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

Mariel Finucane[1], Chris Paciorek*[,1,2], Goodarz Danaei[3], and Majid Ezzati[4]

*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

Outline

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)

World Health Organization

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

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

COPD, chronic obstructive pulmonary disease.

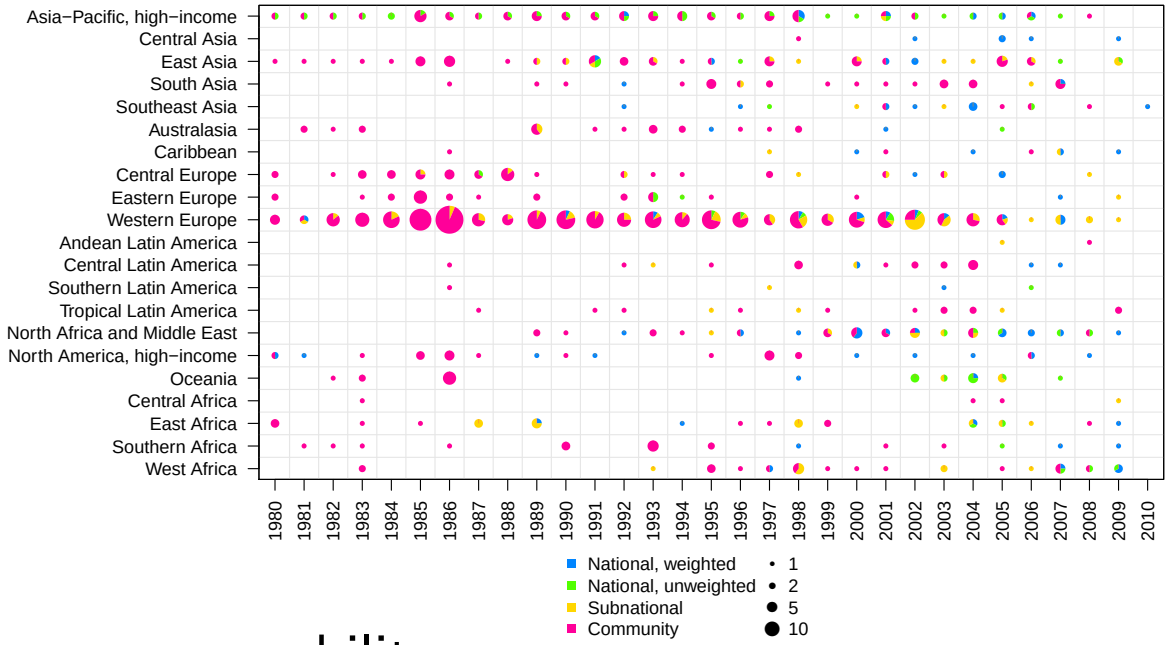
[a] Countries grouped by gross national income per capita – low income ($825 or less), high income ($10 066 or more). Note that

Goals of our work

- Estimate cardiometabolic risk factor means for each country $\times$ year $\times$ adult age group $\times$ 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.

