Published in final edited form as:
Trends Parasitol. 2002 February ; 18(2): 70–74.

Tools from ecology: useful for evaluating infection risk models?

Simon Brooker[*],
Dept of Infectious Disease Epidemiology, Imperial College School of Medicine, Norfolk Place, London, UK W2 1PG

Simon I. Hay, and
Trypanosomiasis and Land Use in Africa (TALA) Research Group, Dept of Zoology, University of Oxford, South Parks Road, Oxford, UK OX1 3PS

Don A.P. Bundy
Human Development Division, The World Bank, Washington DC 20433, USA

Abstract

Despite the increasing number of models to predict infection risk for a range of diseases, the assessment of their spatial limits, predictive performance and practical application are not widely undertaken. Using the example of *Schistosoma haematobium* in Africa, this article illustrates how ecozonation and receiver–operator characteristic analysis can help to assess the usefulness of available models objectively.

The resources targeted at parasite control are finite and often limited. Consequently, when designing control programs, it is essential to know the distribution and abundance of a disease, to devise and target intervention strategies and to optimize the use of available resources. In many African countries, the paucity of epidemiological data hinders the quantification of disease burden for basic planning. In an effort to overcome this problem, environmental data (often derived from satellite sensors) are increasingly used to predict infection risk [1-3]. There is a need, in such approaches, to evaluate objectively the predictive accuracy of the models used to generate risk maps and to consider the spatial extent to which models can be reliably extrapolated [4].

Developing predictive models

Reliable maps of infectious diseases require an understanding of whether models developed for one location can be applied to another because the environmental factors that influence disease transmission are unlikely to be uniform over large geographical areas [5]. Political boundaries are routinely used to define the spatial extent of risk maps, but the ecological heterogeneity, which is usually independent of political boundaries, is ignored. Alternatively, remotely-sensed (RS) derived environmental data can be used to develop ecological zone maps that identify areas of ecological similarity* (Box 1), and are better at defining where existing predictive models can and cannot be applied [6].

The statistical methods commonly used to predict the occurrence or distribution of disease in relation to environmental variables are logistic regression and discriminant analysis, and these approaches have been used to map filariasis[3], malaria [6,7], leishmaniasis [8] and

[*] s.brooker@ic.ac.uk .

*Rogers, D.J. and Wint, W. (1996) *Towards Identifying Priority Areas for Tsetse Control in East Africa*. Consultants' Report to the Food and Agriculture Organisation. Trypanosomiasis and Land Use in Africa, Oxford, UK.

trypanosomiasis [2]. The predictive performance of these models, which is a prerequisite for their refinement [9], is evaluated by examining the agreement between predictions and observations [10] by using data collected from sites other than those used in model development. In logistic regression models, predictions are based on the model outputs that measure the probability of infection occurrence. To assess the predictive performance of a model, a probability threshold needs to be identified that can differentiate locations of relative risk. Often, however, the researcher is confronted with the question: which threshold to select for when discriminating between these different populations. A commonly used method in medical diagnostics, and more recently in ecological studies, is receiver–operator characteristic (ROC) analysis, which can provide an overall measure of model accuracy and can explore the consequences of choosing any given threshold (Box 2).

Schistosomiasis

In common with other infectious diseases, the design of schistosomiasis control operations in Africa is typically constrained by a lack of comprehensive survey data [11]. Climate and other environmental variables influences the distribution of schistosomiasis (either intestinal *Schistosoma mansoni* or urinary *Schistosoma haematobium*) at broad spatial scales [12], hence RS-derived environmental data have potential for predicting transmission patterns [13]. Such an approach has been used to map the likely distribution of *S. mansoni* in Ethiopia [14] and Egypt [15], characterized areas directly relevant to the rational targeting of control.

The WHO recommends mass treatment with praziquantel in areas where infection prevalence of schistosomiasis equals or exceeds 50%. Because schistosomiasis exhibits strong spatial heterogeneity at local levels, there is a need to locate high-risk communities or schools that require mass treatment. For *S. haematobium*, this has been achieved effectively by using morbidity questionnaires [16]. The first step in national planning is to locate areas for which questionnaire surveys are needed, which can be identified by using broad-scale risk maps.

