# Mapping the Probability of Schistosomiasis and Associated Uncertainty, West Africa

Archie C.A. Clements, Amadou Garba,
Moussa Sacko, Seydou Touré, Robert Dembelé,
Aly Landouré, Elisa Bosque-Oliva,
Albis F. Gabrielli, and Alan Fenwick

We aimed to map the probability of *Schistosoma haematobium* infection being >50%, a threshold for annual mass praziquantel distribution. Parasitologic surveys were conducted in Burkina Faso, Mali, and Niger, 2004–2006, and predictions were made by using Bayesian geostatistical models. Clusters with >50% probability of having >50% prevalence were delineated in each country.

Large-scale control programs for tropical infectious diseases have been initiated in recent years (*1,2*), after renewed commitment by governments and international funding agencies to support the control of previously neglected tropical diseases, including parasitic diseases such as malaria, schistosomiasis, onchocerciasis, lymphatic filariasis, and soil-transmitted helminth infections. Schistosomiasis is the second-most important parasitic disease throughout the world, with an estimated 207 million persons infected (*3*).

Success and sustainability of large-scale disease control programs depend on the allocation of resources where they will have maximum benefit (*4*). Given that tropical infectious diseases, such as schistosomiasis, tend to occur in spatially defined foci (i.e., clusters or hot spots) (*5*), efficient resource allocation relies on identifying the location of high-risk populations. Because disease-endemic countries do not have sophisticated surveillance systems that can accurately delineate disease clusters, alternative meth-

ods such as sample-based spatial prediction need to be applied to target control programs.

Among the multinational and multi-institutional partnerships formed to confront the problem of neglected tropical diseases is the Schistosomiasis Control Initiative (SCI; www.schisto.org), which supports national schistosomiasis and soil-transmitted helminth control programs in Burkina Faso, Mali, and Niger (and other African countries). Lengeler et al. (*6*) describe 2 approaches to targeting interventions: one in which the number of recipient schools or communities is determined by available resources and the other in which a prevalence threshold is defined above which all schools or communities benefit from the intervention. SCI takes the latter approach, delineating areas according to the World Health Organization (WHO)–recommended threshold prevalence of 50% for annual mass treatment. However, even this approach needs to take into account factors such as resource availability and decision risk because uncertainties exist when delineating areas based on the selected threshold.

Knowledge of uncertainty regarding the location and spatial dimensions of clusters is important because it makes possible a prior assessment of the risks and potential consequences associated with different resource allocation strategies. Uncertainties in spatial prediction maps originate from factors such as natural random variation and measurement error of the outcome variable and covariates. Bayesian methods are useful because they provide an approach for propagating uncertainty (through a prediction model) in regards to the spatial predictions. Only recently have practical applications of Bayesian methods in large-scale tropical disease control programs been reported (*7–9*).

## The Study

The objective of this study was to produce maps that could be integrated into the SCI-supported national intervention strategies and that explicitly represent uncertainties in spatial predictions so that national control managers could judge the quality of the evidence upon which the strategies will be based. The SCI-supported programs involve mass distribution of praziquantel (for urinary and intestinal schistosomiasis) and albendazole (for soil-transmitted helminths). The parasitic infection with the highest prevalence is urinary schistosomiasis, caused by flukes (*Schistosoma hematobium*), and the programs are planned to control this disease (*2*).

Parasitologic data were collected in coordinated school-based field surveys in Burkina Faso, Mali, and Niger (Figure 1) during 2004–2006 (preintervention) by using standardized protocols (available on request). The collated dataset covered a spatially contiguous area, ≈2,750 km × 850 km, and included the infection status of 27,939 school-age children in 418 randomly selected locations. Infection
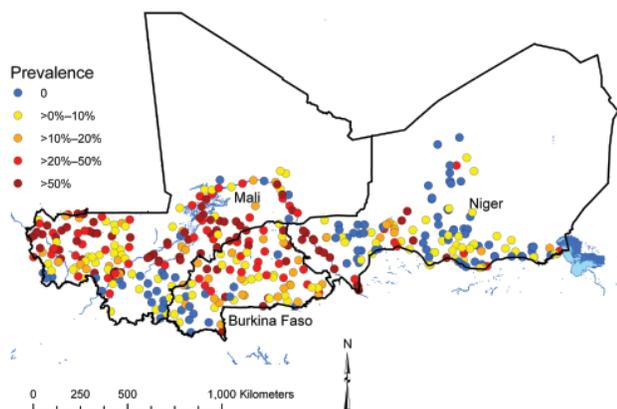
Figure 1. Prevalence of infection with *Schistosoma hematobium* at 418 survey locations in Burkina Faso, Mali, and Niger, 2004–2006.

status was defined according to egg count determined by microscopic examination of urine samples; ≥1 *S. hematobium* eggs indicated infection.