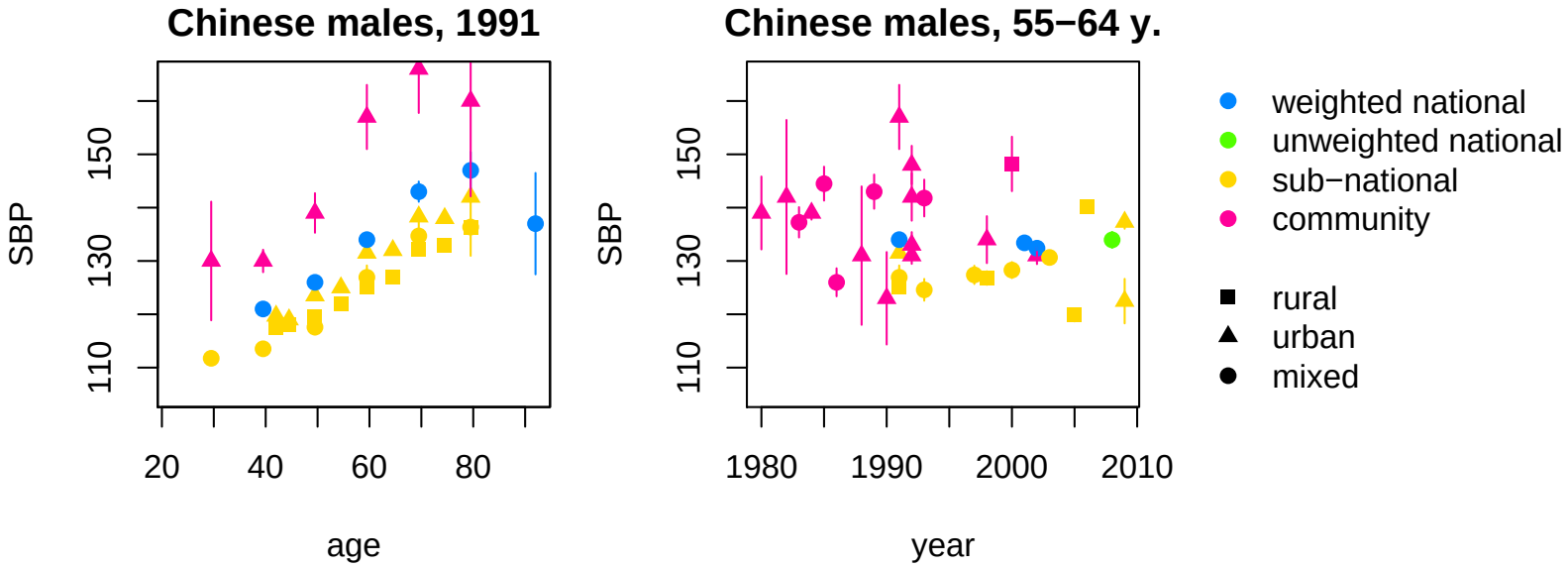
Data

- Systolic Blood Pressure (SBP):
  - 3195 country $\times$ year $\times$ age group observations for males, from 746 study $\times$ country $\times$ 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

$$\bar{y}_{h,i} \sim \mathcal{N}\left(\underbrace{X_i\beta}_{\text{covariate effects}} + \overbrace{a^c_{j[i]} + b^c_{j[i]} t_i}^{\text{country-region hierarchy}} + \underbrace{w_{j[i],t_i}}_{\text{nonlinear change in time}} + \overbrace{\gamma_i(z_h)}^{\text{flexible age model}} + \underbrace{e_i}_{\text{study-specific random effects}}, \overbrace{\mathrm{SD}^2_{h,i}/n_{h,i}}^{\text{sampling variance}} \right)$$

Covariate effects

$$\bar{y}_{h,i} \sim \mathcal{N}\left(\underbrace{X_i \beta}_{\text{covariate effects}} + \overbrace{a^c_{j[i]} + b^c_{j[i]} t_i}^{\text{country-region hierarchy}} + \underbrace{w_{j[i],t_i}}_{\text{nonlinear change in time}} + \overbrace{\gamma_i(z_h)}^{\text{flexible age model}} + \underbrace{e_i}_{\text{study-specific random effects}}, \overbrace{\text{SD}^2_{h,i}/n_{h,i}}^{\text{sampling variance}} \right)$$

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:

$$\beta_1 u_c \, \mathrm{I}\{u_s = \text{rural}\} + \beta_2\{1 - u_c\} \, \mathrm{I}\{u_s = \text{urban}\}$$

	$u_c \approx 0$	$u_c \approx 1/2$	$u_c \approx 1$
$u_s =$ rural	0	$\beta_1/2$	β_1
$u_s =$ mixed	0	0	0
$u_s =$ urban	β_2	$\beta_2/2$	0

Country-region hierarchy

$$\bar{y}_{h,i} \sim \mathcal{N}\left(\underbrace{X_i\beta}_{\text{covariate effects}} + \overbrace{a^c_{j[i]} + b^c_{j[i]} t_i}^{\text{country-region hierarchy}} + \underbrace{w_{j[i],t_i}}_{\text{nonlinear change in time}} + \overbrace{\gamma_i(z_h)}^{\text{flexible age model}} + \underbrace{e_i}_{\text{study-specific random effects}}, \overbrace{\text{SD}^2_{h,i}/n_{h,i}}^{\text{sampling variance}} \right)$$

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:

