

# 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**

contact: [work@samclark.net](mailto:work@samclark.net)

Center for Demography and Ecology  
University of Wisconsin - Madison  
April 28, 2009

# Acknowledgements

- **INDEPTH Member Sites**
- INDEPTH Secretariat
- INDEPTH Funders
- University of Washington Department of Sociology and CSSS, Agincourt HDSS, Farafenni DSS, Kanchanaburi DSS and APHRC UDSS for supporting their scientists to contribute time to this effort
- NIH grants 1 K01 HD057246-01 and 1 R01 HD054511-01 A1
- Greg Matthews (for help compiling data)
- Adrian Raftery (for model-based clustering slides)

- 1 Introduction
  - Motivation
  - Aims
- 2 Data
  - Structure
  - Current Status
- 3 Mortality Model
  - Model
  - Components for Mortality Model
- 4 Clustering
  - Clustering Strategy
  - Clustering Detail
  - Clusters
- 5 Model Life Tables
  - Model
  - Calculation
- 6 Discussion

# Motivation

- Mortality in Africa often measured using indirect techniques that rely on model mortality patterns:

# Motivation

- Mortality in Africa often measured using indirect techniques that rely on model mortality patterns:
  - start with child mortality measured or estimated by various surveys

# Motivation

- Mortality in Africa often measured using indirect techniques that rely on model mortality patterns:
  - start with child mortality measured or estimated by various surveys
  - **extrapolate adult mortality from child mortality**

# Motivation

- Mortality in Africa often measured using indirect techniques that rely on model mortality patterns:
  - start with child mortality measured or estimated by various surveys
  - extrapolate adult mortality from child mortality
  - or, use various indirect methods to estimate adult mortality without reference to child mortality

# Motivation

- Mortality in Africa often measured using indirect techniques that rely on model mortality patterns:
  - start with child mortality measured or estimated by various surveys
  - extrapolate adult mortality from child mortality
  - or, use various indirect methods to estimate adult mortality without reference to child mortality
- Important to have reasonable model mortality patterns fed into these methods

# Motivation

- Mortality in Africa often measured using indirect techniques that rely on model mortality patterns:
  - start with child mortality measured or estimated by various surveys
  - extrapolate adult mortality from child mortality
  - or, use various indirect methods to estimate adult mortality without reference to child mortality
- Important to have reasonable model mortality patterns fed into these methods
- Current model mortality patterns (from model life table systems) based on data from many other parts of the world, *but not Africa*

# Motivation

- Mortality in Africa often measured using indirect techniques that rely on model mortality patterns:
  - start with child mortality measured or estimated by various surveys
  - extrapolate adult mortality from child mortality
  - or, use various indirect methods to estimate adult mortality without reference to child mortality
- Important to have reasonable model mortality patterns fed into these methods
- Current model mortality patterns (from model life table systems) based on data from many other parts of the world, *but not Africa*
- We base estimates of all-age mortality in Africa on comparatively little data using model age patterns of mortality that reflect experience in other parts of the world

# Motivation

- Mortality in Africa often measured using indirect techniques that rely on model mortality patterns:
  - start with child mortality measured or estimated by various surveys
  - extrapolate adult mortality from child mortality
  - or, use various indirect methods to estimate adult mortality without reference to child mortality
- Important to have reasonable model mortality patterns fed into these methods
- Current model mortality patterns (from model life table systems) based on data from many other parts of the world, *but not Africa*
- We base estimates of all-age mortality in Africa on comparatively little data using model age patterns of mortality that reflect experience in other parts of the world

# Motivation

- Mortality in Africa often measured using indirect techniques that rely on model mortality patterns:
    - start with child mortality measured or estimated by various surveys
    - extrapolate adult mortality from child mortality
    - or, use various indirect methods to estimate adult mortality without reference to child mortality
  - Important to have reasonable model mortality patterns fed into these methods
  - Current model mortality patterns (from model life table systems) based on data from many other parts of the world, *but not Africa*
  - We base estimates of all-age mortality in Africa on comparatively little data using model age patterns of mortality that reflect experience in other parts of the world
- ⇒ We must use whatever well-measured data there are on mortality at all ages to construct model mortality patterns that better reflect the mortality experience of Africans

# Specific Aims of the Current Work

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

# Specific Aims of the Current Work

- 1 Evaluate the quality of individual-level data describing mortality from individual DSS sites
- 2 Calculate mortality rates and life tables by site, time, sex and age for all data that pass the evaluation

# Specific Aims of the Current Work

- 1 Evaluate the quality of individual-level data describing mortality from individual DSS sites
- 2 Calculate mortality rates and life tables by site, time, sex and age for all data that pass the evaluation
- 3 Identify commonly observed age patterns of mortality

# Specific Aims of the Current Work

- 1 Evaluate the quality of individual-level data describing mortality from individual DSS sites
- 2 Calculate mortality rates and life tables by site, 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

# Data from INDEPTH Network of DSS Sites

- INDEPTH comprised of  $\sim 40$  DSS sites around the world

# Data from INDEPTH Network of DSS Sites

- INDEPTH comprised of  $\sim$  40 DSS sites around the world
  - mostly Africa

# Data from INDEPTH Network of DSS Sites

- INDEPTH comprised of  $\sim$  40 DSS sites around the world
  - mostly Africa
  - Asia

# Data from INDEPTH Network of DSS Sites

- INDEPTH comprised of  $\sim 40$  DSS sites around the world
  - mostly Africa
  - Asia
  - a few in other areas, Latin America

# Data from INDEPTH Network of DSS Sites

- INDEPTH comprised of  $\sim$  40 DSS sites around the world
  - mostly Africa
  - Asia
  - a few in other areas, Latin America
- DSS = Demographic Surveillance System

# Data from INDEPTH Network of DSS Sites

- INDEPTH comprised of  $\sim$  40 DSS sites around the world
  - mostly Africa
  - Asia
  - a few in other areas, Latin America
- DSS = Demographic Surveillance System
  - monitor geographically defined populations through time

# Data from INDEPTH Network of DSS Sites

- INDEPTH comprised of  $\sim$  40 DSS sites around the world
  - mostly Africa
  - Asia
  - a few in other areas, Latin America
- DSS = Demographic Surveillance System
  - monitor geographically defined populations through time
  - births, deaths, in/out migration and a variety of other things - marriage, household structure, SES, etc.

