CASM-Child:
A Bayesian Hierarchical Model for Multivariate Count Data to Estimate **C**ause- and **A**ge- **S**pecific **M**ortality in **Child**ren.

Austin E Schumacher[1], Tyler H McCormick[2,3], Jon Wakefield[1,2], Yue Chu[4], Jamie Perin[4], and Li Liu[4,5]

[1]Department of Biostatistics, School of Public Health, University of Washington
[2]Department of Statistics, University of Washington
[3]Department of Sociology, University of Washington
[4]Department of International Health, Johns Hopkins Bloomberg School of Public Health
[5]Department of Population, Family, and Reproductive Health, Johns Hopkins Bloomberg School of Public Health

September 19, 2018

Extended Abstract

Background

In 2015, an estimated 5.9 million children worldwide died before age 5 years.[1] As the US government and international community increase their investment in developing and implementing age-targeted, disease-specific childhood interventions and policy,[2–4] their effectiveness requires knowledge of the patterns of child deaths across multiple age groups and causes over time. The majority of these deaths, however, occur in low and middle-income countries (LMICs) without high quality vital registration (VR),[1] creating massive uncertainty about the cause-profiles of child deaths. Development of sample vital registration systems in LMICs is rapidly progressing, requiring statistical modeling developments to utilize this data source.[5]

Limited work in this area (i) uses broad age groups which mask important heterogeneity,[1] (ii) estimates all-cause and cause-specific mortality in two separate frameworks,[1,6,7] (iii) produces estimates separately and independently in each age group,[1,8] and/or (iv) does not account for empirically observed correlations between causes.[1,6]

Previous methods

The two main methods in the literature use a multinomial approach[1] or a squeezing approach.[7] The multinomial approach estimates cause-specific mortality fractions using multinomial regression. Then, these cause-fractions are multiplied to previously estimated all-cause mortality rates from a separate framework[8] to produce cause-specific mortality rates. This is done using a bootstrap-like approach in which all-cause mortality rates are sampled from the posterior distribution, the cause fractions are simulated from the modeling step, and these are combined at the draw level. The squeezing approach models all-cause mortality using a complex regression model and simulates draws from the final stage in order to produce uncertainty. Separately, each cause-specific mortality rate is estimated using an ensemble of regression models. Posterior samples are drawn, each draw is paired with a draw from the previously estimated all-cause mortality rates, and the cause-specific rates are proportionately "squeezed" to add up to the all-cause rate.

These methods have two main shortcomings. First, separate estimation of all-cause and cause-specific mortality actually reuses data in both stages, though this is not explicit, which leads to anti-conservative uncertainty estimation. Second, these models induce a negative correlation between cause-specific mortality due to the constraint that cause fractions must sum to 1. However, cause-specific correlations exist beyond this implied negative correlation, e.g. causes that share an environmental factor such as typhoid and diarrheal diseases which both relate to poor water sanitation. In sample registration data from Bangladesh in the years 2013, 2014, and 2016,[9] empirical covariances between cause-specific death counts range from -157 to 418, whereas a multinomial distribution parametrically estimates covariances ranging between -161 and 0.

To illustrate the severity of this problem, we simulate data with correlated cause-specific mortality rates via a multivariate lognormal distribution, simulate death counts with Poisson distributions, fit a multinomial model, and calculate the bias and uncertainty interval coverage. To set feasible parameters, we calculate the marginal means, variances, and covariances from Bangladesh sample registration data in years 2013, 2014, and 2016,[9] then translate these to the parameters of the lognormal distribution. We find that the estimates from the multinomial model are unbiased, however, the coverage of 95% CIs calculated using the delta method are too narrow (Figure 1).