Following the analysis of detailed data from school-based studies in Cameroon and Tanzania (Fig. 1), we have recently developed risk maps for *S. haematobium* [17,18]. Logistic regression modeling was used to identify environmental variables which were significantly associated with infection patterns, and to develop separate, local models of the probability of having an infection prevalence >50% (WHO recommended treatment threshold) based on a random 50% subset of data from Cameroon and Tanga Region in Tanzania. Evaluation was undertaken using the remaining 50% subset, and ROC analysis showed that the models for both Cameroon and Tanga Region allow reasonable discrimination (Fig. I in Box 2) between high- and low-prevalence schools within the geographical areas in which the surveys were conducted.

Do the models perform reliably in other areas?

The real test of a model lies in applying it to different locations. We therefore validated predictions from the model developed in the Tanga Region by using independent data from the Magu, Kilosa, Mtwara and Tandahimba districts of Tanzania (Fig. 1). The Cameroon model was also validated using all the data from Tanzania. ROC analysis indicated that within Tanzania, the model developed for Tanga Region performed reasonably well in neighboring Kilosa District and further south in the ecologically similar coastal districts of Mtwara and Tandahimb (Fig. Ic and Id in Box 2). However, the models performed poorly in the ecologically distinct Magu District, near Lake Victoria (Fig. Ie in Box 2). By contrast, the ecological model for Cameroon could not be reliably be applied to any region of Tanzania (Fig. If in Box 2).

These results can be explained by reference to the ecological zone map (Fig. I in Box 1). This map shows that Tanzania comprises three ecological zones (Table I in Box 1). The areas where the Tanga Region model reliably predicts a high prevalence of schistosomiasis all fall within the same ecological zone (Zone 1). By contrast, the performance of the model is poor near Lake Victoria – an area represented by a different ecological zone (Zone 2). Cameroon has completely different ecological zones from Tanzania, which could explain why the model for Cameroon does not predict schistosomiasis anywhere in Tanzania. It is also useful to consider the distribution of the snail species involved in local transmission. In Cameroon, the predominant snail species in areas of high prevalence (northern Cameroon) is *Bulinus senegalensis*, which inhabits semi-permanent water bodies and can survive the dry season by aestivation [19]. This distribution supports Wright's original contention [20] that *B. senegalensis* occurs throughout the semi-arid areas of the West African Sudan Savanna, confirmed by available snail survey data [21]. The snail distributions suggest that the ecological model for Cameroon is actually predicting the niche of *B. senegalensis*, which is characterized by limited rainfall and high temperatures (Table I in Box 1). Interestingly, this habitat range broadly corresponds to the purple zone in the ecological zone map.

The distribution of different snail species could also explain why the Cameroon model failed to predict schistosomiasis in Tanzania, where the main snail species are *Bulinus africanus, Bulinus globosus* and *Bulinus nasutus* [22]. In coastal areas, *B. globosus* is the principal snail host responsible for transmission, although *B. africanus* might be important locally. By contrast, *B. nasutus* is the main host in northwestern Tanzania and is found in temporary water bodies. *B. nasutus* does occur in coastal East Africa, but appears to be incompatible with the *S. haematobium* parasite strain found in *B. globosus* in that area [23]. Such differences in the distribution of snail species explain why the model developed for Tanga Region performed well in coastal areas with a common snail species but inadequately in Magu District where a different intermediate host occurs.

These findings suggest that the model for Tanga Region can be extended elsewhere within the same ecological zone, which stretches down to Mozambique and parts of southern Malawi. Ideally, a different model would be needed for other areas of Tanzania. The findings also suggest that the model developed for Cameroon can be applied within common ecological zones throughout West Africa, but not across different ecological zones. Knowledge of snail distributions is also of obvious significance in developing and applying any ecological model for *S. haematobium*, given the differing relationships between ecological factors and different snail species [24]. Sadly, however, detailed snail distribution maps do not exist currently, although there are efforts to map the data in snail databases from Africa held by the Danish Bilharziasis Laboratory and other research institutions (T.K. Kristensen, pers. commun.).

Control applications

The risk maps developed will be unable to capture the well-known foci of schistosomiasis: heterogeneities in water contact patterns [25], and the genetic diversity of *S. haematobium* [23] will also influence patterns on a local scale. The large-area RS-based models, however, are useful for identifying potential areas of high risk (Fig. 2). In particular, these maps can exclude areas where urinary schistosomiasis is unlikely to be prevalent, and so help focus on priority areas where local detailed questionnaire surveys are required to target control more precisely. In addition, these maps can help to estimate the target population for such control.

