Estimating population-level trends in cardiometabolic risk factors using disparate data sources

Mariel Finucane¹, Chris Paciorek*,¹,², Goodarz Danaei³, and Majid Ezzati⁴

*paciorek@stat.berkeley.edu;
www.biostat.harvard.edu/~paciorek

¹ Harvard School of Public Health, Department of Biostatistics
² UC Berkeley, Department of Statistics
³ Harvard School of Public Health, Department of Epidemiology
⁴ Imperial College London

December 8, 2010
Risk Factor Modeling

Outline

1. Background
   Overview
   Data

2. Hierarchical Modeling
   Covariate effects
   Country-region hierarchy
   Nonlinear change in time
   Flexible age model
   Study-specific random effects

3. Computation and Inference

4. Results and Discussion
Global Burden of Disease (GBD)

- The GBD project aims to assess, at the regional and global level, levels of mortality and disability from a wide variety of diseases, injuries, and risk factors.

- Part of the project focuses on estimating levels of diseases and risk factors, while another aspect is to quantify the attribution of mortality and disability to diseases and risk factors.

- Global collaboration, including WHO, the World Bank, and the Institute for Health Metrics and Evaluation (U. of Washington)
# Global Burden of Disease (GBD)

## Background
- Overview
- Data

## Model
- Covariates
- Country-region hierarchy
- Time model
- Age model
- Random effects

## Inference

## Results

### Table 2: Leading causes of death by income group, 2004

<table>
<thead>
<tr>
<th>Disease or Injury</th>
<th>Deaths (millions)</th>
<th>Per cent of total deaths</th>
<th>Disease or Injury</th>
<th>Deaths (millions)</th>
<th>Per cent of total deaths</th>
</tr>
</thead>
<tbody>
<tr>
<td>World</td>
<td></td>
<td></td>
<td>Low-income countries*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 Ischaemic heart disease</td>
<td>7.2</td>
<td>12.2</td>
<td>1 Lower respiratory infections</td>
<td>2.9</td>
<td>11.2</td>
</tr>
<tr>
<td>2 Cerebrovascular disease</td>
<td>5.7</td>
<td>9.7</td>
<td>2 Ischaemic heart disease</td>
<td>2.5</td>
<td>9.4</td>
</tr>
<tr>
<td>3 Lower respiratory infections</td>
<td>4.2</td>
<td>7.1</td>
<td>3 Diarrhoeal diseases</td>
<td>1.8</td>
<td>6.9</td>
</tr>
<tr>
<td>4 COPD</td>
<td>3.0</td>
<td>5.1</td>
<td>4 HIV/AIDS</td>
<td>1.5</td>
<td>5.7</td>
</tr>
<tr>
<td>5 Diarrhoeal diseases</td>
<td>2.2</td>
<td>3.7</td>
<td>5 Cerebrovascular disease</td>
<td>1.5</td>
<td>5.6</td>
</tr>
<tr>
<td>6 HIV/AIDS</td>
<td>2.0</td>
<td>3.5</td>
<td>6 COPD</td>
<td>0.9</td>
<td>3.6</td>
</tr>
<tr>
<td>7 Tuberculosis</td>
<td>1.5</td>
<td>2.5</td>
<td>7 Tuberculosis</td>
<td>0.9</td>
<td>3.5</td>
</tr>
<tr>
<td>8 Trachea, bronchus, lung cancers</td>
<td>1.3</td>
<td>2.3</td>
<td>8 Neonatal infections^</td>
<td>0.9</td>
<td>3.4</td>
</tr>
<tr>
<td>9 Road traffic accidents</td>
<td>1.3</td>
<td>2.2</td>
<td>9 Malaria</td>
<td>0.9</td>
<td>3.3</td>
</tr>
<tr>
<td>10 Prematurity and low birth weight</td>
<td>1.2</td>
<td>2.0</td>
<td>10 Prematurity and low birth weight</td>
<td>0.8</td>
<td>3.2</td>
</tr>
<tr>
<td><strong>Middle-income countries</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 Cerebrovascular disease</td>
<td>3.5</td>
<td>14.2</td>
<td>1 Ischaemic heart disease</td>
<td>1.3</td>
<td>16.3</td>
</tr>
<tr>
<td>2 Ischaemic heart disease</td>
<td>3.4</td>
<td>13.9</td>
<td>2 Cerebrovascular disease</td>
<td>0.8</td>
<td>9.3</td>
</tr>
<tr>
<td>3 COPD</td>
<td>1.8</td>
<td>7.4</td>
<td>3 Trachea, bronchus, lung cancers</td>
<td>0.5</td>
<td>5.9</td>
</tr>
<tr>
<td>4 Lower respiratory infections</td>
<td>0.9</td>
<td>3.8</td>
<td>4 Lower respiratory infections</td>
<td>0.3</td>
<td>3.8</td>
</tr>
<tr>
<td>5 Trachea, bronchus, lung cancers</td>
<td>0.7</td>
<td>2.9</td>
<td>5 COPD</td>
<td>0.3</td>
<td>3.5</td>
</tr>
<tr>
<td>6 Road traffic accidents</td>
<td>0.7</td>
<td>2.8</td>
<td>6 Alzheimer and other dementias</td>
<td>0.3</td>
<td>3.4</td>
</tr>
<tr>
<td>7 Hypertensive heart disease</td>
<td>0.6</td>
<td>2.5</td>
<td>7 Colon and rectum cancers</td>
<td>0.3</td>
<td>3.3</td>
</tr>
<tr>
<td>8 Stomach cancer</td>
<td>0.5</td>
<td>2.2</td>
<td>8 Diabetes mellitus</td>
<td>0.2</td>
<td>2.8</td>
</tr>
<tr>
<td>9 Tuberculosis</td>
<td>0.5</td>
<td>2.2</td>
<td>9 Breast cancer</td>
<td>0.2</td>
<td>2.0</td>
</tr>
<tr>
<td>10 Diabetes mellitus</td>
<td>0.5</td>
<td>2.1</td>
<td>10 Stomach cancer</td>
<td>0.1</td>
<td>1.8</td>
</tr>
</tbody>
</table>

