METHODS PROTOCOL FOR THE UNITED STATES MORTALITY COUNTY DATABASE

Celeste Winant, Monica Alexander, Ameer Dharamshi, Denys Dukhovnov, and Magali Barbieri
with significant contributions by Carl Boe, Ethan Roubenoff, and Rae Willis-Conger
September 28, 2021

Table of contents

INTRODUCTION

The United States Mortality DataBase (USMDB) is developed at the University of California, Berkeley (United States) as part of the larger umbrella project of the Human Mortality Database (HMD), a collaboration between Berkeley, the Max Planck Institute for Demographic Research

(MPIDR, Rostock, Germany), and the French Institute for Demographic Studies (INED, Paris, France). The specific purpose of the USMDB is to provide researchers with easy access to detailed and comparable sub-regional mortality data through its website (at usa.mortality.org). The database contains original period life tables for US geographic entities that are updated every year, when new demographic data become available. Life tables are provided for years since 1959 for the 4 Census Regions, the 9 Census Divisions, and the 50 states plus the District of Columbia using the methods of the HMD. The USMDB also provides a period lifetable series at the county level for years since 1982 using a different approach, rendered necessary because county populations are too small to implement the HMD methods in a meaningful way. This Methods Protocol describes the statistical approach used to construct the county-level life tables. The approach used to construct the life table series at the higher geographic levels is found on the HMD website (at https://www.mortality.org/Public/Docs/MethodsProtocol.pdf).

The USMDB was initially funded by a grant from the National Institute of Aging at the National Institutes of Health (NIA-NIH grant #R01-AG040245, PI: Pr. John Wilmoth) and later supported by funding from the University of California, Berkeley Center on the Economics and Demography of Aging (CEDA), the Society of Actuaries and by another small grant from the NIH (#R03-AG058110).

OVERVIEW

In seeking to estimate underlying age-specific mortality rates for all US counties, we determined that direct estimation is insufficient. The main difficulty is the large number of counties with small populations in which the stochastic variation in death counts is high. For example, 10% (about 300) of US counties have populations below 5,000 and 1% below 1,000. The resulting mortality rates in these small areas are often highly erratic and may have many zero death count cells. When classic demographic methods are applied in this context, the true underlying mortality schedule is unclear and uncertainty is high, with no information on the range of probable outcomes. We thus sought to develop a method based on Bayesian inference for estimating subnational mortality rates and their confidence bounds across geographic areas with a wide variety of population sizes and death counts. We are thereby hoping to foster research on internal variations in mortality, to facilitate monitoring of current, and guide future, policy efforts to improve the health of the population, and to promote investigation in the historical effect of health intervention and changes in the structure of local health programs.

The basic idea behind our chosen statistical approach is to fit observed deaths and population counts to a model of the underlying mortality rate. The model builds on characteristic age patterns in mortality curves, which are constructed using principal components from a set of reference mortality curves. These principal components create an underlying structure of the model in which many different kinds of shapes of mortality curves can be expressed. Within a Bayesian hierarchical framework, information on mortality rates is then pooled across geographic space and smoothed over time. Geographic pooling allows mortality in smaller areas, where the underlying pattern are unclear, to be partially informed by mortality patterns in larger areas. The temporal smoothing component ensures the parameters governing the shape of the mortality curve change gradually and in a relatively regular pattern over time. We add a constraint to the model which ensures that county-level mortality rates, when aggregated to the state or national level, are consistent with the mortality rates observed at the aggregate level. The model was fitted in a Bayesian framework, with samples from the posterior distributions of the parameters obtained via a Markov Chain Monte Carlo (MCMC) algorithm. The original version of the model is fully described in our *Demography* article (Alexander, Zagheni and Barbieri , 2017). This initial version, which yielded estimates for both sexes combined, has been modified to account for variations between the male and female mortality schedules within each county.

Because we found that the Bayesian model operates best on populations of at least 10,000 (around 5,000 men and 5,000 women), we grouped the smallest counties within each state with the most demographically similar counties within the same state to reach this threshold, as explained in more details below (see "County grouping" section).

PREPARING THE INPUTS

County-level mortality estimation requires two main sets of data inputs to estimate underlying mortality rates (see Appendix A for the detailed description of sources and links to providers). We require both death tabulations and mid-year population estimates by county of residence (and state), year, sex, and five-year age-groups up to the highest age possible with a separate 0-1 age group for infants.

The county of residence for each state in the National Center for Health Statistics (NCHS) detailed natality and mortality data is labeled by static FIPS code and/or numerically by NCHS in alphabetical order according to the concurrent roster of counties at the time. FIPS geographic codes were only added to the records starting in 1982. The roster of counties in the NCHS data typically changes three or four years after the Census year. If new counties are added or old counties are removed from a state, then the labeling changes. The FIPS codes for US counties are static and do not change when counties are added or removed. The five boroughs of New York City, each a unique FIPS county, are given the same NCHS county code but unique NCHS city codes. These city codes are also available in the 1981 detailed mortality/natality records. The combined 1981 NCHS county and city codes can be translated to the static FIPS county code using translation programs found in the public domain that apply to this time period. We also spent some time reconciling the NCHS and the Census county identification schemes to create compatible death and population tabulations.

Deaths

The National Center for Health Statistics (NCHS) publishes detailed mortality records for each year included in our series (1982-2019) from which we can compile the mortality tabulations. All required variables are available in the public data for years 1982-2003. Geographic identifiers are suppressed in the public data for years 2004 onward. For purposes of the present analysis, we obtained a restricted dataset for years since 1989 that includes geographic identifiers (down to the county level) for this period. The data were provided by the NCHS through a specific Data User Agreement (DUA). As per the DUA, we are not authorized to release the raw data and must proceed with all analyses (i.e. our Bayesian framework to estimate county-level mortality) on a special restricted-access National Association for Public Health Statistics (NAPHSIS) data server. We downloaded the publicly-available detailed mortality records for 1982-1988 from NCHS and moved these onto our NAPHSIS server.

The tabulations for each set of mortality records (public and restricted) are identical. Variable formats and labels were first unified to match the HMD/USMDB standards for age, sex, year, and geographic identifier (two-letter state acronym and 5-digit FIPS county code). Non-resident records were removed. Age is only provided up to an open interval at 85 years. We combined about one third (1,084) of all US counties into 401 county aggregates to (a) guarantee historical consistency of all geographic areas over the study period (accounting for the splitting or merging of counties over the time period) and (b) obtain a set of areas with populations of 10,000 or greater over the entire period as described next.

County grouping

Counties that at any point between 1982-2019 fell below the threshold were aggregated consistently. We developed a partially automated set of decision rules that were endogenous to the population and geography to minimize any introduced variability. We used three criteria to determine how small counties should be aggregated: total dependency ratio (the ratio of the population 15 to 65 to the population below and above the age bounds, providing a single indicator of the age structure), population density (population per square kilometer), and proportion working in the agriculture sector (to obtain a measure of rurality as research as indicated differences in the mortality pattern of rural versus urban areas). Moreover, we decided that aggregate could not cross state borders and counties aggregated together must be rook contiguous (sharing a border, not just a point) in order for the units to be geographically meaningful.

For small counties, a Mahalanobis distance (Mahalanobis, 1936) was calculated for each adjacent county. Mahalanobis distance, or generalized distance, is a measure of distance for multiple measures simultaneously and is defined as:

$$d(\vec{x},\vec{y}) = \sqrt{(\vec{x}-\vec{y})^{T}S^{-1}(\vec{x}-\vec{y})}$$

where $\vec{x}$ and $\vec{y}$ are vectors of corresponding covariates, S is the covariance between the covariates, T indicates transpose. The generalized distance can more or less be thought of as a form of Euclidian distance that accounts for covariance between variables. That is, a one-unit increase in one variable may have a substantially different meaning that a one-unit increase in another. Mahalanobis distance is used to account for this covariance. For a small county, the adjacent county with the smallest Mahalanobis distance is recommended as the 'most similar' county.

