Prediction of settlement delay in critical illness insurance claims using GB2 distribution

Erengul Dodd *[1] and George Streftaris †[2]

[1]Southampton Statistical Sciences Research Institute and Mathematical Sciences, University of Southampton, Southampton, SO17 1BJ, UK
[2]Maxwell Institute and Department of Actuarial Mathematics and Statistics, Heriot-Watt University, Edinburgh, EH14 4AS, UK

Abstract

In this paper we analyse the delay between illness diagnosis and claim settlement in critical illness insurance using generalised linear-type models under a generalised beta of the second kind family of distributions. A Bayesian approach is employed which allows us to incorporate parameter and model uncertainty and also to impute missing data in a natural manner. We propose methodology involving a latent likelihood ratio test to compare missing data models and a version of posterior predictive p-values to assess different models. Bayesian variable selection is also performed, supporting a small number of models with small Bayes factors, and therefore we base our predictions on model-averaging instead of on a best-fitting model.

Keywords: critical illness insurance, settlement delay, GB2 family of distributions, generalised linear-type models, Bayesian model assessment

1 Introduction

Critical illness insurance (CII) is a type of long-term insurance that pays a lump sum on the diagnosis of a specified list of critical illnesses as specified in the policy. The delay between date of diagnosis of a critical illness and date of settlement of the corresponding claim is the main driver of outstanding claims reserves in the insurance sector, for example the incurred-but-not-settled reserve in CII. This is closely related to insurance capital retention requirements, under EU regulation (Solvency II directive). CII is often subject to long delays between the dates of illness diagnosis and settlement of the resulting claim. In many cases this delay can be measured in months, while often it can be measured in years – as demonstrated in the distribution of observed delay in Figure 1 for the data used in this paper. The frequency of long delays leads us to consider heavy-tailed distributions for modelling their duration. Modelling of this delay has been previously

*Corresponding author. Address: Southampton Statistical Sciences Research Institute, University of Southampton, Highfield, Southampton, SO17 1BJ, UK. Tel: +44 (0)23 80593679. Email: E.Dodd@soton.ac.uk.

†Email: G.Streftaris@hw.ac.uk.

considered (Ozkok et al., 2012), where a 3-parameter Burr distribution was used to model the CII settlement delay.

In this paper we build on earlier work to model the delay between the dates of diagnosis and settlement by fitting a generalised linear-type (GL-type) model to the observed settlement delay under a Bayesian framework using a generalised beta of the second kind (GB2) error distribution. Our work also considers extensive model assessment and comparison involving distributions of the same family. In particular, for model assessment, we propose the use of methodology which is related to posterior predictive checking, and is based on the properties of latent quantities which become available through data augmentation within Markov chain Monte Carlo estimation schemes (e.g. Lau et al. (2014)). Comparison between candidate models is also considered using a latent likelihood ratio-type test (e.g. Streftaris and Gibson (2012)) that is designed to avoid common problems associated with Bayesian model choice. The resulting model offers additional flexibility and better estimation, which are central when aiming at accurate prediction especially in the presence of missing information, as explained later.

The 4-parameter GB2 distribution, sometimes called transformed beta or Pearson Type VI distribution, is a very flexible distribution and encompasses many distributions which are particularly helpful for fitting heavy-tailed data, such as the Burr, Pareto and log-logistic distributions. Since insurance data are generally strongly right-skewed (e.g. claim amounts, number of claims) the GB2 distribution is appropriate for related modelling and is a well-known loss distribution (Klugman et al., 2012). The use of the GB2 distribution has become increasingly popular in relevant literature. To name few, Venter (1983) introduced the GB2 distribution in the actuarial literature as the transformed beta. McDonald and Butler (1990) used covariates under the GB2 model, while Cummins et al. (1990) used it for modelling insurance losses. A hierarchical model for insurance claims was developed by Frees and Valdez (2008), where the long tail of the claims distribution is captured through the GB2 distribution. Recently, Dong and Chan (2013) considered the GB2 distribution in insurance reserving, and Jones et al. (2014) modelled healthcare cost by applying the same distribution to UK data.

The data used in this paper contain individual records related to 19,127 claims settled in the UK during the years 1999-2005. In approximately 17% of these cases the length of the delay period is missing, mainly due to the date of either illness diagnosis or claim settlement not having been recorded. Where only one of these dates is known, the missing date can be estimated using the appropriate distribution of the delay between the two events. Since missing values can be seen as another source of uncertainty, they can naturally be incorporated under a Bayesian analysis. Modelling the distribution of this delay, which we will refer to as the claim delay distribution (CDD), incorporating the cause-specific covariates, is a necessary first step in the estimation of claim rates. Throughout this paper we focus on a Bayesian approach for inference under parameter and model uncertainty. In actuarial research, especially in life-related insurance area, focusing on parameter and model uncertainty is a relatively recent development. There are only few studies in mortality modelling accounting for uncertainty, for example Cairns et al. (2006) or more recently Bennett et al. (2015).

The rest of the paper is organised as follows. The available data are summarised in Section 2. In Section 3 we introduce the GB2 model, construct Bayesian GL-type models by linking the claim related factors possibly affecting the delay to the mean of the data distribution and also estimate relevant parameters. In Section 4 we compare and assess different models using traditional and newly proposed methods. We discuss Bayesian variable selection, particularly Gibbs variable selection, under the GB2 error distribution in Section 5. In Section 6, we explain why we prefer to impute the missing data using Bayesian model-averaging, while final comments and conclusions

Table 1: Information about the covariates.

Covariates	Description
Age	Age last birthday at diagnosis
Sex	**Female** (42.7%) & Male (57.3%)
Benefit type	**Full accelerated (FA)** if the policy covers critical illness and death (88.2%) & Stand alone (SA) if the policy covers only the critical illness (11.8%)
Smoker Status	**Non-smoker** (73.9%) & Smoker (26.1%)
Policy Type	**Joint life** (50.9%) & Single life (49.1%)
Settlement year	1999-2005
Benefit amount	in GBP
Policy duration	Policy duration at diagnosis
Office	13 anonymously coded insurance offices
Cause	See text

are given in Section 7.

2 Data

We use CII claims data settled in the UK between 1999 and 2005, which have been supplied by the Continuous Mortality Investigation (CMI) of the UK. The CMI asks contributing insurance offices to provide four dates related to each claim, these being the dates of: diagnosis of related illness; notification of the insurance company; admission of the claim by the company; and finally settlement. Illness diagnosis signifies the insured event for CII and therefore it is important for financial purposes that the date of this event is available. Out of 19,127 claims in the data, 15,860 cases have both dates of diagnosis and settlement recorded, while the remaining have either no date of settlement (7.9%) or no date of diagnosis (9.1%) available. The observed delay between the dates of diagnosis and settlement has a mean of 185 days with a standard deviation of 263 days (see Figure 1).

The data also contain information on 10 claim-related characteristics, which are used as covariates in our analysis. These are described in Table 1. Covariate 'office' represents the contributing company where the insurance policy was issued. The last covariate in the table gives the cause of claim, related to the following illnesses: coronary artery bypass graft (CABG), cancer, heart attack, kidney failure, major organ transplant (MOT), multiple sclerosis (MS), stroke and total and permanent disability (TPD). These eight causes and death account for 93.4% of the data, while the remaining claims are categorised as 'other' cause. Further details on the data can be found in Ozkok et al. (2012, 2014).

3 The model

We denote the delay period for claim i as D_i and assume it follows a GB2 distribution, i.e.

$$D_i \sim GB2(\alpha, \tau, \gamma, s), \quad i = 1, \ldots, n \tag{1}$$

with probability density function

$$f_D(d_i) = \frac{\Gamma(\alpha+\gamma)}{\Gamma(\alpha)\Gamma(\gamma)} \frac{\tau \left(\frac{d_i}{s}\right)^{\tau\gamma}}{d_i \left[1+\left(\frac{d_i}{s}\right)^{\tau}\right]^{\alpha+\gamma}}, \tag{2}$$

for $d_i > 0$ and parameters $\alpha, \tau, \gamma, s > 0$. Here $\Gamma(.)$ is the gamma function, α, τ and γ are shape parameters and s is the scale parameter. The kth moment of the distribution is given by

$$\mathrm{E}(D^k) = \frac{s^k \Gamma\left(\gamma + \frac{k}{\tau}\right) \Gamma\left(\alpha - \frac{k}{\tau}\right)}{\Gamma(\alpha)\Gamma(\gamma)}, \qquad \alpha\tau > k. \tag{3}$$

Where dates of diagnosis (DoD) or settlement (DoS) are missing, we use other dates related to each claim, i.e. dates of commencement (DoC) of the policy, notification (DoN) and claim admission (DoA) as natural upper and lower limits so we restrict the missing delay values using

$$D_{mis} \sim GB2(\alpha, \tau, \gamma, s_{mis}) I(L_{mis}, U_{mis}). \tag{4}$$

Here the lower and upper bounds are denoted by L_{mis} and U_{mis}, respectively. $I(L_{mis}, U_{mis})$ is the restriction we impose on the missing delay and the boundaries are obtained using the chronological order of the dates. So we have
DoS $-$ DoN $\leq d_{mis} \leq$ DoS $-$ DoC, when DoD is not recorded;
DoA $-$ DoD $\leq d_{mis}$, when DoS is not recorded;
DoN $-$ DoD $\leq d_{mis}$, when both DoS and DoA are not recorded;
$0 < d_{mis}$, when only DoD or DoS is recorded.

