

Simultaneous modelling of multiple traffic safety performance indicators by using a multivariate generalized linear mixed model

Trevor C. Bailey and Paul J. Hewson

University of Exeter, UK

[Received February 2003. Revised December 2003]

Summary. Traffic safety in the UK is one of the increasing number of areas where central government sets targets based on ‘outcome-focused’ performance indicators (PIs). Judgments about such PIs are often based solely on rankings of raw indicators and simple league tables dominate centrally published analyses. There is a considerable statistical literature examining health and education issues which has tended to use the generalized linear mixed model (GLMM) to address variability in the data when drawing inferences about relative performance from headline PIs. This methodology could obviously be applied in contexts such as traffic safety. However, when such models are applied to the fairly crude data sets that are currently available, the interval estimates generated, e.g. in respect of rankings, are often too broad to allow much real differentiation between the traffic safety performance of the units that are being considered. Such results sit uncomfortably with the ethos of ‘performance management’ and raise the question of whether the inference from such data sets about relative performance can be improved in some way. Motivated by consideration of a set of nine road safety performance indicators measured on English local authorities in the year 2000, the paper considers methods to strengthen the weak inference that is obtained from GLMMs of individual indicators by simultaneous, multivariate modelling of a range of related indicators. The correlation structure between indicators is used to reduce the uncertainty that is associated with rankings of any one of the individual indicators. The results demonstrate that credible intervals can be substantially narrowed by the use of the multivariate GLMM approach and that multivariate modelling of multiple PIs may therefore have considerable potential for introducing more robust and realistic assessments of differential performance in some contexts.

Keywords: Generalized linear mixed model; Markov chain Monte Carlo methods; Multivariate modelling; Performance indicators; Road casualties; Traffic safety

1. Background

Performance ‘management’ is currently a high profile activity throughout many parts of the public sector in the UK. Traffic safety is no exception and is typical of activities which are monitored by performance indicators (PIs) based on ‘outcome measures’. In particular, the UK Government has identified three traffic safety targets which are expected to be achieved at a national level by 2010 (relative to a base-line of the mean number of casualties that were reported between 1994 and 1998 inclusively): a 40% reduction in the number of fatally and seriously injured casualties, a 10% reduction in the rate of slight casualties relative to traffic and a 50% reduction in the numbers of children who were fatally and seriously injured. Related performance measures are monitored at the level of each highways authority and published

Address for correspondence: Paul Hewson, Department of Mathematical Sciences, Laver Building, University of Exeter, North Park Road, Exeter, EX4 4QE, UK.
E-mail: p.j.hewson@exeter.ac.uk

in authority league tables under the 'best value' requirements of the Local Government Act of 1999, which specifies that each highways authority sets local performance targets for traffic safety broken down by modal group (pedestrians, pedal cyclists, car occupants, motorcyclists and other vehicle users).

However, current UK traffic safety PIs continue to be expressed simply in the form of crude per capita numbers of reported collisions by type and modal group, with no allowance for geographically differing patterns in road infrastructure and usage or other relevant unmeasured factors, and with no explicit consideration given to the extent to which differences in the raw rates reflect differential performance, rather than just inherent random variability in observed rates. In these respects, current UK traffic safety PIs are depressingly similar to the crude nature of those that are used in relation to many other types of functions carried out by local government in the UK. They reflect an apparent general lack of effort expended on the part of the agencies who are charged with assessing English local authority performance to attempt to distinguish between outlying organizations and organizations which merely happen to occupy the extreme ends of a league table in any given time period.

In general, local government activity does not appear to have received anything like the same level of attention in the literature as that devoted to performance monitoring in other sectors. For example, although local government function in the UK clearly plays a significant role in education, performance monitoring interest in that sector has largely focused on the school as the institutional unit, rather than the local education authority and, unfortunately, the considerable work on modelling variability in school PIs (e.g. Goldstein and Spiegelhalter (1996)) has not migrated into the more general sphere of local government league comparisons.

Although it is of course appreciated that establishing the statistical 'significance' of a departure in traffic safety performance between observational units is simply one part of the wider context of performance management in that sector, it is nevertheless important to make the best statistical inference about differential performance in the first instance and current practice in the UK could undoubtedly be improved in that respect. MacNab (2003), for example, used generalized additive models to smooth year-by-year fluctuations in road-collision-related casualty rates when comparing Canadian Health Board areas and has emphasized the need to distinguish between signal and noise when investigating this kind of data. More broadly, the generalized linear mixed model (GLMM) is now well established in the literature as a means of modelling the uncertainty that is associated with performance judgments based on PIs (e.g. Morris and Christiansen (1996) and Goldstein and Spiegelhalter (1996)). The GLMM approach is an obvious candidate for use in the local government arena and particularly in the traffic safety context.

GLMMs are characterized by inclusion of 'random effects' in addition to the fixed parameters which comprise a conventional generalized linear model. The use of a mixture of both random and fixed effects in statistical modelling to allow for a variety of variance components is now common in many fields and there are several highly accessible introductions to such models for non-statisticians (e.g. Marshall and Spiegelhalter (1998)). In the context of modelling UK traffic safety PIs, the basic idea would be to propose a model for the observed performance outcome measure that incorporates a random effect for each authority. This may be thought of as a latent variable representing a difference in performance outcome for that authority that has not been directly measured and which needs to be estimated. In contrast with an authority-specific fixed parameter, which would simply result in the model estimates exactly reproducing the observed PI value for each authority, the random effects for each authority are considered to be realized values of an underlying random variable having a probability distribution with a zero mean but an unknown variance, the latter being estimated as part of the modelling procedure along

with the estimates of each of the random effects. The model thus acknowledges the presence of a variance component, or source of uncertainty, that is associated with the authority-specific performance effects (possibly relating to unmeasured explanatory factors such as differential exposure, differential recording practices, data quality etc.) and this variance component induces a smoothing into the estimates of the authority-specific effects. Essentially, estimates of these random effects borrow strength from each other and are each 'shrunk' towards their global mean (i.e. zero), the amount of shrinkage for any authority depending on the strength of evidence in the data pertaining to that authority and the estimated size of the variance component that is associated with the random effects.

Mathematical details of a GLMM which is potentially applicable to individual traffic safety indicators are given in Section 3 and results obtained from its use are illustrated in Section 4. However, as seen there, this established GLMM approach does not prove particularly fruitful in differentiating authority performance when straightforwardly applied to any of the individual UK traffic safety PIs which are currently available. The crude nature of the available data means that, when random effects are allowed for, the associated variance component is large and the conventional univariate model for any single traffic safety indicator fails to provide much evidence of differential performance. In other words, the credible intervals on predicted performance rankings simply overlap in the majority of cases. One conclusion from such results might be simply that little can be established about comparative performance of highways authorities from the available data and we suspect that similar conclusions might also apply to PIs which are used in relation to many other types of functions carried out by local government in the UK. However, this not only runs counter to the political ethos of performance management but also provides a statistical challenge about whether better use can be made of the available data.