# Data from INDEPTH Network of DSS Sites

- INDEPTH comprised of  $\sim$  40 DSS sites around the world
  - mostly Africa
  - Asia
  - a few in other areas, Latin America
- DSS = Demographic Surveillance System
  - monitor geographically defined populations through time
  - births, deaths, in/out migration and a variety of other things - marriage, household structure, SES, etc.
  - often set up to conduct clinical or behavioral trials

# Data from INDEPTH Network of DSS Sites

- INDEPTH comprised of  $\sim$  40 DSS sites around the world
  - mostly Africa
  - Asia
  - a few in other areas, Latin America
- DSS = Demographic Surveillance System
  - monitor geographically defined populations through time
  - births, deaths, in/out migration and a variety of other things - marriage, household structure, SES, etc.
  - often set up to conduct clinical or behavioral trials
- Typically monitor tens of thousands of people for decades

# Data from INDEPTH Network of DSS Sites

- INDEPTH comprised of  $\sim 40$  DSS sites around the world
  - mostly Africa
  - Asia
  - a few in other areas, Latin America
- DSS = Demographic Surveillance System
  - monitor geographically defined populations through time
  - births, deaths, in/out migration and a variety of other things - marriage, household structure, SES, etc.
  - often set up to conduct clinical or behavioral trials
- Typically monitor tens of thousands of people for decades
- Wonderful source of data on basic demographic indicators measured on whole populations - all ages and both sexes - through time

# Data from INDEPTH Network of DSS Sites

- INDEPTH comprised of  $\sim 40$  DSS sites around the world
  - mostly Africa
  - Asia
  - a few in other areas, Latin America
- DSS = Demographic Surveillance System
  - monitor geographically defined populations through time
  - births, deaths, in/out migration and a variety of other things - marriage, household structure, SES, etc.
  - often set up to conduct clinical or behavioral trials
- Typically monitor tens of thousands of people for decades
- Wonderful source of data on basic demographic indicators measured on whole populations - all ages and both sexes - through time
- <http://www.indepth-network.org>

# Data

- Individual-level exposure data were requested from sites with the following attributes:
  - site name
  - individual identifier (anonymized)
  - sex
  - date of birth
  - date of death
  - date when observation begin
  - data when observation ended
- Possible for an individual to contribute more than one exposure interval

# Current Status of Data

- 1 Received data from 26 sites

# Current Status of Data

- 1 Received data from 26 sites
- 2 Evaluated the validity and consistency of the data

# Current Status of Data

- ① Received data from 26 sites
- ② Evaluated the validity and consistency of the data
  - valid, meaningful dates

# Current Status of Data

- 1 Received data from 26 sites
- 2 Evaluated the validity and consistency of the data
  - valid, meaningful dates
  - **valid, meaningful codes**

# Current Status of Data

- 1 Received data from 26 sites
- 2 Evaluated the validity and consistency of the data
  - valid, meaningful dates
  - valid, meaningful codes
  - **temporal consistency**

# Current Status of Data

- ① Received data from 26 sites
- ② Evaluated the validity and consistency of the data
  - valid, meaningful dates
  - valid, meaningful codes
  - temporal consistency
  - consistency across multiple records for individuals

# Current Status of Data

- ① Received data from 26 sites
- ② Evaluated the validity and consistency of the data
  - valid, meaningful dates
  - valid, meaningful codes
  - temporal consistency
  - consistency across multiple records for individuals
- ③ Carefully described errors at individual level and communicated those to the relevant sites

# Current Status of Data

- 1 Received data from 26 sites
- 2 Evaluated the validity and consistency of the data
  - valid, meaningful dates
  - valid, meaningful codes
  - temporal consistency
  - consistency across multiple records for individuals
- 3 Carefully described errors at individual level and communicated those to the relevant sites
- 4 Currently waiting for a response from most of those sites

# Current Status of Data

- 1 Received data from 26 sites
- 2 Evaluated the validity and consistency of the data
  - valid, meaningful dates
  - valid, meaningful codes
  - temporal consistency
  - consistency across multiple records for individuals
- 3 Carefully described errors at individual level and communicated those to the relevant sites
- 4 Currently waiting for a response from most of those sites
- 5 Have since discovered that some valid and consistent data are producing anomalous mortality rates

# Data

- Data that pass evaluation are aggregated across time by sex within each site to produce 'site periods' with at least 50,000 person years in each

# Data

- Data that pass evaluation are aggregated across time by sex within each site to produce 'site periods' with at least 50,000 person years in each
- Data used in this preliminary analysis:

# Data

- Data that pass evaluation are aggregated across time by sex within each site to produce 'site periods' with at least 50,000 person years in each
- Data used in this preliminary analysis:
  - 17 sites

# Data

- Data that pass evaluation are aggregated across time by sex within each site to produce 'site periods' with at least 50,000 person years in each
- Data used in this preliminary analysis:
  - 17 sites
  - 82 unique site periods

# Data

- Data that pass evaluation are aggregated across time by sex within each site to produce 'site periods' with at least 50,000 person years in each
- Data used in this preliminary analysis:
  - 17 sites
  - 82 unique site periods
  - 84,000 deaths

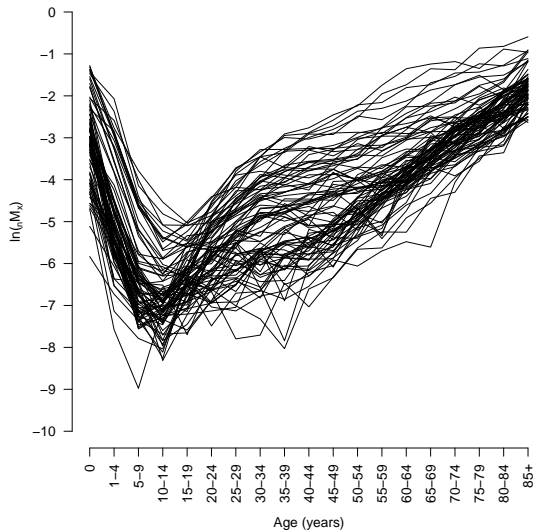
# Data

- Data that pass evaluation are aggregated across time by sex within each site to produce 'site periods' with at least 50,000 person years in each
- Data used in this preliminary analysis:
  - 17 sites
  - 82 unique site periods
  - 84,000 deaths
  - 6.5 million person years

# Data

- Data that pass evaluation are aggregated across time by sex within each site to produce 'site periods' with at least 50,000 person years in each
- Data used in this preliminary analysis:
  - 17 sites
  - 82 unique site periods
  - 84,000 deaths
  - 6.5 million person years
- When corrected data added back into analysis these numbers will increase

# Empirical Mortality Age Profiles



# Mortality Model

# Mortality Model

$$\mathbf{M} = \mathbf{S}\mathbf{B} + \mathbf{C} + \mathbf{R}$$

$19 \times m$  matrix  $\mathbf{M}$  of age-specific mortality rates is a weighted combination of age-varying components  $\mathbf{S}$  ( $19 \times n$ ), with weights  $\mathbf{B}$  ( $n \times m$ ), plus a constant vector  $\mathbf{C}$  ( $19 \times 1$ ), and possibly a vector of age-specific residuals  $\mathbf{R}$  ( $19 \times 1$ ).