Note that the Burr and Pareto distributions are special cases when $\gamma = 1$ and $\tau = \gamma = 1$, respectively. Also the generalised gamma (GG) distribution is a limiting case of the GB2 distribution. The GG distribution has two shape parameters, α and τ, and a scale parameter, s. The gamma, Weibull and exponential distributions are special cases of the GG distribution when $\tau = 1$, $\alpha = 1$ and $\alpha = \tau = 1$, respectively. Also, the log-normal distribution is a limiting distribution. Detailed information on generalised beta distributions and other nested distributions can be found in McDonald (1984) and McDonald and Xu (1995). In this paper we focus our attention on the GB2 model and its performance relative to the previously considered Burr distribution. In addition, comparisons are extended across related models mentioned above and some summary results are presented throughout.

For almost all offices contributing to our data, and for almost all years, the number of CI policies in force increased year on year. Due to the nature of the claims data, for claims settled in any of these seven years, those with relatively short delays relate to claims from policies in force in more recent years; those with relatively longer delays relate to policies in force in earlier years (Ozkok et al., 2014). The growth in the numbers of policies in force means that claims with shorter delays are relatively over-represented in our data. If this is not taken into account, it can introduce bias into the modelling of the CDD. Therefore we should allow for business growth. We do this by introducing weights, representing a growth factor, in the CDD modelling. The growth factor is set at 1 for the most recent year for which the office contributed data. For each earlier year of diagnosis for that office, the growth factor is calculated by multiplying the growth rates for that year of diagnosis and all subsequent years of diagnosis up to the final year for which the office contributed data. Here the growth rate is calculated as the ratio of the average number of policies in force in the following year for that office to the average number of policies in force in the year in question for the same office. In this context, average number in force is the average of the numbers of policies in force at the start and at the end of the year (for more details see Ozkok et al. (2014, Sec. 4.2)).

Here we assume that these office specific weights according to the year of diagnosis are inversely proportional to the variance of the underlying distribution. An example of this can be found in weighted least squares estimation (see for example, Draper and Smith (2014)). The variance of the GB2 distribution is given as

$$\mathrm{Var}(D) = s^2 \frac{\Gamma(\alpha)\Gamma(\gamma)\Gamma\left(\gamma + \frac{2}{\tau}\right)\Gamma\left(\alpha - \frac{2}{\tau}\right) - \left[\Gamma\left(\gamma + \frac{1}{\tau}\right)\right]^2 \left[\Gamma\left(\alpha - \frac{1}{\tau}\right)\right]^2}{\left[\Gamma(\alpha)\right]^2 \left[\Gamma(\gamma)\right]^2}. \tag{5}$$

To introduce weights, w, in the model we define a new scale factor, $s^{'}$, such that

$$s^{'} = s/\sqrt{w}.$$

In Figure 1 we illustrate the fit of the GB2 and related distributions to the observed delay on a log scale before including the covariates under a generalised linear-type setting. It can be seen that the GB2 and Burr have a similar and very close fit to the data at both tails. The GG and log-normal distributions are also close to each other but their fit is not as good as that of the GB2 and Burr distributions. The fit of the Pareto distribution is the least successful among the five distributions, especially for shorter delays.

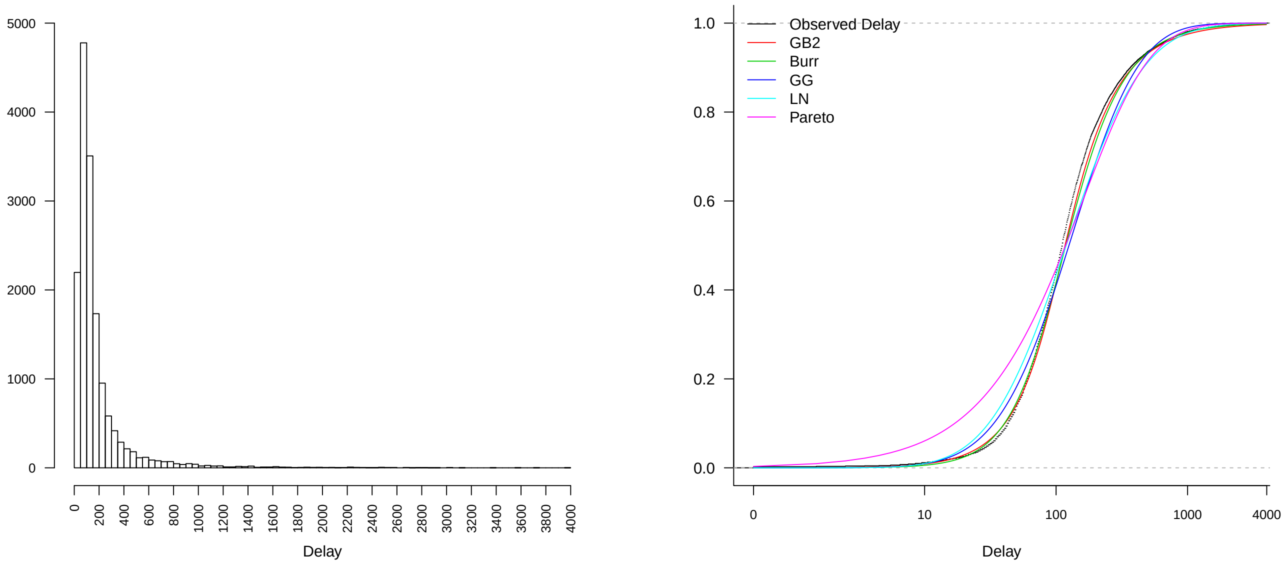


Figure 1: Left panel: histogram of the observed delay (in days). Right panel: CDF of observed delay (in days, on log scale) and fitted distributions.

3.1 GB2 GL-type model

We allow the delay for claim i to depend on the $p \times 1$ vector of standardised covariates, $\boldsymbol{z}_i$, presented in Table 1. We also show the baseline values of the binary covariates in bold font in this table. For the other categorical variables, i.e. office and cause, we use a sum-to-zero restriction. Since the GB2 distribution does not come from an exponential family, we construct a GL-type model (see for example Ozkok et al. (2012)). Although the covariates can be linked to any of the model parameters, we link them to the mean of the GB2 distribution through

$$\log\left(E(D_i)\right) = \eta_i = \beta_0 + \sum_{j=1}^{8} \beta_j z_{ij} + \beta_{9\,O_i} + \beta_{10\,C_i}, \tag{6}$$

where $\boldsymbol{\beta}$ is a 32×1 vector of regression parameters and in particular $\beta_{9\,O_i}$ and $\beta_{10\,C_i}$ denote the coefficients of the office and cause variables, respectively.

For equation (6) to hold, equation (3) implies that

$$s_i^* = \left[\exp(\eta_i)\frac{\Gamma(\alpha)\Gamma(\gamma)}{\Gamma\left(\gamma+\frac{1}{\tau}\right)\Gamma\left(\alpha-\frac{1}{\tau}\right)}\right] \Big/ \sqrt{w_i}. \tag{7}$$

Here w_i denotes the office specific weights according to the year of diagnosis for claim i. For the nested Burr distribution, equation (7) can be adjusted accordingly by substituting $\gamma = 1$. Similar results can be derived for other related distributions (see Section S1 in supplementary material).

