Epidemiol. Infect. (2016), **144**, 1959–1973. © Cambridge University Press 2016
doi:10.1017/S0950268815003234

Variable selection and regression analysis for the prediction of mortality rates associated with foodborne diseases

E. AMENE[1]*, L. A. HANSON[2], E. A. ZAHN[2], S. R. WILD[2] AND D. DÖPFER[1]

[1] *University of Wisconsin-Madison, WI, USA*
[2] *Saint Olaf College, Northfield, MN, USA*

Received 28 June 2015; Final revision 28 November 2015; Accepted 2 December 2015; first published online 20 January 2016

SUMMARY

The purpose of this study was to apply a novel statistical method for variable selection and a model-based approach for filling data gaps in mortality rates associated with foodborne diseases using the WHO Vital Registration mortality dataset. *Correlation analysis* and *elastic net regularization* methods were applied to drop redundant variables and to select the most meaningful subset of predictors. Whenever predictor data were missing, multiple imputation was used to fill in plausible values. *Cluster analysis* was applied to identify similar groups of countries based on the values of the predictors. Finally, a *Bayesian hierarchical regression model* was fit to the final dataset for predicting mortality rates. From 113 potential predictors, 32 were retained after correlation analysis. Out of these 32 predictors, eight with non-zero coefficients were selected using the elastic net regularization method. Based on the values of these variables, four clusters of countries were identified. The uncertainty of predictions was large for countries within clusters lacking mortality rates, and it was low for a cluster that had mortality rate information. Our results demonstrated that, using Bayesian hierarchical regression models, a data-driven clustering of countries and a meaningful subset of predictors can be used to fill data gaps in foodborne disease mortality.

Key words: Bayesian hierarchical regression, cluster analysis, elastic net, foodborne diseases.

INTRODUCTION

Foodborne diseases (FBDs) remain a growing concern for high levels of morbidity and mortality in the human population worldwide [1]. There are many indicators hinting at an increase in the global incidence of FBDs [2, 3]. For industrialized countries, it has been estimated that one-third of the population suffers from foodborne illness every year [4]. A recent study estimated 37·2 million illnesses, 228 744 hospitalizations, and 2612 deaths each year due to FBDs in the United States alone [5]. Furthermore, FBD morbidity and mortality are assumed to be extensive in resource-limited regions of the world although solid data are lacking in these regions [6, 7]. Beyond morbidity and mortality, FBDs also exert an important impact on development and trade [8].

Challenges associated with incomplete data have been emphasized in studies estimating the global burden of pathogen-specific FBDs such as non-typhoidal *Salmonella* gastroenteritis and typhoid fever [9, 10]. The lack of data, particularly from developing countries, makes it difficult to calculate global estimates of disease burden, and this hinders appropriate allocation and prioritization of resources for food safety intervention efforts. The current study explores novel

* Author for correspondence: Dr E. Amene, University of Wisconsin-Madison, 2015 Linden Drive, Madison 53706 WI, USA. (Email: amene@wisc.edu)

Bayesian statistical methods for addressing missing data using the WHO Vital Registration (VR) mortality dataset. The specific aims are twofold: to select a subset of meaningful predictors of FBD mortality and to predict missing FBD mortality rates using the selected variables. This report represents a first attempt to cluster WHO countries based on average values of the selected predictors, and to use the variables and the clusters in a Bayesian hierarchical modelling framework to predict FBD mortality rates for countries with missing data.

METHODS

Dataset

Mortality data

Data regarding mortality rates associated with FBDs were obtained from the WHO VR database (2000–2005). The FBDs in the database include bacterial and viral gastroenteritis, parasitic diseases, and hepatitis A and E. The International Classification of Diseases coding system (ICD-10) was used to classify those diseases; FBDs associated with chemicals and biotoxins were not included in this study due to the lack of specific ICD codes [11, 12]. Mortality rates were averaged over the available years and the mean rate was expressed per 100 000 population based on the 2005 population census. We log-transformed the mortality rate data to stabilize its variance and allow the data to be more normally distributed. Out of 194 WHO countries, only 48 had complete data about national mortality rates associated with FBDs.

Predictor set

We obtained predictors for mortality associated with FBDs from publicly available databases (Food and Agricultural Organization; FAO [13], and World Bank [14]) in two steps. First, we identified 113 predictors from these databases for 48 countries with complete mortality rate data. Next, we selected a meaningful subset of predictors from those 113 predictors. Finally, we retrieved the values for the selected predictors from all 194 WHO countries. The search criteria for the predictors included an established direct and/or indirect association with FBD mortality and the predictor's potential to be a modifiable risk factor, i.e. whether it can be changed through intervention [2, 12, 15].

Statistical analysis

Analyses of the data were performed using the freeware statistical tools R version 3.1.2 and JAGS (Just Another Gibbs Sampler) version 3.4.0 [16, 17]. The models were specified and parameterized in R and the analyses were performed by calling JAGS from R using the *R2jags* package [17]. A stepwise approach was followed for variable selection and estimation of missing mortality rates associated with FBD as follows.

Correlation analysis (CorA)

First, among a total of 113 predictors, we excluded those having missing values and restricted the analyses to complete cases because CorA requires that values must be present for all predictors. This resulted in 46 predictors with complete values. These variables were subjected to pairwise CorA to identify highly correlated and redundant variables. Those pairs of predictors with a high correlation coefficient, i.e. $r \geqslant 0{\cdot}85$ [18], were identified and one member of them was retained based on biological plausibility.

Elastic net regularization (ENR)

Following CorA, we applied ENR. This method offers a statistically appealing regression approach to select meaningful subsets of predictors of mortality associated with FBDs for the 48 countries with complete mortality data. ENR is a flexible variable selection method proposed by Zou & Hastie, which was developed to overcome the flaws of the commonly used Ordinary Least Squares approach with regard to prediction accuracy [19]. The basic form of the linear regression model used to perform variable selection with ENR is shown in equation (1):

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}, \mathbf{e} \sim N\left(0, \sigma^2\right), \tag{1}$$

where $\mathbf{Y}$ is a vector of log-total mortality rates (response variable), $\mathbf{X}$ is an $n \times p$ matrix of predictors, $\boldsymbol{\beta}$ is p vector of regression coefficients and $\mathbf{e}$ is the vector of residual errors. The ENR uses a mixture of the $\mathbf{l}_1$ [Least Absolute Shrinkage and Selection Operator (LASSO)] and $\mathbf{l}_2$ (ridge regression) penalties, which allows both automatic variable selection and shrinkage, respectively [19]. The ENR has two parameters, α and λ. We set α at 0·5 (1 = LASSO and 0 = ridge, 0·5 being for ENR) and performed cross-validation to find the optimal value of regularization parameter λ. The optimal λ value was applied during variable selection. The *glmnet* package in R was used to perform

https://doi.org/10.1017/S0950268815003234 Published online by Cambridge University Press

Table 1. *Description and percent missing of the eight selected variables for predicting log-total mortality associated with foodborne diseases*

Variable name	Number (%) missing[a]	Description (source)
Life expectancy	10 (5·1)	Life expectancy at birth, total (years). Life expectancy at birth indicates the number of years a newborn infant would live if prevailing patterns of mortality at the time of its birth were to stay the same throughout its life (http://data.worldbank.org/indicator/SP.DYN.LE00.IN)
Animalcalpercap	23 (11·8)	Average calorie supply from animal products – *per capita* (http://faostat3.fao.org/search/*/E)
Birthperadolescent	15 (7·7)	Adolescent fertility rate: the number of births per 1000 women ages 15–19 years (http://data.worldbank.org/indicator/SP.ADO.TFRT)
Pctareableland	6 (3·0)	Percent arable land (http://data.worldbank.org/indicator/AG.LND.ARBL.ZS)
Fertilityrate	10 (5·1)	Total fertility rate represents the number of children that would be born to a woman if she were to live to the end of her childbearing years and bear children in accordance with current age-specific fertility rates (http://data.worldbank.org/indicator/SP.DYN.TFRT.IN)
Maternaldeathrisk	15 (7·7)	Maternal mortality ratio (national estimate, per 100 000 live births) (http://data.worldbank.org/indicator/SH.STA.MMRT.NE)
Laborfemmale	18 (9·2)	Labuor force participation rate for females aged 15–24 years: the proportion of the population aged 15–24 years that is economically active (http://data.worldbank.org/indicator/SL.TLF.ACTI.1524.FE.NE.ZS)
Kcalperday	26 (12·0)	Calorie supply *per capita* per day (http://faostat.fao.org/)

[a] Number (%) of countries missing the variable.

the ENR procedure [20]. A detailed description of regression-based ENR as a data-mining technique can be found in the literature [19, 20].