(19 is the number of standard age groups 0, 1-4, 5-9  $\dots$ , 80-84, 85+)

# Mortality Model for Single Mortality Rate Schedule

For a single age-specific mortality rate schedule:

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

- $m_i$  is the age-specific mortality rate for age group  $i$
- $s_{i,j}$  is the value of component vector  $j$  for age  $i$
- $c$  is the constant, non-age-varying component of mortality
- $r_i$  is the residual for age group  $i$
- $b_i$  is the value of the weight given to component vector ( $i = j$ )

# Principal Components

- The 'component' vectors **S** in the mortality model come from a principal component (PC) analysis of the empirical mortality rate schedules

# Principal Components

- The 'component' vectors  $\mathbf{S}$  in the mortality model come from a principal component (PC) analysis of the empirical mortality rate schedules
- Each site period treated as a 'variable' or dimension

# Principal Components

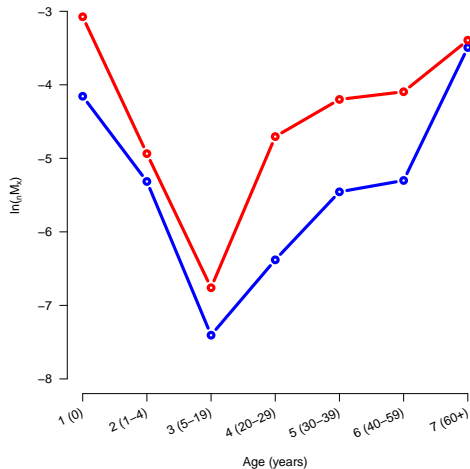
- The 'component' vectors  $\mathbf{S}$  in the mortality model come from a principal component (PC) analysis of the empirical mortality rate schedules
- Each site period treated as a 'variable' or dimension
- PC analysis produces a new set of dimensions that concentrate the information contained in the original site periods in a small number dimensions

# Principal Components

- The 'component' vectors  $\mathbf{S}$  in the mortality model come from a principal component (PC) analysis of the empirical mortality rate schedules
- Each site period treated as a 'variable' or dimension
- PC analysis produces a new set of dimensions that concentrate the information contained in the original site periods in a small number dimensions
- The components we use in the model are the 'scores' produced by PC analysis

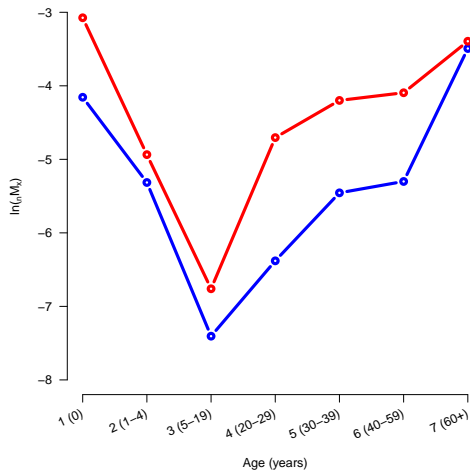
# Example

- Logged mortality rates from Agincourt DSS



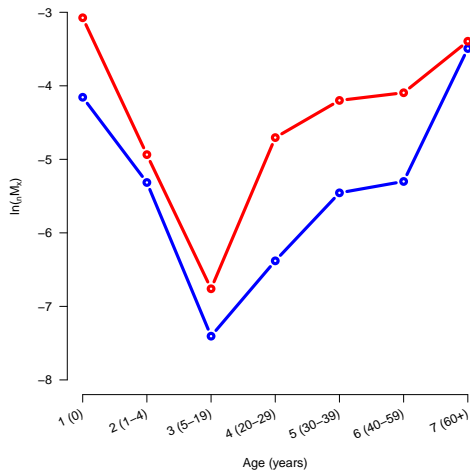
# Example

- Logged mortality rates from Agincourt DSS
  - 1995 - blue



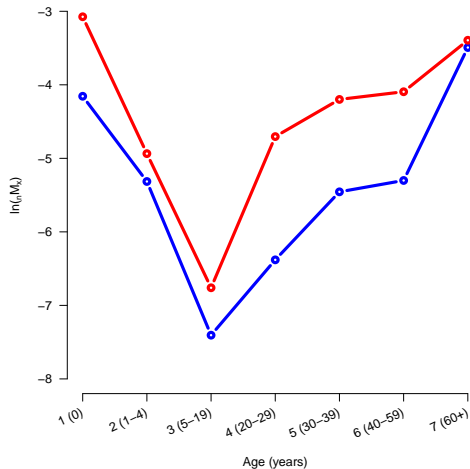
# Example

- Logged mortality rates from Agincourt DSS
  - 1995 - blue
  - 2007 - red



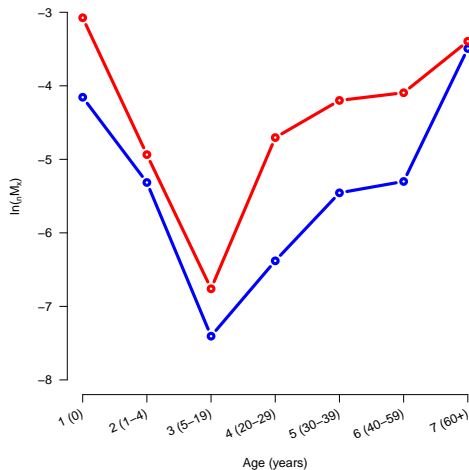
# Example

- Logged mortality rates from Agincourt DSS
  - 1995 - blue
  - 2007 - red
- Wide age groups