Ultimately, the usefulness of such risk maps is subjective but should be evaluated only within the pre-defined context of its application [5] and based on the consequences of correct and incorrect predictions [26]. If, for example, the intended aim of the model is to maximize exclusion of low-prevalence schools then the threshold chosen to define areas of high-risk should favor specificity (Box 2). If, however, the aim is to maximize inclusion of high-prevalence schools then the threshold should favor sensitivity. Examination of the ROC curve identifies the probability threshold that optimizes the preferences for maximal sensitivity or specificity. Here, a probability threshold that maximizes the accurate exclusion of low-prevalence schools (i.e. $p = 0.2$, see Figs Ia,b in Box 2) to define potential areas at high risk for *S. haematobium* was used.

Overlaying this threshold with available population data on schoolchildren[‡] characterizes those at risk of significant schistosomiasis transmission and those as the target of a questionnaire approach to be quantified and associated program costs estimated. For example, 1.9 million children (in 11 out of the 49 districts) in Cameroon and 4.9 million children (in 37 out of 97 districts) in Tanzania were estimated to be the target for a school-based national schistosomiasis control program. Using the available cost data for control from Ghana and Tanzania, the associated program costs for a single treatment of the school population in Tanzania would be US$ 1.0 million–3.2 million, and US$ 0.4 million–1.3 million in Cameroon [17-18].

Concluding remarks

Prediction models are increasingly being developed for a variety of infectious diseases. Validation of such models is a difficult but an essential issue [9] if models are to have practical relevance for control. ROC plots are a useful addition to logistic regressions procedures for disease risk mapping by optimizing the choice of probability threshold required to satisfy stated control objectives. Ecological zone maps are also of value in defining the spatial extent to which such models can be applied, and can identify areas where further survey data and new models are required (see also [6]).

Ecological zonation has great potential in elucidating the biological mechanisms underlying disease distributions. Further work is required on which variables are used to construct ecozones, criteria for defining the optimal number of zones and how these outputs relate to more traditional vegetation and climate classifications. These considerations should be informed by an understanding of parasite–habitat requirements. Consideration should also be given to future work involving the sampling design and spatial scale of field surveys used to develop and validate prediction models [5]. The challenge now is not to demonstrate the possibility of predicting infection risk, but to demonstrate their accuracy, spatial limitations and most importantly, their practical application to specific control objectives.

Acknowledgments

We thank David Rogers and William Wint for valuable discussions on the concept of ecological zoning and its potential application in epidemiology. Raoult Ratard, Joanne Webster, Andrew Hall, Andrew Roddam, Nicholas Lwambo and members of the Tanzania Partnership for Child Development (PCD) are also thanked for their input. The survey work in Cameroon was funded by USAID and the survey work in Tanzania was funded by the PCD, the Edna McConnell Clark Foundation and The Wellcome Trust. S.B. and S.I.H. are supported by a Wellcome Trust Prize Fellowship (#062692) and Wellcome Trust Advanced Training Fellowship (#056642), respectively.

References

1. Hay SI, et al. Earth observation, geographic information systems and *Plasmodium falciparum* malaria in sub-Saharan Africa. Adv. Parasitol. 2000; 47:174–215.

[‡] The data on the schoolchildren population for every district was derived from the 1990 national population forecasts (http://grid2.cr.usgs.gov/globalpop/africa/), and was projected to 2001 using assumptions based on country- and year-specific inter-census growth rates (http://www.census.gov/ipc/www/idbnew.html).

2. Rogers DJ. Satellites, space, time and the African trypanosomiases. Adv. Parasitiol. 2000; 47:130–165.

3. Lindsay SW, Thomas CJ. Mapping and estimating the population at risk from lymphatic filariasis in Africa. Trans. R. Soc. Trop. Med. Hyg. 2000; 94:37–44. [PubMed: 10748895]

4. Kitron U. Risk maps: transmission and burden of vector-borne diseases. Parasitol. Today. 2000; 16:324–325. [PubMed: 10900476]

5. Guisan A, Zimmermann NE. Predictive habitat distribution models in ecology. Ecol. Model. 2000; 135:147–186.

6. Kleinschmidst I, et al. An empirical malaria distribution map for West Africa. Trop. Med. Int. Health. 2001; 6:779–786. [PubMed: 11679126]

