A Bayesian hierarchical model with integrated covariate selection and misclassification matrices to estimate neonatal and child causes of death

Amy R. Mulick

Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London, UK.

E-mail: Amy.Mulick@lshtm.ac.uk

Shefali Oza

Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London, UK.

David Prieto-Merino

Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London, UK.

Francisco Villavicencio

Department of International Health, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD, USA.

Simon Cousens

Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London, UK.

Jamie Perin

Department of International Health, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD, USA.

Summary. Reducing neonatal and child mortality is a global priority. In countries without comprehensive vital registration data to inform policy and planning, statistical modelling is used to estimate the distribution of key causes of death. This modelling presents challenges given that the input data are few, noisy, often not nationally representative of the country from which they are derived, and often do not report separately on all of the key causes. As more nationally representative data come to be available, it becomes possible to produce country estimates that go beyond fixed-effects models with national-level covariates by incorporating country-level random effects. However, the existing frequentist multinomial model is limited by convergence problems when adding random effects, and had not incorporated a covariate selection procedure simultaneously over all causes. We report here on the translation of a fixed effects, frequentist model into a Bayesian framework to address these problems, incorporating a misclassification matrix with the potential to correct for mis-reported as well as unreported causes. We apply the new method and compare the model parameters and predicted distributions of eight key causes of death with those based on the previous, frequentist model.

Keywords: Bayesian hierarchical model, Cause of death, Burden of disease, Neonatal, LASSO, Outcome misclassification

Introduction

Reducing child (under-five) mortality has been a priority for individual countries and the broader international community for decades, but was given added impetus by the Millennium Development Goals (MDGs) established by the United Nations (UN) in 2000. Although dramatic reductions occurred over the MDG period until 2015, an estimated 5.3 million deaths still occurred worldwide in children under five in 2019 (UNICEF et al., 2019). Neonatal deaths (those in the first 28 days of life) account for 47% of under-5 child deaths (UNICEF et al., 2019), and have reduced more slowly than those in the 1–59 month age group (Hug et al., 2019). The Sustainable Development Goals (SDGs; 2015–2030), which followed on from the MDGs, include targets of 25 or fewer under-5 deaths and 12 or fewer neonatal deaths per 1,000 live births for all countries by 2030 (United Nations, 2017).

Understanding the cause-of-death (COD) distribution of neonatal and child deaths is important for selecting appropriate interventions to reduce mortality. Ideally, this information would be available through regularly updated, high-quality data collection systems that are organized at the national level but operate at the local level. At present, only around 70 countries have high-quality vital registration (VR) systems that regularly collect and collate causes of death (World Health Organization, 2017), and the majority of these are high-income countries with low mortality rates. Thus, statistical modelling remains an important tool for estimating the distribution of causes that lead to death for all age groups, including neonates.

Nationally comparable estimates of the neonatal and 1–59 month COD distributions have been produced for 190+ countries since 2005 by the Child Health Epidemiology Reference Group (CHERG) and Maternal Child Epidemiology Estimation (MCEE) group in collaboration with the World Health Organization (WHO). For countries without high-quality VR data, these distributions have been estimated using regression models within a frequentist framework (Liu et al., 2016). This existing approach has several strengths but some important limitations, including the approach to covariate selection, the lack of a mechanism to give additional weight to country-specific data for a country's own estimates and the inability to account for misclassification in the reported CODs.

In this paper, we present our work on translating the existing COD estimation method into the Bayesian framework, extending it to include country-specific random effects and handle COD misclassification, and illustrate its application to neonatal deaths. In Sect. 1 we briefly describe the existing frequentist strategy and outline its limitations; in Sect. 2 we propose methods to address these limitations; in Sect. 3 we outline the model building process; in Sect. 4 we present and evaluate our results for neonatal deaths and compare them with results from the previous method; and we outline the strengths, limitations, and implications of this new modelling approach in the Discussion.

1. Existing strategy and limitations

To produce nationally comparable neonatal COD distributions, we classify each of 194 WHO member states (countries) into one of three groups: those with 1) high-quality VR data, 2) inadequate VR data and low child mortality rates, and 3) inadequate VR data and high child mortality rates (World Health Organization, 2018). For countries in the first group, we use their VR data directly to estimate proportional COD distributions. For countries with inadequate VR, we predict

their COD proportions using a "low mortality model"(group 2 countries) or a "high mortality mode" (group 3 countries). In this paper, we focus on the high mortality model as this model is more technically challenging and requires methodological innovations.

1.1. Data inputs

Cause-of-death data

For the high mortality model, we extracted relevant neonatal COD distributions from studies conducted in high mortality settings since 1980. Details of the literature review and included studies are found elsewhere (Oza et al., 2015). Briefly, these studies typically used verbal autopsy (VA) methods to ascertain the cause of death. This involves interviewing relatives about the symptoms experienced by the deceased individual and using this information to assign a cause of death (World Health Organization, 2016). These studies ranged widely in size and specific study methodologies. Where necessary, we re-classified the recorded CODs into eight key neonatal COD categories: 1) complications of preterm birth ("preterm"), 2) intrapartum-related complications ("intrapartum"), 3) congenital disorders ("congenital"), 4) sepsis and other severe infections ("sepsis"), 5) pneumonia, 6) diarrhoea, 7) neonatal tetanus ("tetanus"), and 8) other causes ("other"). Case definitions are detailed elsewhere (Oza et al., 2015).

Covariate data

We considered 14 explanatory variables for inclusion in our model. Nine of these were continuous metrics: under-five mortality rate (U5MR), neonatal mortality rate (NMR), general fertility rate (GFR), low birthweight rate (LBW), proportion of women delivering with a skilled birth attendant (SBA), adult female literacy rate (FLR), proportion of babies protected at birth against tetanus (PAB), diphtheria/pertussis/tetanus vaccine coverage (DPT), and Bacillus Calmette-Guerin vaccine coverage (BCG). The other five covariates were binary (yes/no) and relate to individual studies: whether reported deaths were from the early neonatal period ("per.early"; 0–6 days, reference 0–27 days); whether reported deaths were from the late neonatal period ("per.late"; 7–27 days, reference 0–27 days); whether the study was conducted in Sub-Saharan Africa (SSA); whether the study was conducted in South Asia (SA); and whether the study distinguished between prematurity and low birth weight ("premvslbw").

1.2. Frequentist modelling approach

We selected "intrapartum" as our "baseline" cause for a multinomial model as all studies reported deaths due to intrapartum-related complications and they represent a relatively high proportion of deaths in high mortality settings. Our frequentist COD modelling approach followed three steps: 1) select covariates using logistic regression for each (non-baseline) COD equation; 2) with selected covariates, build a multinomial regression model for all causes simultaneously and obtain estimated model coefficients; and 3) apply the estimated model coefficients to national-level covariates to produce country-specific COD distributions. The outcome for each regression equation in step 1 was the log of the ratio of deaths attributed to the

given cause relative to deaths attributed to the baseline cause, and we used out-of-sample goodness of fit (GOF) under a jackknife (leave-one-out) procedure to select covariates for each equation.

Not all studies reported CODs in the eight key categories we model. To account for unreported causes, we re-wrote the multinomial likelihood function based on assumptions about which cause category deaths from an unreported cause would have been assigned. For example, if preterm, congenital, or sepsis were unreported, we assumed deaths from these to be in the "other" category. If pneumonia, diarrhoea, or tetanus were unreported, these were assumed to have been included in the sepsis/severe infection category.

Detailed methods have been described by Oza et al. (2015) and are summarised in **Online Supplement Text E1**.

1.3. Limitations of this modelling approach

These multinomial models have been used for fifteen years by the CHERG-MCEE-IMPROVE team, with various minor extensions and modifications over time, to produce neonatal COD estimates for the UN. However, some key issues led us to investigate further improvements and alternative modelling approaches.

First, our current modelling strategy does not give additional weight to input data from a given country for that country's modelled estimates. Therefore, empirical data from a particular country do not influence that country's modelled COD proportions any more than data from other countries. Previously, almost all the studies in our input database were small and not nationally representative, and it was not obvious that the national-level estimates for a country should be particularly influenced by data points from small non-national studies. However, an increasing number of countries now have data from nationally representative VA studies, and efforts are underway to increase the number of such studies (COMSA, 2020).

Second, the current covariate selection approach is an efficient method to search over a large space of covariate combinations. However, two potential limitations with this method are that 1) it does not evaluate all possible covariate combinations and 2) we select covariates for a multinomial model using binomial models. We used individual binomial equations for covariate selection because a similar multinomial approach would be computationally prohibitive.

Finally, apart from the out-of-sample approach to covariate selection, there are no other stability-enhancing components in the modelling process to minimize the impact of noisy data and the risk of overfitting. The input data in our models contains substantial noise due to measurement error in both the outcome (COD) and covariate data, which can compromise model stability. Misclassification of CODs could arise from, e.g., recording errors, differing case definitions and causal hierarchies across studies, or poor interviewee recall in VAs; covariate measurement error can arise from imprecise measurements of difficult-to-measure metrics.

