Open Forum Infectious Diseases

MAJOR ARTICLE

Improved Assessment of *Schistosoma* Community Infection Through Data Resampling Method

David Gurarie,[1,2] Anirban Mondal,[1] and Martial L. Ndeffo-Mbah[3,4]

[1]Department of Mathematics, Applied Mathematics, and Statistics, Case Western Reserve University, Cleveland, Ohio, USA, [2]Center for Global Health and Diseases, School of Medicine, Case Western Reserve University, Cleveland, Ohio, USA, [3]Department of Veterinary and Integrative Biosciences, School of Veterinary Medicine and Biomedical Sciences, Texas A&M University, College Station, Texas, USA, and [4]Department of Epidemiology and Biostatistics, School of Public Health, Texas A&M University, College Station, Texas, USA

Background. The conventional diagnostic for *Schistosoma mansoni* infection is stool microscopy with the Kato-Katz (KK) technique to detect eggs. Its outcomes are highly variable on a day-to-day basis and may lead to biased estimates of community infection used to inform public health programs. Our goal is to develop a resampling method that leverages data from a large-scale randomized trial to accurately predict community infection.

Methods. We developed a resampling method that provides unbiased community estimates of prevalence, intensity and other statistics for *S mansoni* infection when a community survey is conducted using KK stool microscopy with a single sample per host. It leverages a large-scale data set, collected in the Schistosomiasis Consortium for Operational Research and Evaluation (SCORE) project, and allows linking single—stool specimen community screening to its putative multiday "true statistics."

Results. SCORE data analysis reveals the limited sensitivity of KK stool microscopy and systematic bias of single-day community testing versus multiday testing; for prevalence estimate, it can fall up to 50% below the true value. The proposed SCORE cluster method reduces systematic bias and brings the estimated prevalence values within 5%–10% of the true value. This holds for a broad swath of transmission settings, including SCORE communities, and other data sets.

Conclusions. Our SCORE cluster method can markedly improve the *S mansoni* prevalence estimate in settings using stool microscopy.

Keywords. infection intensity; prevalence; resampling; *Schistosoma*.

Schistosomiasis is one of the most prevalent neglected tropical diseases, with >250 million people affected worldwide [1] and the collective burden estimated at 3.3 million disability-adjusted life-years lost [2]. Intensified control efforts over the past 15 years have focused on preventive chemotherapy via mass drug administration (MDA) with praziquantel [2, 3]. Accurate assessment of community infection, along with demographic risk factors and spatial-temporal environmental patterns, are expected to provide essential inputs to guide control interventions.

Commonly used tools for community assessment include Kato-Katz (KK) egg count diagnostics for *Schistosoma mansoni* and urine filtration for *Schistosoma haemotobium* [4]. Both are notoriously uncertain, with high day-to-day variability of egg counts for individual hosts [1–3, 5]. Furthermore, a sizable fraction of repeated tests, combine "zero" and "positive" counts for the same individual. So, a single test could qualify a host as positive or "negative," resulting in highly uncertain and variable outcomes on the community level. Having a multisample host screening, like the Schistosomiasis Consortium for Operational Research and Evaluation (SCORE) data set, can achieve far higher accuracy [6], but such tests could be prohibitively expensive. Realistic control-surveillance programs often rely on single-test/host screening. Such screening can greatly underestimate community infection.

There were several proposals to address the diagnostic variability and uncertainty of KK tests via statistical modeling [3, 4, 6, 7] and mathematical models [8–10]. These models make certain assumptions about putative worm burden distribution in a host community, the ensuing egg release, and KK screening, all expressed through parametric distributions, including negative binomial distributions. By fitting a model to data, one can infer unknown parameters and outputs (eg, prevalence and intensity) from model analysis.

