Data-driven methods for imputing national-level incidence in global burden of disease studies

Scott A McDonald, a Brecht Devleesschauwer, b Niko Speybroeck, c Niel Hens, d Nicolas Praet, e Paul R Torgerson, f Arie H Havelaar, g Felicia Wu, h Marlène Tremblay, i Ermias W Amene j & Dörte Döpfer k

Objective To develop transparent and reproducible methods for imputing missing data on disease incidence at national level for the year 2005.

Methods We compared several models for imputing missing country-level incidence rates for two foodborne diseases – congenital toxoplasmosis and aflatoxin-related hepatocellular carcinoma. Missing values were assumed to be missing at random. Predictor variables were selected using least absolute shrinkage and selection operator regression. We compared the predictive performance of naive extrapolation approaches and Bayesian random and mixed-effects regression models. Leave-one-out cross-validation was used to evaluate model accuracy.

Findings The predictive accuracy of the Bayesian mixed-effects models was significantly better than that of the naive extrapolation method for one of the two disease models. However, Bayesian mixed-effects models produced wider prediction intervals for both data sets.

Conclusion Several approaches are available for imputing missing data at national level. Strengths of a hierarchical regression approach for this type of task are the ability to derive estimates from other similar countries, transparency, computational efficiency and ease of interpretation. The inclusion of informative covariates may improve model performance, but results should be appraised carefully.

Abstract in Arabic, Chinese, Czech, French, Russian and Spanish at the end of each article.

Introduction

An essential prerequisite for estimating global disease burden using summary health metrics, such as the disability-adjusted life year, is the availability of national-level data on the incidence or prevalence of the disease of interest. For some countries such information is not available, due to financial constraints, lack of surveillance data or other factors.

The World Health Organization (WHO) initiative to estimate the global burden of foodborne diseases, launched in 2006, is advised by the Foodborne Disease Burden Epidemiology Reference Group. For applied studies such as this, methods are required to estimate disease burden for countries with missing national-level data. In many studies, extrapolation approaches with little validation have been employed to fill data gaps, e.g. by assigning the value of a certain country or the mean value of neighbouring countries to the country with missing data. Such methods are arbitrary and do not account for uncertainty arising from the imputation of missing data.

Statistical methods are available for the analysis of incomplete datasets. Alternatively, there are numerous methods for imputing missing data and for assessing their validity. Given the availability of information for populations both with and without missing data, empirical models are typically fitted to existing data and predictions are generated for missing values from the fitted model. This method, sometimes termed farcasting, by analogy with forecasting, therefore generates predicted health statistics. An important step in this method is the selection of variables allowing robust predictions.

The imputation methods we investigated require that missing data are missing at random. Resource-poor countries may be less likely to report the incidence of a given disease, despite having a higher incidence compared with resource-rich countries. However, provided that the probability of reporting disease incidence is unrelated to disease incidence within the category of resource-poor countries, such data would be considered missing at random. This observation implies that the imputation model should allow for clustering of countries according to resource-richness. Missing data may also be more frequent for less populous nations. However, true disease incidence is rarely related to population size. Imputed incidences for small national populations usually have minor effects on disease burden estimates at regional or global level.

Our goal was to estimate disease incidence and associated uncertainty at national level. We describe a method for comparing and evaluating relatively simple data-driven imputation approaches with data sets employed in global burden of disease studies, assuming that data are missing at random. We present transparent general approaches that can be applied to all countries and multiple diseases.

Methods

We compared the performance of several imputation approaches using two foodborne diseases – congenital toxoplasmosis and aflatoxin-related hepatocellular carcinoma – for which some published incidence data were available. Congenital toxoplasmosis, caused by the protozoal parasite Toxoplasma gondii, may cause ocular and neurological disorders in the unborn child, possibly leading to stillbirth or neo-
Aflatoxin is a metabolite produced by the fungi *Aspergillus flavus* and *Aspergillus parasiticus* in maize and nuts, and is a known human liver carcinogen. See Appendix A (available at: http://www.cbra.be/publications/imputation-appendix.pdf) for further details regarding these two example data sets. The goal was to produce a complete set of incidence estimates for WHO Member States for the year 2005.

We compared the performance of various methods using leave-one-out cross-validation, because no external incidence estimates for the missing data were available. This allowed us to estimate the expected predictive accuracy of imputation methods in practice. We classified countries according to 17 food consumption clusters, using the Global Environment Monitoring System’s Food Monitoring and Assessment Programme. All except 19 countries had a Food consumption cluster, using the Global classified countries according to 17 food consumption clusters. Using the Global Environment Monitoring System’s Food Monitoring and Assessment Programme. We imputed the expected predictive accuracy of imputation methods in practice. We classified countries according to 17 food consumption clusters, using the Global Environment Monitoring System’s Food Monitoring and Assessment Programme. All except 19 countries had a Food consumption cluster, using the Global Environment Monitoring System’s Food Monitoring and Assessment Programme.