Various modifications to the basic univariate GLMM described above are possible which may potentially improve its discriminatory power. For example, one could introduce some non-exchangeability between groups of authorities (e.g. rural *versus* urban) via separate variance components for these groups. However, in this paper, we also consider a complementary approach which involves multivariate modelling of a range of traffic safety PIs simultaneously. Papageorgiou and Loukas (1988) have considered a bivariate negative binomial model for traffic safety in East Virginia, on the basis that fatality counts should be correlated with injury counts. More generally, it seems intuitively reasonable to believe that there should be some underlying correlation between traffic safety indicators. This opens up the possibility of simultaneous modelling of multiple indicators to improve the strength of the predictions about relative performance that may be obtained by using any one of them in isolation. Accordingly the methods that we suggest here seek to gain more insight into the uncertainty estimates that are associated with individual traffic safety indicators by borrowing strength from related variables. Essentially, authority-specific random effects are allowed to be correlated between variables and their estimates can thus be shrunk across variables to gain increased precision, the degree of shrinkage depending on the correlation structure between variables which is estimated as part of the modelling process. Such a multivariate modelling approach may potentially result in either a narrowing of credible intervals on rankings for any individual PI, thus allowing more discriminatory judgments of performance (in the case where several related PIs agree on differential performance), or it may reinforce the uncertainty of rankings obtained from univariate models (in the case where related PIs contradict each other). Either way, a simultaneous multivariate GLMM would provide valuable additional information when making assessments of relative performance not only in the traffic safety arena but also potentially in other sectors where multiple PIs are employed.

Our proposed approach differs from various other multivariate approaches to the analysis of PIs that have been suggested in the literature and which essentially collapse a range of measures onto a single dimension that is then used to assess the overall performance of an organization. For example Smith (1990) reviewed PIs in use in the public sector in the UK and referred to several studies that had been carried out using classical multivariate approaches. More recently, Stone (2002) reviewed the application of the popular technique of data envelopment analysis in the context of the police service and also discussed 'value-based analysis', which involves a judgmental weighted sum of PIs. There are also various other related approaches.

2. Data

Road casualties per capita by severity and by type of road user are reported by the UK Audit Commission as constituting a 'best value performance indicator' (BVPI). Such indicators are required by central government under the terms of the 1999 Local Government Act (Department for Transport, Local Government and the Regions, 1999). There are nearly 200 such BVPIs applied to various distinct tiers of local government. The road casualty BVPI, catalogued as BVPI 99, is produced for 87 highways authorities throughout England. The Office of the Deputy Prime Minister (2003) have made an interactive Web site available which allows a limited investigation of the published values of the indicator. Under the terms of the Local Government Act 1999, authorities are required to publish details of these indicators in 'best value performance plans' made available to the public. The indicators that are contained in these plans are collated centrally and league tables are drawn up based on them.

Here, we consider a subset of BVPI 99 involving those indicators relating solely to 'vulnerable' road users. It has been suggested that analyses of such indicators can be particularly difficult owing to the small numbers of casualties involved. This problem is illustrated in Table 1 which gives a summary of the nine variables that we shall consider here. There are three road user types, pedestrians, pedal cyclists and motor-cyclists, and three severities of injury for each, namely fatal, serious and slight. One possible solution to the problem of small numbers has been to suggest that casualty counts are aggregated, either by adding together a number of road user types or by adding a number of severities together. Typically, as in BVPI 99, the fatal and serious casualties are aggregated. The Parliamentary Advisory Council for Transport Safety (2003) have argued that BVPI 99 should also merge road user types as a way of limiting the large amount of random fluctuations that are seen with such small numbers. However, there is also a particular and growing interest in examining performance with respect to 'vulnerable' road users in more detail. Central government has promised to reduce congestion (Department for Transport, Local Government and the Regions, 1997) as well as making a commitment to a 10-year plan for integrated transport (Department for Transport, Local Government and the Regions, 1998). Consequently, there is a need to motivate people to travel less by car and more

Table 1. Median and (lower quartile, upper quartile) number of reported casualties per highways authority 2000

<i>Mode</i>	<i>Fatal</i>	<i>Serious</i>	<i>Slight</i>
Pedestrians	4 (2, 10)	39 (22, 70)	149 (73, 237)
Motor-cycle	4 (1, 9)	34 (20.5, 87)	98 (57, 223.5)
Pedal cycle	1 (0, 2)	17 (7, 34)	111 (55, 211)

by other forms of transport. In addition to transportation requirements there are health reasons for promoting life style changes, e.g. encouraging walking and cycling to reduce coronary heart disease (Department of Health, 1999). Both of these objectives would imply an increase in the amount of vulnerable road user traffic. At the same time there are commitments, both in terms of health (Department of Health, 1999) and transport (Department for Transport, Local Government and the Regions, 1998) to reduce injury casualties. Scrutiny of PIs relating to vulnerable road users will therefore become an area of growing priority over the next few years.

The use of casualty counts in reported road collisions as a means of measuring the performance of highways authorities is obviously problematic. One can imagine considerable potential for confounding in the data; for example exposure by mode (rates of walking, cycling and motor-cycling) will vary around the county and enforcement activities by the police may vary. It is therefore particularly difficult to know to what extent departures in observed rates may be due to the nature of the work of the authorities being monitored for performance. Nevertheless, the Cabinet Office (2000) have emphasized the need for 'outcome-focused delivery' rather than 'organization-focused delivery' and argue a case for using PIs to manage this activity. It is therefore clearly necessary to evaluate such indicators in detail both from a statistical point of view in terms of allowing for underlying uncertainty and ultimately to help to identify whether outliers are performance related or due to factors that are beyond the control of the authority.

As described, BVPI 99 is not directly suitable for further modelling because it is published solely as a per capita rate for each authority without the constituent counts. However, one of the advantages of examining this particular data set is that it is also possible to access the source data and to obtain the actual casualty counts and population estimates corresponding to the numerator and denominator of the indicator.

The denominator data that were used in BVPI 99 comprise population estimates for each highways authority taken from Office for National Statistics mid-year estimates for each constituent authority and these are readily available. It is interesting to note in passing that the populations of the 87 highways authorities vary considerably with a median of 284000 and upper and lower quartiles of 163000 and 629000 respectively.

