB or fully Bayesian approach was followed. The empirical statistics D and W^2 did not indicate any cause for concern about the uniformity of the predictive p-values for any of these cross-sectional models.

Several models were considered involving smoothing within but not between units. We present the results from simple independent Poisson regressions with time as a covariate, and also from two more flexible models in which the log-relative-risk in each unit is assumed to follow an independent random walk or AR(1) process. Models of this type have become popular in infectious disease modelling (Hay and Pettitt, 2001; Heisterkamp *et al.*, 2006). Since only a few observations were available on each unit, we assumed constant smoothing and within-unit variability parameters across providers.

We used the WinBUGS software to fit these models, because of the ease with which the implied predictive distributions can be simulated from. However, we note that similar models resulting in smoothing within units only, such as a Poisson–gamma dynamic GLM, could be fitted by using a series of updating equations as outlined by West and Harrison (1997). In addition, Grigg and Spiegelhalter (2007) have developed a 'risk-adjusted exponentially weighted moving average' that is applicable directly to counts, which they have shown performs similarly to a

Table 1. Summary statistics comparing the ability of various models to predict teenage conception rates in 2004 (one step ahead) and 2005 (two steps ahead)†

Smooth	*Model*	*Point predictions*		*Uniformity of p-values*		*Average 90% interval width*	*CRPS*	*Log-score*
		RMSE	χ^2	*D*	W^2			
1 step ahead								
None	Independent Poisson	16.9	777	0.03	0.12	45.8	8.49	4.03
Between providers	Poisson–gamma EB	15.0	586	0.06	0.37	42.1	7.46	3.88
Between providers	Poisson–log-normal full Bayes	14.9	582	0.06	0.25	41.9	7.43	3.87
Within providers	Independent Poisson regressions	17.0	825	0.08‡	0.78‡	41.8	8.65	4.05
Within providers	Independent Poisson random walks	14.6	561	0.11‡	1.65‡	45.0	7.31	3.87
Within providers	Independent Poisson AR(1) processes	14.6	559	0.15‡	3.84‡	46.0	7.32	3.87
Between and within providers	Hierarchical GLM	14.6	581	0.07	0.51‡	38.1	7.29	3.86
Between and within providers	Hierarchical AR(1) + population processes	*13.6*	*510*	0.07	0.38	44.0	*6.89*	*3.82*
2 steps ahead								
None	Independent Poisson	19.1	868	0.06	0.32	46.4	9.22	4.09
Between providers	Poisson–gamma EB	16.9	668	0.09‡	0.86‡	42.5	8.08	3.94
Between providers	Poisson–log-normal full Bayes	16.8	665	0.08‡	0.69‡	42.2	8.04	3.94
Within providers	Independent Poisson regressions	22.0	1233	0.10‡	1.37‡	45.4	10.69	4.28
Within providers	Independent Poisson random walks	17.5	715	0.13‡	2.29‡	50.4	8.30	3.98
Within providers	Independent Poisson AR(1) processes	16.7	696	0.20‡	6.21‡	51.0	8.18	3.98
Between and within providers	Hierarchical GLM	18.5	810	0.10‡	1.25‡	40.0	8.76	4.02
Between and within providers	Hierarchical AR(1) + population processes	*15.5*	*626*	0.09‡	0.91‡	49.2	*7.58*	*3.91*

†The lowest values of RMSE, χ^2, CRPS and log-score for each time period are highlighted in italics. The $\alpha = 0.10$ critical values for $m = 352$ were estimated to be $D = 0.06$ and $W^2 = 0.35$, and the $\alpha = 0.05$ critical values $D = 0.07$ and $W^2 = 0.46$.

‡Rejection of the null hypothesis of uniformity at the 5% level.

Poisson–gamma dynamic GLM. A simple smoothing method such as this is obviously very appealing for routine performance analysis.