2. Proposed new methods

Various statistical methods are available which can address each of the above limitations, but implementing them within the existing classical (frequentist) framework proved challenging.

A mixed effects (ME) model with random country-specific intercepts is a way to give more weight to country-specific empirical data. These models are based on

a hierarchical structure that assumes that some parameters do not vary (i.e. the fixed effects [FE] component) while others are treated as random variables (i.e. the random effects [RE] component) (Snijders, 2005).

Further, regularization techniques are a promising set of methods to simultaneously address the covariate selection and stability issues, by placing a penalty on model complexity. Most of these methods focus on two ways in which instability arises in the context of out-of-sample predictions: 1) increasing the number of covariates increases model complexity and therefore the risk of overfitting the model to the data; and 2) large coefficient values can increase instability. A subset of regularization methods exist that enable covariate selection within a multinomial framework without being computationally prohibitive. Least Absolute Shrinkage and Selection Operator (LASSO) and ridge regressions are examples of such regularization methods (James et al., 2013). Their general approach is based on including all covariates in the model and maximizing the log-likelihood minus a penalization/regularization term, which is a function of the model coefficients. This results in coefficient values with reduced magnitude. Increasing the penalization term in the LASSO regression pushes some covariates to zero (or very close to zero), hence performing a type of covariate selection within the multinomial model itself. We attempted to implement the LASSO regression within our frequentist multinomial modelling framework by adding a penalty term to the likelihood function. The implementation appeared to work in terms of shrinking covariate coefficients towards zero as the LASSO penalty increased in value. However, we consistently ran into convergence problems.

To address these implementation challenges, we propose shifting our multinomial logistic regression model from a frequentist framework implemented in Stata (www.stata.com) to a Bayesian framework in R (R Core Team, 2020), incorporating both country-level random effects and the Bayesian LASSO for covariate selection.

2.1. Shifting to a Bayesian framework

The Bayesian framework is well suited to address the challenges discussed above. First, while mixed-effect models for multinomial logistic regression have been developed for the classical framework, their flexibility is limited (Hedeker, 2003). A similar model for multinomial data has been developed from the Bayesian perspective (Albert and Chib, 1993) and has been widely used (Burda et al., 2008; Jostins and McVean, 2016) and adapted for specific scenarios such as data sparsity (Cawley et al., 2007) and high dimensions (Yau et al., 2003). Moreover, adding in random effects and implementing the LASSO are both straightforward in the Bayesian framework through the use of priors and Markov chain Monte Carlo (MCMC) sampling from the posterior distributions.

We addressed two key methodological issues in order to implement the Bayesian neonatal COD models. First, we implement our method of dealing with unreported CODs (see Sect. 1) in a more flexible framework, so that (as before) all studies can be included in the multinomial model; we extend this further by building in the capability to deal with true COD misclassification that could arise as a result of genuine recording errors. Second, we implement the Bayesian LASSO with selection of the penalty coefficient λ. Approaches exist to select λ during model estimation (Park and Casella, 2008), but since our model is designed to make out-of-sample predictions we require an alternative approach that maximises out-of-sample GOF.

2.2. Derivation of Bayesian multinomial model with random effects

Our proposed statistical model has several components including the basic multinomial model, the misclassification matrix, and the LASSO penalization term. These are described below.

2.2.1. Basic model

Suppose there exist C mutually exclusive causes of death, and that we have a sample of N_s deaths from a given study s, each of which is (correctly) classified into one and only one of the C categories. If we denote the distribution of true CODs in the sample (i.e. our eight key causes) as $T_{1,s}, T_{2,s}, \ldots, T_{C,s}$ and if the sample is random, we can assume that these observations come from a multinomial distribution,

$$\begin{bmatrix} T_{1,s} \\ T_{2,s} \\ \cdots \\ T_{C,s} \end{bmatrix} \sim \text{Multinomial}\left(N_s, \begin{bmatrix} P_{1,s} \\ P_{2,s} \\ \cdots \\ P_{C,s} \end{bmatrix} \right) ,$$

where $P_{c,s}$ represents the probability that a death is due to cause c in the population in which study s is conducted. This can be re-written as

$$\boldsymbol{T_s} \sim \text{Multinomial}\big(N_s, \boldsymbol{P_s}\big) . \tag{1}$$

Because the C causes are mutually exclusive, it follows that $\sum_{c=1}^{C} P_{c,s} = 1$.

2.2.2. Misclassification matrix

Non-reporting or misclassification of CODs can occur in studies as described previously and, to make matters worse, patterns of misclassification may differ across studies. However, we can deal with this, without knowing the true COD, if we know the pattern of misclassification in each study, i.e. the probability that a death due to cause A is actually recorded as being due to cause B. For each study, there is a specific misclassification matrix with the general form

$$\boldsymbol{M_s} := \begin{bmatrix} M^s_{1,1} & M^s_{1,2} & \cdots & M^s_{1,C} \\ M^s_{2,1} & M^s_{2,2} & \cdots & M^s_{2,C} \\ \cdots & \cdots & \cdots & \cdots \\ M^s_{Ds,1} & M^s_{Ds,2} & \cdots & M^s_{Ds,C} \end{bmatrix} , \tag{2}$$

where $M^s_{d,c}$ is the probability that, in study s, a death from *true* cause c is *recorded* as being due to cause d. Since studies record only one cause per death, each column of this matrix must add up to 1: $\sum_{d=1}^{Ds} M^s_{d,c} = 1$ for all $c = 1, \ldots, C$. The number of different *recorded* causes of death in this study is Ds and this can vary between studies. In fact, studies might also differ in the type of recorded causes of death, not only in the number of these. Some recorded causes might appear only in some studies.

Using (2), for a given study s we can express the recorded COD multinomial probability distribution as $\boldsymbol{M_s} \times \boldsymbol{P_s}$, where $\boldsymbol{M_s}$ is the study-specific misclassification matrix and $\boldsymbol{P_s}$ is the study's true probability distribution. Because $\boldsymbol{M_s}$ is not necessarily invertible, we cannot directly estimate $\boldsymbol{P_s}$ from the probability distribution of recorded causes. Nevertheless, we can still use the distribution of the recorded causes of death along with the misclassification matrix to estimate the model coefficients, as follows.

2.2.3. Proposed model

Suppose the probabilities $P_{c,s}$ can be predicted by the values of a set of K explanatory variables $X_{1,s}, X_{2,s}, \ldots, X_{K,s}$. In a multinomial regression framework we assume that the logarithm of the odds of each cause of death relative to a reference cause are linearly dependent on these explanatory variables. This is expressed as a system of $C-1$ linear equations corresponding to each cause of death (excluding the reference category),

$$\begin{aligned} \log(P_{2,s}/P_{1,s}) &= \beta_{2,0} + \beta_{2,1}X_{1,s} + \beta_{2,2}X_{2,s} + \cdots + \beta_{2,K}X_{K,s} \\ \log(P_{3,s}/P_{1,s}) &= \beta_{3,0} + \beta_{3,1}X_{1,s} + \beta_{3,2}X_{2,s} + \cdots + \beta_{3,K}X_{K,s} \\ &\cdots \\ \log(P_{C,s}/P_{1,s}) &= \beta_{C,0} + \beta_{C,1}X_{1,s} + \beta_{C,2}X_{2,s} + \cdots + \beta_{C,K}X_{K,s} \end{aligned} \tag{3}$$

Notice that the β-coefficients (including the intercepts) do not have the study subindex s. This is a fixed-effects model that assumes the associations of the explanatory variables with the causes of death are constant across all studies. We relax the assumption that the baseline log-odds are the same in each study by adding study-specific random effects to the intercepts. Using matrix notation, and including random effects, the model in (3) can be expressed as

$$\begin{bmatrix} \log(P_{2,s}/P_{1,s}) \\ \log(P_{3,s}/P_{1,s}) \\ \cdots \\ \log(P_{C,s}/P_{1,s}) \end{bmatrix} = \begin{bmatrix} U_{2,s} \\ U_{3,s} \\ \cdots \\ U_{C,s} \end{bmatrix} + \begin{bmatrix} \beta_{2,0} & \beta_{2,1} & \cdots & \beta_{2,K} \\ \beta_{3,0} & \beta_{3,1} & \cdots & \beta_{3,K} \\ \cdots & \cdots & \cdots & \cdots \\ \beta_{C,0} & \beta_{C,1} & \cdots & \beta_{C,K} \end{bmatrix} \times \begin{bmatrix} 1 \\ X_{1,s} \\ \cdots \\ X_{K,s} \end{bmatrix} , \tag{4}$$

where the terms $U_{c,s}$ are study-specific and can be modelled as random effects with mean 0 across all studies. The notation in (4) could be simplified to

$$\boldsymbol{LP_s} := \begin{bmatrix} \log(P_{2,s}/P_{1,s}) \\ \log(P_{3,s}/P_{1,s}) \\ \cdots \\ \log(P_{C,s}/P_{1,s}) \end{bmatrix} = \boldsymbol{U_s} + \boldsymbol{\beta} \times \boldsymbol{X_s} , \tag{5}$$

where $\boldsymbol{U_s}$ and $\boldsymbol{X_s}$ are the study-specific vectors of random effects and explanatory variables respectively, and $\boldsymbol{\beta}$ is the matrix of fixed effects common to all studies.