The numerator data for BVPI 99 may be obtained from the Data Archive at Essex University. These data are based on the current system for reporting road collisions within Great Britain. Data on road collisions were first collected in the UK in 1919 but a formalized regime was established in 1949 (Wilding, 2002). Arrangements are made by the local processing authority (which may be the police, local authority or subcontractor, depending on local arrangements) to return these data to the Department for Transport. The resulting raw data are referred to as 'STATs 19 data' (the name of the form which is initially completed) and are summarized for public consumption (Department for Transport, 2001). These raw data are used here to provide the casualty counts by severity and road user type for each highways authority. One of the features of the road casualty PI is that it relates, for a given financial year, to casualties reported in road collisions during the calendar year which ended 15 months before that financial year. In other words, casualties occurring during calendar year 2000 as in the data which we consider here are reported in the 2001–2002 financial year PI. The reliability of the STATs 19 data merits consideration since, in common with much data that are derived from administrative systems, it is not collected to the rigorous standards that are associated with randomized controlled studies and that can lead to various biases as recently illustrated in the context of hospital episode statistics (Spiegelhalter *et al.*, 2002). One of the earlier accounts which considered the reliability of STATs 19 data by comparing them with hospital records is now over three decades old (Bull and Roberts, 1973) and 25 further studies were reviewed by James (1991). The consensus appears

to suggest that single-vehicle collisions, collisions involving more vulnerable road users, young road users and less seriously injured road users tend to be under-reported. In addition, the Transport Research Laboratory have conducted a long-term study in Scotland matching hospital records with STATs 19 reports (Keigan *et al.*, 1999) and similar work has been reported in England (Cryer *et al.*, 2001). These studies tend to suggest that, whereas all fatal collisions are reported, only around 60% of slight collisions are reported. In the context of performance management such patterns of under-reporting may present problems, as lower diligence in recording road collisions would equate to an apparently 'better' set of reported PIs. However, the burden of reporting falls on the police forces and not individual highways authorities. There is no *a priori* reason to believe that reporting patterns differ significantly between police forces and, if they did, a similar pattern of under-reporting would apply to all highways authorities within a given force area. Thus, although acknowledged deficiencies in STATs 19 data will add to variability, these should not introduce systematic bias into PI comparisons between highways authorities.

For a given collision type and road user, an expected casualty count for each authority can be calculated to create a convenient model offset from the data sources that are discussed above based on the product of the overall per capita relevant casualty rate for all authorities and the population estimate for the authority in question. The published PI for each authority is then the ratio of actual to expected casualties. It has been noted, for example, by Woodward (1983) that motor-cycle ownership may be a far more appropriate denominator for certain of these rates. Similarly, one would assume that measures of pedestrian and cyclist activity would also be a more appropriate denominator for those particular collision rates, but the National Travel Survey only measures such activity regionally, and not by authority. Given that, and the fact that the BVPI data set continues to use per capita rates, it would seem appropriate for the purposes of this paper to use the indicator as defined.

3. Models

Given observed count data such as those which were described in Section 2, a natural model to use for any individual casualty count Y_i in authority i ($i = 1, \dots, n$) is the Poisson model: $Y_i \sim \text{Poisson}(e_i \lambda_i)$, where λ_i is the relative performance of authority i and the known offset $e_i = rN_i$ is the corresponding expected casualty count in that authority, formed from the product of a fixed reference performance r (typically the overall casualty rate for all authorities) and the population N_i of authority i . At the simplest level λ_i is then typically modelled as a log-linear function: $\log(\lambda_i) = \beta + \nu_i$ where ν_i are authority-specific random effects such that $\nu_i \sim N(0, \sigma^2)$, leading to a Poisson-log-normal GLMM. It would clearly be possible to include covariates, however structured, in the linear predictor of this kind of GLMM if available and appropriate.

The essential motivation for the use of random effects in modelling PIs and the corresponding interpretation of these effects has already been discussed in Section 1. An alternative view to that is that this simple univariate GLMM provides a means of modelling overdispersion in the response variable (Hinde, 1982). Observed counts often exhibit higher variability than that imposed by the equality of mean and variance which is implicit in a simple Poisson assumption and it is common to include some allowance for this overdispersion when modelling such data. Indeed McCullagh and Nelder (1989) suggested that it may be wise to consider overdispersion the norm with observed count data and it is easy to envisage how this would apply to road casualties. Collisions can be imagined as arising from a clustered process, where the rate varies according to the type of road, hour of day, time of year and specific road lay-out (e.g. a junction). Issues of data quality in published UK traffic PIs, as discussed in the previous section, will also tend to add to this overdispersion. The particular GLMM that was presented earlier is only

one way to allow for such overdispersion. Vogt and Bared (1998) described use of the negative binomial model in considering US road collision data and there are a range of related models which might be used to allow for overdispersion (for example Vistisen (2002) has developed models using the *h*-likelihood approach of Lee and Nelder (1996)). However, in the case of traffic safety data there would appear to be little practical difference between using the Poisson log-normal *versus* negative binomial models (Tunaru, 1999) and we adopt the former approach here.

The Poisson–log-normal model can be fitted by using an EM algorithm, but a Bayesian approach in conjunction with Markov chain Monte Carlo (MCMC) methods facilitates the derivation of interval estimates for ranks predicted from the model (e.g. Goldstein and Spiegelhalter (1996)) and, since that is of particular interest in the analysis of PIs, we adopt the MCMC approach here, using a non-informative gamma hyperprior for the random-effect variance σ^2 and a vague normal prior for the intercept β .

Turning to possible multivariate as opposed to univariate models for traffic safety PIs, Bhattacharya (1967) has considered a bivariate negative binomial model to correct for individual accident proneness and, as discussed in Section 1, Papageorgiou and Loukas (1988) have reported different methods of fitting bivariate negative binomial models to collision data collected on a 50-mile highway segment in Eastern Virginia. Such work on bivariate road collision counts provides good precedents to consider the more general modelling of correlations between sets of traffic safety PIs. Fig. 1 presents pairwise scatterplots on a log-scale for the observed casualty rates on the nine BVPI 99 variables that were introduced in the previous section and exhibits clear evidence of correlation between some of these variables. The higher observed correlations include $\text{corr}_{23}(\text{pedestrian serious, pedestrian slight}) = 0.67$, $\text{corr}_{78}(\text{motor-cycle fatal, motor-cycle serious}) = 0.49$, $\text{corr}_{89}(\text{motor-cycle serious, motor-cycle slight}) = 0.41$ and between variables from different modal groups $\text{corr}_{68}(\text{pedal cycle slight, motor-cycle slight}) = 0.46$ and $\text{corr}_{59}(\text{pedal cycle slight, motor-cycle serious}) = 0.44$. These observed correlations combined with the observation of Aitchison and Ho (1989) that the range of observed correlations (denoted here by *corr*) between counts arising from a Poisson–log-normal model will by definition be lower than the correlations between the log-normal random effects (denoted later by ρ) which induce them provide sound motivation for attempting to model the latent correlation structure between authority-specific random effects relating to different traffic safety PIs.

A particular advantage of adopting the Poisson–log-normal univariate model, rather than a negative binomial model (which equates to a Poisson distribution with mean λ_i combined with a gamma distribution for λ_i (Lawless, 1987)), is that the log-normal assumption for random effects lends itself to a straightforward extension of the univariate model to a multivariate version which encompasses a richer range of possible correlation structures between response variables. If p PIs are being modelled and Y_{ij} denotes the set of casualty counts, where j denotes count type, $j = 1, \dots, p$, and i denotes authority, $i = 1, \dots, n$, then a suitable multivariate model would be $Y_{ij} \sim \text{Poisson}(\lambda_{ij}e_{ij})$, where the known offset $e_{ij} = r_j N_i$ is the expected casualty count of type j and where $\log(\lambda_{ij}) = \beta_j + \nu_{ij}$ with the random effects for any authority, $\nu_{i1} \dots \nu_{ip}$, assumed to be distributed as multivariate normal with a zero mean and an unknown $p \times p$ variance–covariance matrix Σ , i.e. $(\nu_{i1} \dots \nu_{ip}) \sim \text{MVN}(\mathbf{0}, \Sigma)$.

