Appendix to *Data-driven methods for imputing national-level incidence rates in global burden of disease studies*

Table of Contents

eTables

eTable A1. *Assignment of missing GEMS cluster diet for 19 countries.*

Country	Geographical neighbour(s)	Neighbour cluster number
Andorra	Spain	8
Bahrain	Saudi Arabia, Kuwait, and United Arab Emirates	4
Oman	Saudi Arabia, Kuwait, and United Arab Emirates	4
Qatar	Saudi Arabia, Kuwait, and United Arab Emirates	4
Bhutan	India	5
Cook Islands	Other Pacific islands	17
Marshall Islands	Other Pacific islands	17
Micronesia	Other Pacific islands	17
Nauru	Other Pacific islands	17
Niue	Other Pacific islands	17
Palau	Other Pacific islands	17
Tonga	Other Pacific islands	17
Tuvalu	Other Pacific islands	17
Equatorial Guinea	Gabon	16
Eritrea	Ethiopia	13
Lesotho	South Africa	5
Monaco	France	7
San Marino	Italy	10
Singapore	Malaysia	5

eTable A2. *Description of the 51 candidate covariates used in imputation Method 3.*

Variable name (source)	**Description**
agriarea (World Development Indicators, WDI)	http://data.worldbank.org/indicator/AG.LND.AGRI.ZS Agricultural land (% of land area)
agrivalue (WDI)	http://data.worldbank.org/indicator/NV.AGR.TOTL.ZS Agriculture, value added (% of GDP)
animalcalpercap (FAOstat)	http://faostat.fao.org/site/610/DesktopDefault.aspx?PageID=610#ancor Food supply from animal products (kcal/person/day)
birthperadolescent (WDI)	http://data.worldbank.org/indicator/SP.ADO.TFRT Adolescent fertility rate is the number of births per 1,000 women ages 15-19
birthrate (WDI)	http://data.worldbank.org/indicator/SP.DYN.CBRT.IN Birth rate, crude (per 1,000 people)
birthsattended (WDI)	http://data.worldbank.org/indicator/SH.STA.BRTC.ZS births attended by skilled health staff (% of total)
cerealyield (WDI)	http://data.worldbank.org/indicator/AG.YLD.CREL.KG Kg yield per hectare of harvested land, includes wheat, rice, maize, barley, oats, rye, millet, sorghum, buckwheat, and mixed grains
co2emission (WDI)	http://data.worldbank.org/indicator/EN.ATM.CO2E.KD.GD CO2 emissions (kg per USof GDP)
deathmaternalnr (WDI)	http://data.worldbank.org/indicator/SH.MMR.DTHS Number of maternal deaths
deathrate (WDI)	http://data.worldbank.org/indicator/SP.DYN.CDRT.IN Death rate, crude (per 1,000 people)
educexpenditure (WDI)	http://data.worldbank.org/indicator/SE.XPD.TOTL.GD.ZS Expenditure for education (%GDP)
energyuse (WDI)	http://data.worldbank.org/indicator/EG.USE.PCAP.KG.OE Energy use (kg of oil equivalent per capita)
expenditureeducpublic (WDI)	http://data.worldbank.org/indicator/SP.DYN.TFRT.IN Public spending on education total (% of government expenditure)
extremeweather (WDI)	http://data.worldbank.org/indicator/EN.CLC.MDAT.ZS Droughts, floods, extreme, temperatures (% of population, average 1990-2009)
fertilityrate (WDI)	http://data.worldbank.org/indicator/SP.DYN.TFRT.IN Fertility rate total (births per woman)
foodexport (WDI)	http://data.worldbank.org/indicator/TX.VAL.FOOD.ZS.UN Food exports (% of merchandise exports)
forestarea	http://data.worldbank.org/indicator/AG.LND.FRST.ZS