2.3. Bayesian LASSO

In a Bayesian framework, we implement LASSO covariate selection by penalising large β coefficients in a subset of the fixed-effect parameters that could potentially result in overfitting the data. We do this by imposing a double exponential (also referred to as Laplace) prior distribution on them in the model specification (Park and Casella, 2008). Unlike the frequentist LASSO, the Bayesian LASSO shrinks the magnitude of the parameters without completely reducing them to zero, allowing covariates to have negligible effects for some outcomes and non-negligible effects for others. Shrinking the parameters has the additional advantage of stabilising the model if, due to the large number of parameters to be estimated with potentially high uncertainty, model convergence is slow or difficult.

2.3.1. Formal model definition

Our proposed method specifies Eq. (5) in the statistical model and uses an MCMC sampling algorithm to build it in a Bayesian framework. Note that the vector of true

causes of death is unobserved, but with a specification for the true cause distribution, a vector of observed reported causes and a (known) misclassification matrix, we can specify a multinomial distribution for the observed reported causes. Let N_s denote the sample of deaths from a given study s, and $\boldsymbol{M_s}$ the misclassification matrix defined in (2), our Bayesian model can be summarised as follows:

LIKELIHOOD:

$$\begin{aligned}
\boldsymbol{R_s} &\sim \text{Multinomial}\big(N_s, \boldsymbol{M_s} \times \boldsymbol{P_s}\big) && \text{distribution of recorded, observed CODs}\\
\boldsymbol{P_s} &= \frac{\exp(\boldsymbol{LP_s})}{1+\sum_{i\in I}\exp(\boldsymbol{LP_i})} && \text{proportions of true, (un)observed CODs}\\
\boldsymbol{LP_s} &= \boldsymbol{U_s} + \boldsymbol{\beta} \times \boldsymbol{X_s} && \text{log-odds of true, (un)observed CODs}\\
\boldsymbol{U_s} &\sim N\big(0, \Sigma_c\big) && \text{study-specific random effects}
\end{aligned}$$

PRIORS:

$$\begin{aligned}
\Sigma_c &\sim \text{Unif}\big(0, b\big) && \text{for each cause } c\\
\beta_{c,k} &\sim \text{Laplace}\big(0, \lambda\big) && \text{for each cause } c \text{ and LASSO constrained variable } k\\
\beta_{c,k^*} &\sim N\big(0, 0.5\big) && \text{for each cause } c \text{ and unconstrained variable } k^*\\
& && \text{including intercept}
\end{aligned}$$

Note that $\boldsymbol{R_s}$, N_s and $\boldsymbol{X_s}$ are observed data, whereas $\boldsymbol{M_s}$ is assumed to be known. Vector parameter $\boldsymbol{\Sigma} = (\Sigma_2, \ldots, \Sigma_C)$ contains $C-1$ standard deviations of the random effects, one for each cause except for the reference category. These standard deviations have uniformly distributed priors controlled by hyper-parameter b. For simplicity, we assumed there is no correlation between random effects of different causes, although this could be modelled in other ways. The intercepts and any β-coefficients we do not want to be constrained in the LASSO are given normally distributed priors with mean 0 and standard deviation 0.5. The remaining β-coefficients have a Laplace (double exponential) prior with mean 0 and scale $\lambda > 0$. The hyper-parameter λ is the penalty imposed by the LASSO method. We use out-of-sample cross-validation to select optimal λ and b parameters, described in the next section.

2.3.2. Estimation of country-level COD mortality fractions with credible intervals
Once the model has estimated $\boldsymbol{\beta}$ and $\boldsymbol{U_s}$ correcting for potential misclassification we can estimate the expected distribution of true COD in any country for which we have covariate data as

$$P_{c,s} = \frac{\exp\big(U_{c,s} + \boldsymbol{\beta_c} \times \boldsymbol{X_s}\big)}{1 + \exp\big(U_{2,s} + \boldsymbol{\beta_2} \times \boldsymbol{X_s}\big) + \ldots + \exp\big(U_{C,s} + \boldsymbol{\beta_C} \times \boldsymbol{X_s}\big)} \,. \tag{6}$$

When using this model to estimate country-level COD fractions we need to account for two sources of uncertainty: a) uncertainty surrounding the fixed-effects parameter estimates; and b) uncertainty about how to select the most appropriate random effect $U_{c,s}$ for the country that we want to estimate, particularly if that

country was not represented in our input data. In a Bayesian framework, credible intervals around the COD fractions can account for both sources of uncertainty. We compute these credible intervals by repeatedly estimating the COD distribution using values of β and relevant (described below) random effects drawn from MCMC chains generated during model estimation. We use $n = 1{,}000$ sets of estimates from equally spaced MCMC iterations after burn-in, although the selected sets can also be determined by a thinning parameter or even randomly selected. We thus obtain n sets of estimated mortality fractions for each country, from which we find the 2.5^{th} and 97.5^{th} centiles (to obtain 95% credible intervals) and mean and median values (to obtain point estimates) for each COD.

2.3.3. Choice of random effects

An important question is how to select relevant random effects from the matrices $\boldsymbol{U^i}$ in each iteration $i = 1 \ldots n$. Each $\boldsymbol{U^i}$ contains a vector of random effects for each study in the estimation dataset. There are several ways of deciding which is the most appropriate vector for a given country; we illustrate three below:

(a) We could assume $\boldsymbol{U_s} = 0$ in all iterations, which may seem a sensible strategy for countries not represented in the input data. However, this is an extreme option representing a strong prior belief that the country's COD distribution is exactly predicted by the model's fixed effects, when no evidence suggests this. As such it will produce overly narrow credible intervals, with variability determined only by fixed effect estimate uncertainty. We do not explore this option further.

(b) Choose a vector $\boldsymbol{U_s}$ at random from the matrix in each iteration. This strategy relaxes the belief that the country's COD distribution is exactly predicted by the model's fixed effects by drawing from a wide range of random effects at each iteration. In expectation these random effects sum to zero, having little effect on the point estimate, but because many different random effects are drawn the uncertainty in the COD distribution estimate is large. We expect wide credible intervals with variability dependent on both fixed and random effect estimate uncertainty.

(c) An intermediate option is to assume that a subset of the random effects, which may not necessarily sum to zero, are relevant for a given country. We randomly select the $\boldsymbol{U_s}$ vector in each iteration from this subset. This replaces the belief that the country's COD distribution is exactly predicted by the model's fixed effects with the belief that a certain subset of studies are useful for the country's predictions. This might move the point estimate away from the fixed effects predictions and will produce credible intervals of a width somewhere in between options (a) and (b), with variability determined as in (b) but incorporating less uncertainty in the selection of relevant random effects.

We implement options (b) and (c). For countries with no studies or only non-nationally representative studies in the input data, we use option (b) to calculate the credible intervals. For countries with nationally representative studies, we use option (c) to calculate the credible intervals by drawing from their nationally representative random effects.

3. Model building process

3.1. Model building process

There were 95 studies in a previous round of frequentist modelling of neonatal COD estimation (Oza, 2019), and we use the same studies in fitting the Bayesian model. As before, we let "intrapartum" be the reference COD in the multinomial model and build equations for the other seven causes (detailed in Sect. 1) compared to "intrapartum". Each model equation contained an intercept term, 14 covariates (Sect. 1) and a random effect term, all of which were allowed to vary by COD.

3.1.1. Fixed effects

We modelled all covariates linearly, with no polynomial effects or interactions. The beta matrix therefore contained 105 coefficients for estimation.

The early and late period indicators (as well as the intercept) for each COD (21 coefficients) were given $N(0, 0.5)$ priors free from the LASSO penalisation to ensure the two mortality periods can have different COD distributions where appropriate. The remaining 84 coefficients were each subject to LASSO by imposing the Laplace$(0, \lambda)$ prior. We explored λ values of 5, and 10 to 250 in increments of 20. This wide range allows for different values of b to suggest different optimal values of λ.

3.1.2. Random effects

We tagged each study with an indicator variable for the random effects term and identified whether it was nationally representative. This resulted in 95 sets of random effects, of which 16 were nationally representative, and the U matrix therefore contained 665 random effect coefficients for estimation.