Table 1 shows that the models assuming an independent random walk or AR(1) process for each $\log(r_{it})$ over time led to much better predictions than the less flexible regressions with time as a covariate. In fact, independent Poisson regressions with time as a covariate gave worse predictions for both time periods than even the crude 'no-smoothing' model. The summary statistics for the one-step-ahead predictive ability of the two more flexible 'smoothing within' models are slightly better than those from the 'smoothing between' models, but for two steps ahead their performance is slightly worse. Further, for both time periods the D- and W^2-statistics raise concern about deviations of the predictive p-values from uniformity.

The final two rows of Table 1 demonstrate the predictive ability of the two bidirectional smoothing models that were described in Section 4. The one-step-ahead predictive performance of the hierarchical GLM was comparable with the simpler smoothing within models, but this was overshadowed by the hierarchical AR(1) model, which is seen to be preferred over all other models considered. When predicting two steps ahead, the hierarchical GLM was found not to perform at all well, much simpler models like the Poisson–gamma EB model being better, whereas the hierarchical AR(1) model again performed better than all others. Some of the flexible unit-specific curves fitted by this model are illustrated in Fig. 3.

For the MRSA data, results for the same set of models are shown in Table 2. In this instance, the 'smoothing within' models are clearly seen to perform better than the simple hierarchical models in terms of one-step-ahead predictive ability. The hierarchical GLM, smoothing in both directions but assuming simple log-linear time trends, again performed poorly, particularly in predicting two periods ahead. The more flexible hierarchical AR(1) model performed considerably better. The simpler independent Poisson random walks performed as well as the hierarchical AR(1) model in making one-step-ahead predictions, but somewhat worse in predicting two steps ahead. For two-steps-ahead predictions, independent Poisson AR(1) processes gave the best performance, by a small margin over the hierarchical AR(1) model. Taking the predictive ability in both time periods into account, the independent AR(1) models and hierarchical AR(1) model were the best performing models for the MRSA data.

For comparison, we also considered the historical fit of each of the longitudinal models, as measured by the deviance information criterion (DIC) (Spiegelhalter *et al.*, 2002). Like the classical Akaike information criterion, the DIC offers a measure of goodness of fit penalizing for model complexity, models with lower values being preferred. The DIC approximates full cross-validation, so we would expect the resulting order of model preference to be similar to that seen above. However, since the DIC uses different data to evaluate the predictions, some discrepancies are likely. The measures assessing genuine out-of-sample predictive ability might be considered more suitable in this context. However, the DIC makes use of a larger amount of data and should therefore be more stable.

As shown in Tables 3 and 4, the hierarchical AR(1) model is assessed to have the best historical model fit for both data sets, although the independent longitudinal models remain quite competitive for modelling MRSA.

6. Incorporating population level trends

In Section 4.3, we described how we extended the hierarchical AR(1) model of Lin *et al.* (2009) to incorporate a random walk for the population mean log-relative-risk μ_t, and for the logarithm of the population standard deviation τ_t. However, if long-term trends in either of these parameters were observed it would be appropriate to allow for this by assuming instead a