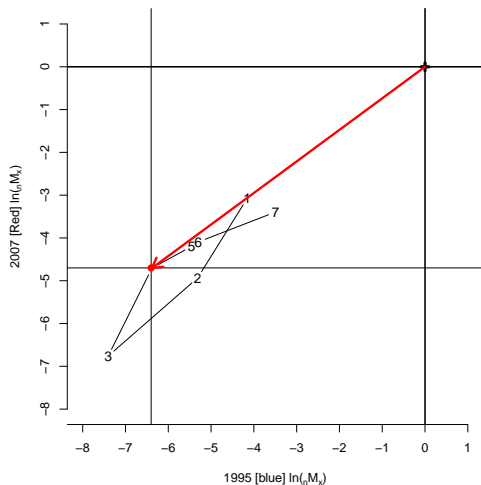
# Example

- Logged mortality rates from Agincourt DSS
  - 1995 - blue
  - 2007 - red
- Wide age groups
- 2007 schedule shows dramatic influence of HIV; provides clear contrast



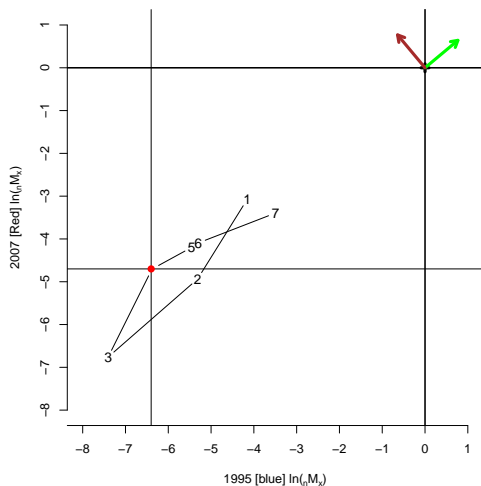
# Data

	age	X:1995	Y:2007
1	0	-4.2	-3.1
2	1-4	-5.3	-4.9
3	5-19	-7.4	-6.8
<b>4</b>	<b>20-29</b>	<b>-6.4</b>	<b>-4.7</b>
5	30-39	-5.5	-4.2
6	40-59	-5.3	-4.1
7	60+	-3.5	-3.4



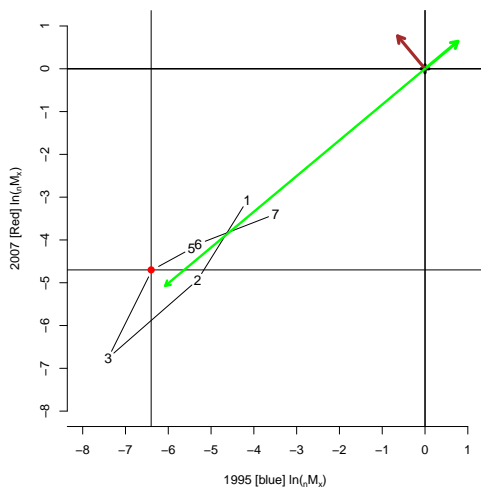
# New Basis: eigenvectors

	X:1995	Y:2007	% $\sigma^2$
E1	0.76	0.64	0.997
E2	-0.64	0.76	0.003



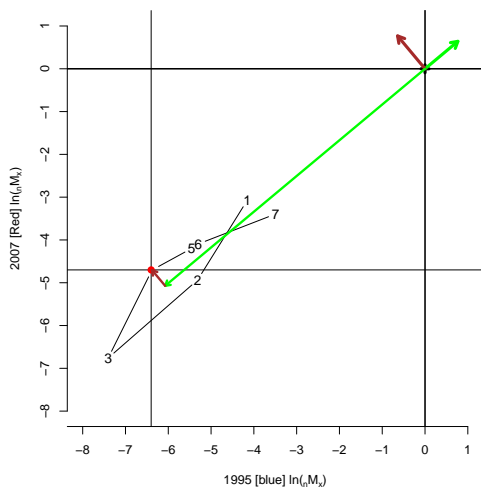
# Scores: along eigenvector 1

	age	S1	S2
1	0	-5.16	0.30
2	1-4	-7.24	-0.38
3	5-19	-10.02	-0.44
<b>4</b>	<b>20-29</b>	<b>-7.91</b>	<b>0.48</b>
5	30-39	-6.88	0.27
6	40-59	-6.69	0.26
7	60+	-4.86	-0.37



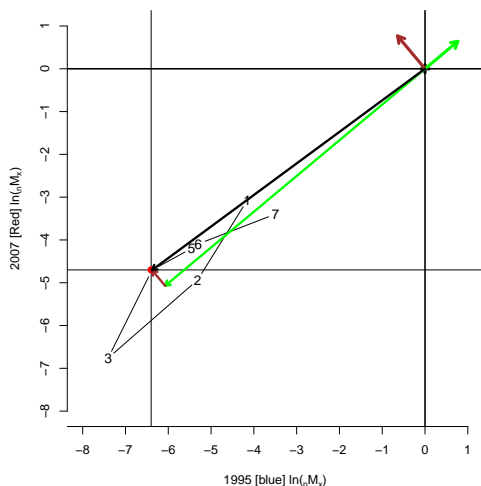
## Scores: along eigenvector 2

	age	S1	S2
1	0	-5.16	0.30
2	1-4	-7.24	-0.38
3	5-19	-10.02	-0.44
<b>4</b>	<b>20-29</b>	<b>-7.91</b>	<b>0.48</b>
5	30-39	-6.88	0.27
6	40-59	-6.69	0.26
7	60+	-4.86	-0.37



$\ln(10\mathbf{M}_{20})$ : sum of components along each eigenvector

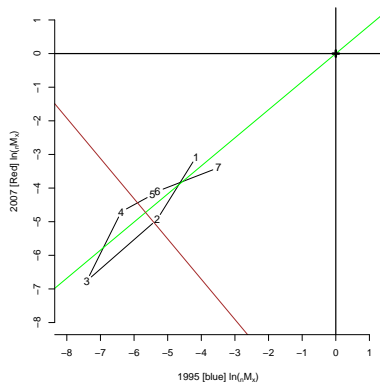
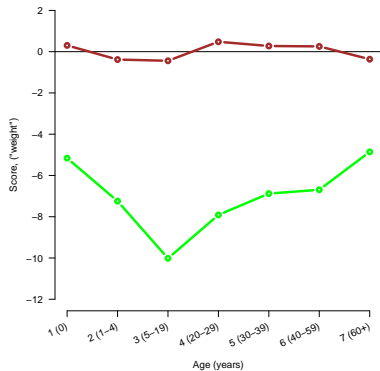
	age	S1	S2
1	0	-5.16	0.30
2	1-4	-7.24	-0.38
3	5-19	-10.02	-0.44
<b>4</b>	<b>20-29</b>	<b>-7.91</b>	<b>0.48</b>
5	30-39	-6.88	0.27
6	40-59	-6.69	0.26
7	60+	-4.86	-0.37