We gave the random effects distributions $N(0, \Sigma_c)$ priors, and gave Σ_c the hyper-prior Unif$(0, b)$. The hyper-prior controlled by b imposes a ceiling on the random effect variances that helps the model to avoid overfitting random effects to noisy data and limits the extent to which study data can influence country predictions. For example, when $b = 0.21$ only 5% of random effects in a particular cause's posterior distribution should affect the fixed-effects odds for that cause by a magnitude greater than 1.51 or its reciprocal. We let this be the upper limit of influence and explored two more restrictive constraints of $b = 0.07$ and $b = 0.14$.

3.2. Formal selection of λ and b using cross-validation

We selected λ and b jointly by partitioning the dataset into k subsets and performing a k-fold cross-validation over a selection of values of both parameters. With $k = 10$ we found that the error curves were sensitive enough to the choice of the partition to suggest different values for λ, providing evidence for high variability but (because k is low relative to sample size) no theoretical guarantee of unbiasedness. Instead, we therefore used leave-one-out cross-validation, in which the error curve is theoretically highly variable but approximately unbiased (Jiang et al., 2002). We set k at 95, the number of studies in our dataset.

For each combination of λ and b, we ran a full leave-one-out analysis in the following way: we remove a study from the input data, run the MCMC model with the remaining 94 studies and predict the distribution of deaths in the out-of-sample study using only the fixed effects. We repeat this 95 times, leaving a different study

out each time so that we obtain an out-of-sample predicted COD distribution for each left-out study. For each study and COD we calculate the squared differences between predicted and observed deaths, weight them by the ratio of the predicted deaths to the total deaths over all studies and CODs, and sum these errors over all studies and CODs to obtain a weighted mean square error (wMSE) for each jackknife sample. We compare the wMSE across jackknife samples (over different values of λ and b) and select the combination that produces the lowest absolute value.

3.2.1. Parameters for the MCMC convergence

To avoid the effects of autocorrelation and ensure convergence, we ran models in four parallel chains of 10,000 iterations each and an initial burn-in sequence of 2,000. We used trace and Gelman plots (Brooks and Gelman, 1998) of the coefficients to determine the number of iterations to burn, choosing the iteration at which the slowest parameter had converged over all four chains (flat locally weighted regression lines on MCMC trace) and the chains showed evidence of convergence (scale reduction factor below 1.1 over majority of iterations). We determined the number of post-burn iterations from the parameter with the strongest autocorrelation, ensuring that the four trace lines cycled through the posterior median at least twice beyond the burn point. For this reason we did not thin our posterior distributions: biasing effects of autocorrelation should average out, given the large number of iterations and chains.

Further modelling details are available in **Online Supplement Text E2**. We used R v3.5.2 for all analyses (R Core Team, 2020). Bayesian analyses were implemented in JAGS (Plummer, 2003) with wrapper functions from the `R2jags` package (Su and Yajima, 2020).

Statistical code is available on `https://github.com/amulick/MCEE-neo.git`.

4. Results and evaluation

We ran leave-one-out models on a 2.8 gigaHertz (GHz) machine with eight parallel processors, each of which completed in ∼16.5 hours using approximately 26 GB of RAM. Total computation time including final model estimation was approximately 700 hours. Further details are available in **Online Supplement Text E3**.

4.1. Cross-validation

We ran leave-one-out analyses at each of the 42 combinations of λ and b. Figure 1 plots the out-of-sample wMSE against values of λ over three values of b. The initial decline of these curves is a reflection of the overfitting of the fixed effect coefficients of the model (i.e. the effect of the covariates used for prediction) to the input data, and the later incline reflects underfitting; thus the value of λ at the minimum wMSE indicates the model with the best out-of-sample fit. These curves were steeper and had lower minima at higher b values, such that the most accurate predictions were achieved with the least restrictive cap on Σ_c ($b = 0.21$).

FIGURE 1 ABOUT HERE

This illustrates that the selection of the best predictive model depends on both λ and b. Among the combinations of λ and b examined, we obtained the worst out-of-sample prediction (wMSE $> 5 \times 10^6$) when the fixed effects are most flexible by both parameters ($\lambda = 5$, $b = 0.07$) – a clear example of overfitting. As we restrict the magnitude of the fixed effects by increasing λ, the out-of-sample prediction initially improves over all three values of b. At the same time, as we "relieve" the fixed effects from having to fit the data by allowing the model to have larger random effects (increasing b; moving from the black to the red to the green line in Fig. 1), we also improve out-of-sample prediction. The combination of these two measures prevents overfitting, but eventually with large enough λ we restrict the fixed effect coefficients too much and predictions worsen again. This limit appears at lower λ if the random effects are allowed to explain more data variability with larger b.

4.2. Final models

To understand the differences between these three best-fit models, we ran a final model for each of them using all data points. Trace and Gelman plots for the covariates and nationally representative random country effects showed good convergence in all three models. Most of the seven Σ_c parameters were sampled consistently at or near b.

The difference in beta estimates was generally very small (Fig. 2), although some estimates were notably weaker in the $\lambda = 190$ and $b = 0.07$ model compared to the other two models, with larger differences in the LASSO-restricted coefficients (white background) than in the unrestricted coefficients (grey background).

FIGURE 2 ABOUT HERE

The difference in random effects was greater (Fig. 3) among the three models. Although some estimates were similar between them, most were markedly different with, as expected, larger effects appearing in models with larger values of b.

FIGURE 3 ABOUT HERE

4.3. Predictions

We used the model with the best out-of-sample fit ($\lambda = 30$ and $b = 0.21$) to predict COD distributions in 80 countries between the years 2000 and 2015 inclusive. For 8 of the 80 countries nationally representative data were available: Bangladesh and Morocco had two nationally representative studies and Afghanistan, India, Indonesia, Mozambique, Nepal and Pakistan had a single study.

4.3.1. Frequentist vs Bayesian

Figure 4 and Table 1 compare predicted proportions of all causes of death between the classical frequentist model and our proposed Bayesian model (fixed effects only). This provides comparable estimates in that the proportions differ only in the method used to develop the predictive model.

FIGURE 4 ABOUT HERE

TABLE 1 ABOUT HERE

Predicted proportions for intrapartum, preterm and sepsis were distributed roughly equally on either side of the line of equivalence; for the other causes one of the models tended to predict larger proportions than the other. The median difference in absolute terms was largest for congenital proportions (Bayesian estimates 3% lower than frequentist); in relative terms the difference was greatest for diarrhoea (Bayesian estimates on average 3 times higher). However, both models assigned relatively small proportions (maximum 0.25) to both of these causes.

Within each cause we highlight outliers (countries whose estimate is greater than 3 standard deviations from the average difference between the two methods) with a triangle. There was only one such difference: Sudan's Bayesian pneumonia estimate was markedly lower than its frequentist estimate.

4.3.2. Bayesian: Fixed vs random effects

Figure 5 compares point estimates from fixed-effects only and fixed- plus random-effects predictions for countries with nationally representative input data (except Morocco, because its studies preceded 2000). For most countries and CODs, the fixed- plus random-effects predicted proportion lies between the empirical proportion than the predicted proportion based on the fixed effects only. There were some exceptions: 1) Bangladesh (2002), congenital; 2) India, diarrhoea; 3) Mozambique, intrapartum and other; 4) Nepal, diarrhoea and tetanus; and 5) Pakistan, congenital. For these, the estimates incorporating the random effects were all further away from the empirical proportion than the estimates based on the fixed effects only.

FIGURE 5 ABOUT HERE

The first exception occurred because Bangladesh has two empirical data points (2002 and 2008), both of which informed its random effects, and these fell on either side of the fixed-effects prediction for congenital. Bangladesh's fixed- plus random-effects prediction equation drew randomly from both random effects to produce the posterior distribution of point estimates, so that the mean prediction averaged out the two random effects. For all other CODs, Bangladesh's empirical data points were on the same side of the fixed-effects estimates.

The second and third exceptions occurred because of non-reporting of key CODs. India did not report pneumonia deaths and Mozambique did not report congenital, diarrhoea or tetanus deaths, so the misclassification matrix for these studies redistributed some of the deaths they did report into these causes. In this way, the random effects for each model COD are not expected to necessarily pull all proportions in the same direction as the empirical data.

The fourth and fifth exceptions likely occurred due to chance. The average random effects (Nepal: diarrhoea and tetanus, Pakistan: congenital) were all very close to zero and the change in predictions due to adding the random effect were very small (Fig. 5). Thus the slight pull in the opposite direction was likely due to the

prediction algorithm sampling, by chance, more random effects on the opposite side of zero than the empirical data would suggest.