We used data from the American Community Survey (ACS) from which we extracted the *Industry by Sex and Median Earnings, 2017 5 year* table (at https://www.census.gov/programs-surveys/acs). This yielded the percentage employed for each county, and the percentage employed in agriculture. Since we were looking at a wide time period, we also downloaded data for the year 2000 and created a variable for employment in that year.

The code protocol was as follows:

1. Read in population files, split by age, sex and year in each county county
2. Combine age and sex groups to generate a dataframe of county-year population totals; note if the county-year is under the 10,000 threshold
3. Load a shapefile of US counties
4. Flag counties that are ever under the 10,000 threshold to be combined
5. Calculate population density, dependency ratio, and percent in agriculture from the US county shapefiles joined with the population dataframe.
6. Generate a sparse adjacency matrix for small counties
7. Join counties with relevant controls , including geographic contiguity
8. Generate Mahalanobis distance for each small county as described above
9. Loop through each state and generate a map of counties for each state.
10. Check the results visually and make informed determination about which counties to aggregate together.
11. Combine counties according to user determination and produce a properly-combined dataframe and shapefile for future usage.
12. Rerun step 2 to confirm that all counties and county aggregates meet the 10,000 population threshold.

All that this program did was *suggest* the best possible combination; note that it did not automatically combine the counties. This was deliberate for a few reasons. First, combinations are not necessarily commutative. County A is recommended to combine with county B, but county B is recommended to combine with county C, etc. An algorithm that chains recommendations together could never converge in a smart way, in for instance the case of Nebraska or Texas, states with a large number of small counties. Second, imagine say that both small county A and small county B are recommended to combine with large county C. It makes more sense to combine A and B than all three. Third, it is easier for a human to make these decisions than a computer, even if it takes much longer.

A comma-separated variable lookup table linking the 1084 individual counties (by FIPS code) to the 401 groups can be found at usa.mortality.org/counties.php. The individual counties in the life table files are listed by FIPS code.

Tabulation

Unlike with the companion Human Mortality Database (HMD) or United States Mortality DataBase (USMDB) State-level data series, we did not need to tabulate deaths by Lexis Triangle. The tabulation by age simply requires relabeling and aggregating the deaths by county (and/or county-grouping), year, sex, and 5-year age groups (0, 1-4, 5-9, ..., 85+ years), setting aside a temporary category for deaths of unknown age by sex/county/year. Deaths of unknown age were then redistributed proportionately over the deaths by known age within a given county/sex/year following the same method as for the HMD. We created a separate input data file for each state.

Population

The Census publishes mid-year (July 1st) population estimates by county of residence, year, sex, and age. These data are available to the public through CDC Wonder. We applied the same county-level groupings as we did for the mortality data.

The age-detail varies from decade to decade. For years since 1990, the detail suffices to create the 5-yr age groups defined above. However, for years 1982-1989, the census grouped infants together with children through age 5. For accurate estimation of infant mortality, we need to disaggregate this first age group in order to separate infants from 1-4 years old children.

We reconstructed mid-year infant populations for these years following a number of steps. We defined a Lexis surface with periods centering on the mid-year. We estimated the July 1st infant population for a given year, Y, by subtracting the lower-triangle infant deaths over the mid-year span, July 1, Y-1 to June 30, Y, from the births spanning that same time period. Detailed mortality and natality data published by the NCHS from years 1981 – 1989 were necessary to create these tabulations of infant deaths and births. Fortunately, the level of detail needed is available in the same publicly available data used for the mortality tabulations described above. We just had to include 1981 data in the analysis and to tabulate the deaths by Lexis triangle.

Once the detailed natality and mortality records for 1981-1989 were labeled with the same standard set of FIPS county and state codes, we applied the same county grouping described above. Detailed natality records are provided by month and year of birth. We tabulated births for mid-year Y-1 aggregating over July 1, Y-1 through June 30, Y.

Mortality records included detailed ages at death. The information provided for a given infant's age at death is determined by the magnitude of each lifespan. For example, if an infant's lifespan is greater than one month but less than one year, the age is recorded in units of month. If the lifespan is greater than one hour but less than 24 hours, the age is recorded in units of hours. The mortality records for these years also provide some detail (month/year) or complete detail (day/month/year) on the date of death. Andreeva describes the method used for the USA HMD to constrain the probability of death for a given record in the respective upper and lower triangle for a given period and age based on the constraints of lifespan and date of death (see https://www.mortality.org/hmd/USA/InputDB/USAcom.pdf, Appendix 2). We used the same method to estimate the number of lower-triangle infant deaths that occurred in the mid-year period spanning July 1, Y-1 through June 30, Y.

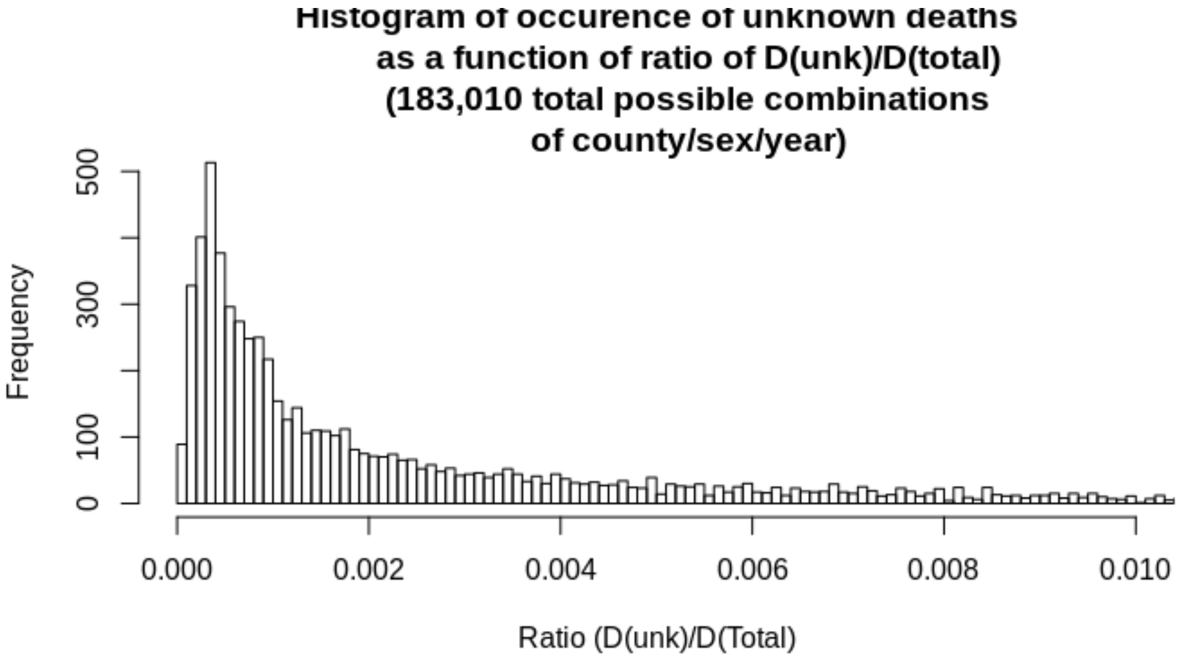
Using these two measures of mid-year natality and lower-triangle infant deaths, we estimated the mid-year infant population and resulting age 1-4 estimates from the official 0-4 estimates for the years where such detail is missing.

Diagnostic checks on input data

Prevalence of deaths of unknown age

Deaths of unknown age were present in 3.66% of all possible combinations of counties, sex and years. The range of recorded deaths by unknown age was 1 to 113. The maximum was found in Los Angeles County (FIPS 06037) in 1986 for males. It represented 0.37% of the total male deaths in that county for that year. For these instances of recorded deaths of unknown age, the median and modal number of deaths by unknown age was 1. The mean number of deaths was 2.13. The median ratio of unknown deaths to total deaths in any given county by year and sex was 0.12%. The mean ratio was 0.3% (Figure 1).

Figure 1



Presented below are qualitative checks looking specifically at our method to reconstruct the infant population by county for years (1982-1989) when they are missing from the official population estimates.