Spatial prediction was based on a logistic regression model (Table), constructed by using the software WinBUGS, version 1.4.2 (MRC Biostatistics Unit, Cambridge and Imperial College, London, UK). The model had infection status as the binary outcome variable, age and sex of the survey participants as individual-level fixed effects, and distance from perennial water body (derived from electronic maps obtained from the Food and Agriculture Organization) and land surface temperature (LST; with a quadratic term; see Hay et al. [10] for details on how these data were derived) as survey location–level fixed effects. Variable selection methods and the model are presented in the online Technical Appendix (available from www.cdc.gov/EID/content/14/10/1629-Techapp.pdf). The model also included a geostatistical random effect for residual spatial clustering of infection prevalence (11).

A prevalence map for the study area was constructed, using the model, by predicting infection prevalence at the centroids of cells of a 0.15 × 0.15 decimal degree (≈18 km × 18 km) grid. This model was implemented with the *spatial.unipred* command of WinBUGS (details are provided in the online Technical Appendix). Estimates from Bayesian models are distributions (termed posterior distributions) that represent the probability of each of a range of plausible values being true for the variable being modeled. To quantify the uncertainties surrounding the model predictions, we plotted the probability of each prediction location having a prevalence >50%, rather than mean predicted prevalence at each location. The probabilities were calculated from the posterior distributions of predicted prevalence at each location (i.e., if 95% of the posterior distribution of predicted prevalence was >0.5, the probability of prevalence >50% at that location was 95%).

Cross-validation was done by randomly allocating survey locations to 3 groups and undertaking 3 separate runs of the model; 1 of the 3 groups was sequentially omitted, and predicted prevalence at the omitted locations was determined by using the model. Predicted prevalence was compared with observed prevalence, dichotomized, according to a 50% observed prevalence threshold. The comparison statistic was the area under the curve (AUC) of the receiver operating characteristic, and a value of >0.7 was considered to indicate acceptable predictive ability. An average AUC was calculated across the 3 model runs.

In the final model (Table), statistically significant correlations suggested that infection prevalence was higher in older boys and increased with proximity to perennial bodies of water, but no association was found between prevalence and LST. The range over which spatial correlation was >5% (chosen to indicate statistically important spatial correlation) was ≈177 km, indicating the approximate radius of clusters. Results of the validation analysis showed an average AUC of 0.86, indicating that the model had an acceptable predictive performance.

Bayesian probability maps were produced for each sex and age group, but for illustrative purposes we present predicted probability of prevalence >50% in boys ages 13–16 years (the group with the highest infection prevalence; Figure 2). Large clusters of prediction locations with a high probability (i.e., >50%; indicative of low uncertainty) of

Table. Bayesian logistic regression model of prevalence of infection with *Schistosoma haematobium* in children in 418 schools in Burkina Faso, Mali, and Niger, 2004–2006*

| Variable | Posterior distribution | |
| --- | --- | --- |
| | Mean (95% CrI) | SD |
| Female gender | 0.70 (0.65–0.76) | 0.03 |
| Age, y | | |
| 9–10 | 1.16 (1.00–1.33) | 0.08 |
| 11–12 | 1.51 (1.31–1.73) | 0.10 |
| 13–16 | 1.79 (1.53–2.06) | 0.14 |
| Distance to perennial water body | 0.34 (0.21–0.54) | 0.08 |
| Land surface temperature | 0.80 (0.51–1.21) | 0.18 |
| Land surface temperature$^2$ | 1.10 (0.85–1.40) | 0.14 |
| Rate of decay of spatial correlation | 2.03 (1.48–2.74) | 0.32 |
| Variance of the spatial random effect (sill) | 7.03 (5.36–9.31) | 1.01 |