In general, though not exclusively, the empirical proportions representing larger numbers of deaths pull the random-effects estimates more strongly towards themselves than proportions representing fewer deaths. This can be seen by comparing Nepal (2004) and Pakistan (2006): Pakistan's study reported on larger numbers of deaths and their random-effects predictions for nearly all CODs are closer to their empirical data than Nepal's predictions are to their empirical data.

4.4. Uncertainty ranges

Figure 6 compares the 2015 mortality proportions and uncertainty ranges across four causes of death between the classical frequentist model and our proposed Bayesian model for selected countries. Most estimates from countries without nationally representative data are close to the frequentist point estimates, or within the uncertainty limits of the frequentist estimates. In general, the Bayesian credible intervals are more symmetric around the point estimates than the bootstrapped frequentist confidence intervals.

FIGURE 6 ABOUT HERE

Table 2 summarises, for 80 countries with available data, differences in the widths of the uncertainty intervals. For countries without nationally representative data, the Bayesian intervals were generally wider than the frequentist, particularly for pneumonia and preterm CODs.

TABLE 2 ABOUT HERE

In contrast, estimates for the countries that did provide nationally representative data generally have narrower Bayesian than frequentist uncertainty intervals. This is because the algorithm calculating the fixed- plus random-effects estimate is drawing from only one or two relevant random effects, rather than all 95 random effects. Fewer random effects add less variability to the posterior distribution; thus its 2.5^{th} and 97.5^{th} centiles are closer together. This has the initially counterintuitive consequence that countries providing more nationally representative data, such as Bangladesh and Morocco with two studies, can have wider uncertainty intervals than they would had they provided only one study.

Discussion

We developed a new approach within a Bayesian framework for estimating the distribution of causes of death at national level based on COD data from national and subnational studies and data on covariates. Compared with the previous frequentist approach, this approach resulted in similar estimates on average for the most common causes of death, but some differences in the estimates for the less common causes. The incorporation of random effects into the model led to somewhat increased levels of uncertainty, with the exception of countries with nationally representative studies, where estimated uncertainty was not unexpectedly decreased. We

believe that the wider uncertainty ranges generated under the new approach better reflect real uncertainty that exists given the substantial scope for misclassification in verbal autopsy studies.

By working within the Bayesian framework we were able to incorporate features which we could not in a frequentist setting. First, we were able to better incorporate empirical data from nationally representative studies through the use of random effects, so that estimates for countries with empirical data from nationally representative studies are "pulled towards" those empirical data. Second, we were able to examine the associations between a set of predictors and all causes of interest simultaneously with the Bayesian LASSO, rather than selecting predictors in a more naïve process with each cause individually or in an exhaustive and computationally prohibitive search across predictors in the multinomial space. Third, we were able to recast our method for mapping between causes as they are reported and specific causes of interest (the misclassification matrix), opening up possibilities for future extensions discussed below.

Including random effects for countries with nationally representative studies has several important advantages. Countries with nationally representative cause-of-death information contribute to estimates for all countries through the model's fixed effects coefficients, but also have greater influence on the estimates for that country which likely reduces bias (Bouwmeester et al., 2013; McCulloch and Neuhaus, 2011). This also is intuitively a pleasing compromise between what countries report, which may have limitations (Menéndez et al., 2020), and what would be expected given the evidence from other areas with similar levels of mortality, intervention coverage, and other health system characteristics. To our knowledge there are not other methods that allow for such a systematic compromise in estimating causes of mortality. Allowing nationally representative cause-of-death measurement to have more influence may also encourage the collection of such data, which would increase the accuracy and usability of future estimates as well as the capacity in low-resource countries for such measurement.

We also incorporated cross-validation to determine the degree of penalization in the Bayesian LASSO. Although the Bayesian LASSO is often implemented with a fixed restriction or with a hyper-prior for the restriction parameter (Park and Casella, 2008), we chose the degree of restriction according to the cross validation error for more robust out of sample prediction (Efron and Tibshirani, 1995).

In addition to its advantages, the proposed method in the Bayesian framework is more flexible than the previous method in the frequentist setting, which we expect to allow for further improvement. For example, cause of death ascertainment can be prone to measurement error, particularly when using verbal autopsy (World Health Organization, 2016). With our new approach, this has potential for resolution through an extension to the misclassification matrix. This framework may also allow increased accounting for uncertainty and measurement error in the covariates for primary data, which often is not available at the same level of resolution for which the causes of mortality were measured (Liu et al., 2016).

We used credible intervals to estimate the uncertainty in predicted cause distributions for all countries by resampling from the MCMC-estimated random effects for each study, and among only those estimates from nationally representative studies for countries with such studies. This may in practice lead to the counterintuitive result that a country with more than one nationally representative study could appear more uncertain than a country with only one nationally representative study.

Studies from the same area may report different causes due to differences in study methods such as cause ascertainment (Murray et al., 2014) or due to epidemiologic changes over time. In such cases, an increase in uncertainty may be warranted. Countries with a single nationally representative study may have uncertainty intervals which are too narrow, because they likely do not account for uncertainty due to possible error in the cause-of-death classification.

The proposed method is in contrast to those used by the Global Burden of Disease (GBD) consortium (Murray et al., 2020) for estimating the causes of mortality. This group looks at many separate causes for different age groups, including data from incomplete vital registration and registries for specific syndromes and aetiologies of disease. After all causes are estimated separately with Gaussian processes, they are then restricted to an age-specific envelope (total number of deaths due to all causes) in a separate process (Murray et al., 2020), although their methods and data sources are not publicly available in detail (Schwab, 2020). The method proposed here is not directly comparable to GBD methods because each are based on different source information. When approaching compositional data such as causes of death, addressing components individually as done by GBD is generally unbiased, but there are caveats for estimating the variance of separately estimated components which are subsequently fitted to an envelope (Begg and Gray, 1984; Fürnkranz, 2002; Hsu and Lin, 2002). The estimated variance for each component is used in the GBD method for harmonizing the many causes to fit the mortality envelope (Murray et al., 2020) and so may introduce bias in the resulting estimates. The vast amount of input data used by GBD allows for the estimation of many different causes, which would be computationally difficult in the multinomial framework. However, the "squeezing" process made necessary by the many different causes may lead to inaccuracies. The method proposed here is computationally complex, but it is executed in a single systematic framework that is used widely in other similar problems with compositional measures (Haan and Uhlendorff, 2006) and with attractive statistical properties that are well documented (Engel, 1988).

The proposed method is, as with many statistical methods, limited by the quality of primary measurements. Causes of death in high mortality and low resource settings are often subject to specific types of measurement error (Adewemimo et al., 2017). Both the amount of information related to causes of death in areas without vital registration (Datta et al., 2020) as well as the methods for measuring causes of death in such areas (Kalter et al., 2020; McCormick et al., 2016) are improving. However, historic cause-of-death measurements are likely unique, as health systems change and mortality among neonates and children declines. So, historic causes of mortality may need bespoke measures when attempting to correct for them (Yadav and Arokiasamy, 2014), which our extension to the misclassification matrix has potential to do. Another important limitation of the proposed method is the amount of time and computational resources necessary for implementation. We used parallel computing on a multi-node high performance computing cluster for these analyses. Such resources are not widely available and are associated with both financial and time related costs, as an analyst must learn both the method and the protocol of the computing cluster. High performance computing, however, is becoming more accessible and may be less of a barrier in the future (Clark, 2020).

Although there have been important advances in vital registration as well as sample registration systems for measuring causes of death for all individuals, there are still gaps in these systems that will likely make modelling causes of death neces-

sary for the foreseeable future (Amouzou et al., 2020). Advances in the accuracy of cause-of-death estimates translate into better knowledge of what contributes most to age-specific mortality and how health systems can be configured for the biggest impact (Walker and Friberg, 2017). Although the proposed method has improved predictions in several aspects relative to previous methods, more improvement is possible and is the subject of further research. Future work related to incorporating measurement error as well as predicting causes of mortality with increased resolution for narrower age groups is ongoing and is likely to increase the usability and reliability of cause-of-death estimates.

Funding

Bill and Melinda Gates Foundation, grant number OPP1096225

Conflicts of interest

None declared.

References

Adewemimo, A., Kalter, H. D., Perin, J., Koffi, A. K., Quinley, J. and Black, R. E. (2017) Direct estimates of cause-specific mortality fractions and rates of under-five deaths in the Northern and Southern regions of Nigeria by verbal autopsy interview. *PloS One*, **12**, e0178129.

Albert, J. H. and Chib, S. (1993) Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, **88**, 669–679.

Amouzou, A., Kante, A., Macicame, I., Antonio, A., Gudo, E., Duce, P. and Black, R. E. (2020) National sample vital registration system: A sustainable platform for COVID-19 and other infectious diseases surveillance in low and middle-income countries. *Journal of Global Health*, **10**, 020368.

Begg, C. B. and Gray, R. (1984) Calculation of polychotomous logistic regression parameters using individualized regressions. *Biometrika*, **71**, 11–18.