Imputation of missing values

After we selected variables using ENR, we searched values of these variables for the remaining 146 countries from publicly available and validated databases (FAO, World Bank) [13, 14]. Whenever multiple values for a given country were available, we took the value for the year closest to 2005. Since not all countries may have full information for the selected predictors, Multiple Imputation (MI) was performed to fill-in missing predictor values using the MICE (Multiple Imputation using Chained Equations) package in R [21]. The percentage of countries missing the variables is indicated in column two of Table 1.

MI helps to handle missing data, where missing values are replaced by random draws from the predictive distribution of the missing data given the observed data [21, 22]. The procedure generates m numbers of complete datasets (also called multiply imputed datasets) ready for further analysis. The optimum number of m varies across studies and may depend on the study design and the proportion of values missing. Prior work suggests that 5–10 multiply imputed datasets are minimally sufficient for generating valid variable estimates [23]. We used 20 multiply imputed datasets in this study. The imputed values were averaged across the number of multiply imputed datasets to provide values for a given missing value. Convergence of the imputation process was assessed by visually examining density plots of each variable to evaluate the plausibility of imputed values across the number of iterations.

Cluster analysis (CA)

Following the imputation step, we carried out CA. The purpose of CA is to aggregate countries into groups based on similar values of predictors such that countries within a cluster have homogenous mortality rates. Although several types of clustering methods exist, we compared four commonly used hierarchical clustering methods to identify the optimal clustering solution for our dataset: *single linkage, complete linkage, UPGMA* (unweighted pair group mean average) and *Ward's minimum variance* methods [24]. We used visual examination of the resulting dendrograms, Gower's distance [25] and cophenetic correlation to select the method of choice [26]. Based on established rules, smaller

https://doi.org/10.1017/S0950268815003234 Published online by Cambridge University Press

Table 2. *Logistic regression of missing indicator (1 = missing log-total mortality, 0 = observed log-total mortality) on predictors of mortality associated with foodborne diseases to test the plausibility of the Missing At Random assumption*

Coefficients	Estimate	Pr(>|z|)
(Intercept)	21·14	0·002
Life expectancy, years	0·18	0·001
Per capita calorie supply from animal origin[a]	−0·16	0·76
Birth per adolescent[a]	−0·89	0·01
Percent of arable land	−0·01	0·49
Fertility rate[a]	1·25	0·20
Maternal mortality ratio[a]	0·05	0·87
Female labour[b]	−0·02	0·06
Kilocalorie per day *per capita*[a]	−0·57	0·71

[a] Log-transformed variables.
[b] Labour force participation rate for females aged 15–24 years.

Gower's dissimilarity coefficients and larger cophenetic correlations indicate that the preferred clustering solution best fits the data. We selected UPGMA as the clustering method of choice for our data. Thereafter, we determined the optimal number of clusters using the gap-statistic, which is one of the most popular methods for estimating the number of clusters in a dataset [27]. In addition, we evaluated the average silhouette width (SW) which is a composite index reflecting the compactness and separation of clusters. A high SW indicates that the clusters are homogenous. A detailed technical description of these methods can be found elsewhere [25–28].

Bayesian hierarchical regression

After having developed a dataset for all 194 countries, we fit a Bayesian hierarchical regression model (BHM) for predicting log-total mortality rate associated with FBDs. We incorporated the clusters obtained from the CA as random effects into our BHM. The regression model fit to the data was formulated as follows (2):

$$\left.\begin{array}{l} Y_i = N\left(\alpha_j[i] + \beta_k X_{ki}, {\sigma_y}^2\right), \text{ for } i = 1, ..., n; \\ k = 1, 2, ..., K, \alpha_j = N\left(\mu_\alpha, {\sigma_\alpha}^2\right), \text{ for } j = 1, ..., J, \end{array}\right\} \quad (2)$$

where Y_i denotes the response variable (log-total mortality); α and β are the intercept and the regression coefficients, respectively; n is the total number of countries; X_{ki} denotes the predictors ($K = 8$); and J is the number of clusters. The variance (σ^2), α's and β's are parameters estimated from the data. In addition to the model constructed using our four cluster solution, we evaluated a new model incorporating the WHO's Global Environmental Monitoring System/Food (GEMS) cluster for comparing the results. The GEMS/Food cluster categorizes the WHO countries into 17 groups based on food consumption and dietary intake of various chemicals [29]. A non-hierarchical Bayesian framework was also fit to the data, which does not take into account any clustering of the data.

Valid inference from the above model assumes that the missingness in the system is Missing At Random (MAR). Missing data is considered MAR whenever the missingness can be explained by one or more predictors in the dataset. Although it is not possible to directly test the MAR assumption based on the data alone, MAR can be demonstrated by showing association between predictors and missingness of the response variable [30]. For testing the validity of the assumption, we created a dummy variable for whether mortality rate is missing or not, and ran a logistic regression to statistically test if any of the variables are associated with missingness (Table 2). A strong statistical association indicates that the MAR assumption is valid. A detailed description of missing data mechanisms can be found in the literature [31].

Some of the predictors in our dataset were not normally distributed, and therefore, we log-transformed them to stabilize their distribution before applying the regression approach. This subset of predictors were '*Per capita* animal calorie consumption', 'Birth per adolescent', 'Fertility rate', 'Maternal death risk' and 'Kilocalories per day'.

In a Bayesian framework all the parameters in the model must have a prior distribution, which is a way

https://doi.org/10.1017/S0950268815003234 Published online by Cambridge University Press

Table 3. *Goodness-of-fit and mean absolute errors (MAE) of three Bayesian hierarchical and non-hierarchical models for predicting foodborne disease mortality rates*

Model	Model fit		MAE[e] (95% CI)[g]	MAE[f] (95% CI)
	DIC[c]	pD[d]		
Non-hierarchical	123·9	11·76	0·53 (0·43–0·69)	0·65 (0·53–0·82)
Hierarchical A[a]	123·7	12·70	0·52 (0·42–0·69)	0·66 (053–0·83)
Hierarchical B[b]	126·1	15·04	0·51 (0·41–0·67)	0·66 (0·53–0·84)

[a] Four cluster random effect; [b] GEMS cluster random effect; [c] Deviance Information Criteria; [d] Effective number of parameters (measure of model complexity); [e] mean absolute error obtained from the fitted model; [f] MAE obtained from the model after 'Leave One Out' cross-validation; [g] Bootstrap 95% confidence interval.

of quantifying lack of knowledge about the parameters [32]. We assigned all the coefficients to have non-informative prior distribution (i.e. a normal distribution with mean = 0 and a precision = 0·01). This implies that the magnitudes of the regression coefficients are expected to lie between −10 and 10. The prior for the precisions, i.e. the inverse variances, were defined in terms of the standard deviation parameters and given uniform prior distributions on the range (0, 10). (The R-JAGS code for the Bayesian framework applied in this project is provided in Supplementary Appendix A.)