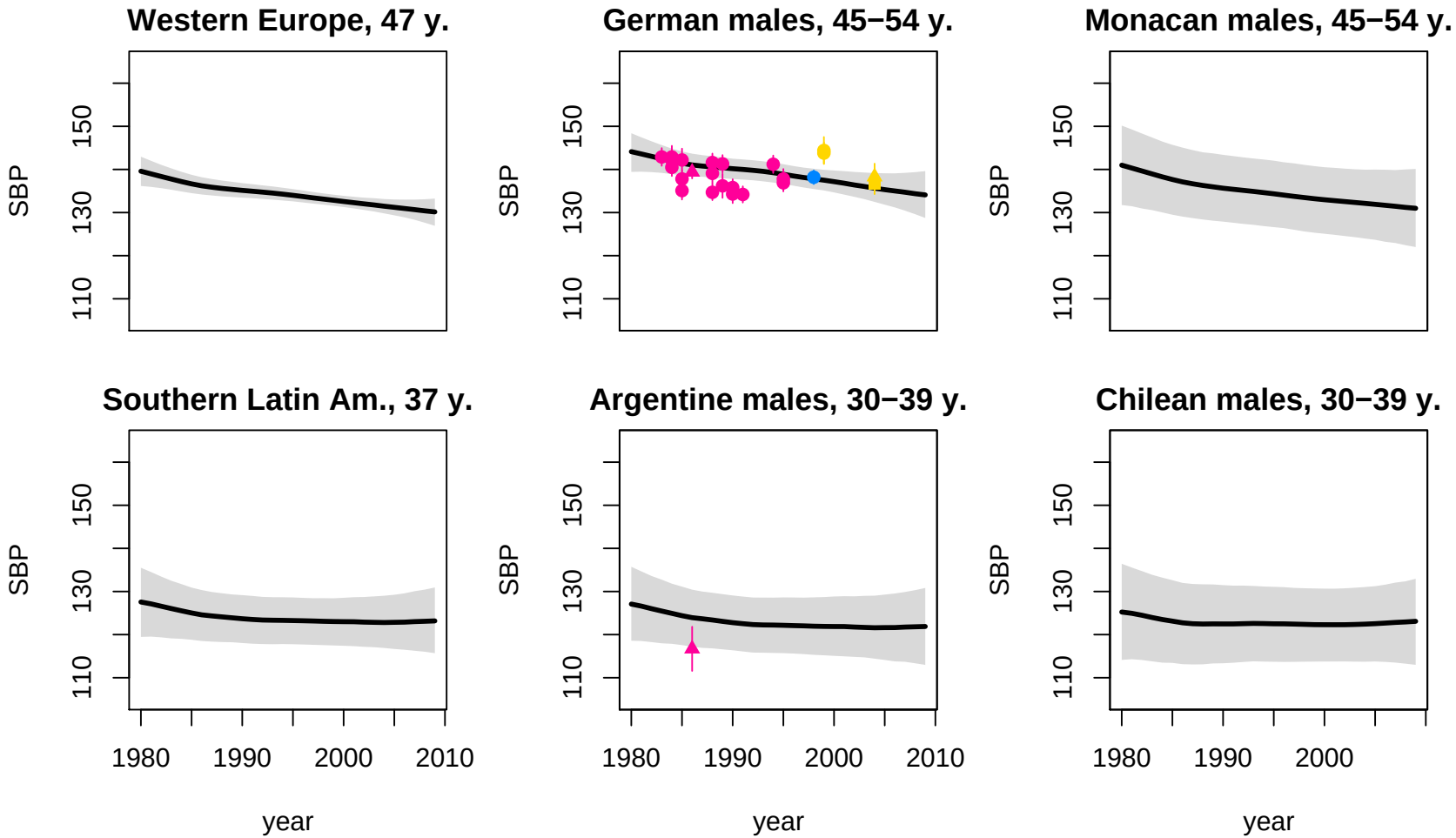
$$a_j^c \sim \mathcal{N}\left(a_{k[j]}^s, \kappa_a^c\right), \quad b_j^c \sim \mathcal{N}\left(b_{k[j]}^s, \kappa_b^c\right),$$

$$a_k^s \sim \mathcal{N}\left(a_{l[k]}^r, \kappa_a^s\right), \quad b_k^s \sim \mathcal{N}\left(b_{l[k]}^r, \kappa_b^s\right),$$

$$a_l^r \sim \mathcal{N}\left(a^g, \kappa_a^r\right), \qquad b_l^r \sim \mathcal{N}\left(b^g, \kappa_b^r\right).$$

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

$$\bar{y}_{h,i} \sim \mathcal{N}\left(\underbrace{X_i\beta}_{\text{covariate effects}} + \overbrace{a^c_{j[i]} + b^c_{j[i]} t_i}^{\text{country-region hierarchy}} + \underbrace{w_{j[i],t_i}}_{\text{nonlinear change in time}} + \overbrace{\gamma_i(z_h)}^{\text{flexible age model}} + \underbrace{e_i}_{\text{study-specific random effects}}, \overbrace{\text{SD}^2_{h,i}/n_{h,i}}^{\text{sampling variance}} \right)$$