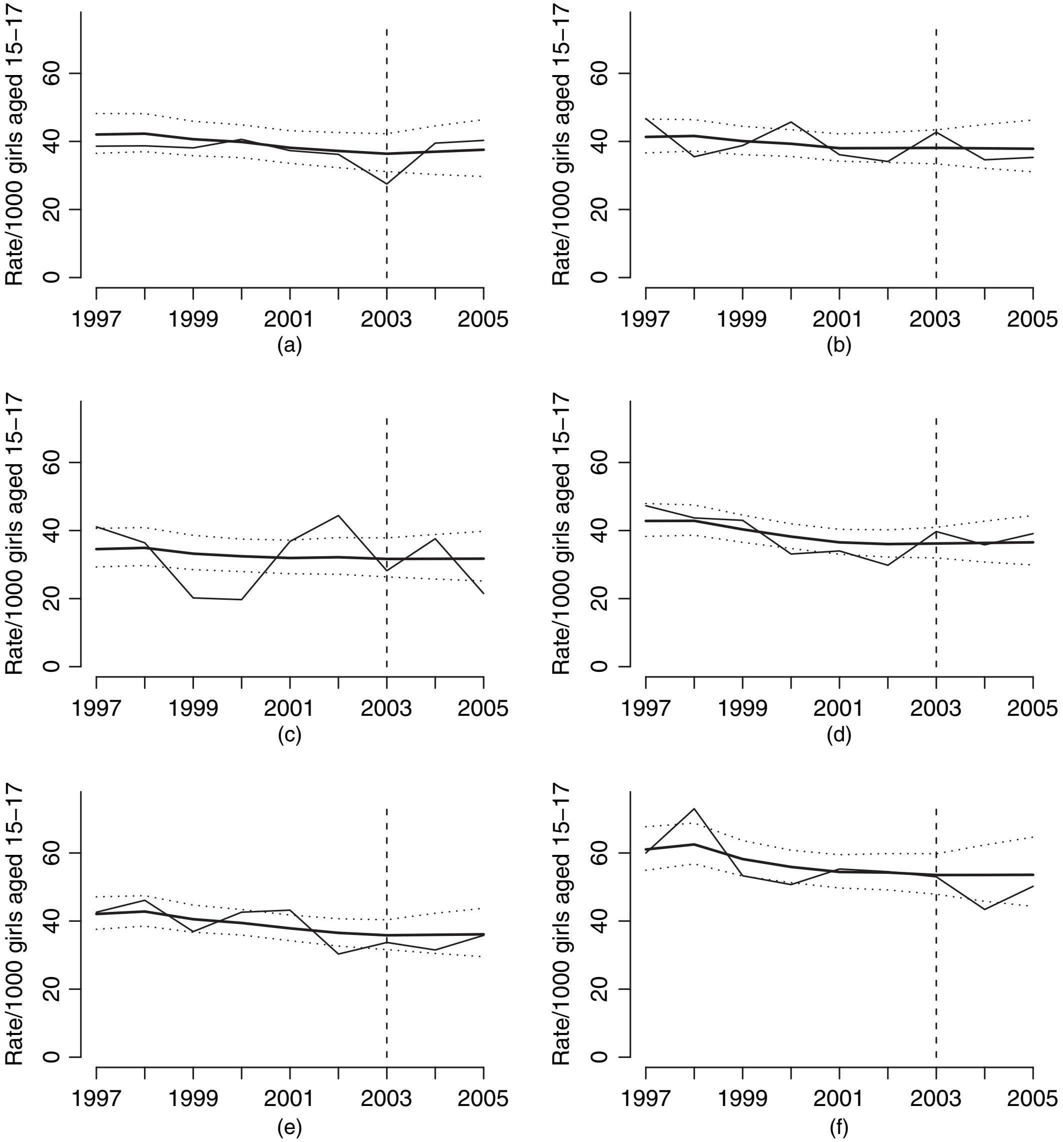

Fig. 3. Plots of teenage conception rates over time (———) for the first six LAs alphabetically, along with the hierarchical AR(1) model posterior means (**———**) and 95% credible intervals (· · · · · · ·) (the final two data points for each LA are reserved for model checking, the plotted values shown for these periods being forecasts for θ_{T-1} and θ_T rather than posterior estimates); this appealing model allows flexible dynamic sequences of rates which are correlated both over time and across providers

second-order dynamic linear model, which is often referred to as a linear growth model (West and Harrison, 1997) for μ_t or $\log(\tau_t)$. This allows observed population trends to be automatically incorporated in predictions. For example, a dramatic population level reduction in MRSA cases was observed between the end of our data period (September 2006) and 2008. We applied an extended version of the model to this updated data set, incorporating a linear growth process for μ_t:

Table 2. Summary statistics comparing the ability of various models to predict MRSA *bacteraemia* rates in October 2005–March 2006 (one step ahead) and April–September 2006 (two steps ahead)†