Variable name (source)	Description
(WDI)	Percentage of forest area
freshwatersources (WDI)	http://data.worldbank.org/indicator/ER.H2O.INTR.K3 Renewable internal freshwater resources, total (billion cubic meters)
GDP (WDI)	http://data.worldbank.org/indicator/NY.GDP.MKTP.CD Gross Domestic Product (current US$)
GDPpercapita (WDI)	http://data.worldbank.org/indicator/NY.GDP.PCAP.CD Gross Domestic Product per capita (current US$)
GNI (WDI)	http://data.worldbank.org/indicator/NY.GNP.PCAP.CD Gross Netto Income (current US$)
Healthexppercapita (WDI)	http://data.worldbank.org/indicator/SH.XPD.PCAP Health expenditure per capita (US$)
healthexppublic (WDI)	http://data.worldbank.org/indicator/SH.XPD.PUBL Health expenditure, public (% of total health expenditure)
healthresource (WDI)	http://data.worldbank.org/indicator/SH.XPD.EXTR.ZS External resources for health (% of total expenditure on health)
hivprev (WDI)	http://data.worldbank.org/indicator/SH.DYN.AIDS.ZS Prevalence of HIV, total (% of population ages 15-49 years)
improvdsanitation (WDI)	http://data.worldbank.org/indicator/SH.STA.ACSN Percentage of households with improved sanitation
improvdwater (WDI)	http://data.worldbank.org/indicator/SH.H2O.SAFE.RU.ZS Improved water source, rural (% of rural population with access)
kcalperday (FAOstat)	http://faostat.fao.org/site/610/DesktopDefault.aspx?PageID=610#ancor Food supply (kcal/person/day)
laborfemmale (WDI)	http://data.worldbank.org/indicator/SL.TLF.CACT.FE.ZS Labor participation rate, female (% of female population ages 15+ years)
laborparticipatn (WDI)	http://data.worldbank.org/indicator/SL.TLF.CACT.ZS Labor force participation rate is the proportion of the population ages 15 and older that is economically active
lifeexp (WDI)	http://data.worldbank.org/indicator/SP.DYN.LE00.IN Life expectancy at birth, total (years)
literacy (WDI)	http://data.worldbank.org/indicator/SE.ADT.LITR.ZS Literacy rate, adult total (% adult >15 years)
literacyyouth (WDI)	http://data.worldbank.org/indicator/SE.ADT.1524.LT.ZS Literacy rate, youth total (% of people ages 15-24 years)
maternaldeathrisk (WDI)	http://data.worldbank.org/indicator/SH.MMR.RISK.ZS Lifetime risk for maternal death

Variable name (source)	Description
maternalmortalitymodel (WDI)	http://data.worldbank.org/indicator/SH.STA.MMRT maternal mortality ratio (modelled estimate, per 100,000 live births)
measlesimm (WDI)	http://data.worldbank.org/indicator/SH.IMM.MEAS Immunization, measles (% of children ages 12023 months)
mortalityinfant (WDI)	http://data.worldbank.org/indicator/SP.DYN.IMRT.IN Mortality rate, infant (per 1,000 live births)
mortalitymale (WDI)	http://data.worldbank.org/indicator/SP.DYN.AMRT.MA Mortality rate, adult, male (per 1,000 male adults).
mortalityneonatal (WDI)	http://data.worldbank.org/indicator/SH.DYN.MORT Mortality rate, under-5 (per 1,000 live births)
pctarableland (WDI)	http://data.worldbank.org/indicator/AG.LND.ARBL.ZS Percentage arable land
pctprimschoolkids (WDI)	http://data.worldbank.org/indicator/SE.PRM.ENRL children in primary school
physdens (WDI)	http://data.worldbank.org/indicator/SH.MED.PHYS.ZS Physicians (per 1,000 people). Physicians include generalist and specialist medical practitioners.
popdens (WDI)	http://data.worldbank.org/indicator/EN.POP.DNST Population density (people per sq. km of land area)
precipannual (WDI)	http://data.worldbank.org/indicator/AG.LND.PRCP.MM Average precipitation in depth (mm per year)
prevundernourishd (WDI)	http://data.worldbank.org/indicator/SN.ITK.DEFC.ZS % of population that is undernourished
primcomplete (WDI)	http://data.worldbank.org/indicator/SE.PRM.CMPT.ZS Primary school completion rate, total (% of relevant age group)
proteinsupply (FAOstat)	http://faostat.fao.org/site/610/DesktopDefault.aspx?PageID=610#ancor Protein supply quantity (g/person/day)
ricepaddy (FAOstat)	http://faostat.fao.org/site/567/default.aspx#ancor Area harvested for rice and paddy
tuberculosisdetection (WDI)	http://data.worldbank.org/indicator/SH.TBS.INCD Percentage of population subject to tuberculosis infection (per 100,000)
under5mort (WDI)	http://data.worldbank.org/indicator/SH.DYN.MORT Mortality rate, under-5 years (per 1,000 live births)
urban (WDI)	http://data.worldbank.org/indicator/SP.URB.TOTL.IN.ZS Urban population (% of total)

