

Global Atlas of Helminth Infection – Technical information

This document outlines the analysis strategy used to develop the risk maps used to (i) estimate the prevalence of helminth infections and the probability that targeted mass drug administration (MDA) is needed across endemic countries, (ii) define intervention districts and (iii) estimate numbers at risk. All maps freely available to download and use at www.thiswormyworld.org.

Analysis outline

The objective of these analyses was to determine the global spatial distribution of soil transmitted helminth (STH) and schistosomiasis infection. Prevalence data for STH (hookworm, *Ascaris lumbricoides* and *Trichuris trichiura*) and schistosomiasis (*Schistosoma haematonium* and *S. mansoni*) were collated using search principles and criteria outlined below, in order to create a robustly geo-located dataset of helminth surveys. This database was used to make a continuous cumulative prevalence surface adopting a Bayesian space-time geostatistics approach,¹ adjusting for environmental covariates; no spatial prediction was made for areas masked as environmentally unsuitable for STH transmission. The resulting models were used to interpolate the probability that cumulative STH prevalence is greater than 20%, and that each schistosome species is greater than 50%, the thresholds recommended by the World Health Organization (WHO) as indicating the need for targeted MDA [1]. The total population at risk of STH and schistosome infection was extracted by district in order to guide targeted intervention strategies.

For further information regarding data sources, the space-time modelling approach and defining populations at risk please contact the Global Atlas of Helminth Infection team at www.thiswormyworld.org/about-us/contact-us

Data sources

Survey data were identified through structured searches of electronic bibliographic databases, complemented with manual searches of local archives and libraries and direct contact with researchers. References from identified publications were checked for additional surveys. Estimates

¹ The space-time model-based geostatistics procedures used were initially developed by the Malaria Atlas Project (<http://www.map.ox.ac.uk/>)

of infection prevalence were included according to pre-defined criteria: only cross-sectional prevalence surveys were included; data were excluded if based on hospital or clinic surveys, post-intervention surveys, or surveys among sub-populations, such as among refugees, prisoners or nomads. In instances where multiple surveys from the same location were surveyed at different times, each survey was included. Abstracted data included details on the source of the data, date and location of survey, characteristics of the surveyed population, survey methodology, method of diagnosis, age range of sampled individuals, and the number of individuals examined and the number positive with hookworm, *A. lumbricoides* and *T. trichiura*. Authors of published data were contacted if relevant information was unclear from the original reports. For the current analysis, survey data were collected between 1974 and present day.

The longitude and latitude of each survey were determined using a combination of resources including a national schools databases, village databases digitised from topographical maps and a range of electronic gazetteers (see Brooker et al. [2]); and contact with authors who used GPS. For purpose of this analysis, data points less than 2km apart were treated as a single location.

Estimating cumulative prevalence of STH

The prevalence of infection with any STH species (i.e. cumulative prevalence of STH) was calculated using a simple probabilistic model of combined infection, incorporating a small correction factor to allow for non-independence between species, following the approach of de Silva and Hall [3]. In brief, when assuming the probability of infection with one species to be independent of infection with others, the cumulative probability of having at least one infection is multiplicative: $P_{HAT} = H + A + T - (HA) - (AT) - (HT) + (HAT)$ where P_{HAT} is the cumulative STH prevalence, H is the prevalence of hookworm infection, A the prevalence of *A. lumbricoides* and T the prevalence of *T. trichiura*. Previous analysis of 60 datasets from 20 countries by de Silva and Hall suggests that, due to non-independence, overestimation of cumulative STH prevalence using simple probability increases by 0.6% for every 10% increase in prevalence [3]. The true cumulative prevalence of STH can therefore be estimated as $P_{HAT} \div 1.06$. This correction factor was thus incorporated into the estimation of cumulative prevalence of STH. This approach to estimating cumulative prevalence is used, both when presenting observed prevalence data and when using species specific models to develop spatial models of cumulative prevalence, as explained below.

Ecological and climatic covariates and limits of transmission

Normalised differenced vegetation index (NDVI; a measure of vegetation density) and land surface temperature (LST) based on the period 1992 to 1996 at 5km resolution were obtained for the National Oceanographic and Atmospheric Administration's (NOAA) Advanced Very High Resolution Radiometer (AVHRR) (<http://noaasis.noaa.gov/NOAASIS/ml/avhrr.html>). Temporal Fourier analysis (TFA) was used, and Fourier Processed products (annual amplitude) were assembled

[4,5]. Population density was derived from adjusted population counts for the year 2000 projected to 2009 by applying national, medium variant, inter-censal growth rates [6]. The annual amplitude products for LST and NDVI were standardised to optimise sampling during MCMC by subtracting the arithmetic mean and dividing by the standard deviation.