7. Thomson MC, et al. Predicting malaria infection in Gambian children from satellite data and bed net use surveys: The importance of spatial correlation in the interpretation of results. Am. J. Trop. Med. Hyg. 1999; 61:2–8. [PubMed: 10432046]

8. Thomson MC, et al. Towards a kala azar risk map for Sudan: mapping the potential distribution of *Phlebotomus orientalis* using digital data of environmental variables. Trop. Med. Int. Health. 1999; 4:105–133. [PubMed: 10206264]

9. Rykiel EJ. Testing ecological models: the meaning of validation. Ecol. Model. 1996; 90:229–244.

10. Fielding AH, Bell JF. A review of methods for the assessment of prediction errors in conservation presence/absence models. Environ. Conserv. 1997; 24:38–49.

11. Brooker S, et al. Towards an atlas of human helminth infection in sub-Saharan Africa: the use of geographical information systems (GIS). Parasitol. Today. 2000; 16:303–307. [PubMed: 10858650]

12. Brown, DS. Freshwater Snails of Africa and their Importance. Taylor & Francis; 1994.

13. Brooker S, Michael E. The potential of geographical information systems and remote sensing in the epidemiology and control of human helminth infections. Adv. Parasitol. 2000; 47:245–288. [PubMed: 10997209]

14. Malone JB, et al. Satellite climatology and the environmental risk of *Schistosoma mansoni* in Ethiopia and East Africa. Acta Trop. 2001; 79:59–72. [PubMed: 11378142]

15. Malone JB, et al. Geographic information systems and the distribution of *Schistosoma mansoni* in the Nile Delta. Parasitol. Today. 1997; 13:112–119. [PubMed: 15275115]

16. Red Urine Study Group. Identification of high risk communities for schistosomiasis in Africa: a multi-country study. World Health Organization; 1995. Social and Economic Research Project Reports, No. 15

17. Brooker S, et al. Predicting the distribution of urinary schistosomiasis in Tanzania using satellite sensor data. Trop. Med. Int. Health. 2001; 6:998–1007. [PubMed: 11737837]

18. Brooker S, et al. Using NOAA–AVHRR data to model human helminth distributions for planning disease control in Cameroon. Photo. Eng. R. Sens. in press.

19. Greer GJ, et al. Human schistosomiasis in Cameroon II. Distribution of the snail hosts. Am. J. Trop. Med. Hyg. 1990; 6:573–580. [PubMed: 2372088]

20. Wright CA. A note on the distribution of *Bulinus senegalensis*. W. Afr. Med. J. 1959; 8:142–148.

21. Betterton C, et al. *Bulinus senegalensis* (Mollusca: Planorbidae) in northern Nigeria. Ann. Trop. Med. Parasitol. 1983; 77:143–149. [PubMed: 6882063]

22. McCullough FS. The distribution of *Schistosoma mansoni* and *S. haematobium* in East Africa. Trop. Geog. Med. 1972; 24:199–207.

23. Stothard JR, et al. The transmission status of *Bulinus* on Zanzibar Island (Unguja) with implications for control of urinary schistosomiasis. Ann. Trop. Med. Parasitol. 2000; 94:87–94. [PubMed: 10723528]

24. Kristensen TK, et al. Use of satellite remote sensing and geographic information systems to model the distribution and abundance of snail intermediate hosts in Africa: a preliminary model for *Biomphalaria pfeifferi* in Ethiopia. Acta Trop. 2001; 79:73–78. [PubMed: 11378143]

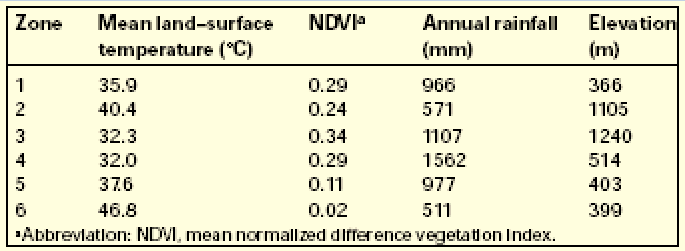
25. Woolhouse, MEJ. Epidemiology of human schistosomes. In: Scott, ME.; Smith, G., editors. Parasitic and infectious diseases: epidemiology and ecology. Academic Press; 1994. p. 197-217.

