Supplementary material for Prediction of settlement delay in critical illness insurance claims using GB2 distribution by Erengul Dodd and George Streftaris

S1 Fitting related distributions

Similar results for s_i^* as in Section 3.1, can be derived for the GG distribution. For the nested Burr and Pareto distributions, equation (7) can be adjusted accordingly by substituting $\gamma = 1$ and $\gamma = \tau = 1$, respectively. On the other hand for the log-normal distribution we employ standard weighted glm where, again, the variance is inversely proportional to the weights.

The posterior distributions of parameters α and τ of the GG distribution are shown in Figure S1 and, as again discussed in Section 4, they provide evidence against other simpler distributions. Under all considered distributions we use the same linear predictor as in Section 3 of the main paper and the same priors for $\boldsymbol{\beta}$, i.e. $\mathrm{N}(0, 10^4)$. The prior distributions for the other model parameters are given below.

GG distribution:

$$\begin{aligned} \alpha &\sim \text{Gamma}(1, 0.01) \\ \tau &\sim \text{Gamma}(1, 0.01) \end{aligned}$$

Burr distribution:

$$\begin{aligned} \alpha &\sim \text{Gamma}(1, 0.01) I(1/\tau, \infty) \\ \tau &\sim \text{Gamma}(1, 0.01) \end{aligned}$$

Log-normal distribution:

$$\sigma^2 \sim \text{Inverse Gamma}(1, 0.1)$$

Pareto distribution:

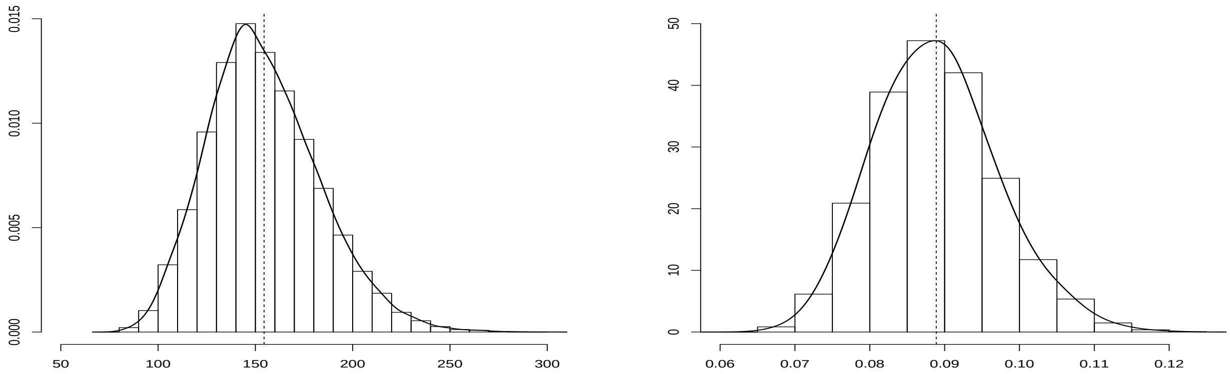
$$\alpha \sim \text{Gamma}(1, 0.01) I(1, \infty)$$

We summarise the posterior estimates of the shape parameters of the Burr and Pareto distributions, and the scale parameter of the log-normal distribution together with 95% credible intervals of these parameters in Table S1.

We note here that the posterior estimates of the GB2 model in Section 3.1, as illustrated in the densities of model parameters shown in Figure 2, also provide evidence against the Pareto model (for which $\gamma = \tau = 1$). Similarly, Figure S1 illustrates that under the GG model, parameters α and τ are far from 1 and therefore related simpler nested models (Gamma, Weibull, exponential) are not supported.

Table S1: Posterior means and 95% credible intervals (CI) of shape parameters of the Burr and Pareto distributions and the scale parameter of the log-normal distribution.

	Parameter	Mean	95% CI
Burr	α	0.5869	(0.5564, 0.6189)
	τ	2.6221	(2.5452, 2.7008)
Pareto	α	6.8084	(6.1727, 7.5150)
Log-normal	σ^2	1.0224	(0.9992, 1.0463)



(a) α

(b) τ

Figure S1: Posterior densities and histograms of model parameters together with their posterior means (dashed line) under the GG distribution.