eMethods

The example foodborne disease datasets

The country-level data on the incidence of congenital toxoplasmosis per 1000 births was originally derived from a systematic literature review (Torgerson & Mastroiacovo, 2013). For countries with two or more available studies with similar quality data, the mean incidence value of the most recent and more valid studies was taken. In the absence of confirmed case data, simple mathematical and statistical modelling approaches were used to estimate the risk, and thus the incidence of seroconversion in pregnant women, which was then adjusted for the risk of maternal transmission and related to national data on annual number of births (see Torgerson & Mastroiacovo, 2013 for further details).

For aflatoxin-related hepatocellular carcinoma, the dataset consisted of country-level national incidence values (Liu & Wu, 2010) that were estimated from: maize and nut consumption patterns in various world regions, the estimated average aflatoxin exposure/contamination levels in maize and nuts in various world regions (via literature searches), and the prevalence of hepatitis B virus infection in these countries (from relevant studies). If estimates of aflatoxin exposure were not available, the authors estimated exposure from food consumption patterns and aflatoxin contamination levels (Liu & Wu, 2010).

The Missing-at-Random (MAR) assumption

What are the conditions by which the MAR assumption would be violated, and what would be the consequences? If reporting probability was inversely correlated with the incidence of infection, then MAR would not hold. This association can however never be confirmed as the necessary incidence data are not reported. However, the MAR assumption (unlike the Missing Completely At Random [MCAR] assumption) requires only that missingness is independent of the unobserved data, given the observed data. Therefore, to use the indicator Gross Domestic Product (GDP) as an example: even if disease incidence is negatively associated with GDP (which is observed), once GDP (and possibly other variables) is taken into account, the missing disease incidence data would be considered MAR. In other words, the relationship between incidence and GDP is assumed to be the same for countries for which incidence is both observed and unobserved. Under MAR, after conditioning on GDP, whether or not incidence data are missing does not depend on the values of the unobserved data.

Equations for Method 3: Bayesian random effects log-Normal regression model

The random effects model assumes that the log-transformed incidence θ_{ij} of country j in region i arises from a Normal distribution with region-specific mean μ_i and within-region variance σ_w^2. The regional mean μ_i is in its turn assumed to follow a Normal distribution with mean μ_0, the global intercept, and between-region variance σ_b^2:

$$\log \theta_{ij} \sim Normal(\mu_i, \sigma_w^2)$$

$$\mu_i \sim Normal(\mu_0, \sigma_b^2)$$

Equations for Method 3: Bayesian mixed effects log-Normal regression model

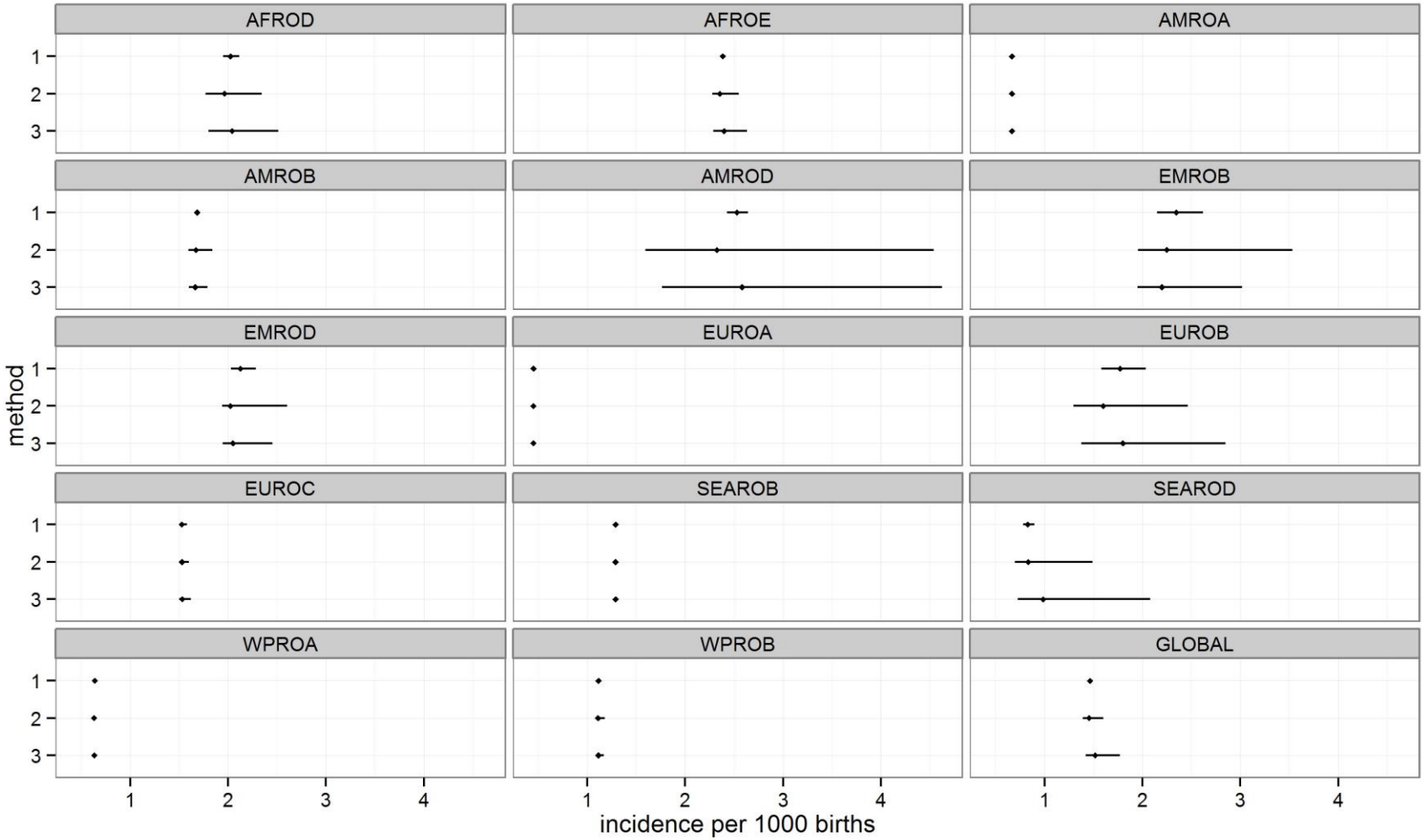
The mixed effects log-Normal regression model assumes that the log-transformed incidence θ_{ij} of country j in region i arises from a Normal distribution with mean $\mu_i + \sum_{k=1}^{p} x_{ijk}\beta_k$ and within-region variance σ_w^2. As before, the regional means μ_i are assumed to follow a Normal distribution with mean μ_0, the global intercept, and between-region variance σ_b^2:

$$\log \theta_{ij} \sim Normal(\mu_i + \sum_{k=1}^{p} x_{ijk}\beta_k, \sigma_w^2)$$