*CrI, Bayesian credible interval. Values for the fixed effects are odds ratios; note the odds ratios for the climate variables are on a common scale, where the variables were standardized to have a mean = 0 and SD = 1. The reference group for sex was boys and for age was 6–8 y. The number of children found to be infected with *S. haematobium* was modeled by using a binomial distribution described by the proportion infected and the total number sampled in each survey location. The proportion infected was modeled by using logistic regression with an intercept, covariates (sex, age, distance to perennial water body, land surface temperature, and a quadratic term for land surface temperature), and a random effect that described spatial correlation (i.e., clustering). Model outputs were distributions (termed *posterior distributions*) that can be summarized by using the mean, SD, and 95% CrI (representing the range of values that contains the true value with a probability of 95%). More details on the model are presented in the online Technical Appendix (available from www.cdc.gov/EID/content/14/10/1629-Techapp.pdf).
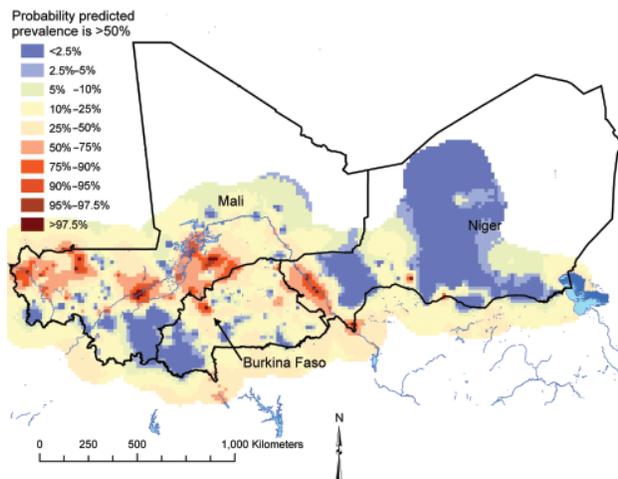
Figure 2. Predicted probability of prevalence of infection with *Schistosoma hematobium* being >50% in Burkina Faso, Mali, and Niger in boys ages 13–16 years; results are based on a Bayesian geostatistical model. The red areas had a low degree of uncertainty that predicted prevalence was >50%, and the blue areas had a high degree of uncertainty that predicted prevalence was >50%.

prevalence being >50% intervention threshold were located in a mid-latitudinal band across Mali, running from western to central regions, and in the Niger River valley region of Niger. Smaller clusters were located in various parts of southern and eastern Mali, northwestern and northeastern Burkina Faso, and south-central Niger.

## Conclusions

Future schistosomiasis control plans should acknowledge uncertainties such as those presented in Figure 2. A possible approach would be to introduce a second threshold for the level of uncertainty that a location is above the intervention prevalence threshold; if the uncertainty is greater than this second threshold, then the location is excluded until new evidence is obtained that confirms prevalence is above or below the intervention prevalence threshold. This second uncertainty threshold should be determined by the quantity of resources available for disease control and the level of decision risk deemed appropriate.

In addition to providing an evidence base for distributing resources in 3 West African countries as part of the SCI-supported national control programs, the maps presented here have a potential role in maintaining sustainability of schistosomiasis control after SCI support ends (SCI is funded through 2009). They can be used as advocacy tools for channeling funds to high-risk populations in the affected countries and, in the likely event that money for schistosomiasis control in these countries becomes more limited after SCI support ends, they can be used to ensure

that scarce governmental resources are distributed as efficiently as possible. National coordinators who might face accountability for targeted (i.e., unequal) distribution of resources will benefit from the defendable, scientifically sound methods presented in this article. By focusing on uncertainty in spatial predictions, more flexible tools for disease control can be developed that allow the geographic dimensions of control programs to be scaled and modified according to available resources and acceptable levels of decision risk.

Dr Clements is an infectious disease epidemiologist with a particular interest in the spatial epidemiology of tropical infectious diseases. A major focus of his research is implementation of Bayesian spatial modeling and prediction to enhance planning of real-world, large-scale tropical infectious disease control programs.