We ran the model for 50 000 iterations with a burn-in of 5000 (i.e. we discarded the first 5000 iterations). We assessed convergence by running two chains of dispersed initial values, and then by observing autocorrelation and density plots of the parameters from the models' outputs. Whenever more than one model was to be evaluated for fit, we used Deviance Information Criteria (DIC) and the effective number of parameters (pD) as model fit comparison tools [33]. The DIC is a Bayesian alternative of Akaike's Information Criteria for comparing competing models, and the pD is a measure of the complexity of the model [33]. A difference in DIC of more than 5–10 units is regarded as strong evidence in favour of the model with smaller DIC [34].

Model validation

In order to assess the predictive performance of our model, we carried out cross-validation using part of the dataset with complete information on mortality rates. We implemented the leave-one-out cross-validation method (LOOCV) used to estimate the generalizability of a model in the absence of external data [35]. This method takes one observation out of the data, sets it aside as a 'testing set', and fits the model using the remaining data, called the 'training set' to assess statistical predictions of the model. The resulting coefficients are then applied to the 'test set', to generate predicted values that are compared to the observed value of that single case. This procedure is performed repeatedly for all observations of the data and the mean absolute error (MAE) of prediction is calculated [equation (3)] and compared to the baseline MAE (i.e. the MAE computed without cross-validation). This comparison allows assessment of 'out-of-sample' predictive performance of the model whenever no external data exist [36]. The MAE is mathematically expressed as follows:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} |f_i - y_i|, \tag{3}$$

where n = total number of test sets, f_i = predicted log-total mortality and y_i = observed log-total mortality. Ninety-five percent confidence intervals (95% CIs) for the MAEs were computed using a non-parametric bootstrapping method with 2000 replications from the *boot.ci* procedure in R. A small MAE indicates a better out-of-sample prediction of the model.

Sensitivity analysis

We assessed robustness of the results by changing the priors. We also examined the model outputs after specifying priors separately for each cluster. Sensitivity analysis regarding priors was conducted by first assigning uninformative priors to means and inverse variances, and then changing the priors of the precision by a factor of 10 as shown in Table 4. Additionally we investigated the stability of predictions by randomly deleting mortality rates and refitting the model, and

https://doi.org/10.1017/S0950268815003234 Published online by Cambridge University Press

Table 4. *Prior parameter values employed on the Bayesian hierarchical model for sensitivity analysis*

Priors	Variances	Means	
Set 1	$1/\sigma^2 \sim \text{dgamma}(10^{-3}, 10^{-3})$	$b0 \sim \text{dnorm}(0, 10^{-3})$	$b \sim \text{dnorm}(0, 10^{-3})$
Set 2	$1/\sigma^2 \sim \text{dgamma}(10^{-2}, 10^{-2})$	$b0 \sim \text{dnorm}(0, 10^{-2})$	$b \sim \text{dnorm}(0, 10^{-2})$
Set 3	$1/\sigma^2 \sim \text{dgamma}(10^{-1}, 10^{-1})$	$b0 \sim \text{dnorm}(0, 10^{-1})$	$b \sim \text{dnorm}(0, 10^{-1})$

$1/\sigma^2$, precision (inverse of the variance). In the JAGS model, priors for variances are specified by precision. Gamma distribution is frequently used to specify priors for precision. b0, Average log-total mortality rate; b, priors for regression coefficients.

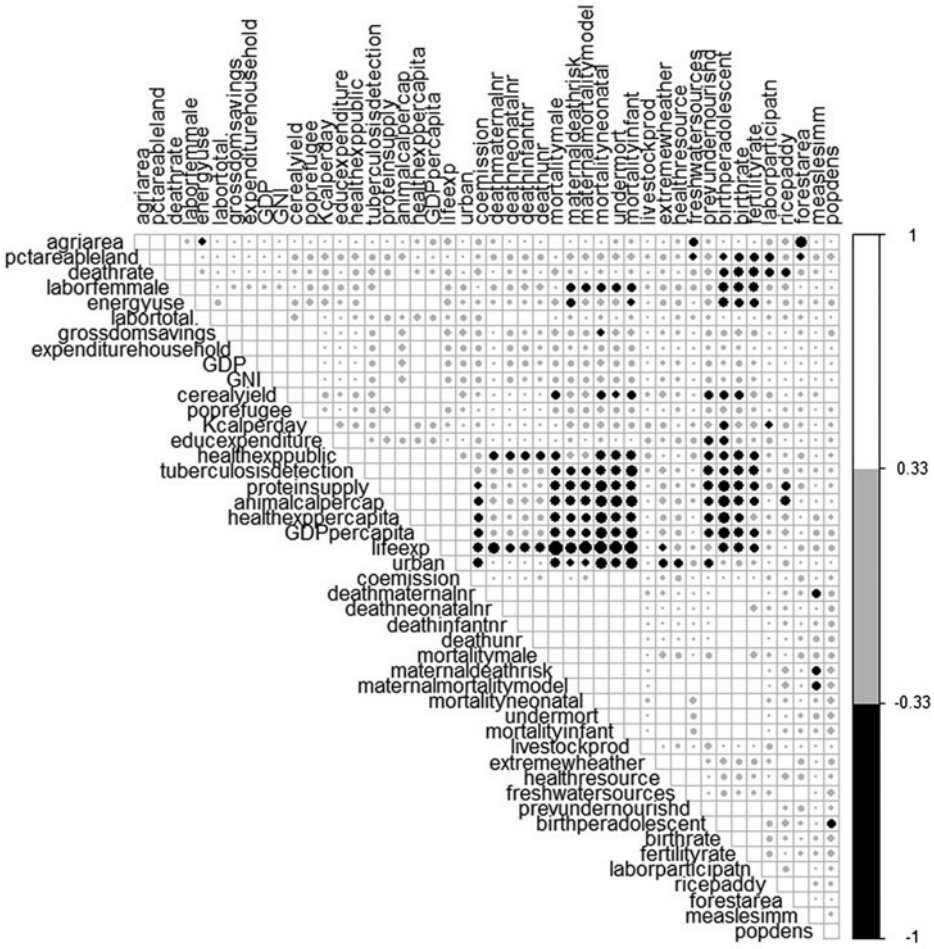

Fig. 1. The correlation matrix of 46 potential predictors of mortality associated with foodborne diseases. The values on the right side of the plot represent correlation coefficients between a pair of variables.

also by randomly adding plausible hypothetical mortality rates for a subset of countries missing the data and refitting the model.

RESULTS

Correlation analysis

Out of 46 predictors screened by means of pairwise CorA, we found 20 of them to be highly correlated, of which we dropped 14 predictors that were redundant. This results in 32 final predictors for further analysis (see the correlation matrix in Fig. 1).

Elastic net regularization

The remaining 32 variables that were retained after CorA were subjected to ENR. Eight non-zero coefficient variables were selected as the final subset of predictors. We used these variables for CA and regression analysis, as indicated in the next sections. Description

https://doi.org/10.1017/S0950268815003234 Published online by Cambridge University Press

of the eight variables and the proportion of missing values for these variables across all 194 countries are indicated in Table 1.

Cluster analysis

The average SW (0·59) and the Mantel optimal cluster methods resulted in two and three clusters, respectively, while the gap statistic suggested four clusters to be optimal. As the gap statistic is the most recommended approach, we decided to partition our dataset into four clusters as shown in the dendrogram (Fig. 2) [27]. One hundred and forty-two countries were grouped into cluster 1, whereas 29, 3, and 20 countries were categorized into clusters 2, 3, and 4, respectively. The majority of the countries in cluster 1 are high- and middle-income countries, whereas many of the countries in the other three clusters are low- or middle-income countries.

