INDEPTH Model Life Tables 2.0

INDEPTH Working Group on All-Cause Mortality:

Samuel J. Clark, Momodou Jasseh, Sureeporn Punpuing,
Eliya Zulu, Ayaga Bawah, Osman Sankoh,

and INDEPTH Network Member Sites

1   Motivation

Age patterns of mortality in Africa are not well documented because few high quality data are available to describe adult mortality. Much of the current description of adult mortality in Africa, and hence expectation of life and levels and trends in anything else that depends on adult mortality, comes from extrapolating child mortality measured by a variety of surveys using indirect estimation techniques. Model age patterns of mortality are necessary both for the indirect estimation of child mortality and the extrapolation of child mortality to adult mortality. Existing widely used model life table systems (Coale and Demeny, 1966; United Nations Department of International Economic and Social Affairs, 1982) are based on mortality observed a number of decades in the past everywhere else in the world *except* Africa. Consequently when model age patterns from existing model life table systems are used to support indirect estimation techniques that measure mortality in Africa, an implicit assumption is made that age patterns of mortality in Africa are similar to the rest of the world 20, 30, or even 50 years ago. There is no reason to expect contemporary age patterns of mortality in Africa to be similar to historical patterns elsewhere in the world, especially in areas with high HIV prevalence where the age pattern of mortality is unique and was not observed anywhere else in the world before the HIV pandemic in Africa.

This work utilizes prospectively collected, individual-level data gathered by demographic surveillance system (DSS) sites that are members of the INDEPTH Network (INDEPTH Network, 2008) to identify commonly observed age patterns of mortality in Africa and Asia and use them to build a set of model life tables that are more appropriate for use in Africa.

2   Specific Aims

1. Evaluate the quality of individual-level data describing mortality from individual DSS sites

2. Calculate mortality rates and life tables by time, sex and age for all data that pass the evaluation

3. Identify commonly observed age patterns of mortality

4. Build an easy-to-use system of model life tables based on the observed patterns

3 Data

The INDEPTH Network coordinates contribution of data from member sites. Individual-level data describing potentially multiple observed intervals for each individual at each site are provided with the following attributes for each record:

- Site name
- Individual ID (anonymized)
- Sex
- Date of birth
- Date of death (if death has occurred)
- Date when observation begins
- Date when observation ends

The final data set for the preliminary analysis presented here includes data from seventeen of the 30+ sites that are members of the INDEPTH Network. The data set includes slightly fewer than four million observed intervals that yield almost 6.5 million person years of exposure and just over 84,000 deaths. These are aggregated as described in section 4.1 to produce 82 unique 'site periods', each site period consisting of sex-age-specific counts of deaths and person years exposed over a defined period of calendar time in a specific site.

4 Methods

4.1 Data Preparation

The raw data are evaluated for validity, consistency and plausibility before they are included in the analysis. Checks for validity include ensuring that all dates are well-formed and in a reasonable historical range, that the structure of each data set conforms to the definition and is complete (no missing items) and interpretable (contains a description of any non-standard codes or conventions that may have been used), and that there are no glaring errors in the data - such as all exposure intervals terminated on the same date in the year 3000.

After data have passed this rudimentary evaluation they are subjected to a more careful check for consistency. The data are composed of possibly multiple observed exposure intervals for each individual. Observation start and stop dates define the duration of observed intervals and dates of birth and possibly death define the duration of individuals' lives. Consistency checks are of two types – verifying that the temporal sequencing of dates within a given interval and across multiple intervals is valid, and ensuring that datum that should be the same across multiple intervals describing the same individual are. In the first category:

1. multiple intervals for an individual must not overlap in time,
2. the DOB must precede the date of death if there is one,
3. the exposure start date must precede the exposure stop date,
4. the exposure start date must occur on or after the date of birth, and
5. the exposure stop date must occur before or on the date of death if there is one.

In the second category, among the intervals for a single individual:

1. dates of birth must be the same,
2. dates of death must be the same if the individual has died, and
3. sexes must be the same.

Data that pass these checks are then aggregated across time by site and sex so that each sex-specific site period contains at least 50,000 person years of exposure (summed across all ages). In a small number of cases this is not possible because a site does not have sufficient exposure, and in that circumstance the site contributes just one site period with all of its exposure.

A life table is then calculated for each resulting site period and examined for plausibility. At this stage site periods or whole sites are rejected that produce clearly implausible mortality patterns. For example, extreme values for the expectation of life (i.e. 5 or 150), or age patterns of mortality that in a gross way do not conform to the standard J-shape of human mortality.

Finally, the resulting life tables are fit using the model described in section 4.4 in order to remove spurious variation and produce a smooth age pattern of mortality.

4.2   Mortality Model

4.2.1   Motivation

There are two requirements for the mortality model. It must represent mortality age patterns in a parsimonious way that helps identify regularities among possibly many empirical age

3

patterns, and it must be able to represent a range of *model* mortality patterns based on the common patterns that emerge from the empirical data.

The general form of the model will be to represent a mortality age pattern as the weighted sum of two or more independent, age-varying components that represent the age-varying nature of the mortality schedule. To this is added a constant at each age to take into account the non-age-varying level of the mortality schedule. Any remaining differences between the modeled and observed age patterns are captured with a residual term.

The independent, age-varying components necessary for this model can be easily derived from a principle components analysis of observed mortality schedules. The resulting score vectors are the independent components we need, and they have the convenient property of encoding the bulk of the variance among the observed mortality schedules in a small number of components. Experience suggests that no more than six components are necessary to capture well over 99% of the variance across observed mortality schedules.

Last, the model must operate on an appropriate scale. Because mortality rates are simply the ratio of deaths to person years lived in a given population, the natural range of mortality rates is $[0, \infty)$. In order to prevent the model from producing mortality rate values less than 0, mortality is modeled on a log scale, and the final mortality rate values are procured by exponentiating the output of the model.

What follows advances similar previous work by the authors (Clark, 2002).