Bouwmeester, W., Twisk, J. W., Kappen, T. H., Van Klei, W. A., Moons, K. G. and Vergouwe, Y. (2013) Prediction models for clustered data: Comparison of a random intercept and standard regression model. *BMC Medical Research Methodology*, **13**, 19.

Brooks, S. P. and Gelman, A. (1998) General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, **7**, 434–455.

Burda, M., Harding, M. and Hausman, J. (2008) A Bayesian mixed logit-probit model for multinomial choice. *Journal of Econometrics*, **147**, 232–246.

Cawley, G. C., Talbot, N. L. and Girolami, M. (2007) Sparse multinomial logistic regression via Bayesian L1 regularisation. In *Advances in Neural Information Processing Systems 19: Proceedings of the 2006 Conference*, 209–216. Cambridge, MA: MIT Press.

Clark, D. (2020) Amazon and Apple are powering a shift away from Intel's chips. *The New York Times*, Dec 1, 2020. URL: `https://www.nytimes.com/2020/12/01/technology/amazon-apple-chips-intel-arm.html`.

COMSA (2020) Countrywide Mortality Surveillance for Action (COMSA) in Mozambique. URL: `https://www.jhsph.edu/research/centers-and-institutes/institute-for-international-programs/current-projects/countrywide-mortality-surveillance-for-action-comsa-in-mozambique/`.

Datta, A., Fiksel, J., Amouzou, A. and Zeger, S. L. (2020) Regularized Bayesian transfer learning for population-level etiological distributions. *Biostatistics*, Advance online publication. URL: `https://doi.org/10.1093/biostatistics/kxaa001`.

Efron, B. and Tibshirani, R. (1995) Cross-validation and the bootstrap: Estimating the error rate of a prediction rule. *Tech. rep.*, Stanford University, Stanford, CA. URL: `https://statistics.stanford.edu/sites/g/files/sbiybj6031/f/BIO176.pdf`.

Engel, J. (1988) Polytomous logistic regression. *Statistica Neerlandica*, **42**, 233–252.

Fürnkranz, J. (2002) Round robin classification. *Journal of Machine Learning Research*, **2**, 721–747.

Haan, P. and Uhlendorff, A. (2006) Estimation of multinomial logit models with unobserved heterogeneity using maximum simulated likelihood. *Stata Journal*, **6**, 229–245.

Hedeker, D. (2003) A mixed-effects multinomial logistic regression model. *Statistics in Medicine*, **22**, 1433–1446.

Hsu, C.-W. and Lin, C.-J. (2002) A comparison of methods for multiclass support vector machines. *IEEE Transactions on Neural Networks*, **13**, 415–425.

Hug, L., Alexander, M., You, D. and Alkema, L. on behalf of the UN Inter-agency Group for Child Mortality Estimation (2019) National, regional, and global levels and trends in neonatal mortality between 1990 and 2017, with scenario-based projections to 2030: a systematic analysis. *The Lancet Global Health*, **7**, e710–e720.

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013) *An introduction to statistical learning with applications in R*. New York, NY: Springer.

Jiang, J., Lahiri, P. and Wan, S. M. (2002) A unified jackknife theory for empirical best prediction with M-estimation. *Annals of Statistics*, **30**, 1782–1810.

Jostins, L. and McVean, G. (2016) Trinculo: Bayesian and frequentist multinomial logistic regression for genome-wide association studies of multi-category phenotypes. *Bioinformatics*, **32**, 1898–1900.

Kalter, H. D., Perin, J., Perin, J., Amouzou, A., Kwamdera, G., Adewemimo, W. A., Nguefack, F., Roubanatou, A. M. and Black, R. E. (2020) Using health facility deaths to estimate population causes of neonatal and child mortality in four African countries. *BMC Medicine*, **18**, 183.

Liu, L., Oza, S., Hogan, D., Chu, Y., Perin, J., Zhu, J., Lawn, J. E., Cousens, S., Mathers, C. and Black, R. E. (2016) Global, regional, and national causes of under-5 mortality in 2000–15: an updated systematic analysis with implications for the Sustainable Development Goals. *The Lancet*, **388**, 3027–3035.

McCormick, T. H., Li, Z. R., Calvert, C., Crampin, A. C., Kahn, K. and Clark, S. J. (2016) Probabilistic cause-of-death assignment using verbal autopsies. *Journal of the American Statistical Association*, **111**, 1036–1049.

McCulloch, C. E. and Neuhaus, J. M. (2011) Misspecifying the shape of a random effects distribution: Why getting it wrong may not matter. *Statistical Science*, **26**, 388–402.

Menéndez, C., Quintó, L., Castillo, P., Carrilho, C., Ismail, M. R., Lorenzoni, C., Fernandes, F., Hurtado, J. C., Rakislova, N., Munguambe, K. et al. (2020) Limitations to current methods to estimate cause of death: a validation study of a verbal autopsy model. *Gates Open Research*, **4**, 55. URL: `https://gatesopenresearch.org/articles/4-55/v1`.

Murray, C. J., Lozano, R., Flaxman, A. D., Serina, P., Phillips, D., Stewart, A., James, S. L., Vahdatpour, A., Atkinson, C., Freeman, M. K. et al. (2014) Using verbal autopsy to measure causes of death: The comparative performance of existing methods. *BMC Medicine*, **12**, 5.

Murray, C. J. L., Aravkin, A. Y., Zheng, P., Abbafati, C., Abbas, K. M., Abbasi-Kangevari, M., Abd-Allah, F., Abdelalim, A., Abdollahi, M., Abdollahpour, I. et al. (2020) Global burden of 87 risk factors in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. *The Lancet*, **396**, 1223–1249.

Oza, S. (2019) *The use of statistical models to estimate the timing and causes of neonatal deaths*. Phd thesis, London School of Hygiene and Tropical Medicine. URL: `https://doi.org/10.17037/PUBS.04655993`.

Oza, S., Lawn, J. E., Hogan, D. R., Mathers, C. and Cousens, S. N. (2015) Neonatal cause-of-death estimates for the early and late neonatal periods for 194 countries: 2000–2013. *Bulletin of the World Health Organization*, **93**, 19–28.

Park, T. and Casella, G. (2008) The Bayesian Lasso. *Journal of the American Statistical Association*, **103**, 681–686.

Plummer, M. (2003) JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd international Workshop on Distributed Statistical Computing*. Vienna, Austria. URL: `https://www.r-project.org/conferences/DSC-2003/`.

R Core Team (2020) R: A language and environment for statistical computing. URL: `https://www.r-project.org/`.

Schwab, T. (2020) Are Bill Gates's billions distorting public health data? *The Nation*, Dec 3, 2020.

Snijders, T. A. B. (2005) Fixed and random effects. In *Encyclopedia of statistics in behavioral science* (eds. B. S. Everitt and D. C. Howell). Chichester, UK: Wiley. URL: `https://doi.org/10.1002/0470013192.bsa234`.

Su, Y.-S. and Yajima, M. (2020) R2jags: Using R to Run 'JAGS'. URL: `https://cran.r-project.org/package=R2jags`.

UNICEF, World Health Organization, World Bank Group and United Nations (2019) Levels and trends in child mortality 2019. *Tech. rep.*, UNICEF, New York, NY. URL: `https://www.unicef.org/reports/levels-and-trends-child-mortality-report-2019`.

United Nations (2017) Annex: Global indicator framework for the Sustainable Development Goals and targets of the 2030 Agenda for Sustainable Development. *Work of the Statistical Commission pertaining to the 2030 Agenda for Sustainable Development*, A/RES/71/313. URL: `https://unstats.un.org/sdgs/indicators/Global%20Indicator%20Framework_A.RES.71.313%20Annex.pdf`.

Walker, N. and Friberg, I. K. (2017) Introduction: Reporting on updates in the scientific basis for the Lives Saved Tool (LiST). *BMC Public Health*, **17**, 774.

World Health Organization (2016) Verbal autopsy standards: The 2016 WHO verbal autopsy instrument. *Tech. rep.*, World Health Organization, Geneva. URL: `http://www.who.int/healthinfo/statistics/verbalautopsystandards/en/`.

— (2017) WHO methods and data sources for country-level causes of death 2000–2015. *Tech. rep.*, World Health Organization, Department of Information, Evidence and Research, Geneva. URL: `https://www.who.int/healthinfo/global_burden_disease/GlobalCOD_method_2000_2015.pdf`. Global Health Estimates Technical Paper WHO/HIS/IER/GHE/2016.3.

— (2018) MCEE-WHO methods and data sources for child causes of death 2000–2016. *Tech. rep.*, World Health Organization, Department of Evidence, Information and Research, Geneva, and Maternal Child Epidemiology Estimation Group. URL: `https://www.who.int/healthinfo/global_burden_disease/childcod_methods_2000_2017.pdf`. Global Health Estimates Technical Paper WHO/HMM/IER/GHE/2018.4.