S2 DIC

In the main paper we consider the following three versions of DIC (keeping the original indices in Celeux et al. (2006)):

$$\begin{aligned}\mathrm{DIC}_4 = &- 4E_{\boldsymbol{\theta},\boldsymbol{D}_{mis}|\boldsymbol{D}_{obs}}\left[\log f(\boldsymbol{D}_{obs},\boldsymbol{D}_{mis}|\boldsymbol{\theta})\right] \\ &+ 2E_{\boldsymbol{D}_{mis}|\boldsymbol{D}_{obs}}\left[\log f\left(\boldsymbol{D}_{obs},\boldsymbol{D}_{mis}|E_{\boldsymbol{\theta}}\left[\boldsymbol{\theta}|\boldsymbol{D}_{obs},\boldsymbol{D}_{mis}\right]\right)\right],\end{aligned} \tag{S1}$$

$$\begin{aligned}\mathrm{DIC}_5 = &- 4E_{\boldsymbol{\theta},\boldsymbol{D}_{mis}|\boldsymbol{D}_{obs}}\left[\log f(\boldsymbol{D}_{obs},\boldsymbol{D}_{mis}|\boldsymbol{\theta})\right] \\ &+ 2\left(\log f\left(\boldsymbol{D}_{obs},\hat{\boldsymbol{D}}_{mis}(\boldsymbol{D}_{obs})|\hat{\boldsymbol{\theta}}(\boldsymbol{D}_{obs})\right)\right),\end{aligned} \tag{S2}$$

$$\begin{aligned}\mathrm{DIC}_8 = &- 4E_{\boldsymbol{\theta},\boldsymbol{D}_{mis}|\boldsymbol{D}_{obs}}\left[\log f(\boldsymbol{D}_{obs}|\boldsymbol{D}_{mis},\boldsymbol{\theta})\right] \\ &+ 2E_{\boldsymbol{D}_{mis}|\boldsymbol{D}_{obs}}\left[\log f\left(\boldsymbol{D}_{obs}|\boldsymbol{D}_{mis},\hat{\boldsymbol{\theta}}(\boldsymbol{D}_{obs},\boldsymbol{D}_{mis})\right)\right].\end{aligned} \tag{S3}$$

In (S2) $\left(\hat{\boldsymbol{D}}_{mis}(\boldsymbol{D}_{obs}),\hat{\boldsymbol{\theta}}(\boldsymbol{D}_{obs})\right)$ is the joint maximum a posteriori estimator of $(\boldsymbol{D}_{mis},\boldsymbol{\theta})$. Following Celeux et al. (2006) we use the pair of $(\boldsymbol{D}_{mis},\boldsymbol{\theta})$ from the MCMC iterations that actually give the highest value of the non-standardised posterior density, i.e. $f(\boldsymbol{D}_{obs},\boldsymbol{D}_{mis}|\boldsymbol{\theta})f(\boldsymbol{\theta})$. In (S3) the $\hat{\boldsymbol{\theta}}(\boldsymbol{D}_{obs},\boldsymbol{D}_{mis})$ denotes the posterior mean of $\boldsymbol{\theta}$.

S3 Latent likelihood ratio test

In more general terms, at each MCMC post-convergence iteration $t = 1, \ldots, N$, we compute the latent value of the likelihood ratio Λ as

$$\Lambda^{(t)} = \frac{L_1\left(\boldsymbol{\theta}^{(t)}; \boldsymbol{D}\right)}{L_2\left(\dot{\boldsymbol{\theta}}; \boldsymbol{D}\right)}$$

where $\boldsymbol{\theta}^{(t)}$ are MCMC posterior estimates at iteration (t) and $\dot{\boldsymbol{\theta}}$ 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 $\mathcal{M}_2$. If models are not nested, we can employ simulation to obtain the empirical sampling distribution of Λ. First we generate a sample from $\mathcal{M}_1$

$$\boldsymbol{D}^* \sim \mathcal{M}_1\left(\boldsymbol{\theta}^{(t)}\right) \tag{S4}$$

and calculate

$$\Lambda^{(t)} = \frac{L_1\left(\boldsymbol{\theta}^{(t)}; \boldsymbol{D}^*\right)}{L_2\left(\dot{\boldsymbol{\theta}}; \boldsymbol{D}^*\right)}. \tag{S5}$$

Then we can either obtain the expectation of π_Λ as an ergodic mean by computing binary $\mathbb{1}(\Lambda \leq \Lambda^{(t)})$ and estimating the expectation of the posterior distribution of the p-values as

$$\begin{aligned} E(\pi_\Lambda|\boldsymbol{D}) &= \int P\left(\Lambda \leq \Lambda^{(t)}\right) f\left(\boldsymbol{\theta}|\boldsymbol{D}\right) d\boldsymbol{\theta} \\ &\approx \frac{1}{N}\sum \mathbb{1}(\Lambda \leq \Lambda^{(t)}), \end{aligned}$$

or we can obtain the entire posterior distribution of p-values using a posterior sample $\pi_\Lambda^{(1)}, \ldots, \pi_\Lambda^{(N)}$, by repeating the steps given in (S4) and (S5) in each MCMC iteration. Here, we prefer the repeated simulation approach when the two models under comparison are not nested.