4.2.2 Model

Assuming the standard 19 age groups (0, 1-4, 5-9, ... 85+), a 19 x m matrix $\mathbf{M}$ composed of m column vectors of age-specific mortality rate schedules can be expressed as a weighted sum of a number of components whose shapes encode the fundamental age pattern of human mortality and a wide range of variations on that:

$$\mathbf{M} = \mathbf{SB} + \mathbf{C} + \mathbf{R} \tag{1}$$

$\mathbf{S}$ is a 19 x n matrix whose columns are the n factor scores used in the model (derived from a principal components analysis of all of the empirical mortality rate schedules). $\mathbf{B}$ is a n x m matrix whose columns are coefficients that multiple each score schedule contained in $\mathbf{S}$ to yield the age-varying component of each mortality schedule. $\mathbf{C}$ is a 19 x m matrix whose columns are constants that are added to the result of the multiplication to modify each mortality schedule in an age-constant way. Finally, $\mathbf{R}$ is a 19 x m matrix of residuals that account for the remaining difference between the modeled and empirical mortality schedules.

$$\mathbf{M} \;=\; \begin{bmatrix} m_{1,1} & m_{1,2} & \cdots & m_{1,m} \\ m_{2,1} & m_{2,2} & \cdots & m_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ m_{19,1} & m_{19,2} & \cdots & m_{19,m} \end{bmatrix} \tag{2}$$

4

$$\mathbf{S} = \begin{bmatrix} s_{1,1} & s_{1,2} & \cdots & s_{1,n} \\ s_{2,1} & s_{2,2} & \cdots & s_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ s_{19,1} & s_{19,2} & \cdots & s_{19,n} \end{bmatrix}$$

$$\mathbf{B} = \begin{bmatrix} b_{1,1} & b_{1,2} & \cdots & b_{1,m} \\ b_{2,1} & b_{2,2} & \cdots & b_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ b_{n,1} & b_{n,2} & \cdots & b_{n,m} \end{bmatrix}$$

$$\mathbf{C} = \begin{bmatrix} c_{\cdot,1} & c_{\cdot,2} & \cdots & c_{\cdot,m} \\ \vdots & \vdots & \vdots & \vdots \\ c_{\cdot,1} & c_{\cdot,2} & \cdots & c_{\cdot,m} \end{bmatrix}$$

$$\mathbf{R} = \begin{bmatrix} r_{1,1} & r_{1,2} & \cdots & r_{1,m} \\ r_{2,1} & r_{2,2} & \cdots & r_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ r_{19,1} & r_{19,2} & \cdots & r_{19,m} \end{bmatrix} \tag{3}$$

Ignoring the residuals, $\mathbf{SB} + \mathbf{C}$ represents the modeled mortality schedules. $\mathbf{SB}$ captures the age-varying component of the mortality schedules and $\mathbf{C}$ represents the non age-varying level of each mortality schedule.

If only one mortality schedule is involved:

$$\begin{bmatrix} m_1 \\ m_2 \\ \vdots \\ m_{19} \end{bmatrix} = b_1 \cdot \begin{bmatrix} s_{1,1} \\ s_{2,1} \\ \vdots \\ s_{19,1} \end{bmatrix} + b_2 \cdot \begin{bmatrix} s_{1,2} \\ s_{2,2} \\ \vdots \\ s_{19,2} \end{bmatrix} + \cdots + b_n \cdot \begin{bmatrix} s_{1,n} \\ s_{2,n} \\ \vdots \\ s_{19,n} \end{bmatrix} + \begin{bmatrix} c \\ c \\ \vdots \\ c \end{bmatrix} + \begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_{19} \end{bmatrix} \tag{4}$$

The score vectors $s_{(\cdot,\cdot)}$ can be thought of as a new basis in age-varying space with the special property that most of the variance in the data is represented by a small subset of the these. The weights $b_{(\cdot)}$ determine a given point in this space, and then c adds a constant amount to each dimension.

The effective parameters in this model are the b's, and as mentioned above when only a small number of score vectors is necessary to account for the bulk of the variance in the empirical data set, then the number of b's necessary is small, making the model effectively parsimonious. Of course in reality the model is much less parsimonious because the score vectors are parameters themselves, albeit fixed. Provided the original data set from which they are calculated is highly varied, they will be capable of representing a wide range of age variation and can be thought of as permanently fixed.

4.3 Identification of Similar Age Patterns of Mortality

Similar age patterns of mortality are identified using a clustering method. First the dimensionality of the data is reduced using factor analysis to concentrate information in a smaller number of dimensions and remove spurious variation.

This is accomplished by conducting a principal components analysis of the empirical age schedules of mortality, including both the female and male schedules together. The score vectors resulting from this are independent of each other and concentrate the information contained in the empirical data set in a small number of new dimensions. As a result all of the information in the original data set can be represented using the first few score vectors, and hence the dimensionality of the data can be reduced from nineteen to roughly four to six.

This reduction in dimensionality is accomplished by regressing (simple OLS linear regression) each empirical mortality schedule on the first ten score vectors and storing the resulting coefficients and constants in a new data set. Keeping ten coefficients provides an opportunity to vary the number of components used in the clustering and choose the number that provides the best clusters.

With this new reduced-dimension data set, the model-based clustering method developed by Fraley and Raftery (2006) is used to identify robust clusters. This is a fully automated, robust clustering method that identifies the number of clusters that maximizes the bayesian information criteria (BIC). The method yields both the BIC values for different numbers of clusters and the classification of the data using the number of clusters with the greatest BIC value.

A priori there is no objective way to choose how many dimensions to include in the clustering, so clustering is performed on the reduced dimension data set using 2–10 dimensions, and the clustering classification from each is saved in a new data set. The best clustering is chosen by calculating a new fit metric, the "total deviation from median" (TDM). This is the sum of the absolute differences between each mortality schedule and the median of all mortality schedules in the cluster to which it is assigned. Lower values of the TDM indicate less variation among mortality schedules in each cluster, and consequently, the best clustering has the lowest TDM value.