Nonlinear change in time

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

$$w_j = w_j^c + w_{k[j]}^s + w_{l[k]}^r + w^g.$$

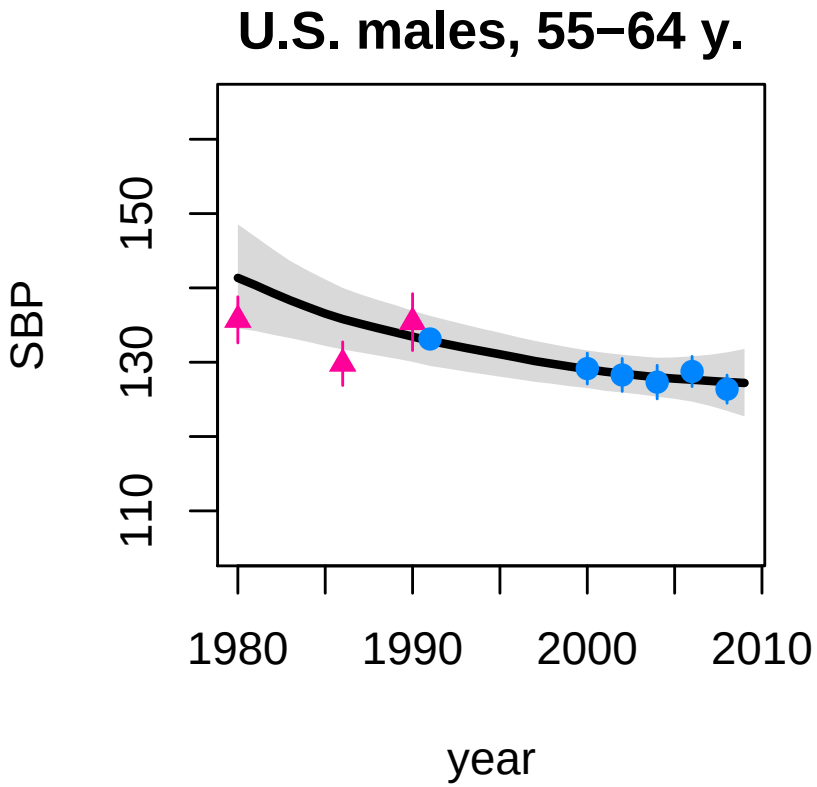
Each component of w_j is assigned a Gaussian autoregressive prior (Breslow & Clayton 1993):

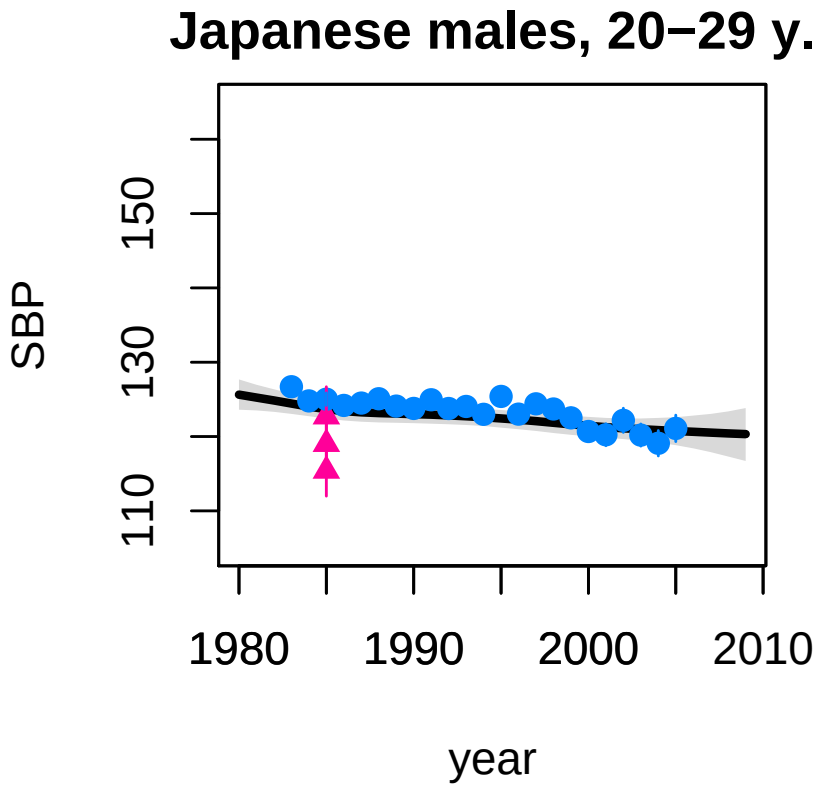
$$\begin{aligned} w_j^c &\sim \mathcal{N}\left(0, (\lambda_c P)^-\right) \quad \text{for } j = 1, \cdots, J \\ w_k^s &\sim \mathcal{N}\left(0, (\lambda_s P)^-\right) \quad \text{for } k = 1, \cdots, K \\ w_l^r &\sim \mathcal{N}\left(0, (\lambda_r P)^-\right) \quad \text{for } l = 1, \cdots, L \\ w^g &\sim \mathcal{N}\left(0, (\lambda_g P)^-\right). \end{aligned}$$

- In the prior:
  $E(w_t^i | w_{s,s\neq t}^i) = \frac{1}{6}\left(4w_{t-1}^i + 4w_{t+1}^i - w_{t-2}^i - w_{t+2}^i\right)$.
- The model-estimated precision parameters λ_c, λ_s,λ_r, and λ_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.