Yadav, S. and Arokiasamy, P. (2014) Understanding epidemiological transition in India. *Global Health Action*, **7**, 23248.

Yau, P., Kohn, R. and Wood, S. (2003) Bayesian variable selection and model averaging in high-dimensional multinomial nonparametric regression. *Journal of Computational and Graphical Statistics*, **12**, 23–54.

Table 1. Differences in predicted cause-of-death fractions for 2015, Bayesian compared to frequentist, for 8 countries

COD	Absolute difference: Median [IQR]	Relative difference: Median [IQR]
congenital	$0.03\ [-0.05, -0.01]$	$0.73\ [0.64, 0.85]$
diarrhoea	$0.01\ [0.01, 0.01]$	$3.24\ [2.80, 3.76]$
intrapartum	$-0.02\ [-0.04, 0.02]$	$0.94\ [0.85, 1.09]$
pneumonia	$0.01\ [0.00, 0.02]$	$1.17\ [1.03, 1.30]$
preterm	$0.01\ [-0.01, 0.03]$	$1.04\ [0.97, 1.10]$
sepsis	$-0.01\ [-0.02, 0.02]$	$0.93, [0.85, 1.09]$
tetanus	$0.00\ [0.00, 0.01]$	$1.59\ [1.05, 2.18]$
other	$0.01\ [0.00, 0.02]$	$1.16\ [1.02, 1.33]$

Table 2. Widths of Bayesian credible intervals compared to widths of bootstrapped frequentist confidence intervals for selected 2015 predicted COD fractions in 80 countries. Values greater than zero indicate how much wider the Bayesian intervals are, in absolute value, than the frequentist intervals

COD	n	Median [IQR]
Countries without nationally representative studies		
congenital	72	$0.01\ [0.00, 0.02]$
intrapartum	72	$0.02\ [-0.01, 0.04]$
pneumonia	72	$0.02\ [0.01, 0.03]$
preterm	72	$0.11\ [0.08, 0.12]$
Countries with nationally representative studies		
congenital	8	$-0.02\ [-0.03, -0.01]$
intrapartum	8	$-0.04\ [-0.06, -0.03]$
pneumonia	8	$0.01\ [-0.02, 0.02]$
preterm	8	$-0.06\ [-0.07, -0.04]$

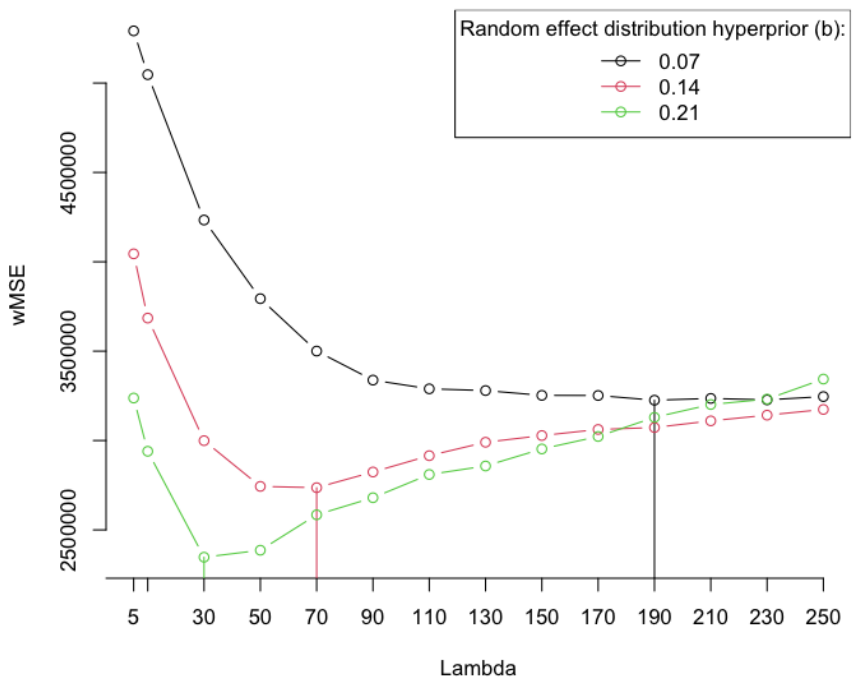

Fig. 1. Weighted Mean Square Error (wMSE) from out-of-sample neonatal cause-of-death predictions plotted against values of fixed-effects shrinkage parameter λ over three values of random effect distribution hyper-priors b

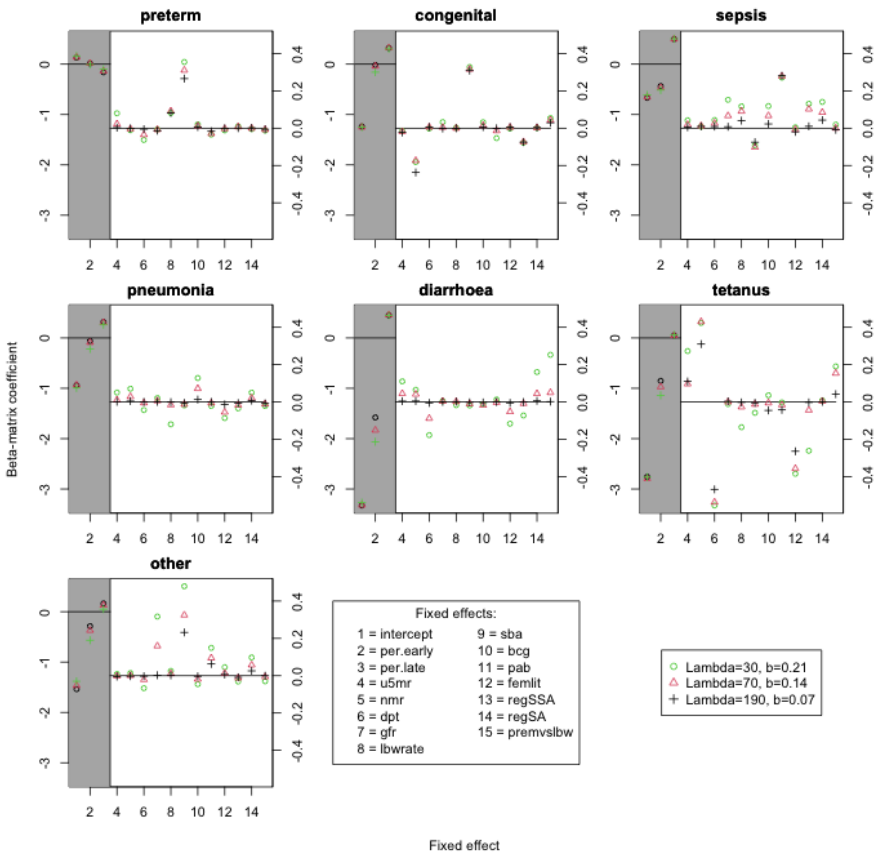

Fig. 2. Bayesian fixed effects coefficients compared between three neonatal cause-of-death models with different values of fixed-effect shrinkage parameters λ and random effect distribution hyper-priors b

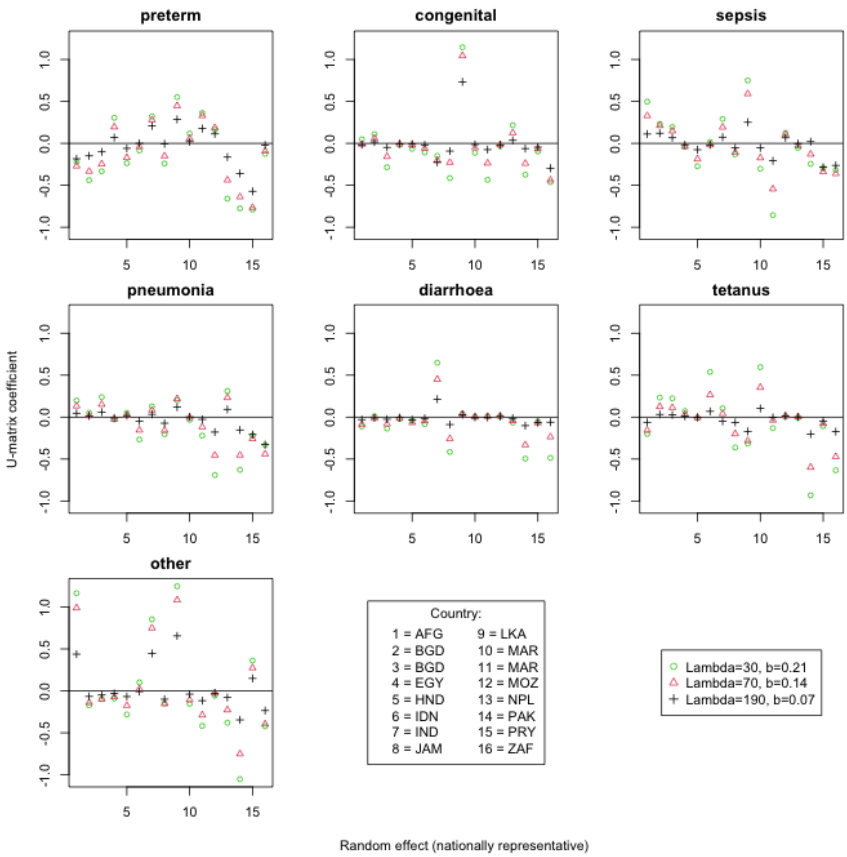

Fig. 3. Bayesian random effects coefficients compared between three neonatal cause-of-death models with different values of fixed-effect shrinkage parameters λ and random effect distribution hyper-priors b. Sixteen sets of random effects are from the studies with nationally representative input data for their country