4.4 Smoothing

The empirical mortality schedules - both unclustered and clustered - naturally contain some stochastic variation that produces small meaningless irregularities in the age profiles. When it is necessary to eliminate this meaningless variation a model-based smoothing procedure is employed.

The eight-parameter deterministic model of mortality as a function of age presented by Heligman and Pollard (1980) is used to represent a smooth curve through the probability of

dying in each age group. This model captures the age-specific shape of a mortality schedule using three components, one for child mortality (three parameters), one for adult mortality (two parameters) and one to represent a 'hump' located somewhere in the age profile (three parameters). The hump was originally designed to represent the slight rise in mortality at young adult ages associated with accidents and/or maternal mortality, and fortunately for us, the parameterization of the hump is sufficiently flexible to also allow it to capture the bulge of increased mortality during adult ages associated with HIV mortality. The age-specific probability of dying $q(a)$ is given by:

$$q(a) = A^{(a+B)^C} + D \exp\left(-E(\ln(a) - \ln(F))^2\right) + \frac{GH^a}{1 + GH^a} \tag{5}$$

<div align="center">Parameters</div>

A	probability of dying at age 1
B	difference between probabilities of dying at ages 0 and 1
C	decline in probability of dying during childhood
D	intensity of hump mortality
E	inverse with spread of hump across age
F	modal age of hump
G	late life mortality (intercept of Gompertz curve at age 0)
H	late life mortality (slope of Gompertz curve)

This model is fit to $_nq_x$ values from a given life table using a maximum likelihood procedure. The log likelihood of observing a given empirical age profile of mortality given a set of parameter values is calculated and maximized. The likelihood is calculated assuming that the probability of dying at each age is binomially distributed and independent of the probabilities of dying at other ages.

$$\mathcal{L} = \prod \binom{P}{d} q^d (1-q)^{(P-d)} \tag{6}$$

where P is the total number of people alive at the beginning of the age group, d is the number of deaths in the age group, and q is the probability of dying in the age group specified by the model with a specific set of parameter values.

4.5   Model Lifetables

Following the tradition set out by existing systems of model life tables, the INDEPTH model life table system 2.0 is composed of four families with different levels of mortality in each. Each family is based on one of the four clusters of mortality patterns identified following the procedure outlined in section 4.3.

4.5.1 Life Table Families

The overall mortality pattern for each cluster is calculated by summing the age-specific deaths and person years across all site periods included in the cluster and dividing them to calculate a new cluster-specific mortality rate schedule for the cluster. The resulting four cluster-specific mortality rate schedules are the underlying mortality age profiles on which the model life table families are based.

4.5.2 Age-Varying Mortality Levels within a Family

There is variation in the overall level of mortality within each of the clusters that underly the model life table families. Within a cluster, some age patterns contain mortality rates that are consistently higher at each age, and some are likewise composed of mortality rates that are consistently lower at each age. The age-specific difference between these generally 'higher' and 'lower' patterns within the cluster must be represented by the mortality model so that it can generate arbitrary mortality patterns at different levels within the family that is based on the cluster that conform to both the underlying age pattern of the cluster and the manner in which it changes as the overall level of mortality changes within the cluster.

Using the model of mortality described in section 4.2.2 this is straightforward observing that:

$$
\begin{aligned}
\mathbf{M}_h - \mathbf{M}_l &= [\mathbf{SB}_h + \mathbf{C}_h] - [\mathbf{SB}_l + \mathbf{C}_l] \\
&= \mathbf{S}\,[\mathbf{B}_h - \mathbf{B}_l] + [\mathbf{C}_h - \mathbf{C}_l] \\
&= \mathbf{S}\boldsymbol{\Delta} + \delta
\end{aligned}
\tag{7}
$$

where h and l indicate the 'high' and 'low' extremes of the mortality patterns within a cluster.

To define an arbitrary level of mortality within a family, a fraction α of the quantity defined in equation 7 that encodes the age-varying deviation associated with moving between the extremes of the levels of mortality in the cluster underlying the family is added to the underlying age profile for the family. With this addition the model that can represent arbitrary mortality levels within a family is:

$$
\mathbf{M} = \mathbf{S}\,[\mathbf{B}_u + \alpha\boldsymbol{\Delta}] + [\mathbf{C}_u + \alpha\delta]
\tag{8}
$$

where u refers to the 'underlying' age pattern of mortality that defines the family.

4.5.3 Calculating Model Life Tables

The final model life tables are constructed by choosing values of α that yield mortality rate schedules that produce life tables that have specific values for the expectation of life at birth. This is accomplished using an optimizer that varies α until the target value of life expectancy is obtained.

A final refinement in the production of model life tables is to smooth the mortality schedules so that the resulting model mortality patterns are smooth and do not contain meaningless stochastic bumps and wiggles. This is done by smoothing all of the site periods using the procedure outlined in section 4.4 and then recalculating the principal components using the smoothed site periods. The resulting smoothed principal components are used in equations 7 and 8 to produce smooth model mortality schedules.

4.6 Software

Data cleaning, counting deaths, calculation of person years and calculation of mortality rates are done using the Structured Query Language (SQL) manipulating data stored in relational databases hosted on the Microsoft SQL Server and MySQL relational database management systems. All statistical analysis is conducted using the statistical package R (The R Foundation for Statistical Computing, 2008) using a combination of customized code and published methods.

5   Preliminary Results

The results presented here are *preliminary* and obtained from a *partial* data set that includes only those data that passed the data quality checks described in section 4.1. Descriptions of the problems with their data have been sent back to sites, and we are waiting for the sites to provide new clean data that addresses the problems we identified. We anticipate that the final data set will be significantly larger than the one used here, and further it is likely that the results will change somewhat when the final data set is assembled.