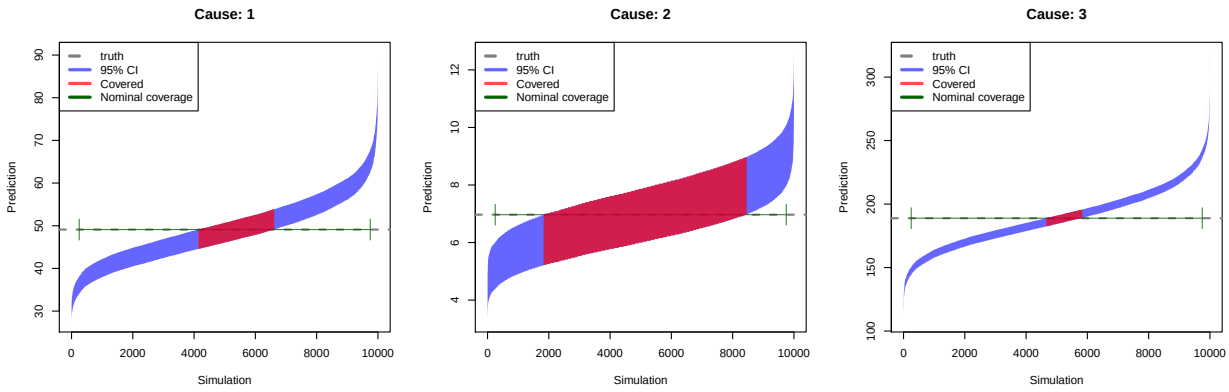

Figure 1: Coverage of 95% CI in each simulation for 3 causes. Variances and covariances are set according to values found empirically in Bangladesh sample registration data.

Theoretical justification for our model

We aim to model cause- and age-specific mortality rates over time. The natural foundation for this work is individual-level failure time analysis with competing risks proposed in Prentice (1978).[10] Suppose we have data $\mathbf{y}_i = \{y_{i1}, \ldots, y_{im}\}$, where y_{ij} is the indicator that individual $i \in \{1, \ldots, n\}$ dies of cause $j \in \{1, \ldots, m\}$, and t_i the time of death/observation. Defining $\boldsymbol{z}_i$ the vector of individual-level covariates and defining cause-specific mortality rates $\lambda_j(t_i; \boldsymbol{z}_i) = \lim_{\Delta t_i \to 0} P\big(t_i \leq T < t_i + \Delta t, J = j | T \geq t_i; \boldsymbol{z}_i\big)/\Delta t$, the data likelihood is

$$\mathcal{L} = \prod_{i=1}^{n} \prod_{j=1}^{m} \left([\lambda_j(t_i; \boldsymbol{z}_i)]^{y_{ij}} \exp\left[- \int_0^{t_i} \lambda_j(u; \boldsymbol{z}_i) du \right] \right).$$

Since data from sample registration systems is commonly only available in tabulated form, we combine the above likelihood with the work of Holford (1980)[11] and Laird and Olivier (1981)[12] for modeling tabulated single-cause survival data. Assume the time axis is partitioned into K time periods and assume individuals fall into H strata with common covariate vectors $\boldsymbol{z}_h$ within each strata. Suppose we now have data y_{hkj}, the number of individuals in strata h and time period k who die from cause j. We assume a constant hazard given $\boldsymbol{z}_h$, $\lambda_{kj}(\boldsymbol{z}_h)$, in each time period. Thus, we can rewrite and summarize the likelihood, with T_{hk} the total exposure time in strata h and time period k, as

$$\mathcal{L}^* = \prod_{h=1}^{H} \prod_{k=1}^{K} \prod_{j=1}^{m} \left[\big(\lambda_{kj}(\boldsymbol{z}_h)\big)^{y_{hkj}} \exp\big(- T_{hk} \lambda_{kj}(\boldsymbol{z}_h)\big) \right].$$

The likelihood factors into a component for each cause j, and the cause-specific likelihoods are identical to the ones in a single-cause survival model. The kernel is identical to that which arises if each $y_{hkj} \sim \text{Poisson}\big(\lambda_{kj}(\boldsymbol{z}_h) T_{hk}\big)$, so we can perform likelihood-based inference using Poisson distributions for each cause where we treat the number of deaths due to each cause as the outcome and the exposure to all causes as the offset. We extend this to age-time-cause tabulations by partitioning the age-time Lexis surface into K tabulation groups and assuming constant hazard in these groups. The above likelihood holds with k now indexing age-time tabulations.

Model description

We propose a novel Bayesian hierarchical model that (1) simultaneously and coherently estimates all-cause and cause-specific under-five mortality rates across multiple age groups in one framework; (2) propagates uncertainty considering correlations across causes; and (3) provides full transparency and reproducibility.