### Imputation approaches

The imputation methods included naive extrapolation and two hierarchical modelling approaches: (i) random effects regression models, and (ii) mixed-effects regression models. The rationale for specifying random effects models is that countries with missing data can “borrow strength” from other countries within the same cluster, and clusters with few or no data can “borrow strength” from the global population.

#### Method 1

Foodborne disease incidences among countries with similar food consumption patterns should be similar. Rather than fitting a statistical model to existing data, the naive extrapolation approach imputes missing incidences as the median of all other countries with data within the same food cluster. If data were missing for all countries within a cluster, the global median was assigned. Country-specific 95% prediction intervals were derived via bootstrapping with replacement. A thousand bootstrap samples of the number of countries with observed data were taken and an identical method applied to each sample. The unit of sampling was the country-level incidence.

#### Method 2

This Bayesian random effects regression model applies a single random effect serving as a clustering variable for different countries. We explored two possible clustering variables: food cluster (n = 17) and WHO subregion (n = 14). To normalize the distribution of the outcome variable, the observed incidences were log-transformed, leading to a log-normal regression model. Other approaches to deal with skewed data were considered such as a normal model with link or a gamma model, but none were found to differ meaningfully from the log-normal model. Equations for this model can be found in Appendix A.

Vague normal priors were specified for the random model coefficients and for hyper-parameters. Posterior distributions were derived through Markov chain Monte Carlo methods using the jags package in R software version 3.0.2 (R Foundation for Statistical Computing, Vienna, Austria). Two chains were initiated of 20,000 posterior samples each, with the first 15,000 discarded. Standard graphical indicators of non-convergence were checked. Predicted values were taken from the posterior predictive distribution. Example code is provided in Appendix A.

#### Method 3

This Bayesian mixed-effects regression model extends Method 2 by including disease-specific covariates, leading to a mixed-effects model. Again, the model was applied to the log-transformed observed incidences, and either food cluster or WHO subregion was included as a random effect. Method 3 relies on predictor variables measured at the national level. Our hypothesis is that covariates capturing the socioeconomic, food production-related and the public health and hygiene situation within a given country will be informative of between-country variation in the incidence of foodborne diseases. We describe below how a common set of potential covariates was derived from publicly available databases.

World development indicators, the data repository of world health statistics and the database of the Food and Agriculture Organization of the United Nations (FAOstat) were the starting point for defining an initial set of 1200 covariates. This was reduced to 194 variables related to food- or waterborne disease, food production and consumption, agriculture, environment, health, demography, economics and development. If the data point for the year 2005 was missing, subsequent years were searched (2006 to 2011 for world development indicators and world health statistics variables and 2006 to 2009 for FAOstat variables), and the nearest year with non-missing data was used. If two variables were highly correlated, one was removed, with the retained member chosen according to relevance as a general indicator (for instance, total mortality rate was chosen in preference to either male-only or female-only mortality.

### Table 1. Comparison of Three Methods for Imputing Missing Incidence Data for Congenital Toxoplasmosis, 2005

<table>
<thead>
<tr>
<th>Method</th>
<th>Mean absolute prediction error (95% CI)</th>
<th>Global incidence per 1000 live births (95% prediction interval)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method 1: median of other countries within same cluster</td>
<td>0.65 (0.55–0.75)</td>
<td>1.47 (1.45–1.49)</td>
</tr>
<tr>
<td>Method 2: Bayesian random effects regression</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Food cluster random effect</td>
<td>0.62 (0.54–0.72)</td>
<td>1.44 (1.38–1.58)</td>
</tr>
<tr>
<td>WHO subregion random effect</td>
<td>0.55 (0.47–0.63)</td>
<td>1.45 (1.39–1.61)</td>
</tr>
<tr>
<td>Method 3: Bayesian mixed effects regression</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Food cluster random effect</td>
<td>0.54 (0.47–0.61)</td>
<td>1.50 (1.42–1.72)</td>
</tr>
<tr>
<td>WHO subregion random effect</td>
<td>0.53 (0.46–0.60)</td>
<td>1.52 (1.42–1.76)</td>
</tr>
</tbody>
</table>

CI: confidence interval; WHO: World Health Organization.

Countries were clustered based on food consumption using the Global environment monitoring system – food contamination monitoring and assessment programme.