We use a Metropolis-Hastings algorithm, more specifically a random walk Metropolis algorithm, to draw samples from the following joint posterior density

$$f(\alpha, \tau, \gamma, \boldsymbol{\beta}|\boldsymbol{D}) \propto f(\boldsymbol{D}|\alpha, \tau, \gamma, \boldsymbol{\beta}) f(\alpha) f(\tau) f(\gamma) f(\boldsymbol{\beta}), \tag{8}$$

where $f(\alpha), f(\tau), f(\gamma)$ and $f(\boldsymbol{\beta})$ denote the prior density functions for α, τ, γ and $\boldsymbol{\beta}$ parameters, respectively. We assign the following non-informative priors to these model parameters:

$$\begin{aligned}
&\alpha \sim \text{Gamma}(1, 0.01) I(1/\tau, \infty)\\
&\tau \sim \text{Gamma}(1, 0.01)\\
&\gamma \sim \text{Gamma}(1, 0.01)\\
&\beta_j \sim \text{N}(0, 10^4), j = 1, \ldots, 8\\
&\beta_{9\,O_i} \sim \text{N}(0, 10^4), O_i = 2, \ldots, 13\\
&\beta_{10\,C_i} \sim \text{N}(0, 10^4), C_i = 2, \ldots, 10.
\end{aligned} \tag{9}$$

Here $I(\text{lower}, \text{upper})$ denotes the restriction we must impose ($\alpha\tau > 1$) in order for the mean of the distribution to be defined. We run this algorithm for a total of 304,000 iterations and obtain inferences after a burn-in phase of 4,000. For efficiency of the algorithm, the parameters are tuned to obtain an acceptance rate between 15% and 40%. By inspecting relevant trace plots (not shown here) we confirmed that all algorithms converged satisfactorily.

For comparison purposes (see also Section 4) we also fit four other related models: Burr, log-normal, generalised gamma and Pareto. The GL-type setting is the same to that presented earlier for the GB2 model, with equation (7) adjusted accordingly for each distribution. Also, similar non-informative priors are assumed for model parameters (see Section S1 in supplementary material).

The posterior densities and histograms of GB2 model parameters (α, τ and γ) can be seen in Figure 2 and support the use of the GB2 distribution, as discussed later in Section 4. Posterior estimates under the other fitted models are shown in supplementary material (Table S1 and Figure S1).

The posterior estimates and 95% credible intervals of the covariate coefficients ($\boldsymbol{\beta}$) under the fitted distributions are shown in Figure 3. The plots illustrate the effect of these claim related characteristics on the duration of the delay period, and the results are generally in agreement with previous work as discussed in (Ozkok et al., 2012). Figure 3 suggests that there are some differences for the coefficients of office and settlement year covariates. This is expected as here we allow for the business growth within each office between successive years.

In terms of variation in parameter estimates, we note that the credible intervals do not show any obvious advantage for either of the two models of main interest (GB2 and Burr). At the same time, the GB2 model is more efficient than the Burr model in dealing with missing values. For this, we compare the relative posterior variance of the β coefficients under the two models when missing values are included and excluded from the analysis (see details in Section S4 in supplementary material).

The plots also show that some covariates are more affected by the tail structure of the distribution. For example, the coefficient estimates of settlement year and policy duration are quite different under different distributions. One of the reasons for this could be that the values of these two covariates are determined using dates of diagnosis (policy duration at diagnosis) and settlement (year of settlement) and hence are most affected from the shape of the underlying delay distribution. Estimates of TPD also appear to be considerably different under different models. As discussed in Ozkok et al. (2012), this may be explained by the large variations in the definition of TPD used by different offices and for different claims, leading to inconsistencies in the recorded date of diagnosis. Similarly, differences in office coefficient estimates (e.g. Offices 4, 12) may be attributed to varying practices in determining the period from claim diagnosis to settlement.

4 Model assessment and comparison

In Section 3 we introduced a GL-type model under the GB2 distribution and briefly discussed several related distributions with varying degrees of complexity. Assessment and comparison of these models is important, especially when aiming at prediction. These tasks are not very straightforward in the presence of missing data, especially with Bayesian models. In this section we present model comparison based on the deviance information criterion and also propose alternative approaches for both model comparison and assessment.

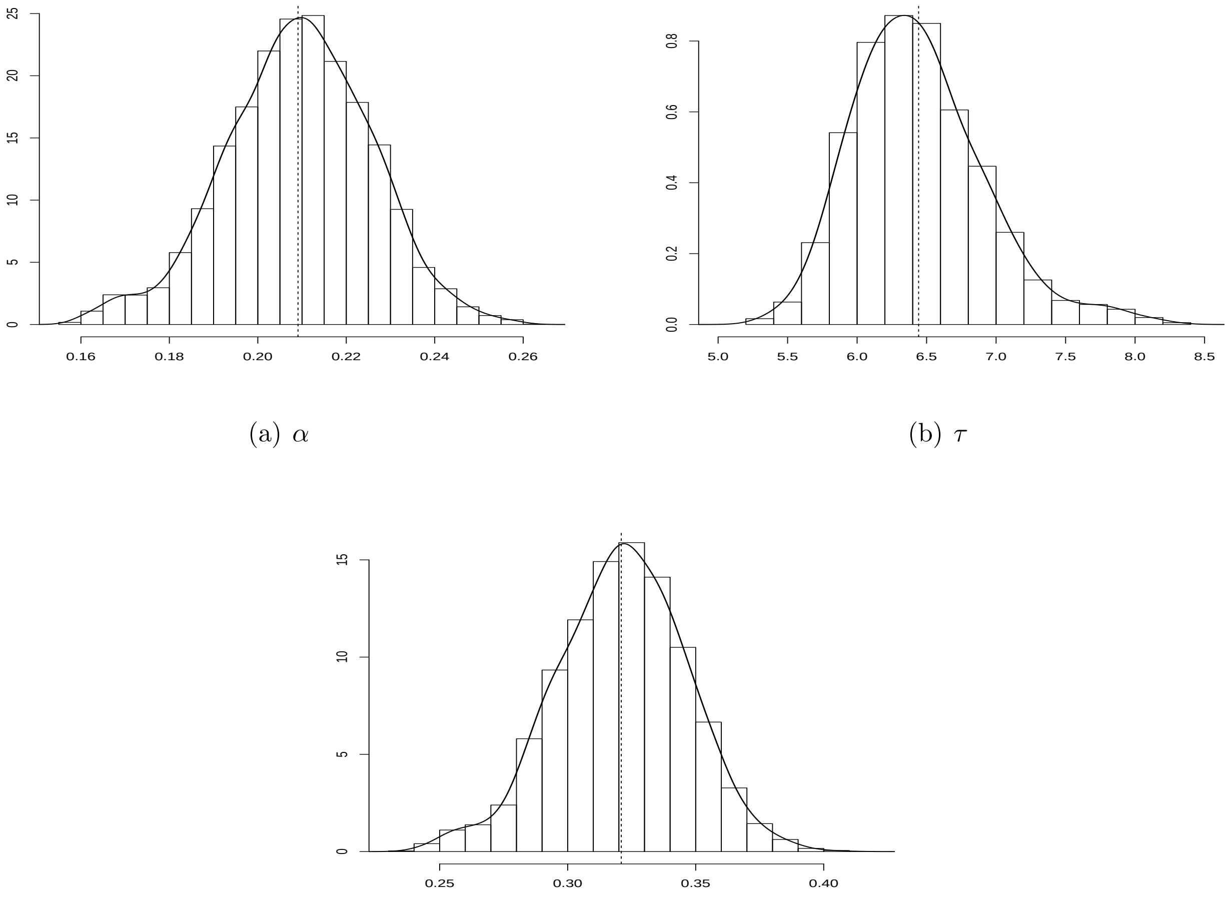
The full Bayesian analysis and inference obtained in Section 3.1 provides solid evidence regarding the suitability of the fitted GB2 model when compared to the nested Burr model (resulting when $\gamma = 1$) which was favoured in earlier work. This evidence can be drawn from the posterior densities of model parameters shown in Figure 2 where γ is significantly different from one, implying that the data suggest moving away from the Burr model. Similar results hold for other related distributions as discussed in the supplementary material (Section S1).

4.1 Deviance information criteria

For formal assessment and comparison of generalised linear models, Spiegelhalter et al. (2002) introduced the deviance information criterion (DIC). Similarly to other information criteria, the DIC is based on a trade-off between the goodness of fit and the complexity of the model. Let $\boldsymbol{\theta}$ denote the parameter vector. Then the DIC can be defined as

$$\mathrm{DIC} = -4E_{\boldsymbol{\theta}|\boldsymbol{D}}\left(\log f(\boldsymbol{D}|\boldsymbol{\theta})\right) + 2\log f(\boldsymbol{D}|\hat{\boldsymbol{\theta}}) \tag{10}$$

where $\hat{\boldsymbol{\theta}}$ is the posterior mean, or alternatively the posterior mode or median of $\boldsymbol{\theta}$. The use of this information criterion can be problematic (DeIorio and Robert, 2002). However, for relatively



(a) α

(b) τ

(c) γ

Figure 2: Posterior densities and histograms of model parameters together with their posterior means (dashed line) under the GB2 distribution.

simple models where there are no missing data, the DIC provides a measure for model comparison.