Here we propose an alternative empirical approach that uses resampling of multiday SCORE community tests and statistical inferences drawn from data analysis. Each community

Received 29 August 2023; editorial decision 14 December 2023; accepted 18 December 2023; published online 21 December 2023

Correspondence: Martial L. Ndeffo-Mbah, PhD, Department of Veterinary and Integrative Biosciences, School of Veterinary Medicine and Biomedical Sciences, Texas A&M University, 660 Raymond Stotzer Parkway, College Station, TX 77843-4458 (m.ndeffo@tamu.edu); David Gurarie, PhD, Department of Mathematics, Applied Mathematics, and Statistics, Case Western Reserve University, 2145 Adelbert Rd., 2083 Martin Luther King Jr Dr, Cleveland, OH 44106-7058 (dxg5@case.edu).

Open Forum Infectious Diseases®

https://doi.org/10.1093/ofid/ofad659

resample consists of random samples from multiple host tests (most individuals in the SCORE study were tested on multiple days, up to 3). Such resampling can be viewed as a putative community snapshot, to be observed in a typical control-surveillance setting with a single test/host. Community statistics (prevalence, intensity) obtained from such resamples exhibit high variability and systematic bias.

Our goal is to apply the SCORE data set to arbitrary single-slide community test data (SCORE or non-SCORE) and estimate its unknown true prevalence-intensity statistics along with their uncertainties (error margins). The analysis of resampled data revealed that single-day KK diagnostics consistently underestimates the multiday "true" prevalence and has a wide range of variability (error margins). We observed this pattern across a broad swath of SCORE communities.

The proposed method and computing tools allow recovery of true community statistics and reduce prediction uncertainty by identifying a suitable cluster of SCORE community tests for given raw data. To validate our method, we applied it to a wide range of single-slide resamples obtained from SCORE and non-SCORE data sets for which multiday testing "truth" was available. It showed that our scheme could produce robust, statistically reliable predictions with reduced error margins.

METHODS

SCORE Data Set

The SCORE project is a large-scale control-surveillance study conducted in several areas endemic for *S mansoni* over the 5-year period [11–14]. The data set contains 450 endemic communities from Kenya, Cote d'Ivoire, and Tanzania. Communities were divided into different control arms with annual screening followed by MDA. On each program cycle a selected host pool (primarily school age) drawn from the local community was given a KK diagnostic test. Some hosts had a single test (day 1), and others had 2 (days 1 and 2) or 3 (not necessarily on consecutive days). The combined diagnostic data set comprises 233 102 individual host tests (single, double, and triple), partitioned into 1744 community pools.

A single (daily) host test had two 42-mg-thick smear slides whose egg counts were reported as (A1, B1) on day 1, (A2, B2) on day 2, and so forth. These counts differ from the standard "eggs per gram" (EPG) by a factor of 12. We combined 2 slides into a single daily count (A + B). The reason for combining (A, B) slides into a single daily score is the high variability of KK smear counts. Eggs are presumably released by worms at a steady pace, but they are unevenly distributed in any specimen. A single random smear gives a distorted view of such uneven distribution, and by combining 2 slides we obtain a more reliable estimate of daily egg release. Multiday SCORE tests arranged into community pools constitute the core ingredient for our modeling and analysis.

Table 1. Schistosomiasis Consortium for Operational Research and Evaluation (SCORE) Test Statistics

	Tests, No.			
Age Group	Single	Double	Triple	Total
CHILDREN	52 824	16 066	141 138	210 028
ADULTS	22 655	140	279	23 074
ALL AGES	75 479	16 206	141 417	233 102

Table 2. Sample Schistosomiasis Consortium for Operational Research and Evaluation (SCORE) Community Test (Triplets) and "True" Means[a]