These ecological data, along with results from previous studies, were used to define the spatial limits for the transmission of STH. Specifically, it has been shown experimentally that the development of free-living infectious stages of *A. lumbricoides* and *T. trichiura* ceases at 38°C and hookworm at 40°C [7-10]. This is supported by an observed relationship between prevalence across sub-Saharan Africa and annual amplitude products for LST and NDVI. On this basis, areas were masked as unsuitable for STH transmission where the annual amplitude products for LST and NDVI exceeded extreme limits (i.e. too hot and/or arid). No spatial prediction was subsequently made for such areas.

Bayesian space-time modelling approach

The probability models used in this study assumed that individuals participating in each sample were egg-positive for helminth infection with a probability that was a continuous function of the time and location of the survey, modified by a set of covariates, and modelled as a Gaussian process [11]. The Bayesian space-time model was implemented in two parts starting with an inference stage in which a Markov Chain Monte Carlo (MCMC) algorithm was used to generate samples from the joint posterior distribution of the parameter set and the space-time random field at the data locations. This was followed by a prediction stage in which samples were generated from the posterior distribution of infection prevalence at each prediction location on a 5 x 5 km grid. Each species was modelled separately. Both the inference and prediction stages were coded using Python (PyMC version 2.0) [12]

For each species, the N_i individuals included in survey i were assumed egg-positive with probability $P(x_i, t_i)$, so that the number positive (Y_i) was distributed binomially:

$$Y_i | N_i, P(x_i, t_i) \sim \text{Binomial}(N_i, P(x_i, t_i))$$

The coefficient $P(x_i, t_i)$ at location x and time t was modelled as the inverse logit function applied to a space-time component, plus an unstructured (random) component. The unstructured component $\epsilon(x_i)$ was represented as a Gaussian process with zero mean and variance V . The space-time

component was represented by a stationary Gaussian process $f(x, t)$ with mean μ and covariance C .

The mean component μ was modelled as a linear function of standardised maximum LST and NDVI, whether the prediction location x was extreme rural (defined as <10 persons per km²) and whether the survey was community-based rather than school-based. The mean component was therefore defined by x parameters:

$$\mu = \beta_x + \sum_{k=1}^K \beta_k X_{i,x,t,k}$$

where $\sum_{k=1}^K \beta_k X_{i,x,t,k}$ denotes the matrix of included covariates and β_x the intercept.

Covariance between spatial and temporal locations was modelled using the space-time covariance function C :

$$C(x_i, t_i, x_j, t_j) = \tau^2 \gamma(0) \frac{(\Delta x)^{\gamma(\Delta t)} K_{\gamma(\Delta t)}(\Delta x)}{2^{\gamma(\Delta t)-1} \Gamma(\gamma(\Delta t) + 1)}$$

$$\gamma(\Delta t) = \frac{1}{2\rho + 2(1 - \rho)[(1 - \nu)e^{-|\Delta t|/\phi t} + \nu \cos(2\pi\Delta t)]}$$

$$\Delta t = |t_i - t_j|$$

K_{γ} is the modified Bessel function of the second kind of order γ , and Γ is the gamma function.

Spatial distances between a pair of points x_i and x_j was computed as the great-circle distance $D_{GC}(x_i, x_j)$ multiplied by a factor that depends on the angle of inclination $\theta(x_i, x_j)$ of the vector pointing from x_i to x_j . θ was computed as if latitude and longitude were Euclidean coordinates (on a cylindrical projection):

$$\Delta x = 2\sqrt{\gamma(\Delta t)} \frac{D_{GC}(x_i, x_j) \sqrt{1 - \psi^2 \cos^2(\theta(x_i, x_j) - \lambda)}}{\phi_x}$$

As temporal separation increases, the covariance approaches a limiting sinusoid

$\tau^2[\rho + (1 - \rho)\nu\cos(2\pi\Delta t)]$ rather than zero. On the other hand, when $\Delta t = 0$ (i.e. points at different locations but the same time), this reduced to a standard exponential form with a range parameter $\phi_x\sqrt{2}$.

The square root of the partial sill τ and the spatial range parameter ϕ_x were assigned skew-normal priors. An exponential, proper prior was assigned to ϕ_τ , and a uniform prior was assigned to the direction of anisotropy parameter λ and to the “eccentricity” parameter ψ^2 , which control the amount of anisotropy. A uniform prior was assigned to the limiting autocorrelation in the temporal direction, ρ , and a standard prior was assigned to the components of the mean.

Model implementation and output

Bayesian inference was implemented using Markov Chain Monte Carlo to generate samples from the posterior distribution of the Gaussian field $f(x_i, t_i)$ at each data location and of the unobserved parameters of the mean, covariance function and Gaussian random noise component.

