Pre-analysis plan: Predicting persistent hotspots of schistosomiasis transmission to guide preventive chemotherapy programs

Benjamin J Singer and Nathan C Lo
Division of HIV, Infectious Diseases, and Global Medicine
University of California, San Francisco

Date: 2022.05.12

Project background

Schistosomiasis is a parasitic disease that infects an estimated 250 million people globally, mostly in low- and middle-income countries. This disease often affects disadvantaged communities in endemic settings leading to inequities in health, and causing a range of symptoms including anemia, malnutrition, gastrointestinal symptoms, chronic abdominal pain, bladder cancer, periportal fibrosis and portal hypertension, pulmonary hypertension, and death. The WHO-recommended public health strategy to control and eliminate schistosomiasis is preventive chemotherapy via mass drug administration, which applies mass empiric treatment with curative praziquantel to populations at-risk for schistosomiasis. The decision to apply mass drug administration in a geographic region depends on the estimated local prevalence of *Schistosoma* spp. infection.

In February 2022, WHO published the first new schistosomiasis guidelines in 10-15 years. These changed the approach to mass drug administration by expanding treatment from only school-aged children to entire communities, lowering the prevalence threshold for treatment, and focusing on more frequent treatment in “persistent hotspots”. These hotspots are a key challenge in the elimination of schistosomiasis, and are defined as high transmission environments where prevalence does not decrease despite mass drug administration. This proposed project focuses on the key scientific gap of how to identify persistent hotspots of schistosomiasis transmission to better target more frequent mass drug administration and improve equity. Current approaches to identify hotspots require waiting 3-5 years after mass drug administration prior to defining a community as a hotspot. Various studies have defined and analysed persistent hotspots of schistosomiasis (Kittur et al. 2019, Kittur et al. 2020, Shen et al. 2019). Our hypothesis is that predictive modeling can identify hotspots at baseline prior to initiation of mass drug administration, allowing more frequent treatment earlier in these high-risk settings. Furthermore, rather than focus on relative changes in prevalence and/or infection intensity over time, we propose to predict whether or not a community will reach the WHO goal of <10% prevalence or <1% moderate-heavy infection prevalence. We will investigate whether a continuous statistical model can more accurately predict persistent hotspots than a categorical model, using baseline epidemiologic data. Hotspot prediction is a key scientific tool needed for implementation of the new WHO guidelines on mass drug administration.

Hypothesis

Statistical models using baseline epidemiologic data can predict whether a community is a persistent hotspot of *Schistosoma* transmission five years after the start of mass drug administration with a classification accuracy of greater than 80%.

Data

We will use data from the Schistosomiasis Consortium for Operational Research and Evaluation (SCORE) randomized trial for mass drug administration against schistosomiasis. We will use available data on *S. mansoni* and *S. haematobium*. For *S. mansoni*, the SCORE project includes a 5-year randomized trial which enrolled approximately 233,000 participants from three countries (Côte d'Ivoire, Kenya, Tanzania). For *S. haemotobium*, the SCORE project includes two distinct 5-year randomized trials in Niger and Mozambique which enrolled a total of approximately 315,000 participants. Each trial evaluated six distinct mass drug administration strategies for schistosomiasis and measured parasitological outcomes over time. We have obtained the de-identified, person-level longitudinal data from these trials. The UCSF IRB has deemed this study exempt given use of secondary de-identified datasets. We may also gain access to schistosomiasis data from the Togo Ministry of Health's longitudinal schistosomiasis survey and other datasets, which we will use to test the generalisability of our predictive model.

The analysis will include only communities in which school-aged children received at least two years of mass empiric treatment with praziquantel, with a coverage of 75% or higher. This inclusion criterion is based on the WHO's definition of persistent hotspots of schistosomiasis, which includes only communities that have been treated with high coverage mass drug administration. The number of villages included from each SCORE country is displayed in Table 1.

Table 1: Villages included

Villages	Niger	Mozambique	Côte d'Ivoire	Kenya	Tanzania
Percent of communities included	93%	16%	52%	88%	84%
Number of included communities	210	21	39	197	125

Study outcomes

The aim of this study is to develop a statistical model that can predict, from baseline infection data, which villages will be categorized as persistent hotspots of *Schistosoma* transmission in year 5 of the SCORE study. We will consider three definitions of persistent hotspot, based on three different target outcomes set by SCORE and the WHO for schistosomiasis control. We will train and test independent models for each outcome.

Table 2: Definitions of study outcomes of a *Schistosoma* persistent hotspot

Number	Outcome	Definition	Persistent hotspot threshold
1	Year 5 prevalence of *Schistosoma* infection	Proportion of persons in a community who have detectable *Schistosoma* eggs in year 5	>10%
2	Year 5 prevalence of *Schistosoma*	Proportion of persons in a community who have greater than 100 *Schistosoma*	>1%

	moderate/heavy infection	*mansoni* eggs per gram of stool or greater than 50 *Schistosoma haematobium* eggs per 10mL of urine	
3	Year 5 relative reduction in *Schistosoma* infection prevalence	Relative reduction in *Schistosoma* infection from year 1 to year 5	Relative prevalence reduction <35%

Variables

We have generated a list of model variables that may be predictive of whether a community is a persistent hotspot of *Schistosoma* transmission. We chose these variables based on prior literature, expert judgement, and available data. The model variables are estimated and assigned at the level of a community. The final model variables will depend on the model selection process. The list of variables defined using the SCORE data set is given in Table 3. The list of variables defined using other data sets is given in Table 4.