Verifying imputation of infant deaths

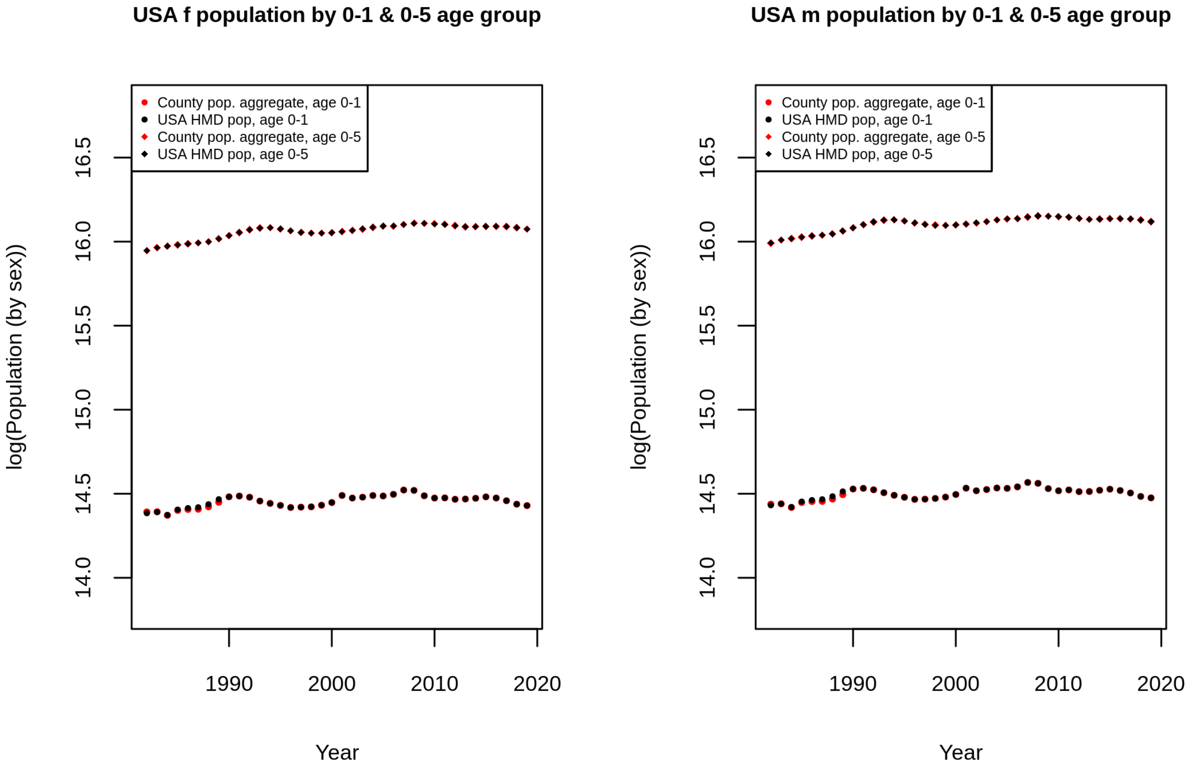
We tabulated county-specific infant deaths by Lexis triangle (with mid-year vertices) using multiple cause of death mortality records for years 1981 -1989, as described above. We compared the national-aggregate of these tabulations by sex and year with USA infant deaths by triangle from the Human Mortality Database (HMD). We cannot carry out a strict comparison since the periods for the HMD tabulations start on January 1st and not on July 1st as is the case for the county-level tabulations. However, we can gauge visually how well the respective tabulations overlap. The comparison of the national-aggregate of county-level infant deaths (red) and the USA HMD deaths (black) by triangle (TL = lower triangle, TU = upper triangle) and the sum of these two triangles (RR = rectangle = TU+TL) are shown below in Figure 2 for USA females (left figure) and males (right figure). The points are positioned time-wise (on the horizontal access) with the starting date of the year-long aggregate. E.g. a point at 1982.5 corresponds to a cell spanning July 1, 1982 to June 30, 1983. A point at 1986.0 corresponds to a cell spanning Jan 1 – Dec 31, 1986.

For both sexes, the national aggregate of county-level infant deaths for the sum of (TL+TU) agrees well with the USA total. For females, there seems to be a systematic underestimation of TU (and corresponding overestimate of TL) of ~200-400 deaths over the period of comparison. For males, the difference is ~300-500 deaths. This corresponds to a ~20-25% difference in estimates of TU, but more like 1-5% for TL. We will factor this uncertainty in our discussion of the final results– since we are thus over-estimating the number of years lived in the interval by those who died, this leads to an under-estimate of mortality – this can wait though].

These estimates of lower-triangle infant mortality factor into our construction of July 1st infant population for years 1982-1989. We can similarly compare the national-aggregate of county-level infant populations (both those which we construct for years 1982-1989 and those tabulated from official estimates for 1990-2019) with infant July 1st populations from the USA HMD series. Note that here we can do a strict comparison.

Shown below in Figure 2 are the 0-1 and 0-4 population estimates for infant populations, female (left figure) and male (right figure). The national aggregate of the county-level estimates are shown in red and the HMD USA estimates in black. On a log scale, the two sets of estimates overlap very well.

Figure 2



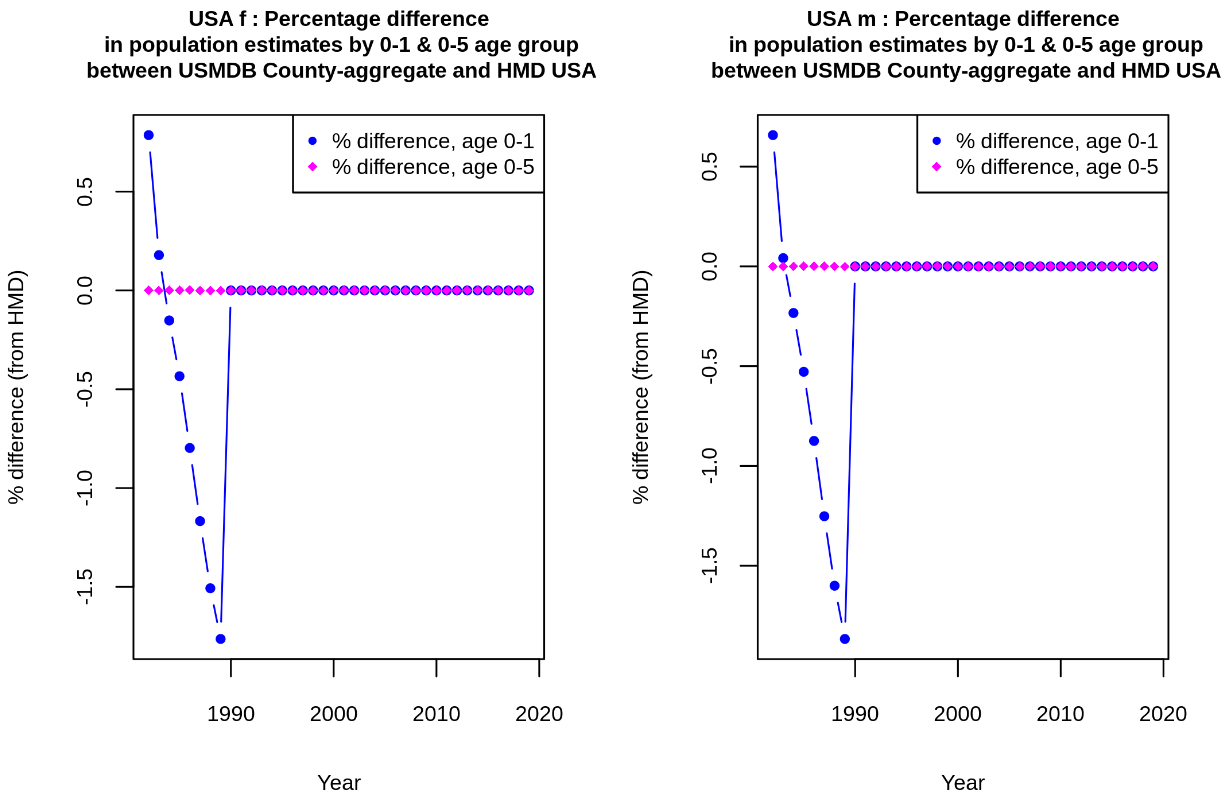
We can look closer at the differences with the plot shown below. The percentage difference is defined as

$$\text{Diff} = 100\% * (P_{agg.county}(x) - P_{USAHMD}(x)) / P_{USAHMD}(x)$$

Where x is either the age-group 0-1 or 0-4 years, P_{USAHMD} is the population in the corresponding age group for the US as a whole and $P_{agg.county}$ the population in the corresponding age group for the aggregate of all county values.