We first consider the DIC, calculated by employing (10) at the posterior means of the parameters under the Bayesian model without considering the missing values. The results are shown in Table 2. The DIC values support the evidence provided by the posterior estimates in Section 3.1, and further suggest that the GB2 gives the best fit to the data among the five considered distributions.

However, for missing data, model comparison based on DIC is more complicated. Due to different interpretations of the missing data (e.g. whether they should be treated as parameters or missing variables) calculation of the DIC is not unique. Celeux et al. (2006) present eight different DIC versions depending on the meaning of the missing data and the type of the model. Here we consider three of these versions, based on the partially observed nature of the likelihood and allowing for different interpretations of the missing information. Keeping the original indices in their paper, we calculate DIC_4, DIC_5 and DIC_8 (for details see Section S2 in supplementary material).

The results are given in Table 2 and once more they indicate that GB2 clearly outperforms the Burr model by providing the best fit to the data. However, with regards to other considered

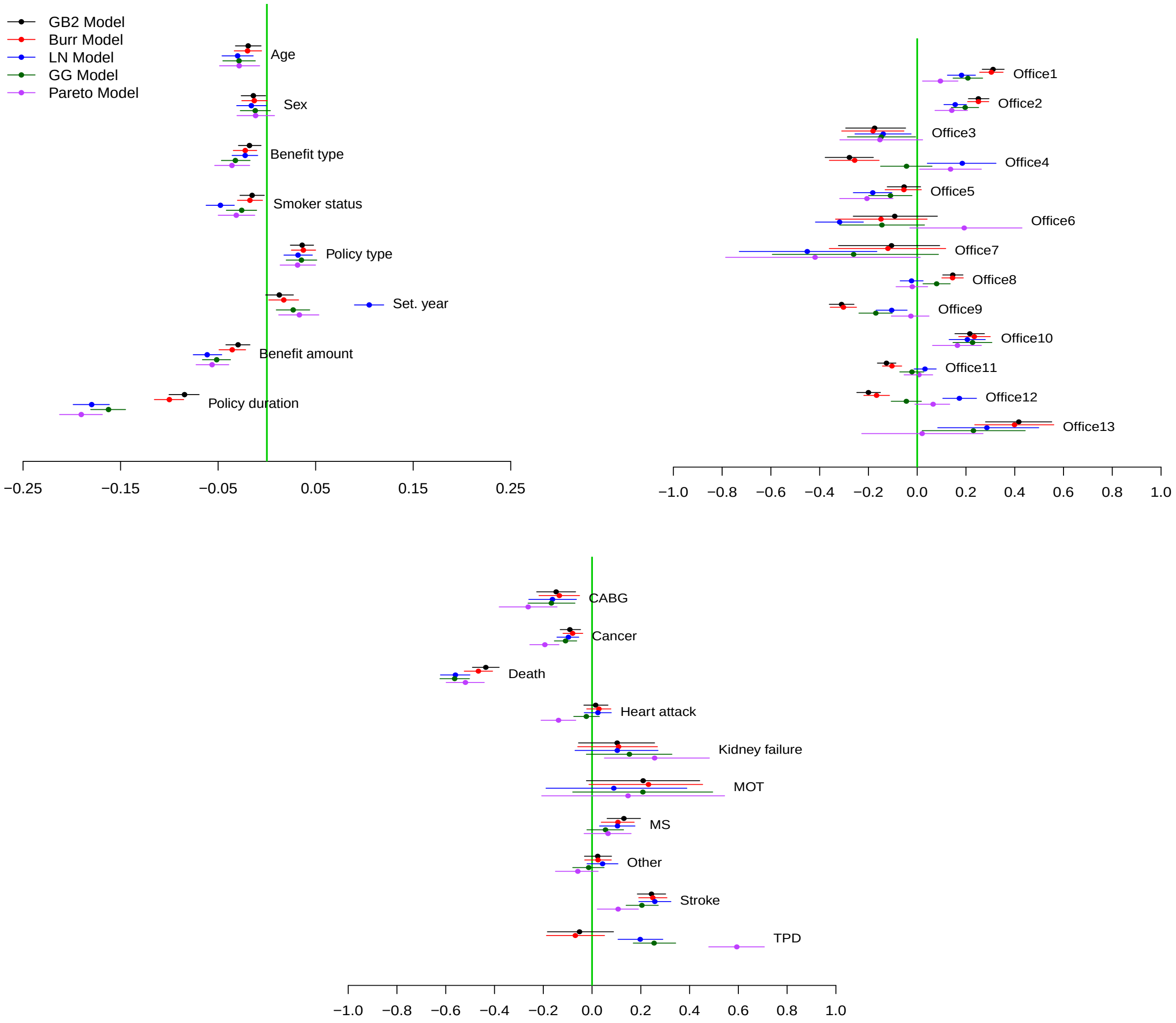


Figure 3: Comparison of the posterior means (dots) and 95% credible intervals (bars) of the $\boldsymbol{\beta}$ coefficients under various distributions.

models, we observe a slight discrepancy in their ranking when missing data are considered; while DIC_8 maintains the ranking suggested by the values of DIC without considering the missing data, DIC_4 and DIC_5 swap the order between the log-normal and Pareto models.

Table 2: Values of DIC of the models.

Without considering missing data

	GB2	Burr	GG	Log-normal	Pareto
DIC	190,983	191,260	193,018	194,356	195,986

Considering missing data

	GB2	Burr	GG	Log-normal	Pareto
DIC_4	230,952	231,251	233,262	237,798	237,665
DIC_5	231,677	232,002	233,835	238,765	238,297
DIC_8	191,037	191,315	193,065	194,796	196,060

4.2 Latent likelihood ratio test

We propose an alternative approach for comparing models under different distributions using a latent likelihood ratio (LLR) test, which can be convenient in cases where Bayesian model selection suffers from problems typically associated with sensitivity to choice of priors within models (e.g. Lindley–Bartlett paradox), chain mixing in reversible-jump MCMC and missing data. The methodology has been originally developed in a different context (e.g. Streftaris and Gibson (2012)), and is particularly relevant to problems with incomplete data. It relies on computing the likelihood ratio of two competing models, where only the hypothesised model, $\mathcal{M}_1$, is fitted under our Bayesian inference process, while for the alternative model, $\mathcal{M}_2$, a maximum likelihood value is calculated. A sample of the posterior distribution of this likelihood ratio can be obtained by employing a posterior sample of the model parameters $\boldsymbol{\theta}$ (and other latent quantities, e.g. missing data) to evaluate the likelihood of $\mathcal{M}_1$, and the maximum likelihood of $\boldsymbol{\theta}$ to evaluate the likelihood of $\mathcal{M}_2$. Then, for each value in the posterior sample of the LLR a relevant tail probability can be computed as explained below and used to provide evidence against the fit of $\mathcal{M}_1$ relative to $\mathcal{M}_2$. Implementation of the method can be achieved utilising the MCMC estimation process. More specifically, when $\mathcal{M}_1$ is Burr and $\mathcal{M}_2$ is GB2, at each MCMC post-convergence iteration $t = 1, \ldots, N$, we compute the latent value of the likelihood ratio Λ as

$$\Lambda^{(t)} = \frac{L_1\left(\alpha^{(t)}, \tau^{(t)}, \boldsymbol{\beta}^{(t)}; \boldsymbol{D}\right)}{L_2\left(\dot{\alpha}, \dot{\tau}, \dot{\gamma}, \dot{\boldsymbol{\beta}}; \boldsymbol{D}\right)}, \tag{11}$$

where $\boldsymbol{\alpha}^{(t)}, \tau^{(t)}, \boldsymbol{\beta}^{(t)}$ are MCMC posterior estimates at iteration (t) and the dotted values in the denominator are the MLEs. To calculate the tail probability $\pi_\Lambda = P\left(\Lambda \leq \Lambda^{(t)}\right)$ we need the sampling distribution of Λ under $(\mathcal{M}_1)$. If models are nested, we can use asymptotic arguments (as demonstrated in Streftaris and Gibson (2012)) implying $-2\log\Lambda^{(t)} \sim \chi^2_{df}$ approximately, and therefore obtain a tail probability as $\pi_\Lambda^{(t)} = P(\chi^2_{df} \geq -2\log\Lambda^{(t)})$, where df is the degrees of freedom of $\mathcal{M}_2$, i.e. the number of estimated parameters in $\mathcal{M}_2$. For non-nested models we can employ simulation to obtain the empirical sampling distribution of Λ (see Section S3 in supplementary material).