Smooth	*Model*	*Point predictions*		*Uniformity of p-values*		*Average 90% interval width*	*CRPS*	*Log-score*
		RMSE	χ^2	*D*	W^2			
1 step ahead								
None	Independent Poisson	7.8	461	0.09	0.29	19.7	4.06	3.23
Between providers	Poisson–gamma EB	7.5	402	0.09	0.32	17.8	3.84	3.18
Between providers	Poisson–log-normal full Bayes	7.4	399	0.10	0.42	18.0	3.83	3.18
Within providers	Independent Poisson regressions	7.8	415	0.12‡	0.68‡	17.6	4.05	3.26
Within providers	Independent Poisson random walks	*6.9*	349	0.05	0.08	19.4	3.54	*3.07*
Within providers	Independent Poisson AR(1) processes	7.1	350	0.07	0.16	19.0	3.60	3.09
Between and within providers	Hierarchical GLM	7.2	353	0.13‡	0.73‡	16.6	3.70	3.17
Between and within providers	Hierarchical AR(1) + population processes	*6.9*	*340*	0.06	0.15	18.6	*3.52*	3.08
2 steps ahead								
None	Independent Poisson	7.3	532	0.10	0.39	19.8	3.86	3.26
Between providers	Poisson–gamma EB	6.3	324	0.08	0.21	17.9	3.28	3.09
Between providers	Poisson–log-normal full Bayes	6.4	322	0.10	0.33	18.0	3.30	3.10
Within providers	Independent Poisson regressions	8.9	663	0.14	0.93	19.1	4.49	3.42
Within providers	Independent Poisson random walks	6.8	384	0.06‡	0.16‡	21.7	3.48	3.14
Within providers	Independent Poisson AR(1) processes	*6.2*	*311*	0.09	0.28	20.5	*3.20*	*3.06*
Between and within providers	Hierarchical GLM	7.6	395	0.13‡	1.01‡	17.5	3.83	3.24
Between and within providers	Hierarchical AR(1) + population processes	6.3	326	0.08	0.24	20.4	3.24	3.09

†The lowest values of the RMSE, χ^2, CRPS and log-score for each time period are highlighted in italics. The $\alpha = 0.10$ critical values for $m = 170$ were estimated to be $D = 0.09$ and $W^2 = 0.34$, and the $\alpha = 0.05$ critical values $D = 0.10$ and $W^2 = 0.47$.

‡Rejection of the null hypothesis of uniformity at the 5% level.

Table 3. Comparison of models for the teenage conceptions data, based on their historical fit to the first $T-2=7$ periods of data, as measured by the DIC

Model	$\bar{D}$	p_D	*DIC*
Independent Poisson regressions	18005	704	18709
Independent Poisson random walks	17762	794	18556
Independent Poisson AR(1)	17732	891	18623
Hierarchical GLM	18015	463	18477
Hierarchical AR(1)	17794	620	18414

Table 4. Comparison of models for the MRSA data, based on their historical fit to the first $T-2=9$ periods of data, as measured by the DIC

Model	$\bar{D}$	p_D	*DIC*
Independent Poisson regressions	8973	338	9311
Independent Poisson random walks	8608	496	9104
Independent Poisson AR(1)	8573	540	9114
Hierarchical GLM	8988	260	9248
Hierarchical AR(1)	8598	490	9088

$$\mu_t = \mu_{t-1} + \lambda_{t-1} + \omega_{t1},$$
$$\lambda_t = \lambda_{t-1} + \omega_{t2},$$

where

$$\begin{pmatrix} \omega_{t1} \\ \omega_{t2} \end{pmatrix} \sim N\left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \nu_1^2 & \rho_\mu \nu_1 \nu_2 \\ \rho_\mu \nu_1 \nu_2 & \nu_2^2 \end{pmatrix} \right\}.$$

The results appeared highly promising: Fig. 4 demonstrates that the model naturally adapts to the observed population trend in the data and appropriately takes this trend into account when making predictions for the next time period.