Table 3: List of model variables to predict whether a community is a persistent hotspot of *Schistosoma* transmission, sourced from SCORE data

Variable name	Definition	Comment
Infection prevalence	Proportion of tested individuals who have detectable *Schistosoma* eggs	
Infection density	Infection prevalence multiplied by population density, i.e. the number of infected persons per unit land area	Population density is sourced separately from SCORE data – see Table 4
Mean infection intensity	Arithmetic mean of *Schistosoma* eggs per gram of stool / per 10mL urine in persons with detectable eggs (i.e. only those infected)	
Dispersion in infection intensity	Dispersion parameter of the maximum likelihood negative binomial distribution fit to intensity (eggs per gram of stool / per 10mL urine) in persons with detectable eggs	This is the best variability measure given the structure of the data
Prevalence, 5-8 years	Proportion of tested individuals between the ages of 5 and 8 years with detectable *Schistosoma* eggs	
Prevalence, 9-12 years	Proportion of tested individuals between the ages of 9 and 12 years with detectable *Schistosoma* eggs	
Mean infection intensity, 5-8 years	Arithmetic mean of *Schistosoma* eggs per gram of stool / per 10mL urine in persons with detectable eggs (i.e. only those infected), ages 5 - 8 years only	
Mean infection intensity, 9-12 years	Arithmetic mean of *Schistosoma* eggs per gram of stool / per 10mL urine in persons with detectable eggs (i.e. only those infected), ages 9 - 12 years only	

Infection prevalence x arithmetic mean infection intensity	Product of the prevalence and the arithmetic mean infection intensity	
Infection prevalence x geometric mean infection intensity	Product of the prevalence and the geometric mean infection intensity	We will use both interaction with arithmetic and geometric infection intensity
Number of MDA rounds in schools	Number of years in which praziquantel was distributed to school-aged children	This variable does not include a coverage requirement
Number of MDA rounds in community	Number of years in which praziquantel was distributed to the general community, not only school-aged children	This variable does not include a coverage requirement

For secondary variables not included in the SCORE data, in some cases, variables may have strong within-country correlation. Previous work (Shen et al. 2019) has shown that a country dummy variable has high importance in predictive hotspot modelling, although our proposed baseline model does not include this in a goal of being more generalizable. There is a risk that an environmental variable that differs greatly between countries could be treated by a statistical model as a country dummy variable. This would effectively reduce the regression problem on this variable to a two- or three-data point regression (for *S. haematobium* and *S. mansoni*, respectively, dependent on the number of countries in the data set). Predictions based on such a regression may not generalize well to other countries, or even other regions of the same country where the variable considered is atypical for that country. We therefore exclude secondary variables that risk becoming country dummy variables based on review of the descriptive data. We will also subject our final model to checks to make sure that no variables are acting as country dummy variables, and we will remove them and re-fit the model if we find evidence of this.

Unless otherwise specified, we take the mean of all time-dependent variables in table 3 over the year 2011, since this is the year in which the SCORE baseline data was gathered.

Table 4: List of model variables from secondary datasets to predict whether a community is a persistent hotspot of *Schistosoma* transmission

Variable name	**Data set, variable**	**Resolution**	**Variable type**	**Comment**
Vegetation	MODIS, NDVI	250m	Continuous	*S. mansoni* only
Proximity to freshwater	CGLS, Water Bodies	60m	Continuous	2020 data used
Annual precipitation	WorldClim, precipitation	2.5 arc-minutes (~5km)	Continuous	*S. mansoni* only
Minimum temperature	WorldClim, temperature	30 arc-seconds (~1km)	Continuous	Minimum of the monthly averages of daily minimum temperature

Population density	GPW, UN-Adjusted Population Density	30 arc-seconds (~1km)	Continuous	

For highly skewed variables, we will use the logarithm of the raw value. Such variables include:

- Mean infection intensity (in all age groups)
- Dispersion of infection intensity (in all age groups)
- Prevalence multiplied by mean infection intensity
- Proximity to freshwater
- Population density

Validation

We will perform the analysis under three independent validation frameworks. The three frameworks differ based on how to use data from different countries for training and testing of the model.

Table 5: Frameworks for model validation

Number	**Validation description**	**Description**	**Data split**
1	Within country	The model is trained and tested independently for each country with more than 100 included villages (Kenya and Tanzania for *S. mansoni*, Niger for *S. haematobium*). Reported model performance is averaged across all country models, weighted by sample size.	70% train 30% test
2	Combined	The model is trained and tested using data from all countries (one model for each species).	70% train 30% test
3	Between country	The model is tested on all the data from one country, and trained on the one or two other countries. This process is repeated for all combinations of country assignments in which the training set is larger than 100 villages (three two-country combinations for *S. mansoni*, training on Niger only for *S. haematobium*) and the model performance is averaged across all scenarios.	N/A

If we gain obtain access to longitudinal data from Togo’s program on schistosomiasis (Bronzan et al. 2018) this will be used as an external test of model performance.