To compare the missing data models, we suggest to impute the missing values under $\mathcal{M}_1$ through data augmentation in MCMC, use the same imputed values in the denominator in (11) when finding the MLEs and follow the same procedure thereafter. When we compare the Burr and the GB2 model using this approach the range of $-2\log\Lambda^{(t)}$ is (239.2, 301.7) when we include the missing data in the analysis. These refer to the χ^2_{33} distribution, and therefore lead all p-values $\pi_\Lambda^{(t)}$ being practically zero, giving overwhelming evidence in favour of GB2. The range of the $-2\log\Lambda$ statistic that we find here is also consistent with the differences between the DIC values presented in Table 2, and therefore confirms our earlier conclusions. Based on these results we do not compare other distributions to the GB2 using the LLR approach.

4.3 Posterior distributions of p-values

We now focus on checking the adequacy of the models, using latent posterior probabilities of the delay variable, whose distribution is known under the assumption that the fit of the assessed model is good. This is related to posterior predictive checking, but leads to model assessment through posterior distributions of p-values (e.g. Streftaris and Gibson (2012); Lau et al. (2014)), similarly to the work in the previous section. As before, we consider the MCMC estimation process and let $f(D_j|\boldsymbol{\theta}^{(t)})$ be a sampling distribution (such as GB2) for delay j at (post-convergence) iteration t, where $j = 1, \dots, k$ and $t = 1, \dots, N$. We also assume that $D_1|\boldsymbol{\theta}, \dots, D_k|\boldsymbol{\theta}$ are conditionally independent and that we are able to compute the cumulative distribution function value $q_j^{(t)} = P(D_j \leq D_j^{\text{obs}}|\boldsymbol{\theta}, \boldsymbol{D})$ given the sampling distribution $f(D_j|\boldsymbol{\theta}^{(t)})$. Then, under the hypothesis that the model fits the data adequately, we should have that $\boldsymbol{q}^t = q_1^{(t)}, \dots, q_k^{(t)} \sim U(0,1)$, and therefore we can obtain a p-value $\pi_Q^{(t)}$ for compliance with $U(0,1)$ at each MCMC iteration (e.g. perform a Kolmogorov-Smirnov goodness-of-fit test). Hence, $\boldsymbol{\pi}_Q = \pi_Q^{(1)}, \dots, \pi_Q^{(N)}$ gives a sample from the posterior distribution of these p-values, and can be used to assess evidence against the fitted model.

4.3.1 Simulation results

To demonstrate the use of the approach described above, we first apply the method to 500 simulated delay values

$$D_i^{(sim)} \sim GB2(\alpha, \tau, \gamma, s_i^*)$$

where s_i^* is given in (7) and the covariates $\boldsymbol{\beta}$ and the remaining model parameters (α, τ and γ) are taken from fitting the real data. For this we randomly select 500 policies to provide the covariate values z (see Section 3.1 for the model specification). Then we fit the GB2 and Burr models to the simulated data $\boldsymbol{D}^{(sim)}$ and perform model assessment by obtaining posterior distributions of p-values as described above (Figure 4) and DIC values (Table 3). For comparison purposes we also present results for the GG, log-normal and Pareto models. Noting that a posterior distribution of p-values concentrated close to zero provides evidence against the fitted model, we observe that Figure 4 shows clear differences among the fit of the five distributions, with increasing evidence against the fitted model as we move away from the assumed (true) GB2 model.

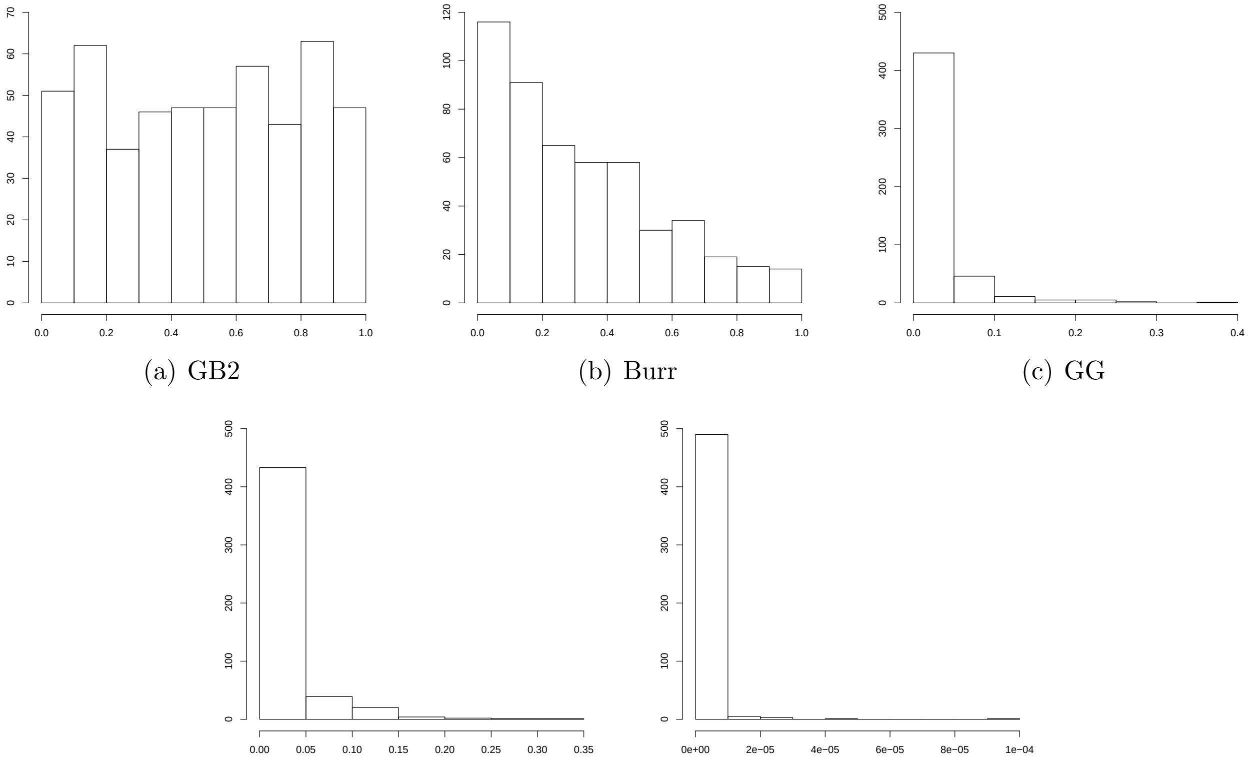
Table 3: Values of DIC under different models for a random sample of 500 delay values.

	GB2	Burr	GG	Log-normal	Pareto
DIC	5951	6039	6090	6139	6198

4.3.2 Real data results

We now apply the method to the entire data set of delay times. Figure 5 shows the distribution of $\boldsymbol{q}^{t'}$ at iteration t' which gives the highest p-value. Visual inspection shows that the distribution under the GB2 model is the closest to $U(0,1)$. However, at this iteration we have $\pi_Q^{t'} = 0.0413$, while for the Burr distribution the corresponding p-value is $\pi_Q^{t'} = 5 \times 10^{-9}$. The distributions of $\boldsymbol{q}^{t'}$ under the GG, log-normal and Pareto models do not show uniformity and indeed all $\boldsymbol{\pi}_Q$ are practically zero under these three models.

For the GB2 and Burr models, the distribution of p-values, $\boldsymbol{\pi}_Q$, are given in Figure 6. The p-values shown here are obtained using a Kolmogorov-Smirnov goodness-of-fit test for $\boldsymbol{q}^t \sim U(0,1)$, but results have also been confirmed using bootstrap techniques for the sampling distribution of the relevant statistic. Both $\boldsymbol{\pi}_Q$ posterior distributions demonstrate very small p-values; however under the Burr model the distribution is more heavily concentrated close to zero, with 100% of $\boldsymbol{\pi}_Q$ values being smaller than 5×10^{-9}. We note here that the magnitude of the evidence against these models, as shown by the small p-values, can be partly explained by the large volume of data