Figure 1 clearly reveals substantial variation in the age patterns of mortality associated with the site periods that are included in this analysis. The first step of our analysis is to identify subsets of the mortality curves displayed in this figure that are similar to each other using the clustering method described in section 4.3.

Underlying both the identification of the clusters and creation of model life tables is the model of mortality presented in section 4.2.2 that relies on a principal components analysis of the raw mortality schedules. The first four factors produced by that analysis are displayed in figure 2. The first component that captures 93% of the variance in the set of empirical mortality schedules has the characteristic J-shape of human mortality, and the remaining three components contain a variety of much smaller amplitude bumps and wiggles that when combined with the first component are able to give the resulting mortality pattern a nuanced age dependance.

Figure 3 and 4 display the results of the clustering described in section 4.3. Four clusters using four components in the clustering produced the best result. We will not provide detailed interpretations of these clusters based on the site periods they contain until we have
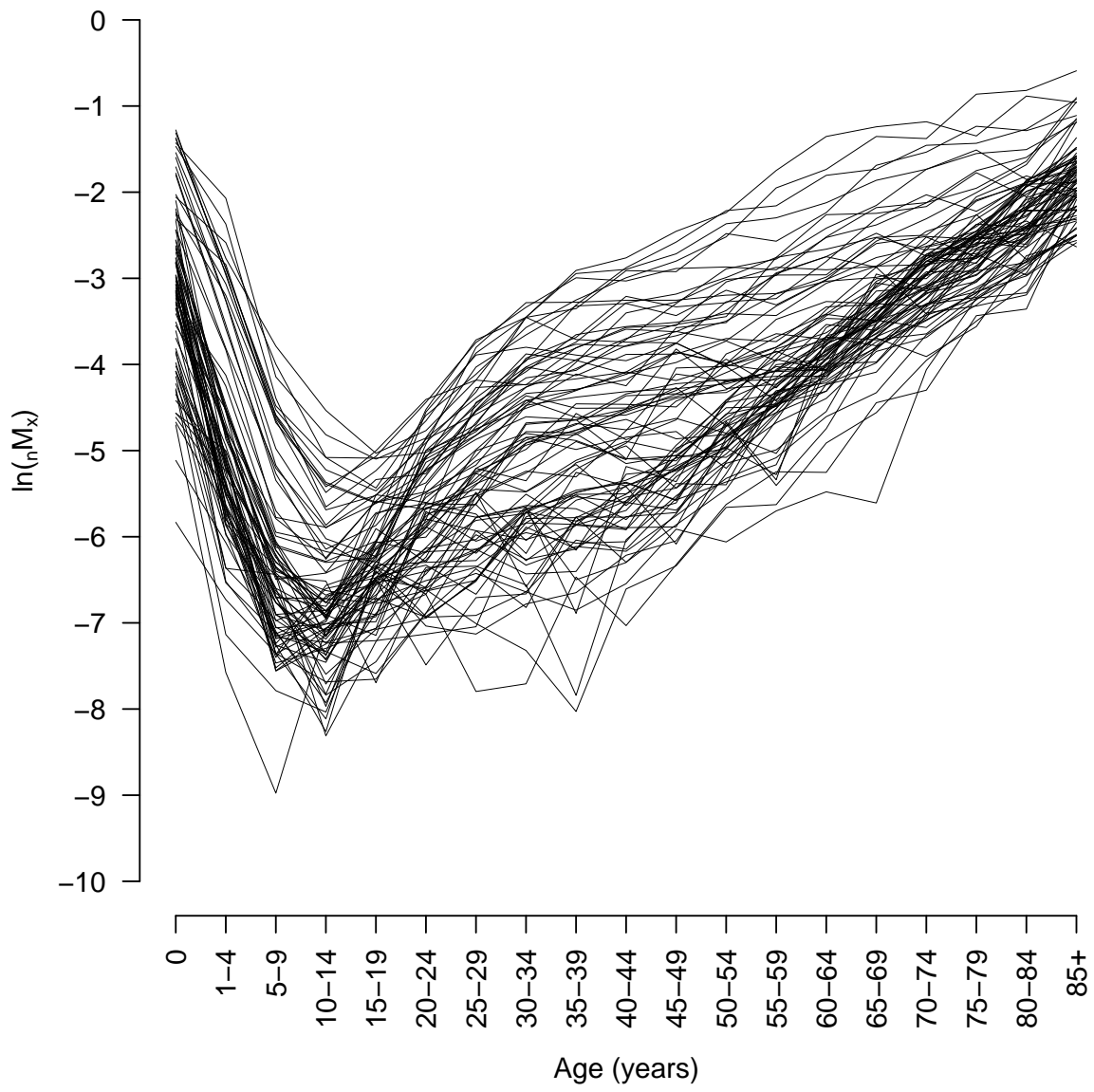
Figure 1: INDEPTH Mortality Schedules
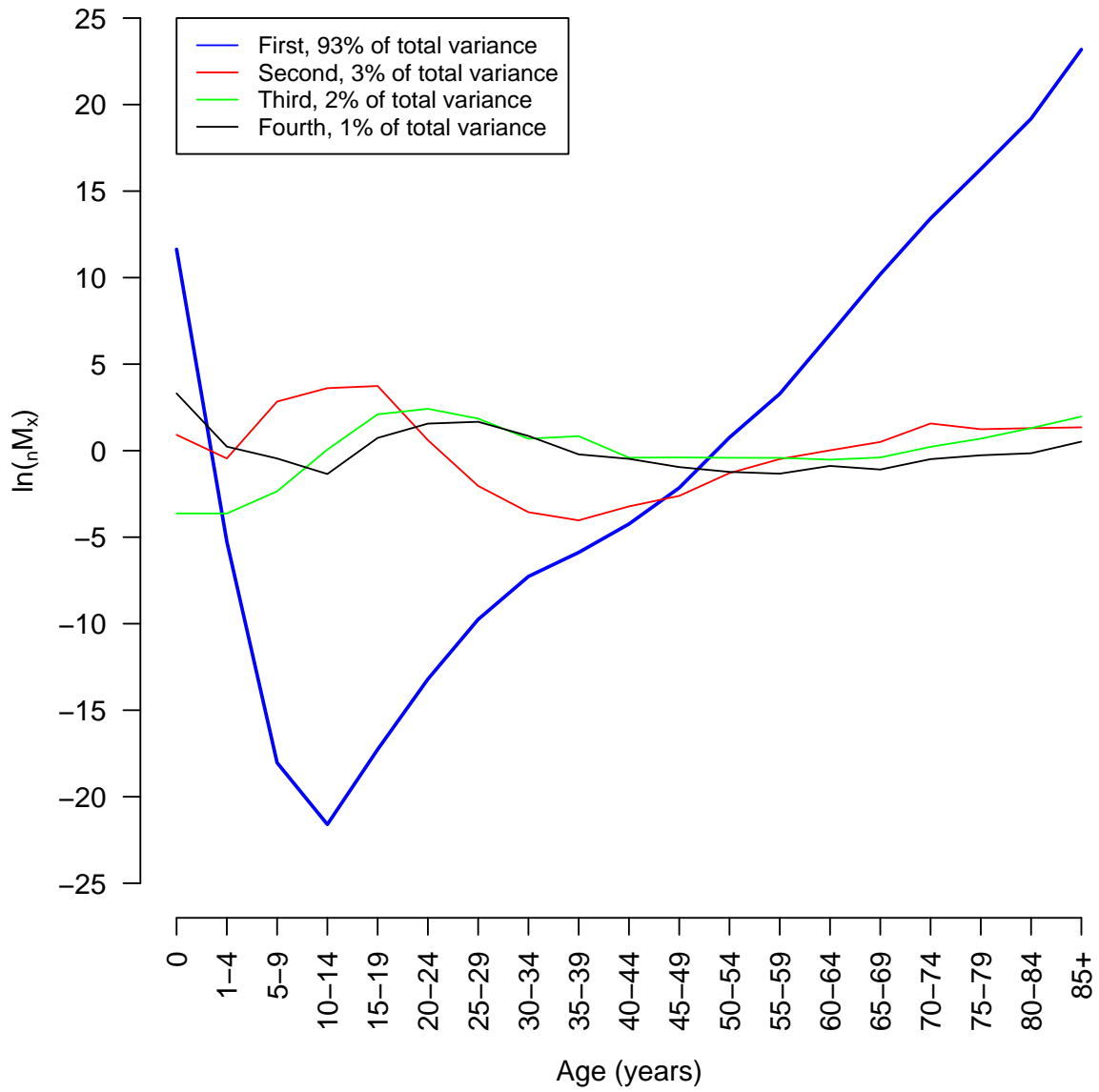
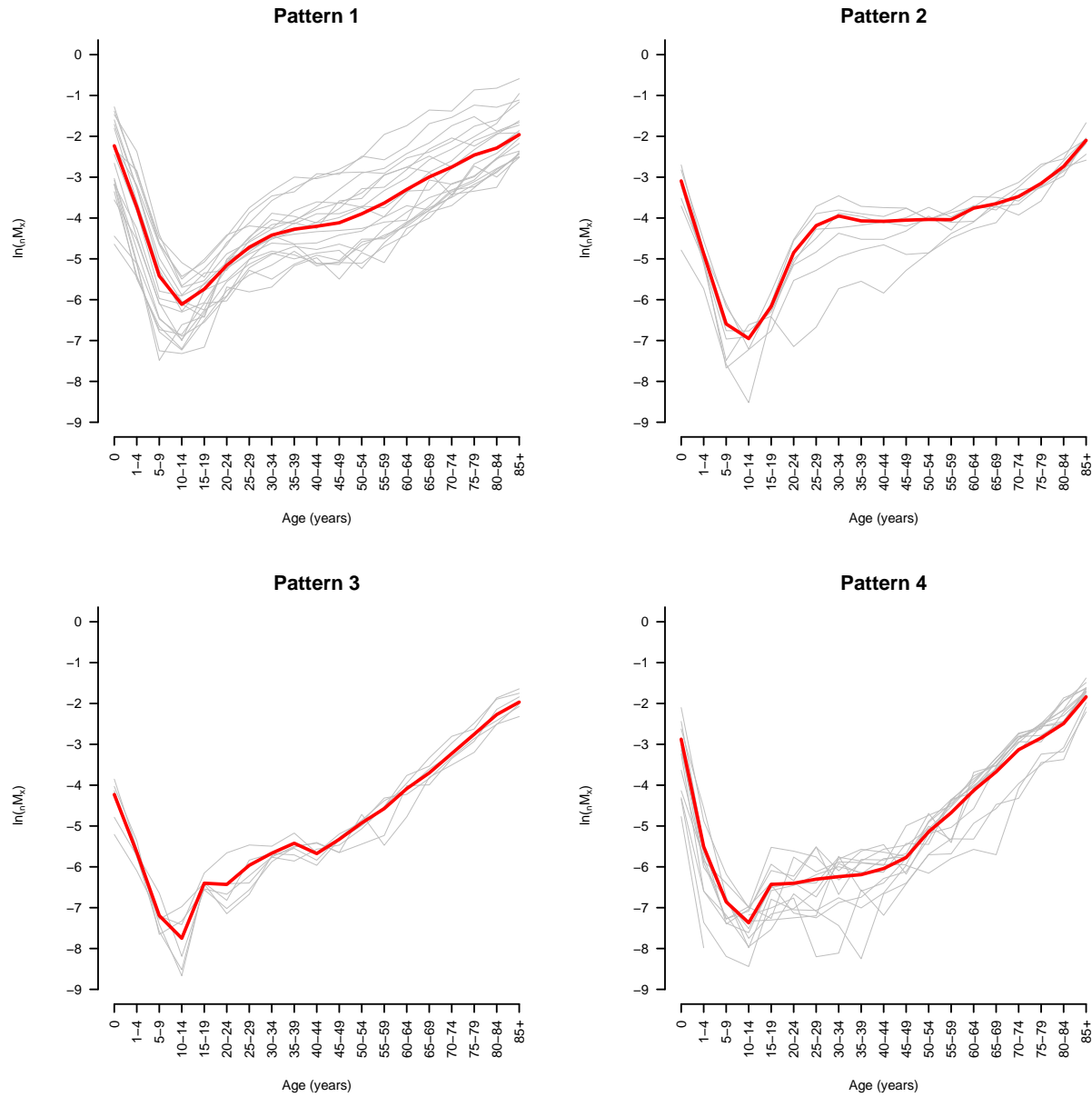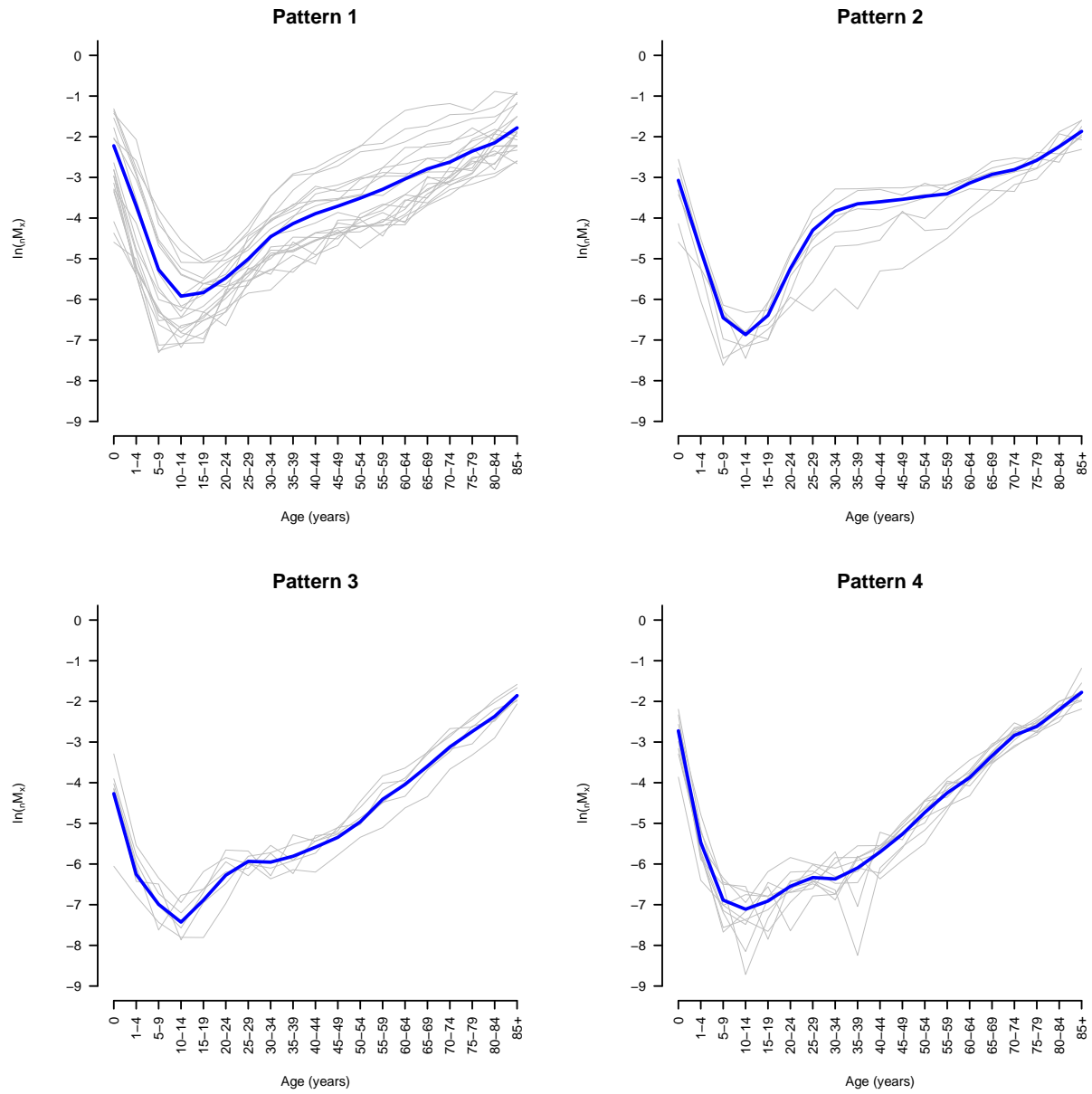Figure 2: First Four Principal Components of Mortality