Regression and model validation

Model validation and fit

The results of a statistical test to assess the validity of the MAR assumption are shown in Table 2. Three of the eight predictors were significantly associated with missingness in the data indicating that the MAR assumption holds for the current analysis. Table 3 shows the results of goodness-of-fit assessment (DIC and pD) and the MAEs for the three models having different structures. No substantial differences were observed regarding either DIC or MAE in the three models (Table 2). However, the model fitted with the GEMS cluster (hierarchical B in Table 2) has the highest pD due to the large number of clusters in the data. This model also did not converge well even after a higher number of iterations. We selected the BHM because its structure allows it to 'borrow strength' (i.e. to pool information) across clusters. The latter characteristic is very helpful whenever data are lacking within clusters. Assessment of autocorrelation, density plots and trace plots showed that convergence criteria for the BHM were met.

Sensitivity analysis and prediction of mortality rates

Table 4 indicates the range and specifications of priors for the variance and mean components used to evaluate robustness of the BHM. Varying the priors to become more informative did not substantially change the predictions of log-total mortality rates for countries within cluster 1 (Fig. 3*a*) and cluster 4 (Fig. 3*d*). These clusters contain countries with information regarding mortality rates. Conversely, the priors substantially influenced the uncertainty of predictions within the clusters lacking any mortality rate data, i.e. cluster 2 (Fig. 3*b*) and cluster 3 (Fig. 3*c*).

We set the same prior distribution for all clusters (also called exchangeable priors) and specified smaller values for the variance of the means and precision parameters (see prior set 3 in Table 4) for the final prediction. Constraining the parameter values to be within –10 and 10 [e.g. specifying the prior of $b0$ at dnorm (0, 0·1)] was not a serious restriction. Since the model is on a log-scale, it is not possible to see values as extreme as –10 or 10, which correspond to mortality rate of e^{-10} or e^{10} per 100 000 population.

Deleting the observed mortality rate from cluster 4 (i.e. removing the mortality rate value of Guatemala) and refitting the model resulted in a very large uncertainty of predicted log-total mortality rates for all countries in this cluster (Fig. 4*d*). On the other hand, randomly adding plausible values for a subset of countries in cluster 2 (Fig. 4*b*) and cluster 3 (Fig. 4*c*), i.e. clusters that lack any observed mortality rate data, considerably reduced the uncertainty of predictions. This indicates that uncertainty is highest for clusters with little or no information and the predicted mortality rate for countries lacking the data can be substantially improved if mortality rate values are obtained for a few countries in the cluster. The change in predicted log-total mortality was minimal for countries in cluster 1 (Fig. 4*a*) whenever mortality rate values were added or deleted from the other clusters as part of the sensitivity analysis.

Supplementary Appendix B shows the final predicted log-total mortality rates for all countries using the BHM. None of the countries in clusters 2 and 3 had observed mortality rates. For countries within these clusters, the BHM predicted wide 95% CIs for the median of log-total mortality rates indicating large uncertainty. On the other hand, the uncertainty of predictions was very small for countries within clusters 1 and 4. The overall median (95% CI) of the predicted log-total mortality rate ranged from −1·23 (−2·03 to −0·44) for Greece to 5·04 (2·68–7·36) for Afghanistan, which yields median mortality rates of 0·29 (0·13–0·63) and 155·19 (14·66–1572·85) per 100 000 population, respectively. As indicated in Supplementary Appendix B, some of the 95% CIs of predictions did not contain the observed mortality rates.

DISCUSSION

Lack of sufficient and complete data for mortality and morbidity from many countries poses an ongoing

https://doi.org/10.1017/S0950268815003234 Published online by Cambridge University Press

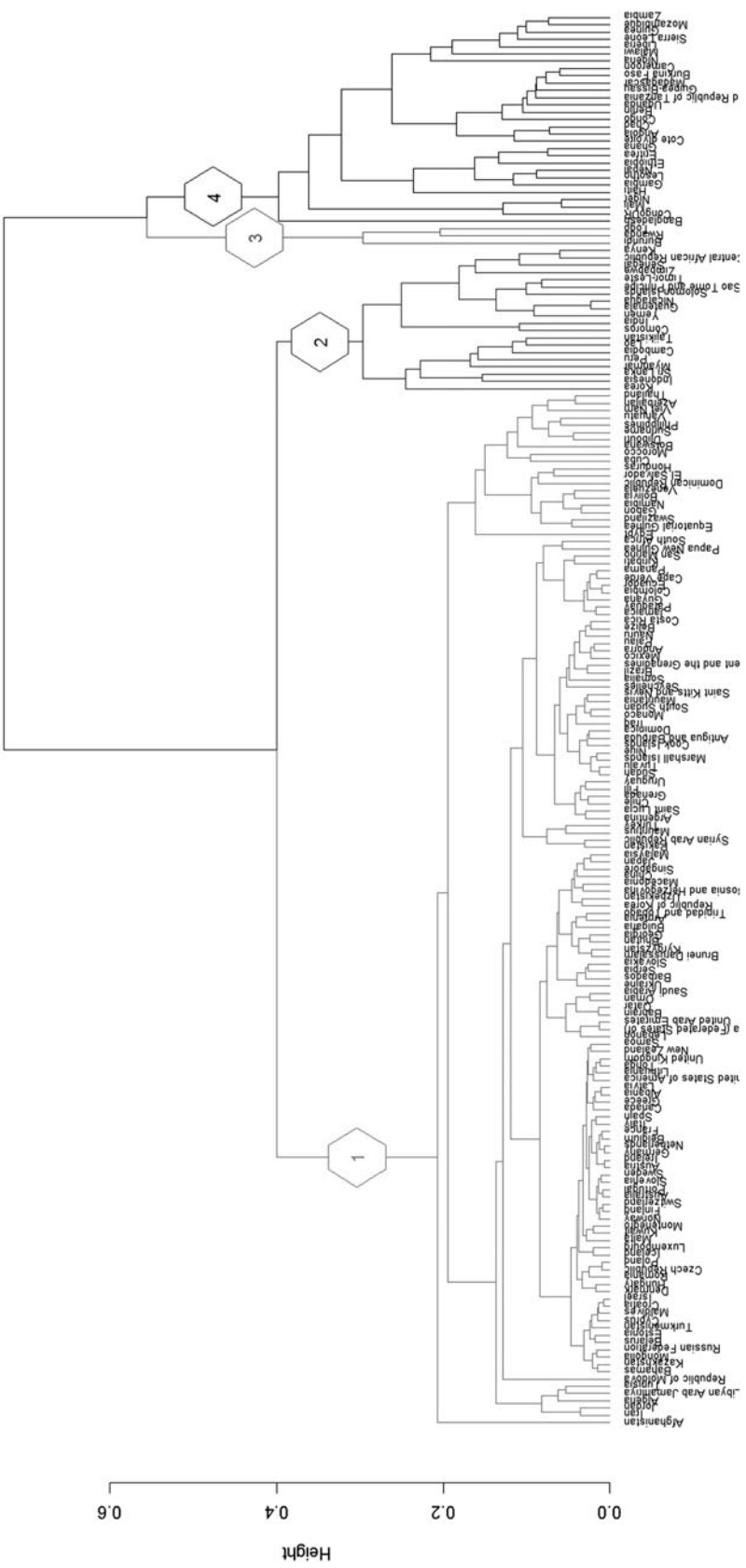

Fig. 2. Hierarchical cluster analysis of all 194 WHO countries using the UPGMA method. Eight predictors of mortality associated with foodborne diseases were used to construct the dendrogram. The numbers shown at the top of the dendrogram indicate the cluster identification.