In our main MRSA data set (from April 2001 to September 2006), the posterior parameter estimates of τ_t gave an impression of decreasing slightly over time, indicating that rates across NHS trusts may have been becoming more similar to each other. Fig. 1(b), although only showing rates for the first eight NHS trusts, also gives some indication that this may have been so in the final two periods. It is possible that the predictive ability of the hierarchical AR(1) model for this data set could have been improved by incorporating a linear growth model for $\log(\tau_t)$. Exploratory analyses of this type resulted in convergence problems, but it might be possible to overcome these by assuming some fixed correlation between the two random errors involved in the linear growth model.

7. Conclusions

Our analysis has illustrated the relatively poor predictive ability of independent Poisson regressions with time as a covariate, and also of an extended version of this in which the intercepts and gradients are assumed to be random effects. For these two data sets at least, log-linear

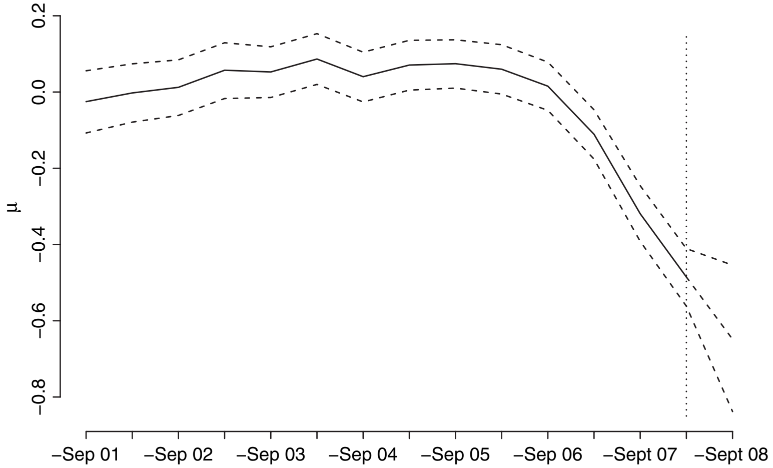

Fig. 4. Hierarchical AR(1) model with linear growth: posterior estimates and 95% intervals for the population mean log-relative-risk μ_t for updated MRSA data; a prediction for μ_t in the next period is plotted beyond the dotted vertical line

time trends for the relative risks do not seem to be realistic, and more flexible alternatives are required. The poor performance of the hierarchical GLM is of particular interest given recommendations of such models in the performance monitoring literature (Gray *et al.*, 1996; Daniels and Gatsonis, 1999; Bronskill *et al.*, 2002).

We note that it is possible to incorporate auto-correlated residuals into a hierarchical GLM (Diggle, 1988; Chi and Reinsel, 1989; Goldstein *et al.*, 1994), thereby relaxing the assumption of conditional independence given the regression line. This is suitable for repeated observations that are measured quite closely together in time, in which the correlations are assumed to be due to temporary common impacts. However, the underlying trend is still assumed to follow a simple polynomial form. For longitudinal performance data such as our examples, we expect some true underlying deviations from a simplistic trend line, and therefore we believe that it is more appropriate to incorporate this explicitly in the model.

The hierarchical time series models that have been suggested in the literature, allowing flexible underlying trends in each unit, appear to be much more promising. We have described a highly intuitive model in which a marginal hierarchical structure is assumed to hold in each time period, whereas the standardized log-relative-risks follow an AR(1) process (Lin *et al.*, 2009). This model was extended to incorporate random walks for the population mean and a measure of between-provider variability, therefore also allowing these parameters to change smoothly over time and predictions to be made automatically. We also discussed how it can be extended further, to incorporate longer-term population level trends.

For the teenage conceptions data, measures of both one and two steps ahead out-of-sample predictive ability and the DIC each demonstrated the increased accuracy of this bidirectional smoothing model relative to any other model considered. For the MRSA data set, the out-of-sample predictive checks suggested comparable performance of the hierarchical AR(1) and simpler independent Poisson AR(1) processes, with the DIC comparisons again slightly favouring the hierarchical AR(1) model.