There is no quantitative difference between the two sets of estimates both for ages 0-1 and 1-4 for years after 1989. Our reconstruction of the 0-1 population for years 1982 – 1989 yields a net -1.9% to +.8% difference which shifts monotonically with time. The increase in magnitude of this difference from years 1894-1989 does not correspond with the trends seen in the differences of lower-triangle infant deaths.

Figure 3


USA f : Percentage difference in population estimates by 0-1 & 0-5 age group between USMDB County-aggregate and HMD USA

USA m : Percentage difference in population estimates by 0-1 & 0-5 age group between USMDB County-aggregate and HMD USA

Bayesian Estimation of County-Level Mortality Rates

This section presents a summary and discussion of the current iteration of the joint model for county-level mortality estimation in the US. The Bayesian model used to estimate county-level mortality is described in detail in Alexander, Zagheni, and Barbieri, 2017. The core of the model remains intact and changes focus on fine tuning the edges of the model to recognize patterns in the county-year-age level error distributions to account more specifically for separate mortality patterns by sex.

Model Summary

We begin by defining $y_{\text{x,s,a,t}}$ as the observed number of deaths in age bucket x, sex s, in county a, and in year t. Then,

$$y_{x,s,a,t} = \text{Poisson}\left(\lambda_{\text{x,s,a,t}} = m_{\text{x,s,a,t}} \times P_{x,s,a,t}\right)$$

where $P_{x,s,a,t}$ is the population corresponding to age bucket x, sex s, county a, and in year t, and $m_{x,s,a,t}$ is the mortality rate to be estimated for age bucket x, sex s, county a, and in year t.

The mortality rates are estimated on the log scale using principal components generated by the singular value decomposition of state level mortality rates in the US as follows:

$$\log\left(m_{x,s,a,t}\right) = \beta_{1s,a,t} \cdot Y_{1s} + \beta_{2s,a,t} \cdot Y_{2s} + \beta_{3s,a,t} \cdot Y_{3s} + u_{x,s,a,t}$$

where $Y_{i,s}$ is the i^{th} principal component, β is the estimated coefficients, and u is the error. Here we continue to use 3 principal components though we acknowledge that the results of other analyses indicate that increasing the number from 3 to 5 may yield a more flexible model though that comes with a fairly significant computational burden. For the current work we will leave it at 3 to avoid introducing an additional layer of sensitivity.

The sets of β coefficients for each sex-county-year are pooled geographically such that

$$\beta_{i,\mathrm{s},\mathrm{a},t} \sim N\left(\mu_{\beta_{i,s,t}}, \sigma^2_{\beta_{i,t}}\right), i = 1, 2, 3$$

$$\mu_{\beta_{i,s,t}} \sim N\left(2 \cdot \mu_{\beta_{i,s,,t-1}} - \mu_{\beta_{i,s,,t-2}},\ \sigma^2_{\mu_{\beta_i}}\right)$$

$$\sigma_{\beta_{i,t}} \sim \mathcal{LN}(-1.5, 0.5)$$

$$\sigma_{\mu_{\beta_i}} \sim \mathcal{N}(0, 1)$$

This setup generates β coefficients for each county from a common state-level distribution. Further, at the state-level, we smooth over time at the second differences level to produce gradual changes instead of erratic behavior.

The error distribution that completes the log-mortality rate estimation is generated in a bivariate setup as follows:

$$\left(\frac{u_{x,1,a,t}}{u_{x,2,a,t}}\right) \sim \left(O_2, \sigma_x\ 1_2\ L_{x,t} L^T_{x,t} 1_2 \sigma_x\ \right)$$

$$\sigma_a \sim \mathcal{N}(0, 0.25)$$

$$L_{x,t} \sim LKJ_{cholesky}(1)$$

The above setup captures the relationship between male/female errors using an age-year correlation matrix that is assigned the uninformative LKJ(1) prior. This allows for the tracking of correlations in errors over time and across age buckets. Given that the log-mortality values for different age buckets occurs in substantially different parts of the log curve, we add age specific error scaling to recognize the differences in the log distortion.

An intuitive way of thinking about this is that the principal components produce the expected mortality derived from state and local patterns while the u errors are a sort of "surprise" mortality. These "surprise" deaths depart from the expectation in an often correlated matter.

Summary of Changes

The model described above departs from the original model motivated and described in "A Flexible Bayesian Model for Estimating Subnational Mortality" (Alexander et al, 2017). We note that the core of the model is robust and does not need to be modified and instead focused on features at the edge of the model that were open to minor improvements:

- The principal components used here are derived from the stacked male and female state-level mortality data as opposed to a combined (or totaled) state-level mortality data. This change was made to allow the principal components to capture sex-specific differences in mortality.

- The primary change between the original model and the joint model is the error setup (the u parameter). If u is left to be an independent variable, it can flexibly adapt to any observation as there is a one-to-one correspondence in $u_{x,s,a,t}$ terms and observations. Flexibility is certainly a positive quality but it must be balanced with concerns of overfitting the data. In this version, we anchor the errors in the two sexes for a given age-county-year together such that we avoid capturing all noise and instead focus on picking up consistent patterns. As a concrete example, if we were to observe zero deaths in a given age-sex-county-year, the error term is incentivized to be as negative as possible to pull the mortality rate to zero. However, the joint version is less susceptible to this as now two values will need to be pulled down together which is often not consistent with the data as there are far more zero-values in the female data versus the male data. We find that for many age groups, there are substantial correlations between the error terms thus validating the use of a more structured approach here.

- By adding unique error scaling for each age bucket, we now adapt to the differences in the log curve's stretching effect.

- Some of the scaling parameters have had minor adjustments to their prior distributions to reflect differences in input scaling as well as the transition from JAGS to Stan. Since Stan prefers to model using the sigma normal parameterization instead of the tau parameterization, the scaling term prior variances have been adjusted accordingly. In addition, uniform priors are discouraged in Stan, excluding situations involving parameters with natural bounds like correlation, and so uniform priors have been replaced with other standard priors. The second order smoothing variance has been assigned a log- normal prior due to the sampling challenges associated with the funnel geometry produced by the smoother. The intention is to avoid a variance approaching zero which corresponds to straight line in the state-level β coefficients but also to avoid unnecessarily large smoothing variances that fail to constrain the parameters.

Model implementation

The model was implemented in Stan (mc-stan.org), a probabilistic programming language for Bayesian inference (Stan Development Team, 2021). We run the model in R using the Rstan interface (Stan Development Team, 2020). We run the model on each set of counties for a given state independent from the counties from other states. Sampling is carried out with four independent Monte Carlo Markov chains each run for 3000 iterations (with a warm up of 500 iterations), with the target acceptance (adapt_delta) set to 0.9 and tree depth of 12. This yields a total of 10,000 independent samples of each of the model parameters. The sampling for each state was run on five separate cores of a Linux server.

Diagnostic checks on convergence

We checked for each run for divergence in any iterations, whether any iterations saturated the maximum tree depth of 12, or whether the energy Bayesian fraction of missing information (E-BFMI) was low. The runs on all fifty states passed these three tests. We also checked whether any individual parameter yielded a high value for the potential scale reduction factor defined as R-hat, checking for values greater than 1.01. If so, we look more closely at the trace plots and histograms for the 10,000 samples for the given parameter to look for pathological behavior. For our inaugural estimation of county-level mortality schedules covering the time period 1982-2018, we found in nine out of the fifty states isolated instances of parameters with large R-hat (ranging from 1 to 10 parameters out of a total numbering in the ten- to few-hundred-thousand for a given state). In all but a few of these cases, the large R-hat were under a value of 1.02. We made note but otherwise accepted these exceptions.