https://doi.org/10.1017/S0950268815003234 Published online by Cambridge University Press

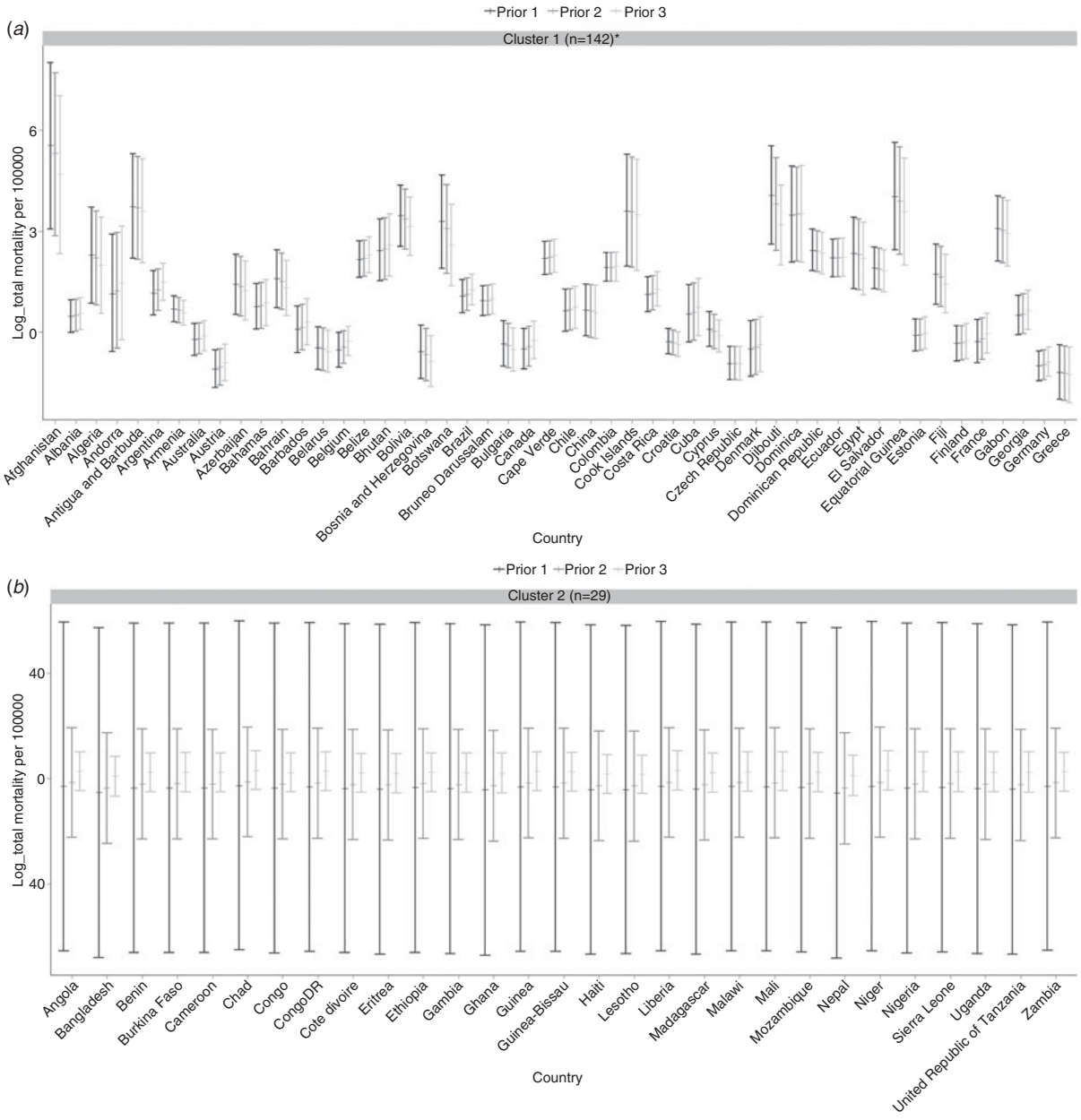

Fig. 3. For legend see next page.

challenge for the estimation of global burden of FBDs [15].

The eight predictors we selected in this report are proxy attributes to capture the socioeconomic, food-production, hygiene, and health status of countries. This is in agreement with a previous study regarding variable selection to estimate the missing incidence of specific FBDs [37]. In addition, a frequentist-based analysis of part of the current dataset highlighted that both health and non-health-related variables can be used as proxy predictors to measure mortality associated with FBDs [12].

Grouping of WHO countries based on the main predictors of FBD mortality was a novel attempt to create data-driven clusters of countries with relatively homogenous mortality rates. This clustered structure will help 'borrow strength' from similar countries while predicting missing FBD mortality rates. The WHO countries have been previously grouped based on geographical attributes (e.g. WHO subregions) or

https://doi.org/10.1017/S0950268815003234 Published online by Cambridge University Press