26. Zweig MH, Campbell G. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. Clin. Chem. 1993; 39:561–577. [PubMed: 8472349]

Box 1

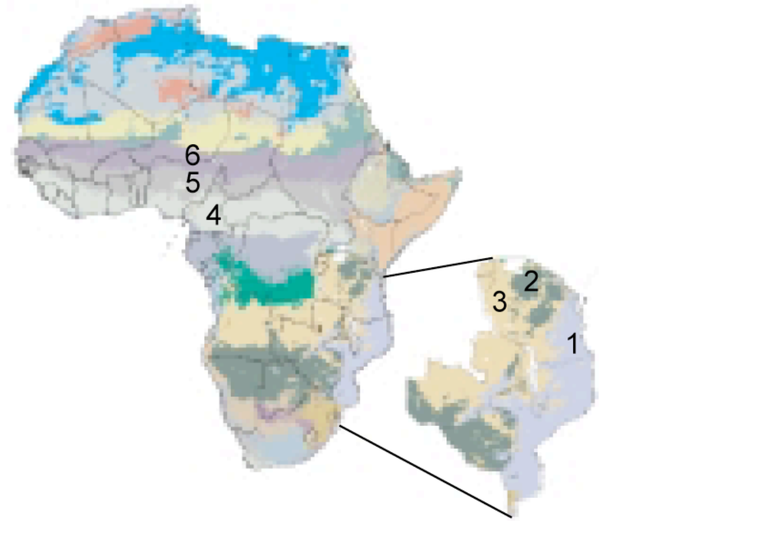
The ecological zones in Africa based on remotely-sensed derived ecological variables The map in Fig. I is based on the mean annual summaries (1982–2000) of multi-temporal remotely-sensed (RS) derived data from the advanced very high resolution radiometer (AVHRR). These data were processed using standard procedures to provide middle-infrared brightness, temperature, land–surface temperature (LST) and photosynthetic activity estimates (expressed as the normalized difference vegetation index, NDVI) for the African continent [a]. Ref. [a] also discusses the several problems involved with using satellite-derived imagery. These data in combination with a digital elevation model (DEM) of Africa (http://edcwww.cr.usgs.gov/landdaac/gtopo30) generated twenty ecological zones using the unsupervised classification procedures of Earth Resources Data Analysis System (ERDAS) Imagine 8.4™ software. ERDAS implements the Iterative Self-Organizing Data Analysis Technique (ISODATA), an iterative method that uses the Euclidean distance as a similarity measure for clustering data into different classes [b]. A complete synopsis of the environmental criteria defined and hence separated each zone is not covered here, but Table I shows the mean values of the clusters that defined the zones mentioned in the text. Visual comparison indicated a good pattern that matches to the widely used White’s vegetation map and agro-ecological zones of Africa [c,d] by the Food Agriculture Organization, which was based on physiognomy and floristic composition, and length of growing season, respectively.

Table I. Mean values of key environmental variables according to zones



Zone	Mean land–surface temperature (°C)	NDVI[a]	Annual rainfall (mm)	Elevation (m)
1	35.9	0.29	966	366
2	40.4	0.24	571	1105
3	32.3	0.34	1107	1240
4	32.0	0.29	1562	514
5	37.6	0.11	977	403
6	46.8	0.02	511	399

[a]Abbreviation: NDVI, mean normalized difference vegetation index.



TRENDS in Parasitology

References

a. Hay SI. An overview of remote sensing and geodesy for epidemiology and public health applications. Adv. Parasitol. 2000; 47:2–27.

b. Jensen, JR. Introductory Digital Image Processing. A Remote Sensing Perspective. Prentice-Hall; 1996. Thematic information extraction: image classification; p. 197-256.

c. White, F. The vegetation of Africa – a descriptive memoir to accompany the UNESCO/AETFAT/UNSO vegetation map of Africa. United Nations Educational, Scientific and Cultural Organization; 1983. Natural Resources Research Report XX

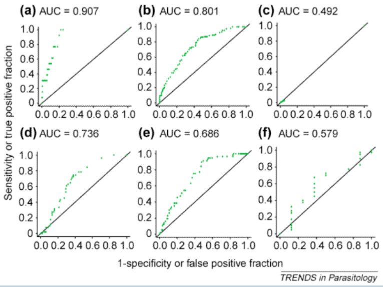
d. FAO. Report on the argo-ecological zones project, Vol. 1: Methodology and results for Africa. World Soil Res. Rep. 1978; 48:32–71.