(a) GB2 (b) Burr (c) GG

(d) Log-normal (e) Pareto

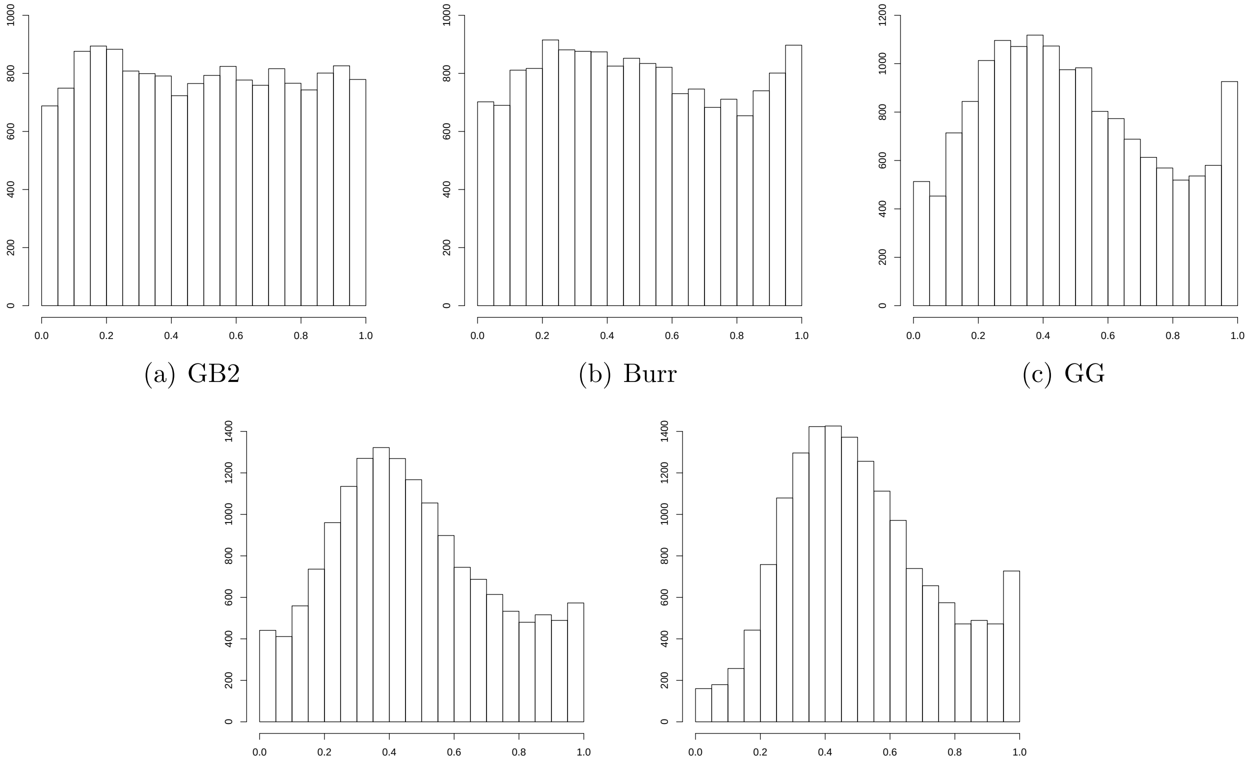
Figure 4: Posterior distributions of p-values $\boldsymbol{\pi}_Q$ for assessment of different models using a random sample of 500 delay values. Note that, for presentation purposes, the scale on the x axis is not the same in all plots.

and extended number of covariate combinations implying that certain clusters of observations may not be fitted well. Nevertheless, the results here reveal that in terms of adequacy of predictive behaviour, the evidence against the Burr model is relatively much stronger. This is in agreement with the model comparison results in Sections 4.1 and 4.2. Furthermore, 95% credible intervals of the fitted delay contained 95.3% of the observed delay values under the GB2 model, and 94.4% of values under the Burr model. Overall, our analysis provides solid Bayesian arguments against the examined simpler distributions when considering the more general GB2.

5 Bayesian variable selection with GB2

For prediction purposes it is important to choose the most suitable model. In our case the 10 covariates lead to $2^{10} = 1024$ possible models, denoted as $m_1, \ldots, m_{1024}$, and we want to choose the 'best' model which includes a selected subset of covariates. Since posterior model probabilities are not always analytically tractable, we use an MCMC approach (specifically Gibbs variable selection, see Dellaportas et al. (2002)) to generate a sample from the joint posterior distribution and observe the relative frequencies of the 1024 competing models. Let $\boldsymbol{\kappa}$ be a $p \times 1$ indicator vector, where $\kappa \in \{0, 1\}^p$, showing which covariates are in the model. For example if $\kappa_i = 1$ then the ith covariate is present in the model.

We consider the GB2 GL-type model as discussed in Section 3. Under Bayesian variable



(a) GB2

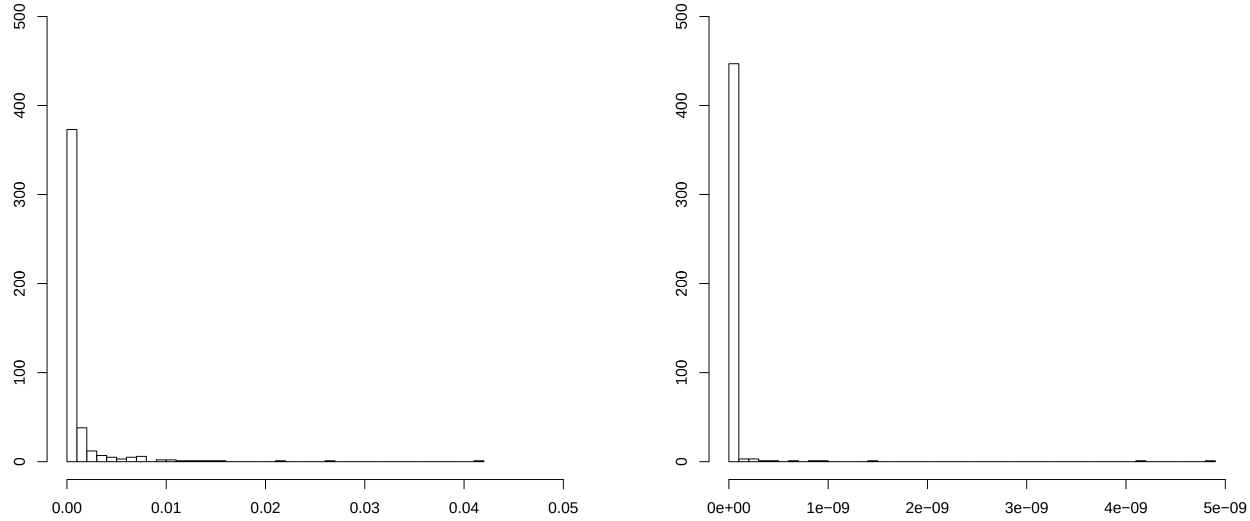
(b) Burr

(c) GG

(d) Log-normal

(e) Pareto

Figure 5: Distributions of $\boldsymbol{q}^{t'}$ for different models at the iteration t' where the p-value takes its maximum.



(a) GB2

(b) Burr

Figure 6: Posterior distributions of p-values $\boldsymbol{\pi}_Q$ under the GB2 and Burr models. Note that, for presentation purposes, the scale on the x axis is not the same in two plots.

selection the linear predictor becomes

$$\eta_i = \beta_0 + \sum_{j=1}^{8} \kappa_j \beta_j z_{ij} + \kappa_9 \beta_{9\,O_i} + \kappa_{10} \beta_{10\,C_i},$$

The prior we specify for $(\boldsymbol{\kappa}, \boldsymbol{\beta})$ is in the form of $f(\boldsymbol{\kappa}, \boldsymbol{\beta}) = f(\boldsymbol{\kappa}) f(\boldsymbol{\beta}|\boldsymbol{\kappa})$. That is, the prior of parameter vector $\boldsymbol{\beta}$ is conditional on the model including a subset of covariates $\boldsymbol{\kappa}$. It is common to use $f(\kappa_j) = Bernoulli(0.5)$ as a prior for indicators when there are no restrictions on the model space (see Chipman (1996) and George and McCulloch (1993)). In other words, we give equal probabilities to each of the possible models. We note that different Bernoulli(p) priors could be employed here to reflect stronger preference towards inclusion of certain covariates in the model, for example to accommodate expert beliefs in the field. On the other hand, there are various approaches in the literature for the choice of $f(\boldsymbol{\beta}|\boldsymbol{\kappa})$. Here we use empirical Bayes priors which use information equivalent to that contained in one data point (Kass and Wasserman, 1995).

The estimates of the posterior inclusion probabilities, $\hat{f}(\kappa|\boldsymbol{D})$, and their standard deviations are given in Table 4. According to this, policy type, benefit amount, policy duration, office and cause of claim are found to be important with more than 95% inclusion probability. Age and benefit type have inclusion probabilities that are relatively high but less than 50%, while sex, smoker status and settlement year are only included with less than 20% probability. Table 5 shows the highest five posterior model probabilities, $\hat{f}(m|\boldsymbol{D})$. The first model, m_{977}, excludes all the covariates with less than 50% posterior inclusion probability in Table 4. The differences between m_{977} and m_{981}, m_{978} and m_{982} are due to the inclusion of benefit type, age, or both benefit type and age, respectively. The relatively high posterior inclusion probabilities of benefit type and age leads to very close posterior model probabilities.

Table 4: Posterior inclusion probabilities under the GB2 distribution.