The covariate set consisted of percent arable land, percent urban population, annual precipitation, CO₂ emissions, rice supply, agricultural value added, neonatal mortality rate and fresh-water sources.

---

*DOI: 10.2471/BLT.14.139972*
To the remaining 112 variables, an arbitrary missingness threshold of 26% was applied to limit the number of missing values, reducing the data set to 65 variables. A final set of 51 variables resulted when non-numeric variables were removed (Appendix A). Any missing values in these remaining 51 variables were imputed using the mice package for R. One hundred imputed data sets were generated using the predictive mean matching method. This is useful for bounded variables such as proportions, since imputed values are sampled only from observed values. The simple mean of each set of 100 imputed data sets was calculated, implying that uncertainty in imputed values was ignored. Nine covariates were log-transformed following inspection of normal quantile-quantile plots.

The use of principal components analysis was explored to further reduce the set of 51 covariates. However, as this step did not increase model performance, it was excluded from the reported results.

### Analysis

Fixed effects were selected from the set of potential covariates (Appendix A) in a data-driven, stepwise manner. First, least absolute shrinkage and selection operator regression was used to select covariates with non-zero regression coefficients. We optimized the covariate estimates and model fit using the tuning parameter lambda. Least absolute shrinkage and selection operator models were fitted using the glmnet package for R.
Subsequently, backward-stepwise elimination based on Akaike’s information criterion further reduced the covariate set resulting from the least absolute shrinkage and selection operator step, yielding a subset of covariates that were significantly associated with disease incidence.\(^{23,25}\)

Model implementation was the same as for the random effects model (Method 2). As before, vague normal priors were specified for the model coefficients and hyper-parameters. Example JAGS code is provided in Appendix A.

**Predictive accuracy**

The accuracy of the model predictions was evaluated using leave-one-out cross-validation.\(^{26}\) Data for a single country were temporarily deleted and the remaining data used in an attempt to recover this deleted value. The procedure was repeated by holding out and predicting each country in turn. The mean absolute prediction error was computed as the prediction error averaged over all held-out countries. Ninety-five percent confidence intervals (CI) around mean absolute prediction error values were computed via bootstrapping methods; 10,000 samples with replacement were used, and the mean absolute prediction error was calculated for each sample separately.

To compare the predictive performance of the three models, we applied the Wilcoxon signed-ranks test to the paired absolute prediction errors obtained from each imputation method. We used Bonferroni’s correction for multiple comparisons.

**Effect of database size**

We estimated the effect of varying the number of observations in the database on the central estimate and 95% CI of the mean absolute prediction error. As an example, we used a mixed-effects regression model of toxoplasmosis. To simplify the analysis, a frequentist version was used, yielding virtually identical results. The database size was reduced from 115 to 15 countries in steps of five. Mean absolute prediction error values for each reduced database size and bootstrapped 95% CI were computed by taking 100 random samples of the specified size from the set of countries with data, and calculating mean absolute prediction error separately for each sample.
Results

**Congenital toxoplasmosis**

We obtained the toxoplasmosis incidence per 1000 live births for 118 countries, implying 74 countries without data. Table 1 shows the mean absolute prediction error for each method and variant. We compared the predictive accuracy between the three methods. For Method 2 and Method 3 we also assessed the effect of including each of the two possible clustering variables as random effect. Compared with Method 1, predictive accuracy was greater for both Method 3 variants ($P < 0.05$), while both Method 2 variants were statistically indistinguishable. In general, Method 3 generated larger prediction intervals (Table 1 and Appendix A). Eight covariates were retained for Method 3, based on an Akaike’s information criterion penalty threshold of six. These were: percent arable land, percent urban population, annual precipitation, CO$_2$ emissions, rice supply, agricultural value (as percentage of total gross domestic product), neonatal mortality rate and fresh-water sources.

Fig. 1 compares original and imputed incidence values, derived using Method 3, with WHO subregion as random effect, for the 118 countries with non-missing data. Imputed incidence values for the 74 countries with missing data, also derived using Method 3, are displayed in Fig. 2. A comparison of the three methods in terms of the global incidence per 1000 live births is provided in Table 1 and by WHO subregion in Appendix A.

**Aflatoxin-related hepatocellular carcinoma**

We obtained the incidence of hepatocellular carcinoma per 100 000 population for 33 countries, implying 159 countries without data. Table 2 shows the mean absolute prediction error for each method. The performances of both Method 2 and Method 3 variants were not statistically distinguishable from that of Method 1. As seen for toxoplasmosis, Method 3 generated larger prediction intervals (Table 2 and Appendix A). Three covariates were retained for Method 3, i.e. food supply from animal products, percentage of population subject to tuberculosis infection and energy use (defined as kg of oil equivalent per capita).