| High-income countries                |                  |                          |                                       |                  |                          |

COPD, chronic obstructive pulmonary disease.

^ Countries grouped by gross national income per capita – low income ($825 or less), high income ($10,066 or more). Note that these high-income groups differ slightly from those used in the Disease Control Priorities Project (see Annex C, Table C2).
Goals of our work

- Estimate cardiometabolic risk factor means for each country × year × adult age group × sex.
  - Systolic blood pressure
  - Total cholesterol
  - Body mass index
  - Fasting plasma glucose

- Estimate age-standardized sub-regional, regional, and global risk factor trends over time by sex.

- Quantify and emphasize the uncertainty of the estimates.
Data collection

- Our colleagues did a systematic literature search for health surveys and epidemiological studies.

- Outcome comparability:
  - In some cases prevalences were reported rather than mean.
  - Regressions were developed to estimate missing study means with (bootstrapped) uncertainty.
  - Uncertainty was reported in various ways (SD, SE, CI) and in some cases was missing.
Data

- Systolic Blood Pressure (SBP):
  - 3195 country × year × age group observations for males, from 746 study × country × years.
  - 3167 observations for females, from 722 studies.

- Total Cholesterol (TC):
  - 1527 observations for males, from 356 studies.
  - 1492 observations for females, from 337 studies.

- Body Mass Index (BMI):
  - 3211 observations for males, from 697 studies.
  - 3589 observations for females, from 815 studies.

- Fasting Plasma Glucose (FPG):
  - 1751 observations for males, from 345 studies.
  - 1752 observations for females, from 344 studies.

A 'full' dataset would have $\sim 200 \times 29 \times 6 \approx 36000$ data points from nationally-representative surveys.
Challenges

- Data are sparse geographically and in time.
- Changes in time and in age may be nonlinear.
- There may be high order interactions.
- Some study means are representative only of a particular community or province, not of the entire country.
- Some studies include only urban or only rural populations.
- Sampling variability (standard errors) differs across studies.
Modeling strategy

- Borrow strength between countries and regions based on predetermined country clusters.
  - Estimate degree of pooling via hierarchical modeling.
- Include country-level covariates to improve prediction.
- Model changes with time and age in a nonlinear, but smooth fashion.
  - Estimate smoothing parameters for data-informed borrowing of strength.
- Include subnational and community data but adjust/discount using offset/variance terms.
- Include rural-only and urban-only studies, but account for differences between country- and study-level urbanization.
- Model males and females separately.
The likelihood

\[ \bar{y}_{h,i} \sim \mathcal{N} \left( X_i \beta + a^c_{j[i]} + b^c_{j[i]} t_i + w_{j[i],t_i} + \gamma_i(z_h) + e_i, SD_{h,i}^2 / n_{h,i} \right) \]
Covariate effects

\[ y_{h,i} \sim N \left( X_i \beta + a_{j[i]}^c + b_{j[i]}^c t_i + w_{j[i],t_i} + \gamma_i(z_h) + e_i, \frac{SD_{h,i}^2}{n_{h,i}} \right) \]