We note that the MRSA data were the more volatile over time of the two examples that are considered in this paper; the results therefore give some indication that smoothing over time might be more beneficial in more volatile data sets. This seems somewhat counterintuitive, although we note that smoothing a process that is already smooth will have little effect and therefore perhaps not be beneficial. Further, the MRSA data consist of the longer time series out of the two data sets, perhaps therefore offering more potential for benefits.

The results from our two running examples do not provide enough evidence to formulate guidelines on when to smooth within and when between units. In the absence of such well-tested rules and in the light of our results, it seems natural to recommend the hierarchical AR(1) model as a default choice. This model adapts automatically to characteristics of the particular data set, selecting an appropriate degree of smoothing in each direction. It is straightforward to programme in WinBUGS and has the advantage of allowing inferences to be made simultaneously about individual units and population parameters, each of which is allowed to vary smoothly in a non-linear fashion. Long-term trends in these parameters can be incorporated into predictions by using the extension that is outlined in Section 6. We suggest that this intuitive model also seems likely to be applicable to other areas in which multiple correlated time series are observed, such as in economic analyses and spatiotemporal analyses of environmental or epidemiological data.

A major drawback of the models involving a latent auto-regressive process (either the independent or hierarchical versions) is that these each took a long time to fit in WinBUGS, due to high auto-correlations in the Markov chains sampled. This was particularly so for the MRSA data, for which it took several hours (on a Q9500 quad-core central processor unit with 3 Gbytes of random-access memory) for the posterior mean of the deviance from the two chains to achieve a reasonable degree of agreement. Reparameterization of the latent processes in a multivariate normal form is likely to be beneficial, since this would result in the sampler performing block updates which should be more efficient. However, an additional drawback does remain, in that specialist software should ideally not be needed for routine performance monitoring. With this in mind, alternative approaches to fitting similar models (van Houwelingen *et al.*, 2000; McClellan and Staiger, 1999) are of much interest. Ideally, appropriate bidirectional smoothing would be achieved by using only closed form formulae. In particular, the simple two-stage approach of Martz *et al.* (1999) deserves further attention.

Acknowledgement

This work was funded by a Medical Research Council doctoral studentship at the Medical Research Council Biostatistics Unit, Cambridge.

References

Bird, S. M., Cox, D., Farewell, V. T., Goldstein, H., Holt, T. and Smith, P. C. (2005) Performance indicators: good, bad, and ugly. *J. R. Statist. Soc.* A, **168**, 1–27.

Bröcker, J. and Smith, L. A. (2008) From ensemble forecasts to predictive distribution functions. *Tellus*, **60**, 663–678.

Bronskill, S. E., Normand, S.-L. T., Landrum, M. B. and Rosenheck, R. A. (2002) Longitudinal profiles of health care providers. *Statist. Med.*, **21**, 1067–1088.

Burgess, J. F., Christiansen, C. L., Michalak, S. E. and Morris, C. N. (2000a) Medical profiling: improving standards and risk adjustment using hierarchical models. *J. Hlth Econ.*, **19**, 291–309.

Burgess, J. F., Lourdes, V. and West, M. (2000b) Profiling substance abuse provider trends in health care delivery systems. *Hlth Serv. Outcms Res. Methodol.*, **1**, 253–276.

Chi, E. M. and Reinsel, G. C. (1989) Models for longitudinal data with random effects and AR(1) errors. *J. Am. Statist. Ass.*, **84**, 452–459.

Daniels, M. J. and Gatsonis, C. (1999) Hierarchical generalized linear models in the analysis of variations in health care utilization. *J. Am. Statist. Ass.*, **94**, 29–42.

Daniels, M. J. and Normand, S.-L. T. (2006) Longitudinal profiling of health care units based on continuous and discrete patient outcomes. *Biostatistics*, **7**, 1–15.