Box 2

Receiver-operator characteristic analysis for the epidemiology of infectious diseases

The predictive accuracy of prediction models based on logistic regression can be assessed in terms of sensitivity (the percentage of locations with infection and or disease correctly predicted as such) and specificity (the percentage of locations without infection and or disease correctly predicted as such). The predictions of disease occurrence are based on whether the predicted probability arising from the logistic regression model is above or below a chosen probability threshold. Varying the probability threshold across a range of values will generate a series of pairs of sensitivity and specificity. This series of points defines a receiver-operator curve (ROC), and describes the compromise that a model attains between sensitivity and specificity [a-d]. A model with a perfect discrimination between occurrence and absence of disease or infection has a ROC curve that passes through the upper left corner of the graph (100% sensitivity and 100% specificity). The closer the ROC curve is to the upper left corner of the graph, the higher the overall predictive accuracy of the model.

A useful index describing the overall accuracy of models that is independent of a single probability cut-off point is the area under the curve (AUC)[b]. For example, an AUC of 0.89 indicates that, for 89% of the time, a randomly selected location with disease or infection has a probability that is greater than that for a randomly selected location without disease or infection. An AUC between 0.5 and 0.7 indicates a poor discriminative capacity in distinguishing between different populations, whereas an AUC from 0.7–0.9 indicates a reasonable capacity and >0.9 indicate a very good capacity.



TRENDS in Parasitology

Fig. I Receiver–operator curves for different schistosomiasis ecological models: (a) Cameroon model applied to validation data in Cameroon (n = 170); (b) Tanga Region model applied to validation data in Tanga Region (n = 290); (c) Cameroon model applied to Tanzania (n = 980); (d) Tanga Region model applied to Kilosa District (n = 164); (e) Tanga Region model applied to Mtwara Region (n = 176), and (f) Tanga Region model applied to Magu District (n = 49). Dots represent pairs of sensitivity and false positive values, and the solid line represents the values expected by chance alone for each probability threshold.

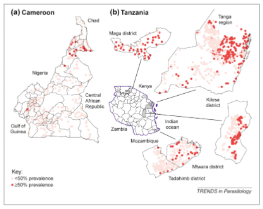
References

a. Metz CE. Basic principles of ROC analysis. Semin. Nucl. Med. 1978; 8:283–298. [PubMed: 112681]

b. Fielding AH, Bell JF. A review of methods for the assessment of prediction errors in conservation presence/absence models. Environ. Conserv. 1997; 24:38–49.

c. Pearce J, Ferrier S. Evaluating the predictive performance of habitat models developed using logistic regression. Ecol. Model. 2000; 133:225–245.

d. Greiner M, et al. Principles and practical application of the receiver-operating characteristic analysis for diagnostic tests. Prev. Vet. Med. 2000; 45:23–41. [PubMed: 10802332]



TRENDS in Parasitology

Fig. 1.

Spatial distribution of urinary schistosomiasis in Cameroon (a) and Tanzania (b). The infection prevalence was assessed by microscopy of urine in Cameroon, and the data was available for 19 524 children from 333 schools. In Tanzania, the infection prevalence was estimated from carefully validated questionnaire surveys (schoolchildren were asked whether they have urinary schistosomiasis or blood in their urine, termed locally as kichocho). The data for Tanzania are available for 166 099 children from 1960 schools. Although the prevalence of kichocho in schools underestimates the parasitological prevalence of infection, the parasite prevalence can be calibrated for each locality to define the extrapolated risk of having infection prevalence (>50%), and thus comparable to the parasitological data from Cameroon (see Ref. [17] for calibration and data sources).

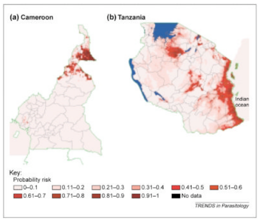


Fig. 2.

Prediction models for *Schistosoma haematobium* transmission in Cameroon (a) and Tanzania (b). The map shows the probability risk (0-1) of a particular area having an infection prevalence that exceeds the >50% threshold. Reproduced with permission, from Refs [17,18].