Statistical modeling based on our theoretical justification provides the flexibility to choose a model for the cause-specific mortality rates that is driven by specific data. As a working example, we assume proportional constant cause-specific hazards, which is common in the literature,[11,12] in each age-time group such that

$$\lambda_{kj}(\boldsymbol{z}_h) = \bar{\lambda}_{kj} \exp(\mathbf{Z}_h \boldsymbol{\theta}).$$

$\mathbf{Z}_h$ is a model matrix for each subgroup and $\boldsymbol{\theta}$ a vector of corresponding covariates. To account for variation in age, time, and cause, we model the cause-specific hazards as

$$\log(\bar{\lambda}_{kj}) = \mathbf{X}_k \boldsymbol{\gamma} + b_{jk}, \quad \mathbf{b}_k \sim \text{Normal}(\mathbf{0}, \boldsymbol{\Sigma}).$$

$\mathbf{X}_k$ is a model matrix for each age-time group, $\boldsymbol{\gamma}$ is a vector of corresponding covariates, and b_{jk} is an intercept for each age-time-cause. To account for correlations between causes, we specify a multivariate Normal distribution on the age-time-specific vector $\mathbf{b}_k = \{b_{1k}, \ldots, b_{mk}\}$, similar to that used in Chib and Winkelmann (2001).[13] This formulation can be simplified by combining $\mathbf{X}_h$ and $\mathbf{Z}_k$ into one matrix and $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ into one vector of covariates, leading to

$$y_{khj} \sim \text{Poisson}\big(\exp(\tilde{\mathbf{X}}_{hk}\boldsymbol{\beta} + b_{jk})T_{kh}\big).$$

This is one choice of mean model. The flexibility of our modeling framework allows for more specialized models as data allow.

We fit this as a Bayesian hierarchical model, parameterizing $\boldsymbol{\Sigma}$ by its standard deviation and correlation as $\sigma\Omega\sigma'$ and placing an LKJ(1) prior[14] on Ω, diffuse half-Cauchy$(0, 5)$ priors on σ, and diffuse Normal$(0, 10)$ priors on β. We can choose informative priors depending on available information. The model is fit using `STAN`[15] interfaced with `R` version 3.5.0 using the `rstan` package version 2.17.3 for fast, accurate Bayesian inference based on Hamiltonian Monte Carlo.[16]

Simulation study

As a simulation study to examine properties of the model, we fit it to 1000 data sets generated from the above model and calculate bias and coverage of 95% credible intervals. We do this with a simple example with $m = 2$ causes, $K = 25$ age-time groups, $H = 200$ strata, $\beta = -1$, $\sigma = (1, 2)$, and $\Omega = \begin{bmatrix} 1 & 0.2 \\ 0.2 & 1 \end{bmatrix}$. We then perform a complex simulation study with parameter dimensionality equal to that found in sample registration data from China[17] described below and parameter values set to various scenarios in plausible ranges from the observed data.

In the simple simulation, we find that parameter estimates are unbiased with nominal coverage of the 95% credible intervals for all parameters and also for the predicted age-time-cause-strata-specific mortality rates. In the complex simulation study, we expect results to remain unbiased with coverage of credible intervals less than 95% due to high dimensionality of age-time groups and low number of strata.