**Effect of database size**

We used Method 3 to impute missing values for toxoplasmosis after deleting some observations from the database. With decreasing database size, both the mean and the variability of the mean absolute prediction error tended to increase (Fig. 4).

**Discussion**

We compared the performance of a simple extrapolation method to imputation approaches using regression models. Variation in predictive accuracy across regression-based methods was small. For only one of the two datasets, the Bayesian mixed-effects regression models performed significantly better than the baseline Method 1. 

The data-driven approach to selection of covariates for the mixed-effects model retained eight covariates for toxoplasmosis and three for hepatocellular carcinoma. For toxoplasmosis, these included percent urban population and neonatal mortality rate, which are proxy variables for socioeconomic development and general population health. The retained covariates for hepatocellular carcinoma included food supply from animal products and energy use. Transmission of foodborne infection or the risk of contamination may be associated with non-health related variables which serve as proxies for variations in public health and hygiene between countries.

To quantitatively evaluate the performance of any prediction method, validation is necessary. For our example diseases, internal validation was the only feasible choice, because no external data sources were available. However, selection of imputation approaches should not be based exclusively on numerical criteria such as prediction error, but also on biological plausibility, computational considerations, measures of uncertainty and user-friendliness. Such criteria become even more relevant when dealing with sparse data sets. Imputed values should be reported with an assessment of their uncertainty and interpreted by at least one disease expert.

It seems intuitive to include biologically plausible covariates in the model, either selected through a data-driven approach or directly provided by expert opinion. Our results show that inclusion of disease-specific covariates may be associated with greater prediction error but increased predictive performance compared to naive Method 1. However, as evaluations were only based on internal validation, these results should be interpreted with caution. Indeed, better predictive performance may result merely from overfitting of available data.

**Table 2. Comparison of three methods for imputing missing incidence data for aflatoxin-related hepatocellular carcinoma, 2005**

<table>
<thead>
<tr>
<th>Method</th>
<th>Mean absolute prediction error (95% CI)</th>
<th>Global incidence per 100 000 population (95% prediction interval)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method 1: median of other countries within same cluster$^*$</td>
<td>1.27 (0.93–1.64)</td>
<td>1.13 (1.10–1.16)</td>
</tr>
<tr>
<td>Method 2: Bayesian random effects regression</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Food cluster random effect</td>
<td>1.24 (0.97–1.54)</td>
<td>1.00 (0.91–1.44)</td>
</tr>
<tr>
<td>WHO subregion random effect</td>
<td>1.23 (0.95–1.55)</td>
<td>1.05 (0.91–1.64)</td>
</tr>
<tr>
<td>Method 3: Bayesian mixed effects regression$^a$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Food cluster random effect</td>
<td>1.08 (0.87–1.31)</td>
<td>1.17 (0.94–3.77)</td>
</tr>
<tr>
<td>WHO subregion random effect</td>
<td>1.08 (0.85–1.32)</td>
<td>1.19 (0.94–3.50)</td>
</tr>
</tbody>
</table>

$^*$ CI: confidence interval; WHO: World Health Organization.

$^*$ Countries were clustered based on food consumption using the global environment monitoring system – food contamination monitoring and assessment programme.

$^a$ The covariate set consisted of food supply from animal products, percentage of population subject to tuberculosis infection and energy use.
especially with sparse data sets. Also, out-of-range predictions become more likely with decreasing dataset size. As a result, the inclusion of covariates can lead to unexpected results and should be appraised carefully by disease experts.

A high proportion of missing data might limit the application of these methods. Deleting observations from the toxoplasmosis database demonstrated the degree of bias and variability that may occur when applying imputation methods to small numbers of existing data points. This issue needs to be considered when presenting and interpreting health statistics that draw upon model-based predictions.

We applied the imputation techniques to published point estimates of disease incidence, although estimates of uncertainty were also provided with the original data sets. To fully represent the uncertainty in predictions, Monte Carlo sampling could be applied to take into account uncertainty in national-level incidence. However, for purposes of comparing the degree of bias across imputation methods, the point estimate was considered adequate.

The quality of any imputation model depends on the quality of available data and underlying study designs. In the current case, imputed incidences and estimated uncertainty depend on existing national-level data, while for the mixed-effects model, appropriate data on covariates were also required. If any of these indicators are unreliable, or have a non-monotonic relationship with disease incidence, then predictive abilities of models would be compromised.\textsuperscript{11} The degree of bias introduced by the preliminary multiple imputation step to fill in missing covariate values is not known.