Denuit, M. and Lambert, P. (2005) Constraints on concordance measures in bivariate discrete data. *J. Multiv. Anal.*, **93**, 40–57.

Department for Children, Schools and Families (2010) Teenage pregnancy strategy: beyond 2010. *Report*. (Available from `http://www.education.gov.uk/consultations/downloadableDocs/4287_Teenage pregnancy strategy_aw8.pdf`.)

Diamond, I., Clements, S., Stone, N. and Ingham, R. (1999) Spatial variation in teenage conceptions in south and west England. *J. R. Statist. Soc.* A, **162**, 273–289.

Diebold, F. X., Gunther, T. A. and Tay, A. S. (1998) Evaluating density forecasts with applications to financial risk management. *Int. Econ. Rev.*, **39**, 863–883.

Diggle, P. J. (1988) An approach to the analysis of repeated measurements. *Biometrics*, **44**, 959–971.

Efron, B. and Morris, C. (1975) Data analysis using Stein's estimator and its generalizations. *J. Am. Statist. Ass.*, **70**, 311–319.

Elder, R., Kapetanios, G., Taylor, T. and Yates, T. (2005) Assessing the MPC's fan charts. *Bnk Engl. Q. Bull.*, autumn, 326–348.

Gneiting, T., Balabdaoui, F. and Raftery, A. E. (2007) Probabilistic forecasts, calibration and sharpness. *J. R. Statist. Soc.* B, **69**, 243–268.

Gneiting, T. and Raftery, A. E. (2007) Strictly proper scoring rules prediction and estimation. *J. Am. Statist. Ass.*, **102**, 359–378.

Gneiting, T., Raftery, A. E., Westveld III, A. H. and Goldman, T. (2005) Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Mnthly Weath. Rev.*, **133**, 1098–1118.

Goldstein, H., Healy, M. J. R. and Rasbash, J. (1994) Multilevel time series models with applications to repeated measures data. *Statist. Med.*, **13**, 1643–1655.

Goldstein, H. and Spiegelhalter, D. J. (1996) League tables and their limitations: statistical issues in comparisons of institutional performance (with discussion). *J. R. Statist. Soc.* A, **159**, 385–443.

Gray, J., Goldstein, H. and Jesson, D. (1996) Changes and improvements in schools' effectiveness: trends over five years. *Res. Pap. Educ.*, **11**, 35–51.

Gray, J., Jesson, D., Goldstein, H., Hedger, K. and Rasbash, J. (1995) A multi-level analysis of school improvement: changes in schools' performance over time. *School Effectiv. School Imprvmnt*, **6**, 97–114.

Greenland, S. and Robins, J. M. (1991) Empirical-Bayes adjustments for multiple comparisons are sometimes useful. *Epidemiology*, **2**, 244–251.

Grigg, O. A. and Spiegelhalter, D. J. (2007) A simple risk-adjusted exponentially weighted moving average. *J. Am. Statist. Ass.*, **102**, 140–152.

Hay, J. L. and Pettitt, A. N. (2001) Bayesian analysis of a time series of counts with covariates: an application to the control of an infectious disease. *Biostatistics*, **2**, 433–444.

Heisterkamp, S. H., Dekkers, A. L. M. and Heijne, J. C. M. (2006) Automated detection of infectious disease outbreaks: hierarchical time series models. *Statist. Med.*, **25**, 4179–4196.

Hersbach, H. (2000) Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weath. Forecast.*, **15**, 559–570.

van Houwelingen, H. C., Brand, R. and Louis, T. A. (2000) Empirical Bayes methods for monitoring health care quality. *Technical Report*. Leiden University Medical Center, Leiden. (Available from `http://www.msbi.nl/dnn/People/Houwelingen/Publications/tabid/158/Default.aspx`.)

Howley, P. P. and Gibberd, R. (2003) Using hierarchical models to analyse clinical indicators: a comparison of the gamma-Poisson and beta-binomial models. *Int. J. Qual. Hlth Care*, **15**, 319–329.