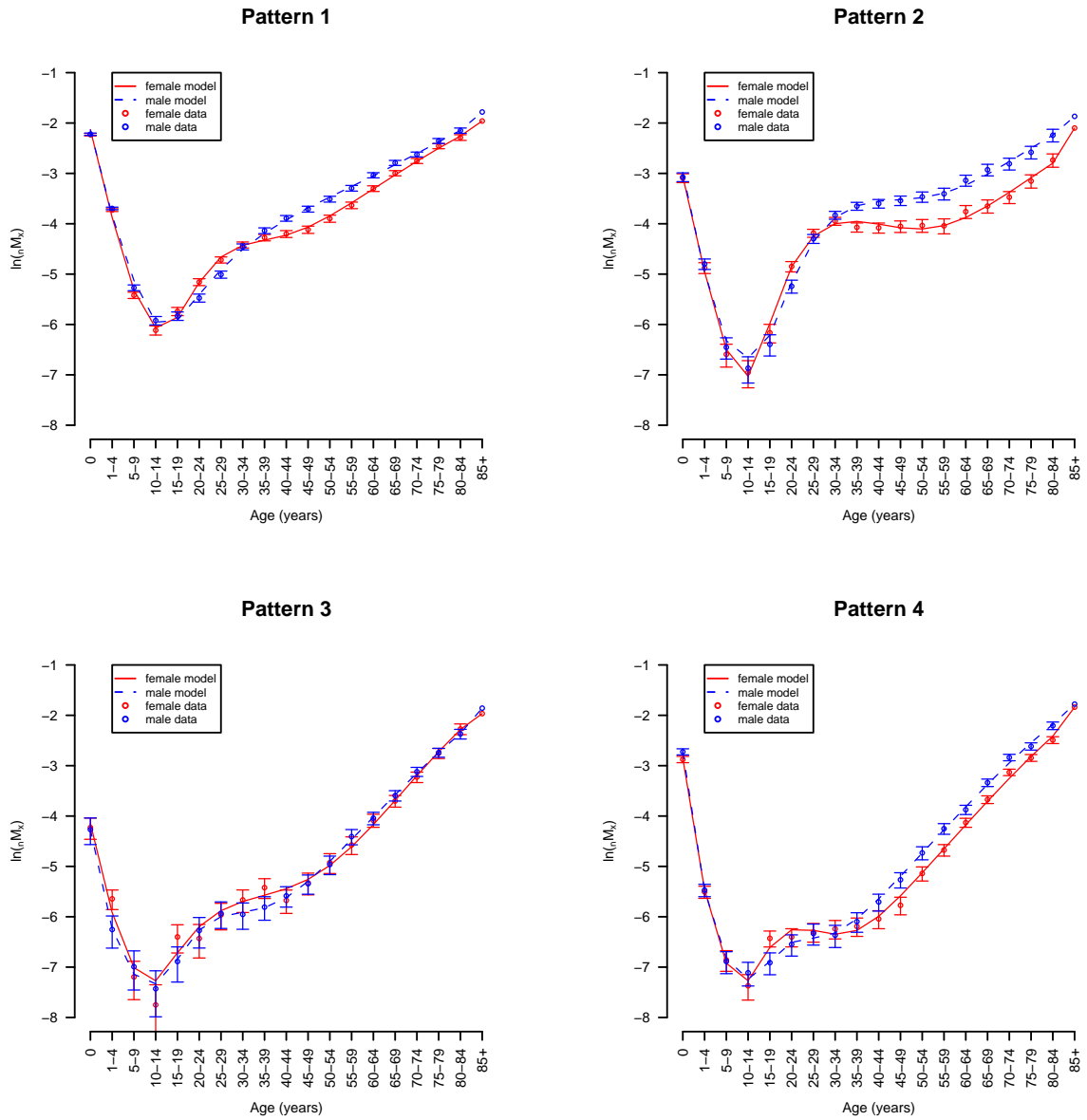Figure 3: Female Clusters.