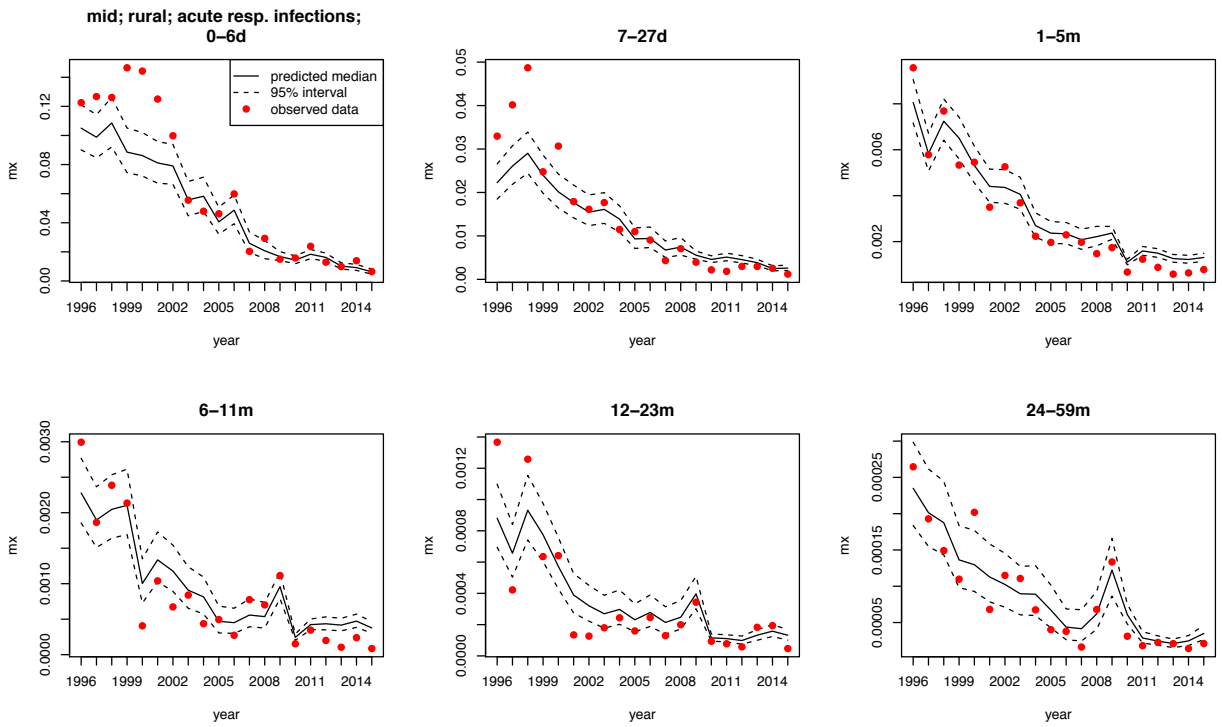

Figure 2: Predicted mortality rate with 95% credible interval over time for acute repiratory infections in rural middle region vs. observed mortality rates.

CASM-Child in China

We fit our model to tabulated death counts and exposure time from China's sample registration system for single years between 1996-2015, six age groups (0-6d, 7-27d, 1-5m, 6-11m, 12-23m, 24-59m), eight mutually exclusive and collectively exhaustive causes of death, and six strata (three regions split into urban/rural). This data is described in detail in He et. al. (2017).[17]

We fit the above model with the model matrix having a linear time trend, an indicator for each age group, an indicator for urban/rural, and an indicator for region. This simple mean model is a proof of concept that will be expanded upon in future work.

Figure 2 shows results for acute respiratory infections in China's rural middle region. Standard diagnostic plots (trace, $\hat{R}$, effective sample size, and acceptance probabilities) showed no issues. Predicted mortality rates exhibit both positive and negative bias in certain time periods for some causes. More work on smoothing the mean model is necessary as year-to-year fluctuations can be extreme, though estimates are generally plausible. These results are quite promising for a simple mean model and covariates. Further work will develop a sophisticated mean model to estimate age and time trends more accurately. Possible routes include splines[6] or singular value decomposition.[18] We will also explore copula modeling[19] as a potential way to model cause-correlations more efficiently and flexibly.

References

[1] Li Liu, Shefali Oza, Dan Hogan, Yue Chu, Jamie Perin, Jun Zhu, Joy E Lawn, Simon Cousens, Colin Mathers, and Robert E Black. 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(10063):3027–3035, 2016.

[2] Roger I Glass, Alan E Guttmacher, and Robert E Black. Ending preventable child death in a generation. *Jama*, 308(2):141–142, 2012.

[3] John J Aponte, David Schellenberg, Andrea Egan, Alasdair Breckenridge, Ilona Carneiro, Julia Critchley, Ina Danquah, Alexander Dodoo, Robin Kobbe, Bertrand Lell, et al. Efficacy and safety of intermittent preventive treatment with sulfadoxine-pyrimethamine for malaria in african infants: a pooled analysis of six randomised, placebo-controlled trials. *The Lancet*, 374(9700):1533–1542, 2009.