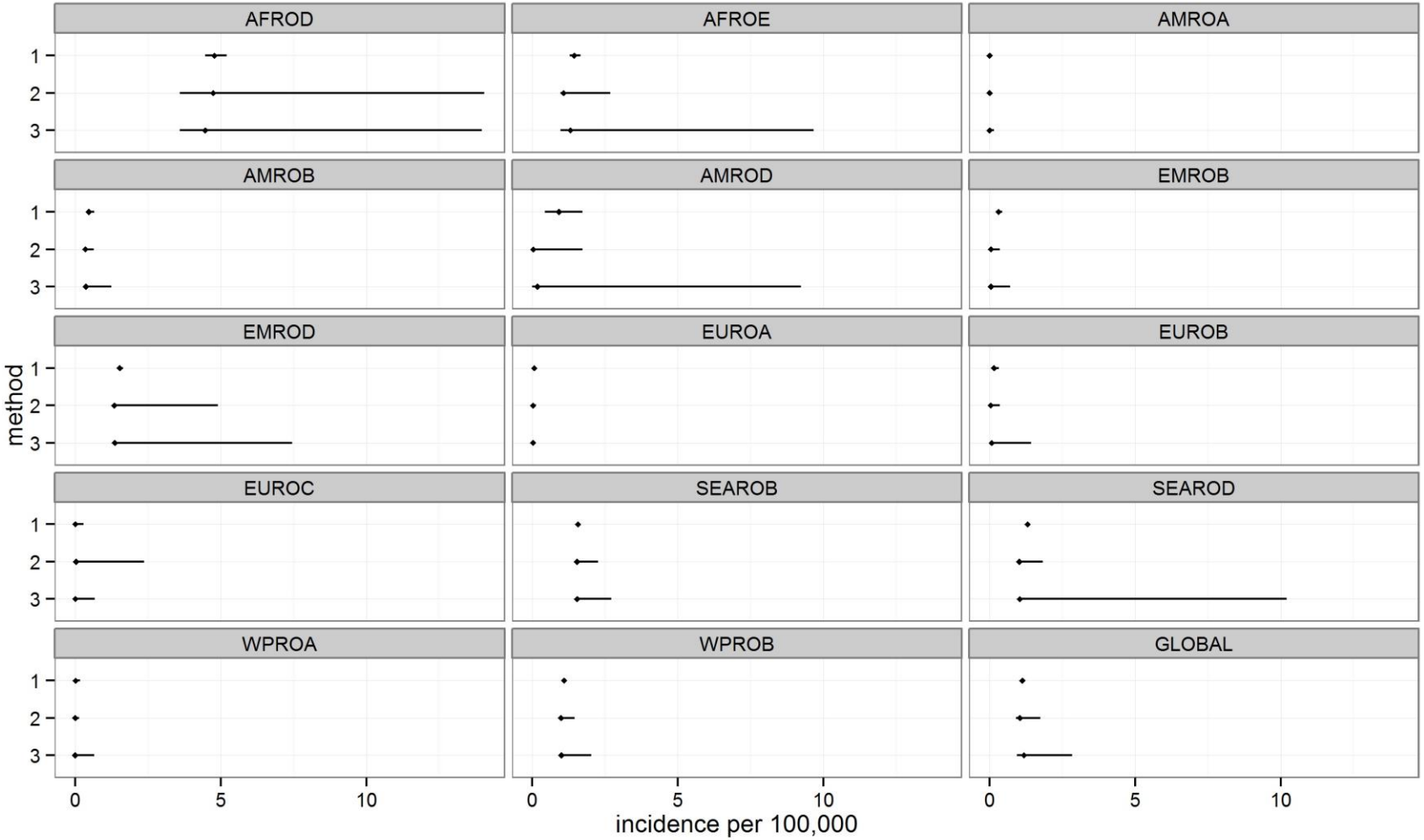
$$\mu_i \sim Normal(\mu_0, \sigma_b^2)$$