- Covariate effects
- Country-region hierarchy
- Flexible age model
- Nonlinear change in time
- Study-specific random effects
- Sampling variance
Covariate effects

Country-level covariates (moving average of previous five years):

- national income (log per capita GDP)
- country-level urbanization ($u_c$) (%)
- national availability of multiple food types, summarized via PCA

Study-level covariates for study bias adjustment:

- a three-category study-level urbanization variable ($u_s$):
  - urban,
  - rural,
  - mixed (baseline),
- a four-category variable indicating whether the study was:
  - nationally-representative, weighted (baseline),
  - nationally-representative, unweighted,
  - sub-national,
  - community.
Covariate effects: urbanization

In addition to a time-varying main effect of country-level urbanization \( (u_c) \), we add the following offset for studies whose urbanization level \( (u_s) \) differs from that of their country \( \times \) year:

\[
\beta_1 u_c I\{u_s = \text{rural}\} + \beta_2 \{1 - u_c\} I\{u_s = \text{urban}\}
\]

<table>
<thead>
<tr>
<th>( u_s )</th>
<th>( u_c \approx 0 )</th>
<th>( u_c \approx 1/2 )</th>
<th>( u_c \approx 1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>rural</td>
<td>0</td>
<td>( \beta_1/2 )</td>
<td>( \beta_1 )</td>
</tr>
<tr>
<td>mixed</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>urban</td>
<td>( \beta_2 )</td>
<td>( \beta_2/2 )</td>
<td>0</td>
</tr>
</tbody>
</table>
Country-region hierarchy

\[ \bar{y}_{h,i} \sim \mathcal{N} \left( \begin{array}{c} X_i \beta + a_{j[i]}^c + b_{j[i]}^c t_i + w_{j[i],t_i} + \gamma_i(z_h) + e_i \\ \text{covariate effects} \\ \text{nonlinear change in time} \\ \text{study-specific random effects} \\ \text{country-region hierarchy} \\ \text{flexible age model} \\ \text{sampling variance} \end{array} \right), SD_{h,i}^2/n_{h,i} \]
Country-region hierarchy

We implement an exchangeable hierarchical model for:

- the country intercepts and slopes around their sub-regional counterparts,
- the sub-region intercepts and slopes around their regional counterparts,
- the region intercepts and slopes around their global counterparts:

\[
\begin{align*}
    a_j^c & \sim \mathcal{N} \left( a_k^s[j], \kappa_{a}^c \right), & b_j^c & \sim \mathcal{N} \left( b_k^s[j], \kappa_{b}^c \right), \\
    a_k^s & \sim \mathcal{N} \left( a_l^r[k], \kappa_{a}^s \right), & b_k^s & \sim \mathcal{N} \left( b_l^r[k], \kappa_{b}^s \right), \\
    a_l^r & \sim \mathcal{N} \left( a_g, \kappa_{a}^r \right), & b_l^r & \sim \mathcal{N} \left( b_g, \kappa_{b}^r \right).
\end{align*}
\]
Hierarchy → shrinkage

- This hierarchical structure compromises between overly noisy within-unit and overly simplified cross-unit estimates.
- More shrinkage in units where the data are sparse or noisy and less in data-rich units.
Nonlinear change in time

\[ y_{h,i} \sim \mathcal{N} \left( X_i \beta + a_{j[i]}^c + b_{j[i]}^c t_i + w_{j[i],t_i}^c + \gamma_i(z_h) + e_i \right), \text{SD}^2_{h,i}/n_{h,i} \]
Nonlinear change in time

In country $j$, we capture nonlinearity using the $T$-vector $w_j$.

$$w_j = w_j^c + w_k^s[j] + w_l^r[k] + w^g.$$  

Each component of $w_j$ is assigned a Gaussian autoregressive prior (Breslow & Clayton 1993):

- $w_j^c \sim \mathcal{N}(0, (\lambda_c P)^-)$ for $j = 1, \ldots, J$
- $w_k^s \sim \mathcal{N}(0, (\lambda_s P)^-)$ for $k = 1, \ldots, K$
- $w_l^r \sim \mathcal{N}(0, (\lambda_r P)^-)$ for $l = 1, \ldots, L$
- $w^g \sim \mathcal{N}(0, (\lambda_g P)^-)$.