## Putting it together with equations

$$\begin{aligned}\ln({}_{10}\mathbf{M}_{20}) &= S1 \cdot \mathbf{E1} + S2 \cdot \mathbf{E2} \\ &= S1 \cdot (E1_1, E1_2) + S2 \cdot (E2_1, E2_2) \\ &= (S1 \cdot E1_1 + S2 \cdot E2_1, S1 \cdot E1_2 + S2 \cdot E2_2)\end{aligned}$$

# Scores with graphs

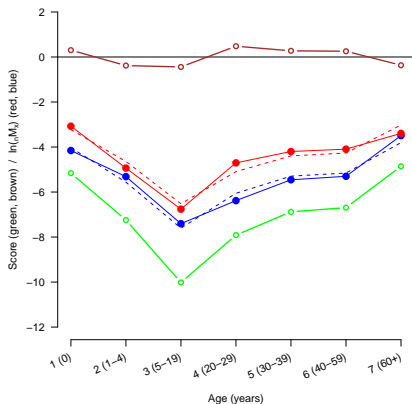


# Using scores as components in model of mortality

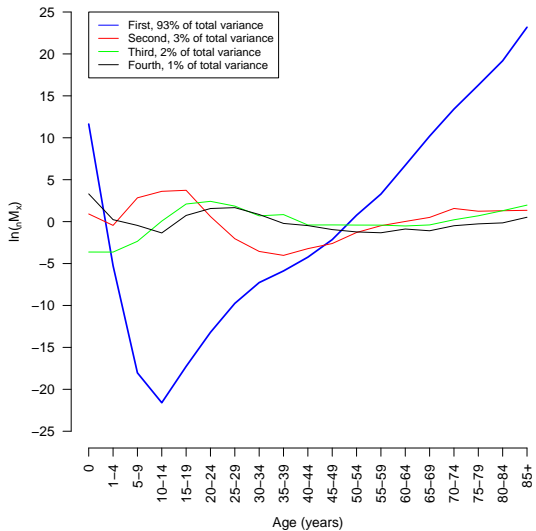
Regressions of empirical mortality schedules (red and blue dots) on:

- 1 both score vectors (green and brown);  
predicted values = solid lines
- 2 only the first score vector associated with the large variance direction (green);  
predicted values = dotted lines

# Comps	$\alpha$	$\beta_0$	$\beta_1$	$R^2$
1995				
2	0.00	0.077	-0.064	1.00
1	-0.23	0.74	-	0.96
2007				
2	0.00	0.064	0.077	1.00
1	0.27	0.68	-	0.94



# Components of INDEPTH Mortality



# Identification of Similar Age Patterns of Mortality

- Goal: group site periods into a small number of 'clusters' that each contain very similar age patterns

# Identification of Similar Age Patterns of Mortality

- Goal: group site periods into a small number of 'clusters' that each contain very similar age patterns
- Ignore 'level' of mortality in this clustering, focus on age variation

# Identification of Similar Age Patterns of Mortality

- Goal: group site periods into a small number of 'clusters' that each contain very similar age patterns
- Ignore 'level' of mortality in this clustering, focus on age variation
- Clustering

# Identification of Similar Age Patterns of Mortality

- Goal: group site periods into a small number of 'clusters' that each contain very similar age patterns
- Ignore 'level' of mortality in this clustering, focus on age variation
- Clustering
  - ④ Simplify data - reduce number of dimensions and concentrate variance

# Identification of Similar Age Patterns of Mortality

- Goal: group site periods into a small number of 'clusters' that each contain very similar age patterns
- Ignore 'level' of mortality in this clustering, focus on age variation
- Clustering
  - 1 Simplify data - reduce number of dimensions and concentrate variance
  - 2 Regress each empirical mortality schedule on the first few components derived from the PC analysis

# Identification of Similar Age Patterns of Mortality

- Goal: group site periods into a small number of 'clusters' that each contain very similar age patterns
- Ignore 'level' of mortality in this clustering, focus on age variation
- Clustering
  - 1 Simplify data - reduce number of dimensions and concentrate variance
  - 2 Regress each empirical mortality schedule on the first few components derived from the PC analysis
  - 3 **Keep the coefficients and constants from these regressions → new 'reduced' data set**

# Identification of Similar Age Patterns of Mortality

- Goal: group site periods into a small number of 'clusters' that each contain very similar age patterns
- Ignore 'level' of mortality in this clustering, focus on age variation
- Clustering
  - ① Simplify data - reduce number of dimensions and concentrate variance
  - ② Regress each empirical mortality schedule on the first few components derived from the PC analysis
  - ③ Keep the coefficients and constants from these regressions → new 'reduced' data set
- Perform a cluster analysis on the coefficients (ignoring the constants that correspond to non-age-varying level) in this reduced data set using the *model-based clustering* method

# Model-Based Clustering (courtesy of Adrian Raftery)

- Model-based clustering puts cluster analysis on a solid statistical basis and answers questions such as:

# Model-Based Clustering (courtesy of Adrian Raftery)

- Model-based clustering puts cluster analysis on a solid statistical basis and answers questions such as:
  - How many groups?

# Model-Based Clustering (courtesy of Adrian Raftery)

- Model-based clustering puts cluster analysis on a solid statistical basis and answers questions such as:
  - How many groups?
  - Which clustering method to use?

# Model-Based Clustering (courtesy of Adrian Raftery)

- Model-based clustering puts cluster analysis on a solid statistical basis and answers questions such as:
  - How many groups?
  - Which clustering method to use?
  - How certain can we be about the clustering?