[4] Melissa A Penny, Robert Verity, Caitlin A Bever, Christophe Sauboin, Katya Galactionova, Stefan Flasche, Michael T White, Edward A Wenger, Nicolas Van de Velde, Peter Pemberton-Ross, et al. Public health impact and cost-effectiveness of the rts, s/as01 malaria vaccine: a systematic comparison of predictions from four mathematical models. *The Lancet*, 387(10016):367–375, 2016.

[5] Carla AbouZahr, Don De Savigny, Lene Mikkelsen, Philip W Setel, Rafael Lozano, and Alan D Lopez. Towards universal civil registration and vital statistics systems: the time is now. *The Lancet*, 386(10001):1407–1418, 2015.

[6] Leontine Alkema and Jin Rou New. Global estimation of child mortality using a bayesian b-spline bias-reduction model. *The Annals of Applied Statistics*, pages 2122–2149, 2014.

[7] Mohsen Naghavi, Amanuel Alemu Abajobir, Cristiana Abbafati, Kaja M Abbas, Foad Abd-Allah, Semaw Ferede Abera, Victor Aboyans, Olatunji Adetokunboh, Ashkan Afshin, Anurag Agrawal, et al. Global, regional, and national age-sex specific mortality for 264 causes of death, 1980–2016: a systematic analysis for the global burden of disease study 2016. *The Lancet*, 390(10100):1151–1210, 2017.

[8] Danzhen You, Lucia Hug, Simon Ejdemyr, Priscila Idele, Daniel Hogan, Colin Mathers, Patrick Gerland, Jin Rou New, Leontine Alkema, et al. Global, regional, and national levels and trends in under-5 mortality between 1990 and 2015, with scenario-based projections to 2030: a systematic analysis by the un inter-agency group for child mortality estimation. *The Lancet*, 386(10010):2275–2286, 2015.

[9] Bangladesh Bureau of Statistics. Report on Bangladesh vital statistics 2016. Technical report, Bangladesh Bureau of Statistics, 2017.

[10] Ross L Prentice, John D Kalbfleisch, Arthur V Peterson Jr, Nancy Flournoy, Vern T Farewell, and Norman E Breslow. The analysis of failure times in the presence of competing risks. *Biometrics*, pages 541–554, 1978.

[11] Theodore R Holford. The analysis of rates and of survivorship using log-linear models. *Biometrics*, pages 299–305, 1980.

[12] Nan Laird and Donald Olivier. Covariance analysis of censored survival data using log-linear analysis techniques. *Journal of the American Statistical Association*, 76(374):231–240, 1981.

[13] Siddhartha Chib and Rainer Winkelmann. Markov chain monte carlo analysis of correlated count data. *Journal of Business & Economic Statistics*, 19(4):428–435, 2001.

[14] Daniel Lewandowski, Dorota Kurowicka, and Harry Joe. Generating random correlation matrices based on vines and extended onion method. *Journal of multivariate analysis*, 100(9):1989–2001, 2009.

[15] Bob Carpenter, Andrew Gelman, Matthew D Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan: A probabilistic programming language. *Journal of statistical software*, 76(1), 2017.

[16] Michael Betancourt and Mark Girolami. Hamiltonian monte carlo for hierarchical models. *Current trends in Bayesian methodology with applications*, 79:30, 2015.

[17] Chunhua He, Li Liu, Yue Chu, Jamie Perin, Li Dai, Xiaohong Li, Lei Miao, Leni Kang, Qi Li, Robert Scherpbier, et al. National and subnational all-cause and cause-specific child mortality in china, 1996–2015: a systematic analysis with implications for the sustainable development goals. *The Lancet Global Health*, 5(2):e186–e197, 2017.

[18] David J Sharrow, Samuel J Clark, and Adrian E Raftery. Modeling age-specific mortality for countries with generalized hiv epidemics. *PloS one*, 9(5):e96447, 2014.

[19] Joanna H Shih and Thomas A Louis. Inferences on the association parameter in copula models for bivariate survival data. *Biometrics*, pages 1384–1399, 1995.