Triplet	Mean	Triplet	Mean	Triplet	Mean
{0, 0, 0}	0.	{0, 0, 0}	0.	{0, 0, 0}	0.
{0, 0, 0}	0.	{0, 0, 0}	0.	{0, 0, 0}	0.
{0, 0, 0}	0.	{0, 0, 0}	0.	{0, 0, 0}	0.
{0, 0, 0}	0.	{0, 0, 0}	0.	{0, 0, 0}	0.
{0, 0, 0}	0.	{0, 0, 0}	0.	{0, 0, 0}	0.
{0, 0, 0}	0.	{0, 0, 0}	0.	{0, 0, 0}	0.
{0, 1, 0}	0.33	{0, 2, 0}	0.67	{5, 0, 0}	1.67
{0, 9, 0}	3.	{0, 9, 0}	3.	{0, 9, 0}	3.
{0, 11, 0}	3.67	{0, 12, 0}	4.	{0, 0, 12}	4.
{10, 3, 0}	4.33	{0, 20, 0}	6.67	{27, 1, 2}	10.
{24, 40, 0}	21.33	{84, 0, 0}	28.	{84, 0, 0}	28.
{84, 30, 4}	39.33	{84, 42, 39}	55.	{68, 31, 84}	61.

[a]The community prevalence-intensity values ("truth") are (0.5, 1.86).

The bulk of SCORE subjects were school-aged children, considered a sentinel pool for community assessment; adults were added in some cases, but their contribution was marginal. The aggregate SCORE data statistics are given in Table 1. We therefore dropped adults from our community pools and the model. This study did not include factors necessitating patient consent or ethical approval.

A distinctive feature of SCORE data is the high variability of daily counts for individuals. Furthermore, a sizable fraction of multiple tests combine zero and positive counts (Table 2), which confounds community assessment. Our method aims to rectify such deficiencies, via systematic resampling and comparison to truth. Multiday SCORE data have many potential applications, among them estimates of the overall KK test sensitivity based on the aggregate data (Supplement A; Supplementary Tables 1 and 2). Our primary goal, however, is community-level assessment, where general sensitivity estimates have little practical use.

Prevalence and Geometric Mean Intensity for Community Assessment

We want to infer true prevalence-intensity statistics from a collection of multiday community tests. The conventional approach consists of replacing each multiple count by its arithmetic mean:

$$(e_1, e_2, e_3) \to \bar{e} = \frac{e_1 + e_2 + e_3}{3} \qquad (1)$$

(see, eg, [4, 14–17]). The resulting community statistics, based on mean scores, will be considered true values. Table 2 illustrates the procedure for a typical SCORE community.

Our modeling method makes extensive use of 2 community statistics: conventional prevalence P (fraction of positive counts), and infection intensity, measured by the geometric mean G of positive counts. So a test data of size n is split into zero and positive EPG counts:

$$\{\underbrace{0, \ldots, 0}_{n-m} \mid \underbrace{e_1, e_2, \ldots}_{m}\},$$

and P and G are defined as $P = \frac{m}{n}$ and $G = \left(\prod_1^m e_k\right)^{1/m}$; in most applications below, G is replaced by $\log(G)$.

The choice of G for intensity can be justified on mathematical and empirical grounds (see Supplement B; Figure 2). Two statistics, P and G, are largely independent over a wide range of values for P, as demonstrated by SCORE data analysis below.

Our goal is to explore the variability of conventional community screening (single test/host) and their P-G statistics, in relation to putative (multiday) truth. We do it via extensive analysis of the SCORE data set. A typical SCORE community test would combine singlet, doublet, and triplet counts (1-, 2-, or 3-day testing/host), whose mean counts (true values) have different statistical significance; a "triplet mean" should carry higher significance than a singlet. One way to account for significance (singlet vs doublet vs triplet) is to assign them different weights. Another approach, adopted here, is to confine analysis to triplets alone (dropping the rest). Indeed, triplets comprise the bulk of SCORE tests, about two-thirds of all tests (Table 1). Furthermore, they dominate test data among well-sampled communities (pool size $50 \le m \le 120$), as illustrated in Figure 1. Such pruning of community test data to triplets alone may distort their true P-G values, but our goal is not accurate assessment of realistic SCORE communities. The resulting collection of SCORE-like "triple-screened" communities will serve a basis (test bed) for the cluster selection method developed below.

Resampling Method for the SCORE Test Bed

The SCORE test bed used in our model is confined to young ages (1–15 years) and triplet test samples. Altogether, we obtained 1624 community tests, representing a broad range of infection levels, from near-zero prevalence ($P \approx 0.01$) to fully infected ($P \approx 0.99$).