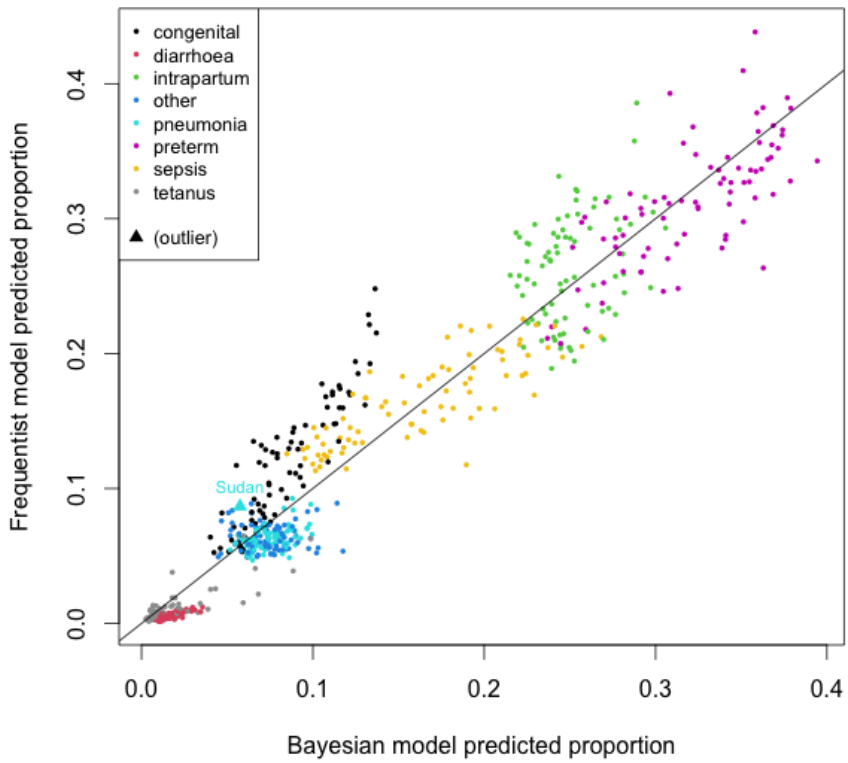

Fig. 4. Predicted Bayesian fixed-effects cause-of-death proportions compared with predicted frequentist proportions in 2015 for 80 countries

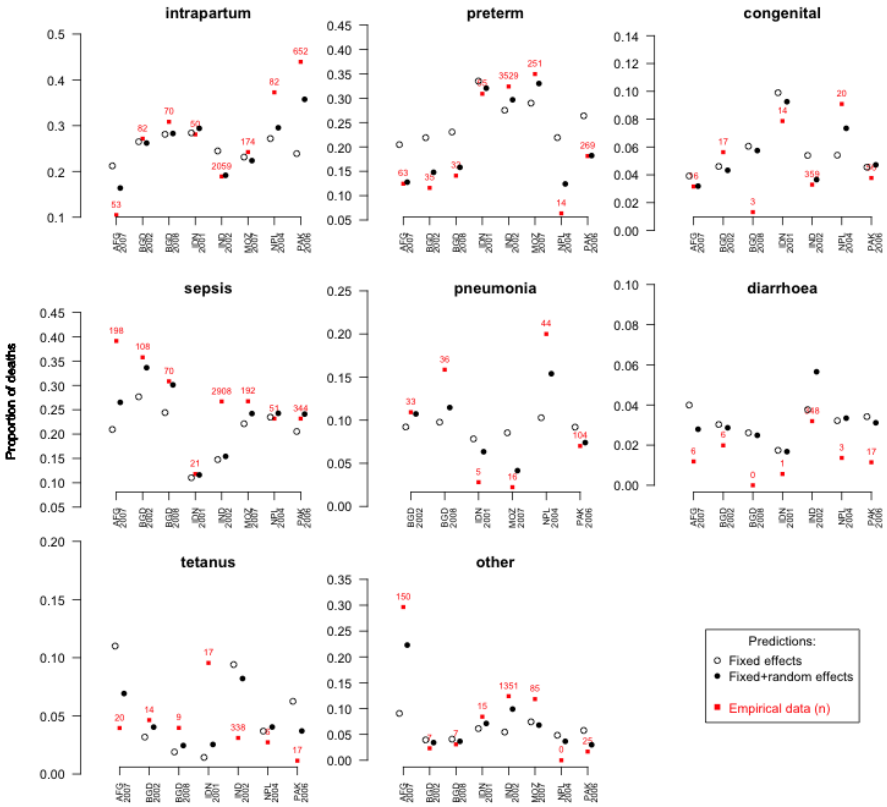

Fig. 5. Fixed-effects only versus fixed- plus random-effects estimates of cause-of-death distributions in countries contributing nationally representative data to the Bayesian model. Red squares mark the study COD proportion and indicate the absolute number of deaths it represents. Bangladesh is represented twice because it provided two nationally representative studies. Afghanistan, India and Mozambique do not appear in all causes because their studies did not report deaths in the same categories that we report

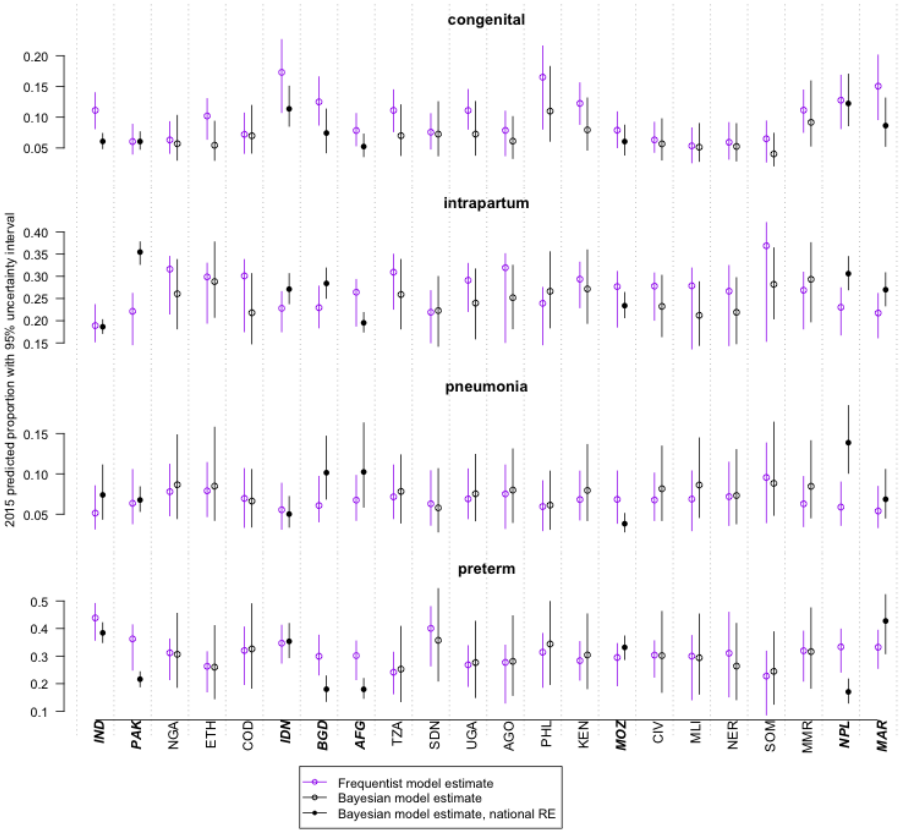

Fig. 6. Mortality proportions and uncertainty ranges for select neonatal causes of death in 2015 compared between a classical frequentist model with bootstrapped confidence intervals and our proposed Bayesian model (fixed- plus random-effects estimates) with credible intervals. In descending order from left to right are a union of 20 countries with the highest burden of neonatal mortality in 2015, and eight countries that contributed nationally representative studies to the Bayesian model. Countries contributing nationally representative data are highlighted in bold italic