This model has an advantage over previous studies employing bivariate negative binomial models in that negative as well as positive correlation can be modelled. Such multivariate Poisson–log-normal models were first proposed in the literature by Aitchison and Ho (1989) who then used Gauss–Hermite quadrature to fit a marginal model. Gueorguieva (2001) published an example using Joe and Xu’s (1996) inference function for margins. However, as in the univariate case, MCMC sampling is also a powerful alternative fitting algorithm. Chib and Winkelmann

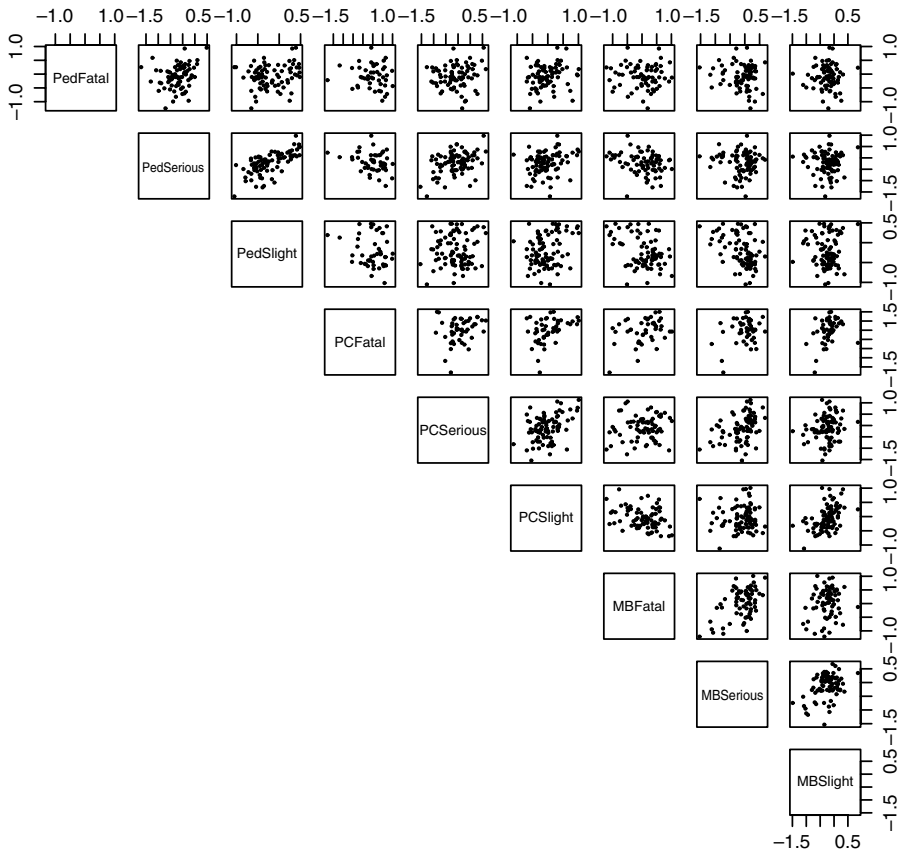


Fig. 1. Bivariate log-log-scatterplots denoting casualty rates for each of the nine variables

(2001) proposed a Metropolis–Hastings algorithm for modelling bivariate Poisson–log-normal data; similarly bivariate data have also been modelled in WinBUGS (Congdon, 2001). Here we used the MCMC approach and employed WinBUGS (Spiegelhalter *et al.*, 1998). We took independent vague normal priors for β_j and a non-informative Wishart hyperprior for Σ , i.e. $\Sigma \sim \text{Wishart}(\mathbf{B}, d)$, where \mathbf{B} is a prespecified $p \times p$ matrix reflecting vague prior correlation beliefs and d denotes the associated degrees of freedom. The degrees of freedom for the Wishart hyperprior must be equal to or greater than the number of variables being modelled and we adopted the least informative choice corresponding to $d = p$. A range of Wishart hyperpriors were used in fitting each of the models, but all gave similar results. Correlation estimates that are reported in this paper are based on a hyperprior with \mathbf{B} having values of 0.01 on the diagonals and 0.005 on the off-diagonals. An alternative parameterization of the multivariate normal hyperprior in terms of a series of conditional univariate normals has been suggested by Congdon (2001). This allows the correlation parameters between each pair of outcomes to be modelled directly, making it potentially easier to specify a prior structure. We experimented with this formulation to provide a useful check on the sensitivity of the model to the specification of the Wishart hyperprior. However, in practice we found that, as the number of dimensions increased, a model adopting this parameterization was increasingly cumbersome to fit and that with our number of variables the Wishart approach was preferred as more computationally viable.

Within the above framework, various possible multivariate model formulations were considered. First, we considered a model involving all variables ($p=9$) and also one involving only the fatal and serious casualty rates ($p=6$). The motivation here was to provide a check on whether predicted performance rankings that were obtained from the nine-variable model were being dominated by what might arguably be the least reliable data. On the one hand, the reliability of the data may decrease as the severity of casualty decreases, but, on the other, the relatively larger numbers of slight injuries may tend to dominate the model fitting. Second, we explored the exchangeability assumptions that are implicit in the model. It may not be the case that all authorities can be regarded as being drawn from an essentially comparable population. One of the most obvious relevant categorizations is a rural–urban split and we therefore considered a model using two variance–covariance matrices, i.e. $(\nu_{i1} \dots \nu_{ip}) \sim \text{MVN}(\mathbf{0}, \Sigma_{\text{rural}})$ for predominantly rural authorities and $(\nu_{i1} \dots \nu_{ip}) \sim \text{MVN}(\mathbf{0}, \Sigma_{\text{urban}})$ for predominantly urban authorities, with respective hyperpriors $\Sigma_{\text{rural}} \sim \text{Wishart}(\mathbf{B}_{\text{rural}}, p)$ and $\Sigma_{\text{urban}} \sim \text{Wishart}(\mathbf{B}_{\text{urban}}, p)$.

4. Results

4.1. Model diagnostics

Model fitting in each case was based on running three chains to check for mixing, thinning by a factor of 10 and running the sampler for a burn-in period of 5000 iterations (chosen according to convergence checks including Gelman's R (Gelman *et al.*, 2003)). Model fit was assessed by means of the deviance information criterion (DIC) that was suggested by Spiegelhalter *et al.* (2002) which attempts to penalize the model fit (deviance) for the model complexity by estimating the effective number of parameters (pD). Generally, smaller values of the DIC suggest better fitting models. The nine-variable model with full exchangeability between authorities had a DIC of 4986.4 whereas the nine-variable model with exchangeability only between rural and urban authorities had a DIC of 5187.6, suggesting a preference for the simpler model. The six-variable model, comprising only the fatal and serious casualty rates and with full exchangeability between authorities, had a DIC of 2860.1. Although information criteria do not allow direct comparisons between models with different numbers of response variables, these results do provide some reassurance about the consistency of the various models.