We first examine the variability of single-slide KK diagnostics by generating random resamples (snapshots) of test-bed communities. Then we develop a prediction method by linking such resamples to truth.

Each community resample (a single-slide random snapshot) gives prevalence intensity values represented by a point $z_i = (P_i, G_i)$ in the P-G plane. An ensemble of resamples generates an uncertainty cloud, $C = \{z_i : i = 1, 2, \ldots\}$, with center

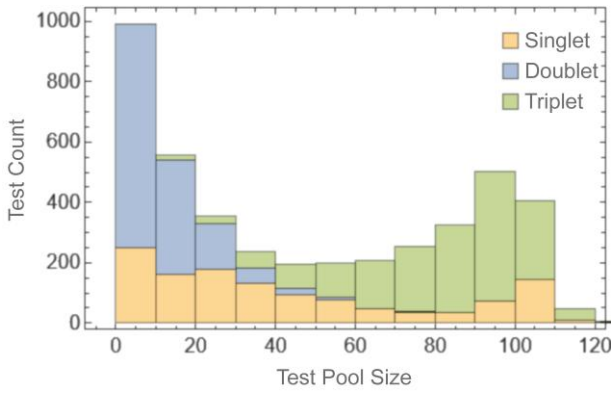

Figure 1. Test multiplicity distribution by test pool size. Triplets dominate well-sampled communities (pool size >50 individuals).

$z_0 = (P_0, G_0)$—expected (mean) PG values (Figure 2*B*). We approximate such a cloud by a normal distribution $D_0(z) = N(z_0, \Sigma_0)$ with mean (center) z_0 and covariance matrix Σ_0. These resample clouds and their distributions play a crucial role in our method.

SCORE Cluster Method

The main goal is for a given singlet community test (snapshot) $T = \{e_1, e_2, ..\}$ to reconstruct its putative multiday SCORE-like truth. Raw data T could come from a resampled SCORE community or from test data collected elsewhere. In either case, we aim to identify a cluster of SCORE-like communities $\{T_j\}$ from the test bed that are "similar" to T, in terms of PG test statistics. We call that a "local SCORE cluster." We start by computing PG values of raw data T, to get a reference point $z_0 = (P_0, G_0)$. We then ask for a SCORE test-bed cluster $\{T_j\}$, made of triplet tests that are likely of reproducing point z_0 via resampling.

The proposed procedure, called *SCORE cluster selection,* identifies such cluster pool. It uses a collection of normal distributions $\{D_j(z)\}$ made of 1624 triplet SCORE communities $\{T_i\}$—the test bed. Each test-bed distribution $D_i(z)$ is evaluated at the reference point z_0, and a few highest likelihood choices $\{w_j = D_j(z_0) : j = 1, \ldots, m\}$ (highest w_j) are selected as local SCORE cluster of z_0. Such SCORE-like communities are most likely to generate raw data T via random resampling, based on their PG statistics. Likelihood values $\{w_j\}$ can serve as weights of a virtual "SCORE-like composite community," consistent with raw data T. We shall consistently use such weighted SCORE clusters $\{T_i, w_i\}$ to infer the unknown true statistics of raw data T. For instance, each cluster $\{T_i, w_i\}$ has its true PG values $Z_i = (P_i, G_i)$, and then the estimated truth $Z_0 = (P_T, G_T)$ for T is given by weighted mean "cluster truth" $Z_0 = \left(\sum_i w_i Z_i\right) / \left(\sum_i w_i\right)$.