Verifications of the model output

In addition to checking convergence, we performed a set of verifications on the model output before computing the life tables.

Comparing mortality rates between fitted and direct estimates

The model generates distributions of the natural logarithm of age-specific mortality rate across 10,000 samples for each county, year, sex, and age group combination. We first checked visually, comparing the summary statistics of each parameter (the median and the pair of 0.025- and 0.975 quantile-values forming the 95% confidence interval) with the direct estimates taken from the natural logarithm of the ratio of observed deaths and population estimates. While the direct estimate is only meaningful if the sample size of both the deaths and population estimates is large, we verified that the age-specific mortality curve generated by the fit did not spuriously deviate from the data. However, when the sample size of the numerator is small, especially in the extreme case where there are zero observed deaths, the comparison for any given mortality-rate between the fitted and direct estimates is not meaningful and cannot be transformed into a quantitative gauge of goodness of fit. We also cannot publish these comparisons without violating the data agreement that prohibits us from publishing and identifying death tabulations with cells under a threshold of 10 from the detailed mortality data. We thus also resorted to an alternative approach.

Post-predictive check – simulating deaths from estimated mortality rates and comparing with observed

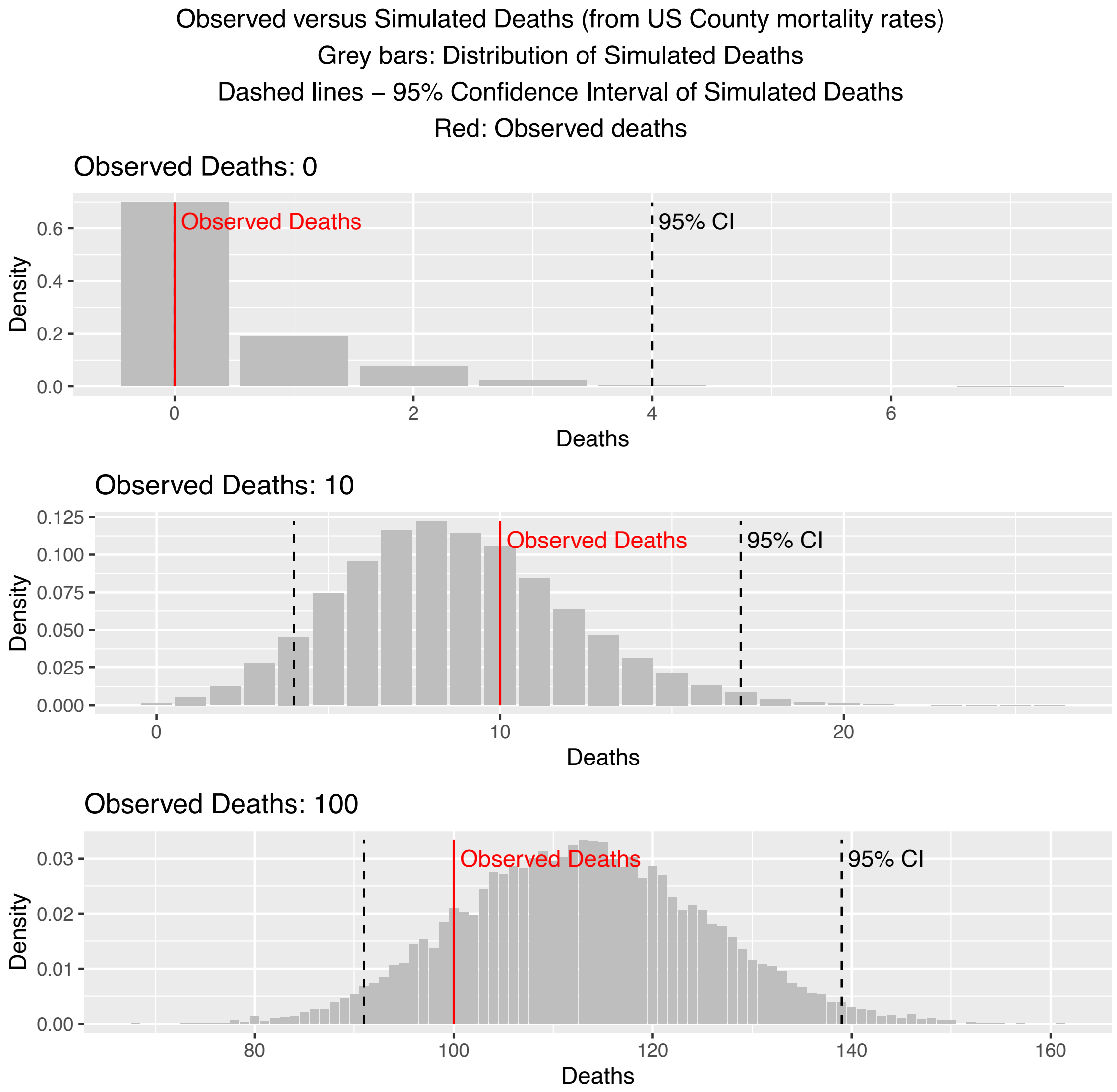
We compared the observed distribution of deaths to the distribution of deaths simulated from the estimated mortality rate. The Bayesian model assumes that for a group, i, the observed deaths, $dobs_i$, for the given population, p_i, are the outcome of a Poisson process determined by the lambda, λ_i, which is a function of the underlying mortality rate, m_i, experienced by that group.

$$dobs_i \sim Poisson(\lambda_i = m_i \times p_i)$$

We can gauge how well the model generates each measure of m_i with a simple post-predictive check. We sampled the model N= 10,000 times per parameter. From the ensemble of N measures of the underlying mortality rate, $\{m_j\}_i$, we simulated N deaths, $\{dsim_j\}_i$ using the formula above, where j is the sampling index $j = \{1,..,N\}$. We conservatively investigated whether the single observed death-count fell within the 95% confidence interval of the distribution of simulated deaths. To do so, we assigned a quantile, q_i, to each $dobs_i$ with respect to $\{dsim_j\}_i$ and looked at the statistics of the distribution of $\{q_i\}$ over all (or select) groups in the series.

In the panels in Figure 4 below, we identify from the input data individual state/county/age/sex/year cells with respectively 0, 10, and 100 observed deaths. We compare the observed deaths, *dobs*, with the distribution of simulated deaths *{dsim}* generated from the estimated underlying mortality rate for that cell. We suppress the state/county/age/sex/year identifiers since those are not relevant and would breach confidentiality. In all three cases, *dobs* falls within the 95% confidence interval of the distribution of *{dsim}*, which is true for 98.9% of all cases in the results. In the top panel (*dobs* = 0) The observed death aligns with the median and mode of the distribution of simulated deaths. In the middle panel, the number of observed deaths is greater than the median value of simulated deaths, and in the bottom panel, the number of observed deaths is lower. Case by case deviations from the median are to be expected. On whole, 38.3% of the values observed deaths were less than the respective median of simulated deaths, 39.0% of the values of observed deaths were greater than the respective median of simulated deaths, and 22.7% of observed deaths equaled the respective median of simulated deaths.