Individual female mortality patterns included in each cluster in gray and overall female pattern for each cluster in red.
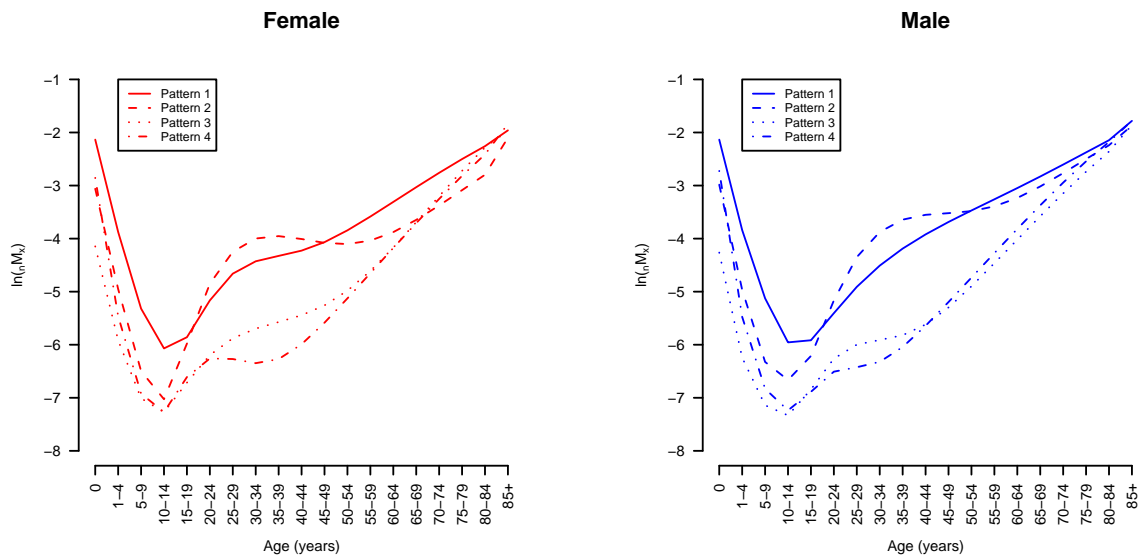
Figure 4: Male Clusters.

Individual male mortality patterns included in each cluster in gray and overall male pattern for each cluster in blue.

Figure 5: INDEPTH Mortality Patterns.

Female in red and male in blue. Overall mortality rates by sex for each cluster with 95% confidence intervals.

Figure 6: INDEPTH Mortality Patterns by Sex.

conducted the final analysis. However, Pattern 1 appears to be mainly an African pattern without HIV; Pattern 2 is mainly an African pattern composed of site periods with HIV; Pattern 3 is mainly an Asian Pattern that does not include Bangladesh; and Pattern 4 is essentially a pattern that includes Bangladesh, Indonesia and Vietnam.

Figure 5 displays the underlying mortality pattern associated with each cluster including 95% confidence intervals, and figure 6 displays all four patterns together for each sex separately. This figure makes clear how substantially different from one another the patterns are.

Finally figure 7 displays an arbitrary range of model patterns generated using the method described in section 4.5.2 for female Pattern 2. The model patterns in gray do a good job of covering the range occupied by the empirical mortality schedules in the female Pattern 2 cluster.