Parameter		$\hat{f}(\kappa\|\boldsymbol{D})$	sd
Age	β_1	0.3738	0.4838
Sex	β_2	0.1370	0.3438
Benefit type	β_3	0.4440	0.4969
Smoker status	β_4	0.1758	0.3806
Policy type	β_5	0.9997	0.0176
Settlement year	β_6	0.0617	0.2406
Benefit amount	β_7	0.9687	0.1742
Policy duration	β_8	1	0
Office	β_9	1	0
Cause	β_{10}	1	0

To interpret these probabilities we use posterior odds, defined as $PO = \frac{f(m_j|\boldsymbol{D})}{f(m_k|\boldsymbol{D})}\frac{f(m_j)}{f(m_k)}, j \neq k$, which under the same prior distribution for each model are equal to the Bayes factor. Models with Bayes factor less than 3 are barely different (Kass and Raftery, 1995), and therefore we can not distinguish between the first four models in Table 5. These results were also confirmed using an approximation to the marginal likelihood for the GB2 model, as shown in Section S5 in supplementary material.

We can also compare the first five models given in Table 5 employing the LLR test we introduced in Section 4.2. We compare (a) m_{977} to m_{979}; and (b) m_{977} to m_{982}. In both cases, at each iteration

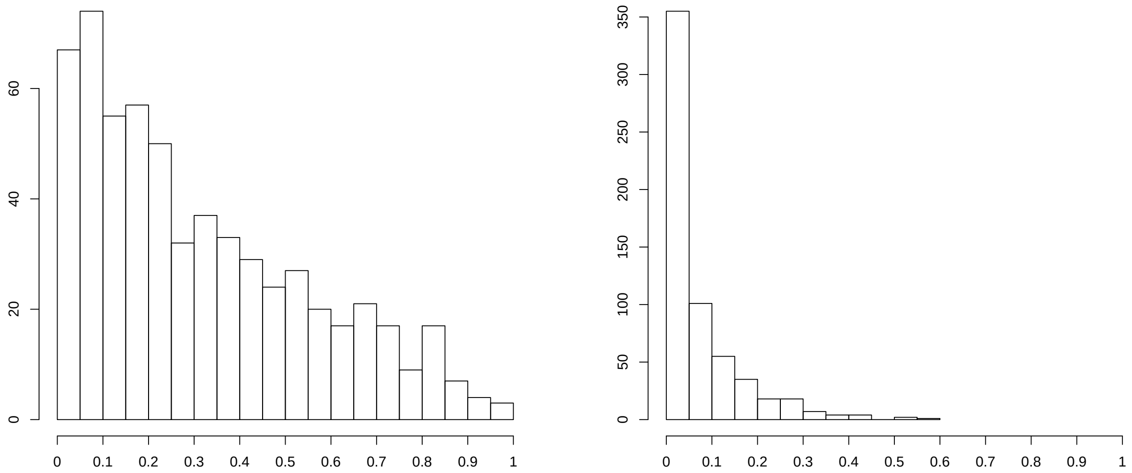
Table 5: Posterior model probabilities under the GB2 distribution using the same covariate subscripts as presented in Table 4.

	Model		$\hat{f}(m\|\boldsymbol{D})$	$PO(m_{977}/.)$
1	m_{977}	$z_5 + z_7 + z_8 + z_9 + z_{10}$	0.1996	1.00
2	m_{981}	$z_3 + z_5 + z_7 + z_8 + z_9 + z_{10}$	0.1843	1.08
3	m_{978}	$z_1 + z_5 + z_7 + z_8 + z_9 + z_{10}$	0.1503	1.33
4	m_{982}	$z_1 + z_3 + z_5 + z_7 + z_8 + z_9 + z_{10}$	0.1007	1.98
5	m_{979}	$z_2 + z_5 + z_7 + z_8 + z_9 + z_{10}$	0.0449	4.44

(t) we need to compute

$$\Lambda_j^{(t)} = \frac{L_{m_{977}}\left(\alpha^{(t)}, \tau^{(t)}, \gamma^{(t)}, \boldsymbol{\beta}^{(t)}; \boldsymbol{D}\right)}{L_i\left(\dot{\alpha}, \dot{\tau}, \dot{\gamma}, \dot{\boldsymbol{\beta}}; \boldsymbol{D}\right)}$$

where $i = m_{979}, m_{982}$ and $j = a, b$, respectively. To obtain the posterior p-values, π_Λ, we calculate the tail probabilities as described in Section 4.2 with $df = 29, 30$ for cases (a) and (b). The posterior distributions of these p-values can be seen in Figure 7. In case (a) the proportion of p-values that are smaller than 0.01 is 2%, giving no evidence at all against m_{977} when compared to m_{979}. In case (b), the corresponding proportion is 23%, and the plot now suggests some weak evidence against m_{977} when compared to the larger model m_{982}. Again, these findings are consistent with the results in Table 5.



(a) Comparison of m_{977} and m_{979}.

(b) Comparison of m_{977} and m_{982}.

Figure 7: Posterior distributions of p-values π_Λ under the GB2 distribution with different covariates.

6 Bayesian model averaging and prediction

In Section 5, we showed that under the GB2 GL-type setting there is a number of candidate models with relatively small difference in posterior model probability. Although a highest probability model can be identified, the small differences in relative posterior odds point towards the use of model averaging across models. This will enable us to incorporate model uncertainty in a natural manner and will provide better average predictive ability than using a single model (Madigan and Raftery, 1994). Therefore, we perform model-averaging across models for which the posterior odds (or Bayes factor here) are less than 3, following Kass and Raftery (1995). We estimate the missing delay using the first four models in Table 5 separately, and average over these selected models. Under model uncertainty, the posterior probability distribution for the quantity of interest (e.g. parameters, missing delay), $\boldsymbol{\Delta}$ can be given as $f(\boldsymbol{\Delta}|\boldsymbol{D}) = \sum\limits_{k} f(\boldsymbol{\Delta}|m_k, \boldsymbol{D})f(m_k|\boldsymbol{D})$ which is a mixture of the posterior probability distributions for $\boldsymbol{\Delta}$, obtained from the individual models in the usual way, weighted by their posterior probabilities $f(m_k|\boldsymbol{D}) \propto f(\boldsymbol{D}|m_k)f(m_k)$. Since we are only interested in the first four models, all probabilities are implicitly conditional on this set of models and we need to reassign the posterior probabilities of these models. So, for example, the reassigned posterior probability of m_{977} is

$$\hat{f}^{'}(m_{977}|\boldsymbol{D}) = 0.1996/(0.1996 + 0.1843 + 0.1503 + 0.1007) = 0.3144.$$

To illustrate prediction of the delay using model averaging, we identify eight observations with different characteristics which have missing delay. These are given in Table 6. In each case we choose a covariate with a different value (or level) in order to explore the effect of these claim-related characteristics on delay prediction. In Table 6, missing observation 1 serves as a base case for missing observations 2-6, while missing observation 6 serves as a base case for missing observations 7 and 8. The value or level changes in observations are highlighted in boldface. For each observation, posterior estimates of the mean of the missing delay distribution under the 'best' model (m_{977}) and average model are given in Table 7. According to these, for shorter policy durations (observation 2) the delay is longer. Having a joint life policy (observation 6) and increasing benefit amount (observations 3 and 4) lead to shorter delay duration. This might be due to shorter notification period. On the other hand, MS (observation 5) or cancer claims (observations 5 and 8) have longer delay predictions compared to death claims. Also some offices appear to settle the claims significantly faster than others (observation 7).

Note that the missing delay predictions for observations 1-7 are very close under model m_{977} and the average model (see Table 7), whereas there is a small difference between the predictions under these two approaches for observation 8. This difference could be explained by covariate benefit type; for missing observations 1-7 its level is 'full accelerated' (baseline category – see Table 1), while for observation 8 it is 'stand alone'. Note that under m_{981} and m_{982}, benefit type (z_3) is retained in the model and the coefficient is estimated as -0.0199 (CI: -0.0301, -0.0099) and -0.0215 (CI: -0.0311, -0.0130), respectively. The negative coefficient for benefit type causes shorter delay predictions for observation 8 under m_{981} (305.57; CI: 280.78, 328.00) and m_{982} (308.75; CI: 284.81, 331.92) compared to the models which do not include this covariate, i.e. m_{977} (323.44; CI: 298.88, 350.99) and m_{978} (322.94; CI: 302.26, 343.12). The average model provides an average prediction (315.81; CI: 292.19, 339.43) and demonstrates how model uncertainty could be incorporated under the Bayesian model averaging.

Table 6: Characteristics of the missing observations considered for prediction.