4.2. Estimates of correlation

As outlined in Section 1, the success of a multivariate as opposed to a univariate approach relies on shrinkage of random effects across variables, and the degree to which that occurs will be related to the estimates of the variance–covariance structure. Estimates of the model correlation matrix therefore merit some consideration along with their sensitivity to prior assumptions.

Posterior mean estimates of correlations between random effects for the reduced variable set (six variables relating to fatal and serious casualties) are presented in Table 2 and those for the nine-variable fully exchangeable model are presented in Table 3.

Given the non-informative nature of the hyperprior, posterior correlation estimates which embrace both high positive and fairly high negative values, as well as low correlations, provide some reassurance that the hyperprior has not dominated the model to an unacceptable extent. Comparing the correlation estimates for the six- and nine-variable models also reveals some consistency between the correlation estimates which would tend to suggest that correlation estimates in the nine-variable model are not unreasonably dominated by the most unreliable data—the slight casualty rates.

Table 2. Means of estimates of the latent correlation ρ derived from the posterior of the variance–covariance matrix for the reduced model (only including variables for fatal and serious casualty rates)

Variable	Estimates for the following variables:					
	1	2	4	5	7	8
1 (pedestrian fatal)	1.00	0.87	−0.06	0.51	−0.24	0.31
2 (pedestrian serious)	0.87	1.00	−0.21	0.55	−0.47	0.17
4 (pedal cycle fatal)	−0.06	−0.21	1.00	0.55	0.79	0.78
5 (pedal cycle serious)	0.51	0.55	0.55	1.00	0.10	0.61
7 (motor-cycle fatal)	−0.24	−0.47	0.79	0.10	1.00	0.71
8 (motor-cycle serious)	0.31	0.17	0.78	0.61	0.71	1.00

Table 3. Means of estimates of the latent correlation ρ derived from the posterior of the variance–covariance matrix with all nine variables where all authorities are assumed exchangeable

Variable	Estimates for the following variables:								
	1	2	3	4	5	6	7	8	9
1 (pedestrian fatal)	1.00	0.92	0.78	−0.07	0.49	0.36	−0.46	−0.07	0.14
2 (pedestrian serious)	0.92	1.00	0.78	−0.20	0.49	0.25	−0.48	−0.14	−0.06
3 (pedestrian slight)	0.78	0.78	1.00	−0.34	0.05	0.34	−0.79	−0.56	0.05
4 (pedal cycle fatal)	−0.07	−0.20	−0.34	1.00	0.63	0.65	0.24	0.64	0.63
5 (pedal cycle serious)	0.49	0.49	0.05	0.63	1.00	0.60	−0.01	0.53	0.24
6 (pedal cycle slight)	0.36	0.25	0.34	0.65	0.60	1.00	−0.52	0.01	0.51
7 (motor-cycle fatal)	−0.46	−0.48	−0.79	0.24	−0.01	−0.52	1.00	0.77	0.11
8 (motor-cycle serious)	−0.07	−0.14	−0.56	0.64	0.53	0.01	0.77	1.00	0.48
9 (motor-cycle slight)	0.14	−0.06	0.05	0.63	0.24	0.51	0.11	0.48	1.00

A more detailed examination of the correlation matrix from the nine-variable model in Table 3 suggests that correlation tends to be much higher among variables relating to the same type of road user. There are some notable exceptions to this, and there are clearly some estimates of negative correlation. There may be some suggestion of separate component structures for non-motorized transport and for two-wheeled transport and this is supported by a principal components analysis of the correlation matrix.

An interesting comparison is that between correlation estimates when rural and urban authorities are split as presented in Tables 4 and 5. There are some similarities between the two sets of estimates. However, it appears that there is higher correlation between pedestrian casualty rates in the urban authorities, a higher correlation between motor-cycle casualty rates in rural authorities and a striking contrast between pedal cycle fatal and serious casualty rates when comparing the two. The overall model DIC is the sum of the DIC for each observation, and it is apparent that the difference between the DIC for the two models is largely due to the influence of the pedal cycle casualty rates, the rural–urban model having values of 681.449 for serious and 759.371 for slight pedal cycle rates whereas the fully exchangeable model had values of 529.394 and 733.681 respectively. It seems quite likely that cycling rates vary the greatest, and

Table 4. Means of estimates of the latent correlation ρ derived from the posterior of the variance–covariance matrix for the urban authorities where partial exchangeability is assumed

Variable	Estimates for the following variables:								
	1	2	3	4	5	6	7	8	9
1 (pedestrian fatal)	1.00	0.92	0.89	-0.52	0.52	0.57	-0.67	-0.02	0.33
2 (pedestrian serious)	0.92	1.00	0.93	-0.62	0.47	0.39	-0.66	-0.09	0.23
3 (pedestrian slight)	0.89	0.93	1.00	-0.67	0.23	0.41	-0.75	-0.23	0.28
4 (pedal cycle fatal)	-0.52	-0.62	-0.67	1.00	0.12	-0.06	0.84	0.72	0.28
5 (pedal cycle serious)	0.52	0.47	0.23	0.12	1.00	0.54	-0.07	0.62	0.34
6 (pedal cycle slight)	0.57	0.39	0.41	-0.06	0.54	1.00	-0.49	0.32	0.66
7 (motor-cycle fatal)	-0.67	-0.66	-0.75	0.84	-0.07	-0.49	1.00	0.53	-0.05
8 (motor-cycle serious)	-0.02	-0.09	-0.23	0.72	0.62	0.32	0.53	1.00	0.63
9 (motor-cycle slight)	0.33	0.23	0.28	0.28	0.34	0.66	-0.05	0.63	1.00

Table 5. Means of estimates of the latent correlation ρ derived from the posterior of the variance–covariance matrix for the rural authorities where partial exchangeability is assumed

Variable	Estimates for the following variables:								
	1	2	3	4	5	6	7	8	9
1 (pedestrian fatal)	1.00	0.59	0.69	0.02	0.05	0.13	-0.16	-0.08	0.20
2 (pedestrian serious)	0.59	1.00	0.62	-0.03	0.42	0.19	-0.47	-0.16	-0.24
3 (pedestrian slight)	0.69	0.63	1.00	-0.08	-0.05	0.29	-0.60	-0.52	0.08
4 (pedal cycle fatal)	0.02	-0.03	-0.08	1.00	0.73	0.80	0.25	0.57	0.68
5 (pedal cycle serious)	0.05	0.42	-0.05	0.73	1.00	0.69	-0.05	0.44	0.14
6 (pedal cycle slight)	0.13	0.19	0.29	0.80	0.69	1.00	-0.30	0.06	0.42
7 (motor-cycle fatal)	-0.16	-0.47	-0.60	0.25	-0.05	-0.30	1.00	0.82	0.50
8 (motor-cycle serious)	-0.08	-0.16	-0.52	0.57	0.44	0.06	0.82	1.00	0.51
9 (motor-cycle slight)	0.21	-0.24	0.08	0.68	0.14	0.42	0.50	0.52	1.00

estimates of pedal cycle casualty rates are the worst affected by the use of a PI that calculates rates relative to population rather than exposure to transport mode. Cycling is possibly more common in flat areas of the eastern part of England, and neither the rural–urban model nor the simple nine-variable model account adequately for this. It is also possible that the rural–urban classification should be considered as a continuum rather than a sharp boundary.