# Model-Based Clustering (courtesy of Adrian Raftery)

- Model-based clustering puts cluster analysis on a solid statistical basis and answers questions such as:
  - How many groups?
  - Which clustering method to use?
  - How certain can we be about the clustering?
- **Model-based clustering:**

# Model-Based Clustering (courtesy of Adrian Raftery)

- Model-based clustering puts cluster analysis on a solid statistical basis and answers questions such as:
  - How many groups?
  - Which clustering method to use?
  - How certain can we be about the clustering?
- Model-based clustering:
  - *A framework for cluster analysis*

# Model-Based Clustering (courtesy of Adrian Raftery)

- Model-based clustering puts cluster analysis on a solid statistical basis and answers questions such as:
  - How many groups?
  - Which clustering method to use?
  - How certain can we be about the clustering?
- Model-based clustering:
  - A *framework* for cluster analysis
  - **Bases cluster analysis on a statistical (mixture) model:**  
 $y \sim \sum_{g=1}^G \tau_g f_g(y)$ , where  $y$  is data and  $f_g(\cdot)$  are distributions

# Model-Based Clustering (courtesy of Adrian Raftery)

- Model-based clustering puts cluster analysis on a solid statistical basis and answers questions such as:
  - How many groups?
  - Which clustering method to use?
  - How certain can we be about the clustering?
- Model-based clustering:
  - A *framework* for cluster analysis
  - Bases cluster analysis on a statistical (mixture) model:  
 $y \sim \sum_{g=1}^G \tau_g f_g(y)$ , where  $y$  is data and  $f_g(\cdot)$  are distributions
  - Gives answers to questions based on standard statistical principles

# Model-Based Clustering (courtesy of Adrian Raftery)

- Model-based clustering puts cluster analysis on a solid statistical basis and answers questions such as:
  - How many groups?
  - Which clustering method to use?
  - How certain can we be about the clustering?
- Model-based clustering:
  - A *framework* for cluster analysis
  - Bases cluster analysis on a statistical (mixture) model:  
 $y \sim \sum_{g=1}^G \tau_g f_g(y)$ , where  $y$  is data and  $f_g(\cdot)$  are distributions
  - Gives answers to questions based on standard statistical principles
- Software: R packages available at <http://cran.r-project.org>:

# Model-Based Clustering (courtesy of Adrian Raftery)

- Model-based clustering puts cluster analysis on a solid statistical basis and answers questions such as:
  - How many groups?
  - Which clustering method to use?
  - How certain can we be about the clustering?
- Model-based clustering:
  - A *framework* for cluster analysis
  - Bases cluster analysis on a statistical (mixture) model:  
 $y \sim \sum_{g=1}^G \tau_g f_g(y)$ , where  $y$  is data and  $f_g(\cdot)$  are distributions
  - Gives answers to questions based on standard statistical principles
- Software: R packages available at <http://cran.r-project.org>:
  - **model-based clustering package: mclust**

# Model-Based Clustering (courtesy of Adrian Raftery)

- Model-based clustering puts cluster analysis on a solid statistical basis and answers questions such as:
  - How many groups?
  - Which clustering method to use?
  - How certain can we be about the clustering?
- Model-based clustering:
  - A *framework* for cluster analysis
  - Bases cluster analysis on a statistical (mixture) model:  
 $y \sim \sum_{g=1}^G \tau_g f_g(y)$ , where  $y$  is data and  $f_g(\cdot)$  are distributions
  - Gives answers to questions based on standard statistical principles
- Software: R packages available at <http://cran.r-project.org>:
  - model-based clustering package: `mclust`
- **References:**

# Model-Based Clustering (courtesy of Adrian Raftery)

- Model-based clustering puts cluster analysis on a solid statistical basis and answers questions such as:
  - How many groups?
  - Which clustering method to use?
  - How certain can we be about the clustering?
- Model-based clustering:
  - A *framework* for cluster analysis
  - Bases cluster analysis on a statistical (mixture) model:  
 $y \sim \sum_{g=1}^G \tau_g f_g(y)$ , where  $y$  is data and  $f_g(\cdot)$  are distributions
  - Gives answers to questions based on standard statistical principles
- Software: R packages available at <http://cran.r-project.org>:
  - model-based clustering package: `mclust`
- References:
  - Model-based clustering: Banfield and Raftery (1993, *Biometrics*), Fraley and Raftery (2002, *J. Amer. Statist. Ass.*)

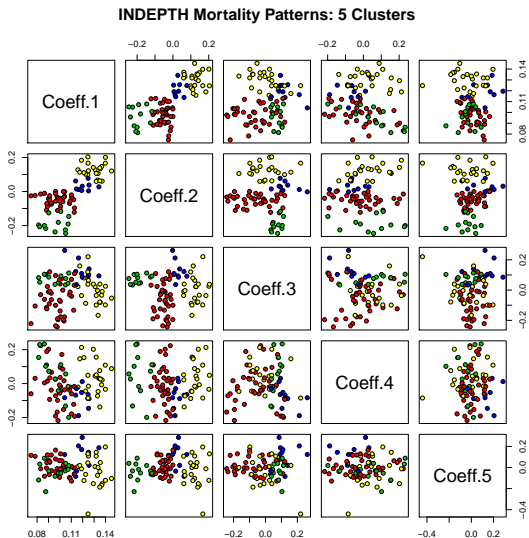
# Model-Based Clustering (courtesy of Adrian Raftery)

- Model-based clustering puts cluster analysis on a solid statistical basis and answers questions such as:
  - How many groups?
  - Which clustering method to use?
  - How certain can we be about the clustering?
- Model-based clustering:
  - A *framework* for cluster analysis
  - Bases cluster analysis on a statistical (mixture) model:  
 $y \sim \sum_{g=1}^G \tau_g f_g(y)$ , where  $y$  is data and  $f_g(\cdot)$  are distributions
  - Gives answers to questions based on standard statistical principles
- Software: R packages available at <http://cran.r-project.org>:
  - model-based clustering package: `mclust`
- References:
  - Model-based clustering: Banfield and Raftery (1993, *Biometrics*), Fraley and Raftery (2002, *J. Amer. Statist. Ass.*)
- Website: [www.stat.washington.edu/raftery](http://www.stat.washington.edu/raftery)  
→ Research → Model-based clustering

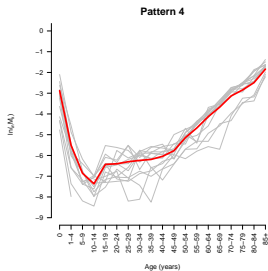
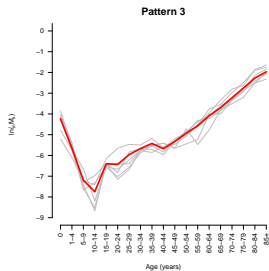
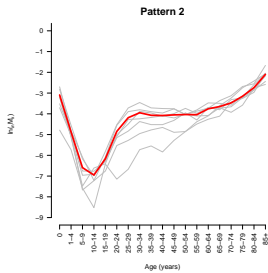
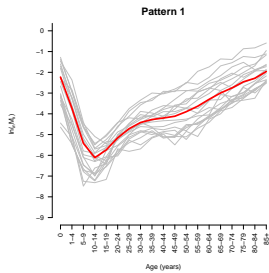
## Reduced Data Set