Missing observation	1	2	3	4	5	6	7	8
Policy type	JL	JL	JL	JL	JL	**SL**	SL	SL
Benefit amount (GBP)	50 000	50 000	**5 000**	**115 000**	50 000	50 000	50 000	50 000
Policy duration (years)	> 5	$<$**1**	> 5	> 5	> 5	> 5	> 5	> 5
Office	2	2	2	2	2	2	**11**	2
Cause	Death	Death	Death	Death	**MS**	Death	Death	**Cancer**
Benefit type	FA	FA	FA	FA	FA	FA	FA	SA

Table 7: Posterior estimates of the mean of the missing delay distribution under the 'best' model (m_{977}) and average model for the observations summarised in Table 6.

		1	2	3	4	5	6	7	8
m_{977}	Mean	221.22	271.60	234.59	214.06	403.36	244.60	152.62	323.44
	95%CI	(203.58, 240.24)	(251.39, 292.70)	(216.81, 254.16)	(196.23, 232.72)	(368.40, 440.64)	(225.47, 265.05)	(141.08, 163.55)	(298.88, 350.99)
Average model	Mean	220.26	270.40	233.49	213.25	402.44	244.06	152.68	315.81
	95%CI	(202.87, 236.22)	(249.89, 289.09)	(214.97, 250.38)	(196.18, 228.91)	(366.83, 439.28)	(224.84, 262.75)	(141.24, 163.64)	(292.19, 339.43)

7 Discussion

Accurate prediction of the settlement delay for CII claims, that also accounts for the associated uncertainty, is very important for premium pricing, estimation of future liabilities and solvency of insurance companies. The research presented in this paper can therefore contribute towards the practical implementation of robust statistical methodology in this field. At the same time, our work brings forward the application of a number of methodological advances into the area of CII, including the use of the GB2 distribution in a GL-type setting and relevant variable selection, and the assessment of such models, also in the presence of incomplete data, using recently developed techniques. The use of these methodological advances, as presented here, can also be generalised in a range of applications concerning lifetime, survival, or other heavy-tailed data.

The GB2 distribution is a very flexible distribution which encompasses various distributions relevant in modelling insurance data. Previously, CMI (2008) and Ozkok et al. (2012) found that a 3-parameter Burr distribution provides a very good fit to the delay between diagnosis and settlement in critical illness insurance. In this paper different model comparison approaches suggested that the 4-parameter GB2 distribution improves the fit to the data considerably, when compared to the fit under the nested Burr distribution. While the Burr model runs (as perhaps expected) faster, at approximately 79% of the CPU time required for the GB2 model, our analysis demonstrates that the latter does not increase the variation of parameter estimates, and is also more efficient in dealing with missing values. These considerable improvements also suggest that there are marked differences in the fit of the models, thus highlighting the challenge in fitting detailed models to such long-tailed data.

When missing observations are considered in the analysis, we compared models using different versions of DIC. While the DIC approach strongly demonstrated that the GB2 performs better than competing distributions, the comparison outcome among less well performing models was not clear due to the complexities associated with incomplete data. Therefore, we have provided here methodology relying on a latent likelihood ratio test within a hybrid Bayesian-classical framework. This has the advantage that does not involve prior distributions under both competing models, while also being designed to deal with missing data models and can therefore provide an alternative to more commonly used model comparison techniques. A possible limitation of this approach can appear under an excessive volume of missing data, as these are only estimated under the hypothesised model. This may potentially lead to bias in favour of the hypothesised model, but mainly in cases of severe under-observation of the data generation process which are not typical of the applications presented here.

Model assessment was also performed, using posterior distributions of p-values related to posterior predictive checking. Similarly to other methods based on p-values, this approach is also affected by the sample size, where typically as the sample size increases, the p-value decreases. Nevertheless, we confirmed the ability of the method to correctly assess candidate models using a simulation study based on 500 random samples. Using the full data set (more than 19,000 observations), results based on p-values related to posterior predictive checking strengthened earlier conclusions in the paper, by showing that evidence against the predictive adequacy GB2 model is weaker than for the Burr and other considered models.

Finally, Bayesian variable selection results suggested that it is not straightforward to distinguish between the four highest probability models. Hence, instead of choosing the 'best' model, we opted to use a Bayesian model averaging approach. Missing data were imputed using the average model. Predictions under the 'best' and the average model were presented for a set of missing observations, demonstrating an advantage for the model averaging approach in certain cases.

Acknowledgement

This research has been partly funded by the Institute and Faculty of Actuaries, UK, within the Pump-priming Research Grant scheme. The authors are also grateful to the Continuous Mortality Investigation of the UK for supplying the data.

References

Bennett, J. E., Li, G., Foreman, K., Best, N., Kontis, V., Pearson, C., Hambly, P. and Ezzati, M. (2015) The future of life expectancy and life expectancy inequalities in England and Wales: Bayesian spatiotemporal forecasting. *The Lancet*, **386**, 163 – 170.

Cairns, A. J., Blake, D. and Dowd, K. (2006) A two-factor model for stochastic mortality with parameter uncertainty: Theory and calibration. *Journal of Risk and Insurance*, **73**, 687–718.

Celeux, G., Forbes, F., Robert, C. P., Titterington, D. M. et al. (2006) Deviance information criteria for missing data models. *Bayesian analysis*, **1**, 651–673.

Chipman, H. (1996) Bayesian variable selection with related predictors. *Canadian Journal of Statistics*, **24**, 17–36.

CMI (2008) A new methodology for analysing CMI critical illness experience. *Tech. Rep. WP33*, The Faculty of Actuaries and Institute of Actuaries, UK.

Cummins, J. D., Dionne, G., McDonald, J. B. and Pritchett, B. M. (1990) Applications of the GB2 family of distributions in modeling insurance loss processes. *Insurance: Mathematics and Economics*, **9**, 257–272.

DeIorio, M. and Robert, C. P. (2002) Discussion of Spiegelhalter et al. *Journal of the Royal Statistical Society, Series B*, **64**, 629–630.

Dellaportas, P., Forster, J. J. and Ntzoufras, I. (2002) On Bayesian model and variable selection using MCMC. *Statistics and Computing*, **12**, 27–36.

Dong, A. and Chan, J. (2013) Bayesian analysis of loss reserving using dynamic models with generalized beta distribution. *Insurance: Mathematics and Economics*, **53**, 355–365.

Draper, N. R. and Smith, H. (2014) *Applied regression analysis*. John Wiley & Sons.

Frees, E. W. and Valdez, E. A. (2008) Hierarchical insurance claims modeling. *Journal of the American Statistical Association*, **103**, 1457–1469.

George, E. I. and McCulloch, R. E. (1993) Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, **88**, 881–889.

Jones, A. M., Lomas, J. and Rice, N. (2014) Applying beta-type size distributions to healthcare cost regressions. *Journal of Applied Econometrics*, **29**, 649–670.

Kass, R. E. and Raftery, A. E. (1995) Bayes factors. *Journal of the American Statistical Association*, **90**, 773–795.

Kass, R. E. and Wasserman, L. (1995) A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *Journal of the American Statistical Association*, **90**, 928–934.

Klugman, S. A., Panjer, H. H. and Willmot, G. E. (2012) *Loss models: from data to decisions*, vol. 715. John Wiley & Sons.

Lau, M., Marion, G., Streftaris, G. and Gibson, G. (2014) New model diagnostics for spatio-temporal systems in epidemiology and ecology. *Journal of the Royal Society Interface*, **11: 20131093**.

Madigan, D. and Raftery, A. E. (1994) Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association*, **89**, 1535–1546.

McDonald, J. B. (1984) Some generalized functions for the size distribution of income. *Econometrica: Journal of the Econometric Society*, 647–663.

McDonald, J. B. and Butler, R. J. (1990) Regression models for positive random variables. *Journal of Econometrics*, **43**, 227–251.

McDonald, J. B. and Xu, Y. J. (1995) A generalization of the beta distribution with applications. *Journal of Econometrics*, **66**, 133 – 152.

Ozkok, E., Streftaris, G., Waters, H. R. and Wilkie, A. D. (2012) Bayesian modelling of the time delay between diagnosis and settlement for Critical Illness Insurance using a Burr generalised-linear-type model. *Insurance: Mathematics and Economics*, **50**, 266–279.

— (2014) Modelling critical illness claim diagnosis rates I: methodology. *Scandinavian Actuarial Journal*, **2014**, 439–457.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and Van Der Linde, A. (2002) Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **64**, 583–639.

Streftaris, G. and Gibson, G. J. (2012) Non-exponential tolerance to infection in epidemic systems—modeling, inference, and assessment. *Biostatistics*, **13**, 580–593.

Venter, G. G. (1983) Transformed beta and gamma distributions and aggregate losses. In *Proceedings of the Casualty Actuarial Society*, vol. 70, 156–193.