Figure 4



Observed versus Simulated Deaths (from US County mortality rates)
Grey bars: Distribution of Simulated Deaths
Dashed lines – 95% Confidence Interval of Simulated Deaths
Red: Observed deaths

Observed Deaths: 0

Observed Deaths: 10

Observed Deaths: 100

Post- predictive check – simulating state-level deaths from the aggregate of estimated mortality rates and comparing with observed

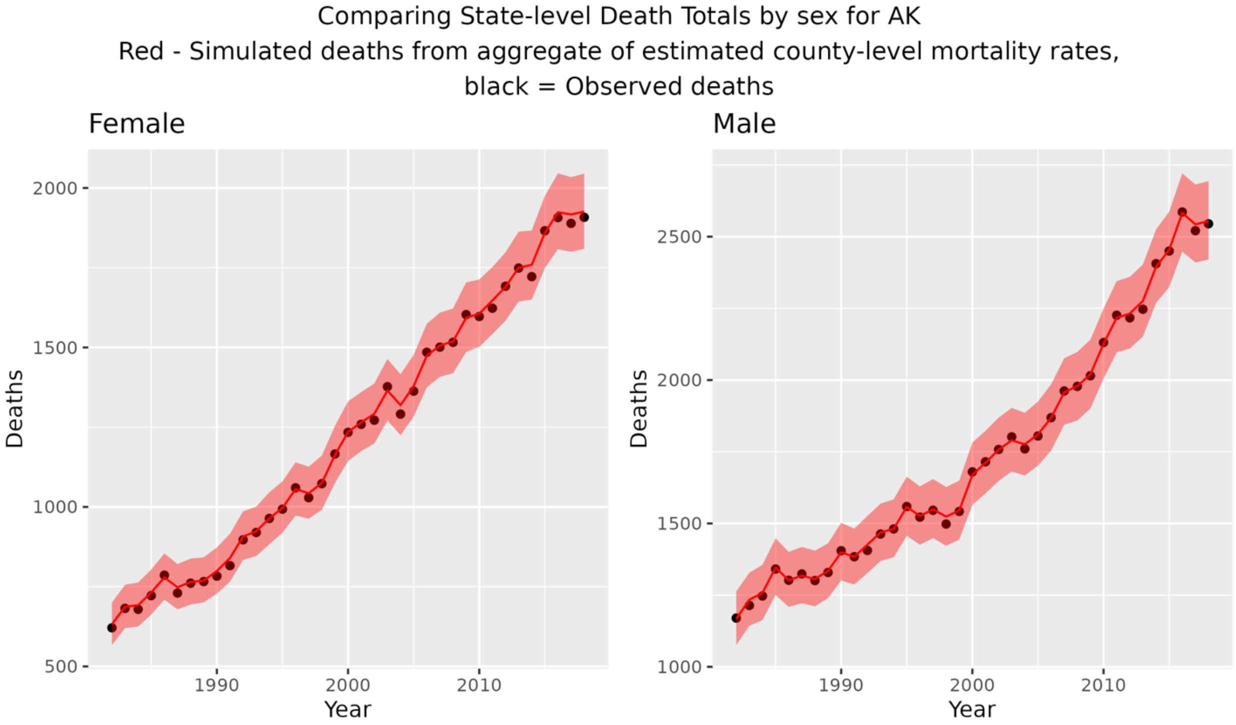
That last measure hints at a constraint that was suggested in the initial presentation of this class of Bayesian models (Alexander et al, 2017). In the case we are exploring – determining underlying mortality rates for all counties in a given state -, it would be fair to assume that the total number of simulated deaths in the state aggregate to the total of observed deaths in that state. We can further constrain that this equivalence be met for each sex, s, and each year, *t*. Conversely, we can construct a lambda that is derived from the aggregate of mortality rates over all ages, *x*, and counties, a, in the state, holding sex and year fixed, and test whether the sum of observed deaths over the chosen groups is represented by the Poisson distribution parameterized by the sum of lambdas (mortality rates and populations) for each group.

$$\sum_{a,x} dobs_{a,x,s,t}\Big|_{s,t} \sim Poisson\left(\sum_{a,x} \lambda_{a,x,s,t} = m_{a,x,s,t} \times p_{a,x,s,t}\right)\Big|_{s,t}$$

In this formulation, we sum over all counties and ages holding sex and year fixed. While this constraint was not implemented in the original model, we can use it as an overall post-predictive check of how well the model performed.

Shown below in Figure 5 are the comparisons of the sum of observed deaths by sex and year for Alaska (AK) over all counties and ages with the distribution of simulated deaths generated with the formula above. We plot the median (red line) and the 95% CI of the distribution (pink area) of simulated deaths for each year and sex for the given state and the observed death total for the corresponding sex and year for the state as a whole (black dots).

Figure 5



LIFE TABLE CALCULATIONS

We calculated life tables beginning with the output of the modelled distribution of the natural logarithm of age-specific mortality rates across 10,000 samples, each covering a combination of counties, years, age and sexes. We followed the widely adopted Chiang method (1984), with several modifications. We omit the indices for samples, counties, years, and sexes in the equations that follow for clarity, as we conducted this procedure for every combination of counties, years, sexes across all 10,000 samples.

The first step in the construction of period life tables involves a conversion of discrete and exponentiated logarithms of age-specific mortality rates ${}_{n}M_{x}$ to the probabilities of survival ${}_{n}q_{x}$, where x represents the lower bound of an age interval and n its width. We used a modified Chiang formula for the conversions (Wachter 2014):

$${}_{n}q_{x} = \frac{n\ {}_{n}M_{x}}{1\ +\ (n\ -\ {}_{n}a_{x})_{n}M_{x}}$$

where the value for ${}_{n}a_{x}$, or the average number of years lived by those dying within the age interval from age x to $x + n$, has been fixed to $n/2$ for most age groups, except for infants, children of ages 1–4 and adults 85 and over. For the latter 3 age groups, the ${}_{n}a_{x}$ values were determined using Keyfitz and Flieger (1968) algorithms, as follows:

$${}_{1}a_{0} = 0.07 + 1.7{}_{1}M_{0}$$

$${}_{4}a_{1} = 1.5$$

$${}_{\infty}a_{85} = \frac{1}{{}_{\infty}M_{85}}$$

The probability of death in the open-age interval, or ${}_{\infty}q_{85}$, is necessarily 1, as all who have crossed the age 85 will have eventually deceased.

Next, we constructed the life table survivorship column, denoted l_{x}, using the conjugate of the probability of death, or survival probability. Starting with a fixed standard population (radix, or l_{0}) of 100,000 at age 0, we calculated the proportion surviving at the beginning of each age interval for each successive age group as the cumulative product of survival probabilities of all preceding age groups:

$$l_{x} =\ l_{0}\prod_{x=0}^{x-n}(1 - {}_{n}q_{x})$$

Next, we calculated the differences in survivorship representing life table deaths between the adjacent age groups, denoted ${}_{n}d_{x}$.

$${}_{n}d_{x} = l_{x} - l_{x+n}$$

The following key column of the life table, the total number of person-years lived (PYL) within a synthetic cohort in an age interval between age x and $x + n$, denoted ${}_{n}L_{x}$, has been computed as:

$${}_{n}L_{x} =\ n\ l_{x+n} +\ {}_{n}a_{x}\ {}_{n}d_{x}$$

In turn, we calculated the reverse cumulative sum of PYL, defined as the total years of remaining life in a cohort at age x, starting with the ultimate age group, denoted as ω:

$$T_x = \sum_{\omega}^{x} {}_nL_x$$

And finally, we computed the average number of years of remaining life at age x, or life expectancy, by dividing the total number of remaining PYL in a cohort by the survivorship at the beginning of a corresponding age interval:

$$e_x = \frac{T_x}{l_x}$$

To compute both-sex combined life tables, we rescaled the implied age-specific mortality rates by weighing each set by the population for each sex, as follows:

$${}_nM_x^{BS} = \frac{{}_nM_x^F \; {}_nN_x^F + {}_nM_x^M \; {}_nN_x^M}{{}_nN_x^{BS}}$$

where ${}_nN_x$ is the population in age group x to $x + n$, and superscripts denoting both-sex (BS), females (F) and males (M).

The remaining steps leading to the completion of both-sex life tables were identical to the aforementioned procedure.

In addition to computing the central estimates of life table quantities, we characterized the uncertainty of each life table column of interest. Given that the modelled age-specific mortality rate distribution was the result of a Poisson simulation with 10,000 random draws, the construction of confidence intervals for life table quantities did not require closed-form solutions and followed directly from the simulated distribution, equivalent to the Monte Carlo simulation solution proposed by Silcocks, Jenner and Reza (2001). Instead, the 95% confidence bounds were determined as 2.5th and 97.5th percentiles of the distribution of life table quantities, calculated on the basis of each sample of age-specific mortality rates for every county, year, and sex.