Coeff.1	Coeff.2	Coeff.3	Coeff.4	Coeff.5	clust
0.09	-0.07	-0.06	0.07	-0.03	1
0.09	-0.04	-0.09	0.07	-0.07	1
0.09	-0.04	-0.01	0.20	0.07	1
.	.	.	.	.	.
0.11	-0.25	0.08	0.06	0.04	2
0.09	-0.20	0.10	0.23	0.09	2
0.08	-0.21	0.05	0.23	0.03	2
.	.	.	.	.	.
0.11	0.06	0.12	-0.07	-0.08	3
0.13	0.04	0.14	-0.09	0.02	3
0.12	0.08	0.09	-0.06	0.12	3
.	.	.	.	.	.
0.14	0.15	-0.03	-0.01	-0.17	4
0.14	0.17	-0.10	0.11	0.19	4
0.12	0.20	0.04	0.05	0.13	4

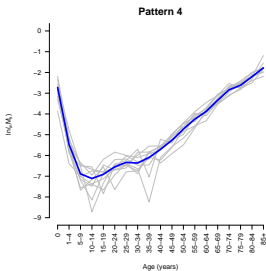
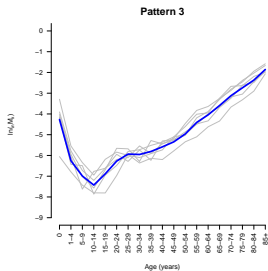
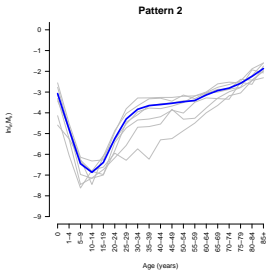
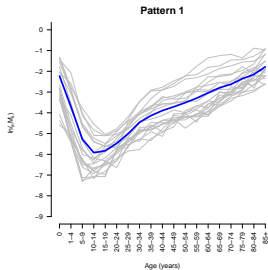
# Clustered Coefficients Scatter Plots - Five Coefficients



# Female Clusters

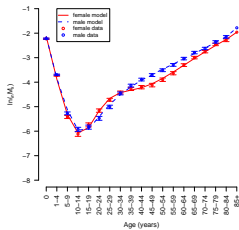


# Male Clusters

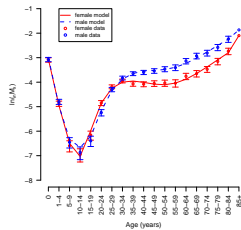


# INDEPTH Mortality Patterns

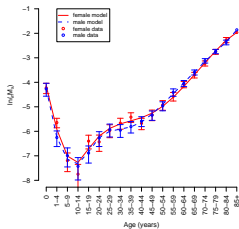
Pattern 1



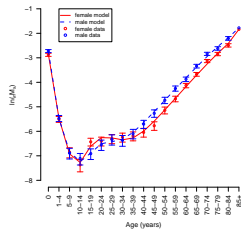
Pattern 2



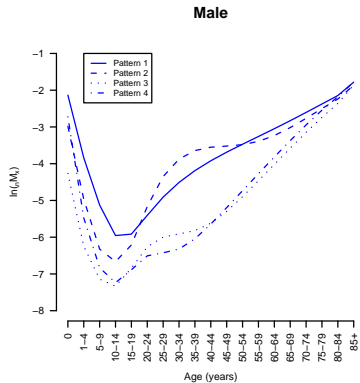
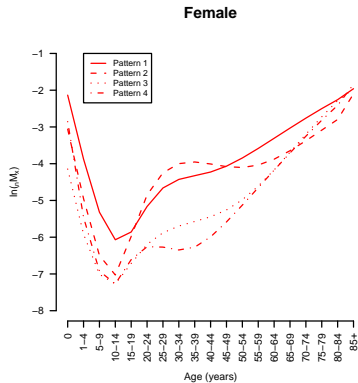
Pattern 3



Pattern 4



# INDEPTH Mortality Patterns by Sex



# Constructing Model Life Tables

- ④ Model life table system organized into families with various (arbitrary) levels of mortality within each family

# Constructing Model Life Tables

- 1 Model life table system organized into families with various (arbitrary) levels of mortality within each family
- 2 Each family:

# Constructing Model Life Tables

- 1 Model life table system organized into families with various (arbitrary) levels of mortality within each family
- 2 Each family:
  - represents a different underlying age pattern of mortality

# Constructing Model Life Tables

- 1 Model life table system organized into families with various (arbitrary) levels of mortality within each family
- 2 Each family:
  - represents a different underlying age pattern of mortality
  - is based on one of the clusters identified in the empirical data

# Constructing Model Life Tables

- 1 Model life table system organized into families with various (arbitrary) levels of mortality within each family
- 2 Each family:
  - represents a different underlying age pattern of mortality
  - is based on one of the clusters identified in the empirical data
- 3 The structure of the model is:

$$[\text{underlying age pattern}] + \alpha \cdot [\text{family-age-specific deviation}]$$

where  $\alpha$  varies to create different mortality levels within the family

# Underlying Mortality Pattern for each Family

- Each empirical cluster defines a characteristic age pattern that gives rise to a family in the model life table system

# Underlying Mortality Pattern for each Family

- Each empirical cluster defines a characteristic age pattern that gives rise to a family in the model life table system
- The 'underlying' age pattern for each family is obtained from the empirical data by aggregating across site periods in a cluster:

# Underlying Mortality Pattern for each Family

- Each empirical cluster defines a characteristic age pattern that gives rise to a family in the model life table system
- The 'underlying' age pattern for each family is obtained from the empirical data by aggregating across site periods in a cluster:
  - sum deaths and person years for each age group in the cluster

# Underlying Mortality Pattern for each Family

- Each empirical cluster defines a characteristic age pattern that gives rise to a family in the model life table system
- The 'underlying' age pattern for each family is obtained from the empirical data by aggregating across site periods in a cluster:
  - sum deaths and person years for each age group in the cluster
  - **divide these to create a new cluster/family-specific set of age-specific mortality rates**