## References

1. Kabatereine NB, Fleming FM, Nyandindi U, Mwanza JC, Blair L. The control of schistosomiasis and soil-transmitted helminths in East Africa. Trends Parasitol. 2006;22:332–9. DOI: 10.1016/j.pt.2006.05.001
2. Garba A, Toure S, Dembele R, Bosque-Oliva E, Fenwick A. Implementation of national schistosomiasis control programmes in West Africa. Trends Parasitol. 2006;22:322–6. DOI: 10.1016/j.pt.2006.04.007
3. Steinmann P, Keiser J, Bos R, Tanner M, Utzinger J. Schistosomiasis and water resources development: systematic review, meta-analysis, and estimates of people at risk. Lancet Infect Dis. 2006;6:411–25. DOI: 10.1016/S1473-3099(06)70521-7
4. Rabinovich NR. The renaissance in global health. Trends Parasitol. 2006;22:277. DOI: 10.1016/j.pt.2006.05.008
5. Brooker S. Spatial epidemiology of human schistosomiasis in Africa: risk models, transmission dynamics and control. Trans R Soc Trop Med Hyg. 2007;101:1–8. DOI: 10.1016/j.trstmh.2006.08.004
6. Lengeler C, Utzinger J, Tanner M. Screening for schistosomiasis with questionnaires. Trends Parasitol. 2002;18:375–7. DOI: 10.1016/S1471-4922(02)02318-8

7. Clements ACA, Moyeed R, Brooker S. Bayesian geostatistical prediction of the intensity of infection with *Schistosoma mansoni* in East Africa. Parasitology. 2006;133:711–9. DOI: 10.1017/S0031182006001181

8. Clements AC, Lwambo N, Blair L, Nyandindi U, Kaatano G, Kinunghi S, et al. Bayesian spatial analysis and disease mapping: tools to enhance planning and implementation of a schistosomiasis control programme in Tanzania. Trop Med Int Health. 2006;11:490–503. DOI: 10.1111/j.1365-3156.2006.01594.x

9. Clements ACA, Brooker S, Nyandindi U, Fenwick A, Blair L. Bayesian spatial analysis of a national urinary schistosomiasis questionnaire to assist geographic targeting of schistosomiasis control in Tanzania, East Africa. Int J Parasitol. 2008;38:401–5. DOI: 10.1016/j.ijpara.2007.08.001

10. Hay SI, Tatem AJ, Graham AJ, Goetz SJ, Rogers DJ. Global environmental data for mapping infectious disease distribution. Adv Parasitol. 2006;62:37–77. DOI: 10.1016/S0065-308X(05)62002-7

11. Diggle P, Tawn J, Moyeed R. Model-based geostatistics. J R Stat Soc Ser C Appl Stat. 1998;47:299–350. DOI: 10.1111/1467-9876.00113

Address for correspondence: Archie C.A. Clements, School of Population Health, University of Queensland, Herston, Queensland, Australia; email: a.clements@uq.edu.au
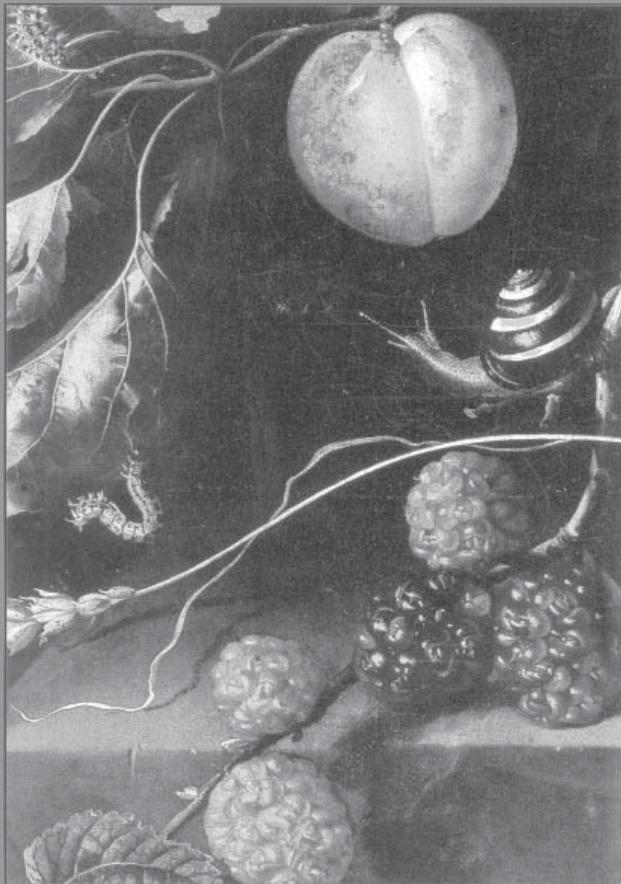
EMERGING INFECTIOUS DISEASES

Tracking trends and analyzing new and reemerging infectious disease issues around the world