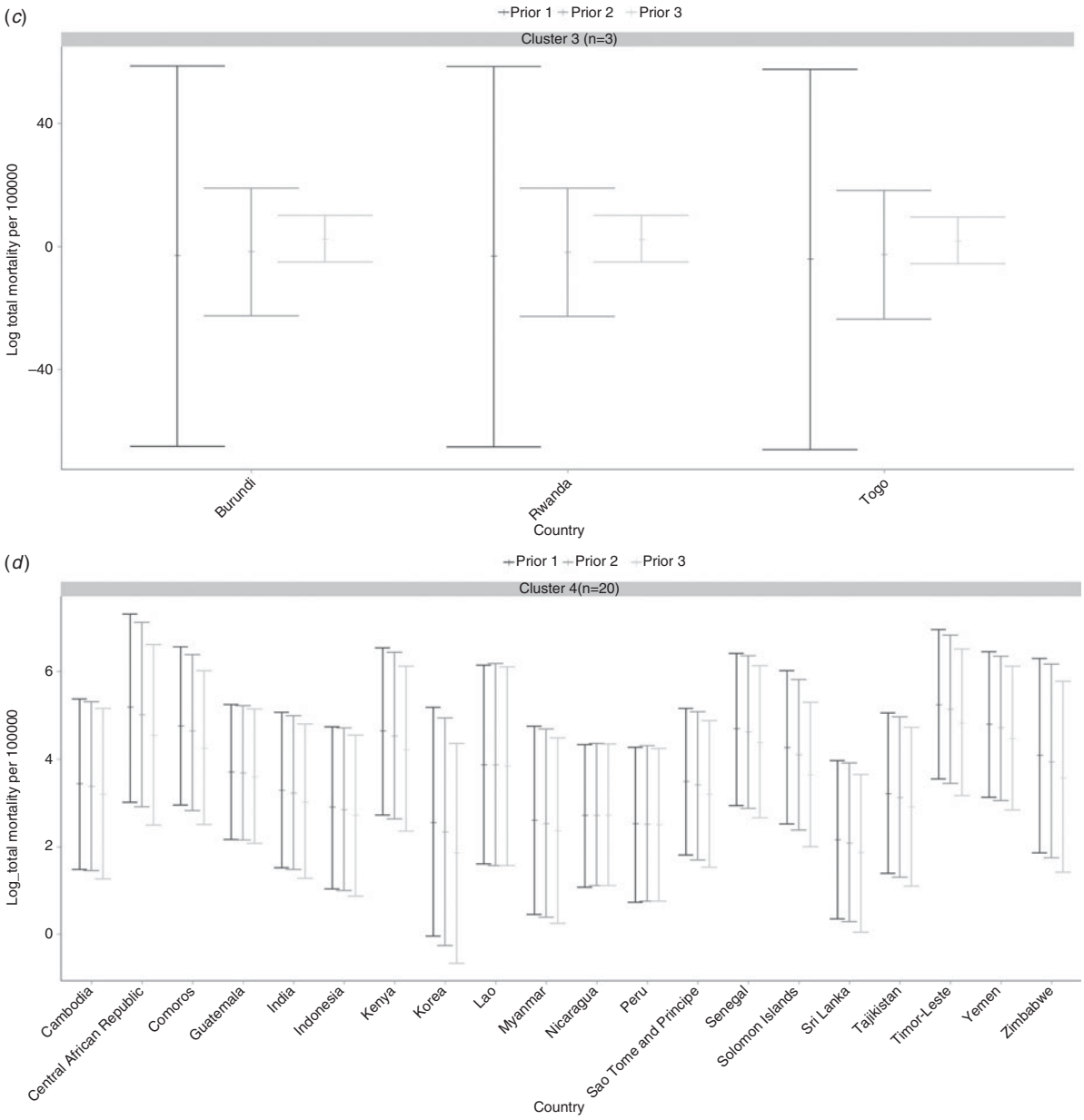

Fig. 3. Sensitivity analyses of the median and 95% credible intervals of log-total mortality predictions of the Bayesian hierarchical model using three priors. Panels (*a*)–(*d*) represent the four clusters. Since the predictions are stable for all countries in (*a*) cluster 1 while using the three priors, only the values of 50 countries are shown for optimal display of this cluster. * The numbers in parentheses indicate the number of countries that belong to each cluster. Prior 1: precision~dgamma(10^{-3}, 10^{-3}), mean~dnorm(0, 10^{-3}); prior 2: precision~dgamma(10^{-2}, 10^{-2}), mean~dnorm(0, 10^{-2}); prior 3: precision~dgamma(10^{-1}, 10^{-1}), mean~dnorm(0, 10^{-1}).

other parameters depending on the goal of the classification scheme. For example, the GEMS Food cluster is designed to group countries based on food consumption and risk assessment [29]. In a recent study to predict missing national-level incidence of specific FBDs, the WHO subregions and the GEMS regions were used as random effects [37]. In our study, comparison of predictions and assessment of model fit using the GEMS cluster *vs.* our data-driven clusters indicated that the model with the GEMS cluster lacks convergence while our four-cluster solution has converged quite well. In our study, 47 of the 48 countries with complete mortality rate data grouped into cluster 1, which indicates that those countries that routinely report mortality rates have similarities in the eight predictor values.

https://doi.org/10.1017/S0950268815003234 Published online by Cambridge University Press

Most missing-data analysis approaches require that the MAR assumption is fulfilled for the results to be valid [38]. Although the MAR assumption, as such, is not testable, we justified its validity for our dataset by demonstrating a statistical association of the predictors with the missingness of mortality rates. Consequently, three predictors were significantly associated with missingness, implying that the missing data can be partly explained by the predictors in the dataset. A previous simulation study has shown that an erroneous assumption of MAR will often have only a slight impact on the estimates [39]. Meanwhile, it is also important to note that all other missing-data analysis approaches require assumptions that are just as difficult to justify.

In our study, the choice of the best-fit model used for prediction of mortality associated with FBDs was based on comparison of model fit parameters, out-of-sample prediction performance, and the method's suitability for analysis of missing data with hierarchical structure. To evaluate predictive performance of a model, using the same data as were initially used to build the model may introduce over-fitting problems [40]. Furthermore, collecting new data to validate the model for predictive strength is not feasible and, therefore, the LOOCV method solves an over-fi tting problem and helps to assess the predictive accuracy. Although we did not observe a substantial difference between the three models regarding MAEs, we preferred the BHM for a number of reasons. The structure of a BHM enables 'borrowing strength' across clusters which improves prediction of mortality rates. The latter is particularly essential whenever data are missing [41]. Moreover, a BHM facilitates the estimation of several parameters over similar units (e.g. countries within clusters) in order to improve the precision of the estimated effects for each unit. It has also been described previously that a Bayesian approach allows for a more efficient use of data as the method does not depend on the asymptotic theory of large sample approximation [42]. This is essential whenever there are few observations and a high proportion of missing values in the dataset.

Part of the dataset used herein has been analysed previously using a classical frequentist framework [12]. Our Bayesian approach, however, has several important advantages over this likelihood-based frequentist method. A Bayesian approach includes an opportunity to assign pertinent information (prior) to unknown parameters (including missing values distribution) [32]. This is particularly useful for analysis of a dataset with missing values. Second, Bayesian models can be easily updated rationally when new data become available. As such, future research on FBDs can directly utilize the current results as priors. Additionally, BHM provides a convenient setting for a dataset with inherent hierarchical structure. In our dataset, countries within clusters are assumed to have more similar mortality rates compared to the rates between countries across clusters. Furthermore, implementing BHM allows the pooling of information across clusters, such that clusters with little or no data 'borrow strength' of the log-mortality rates from other clusters. In our analyses, the predicted median mortality rates for countries in clusters 2 and 3 were smoothed towards the overall average population estimate (Fig. 4). Although the predicted median mortality rates are close to the overall average log-mortality rate, the uncertainty is large as a direct result of the data quality or lack thereof. The reduction in uncertainty of the predictions was achieved by adding hypothetical but plausible data for a subset of countries and this has a practical implication. For example, data collection strategies for mortality rates can be based on cluster information. If mortality data can be obtained from a proportion of countries from a properly defined cluster, one can use BHM to predict mortality rates for the remaining countries missing the data and thereby make optimal usage of the data available.

A limitation of this study is that the predictors were selected from countries with complete FBD mortality rate data (i.e. from 48 countries that are mostly middle- and high-income countries). At the same time, the clustering of all WHO countries was based on the values of these selected predictors. Since country-level socioeconomic factors may determine the incidence of mortality associated with FBDs, the risk distribution regarding mortality rates due to FBDs may be different between countries. On the other hand, it is possible that similar potential risk factors (predictors) may be shared between developed and developing countries despite the differences in disease burden [2, 43]. The other limitation is that there might be a risk of excluding strong predictors while dropping variables containing missing values to fulfil the statistical requirements of CorA. However, the eight selected predictors are considered to be meaningful predictors for FBD mortality.

The high proportion of missing values in the dataset might be the cause for some of the predictions to be outside the observed range. Therefore, the estimates from the final model in this report should be

https://doi.org/10.1017/S0950268815003234 Published online by Cambridge University Press