Time model fits

The fitted time effects compromise between the data and the smoothing specified in the autoregressive prior:





Flexible age model

$$\bar{y}_{h,i} \sim \mathcal{N}\left(\underbrace{X_i\beta}_{\text{covariate effects}} + \overbrace{a^c_{j[i]} + b^c_{j[i]} t_i}^{\text{country-region hierarchy}} + \underbrace{w_{j[i],t_i}}_{\text{nonlinear change in time}} + \overbrace{\gamma_i(z_h)}^{\text{flexible age model}} + \underbrace{e_i}_{\text{study-specific random effects}}, \overbrace{\text{SD}^2_{h,i}/n_{h,i}}^{\text{sampling variance}} \right)$$

Flexible age model

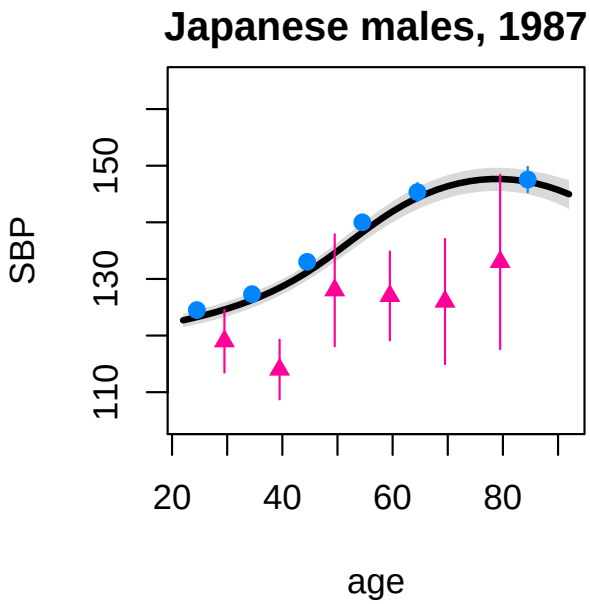
We use a cubic spline model with knots at ages 45 and 60:

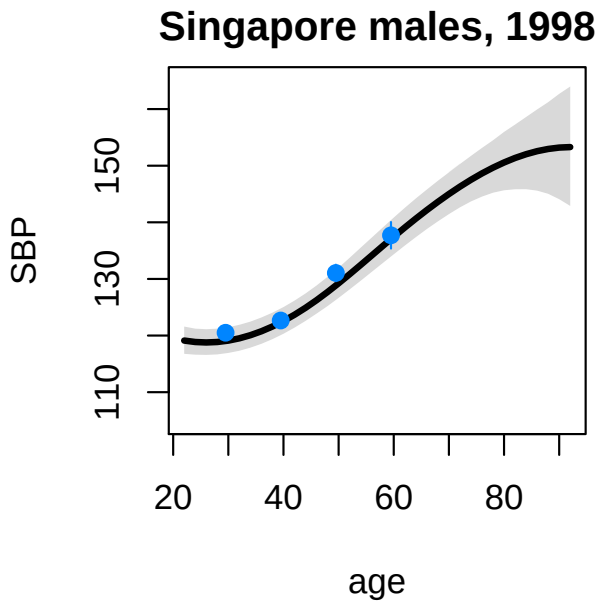
$$\gamma_i(z_h) = \gamma_{1i} z_h + \gamma_{2i} z_h^2 + \gamma_{3i} z_h^3 + \gamma_{4i}(z_h - 45)_+^3 + \gamma_{5i}(z_h - 60)_+^3.$$

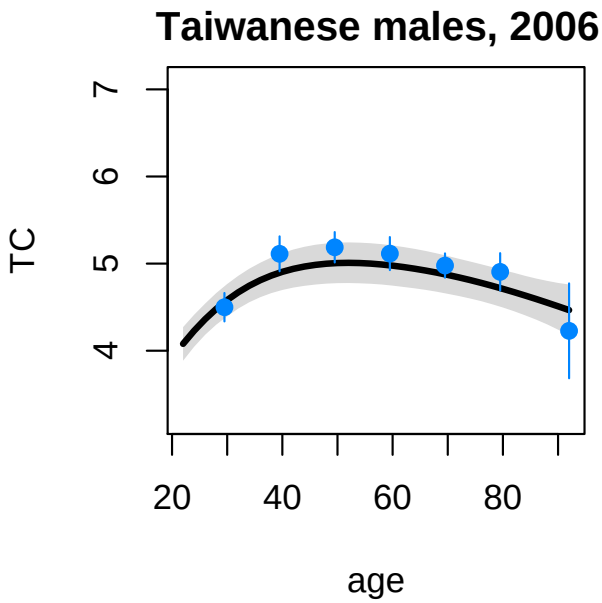
$$\begin{aligned}
\gamma_{1i} &= \psi_1 + \phi_1 \mu_i + c_{1j[i]} \\
\gamma_{2i} &= \psi_2 + \phi_2 \mu_i + c_{2j[i]} \\
\gamma_{3i} &= \psi_3 + \phi_3 \mu_i + c_{3j[i]} \\
\gamma_{4i} &= \psi_4 + \phi_4 \mu_i + c_{4j[i]} \\
\gamma_{5i} &= \psi_5 + \phi_5 \mu_i + c_{5j[i]}.
\end{aligned}$$