## Age-Varying Levels within each Family

Within a cluster, the age-specific change (**D**) from low to high levels within the cluster can be captured by:

$$\begin{aligned}\mathbf{D} &= \mathbf{M}_h - \mathbf{M}_l \\ &= [\mathbf{S}\mathbf{B}_h + \mathbf{C}_h] - [\mathbf{S}\mathbf{B}_l + \mathbf{C}_l] \\ &= \mathbf{S}[\mathbf{B}_h - \mathbf{B}_l] + [\mathbf{C}_h - \mathbf{C}_l] \\ &= \mathbf{S}\mathbf{\Delta} + \delta\end{aligned}$$

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

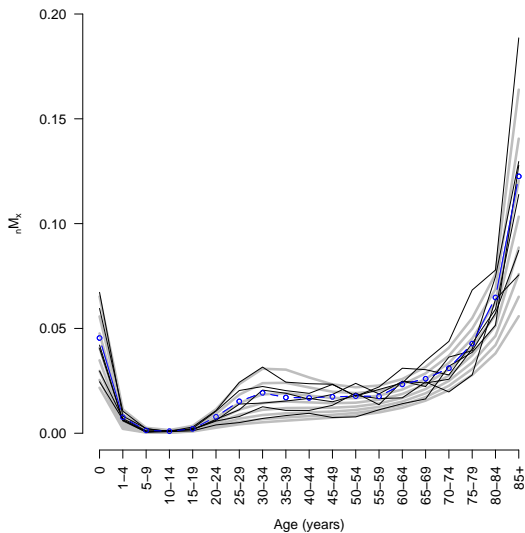
# Model Life Table Calculation

Individual model life tables can be calculated by varying  $\alpha$  in:

$$\mathbf{M} = \mathbf{S} [\mathbf{B} + \alpha \mathbf{\Delta}] + [\mathbf{C} + \alpha \delta]$$

where  $\mathbf{B}$ ,  $\mathbf{\Delta}$  and  $\delta$  are indexed by family.

## Example Model Mortality Patterns Female Pattern 2



## Discussion - where are we going ?

- *This has been a preview !*

## Discussion - where are we going ?

- *This has been a preview !*
- **Waiting for sites to address data issues**

## Discussion - where are we going ?

- *This has been a preview !*
- Waiting for sites to address data issues
- Complete a final analysis along the lines suggested here

## Discussion - where are we going ?

- *This has been a preview !*
- Waiting for sites to address data issues
- Complete a final analysis along the lines suggested here
- Create a set of model life tables based on these patterns that can be printed

## Discussion - where are we going ?

- *This has been a preview !*
- Waiting for sites to address data issues
- Complete a final analysis along the lines suggested here
- Create a set of model life tables based on these patterns that can be printed
- Create a more flexible electronic version that can produce lifetables at arbitrary levels of life expectancy:

## Discussion - where are we going ?

- *This has been a preview !*
- Waiting for sites to address data issues
- Complete a final analysis along the lines suggested here
- Create a set of model life tables based on these patterns that can be printed
- Create a more flexible electronic version that can produce lifetables at arbitrary levels of life expectancy:
  - **Excel spreadsheet**

## Discussion - where are we going ?

- *This has been a preview !*
- Waiting for sites to address data issues
- Complete a final analysis along the lines suggested here
- Create a set of model life tables based on these patterns that can be printed
- Create a more flexible electronic version that can produce lifetables at arbitrary levels of life expectancy:
  - Excel spreadsheet
  - R package

# Discussion - where are we going ?

- *This has been a preview !*
- Waiting for sites to address data issues
- Complete a final analysis along the lines suggested here
- Create a set of model life tables based on these patterns that can be printed
- Create a more flexible electronic version that can produce lifetables at arbitrary levels of life expectancy:
  - Excel spreadsheet
  - R package
  - **someone - not me - can produce a stata .ado implementation**

# Discussion - where are we going ?

- *This has been a preview !*
- Waiting for sites to address data issues
- Complete a final analysis along the lines suggested here
- Create a set of model life tables based on these patterns that can be printed
- Create a more flexible electronic version that can produce lifetables at arbitrary levels of life expectancy:
  - Excel spreadsheet
  - R package
  - someone - not me - can produce a stata .ado implementation
- Publish these electronic materials on the web

# Where else are we going ?

- Complete other sections of *INDEPTH Monograph on Mortality*

# Where else are we going ?

- Complete other sections of *INDEPTH Monograph on Mortality*
  - site chapters

# Where else are we going ?

- Complete other sections of *INDEPTH Monograph on Mortality*
  - site chapters
  - description of mortality levels and trends

# Where else are we going ?

- Complete other sections of *INDEPTH Monograph on Mortality*
  - site chapters
  - description of mortality levels and trends
  - comparison of mortality indicators across sites

# Where else are we going ?

- Complete other sections of *INDEPTH Monograph on Mortality*
  - site chapters
  - description of mortality levels and trends
  - comparison of mortality indicators across sites
  - . . .

# Where else are we going ?

- Complete other sections of *INDEPTH Monograph on Mortality*
  - site chapters
  - description of mortality levels and trends
  - comparison of mortality indicators across sites
  - ...
- Publish monograph and a summary article

# Where else are we going ?

- Complete other sections of *INDEPTH Monograph on Mortality*
  - site chapters
  - description of mortality levels and trends
  - comparison of mortality indicators across sites
  - ...
- Publish monograph and a summary article
- Investigate the potential for a more comprehensive analysis of mortality including *cause* and possibly covariates:

# Where else are we going ?

- Complete other sections of *INDEPTH Monograph on Mortality*
  - site chapters
  - description of mortality levels and trends
  - comparison of mortality indicators across sites
  - ...
- Publish monograph and a summary article
- Investigate the potential for a more comprehensive analysis of mortality including *cause* and possibly covariates:
  - HIV prevalence

# Where else are we going ?

- Complete other sections of *INDEPTH Monograph on Mortality*
  - site chapters
  - description of mortality levels and trends
  - comparison of mortality indicators across sites
  - ...
- Publish monograph and a summary article
- Investigate the potential for a more comprehensive analysis of mortality including *cause* and possibly covariates:
  - HIV prevalence
  - **SES**

# Where else are we going ?

- Complete other sections of *INDEPTH Monograph on Mortality*
  - site chapters
  - description of mortality levels and trends
  - comparison of mortality indicators across sites
  - ...
- Publish monograph and a summary article
- Investigate the potential for a more comprehensive analysis of mortality including *cause* and possibly covariates:
  - HIV prevalence
  - SES
  - family structure, etc.