4.3. Comparison of estimates of performance rankings

A major interest in comparing models obviously concerns the predictions of the authority-specific random effects and their credible intervals, since these pertain to evidence of differential performance. At the same time, whether intended by those commissioning performance measurement or not, it remains the case that ranking of institutions by using published PIs will inevitably take place. Given that, and Audit Commission notions that the top quartile represents the level of performance to which all authorities should aspire, we compare model predictions

concerning rankings of random effects with particular reference to which authorities have 95% credible intervals for such ranks which are entirely within either the top or the bottom quartiles. In the absence of any 'gold standard' for casualty rates, evidence that an authority is very likely to be in either of these quartiles would perhaps justify further effort to ascertain the reason for such a ranking.

First, it is informative to compare rankings that are obtained from the six-variable model (which excluded the slight casualty rates) with those resulting from the full nine-variable model. Taking, as an illustration, the case of pedestrian serious casualty rates, Devon, Leicestershire, Norfolk, Blackpool and North Somerset have 95% credible intervals on ranks which place them among the top quartile of all highways authorities in England (i.e. the lowest pedestrian serious casualty rates) regardless of which model is considered. However, with the nine-variable model West Sussex and Thurrock have credible intervals shrunk so that they are also ranked entirely within the top quartile. At the other extreme, Southampton, West Yorkshire, Torbay, Cumbria, South Yorkshire and Suffolk all have 95% credible intervals for ranks which place them entirely in the 'bottom' quartile (i.e. the highest pedestrian serious casualty rates) regardless of the model used, whereas with the nine-variable model Merseyside and Essex also have credible intervals on their ranks that are entirely within the bottom quartile. Similar types of minor adjustments, albeit involving different individual authorities, are seen for pedal cycle and motor-cycle serious casualty rates and the overall picture that is indicated by these kinds of comparisons is that the additional shrinkage that is seen when the three slight casualty rate variables are included in the model is not accompanied by any radical change in the ranks that are predicted by the model which excludes slight casualty rates. Most of the authorities that have credible intervals that are shrunk into either of the extreme quartiles when adding these additional variables were already very close in the six-variable model and all except one (Dorset serious motor-cycle casualty rate) were entirely within the 50th percentile.

We can make similar comparisons between the rural-urban model and the nine-variable model. In addition to those authorities that are found in the extreme quartiles in the fully exchangeable model some additional shrinkage is seen. For the pedestrian serious casualty rates, North East Lincolnshire and the City of York have 95% credible intervals that are shrunk such that they are entirely in the top quartile in the rural-urban model and at the other extreme Blackpool and North Somerset have 95% credible intervals that are shrunk such that they are entirely in the bottom quartile. For the motor-cycle serious casualty rate, authorities with 95% credible intervals in the top quartile are unchanged when comparing the models, but Brighton and Hove ceases to be in the bottom quartile when considering the rural-urban model. Considering the uncertainty in rank for the pedal cycle casualty rate introduces the greatest inconsistencies. Norfolk, Devon, Derby, Cheshire, Hampshire, Cornwall, Southampton and Bath and North East Somerset have totally non-overlapping 95% credible intervals when comparing the two models. Norfolk had a 95% credible interval that was in the top quartile in the nine-variable model, but it is somewhere in the middle 50th percentile when the rural-urban model is considered. Bath and North East Somerset and North Lincolnshire have a 95% credible interval that is in the bottom quartile in the rural-urban model. The major differences between predictions from the rural-urban and fully exchangeable nine-variable models would therefore appear to relate largely to pedal cycle casualty rates. It was noted earlier that the largest discrepancies in posterior estimates of correlation between the two models also occurred for these variables and that neither the rural-urban model nor the simple nine-variable model accounts particularly well for variations in pedal cycle casualty rates.

It is obviously also of considerable interest to compare results that are achieved from multivariate models with those obtained from separate univariate GLMMs to establish whether

the multivariate approach has been successful in its objective of increasing the precision of comparisons of performance. In this respect a similar picture emerges from the use of any of the multivariate formulations and we illustrate that by a graphical comparison of results from the nine-variable fully exchangeable model with those obtained from univariate models on each of the variables. In Fig. 2 we superimpose the median, 2.5th percentile and 97.5th percentiles of the posterior for the rank from both the nine univariate models and the multivariate model for each of the 87 highways authorities. This 'caterpillar' plot differs somewhat from the more usual bar-chart presentation of PIs. The dots within the plot denote the mean of the estimated posterior distribution for each of the ranks and the 'error bars' denote the 95% credible intervals from this posterior distribution. As has been suggested by Goldstein and Myers (1996) we not only have an estimate of performance ranking, but importantly we also have an estimate of the likely uncertainty that is associated with that estimate. It is apparent that all credible intervals have been considerably narrowed by the use of the multivariate model. This is particularly so in the case of variables with the smallest counts, namely fatal casualties. However, some narrowing is also apparent in credible intervals for indicators concerning some slight casualty types where counts were lower. The posterior estimates of correlation have suggested that there are some notably related variables in this data set and the multivariate model has been able to suggest that the corresponding univariate models may exaggerate the uncertainty that is associated with the assessment of a PI when a single variable is examined in isolation.

Although these comparisons have indicated that the multivariate GLMM approach has achieved considerable narrowing of credible intervals on the estimates of the ranking of the random effects, it is nevertheless interesting to note how 'fuzzy' the resulting league table remains. Despite the substantial narrowing of credible intervals a large degree of overlap remains between many authorities. However, using the multivariate approach it is possible to identify a small number of authorities that are more likely to have substantively meaningful departures from average performance levels whereas this is not generally the case using univariate models. These discrepancies may then merit further qualitative research to establish reasons for the departure.

Finally, as mentioned earlier, the Parliamentary Advisory Council for Transport Safety (2003) have suggested that one simple method of reducing the problem of variability in some traffic safety PIs would be to aggregate crude PIs across say vulnerable road user types. It is clear from the presence of some negative values in the estimated correlation matrix in our multivariate models that such aggregation is inadvisable and will result in masking of information. By way of an illustration of that, we consider three authorities who would occupy a bottom quartile position by using aggregated crude PIs for fatally and seriously injured casualties. Despite the lower aggregate ranking, Blackpool is one of a few authorities to have 95% credible intervals for rank of the serious pedestrian casualty rate entirely within the top quartile. Cumbria has a 95% credible interval for rank of the serious pedal cycle casualty rate entirely within the top quartile and Poole has a 95% quartile for rank of the motor-cycle serious casualty rate entirely within the top quartile when modelled multivariately (and this is in line with the raw data for these isolated variables). Numerous other examples of such masking can be found by careful examination and highlight the potentially interesting information that could be lost by aggregating data across variables. It clearly seems preferable, where the models are sufficiently robust, to be able to present the information that authorities may have better performance outcomes in certain modal groups than it would be to use a crude aggregated PI which may mask the presence of high levels of relative performance on some of the outcomes that are aggregated.