6   Discussion

Although the preliminary analysis presented here uses only approximately one third of the data available in the INDEPTH Network, the results are exciting. A small number of robust age patterns of mortality emerge from the empirical data, they are clearly different from each other and, especially for Pattern 2, define age-patterns of mortality that are new and different from those that exist already. Further the model proposed here appears to work

Figure 7: Representative Model Patterns, Female Pattern 2.
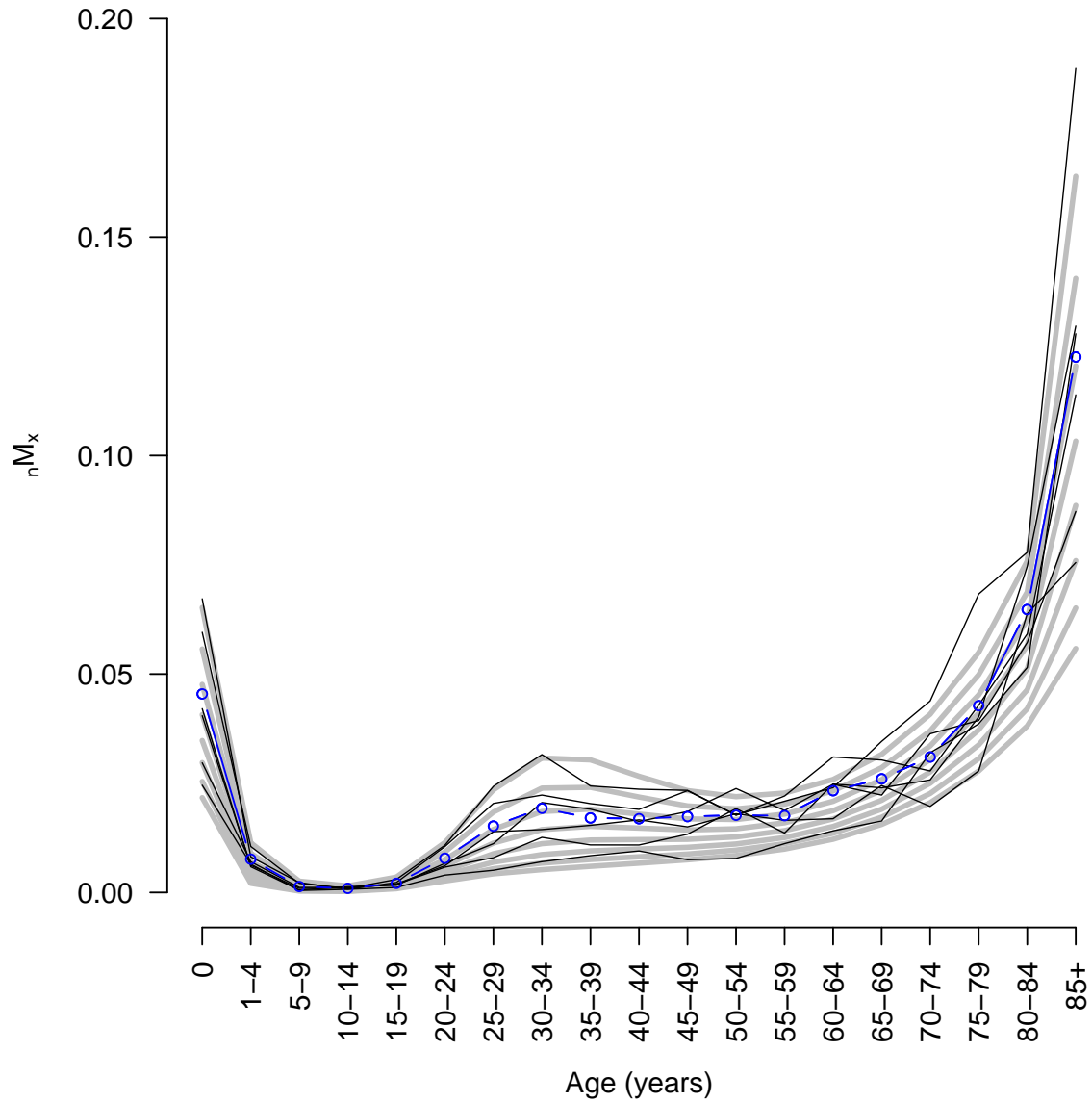
Model patterns in gray, empirical schedules in female Pattern 2 in black, overall female Pattern 2 in blue, α from -1.25 to 0.5 in 0.25 increments.

well to manipulate age-specific mortality profiles. Crucially, the simple formulation used to generate model mortality patterns works well and appears able to generate arbitrary levels of mortality within each model family using only one varying parameter. Together these demonstrations augur well for analysis of the full data set and the creation of a simple, powerful set of model life tables that are more appropriate for applications in Africa.

References

Clark, S. J. (2002). Indepth mortality patterns for africa. In INDEPTH Network (Ed.), *Population, Health, and Survival at INDEPTH Sites*, Volume 1 of *Population and Health in Developing Countries*. Ottawa: IDRC Press.

Coale, A. J. and P. Demeny (1966). *Regional Model Life Tables and Stable Populations*. Princeton, New Jersey: Princeton University Press.

Fraley, C. and A. E. Raftery (2006). Mclust version 3 for R: Normal mixture modeling and model-based clustering. Technical Report No. 504, Department of Statistics, University of Washington, USA. http://www.stat.washington.edu/mclust.

Heligman, L. and J. Pollard (1980). The age pattern of mortality. *Journal of the Institute of Actuaries 107*, 49–80.

INDEPTH Network (2008). An international network of field sites with continuous demographic evaluation of populations and their health in developing countries. http://www.indepth-network.org.

The R Foundation for Statistical Computing (2008). R. http://www.R-Project.org.

United Nations Department of International Economic and Social Affairs (1982). *Model Life Tables for Developing Countries*. Population Studies, No. 77. New York: United Nations.