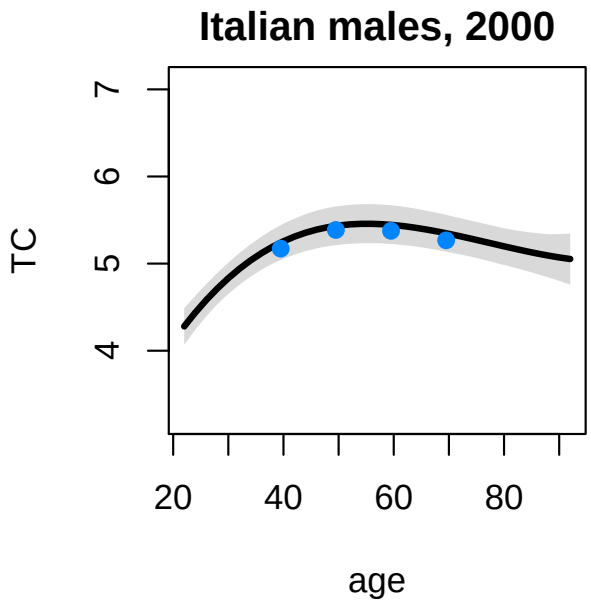
- The ϕ's allow each component of the age trend to depend on $\mu_i = a^c_{j[i]} + b^c_{j[i]} t_i + X_i \beta + w_{j[i],t_i} + e^s_i$, the predicted mean outcome value for that study at a baseline age.
- The c's produce country-specific random age curves.