James, W. and Stein, C. (1961) Estimation with quadratic loss. In *Proc. 4th Berkeley Symp. Mathematical Statistics and Probability*, vol. 1 (ed. J. Neyman), pp. 361–379. Berkeley: University of California Press.

Jensen, S. T., McShane, B. and Wyner, A. J. (2009) Hierarchical Bayesian modeling of hitting performance in baseball. *Baysn Anal.*, **4**, 191–212.

Jones, H. E. and Spiegelhalter, D. J. (2009) Accounting for regression-to-the-mean in tests for recent changes in institutional performance: analysis and power. *Statist. Med.*, **28**, 1645–1667.

Leckie, G. and Goldstein, H. (2009) The limitations of using school league tables to inform school choice. *J. R. Statist. Soc.* A, **172**, 835–851.

Liang, K.-Y. and Zeger, S. L. (1986) Longitudinal data analysis using generalized linear models. *Biometrika*, **73**, 13–22.

Lin, R., Louis, T. A., Paddock, S. M. and Ridgeway, G. (2009) Ranking USRDS provider-specific SMRs from 1998-2001. *Hlth Serv. Outcms Res. Methodol.*, **9**, 22–38.

Marshall, G., Shroyer, A. L. W., Grover, F. L. and Hammermeister, K. E. (1998) Time series monitors of outcomes: a new dimension for measuring quality of care. *Med. Care*, **36**, 348–356.

Martz, H. F., Parker, R. L. and Rasmuson, D. M. (1999) Estimation of trends in the scram rate at nuclear power plants. *Technometrics*, **41**, 352–364.

McClellan, M. and Staiger, D. (1999) The quality of health care providers. *Technical Report*. National Bureau of Economic Research, Cambridge. (Available from `http://www.nber.org/papers/w7327`.)

Morris, C. N. and Christiansen, C. L. (1996) Hierarchical models for ranking and for identifying extremes, with applications. In *Bayesian Statistics 5* (eds J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith), pp. 277–296. Oxford: Oxford University Press.

Normand, S.-L. T., Glickman, M. E. and Gatsonis, C. A. (1997) Statistical methods for profiling providers of medical care: issues and applications. *J. Am. Statist. Ass.*, **92**, 803–814.

Office of the Deputy Prime Minister (2004) The English indices of deprivation 2004: summary (revised). *Report*. Office of the Deputy Prime Minister, London. (Available from `http://www.communities.gov.uk/documents/communities/pdf/131209.pdf`.)

Spiegelhalter, D. J. (1986) Probabilistic prediction in patient management and clinical trials. *Statist. Med.*, **5**, 421–433.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and van der Linde, A. (2002) Bayesian measures of model complexity and fit (with discussion). *J. R. Statist. Soc.* B, **64**, 583–639.

Stephens, M. A. (1974) EDF statistics for goodness of fit and some comparisons. *J. Am. Statist. Ass.*, **69**, 730–737.

West, M. and Aguilar, O. (1998) Studies of quality monitor time series: the V.A. hospital system. *Technical Report*. Duke University, Durham. (Available from `http://ftp.isds.duke.edu/WorkingPapers/97-22a.pdf`.)

West, M. and Harrison, J. (1997) *Bayesian Forecasting and Dynamic Models*, 2nd edn. New York: Springer.

Wilkinson, P., French, R., Kane, R., Lachowycz, K., Stephenson, J., Grundy, C., Jacklin, P., Kingori, P., Stevens, M. and Wellings, K. (2006) Teenage conceptions, abortions, and births in England, 1994-2003, and the national teenage pregnancy strategy. *Lancet*, **368**, 1879–1886.

Willms, J. D. and Raudenbush, S. W. (1989) A longitudinal hierarchical model for estimating school effects and their stability. *J. Educ. Measmnt*, **26**, 209–232.

Winkler, R. L. (1971) Probabilistic prediction: some experimental results. *J. Am. Statist. Ass.*, **66**, 675–685.