Life table diagnostics

At this stage, it was necessary to establish the degree of accuracy attained by our model. As we had no systematic and authoritative source of comparative life table quantities at a county level, we chose to aggregate the deaths implied by the simulated age-specific mortality rates to the state level. This presented us with an opportunity to test our results against the state level United States Mortality Database (USMDB) tables which had been constructed with a different set of methods (see the HMD Methods Protocol at https://www.mortality.org/Public/Docs/MethodsProtocol.pdf).

Our initial comparison demonstrated the lack of fit and a consistent overestimation of life expectancy by our model, all while the cumulative number of deaths fell within the 95% confidence interval. Our investigation concluded that there was near perfect agreement between observed and estimated deaths at the state level for all ages below 85 and that the main cause of the lack of fit was attributable to mortality in the open age interval (85+ years). The value of ${}_{\infty}a_{85}$ we adopted following classic demographic approximation assumes a constant hazard over the open age interval while we know that the hazard should be increasing with age. This is not a big problem where most of the deaths in a population occur below the open age interval but there is increasing divergence as longevity increases. This assumption is generally not adequate for modern populations, because the survivorship at advanced ages has increased considerably in the recent decades and the deaths are no longer decreasing monotonically with age past the age 85. However, as we did not have detailed age at death by county over the age 85 and of all the possible ways to estimate mortality for ages above 85, we have assumed that the classic method is the most robust.

References

Alexander, M., Zagheni, E., and Barbieri, M. (2017). *A Flexible Bayesian Model for Estimating Subnational Mortality*. Demography 2017, 54(6): 2025–2041.

Chiang, C. L. (1984). *Life table and its applications*. Robert E. Krieger Publishing, Malabar, FL. https://doi.org/10.2307/2529814

Keyfitz, N., and Flieger, W. (1968) *World population: an analysis of vital data*. https://doi.org/10.2307/1530261

Mahalanobis, P.C. (1936) *On the Generalised Distance in Statistics.* Sankhya A 80, 1-7 (2018). https://doi.org/10.1007/s13171-019-00164-5

Silcocks, P. B. S., Jenner, D. A., and Reza, R. (2001). Life expectancy as a summary of mortality in a population: statistical considerations and suitability for use by health authorities. *Journal of Epidemiology & Community Health*, *55*(1), 38-43. https://doi.org/10.1136/jech.55.1.38

Stan Development Team (2020). Rstan: the R interface to Stan. R package version 2.21.2. http://mc-stan.org/

Stan Development Team (2021). Stan Modeling Language Users Guide and Reference Manual, version 2.27. http://mc-stan.org/

Wachter, K. W. (2014). *Essential demographic methods*. Harvard University Press. https://doi.org/10.4159/9780674369757

Appendix A – Description of input data and sources of information

Mortality

1981 – 1988

- NCHS detailed mortality files – public files
- National Center for Health Statistics. (1981–1988). Mortality Multiple Cause of Death.
- Data & Documentation (Mortality Multiple Cause Files): https://www.nber.org/research/data/mortality-data-vital-statistics-nchs-multiple-cause-death-data (link valid February 21, 2021)

1989-2018

- NCHS detailed mortality files – restricted access through NAPHSIS agreement
- National Center for Health Statistics. Natality and detailed multiple cause mortality files, 1989-2018, as compiled from data provided by the 57 vital statistics jurisdiction through the Vital Statistics Cooperative Program

Population

1980 – 1989

- US Census Bureau - Estimates of the Population of Counties in the United States by Age, Sex, and Race: July 1, 1980
- Source: Intercensal Population Estimates by Age, Sex, and Race: 1980-1989
- https://www.census.gov/data/tables/time-series/demo/popest/1980s-county.html
- Internet Release date: October 22, 2004
- Revised May 12, 2009
- https://www2.census.gov/programs-surveys/popest/datasets/1980-1990/counties/asrh/pe-02.xls

1990-1999

- CDC National Center for Health Statistics
- Source: Bridged-Race Intercensal Population Estimates for July 1, 1990 – July 1, 1999
- https://www.cdc.gov/nchs/nvss/bridged_race/data_documentation.htm#july1990-1999
- Internet Release date: July 26, 2004
- Data:
  ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/datasets/nvss/bridgepop/icenA1_1.zip
  ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/datasets/nvss/bridgepop/icenA2_1.zip
  ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/datasets/nvss/bridgepop/icenA3_1.zip
  ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/datasets/nvss/bridgepop/icenA4_1.zip
- Documentation: ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/datasets/nvss/bridgepop/DocumentationBridgedIntercenA1.doc

2000-2009

- CDC National Center for Health Statistics

- Source: July 1, 2000 – July 1, 2009 Revised Bridged-Race Intercensal Population Estimates
- https://www.cdc.gov/nchs/nvss/bridged_race/data_documentation.htm#july2000-2009
- Internet Release date: October 26, 2012
- Data: ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/datasets/nvss/bridgepop/2000_09/icen_2000_09_y0509.zip and ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/datasets/nvss/bridgepop/2000_09/icen_2000_09_y0004.zip
- Documentation: ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/datasets/nvss/bridgepop/2000_09/DocumentationRevisedBridgedIntercen2000_09.pdf

2010-2019

- CDC National Center for Health Statistics
- Source: Vintage2019 Bridged-Race Postcensal Population Estimates
- https://www.cdc.gov/nchs/nvss/bridged_race/data_documentation.htm#Vintage2019
- Internet Release date: July 9, 2020
- Data: https://www.cdc.gov/nchs/nvss/bridged_race/pcen_v2019_y1019_txt.zip
- Documentation: (Note, vintage 2019 release links to vintage 2018 documentation) https://www.cdc.gov/nchs/nvss/bridged_race/Documentation-Bridged-PostcenV2018.pdf

Natality

1981 – 1988

- National Center for Health Statistics (NCHS)
- Birth Data Files
- Data & Documentation: https://www.cdc.gov/nchs/data_access/vitalstatsonline.htm

1989

- NCHS detailed natality files
- National Center for Health Statistics. Natality and detailed multiple cause mortality files, 1989, as compiled from data provided by the 57 vital statistics jurisdiction through the Vital Statistics Cooperative Program
- National Center for Health Statistics (1989). Data File Documentations, Natality, January 2002 (machine readable data file and documentation, CD-ROM Series 21, No. 23H - ASCII), National Center for Health Statistics, Hyattsville, Maryland.
- Data provided by NAPHSIS

Appendix B - Detail and format of the input data

The Bayesian model, implemented in the Stan framework invoked through the Rstan package in R – (Stan Development Team, 2020), expects input series of deaths and population estimates, which we created from the data listed in Appendix A. Below, we described the input data necessary to run the R-code (soon available at usa.mortality.org/counties.php) producing our estimated mortality rates from the Bayesian framework.

County-level mortality rates were computed separately for each state, so we created 50 separate respective input death-tabulation and population-tabulation files, a pair for each state. Each input file is a comma-separated-variable (CSV) file. The first row is a descriptive text header followed by subsequent rows containing the data and their descriptors. The column-wise order of variables is not fixed since the data are read in by header.

Detail and format of the death input file – The files should contain the tabulated deaths for each year, county, sex, and age-group. If there are zero tabulated deaths, these lines do not need to be included. The wrapper-script for the Bayesian model will impute zeros where data are missing. The first four lines of the input file for California are shown below (the death counts include decimals because deaths of unknown ages have already been redistributed):

```
Year,State,PopCode,Sex,Age,AgeInterval,Deaths
1982,CA,06001,f,0,1,76.1984766050054
1982,CA,06001,f,1,4,10.0261153427639
1982,CA,06001,f,5,5,16.0417845484222
```

…

Of the listed columns, the critical ones are the year (“Year”), county (“PopCode”), sex (“Sex”), age (“Age”), and deaths (“Deaths”). We include the state (“State”) and the age interval (“Age Interval”) for reference but these two variables are not used by the computer script. The single file should include tabulations for each year in the series, each county or county-grouping in the state, each age-group, and each sex.