eFigures

eFigure 1. *Predicted WHO subregion-level incidence (per 1000 births) of congenital toxoplasmosis, comparing Methods 1, 2 (Bayesian random effects regression with WHO subregion as random effect), and 3 (Bayesian mixed effects regression with WHO subregion as random effect; LASSO-reduced set of nine covariates). Lines indicate 95% prediction intervals.*



eFigure 2. *Predicted WHO subregion-level incidence of aflatoxin-related hepatocellular carcinoma, comparing Methods 1, 2 (Bayesian random effects regression with WHO subregion as random effect), and 3 (Bayesian mixed effects regression with WHO subregion as random effect; LASSO-reduced set of nine covariates). Lines indicate 95% prediction intervals.*



eCode

eCode A1. *Example JAGS/BUGS code to fit a Bayesian random effects log-Normal regression model.*

```
model {

  ## Likelihood, specified using nested indexing
  ## .. n     = number of countries
  ## .. y[]   = country-specific log-transformed incidence rates
  ## .. reg[] = integer indicating to which region a country belongs
  ## .. mu[]  = regional intercept per country
  ## .. tau   = within-region precision
  for(j in 1:n) {
    y[j]   ~ dnorm(mu[j], tau)
    mu[j] <- phi[reg[j]]
  }

  ## Distribution of regional intercepts
  ## .. N       = number of regions
  ## .. phi[]   = regional intercepts
  ## .. phi.c   = global intercept
  ## .. phi.tau = between-region precision
  for(i in 1:N) {
    phi[i] ~ dnorm(phi.c, phi.tau)
  }

  ## Prior distributions
  tau ~ dgamma(0.005, 0.005)
  phi.c ~ dnorm(0, 0.00001)
  phi.tau ~ dgamma(0.005, 0.005)
}
```

eCode A2. *Example JAGS/BUGS code to fit a Bayesian mixed effects log-Normal regression model, specifying two covariates as fixed effects.*

```
model {

  ## Likelihood, specified using nested indexing
  ## .. n          = number of countries
  ## .. y[]        = country-specific log-transformed incidence rates
  ## .. reg[]      = integer indicating to which region a country belongs
  ## .. mu[]       = regional intercept per country
  ## .. tau        = within-region precision
  ## .. beta[]     = covariate coefficients
  ## .. x1[], x2[] = country-specific covariate values (note the centering)
  for(j in 1:n) {
    y[j]   ~ dnorm(mu[j], tau)
    mu[j] <- phi[reg[j]] +
             beta[1] * (x1[j] - mean(x1[])) +
             beta[2] * (x2[j] - mean(x2[]))
  }

  ## Distribution of regional intercepts
  ## .. N       = number of regions
  ## .. phi[]   = regional intercepts
  ## .. phi.c   = global intercept
  ## .. phi.tau = between-region precision
  for(i in 1:N) {
    phi[i] ~ dnorm(phi.c, phi.tau)
  }

  ## Prior distributions
  for(k in 1:2) {
    beta[k] ~ dnorm(0, 0.00001)
  }
  tau ~ dgamma(0.005, 0.005)
  phi.c ~ dnorm(0, 0.00001)
  phi.tau ~ dgamma(0.005, 0.005)
}
```

eReferences

Torgerson PR, Mastroiacovo P. The global burden of congenital toxoplasmosis: a systematic review. Bull World Health Organ 2013; 91: 501-8.

Liu Y, Wu F. Global burden of aflatoxin-induced hepatocellular carcinoma: a risk assessment. *Environ Health Perspect* 2010; 118: 818-24.