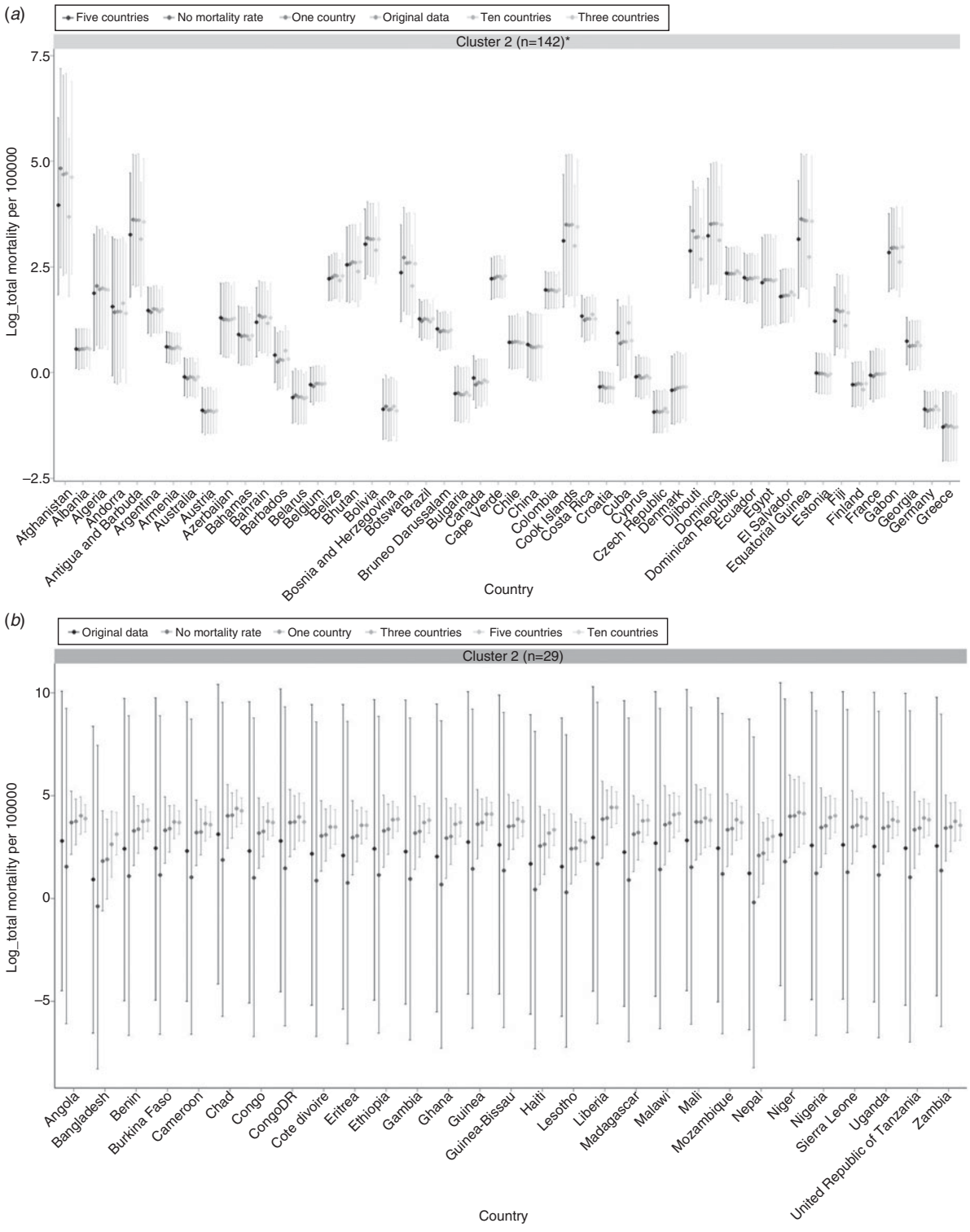

Fig. 4. Comparison of the median and 95% credible intervals of log-total mortality predictions of the Bayesian hierarchical model with regard to deleting or randomly adding mortality rates for a subset of countries. Panels (*a*)–(*d*) represent the four clusters. Since the predictions are stable for all countries in (*a*) cluster 1 regardless of deleting and adding values, only the values of 50 countries out of 142 are shown for optimal display of this cluster. The four panels depict the change in uncertainty of predictions when new information is added or deleted from the dataset. Subsets of countries were randomly selected from (*b*) cluster 2 and (*c*) cluster 3 that lack mortality rate data. Then a log-total mortality

[*Legend continues on next page*

https://doi.org/10.1017/S0950268815003234 Published online by Cambridge University Press

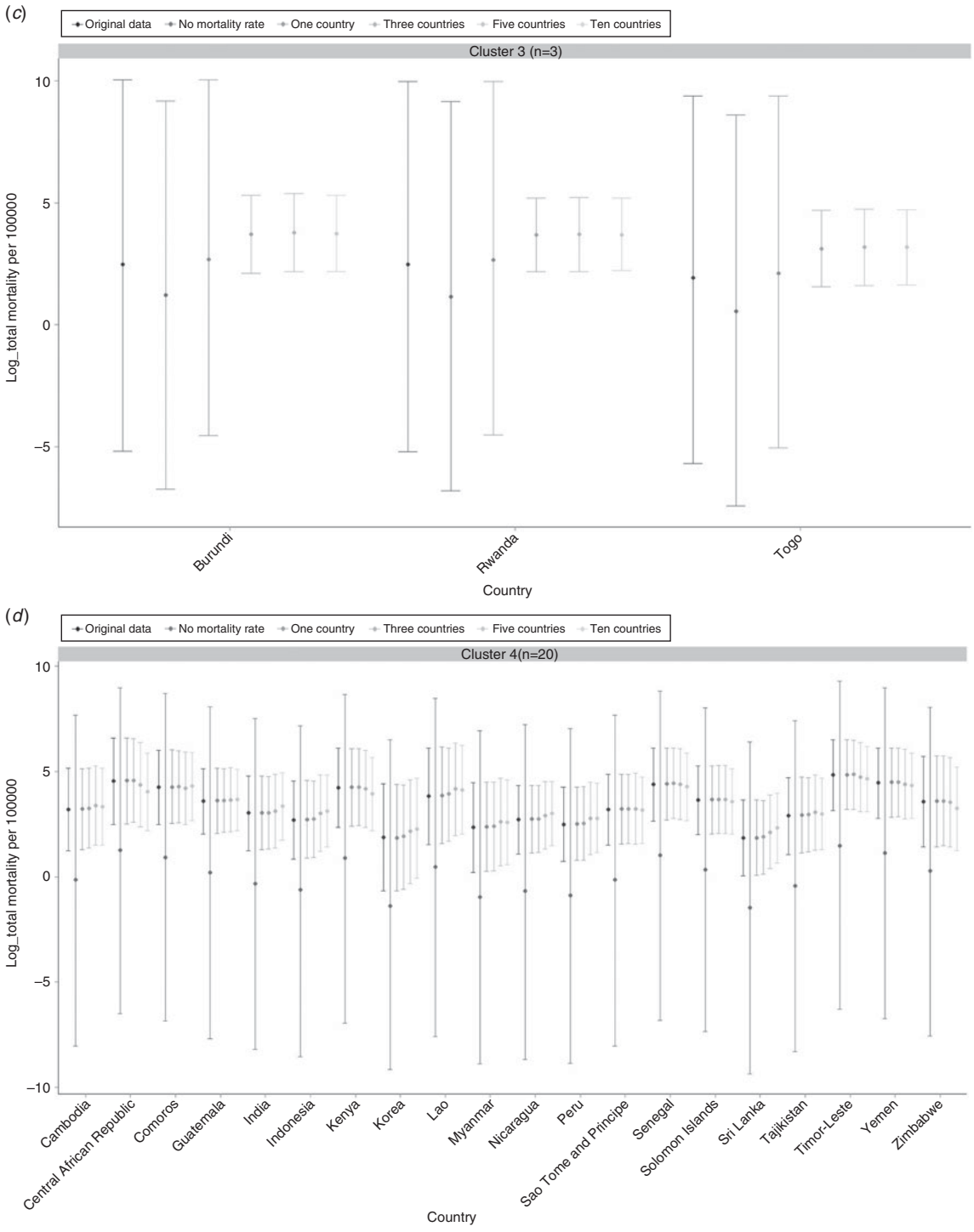

Fig. 4 (*cont.*)**.** rate of 3·75 was assigned to each of them before a model was fitted. The 3·75 value is the observed mortality rate of Guatemala, which has the largest value among all countries with information on mortality rates. Note that each panel in the above figure has different scales on the y-axis to optimally display the 95% credible intervals. Explanation of keys: *Original data*: all observed data included in model predictions. *No mortality rate*: model predictions after deleting the value of Guatemala from cluster 4. *One country*: model predictions after a hypothetical mortality rate was assigned to one country (Angola). *Three countries*: model predictions after a hypothetical mortality rate was assigned to three countries (Angola, Guinea, Rwanda). *Five countries*: model predictions after a hypothetical mortality rate was assigned to five countries (Angola, Guinea, Rwanda, Uganda, Nepal). *Ten countries*: model predictions after a hypothetical mortality rate was assigned to ten countries (Angola, Guinea, Rwanda, Uganda, Nepal, Benin, Bangladesh, Madagascar, Chad, Burkina Faso).

https://doi.org/10.1017/S0950268815003234 Published online by Cambridge University Press

interpreted with caution. The FBD mortality predictions obtained from our suggested method can be improved by obtaining data across regions of diverse social and geographical characteristics of more countries. Such valid additional data would decrease the uncertainty of estimates and possibly change the clustering of the countries. The general approach presented in this paper, however, would not change. Finally, whenever resources are limited, the selected variables can provide suggestions for future data collection regarding risk factors of FBD mortality.