If missingness and disease incidence are associated, the missing at random assumption is invalid. However, establishing such an association is difficult, since the necessary incidence data are not available (Appendix A). We were unable to estimate the degree of potential bias resulting from fitting models to disease incidence data from resource-rich countries to predict incidence rates for countries with fewer resources.

We have restricted our analysis to Bayesian methods. However, if uninformative priors are specified, Bayesian and frequentist statistical frameworks generate similar results. The advantage of Bayesian methods is that the within- and between-region variability can be modelled separately. Furthermore, informative priors for regression coefficients could be specified based on previous analyses of other foodborne diseases with similar incidence distributions. This can be particularly relevant for sparse data sets, with the caveat that in such settings the priors can strongly influence the results.

Selection of covariates required several assumptions, and thresholds for selection were not guided by evidence alone. Ideally, imputations should be done in a context-specific manner, with disease experts introducing specific covariates for which sufficient global data are available or eliminating certain possibilities depending, for instance, on food consumption habits. Although time-consuming, this would improve the biological plausibility of the models.
Finally, the least absolute shrinkage and selection operator method is known for instability with respect to the set of non-zero covariates retained. Slight changes in the data can result in very different sets of retained covariates. Ideally, covariate sets would be selected during each cross-validation run separately; however, this was impractical. We had to assume that the data are independent. Least absolute shrinkage and selection operator penalized regression methods ignore dependency, although recent developments try to accommodate for this. Nevertheless, least absolute shrinkage and selection operator methods are relatively robust to collinearity of covariates.

Conclusion

Imputation will never be a perfect substitute for actual data. We evaluated the predictive accuracy of various methods to impute missing national-level disease parameters. We described data-driven methods for reducing a large data set of socioeconomic, food production, and demographic indicator variables, which may be used to complement imputation models. Hierarchical models, specifying structural relationships between countries, can be a useful approach to the problem of estimating missing national incidence.

Acknowledgements

SAM, BD, NS, NP, PRT, AH, FW, and DD are members of the Foodborne Disease Burden Epidemiology Reference Group of the World Health Organization. BD is also affiliated with the Institute of Health and Society (IRSS), Université catholique de Louvain, Brussels, Belgium. NH is also affiliated with the Centre for Health Economic Research and Modelling Infectious Diseases, University of Antwerp, Antwerp, Belgium. AHH is also affiliated with the Institute for Risk Assessment Sciences, Utrecht University, Utrecht, the Netherlands.

Funding: This work was co-funded by grants awarded by the Special Research Fund of Ghent University and the National Institution of Health, Ruth L. Kirschstein National Research Service Award, Institutional Training Grant T32 RR023916 and T32 OD010423.

Competing interests: None declared.
Aim filled national-level incidence data-driven methods in global disease burden research

We compared two foodborne disease models to fill in missing national-level incidence data for 2005: congenital toxoplasmosis and aflatoxin-related hepatocellular carcinoma. We assumed that missing values are missing at random. We used least absolute shrinkage and selection operator (LASSO) regression to select predictive variables. We compared natural extrapolation methods and Bayesian random and mixed-effects regression models for their predictive performance. We used leave-one-out cross-validation to evaluate the model accuracy.

Results In one of the disease models, the Bayesian mixed-effect model had significantly better predictive accuracy than the naive extrapolation method. However, the Bayesian mixed-effect models produced wider predictive intervals for both data sets.

Conclusion There are several methods available for filling in missing national-level data. The advantages of a hierarchical regression approach for this task include deriving estimates from similar countries, transparency, computational efficiency, and ease of interpretation. Including informative covariates can improve model performance, but results should be carefully evaluated.
por medio de la contracción absoluta mínima y la regresión del operador de la selección. Se comparó el rendimiento predictivo de los enfoques de extrapolación ingena y los modelos bayesianos de regresión al azar y de efectos mixtos, y se empleó una validación cruzada dejando uno fuera para evaluar la exactitud del modelo.

**Resultados** La exactitud predictiva de los modelos bayesianos de efectos mixtos fue significativamente mejor que la del método de extrapolación ingena para uno de los dos modelos de enfermedad. Sin embargo, los modelos bayesianos de efectos mixtos generaron intervalos de predicción más amplios para ambos conjuntos de datos.

**Conclusión** Son numerosos los enfoques disponibles para la imputación de datos ausentes a nivel nacional. Los puntos fuertes de un enfoque de regresión jerárquica para este tipo de tareas son la capacidad de obtener estimaciones de otros países similares, la transparencia, la eficiencia informática y la facilidad de interpretación. La inclusión de covariables informativas puede mejorar el rendimiento del modelo, aunque se deben evaluar atentamente los resultados.

**Referencias**