Data type & format (* denotes required fields)

- ***year** (Year) - four-digit numeric (integer)
- state (State) – two letter acronym (character)
- ***county** (PopCode) – five-character FIPS code and/or county-grouping code (character)
- ***sex** (Sex) – “m” for male, “f” for female (character)
- ***age** (Age) – the lower bound of the age group, one or two-digit numeric (e.g. 0,1,5,10,15,20,25,30,35,40,45,50,65,70,75,80,85) (integer)
- age interval (Age Interval) – the interval of the age group in years, number or ‘+’ for open age interval (character)
- ***deaths** (Deaths) – the tabulated deaths for the county/sex/year/age-group (numeric, or float) (Note that the Bayesian estimator can only handle integer deaths so the wrapper script will first round the given tabulation to the nearest integer value)

Detail and format of the population input file – These files should contain the tabulated population counts for each year, county, sex, and age-group. If there is no population of a certain age and sex for a given county and year, a zero line should be included in the input file (the Bayesian estimation model and script does not impute zero values). The first four lines of the input file for California are shown below:

```
PopCode,Sex,Age,AgeInterval,Day,Month,Year,Population
06001,f,0,1,1,7,1982,8712.55919043289
06001,f,1,4,1,7,1982,29998.4408095671
06001,f,5,5,1,7,1982,34029
...
```

The order of the columns does not matter since the wrapper script for the Bayesian model reads in the data by header. Of the listed columns, the critical ones are the year ("Year"), county ("PopCode"), sex ("Sex"), age ("Age"), and population ("Population"). We include the age interval ("Age Interval"), the day ("Day") and month ("Month") of the official population estimate. For idiosyncratic reasons, we do not include the state as it can be inferred from the first two digits of the county FIPS code or county grouping code. The single file (one per state) should include tabulations for each year in the series, each county or county-grouping in the state, each age-group, and each sex.

Data type and format (* denotes required fields)

- ***county** (PopCode) – five-character FIPS code and/or county-grouping code (character)*year (Year) - should be a four-digit numeric (integer)
- ***sex** (Sex) – "m" for male, "f" for female (character)
- ***age** (Age) – the opening age of the age group, one or two-digit numeric (e.g. 0,1,5,10,15,20,25,30,35,40,45,50,65,70,75,80,85) (integer)
- day (Day) – day of the month of population estimate - one or two-digit numeric (integer)
- month (Month) – numeric code for month of population estimate - one or two-digit numeric (integer)
- ***year** (Year) – year of population estimate - four-digit numeric (integer)
- age interval (Age Interval) – the age interval of the age group, number or '+' for open age interval (character)
- ***population** (Population) – the tabulated population estimates for the county/sex/year/age-group (numeric, or float) (Note that the Bayesian estimator can process non-integer values for the population estimates)

Detail and format of the matching file for the county aggregates – As discussed above, the model runs best for population of at least 10,000 (about 5,000 males and 5,000 females). We thus had to group together small counties and we did so within each state (see the section titled "County grouping"). We provide a text file listing all the counties, which have been aggregated together with the group identification code. The R code will however run just as well with any other type of aggregation the user might choose as long as the format of the matching table conforms to what the computer script is expecting.

Appendix C - Format of the county-level life tables

Below we describe the format of the county-level life tables available on the USMDB website (https://usa.mortality.org). For each state, we publish sex-specific life tables. Each file contains life tables for all years in the series and for all counties and county groupings within the state. File size thus varies as a function of the number of counties and county-groupings in a given state.

All life tables are in the 5x1 (by 5-year age-group and single calendar year) format. Five-year age groups means 0, 1-4, 5-9, 10-14,..., 80-84, 85+. Age groups are defined in terms of completed age, so "5-9" extends from exact age 5 to just before the 10th birthday (sometimes written elsewhere as "5-10"). For brevity, we indicate age-group by the the lower bound of the age bucket (e.g. “0” for ages 0-1).

The core variables in the life tables are listed in the table below

Variable	*Variable name*	*Description*
County	`fips`	5-digit FIPS code for single counties ; 5-character alpha-numeric code for county aggregates
Year	`year`	Calendar year
Age	`age`	Age group for *n*-year interval from exact age *x* to just before exact age *x*+*n*, where *n*=1, 4, 5, or NA (open age interval, ∞) years
Sex	`sex`	“b” for both sexes, “f” for females, “m” for males
m(x)	`mx`	Central death rate between ages *x* and *x*+*n*
n(x)	`n`	Age interval; *n* = 1, 4, 5, or NA (open age interval, ∞) years
a(x)	`ax`	Average length of survival between ages *x* and *x*+*n* for persons dying in the interval
q(x)	`qx`	Probability of death between ages *x* and *x*+*n*
l(x)	`lx`	Number of survivors at exact age *x*, assuming l(0) = 100,000
d(x)	`dx`	Number of lifetable deaths between ages *x* and *x*+*n*
L(x)	`Lx`	Number of person-years lived between ages *x* and *x*+*n*
T(x)	`Tx`	Number of person-years remaining after exact age *x*
e(x)	`ex`	Life expectancy at exact age *x* (in years)

We also include the variance and lower- and upper- bound of the 95% confidence interval for m(x), q(x), and e(x), making for a total of 22 columns.

We use the following naming convention for the lifetable files: *XX*_*s*_county_lt, where *XX* ia the two-letter state acronym (e.g. "DE" for Delaware) and *s* is sex ("b" for both sexes, "f" for female, "m" for male), The life tables are shared in two formats: as single-object R-binaries (with file extension ".rds") and text files with comma-separated variable format (with file extension ".csv"). Both files make use of the same column headers. Below we share a screen print (from the R programming console) of the first six lines of the both-sex combined lifetable file for Delaware (DE) starting with fips county "10001" and year 1982 (with file header DE_b_county_lt)

```
    fips year age sex              mx n            ax             qx           lx           dx
1: 10001 1982   0   b 0.0117858096186 1 0.09003587635 0.011660751973 100000.00000 1166.0751973
2: 10001 1982   1   b 0.0004663159197 4 1.50000000000 0.001863091705  98833.92480  184.1366655
3: 10001 1982   5   b 0.0002298135998 5 2.50000000000 0.001148408200  98649.78814  113.2902256
4: 10001 1982  10   b 0.0002339706051 5 2.50000000000 0.001169169147  98536.49791  115.2058333
5: 10001 1982  15   b 0.0008716813621 5 2.50000000000 0.004348929608  98421.29208  428.0272712
6: 10001 1982  20   b 0.0012758513538 5 2.50000000000 0.006358974005  97993.26481  623.1366236


             Lx          Tx          ex               var_mx     mx_CI_lower     mx_CI_upper
1:  98938.91341 7413661.607 74.13661607 0.000003058928580800 0.0086061407052 0.0153844499079
2: 394875.35755 7314722.694 74.01024201 0.000000007574898838 0.0003165120882 0.0006638320105
3: 492965.71512 6919847.336 70.14558741 0.000000003363176032 0.0001474837350 0.0003797521505
4: 492394.47498 6426881.621 65.22336147 0.000000001713634289 0.0001624151887 0.0003270936513
5: 491036.39221 5934487.146 60.29678153 0.000000023962192284 0.0006483652475 0.0012567481859
6: 488408.48248 5443450.754 55.54923356 0.000000041514406575 0.0009554889956 0.0017502073894


                 var_qx     qx_CI_lower    qx_CI_upper       var_ex ex_CI_lower ex_CI_upper
1: 0.000002930401674950 0.0085388731762 0.015173460209 0.2050410475 73.23823369 75.02154589
2: 0.000000120602969470 0.0012650473460 0.002650928614 0.1830033891 73.17099081 74.85335254
3: 0.000000083850654540 0.0007371468821 0.001896959816 0.1815210884 69.30999671 70.98359091
4: 0.000000042733673780 0.0008117463435 0.001634131971 0.1806225916 64.39274684 66.05703386
5: 0.000000593329455910 0.0032365800225 0.006264060064 0.1798985097 59.46406987 61.12688868
6: 0.000001023873157370 0.0047660601826 0.008712913433 0.1722978747 54.73690588 56.37831637
```