- In the prior:
  $$E(w_t | w_{s,s\neq t}) = \frac{1}{6} \left( 4w_{t-1} + 4w_{t+1} - w_{t-2} - w_{t+2} \right).$$
- The model-estimated precision parameters $\lambda_c, \lambda_s, \lambda_r$, and $\lambda_g$ determine the degree of smoothing at each level.
- In order to achieve identifiability of the $a^c$'s, $b^c$'s, and $w$'s, we constrain the mean and slope of $w^g$ and of each $w^c$, $w^s$, and $w^r$ to be zero.
The fitted time effects compromise between the data and the smoothing specified in the autoregressive prior:

**U.S. males, 55–64 y.**

**Japanese males, 20–29 y.**
Flexible age model

\[
\bar{y}_{h,i} \sim N \left( X_i \beta + a_{j[i]}^c + b_{j[i]}^c t_i + w_{j[i],t_i} + \gamma_i(z_h) + e_i, SD_{h,i}^2 / n_{h,i} \right)
\]
Flexible age model

We use a cubic spline model with knots at ages 45 and 60:

\[ \gamma_i(z_h) = \gamma_1 i z_h + \gamma_2 i z_h^2 + \gamma_3 i z_h^3 + \gamma_4 i (z_h - 45)^3 + \gamma_5 i (z_h - 60)^3. \]

\[ \gamma_1 i = \psi_1 + \phi_1 \mu_i + c_{1i} \]
\[ \gamma_2 i = \psi_2 + \phi_2 \mu_i + c_{2i} \]
\[ \gamma_3 i = \psi_3 + \phi_3 \mu_i + c_{3i} \]
\[ \gamma_4 i = \psi_4 + \phi_4 \mu_i + c_{4i} \]
\[ \gamma_5 i = \psi_5 + \phi_5 \mu_i + c_{5i}. \]

- The \( \phi \)'s allow each component of the age trend to depend on \( \mu_i = a^c_{ji} t_i + b^c_{ji} t_i + X_i \beta + w_{ji}, t_i + e_i, \) the predicted mean outcome value for that study at a baseline age.
- The \( c \)'s produce country-specific random age curves.
Age model fits

Japanese males, 1987

Taiwanese males, 2006

Singapore males, 1998

Italian males, 2000
• The distribution of estimated country-specific age trends.
• The estimated global mean age trend.
Study-specific random effects

\[
\bar{y}_{h,i} \sim N\left( X_i \beta + a^c_{j[i]} + b^c_{j[i]} t_i + w_{j[i],t_i} + \gamma_i(z_h) + e_i, SD^2_{h,i}/n_{h,i} \right)
\]

- **Covariate effects**
- **Country-region hierarchy**
- **Flexible age model**
- **Nonlinear change in time**
- **Study-specific random effects**
- **Sampling variance**
Study-specific random effects

- $e_i = e_i^s + e_{h,i}^{s \times a}$
- Each $e_i^s$ is assigned a normal prior with variance depending on the coverage of study $i$:
  \[
  \text{Var}(e_i^s) = \begin{cases} 
  \nu_w & \text{if study } i \text{ is weighted national} \\
  \nu_u & \text{if study } i \text{ is unweighted national} \\
  \nu_s & \text{if study } i \text{ is sub-national} \\
  \nu_c & \text{if study } i \text{ is community},
  \end{cases}
  \]
  \[
  \nu_w < \nu_u < \nu_s < \nu_c.
  \]
- Structure is analogous for $e_{h,i}^{s \times a}$, the study-age-specific random effects.

Chinese males, 1991

Chinese males, 55–64 y.
MCMC

- Fairly vanilla Metropolis-Hastings, with some exact conditional sampling
- Cross-level dependence of random effects and their hyperparameters slows mixing, e.g., $\{w^c, \lambda^c\}$
- Solution: jointly sample random effects + associated hyperparameter(s)
Computation

- Fast linear algebra implementation (GotoBLAS linked to R)
- Sparse matrix manipulations (spam package in R)
  - Recall the precision matrix for $w^\cdot$, which is sparse
- Combination of multiple MCMC chains from a Linux cluster (i.e., embarrassingly parallel)
Inferential Output

- Core products:
  - Country × year × age × sex mean levels
  - Age-standardized country, sub-region, region, and global mean levels by year and sex
  - Linear trends in age-standardized country mean levels by sex

- Comments:
  - All inferential products are calculated for each MCMC sample and then summarized across samples to propagate uncertainty properly.
  - Aggregate across countries/regions, etc., then age-standardize at the level of interest.
Model assessment

- Posterior predictive checks suggest missing age × time × country interaction.
- Data plotted against predictions [see Web6 pdf]
- Cross-validation (10-fold)
  - Assess performance at various points in the covariate/cluster space
  - Assess predictions for countries with data, without data, and for extrapolation over time
  - Unit of consideration was the study, not study-age observation
## Cross-validation (cholesterol)