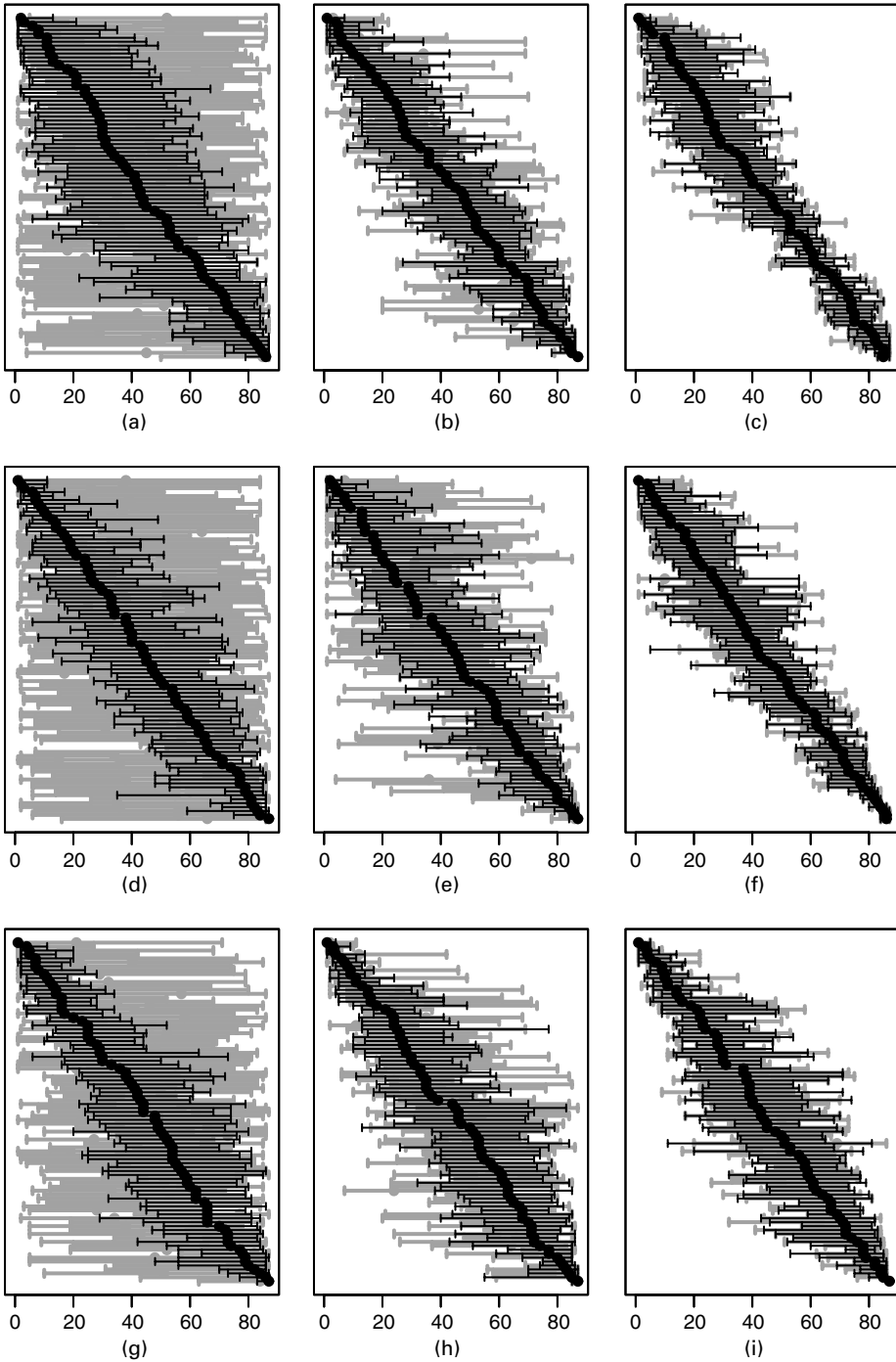


Fig. 2. Summaries of rankings of random effects from posterior distributions from nine univariate models (—) and one multivariate model (—) superimposed: (a) pedestrian fatalities; (b) pedestrian serious casualties; (c) pedestrian slight casualties; (d) pedal cyclist fatalities; (e) pedal cycle serious casualties; (f) pedal cyclist slight casualties; (g) motor-cycle fatalities; (h) motor-cycle serious casualties; (i) motor-cycle slight casualties

5. Discussion

In this paper we have proposed a multivariate GLMM that can be readily applied to groups of related traffic safety PIs to improve the performance inference that can be drawn from any one of them used in isolation. With the traffic safety data studied, use of a multivariate rather than a univariate approach substantially narrows the credible intervals that are associated with estimates of both random effects and their ranks and so allows a more precise identification of differential performance on any of the variables modelled. The approach has general application and may also be of use in the analysis of multiple PIs relating to many other areas of performance evaluation. Multivariate modelling of multiple PIs may therefore have considerable potential for introducing more robust and realistic assessments of differential performance in a variety of contexts.

Turning to particular aspects of the results, we would highlight differences between the nine-variable fully exchangeable model and the rural–urban model as evidence of a general future need to correct for differential exposure in the analysis of traffic safety PIs. We believe that the rural–urban divide that was used in our models is primarily a crude surrogate for such exposure differences and that these need to be allowed for more explicitly. A more general point that our results emphasize is the deficiencies of the crude league table in performance ‘management’. When, as here, the source data are available and modelling can be carried out, the ranking of the majority of the units of analysis remains very uncertain, even when a multivariate approach is used to extract maximum information from the data. The inference is that in most cases performance differences simply cannot be identified from these data, despite what the raw league table suggests. It should further be noted that this study has only been possible because the PI source data could be obtained elsewhere and neither it nor univariate modelling could have been carried out on the published per capita PIs. This highlights the need for a greater understanding of the role of statistical modelling and associated relevant data in performance management in local government.

On a more general front, multivariate GLMMs may be of particular use in cases where there has been a recognition of excessive year-by-year random fluctuation and a subsequent desire to deal with this by aggregating data across PI variables. By using multivariate models we suggest that it is possible to share strength across variables at the same time as retaining discriminatory information that is associated with particular indicators. In this paper we have been able to maintain a distinction between three discrete types of road user and it would remain possible to identify significant outliers with respect to one type of road user that may not be apparent if all types of road user were aggregated. Indeed evidence from the correlation estimates that we obtained suggests that different modal groups tend not to be highly correlated and so may present very different patterns of behaviour which would be masked by aggregating the PIs.