For each species, samples were generated from the mid-year 2009 mean of the posterior distribution of $f(x_i, t_i)$ at each prediction location at points on a regular 5 × 5 km spatial grid across sub-Saharan Africa. Model output therefore consisted of samples from the predicted posterior distribution of the 2009 infection prevalence at each grid location, which were used to generate point estimates of infection prevalence. Cumulative prevalence of STH was then estimated using the probabilistic model given above, producing a posterior probability distribution for cumulative STH at each prediction location. Probability contour maps were subsequently developed by calculating the observed proportion of the STH posterior probability distribution at each prediction location that exceeded WHO policy intervention thresholds (20% and 50% prevalence) [13].

Defining intervention districts

Prediction locations were classified as endemic if the probability that STH prevalence exceeded 20% (the MDA once-yearly intervention threshold) was > 0.5. These locations were further classified as



Technical Information – Development of predictive risk models

hyper-endemic if the probability that STH prevalence exceeded 50% (the MDA twice-yearly intervention threshold) was > 0.5 . Administrative boundaries were derived from the Global Administrative Unit Layers (GAUL), an initiative implemented by FAO within the EC-FAO Food Security Programme funded by the European Commission (http://www.foodsecinfoaction.org/News/news_06_06.htm). Digital administration level 2 (district) boundaries and a population distribution map (adjusted population counts for the year 2000 projected to 2009 by applying national, medium variant, inter-censal growth rates, as previously described [6], at 5 km² resolution) were overlaid on the endemicity-class surface to extract population-adjusted proportions of each district in each endemicity class. Districts were defined as MDA intervention districts if over 33% of the population were in endemic (once-yearly MDA) or hyperendemic (twice yearly MDA) endemicity classes.

Defining populations at risk

Individuals were classified “at risk” if they lived in a prediction location classified as exceeding 20% prevalence (i.e. endemic for STH). Population counts were derived using the population distribution maps described above. The proportion of the population of school-going age (5-14 years) for each country was based on the World Population Prospects: 2008 Revision Population Database (<http://esa.un.org/unpp/index.asp?panel=3>; accessed 11th May 2010) and the primary-school net enrolment rate was derived from World Bank Indicators (<http://data.worldbank.org/indicator/SE.PRM.NENR>; accessed 11th May 2010; [14]).

References

1. WHO (2006) Preventive chemotherapy in human helminthiasis. Coordinated use of anthelmintic drugs in control interventions: a manual for health professionals and programme managers. World Health Organization, Geneva.
2. Brooker S, Kabatereine NB, Smith JL, Mupfasoni D, Mwanje MT, et al. (2009) An updated atlas of human helminth infections: the example of East Africa. *Int J Health Geogr* 8: 42.
3. de Silva N, Hall A (2010) Using the prevalence of individual species of intestinal nematode worms to estimate the combined prevalence of any species. *PLoS Negl Trop Dis* 4: e655.
4. Hay SI, Tatem AJ, Graham AJ, Goetz SJ, Rogers DJ (2006) Global Environmental Data for Mapping Infectious Disease Distribution. *Adv Parasitol* 62: 37-77.
5. Scharlemann JPW, Benz D, Hay SI, Purse BV, Tatem AJ, et al. (2008) Global data for ecology and epidemiology: a novel algorithm for temporal Fourier processing MODIS data. *PLoS ONE* 3: e1408.
6. Hay SI, Guerra CA, Gething PW, Patil AP, Tatem AJ, et al. (2009) A world malaria map: *Plasmodium falciparum* endemicity in 2007. *PLoS Med* 24: e1000048.
7. Beer RJ (1976) The relationship between *Trichuris trichiura* (Linnaeus 1758) of man and *Trichuris suis* (Schrunk 1788) of the pig. *Research in Veterinary Science* 20: 47-54.
8. Seamster AP (1950) Developmental studies concerning the eggs of *Ascaris lumbricoides* var. *suum*. *The American Midland Naturalist* 43: 450-468.
9. Udonsi JK, Atata G (1987) *Necator americanus* : Temperature, pH, Light, and Larval Development, Longevity, and Desiccation Tolerance. *Experimental Parasitology* 63: 136-142.
10. Smith G, Schad GA (1989) *Ancylostoma duodenale* and *Necator americanus*: effect of temperature on egg development and mortality. *Parasitology* 99: 127-132.
11. Banderjee S, Carlin BP, Gefland AE (2004) Hierarchical modeling and analysis for spatial data. Boca Raton, Florida, USA: Chapman and Hall / CRC Press LLC.
12. Patil AP, Huard D, Fonnesebeck CJ (2010) PyMC: Bayesian stochastic modelling in Python. *J Stat Soft* 35.
13. Diggle P, Thomson MC, Christensen OF, Rowlingson B, Osmer V, et al. (2007) Spatial modelling and the prediction of *Loa loa* risk: decision making under uncertainty. *Annals of Tropical Medicine & Parasitology* 101: 499-509.
14. World Development Indicators (WDI) 2009 Washington D.C.: World Bank; 2009.