A peer-reviewed journal published by the National Center for Infectious Diseases

Vol. 3, No. 4, Oct–Dec 1997

Search past issues of EID at www.cdc.gov/eid

Conference on Emerging Foodborne Pathogens

Pathogen Emergence

Host Susceptibility

Chronic Sequelae

Risk Assessment

Consumer Concern

# Mapping the Probability of Schistosomiasis and Associated Uncertainty, West Africa

**Technical Appendix**

**Variable Selection**

Covariates were initially selected on the basis of biological plausibility. Age and sex have both been demonstrated to be associated with urinary schistosomiasis prevalence likely due to physiologic differences in susceptibility. Transmission of *Schistosoma haematobium* involves contact with cercariae in water in which the intermediate host snail, *Bulinus* spp., is present. Proximity to a perennial water body is, therefore, a plausible risk factor for exposure and subsequent infection. The intermediate host snail has an optimal temperature range above and below which survival is impeded, justifying the inclusion of a quadratic term for land surface temperature (LST). Precipitation levels might indicate the availability of snail habitats and (in high-precipitation areas) the risk of washout of snail populations; normalized difference vegetation index (NDVI) is a proxy measure of rainfall that was also considered for inclusion in the model. Elevation was not considered for inclusion because of collinearity with temperature and precipitation. Therefore, the initial candidate set of variables included sex, age, proximity to perennial water bodies, LST (with and without a quadratic term) and NDVI. Variable selection was done in Stata/SE 10.0 (StataCorp, College Station, TX, USA) by using a fixed-effects logistic regression model with backwards variable selection, with an exclusion criterion of Wald p>0.2. NDVI was excluded and all remaining variables were selected for inclusion in the spatial model. The quadratic term for LST was significant.

**Model Building and Assessment**

The model was of the form

$$Y_{i,j} \sim Binomial(n_{i,j}, p_{i,j})$$

where $Y_{i,j}$ was the number of positive infection status individuals, $n_{i,j}$ was the number tested and $p_{i,j}$ was the risk for positive infection status in age-gender group $i$, location $j$;

$$logit(p_{i,j}) = \alpha + \beta \times girl_{i,j} + \delta1 \times age1_{i,j} + \delta2 \times age2_{i,j} + \delta3 \times age3_{i,j} + \lambda_j$$

where $\alpha$ was the intercept, $\beta$ was the coefficient for female gender, $\delta1$-3 were the coefficients for age groups 9–10, 11–12, and 13–16 years;

$$\lambda_j = \varepsilon \times dist_j + \gamma \times LST_j + \kappa \times LST_j^2 + \theta_j$$

where $\varepsilon$ was the coefficient for distance to perennial water body, $\gamma$ was the coefficient for land surface temperature (LST), $\kappa$ was the coefficient for the quadratic term of LST; and $\theta_j$ was defined by the isotropic, exponentially decaying correlation function

$$f(d_{ij};\phi) = \exp\left[-(\phi d_{ij})\right]$$

where $d_{ij}$ are the distances between pairs of points $i$ and $j$, and $\phi$ is the rate of decline of spatial correlation per unit of distance. Noninformative priors were specified for the intercept (uniform prior with bounds $-\infty$ and $\infty$) and the coefficients (normal prior with mean $= 0$ and precision, the inverse of variance, $= 1 \times 10^{-4}$). The prior distribution of $\phi$ was also uniform with upper and lower bounds set at 0.1 and 50. The precision of $\theta_j$ was given a noninformative prior gamma distribution.

**Implementation of the *spatial.unipred* Command in WinBUGS**

In WinBUGS, interpolation of a variable to nonsampled locations can be done using the *spatial.unipred* command, where the spatial correlation structure of that variable has been modeled by using an exponentially decaying correlation function fitted to the observed data. The *spatial.unipred* interpolation function uses kriging, where predicted values are a weighted average of observed data obtained from nearby locations with weights being apportioned according to distance and direction from the prediction location. In our models, the interpolated variable was the spatial random effect. Predicted prevalences were calculated by adding the interpolated random effect to the sum of the products of the coefficients for the fixed effects and the values of the fixed effects at each prediction location. For the individual-level fixed effects (sex and age), separate calculations were done, where the coefficient for the relevant age and sex were added to the sum. The overall sum was then back-transformed from the logit scale to the prevalence scale, giving prediction surfaces for prevalence of infection in each age and sex group.