Age model fits

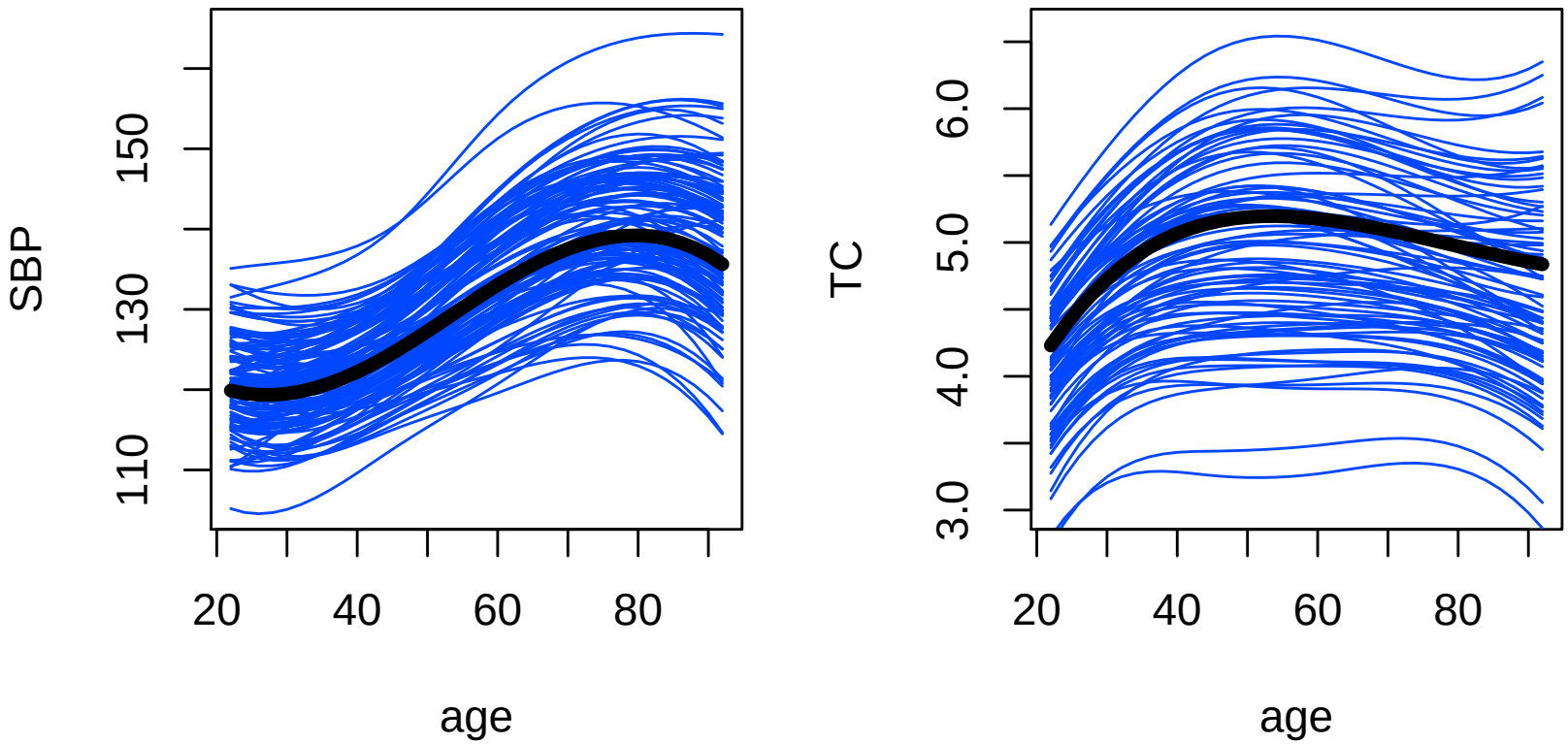








Age model fits



- The distribution of estimated country-specific age trends.
- The estimated global mean age trend.

Study-specific random effects

$$\bar{y}_{h,i} \sim \mathcal{N}\left(\underbrace{X_i\beta}_{\text{covariate effects}} + \overbrace{a^c_{j[i]} + b^c_{j[i]} t_i}^{\text{country-region hierarchy}} + \underbrace{w_{j[i],t_i}}_{\text{nonlinear change in time}} + \overbrace{\gamma_i(z_h)}^{\text{flexible age model}} + \underbrace{e_i}_{\text{study-specific random effects}}, \overbrace{\mathrm{SD}^2_{h,i}/n_{h,i}}^{\text{sampling variance}} \right)$$

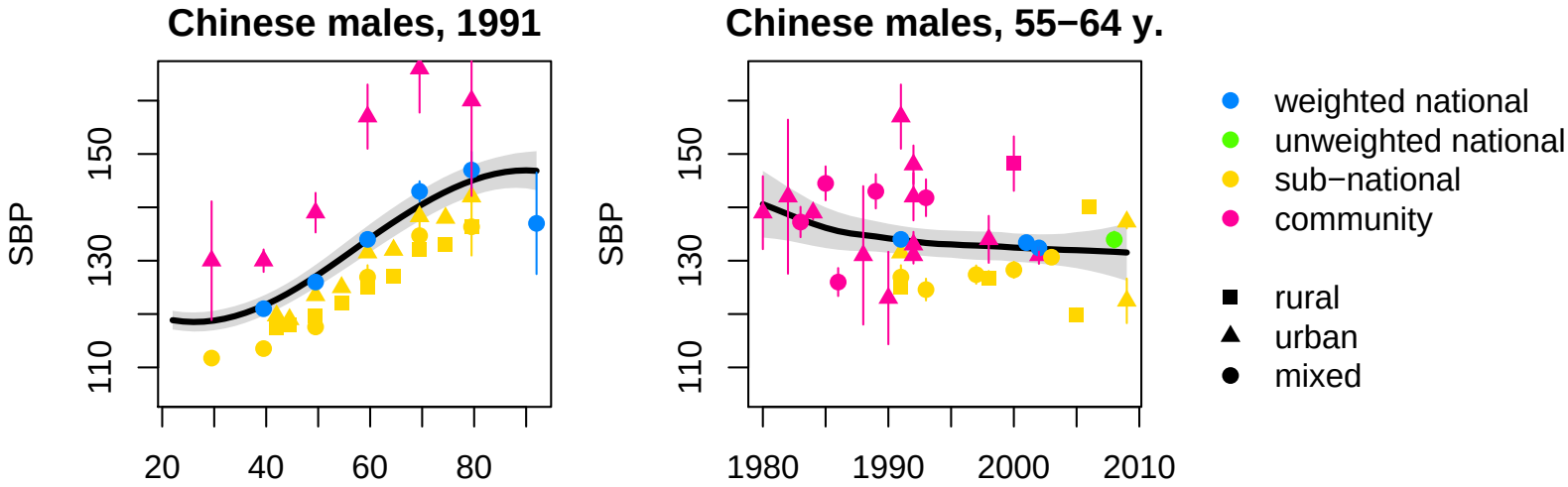
Study-specific random effects

- $e_i = e_i^s + e_{h,i}^{s\times a}$
- Each e_i^s is assigned a normal prior with variance depending on the coverage of study i:

$$\mathrm{Var}(e_i^s) = \begin{cases} \nu_w & \text{if study } i \text{ is weighted national} \\ \nu_u & \text{if study } i \text{ is unweighted national} \\ \nu_s & \text{if study } i \text{ is sub-national} \\ \nu_c & \text{if study } i \text{ is community,} \end{cases}$$