We will measure model performance with the primary outcome of classification accuracy, defined as the proportion of study villages with a correct classification as a persistent hotspot. We will also report the sensitivity and specificity of our models.

Statistical modelling

We will investigate the use of multiple statistical modelling approaches. For each row in table 2, we will formulate both a regression model to predict the quantity in the ‘Outcome’ column, and a categorization model to predict the binary outcome for each village according to the ‘Persistent hotspot threshold’ column.

For regression models of outcomes as continuous variables, we will use:

- Linear regression with forward stepwise variable selection
- Elastic-net linear regression
- Random forests
- Boosted decision trees

For categorisation models of outcomes as binary variables, we will use:

- Logistic regression with forward stepwise variable selection
- Elastic-net logistic regression
- Random forests
- Boosted decision trees

For the within-country and combined validation schemes (see table 5), we will perform five-fold cross-validation using the training data to set the free hyperparameters of these models. We will test each model on a separate validation set. For the between-country validation scheme we will perform group-wise cross-validation using the two countries in the training set for *S. mansoni*, and five-fold cross-validation in the single-country training set for *S. haematobium*. We will test each model using a separate country as the validation set.

In both the continuous and binary cases, we will also investigate the use of ensemble modelling, which combines the top performing models as an arithmetic mean. We will select the number of models to combine based on their incremental improvement in classification accuracy with a minimum requirement of 2% increase in classification accuracy.

Contingency

If the models described above do not reach our target of 80% classification accuracy, we will consider additional methodologies including the following:

- Mechanistic models of *Schistosoma* transmission
- More advanced statistical learning methodologies, including neural networks and Bayesian models (e.g., Bayesian linear regression, Bayesian additive regression trees)
- Including data from later years in the SCORE dataset, i.e. not only baseline data.
- Including county level variables and additional secondary dataset variables

These can be combined in an ensemble with the models above.

We will compare our final models with a null model and a model including a categorical country variable, to investigate whether any variables are being treated as country variables. If a single variable is absorbing a large amount of the variance between countries, we will remove it from the analysis.

Contact:

Benjamin John Singer

Digitally signed by Benjamin John Singer
Date: 2022.05.13
10:40:49 -07'00'

Benjamin Singer, DPhil
Email: Benjamin.singer@ucsf.edu

Nathan C Lo

Digitally signed by Nathan C Lo
DN: cn=Nathan C Lo, o, ou=UCSF,
email=nathan.lo@ucsf.edu, c=US
Date: 2022.05.12 21:53:07 -07'00'

Nathan Lo, MD PhD
Email: Nathan.lo@ucsf.edu

Division of HIV, Infectious Diseases, and Global Medicine
University of California, San Francisco

Citations

Kittur N, King CH, Campbell CH, Kinung'hi S, Mwinzi PNM, Karanja DMS, N'Goran EK, Phillips AE, Gazzinelli-Guimaraes PH, Olsen A, Magnussen P, Secor WE, Montgomery SP, Utzinger J, Walker JW, Binder S, Colley DG. Persistent Hotspots in Schistosomiasis Consortium for Operational Research and Evaluation Studies for Gaining and Sustaining Control of Schistosomiasis after Four Years of Mass Drug Administration of Praziquantel. *Am J Trop Med Hyg*. 2019 Sep;101(3):617-627. doi: 10.4269/ajtmh.19-0193. PMID: 31287046; PMCID: PMC6726953.

Kittur N, Campbell CH, Binder S, Shen Y, Wiegand RE, Mwanga JR, Kinung'hi SM, Musuva RM, Odiere MR, Matendechero SH, Knopp S, Colley DG. Discovering, Defining, and Summarizing Persistent Hotspots in SCORE Studies. *Am J Trop Med Hyg*. 2020 Jul;103(1_Suppl):24-29. doi: 10.4269/ajtmh.19-0815. PMID: 32400365; PMCID: PMC7351310.

Shen Y, Sung MH, King CH, Binder S, Kittur N, Whalen CC, Colley DG, Modeling Approaches to Predicting Persistent Hotspots in SCORE Studies for Gaining Control of Schistosomiasis Mansoni in Kenya and Tanzania, *The Journal of Infectious Diseases*, 2020 Mar;221(5):786-803 doi: 10.1093/infdis/jiz529.

Bronzan RN, Dorkenoo AM, Agbo YM, Halatoko W, Layibo Y, Adjeloh P, Teko M, Sossou E, Yakpa K, Tchalim M, Datagni G, Seim A, Sognikin KS. Impact of community-based integrated mass drug administration on schistosomiasis and soil-transmitted helminth prevalence in Togo. *PLoS Negl Trop Dis*. 2018 Aug 20;12(8):e0006551. doi: 10.1371/journal.pntd.0006551. PMID: 30125274; PMCID: PMC6124778.