One extension to this model would be to develop it to handle longitudinal observations on a range of related indicators. The simplest way of introducing this would be to expand the multivariate normal distribution into a third, temporal, dimension but that would require great care in the specification and implementation owing to the large number of potential intervariable correlations. It may be possible to simplify the correlation structure by incorporating the temporal data in a simpler autoregressive manner. Another extension may be to include hospital episode statistics data on road collisions (if that could be made available in a format that enumerated road collisions by the highways authority in which they occurred). If the intervariable correlation assumptions in the STATS 19 data are reasonable, then that would allow a degree of validation of the STATS 19 data. Finally, if we have disparate groups of authorities, then we may also have disparate variance–covariance structures for PIs. For example if motor-cycle

collisions were more likely to be fatal or serious on minor rural roads even a rural–urban split is inadequate. It would therefore be useful to extend this modelling approach to look at more complex correlation structures.

While acknowledging that a focus on ‘outcome’-oriented PIs may simply amount to a focus on underlying phenomena that organizations can do little about, it remains the case that such performance monitoring is a rapidly expanding area in local government in the UK as well as elsewhere. Given that, we feel that a better appreciation of the potential for modelling to understand variation in PIs is needed in the public sector and simultaneous multivariate GLMMs offer a useful framework for investigating groups of related PIs which could contribute to increasing the robustness that is associated with performance assessments.

Acknowledgements

Data on road accidents were supplied by the Data Archive at Essex University. This work was completed while one of the authors was employed in the Environment Directorate at Devon County Council. We thank the Joint Editor and referees for full, helpful and useful comments on earlier versions of the paper.

References

- Aitchison, J. and Ho, C. H. (1989) The multivariate Poisson-log normal distribution. *Biometrika*, **76**, 643–653.
- Bhattacharya, S. (1967) A result on accident proneness. *Biometrika*, **54**, 324–325.
- Bull, J. and Roberts, B. (1973) Road accident statistics—a comparison of police and hospital information. *Accid. Anal. Prev.*, **5**, 45–53.
- Cabinet Office (2000) *Wiring It Up*. London: Cabinet Office.
- Chib, S. and Winkelmann, R. (2001) Markov Chain Monte Carlo analysis of correlated count data. *J. Bus. Econ. Statist.*, **19**, 428–435.
- Congdon, P. (2001) *Bayesian Statistical Modelling*. Chichester: Wiley.
- Cryer, P., Westrup, S., Cook, A., Ashwell, V., Bridger, P. and Clarke, C. (2001) Investigation of bias after data linkage of hospital admissions data to police road crash reports. *Inj. Prev.*, **7**, 234–241.
- Department of Health (1999) *Saving Lives: Our Healthier Nation*. London: Department of Health.
- Department for Transport (2001) *Road Accidents Great Britain*. London: Department for Transport.
- Department for Transport, Local Government and the Regions (1997) *The Road Traffic Reduction Act 1997*. London: Department for Transport, Local Government and the Regions.
- Department for Transport, Local Government and the Regions (1998) *A New Deal for Transport: Better for Everyone*. London: Department for Transport, Local Government and the Regions.
- Department for Transport, Local Government and the Regions (1999) *Local Government Act 1999*. London: Department for Transport, Local Government and the Regions.
- Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (2003) *Bayesian Data Analysis*. Boca Raton: CRC Press.
- Goldstein, H. and Myers, K. (1996) Freedom of Information: towards a code of ethics in performance indicators. *Res. Intell.*, **57**, 12–16.
- 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.
- Gueorguieva, R. (2001) A multivariate generalized linear mixed model for joint modelling of clustered outcomes in the exponential family. *Statist. Modelling*, **1**, 177–193.
- Hinde, J. (1982) Compound Poisson regression models. In *GLIM 82* (ed. R. Gilchrist), pp. 109–121. New York: Springer.
- James, H. (1991) Underreporting of road traffic accidents. *Traffic Engng Control*, **32**, 574–583.
- Joe, H. and Xu, J. J. (1996) The estimation method of inference functions for margins for multivariate models. *Technical Report*. Department of Statistics, University of British Columbia, Vancouver.
- Keigan, M., Broughton, J. and Tunbridge, R. J. (1999) Linkage of STATs 19 and Scottish Hospital In-patient data—analysis for 1980–1995. *Report TRL420*. Transport Research Laboratory, Crowthorne.
- Lawless, G. F. (1987) Negative binomial and mixed Poisson regression. *Can. J. Statist.*, **15**, 209–225.
- Lee, Y. and Nelder, J. A. (1996) Hierarchical generalized linear models (with discussion). *J. R. Statist. Soc. B*, **58**, 619–678.
- MacNab, Y. C. (2003) A Bayesian hierarchical model for accident and injury surveillance. *Accid. Anal. Prev.*, **35**, 91–102.

- Marshall, E. C. and Spiegelhalter, D. J. (1998) Reliability of league tables of in vitro fertilisation clinics; retrospective analysis of live birth rates. *Br. Med. J.*, **316**, 1701–1705.
- McCullagh, P. and Nelder, J. A. (1989) *Generalized Linear Models*, 2nd edn. London: Chapman and Hall.
- Morris, C. and Christiansen, C. (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.
- Office of the Deputy Prime Minister (2003) *Best Value Performance Indicators*. London: Office of the Deputy Prime Minister. (Available from www.odpm.gov.uk.)
- Papageorgiou, H. and Loukas, S. (1988) On estimating the parameters of a bivariate model applicable to traffic accidents. *Biometrics*, **44**, 495–504.
- Parliamentary Advisory Council for Transport Safety (2003) *Best Value, Local Transport Plans and Road Safety*. London: Parliamentary Advisory Council for Transport Safety.
- Smith, P. (1990) The use of performance indicators in the public sector. *J. R. Statist. Soc. A*, **153**, 53–72.
- Spiegelhalter, D. J., Aylin, P., Best, N. G., Evans, S. J. and Murray, G. D. (2002) Commissioned analysis of surgical performance using routine data: lessons from the Bristol inquiry (with discussion). *J. R. Statist. Soc. A*, **165**, 191–231.
- Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and van der Linde, A. (2002) Bayesian measures of complexity and fit (with discussion). *J. R. Statist. Soc. B*, **64**, 583–639.
- Spiegelhalter, D., Thomas, A. and Best, N. (1998) WinBUGS version 1.1.1 user manual. *Technical Report*. Medical Research Council Biostatistics Unit, Cambridge.
- Stone, M. (2002) How not to measure the efficiency of public services (and how one might). *J. R. Statist. Soc. A*, **165**, 405–434.
- Tunaru, R. (1999) Hierarchical Bayesian models for road accident data. *Traff. Engng Control*, **40**, 318–324.
- Vistisen, D. (2002) A consistent method for estimating the effect of hot spot safety work. *Traff. Engng Control*, **43**, 96–100.
- Vogt, A. and Bared, J. G. (1998) Accident models for two-lane rural roads: segments and intersections. *Report FHWA-RD-98-133*. National Transportation Library, Washington DC.
- Wilding, P. (2002) The 2002 quality review of road accident statistics. In *Road Accidents Great Britain 2001: the Casualty Report*, pp. 32–37. London: Department for Transport.
- Woodward, A. (1983) Time trends in motorcycle accidents in Britain. *J. Epidem. Commty Hlth*, **37**, 66–69.