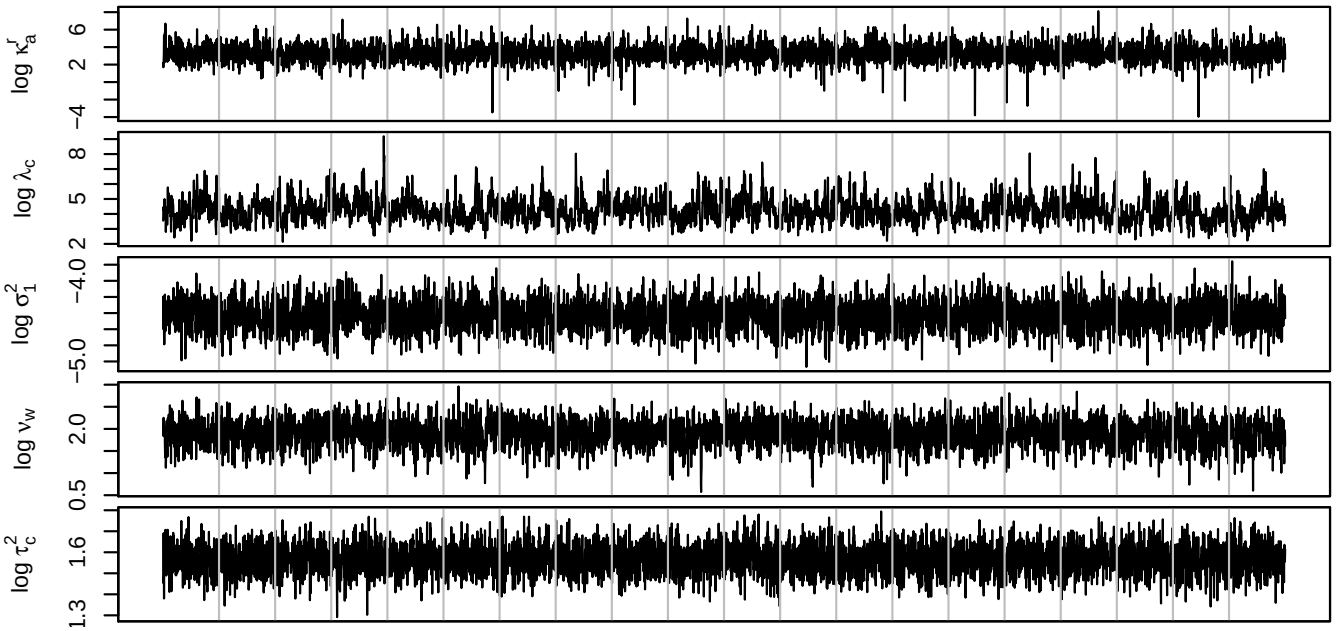
$$\nu_w < \nu_u < \nu_s < \nu_c.$$

- Structure is analogous for $e_{h,i}^{s\times a}$, the study-age-specific random effects.



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 $\times$ year $\times$ age $\times$ 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 $\times$ time $\times$ 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)

	Female model		Male model	
	No. of held-out observations	Percent covered	No. of held-out observations	Percent covered
Region				
Western high-income regions	223	0.94	265	0.92
Central/East Europe and Central Asia	67	0.93	53	0.96
Sub-Saharan Africa	72	1.00	67	1.00
North Africa and Middle East	76	0.97	68	0.97
South Asia	30	0.97	49	0.92
East and Southeast Asia and Pacific	115	0.97	104	0.92
Latin America and Caribbean	66	0.97	73	0.99
Scope				
Rural	54	0.93	63	0.98
Urban	167	0.98	173	0.95
Mixed	428	0.96	443	0.94
Coverage				
Community	260	0.97	290	0.97
Sub-national	111	0.91	120	0.85
Unweighted national	109	0.98	121	0.98
Weighted national	169	0.96	148	0.94
Age quartile				
(20,40]	226	0.96	235	0.94
(40,50]	132	0.96	138	0.99
(50,60]	130	0.97	136	0.97
(60,100]	161	0.94	170	0.90
Hold-out algorithm				
All of the country's studies	272	0.97	254	0.99
All of the country's 2000-2009 studies	143	0.95	253	0.89
A random 1/3 of the country's studies	234	0.95	172	0.95
Year quartile				
[1980,1995]	158	0.92	159	0.97
(1995,2002]	173	0.98	200	0.94

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