SUPPLEMENTARY MATERIAL

For supplementary material accompanying this paper visit http://dx.doi.org/10.1017/S0950268815003234.

ACKNOWLEDGEMENTS

This work was supported by the NIH Ruth L. Kirschstein National Research Service Award Institutional Training Grant T32 RR023916 and T32 OD010423. The views expressed in this document are solely those of the authors and do not represent the views of the World Health Organization.

DECLARATION OF INTEREST

None.

REFERENCES

1. **Havelaar AH, *et al.*** WHO initiative to estimate the global burden of foodborne diseases. *Lancet* 2013; **381**: S59.
2. **Todd EC.** Epidemiology of foodborne diseases: a worldwide review. *World Health Statistics Quarterly. Rapport Trimestriel De Statistiques Sanitaires Mondiales* 1997; **50**: 30–50.
3. **Käferstein FK.** Actions to reverse the upward curve of foodborne illness. *Food Control* 2003; **14**: 101–109.
4. **WHO.** Fact sheets (http://www.who.int/mediacentre/factsheets/en/). World Health Organization.
5. **Scallan E, *et al.*** Foodborne illness acquired in the United States – major pathogens. *Emerging Infectious Diseases* 2011; **17**: 7–15.
6. **Motarjemi Y, Käferstein FK.** Global estimation of foodborne diseases. *World Health Statistics Quarterly. Rapport Trimestriel De Statistiques Sanitaires Mondiales* 1997; **50**: 5–11.
7. **WHO.** Initiative to estimate the global burden of foodborne diseases (http://www.who.int/foodsafety/foodborne_disease/ferg/en/index.html). World Health Organization.
8. **Kuchenmüller T, *et al.*** Estimating the global burden of foodborne diseases – a collaborative effort. *Eurosurveillance* 2009; **14**.
9. **Majowicz SE, *et al.*** The global burden of nontyphoidal *Salmonella* gastroenteritis. *Clinical Infectious Disease* 2010; **50**: 882–889.
10. **Crump JA, Luby SP, Mintz ED.** The global burden of typhoid fever. *Bulletin of the World Health Organization* 2004; **82**: 346–353.
11. **ICD-10.** International Statistical Classification of Diseases and health related problems. World Health Organization, 2004, pp. 824.
12. **Hanson LA, *et al.*** Estimating global mortality from potentially foodborne diseases: an analysis using vital registration data. *Population Health Metrics* 2012; **10**: 5.
13. **Food and Agriculture Organization of the United Nations Statistics Division (FDASTAT).** 2015 (http://faostat3.fao.org/home/E). Accessed 30 October 2015.
14. **The World Bank.** Data 2015 (http://data.worldbank.org/). Accessed 30 October 2015.
15. **Jahan S.** Epidemiology of foodborne illness. In: Valdez B, ed. *Scientific, Health and Social Aspects of the Food Industry. InTech*, 2012, pp. 321–342. doi:10.5772/31038.
16. **R Core Team.** R: A language and environment for statistical computing. R Foundation for statistical computing, Vienna, Austria, 2014 (http://www.r-project.org/).
17. **Plummer M, *et al.*** JAGS: a program for analysis of Bayesian graphical models using Gibbs sampling. *Proceedings of the Third International Workshop on Distributed Statistical Computing 2003*, pp. 20–22 (http://www.ci.tuwien.ac.at/Conferences/DSC-2003/Drafts/Plummer.pdf).
18. **Taylor R.** Interpretation of the correlation coefficient: a basic review. *Journal of Diagnostic Medical Sonography* 1990; **6**: 35–39.
19. **Zou H, Hastie T.** Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society* 2005; **67**: 301–320.
20. **Friedman J, Hastie T, Tibshirani R.** Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* 2010; **33**: 1.
21. **Buuren S van.** *Flexible Imputation of Missing Data*. Boca Raton, FL: CRC Press, 2012.
22. **Schafer JL.** Multiple imputation: a primer. *Statistical Methods in Medical Research*. 1999; **8**: 3–15.
23. **Graham JW, Olchowski AE, Gilreath TD.** How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention Science* 2007; **8**: 206–213.
24. **Borcard D, Gillet F, Legendre P.** Numerical ecology with R. Springer; 2011, pp 315.
25. **Gower JC.** A general coefficient of similarity and some of its properties. *Biometrics* 1971; **27**: 857–871.
26. **Saraçli S, Doğan N, Doğan İ.** Comparison of hierarchical cluster analysis methods by Cophenetic correlation. *Journal of Inequalities and Application* 2013; **2013**: 1–8.
27. **Tibshirani R, Walther G, Hastie T.** Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society* 2001; **63**: 411–423.
28. **Rousseeuw PJ.** Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* 1987; **20**: 53–65.

29. **Global Environmental Monitoring System (GEMS).** Food: Report of the WHO working group on collection of food consumption data (COFOCO), 2012 (http://apps.who.int/iris/bitstream/10665/75228/1/9789241504119_eng.pdf).
30. **Mason A, *et al.*** Strategy for modelling non-random missing data mechanisms in observational studies using Bayesian methods. 2010 (http://eprints.ncrm.ac.uk/1776/).
31. **Rubin DB.** Inference and missing data. *Biometrika* 1976; **63**: 581–592.
32. **Gelman A, Hill J.** *Data Analysis Using Regression and Multilevel/Hierarchical Models*, 1st edn. Cambridge University Press, 2006, pp. 648.
33. **Spiegelhalter DJ, Best NG, Carlin BP.** Bayesian deviance, the effective number of parameters, and the comparison of arbitrarily complex models. Division of Biostatistics, University of Minnesota, 1998. Report No.: TR 98-009.
34. **Spiegelhalter DJ, *et al.*** Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society* 2002; **64**: 583–639.
35. **Stone M.** Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society* 1974; **36**: 111–147.
36. **Willmott CJ, Matsuura K.** Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research* 2005; **30**: 79–82.
37. **McDonald SA, *et al.*** Data-driven methods for imputing national-level incidence in global burden of disease studies. *Bulletin of the World Health Organization* 2015; **93**: 228–236.
38. **Little RJA, Rubin DB.** *Statistical Analysis with Missing Data*, 2nd edn. Hoboken, NJ: Wiley, 2002, pp 381.
39. **Collins LM, Schafer JL, Kam CM.** A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*. 2001; **6**: 330–351.
40. **Arlot S, Celisse A.** A survey of cross-validation procedures for model selection. *Statistics Surveys* 2010; **4**: 40–79.
41. **Hong H, *et al.*** A Bayesian missing data framework for generalized multiple outcome mixed treatment comparisons. *Research Synthesis Methods* (in press).
42. **Congdon PP.** *Bayesian statistical modelling*. John Wiley & Sons; 2007, pp. 598.
43. **Käferstein FK, Motarjemi Y, Bettcher DW.** Foodborne disease control: a transnational challenge. *Emerging Infectious Diseases* 1997; **3**: 503–510.