<table>
<thead>
<tr>
<th>Region</th>
<th>Female model</th>
<th>Male model</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of held-out observations</td>
<td>Percent covered</td>
<td>No. of held-out observations</td>
</tr>
<tr>
<td>Western high-income regions</td>
<td>223</td>
<td>0.94</td>
</tr>
<tr>
<td>Central/East Europe and Central Asia</td>
<td>67</td>
<td>0.93</td>
</tr>
<tr>
<td>Sub-Saharan Africa</td>
<td>72</td>
<td>1.00</td>
</tr>
<tr>
<td>North Africa and Middle East</td>
<td>76</td>
<td>0.97</td>
</tr>
<tr>
<td>South Asia</td>
<td>30</td>
<td>0.97</td>
</tr>
<tr>
<td>East and Southeast Asia and Pacific</td>
<td>115</td>
<td>0.97</td>
</tr>
<tr>
<td>Latin America and Caribbean</td>
<td>66</td>
<td>0.97</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Scope</th>
<th>Female model</th>
<th>Male model</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of held-out observations</td>
<td>Percent covered</td>
<td>No. of held-out observations</td>
</tr>
<tr>
<td>Rural</td>
<td>54</td>
<td>0.93</td>
</tr>
<tr>
<td>Urban</td>
<td>167</td>
<td>0.98</td>
</tr>
<tr>
<td>Mixed</td>
<td>428</td>
<td>0.96</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Coverage</th>
<th>Female model</th>
<th>Male model</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of held-out observations</td>
<td>Percent covered</td>
<td>No. of held-out observations</td>
</tr>
<tr>
<td>Community</td>
<td>260</td>
<td>0.97</td>
</tr>
<tr>
<td>Sub-national</td>
<td>111</td>
<td>0.91</td>
</tr>
<tr>
<td>Unweighted national</td>
<td>109</td>
<td>0.98</td>
</tr>
<tr>
<td>Weighted national</td>
<td>169</td>
<td>0.96</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Age quartile</th>
<th>Female model</th>
<th>Male model</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of held-out observations</td>
<td>Percent covered</td>
<td>No. of held-out observations</td>
</tr>
<tr>
<td>(20,40]</td>
<td>226</td>
<td>0.96</td>
</tr>
<tr>
<td>(40,50]</td>
<td>132</td>
<td>0.96</td>
</tr>
<tr>
<td>(50,60]</td>
<td>130</td>
<td>0.97</td>
</tr>
<tr>
<td>(60,100]</td>
<td>161</td>
<td>0.94</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Hold-out algorithm</th>
<th>Female model</th>
<th>Male model</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of held-out observations</td>
<td>Percent covered</td>
<td>No. of held-out observations</td>
</tr>
<tr>
<td>All of the country’s studies</td>
<td>272</td>
<td>0.97</td>
</tr>
<tr>
<td>All of the country’s 2000-2009 studies</td>
<td>143</td>
<td>0.95</td>
</tr>
<tr>
<td>A random 1/3 of the country’s studies</td>
<td>234</td>
<td>0.95</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Year quartile</th>
<th>Female model</th>
<th>Male model</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of held-out observations</td>
<td>Percent covered</td>
<td>No. of held-out observations</td>
</tr>
<tr>
<td>[1980,1995]</td>
<td>158</td>
<td>0.92</td>
</tr>
<tr>
<td>(1995,2002]</td>
<td>173</td>
<td>0.98</td>
</tr>
</tbody>
</table>
Results

• SBP: clear decreases in developed countries; uncertainty elsewhere but some indications of no change or increases
• TC: clear decreases in developed countries and former Soviet bloc; little change apparent elsewhere but high uncertainty
• BMI: clear increases everywhere, with possible male/female differences by sub-region

[see pdfs]
Shortcomings

• Various interactions are not included, in particular:
  - age $\times$ time $\times$ country effects
  - study-level biases likely vary with other factors
• Data points are associated with age group midpoints.
• Aggregation loses information:
  - We estimate only population means and not full distributions or exceedances
  - Prevalence data is 'converted' to means via pre-processing (with similar manipulations for missing uncertainty information)
Current work

- Goal: estimate full distributions of various malnutrition indicators: hemoglobin, chronic and acute malnutrition in children, vitamin A
- Data: individual-level data, sample means, and sample prevalences
- Approach: extend this work to mixture models, either finite mixtures or Dirichlet process style models