S4 Efficiency in dealing with missing values

In the context of the work presented in the main paper, assessment of model efficiency in dealing with missing values is not as straightforward as in multiple imputation analysis (e.g. Rashid et al. (2015)). However, we have compared the relative posterior variance of model β parameters, including and excluding missing values under the GB2 and Burr model, using the following relative efficiency measure

$$\frac{sd^{MV}_{GB2}}{sd_{GB2}} - \frac{sd^{MV}_{Burr}}{sd_{Burr}}, \tag{S6}$$

where sd_{GB2}^{MV} denotes the posterior standard deviation of each parameter under the GB2 model when missing values are included in the analysis - and similarly for the other cases. Therefore, negative values indicate that the GB2 distribution is more efficient in dealing with missing values. A barplot of the values of this measure for the 30 β coefficients is shown in Figure S2 and suggests that the GB2 model outperforms the Burr model (note that this difference has also a negative mean of 0.024).

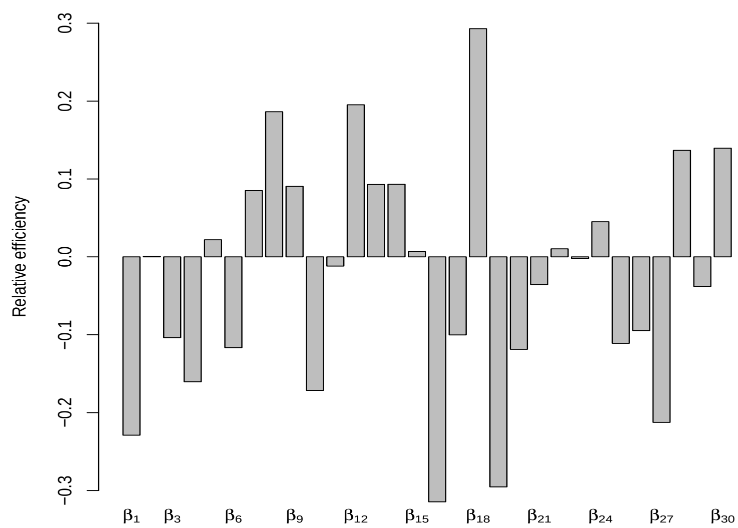


Figure S2: Barplot comparing the efficiency of the GB2 and Burr model in dealing with missing values. The relative efficiency measure (S6) has been computed for the β coefficients of the models.

S5 Variable selection using marginal likelihoods

We can confirm the results found in Section 5 using an approximation to the marginal likelihood for the GB2 model, since its exact analytical calculation is not tractable. If we denote these models by m where $m \in (m_1, \ldots, m_{1024})$ and the parameter vector of model m by $\boldsymbol{\theta}_m$, then the marginal likelihood can be expressed as

$$f(\boldsymbol{D}|m) = \int f(\boldsymbol{D}|\boldsymbol{\theta}_m, m) f(\boldsymbol{\theta}_m|m) d\boldsymbol{\theta}_m. \tag{S7}$$

We consider the Laplace approximation to obtain (S7), that is

$$\tilde{f}(\boldsymbol{D}|m) = (2\pi)^{d_m/2} |\tilde{\Sigma}_m|^{1/2} f(\boldsymbol{D}|\tilde{\boldsymbol{\theta}}_m, m) f(\tilde{\boldsymbol{\theta}}_m|m) \tag{S8}$$

where $\tilde{\boldsymbol{\theta}}_m$ denotes the parameter vector of model m calculated at the posterior modes, d_m is the dimension of model m and $\tilde{\Sigma} = (-H(\tilde{\boldsymbol{\beta}}))^{-1}$ is the covariance matrix with the Hessian matrix of the likelihood, H, evaluated at the posterior modes of the parameters. The resulting posterior model probabilities based on the approximated marginal likelihood under the GB2 distribution are given in Table S2 and are very similar to those calculated employing the MCMC approach (see Section 5; Table 5). This is not surprising as this approximation to marginal likelihoods is known to give

Table S2: Laplace approximation for the marginal likelihood.

	Model	$\tilde{f}(m\|\boldsymbol{D})$	$PO(m_{981}/.)$
1	m_{981}	0.2129	1.00
2	m_{977}	0.2098	1.01
3	m_{978}	0.1666	1.27
4	m_{982}	0.1120	1.90
5	m_{979}	0.0412	5.16

good results with large sample sizes (Gelman et al., 2000). More information on this approach can be found in Kass and Raftery (1995).

References

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.

Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (2000) *Bayesian Data Analysis*. Chapman and Hall/CRC.

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

Rashid, S., Mitra, R. and Steele, R. (2015) Using mixtures of t densities to make inferences in the presence of missing data with a small number of multiply imputed data sets. *Computational Statistics & Data Analysis*, **92**, 84–96.

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