To avoid confusion, we stress that the local SCORE clusters used in our model are not linked physically (geospatially or

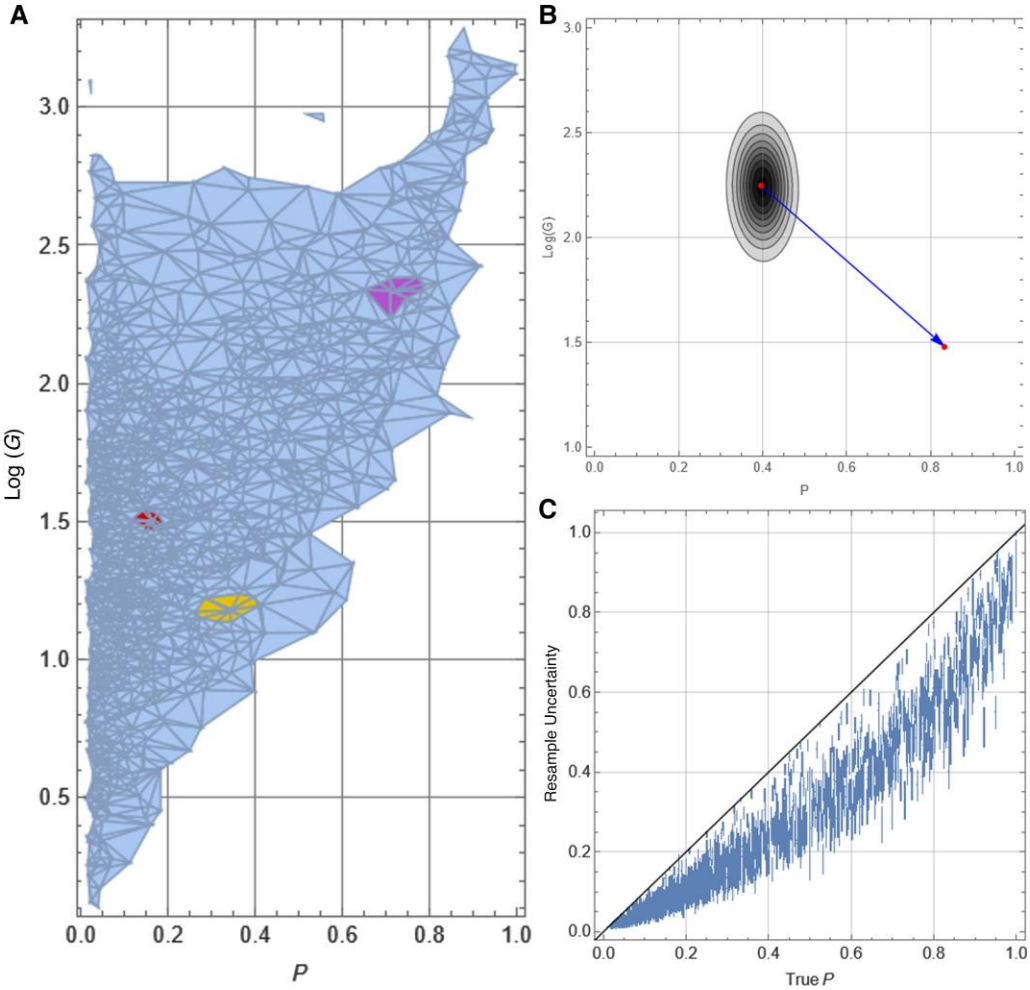

Figure 2. *A,* Resampled cloud centers of 1624 Schistosomiasis Consortium for Operational Research and Evaluation (SCORE) community tests in the *P-G* plane (*P* for prevalence and *G* for geometric mean), as nodes of the mesh region with 3 highlighted local SCORE clusters. *B, PG* cloud approximated by normal distribution $D_0(z)$ centered at $z_0 = (0.41, 2.22)$, with arrow linking z_0 to its multislide SCORE "truth" $z_T = (0.85, 1.48)$. *C,* Prevalence distributions of 1624 community tests versus their SCORE true values (P_T); bars represent means with standard deviations

otherwise) to real SCORE communities. Starting with raw data (*P, G*), we extract the local SCORE clusters by scanning the complete test bed and identifying a dozen clusters most likely to generate raw data (*P, G*) via random resampling.

RESULTS

KK Uncertainty and Sensitivity

Multiday SCORE tests exhibit wide day-to-day variability of egg counts for individual hosts (Supplementary Figure 1). Furthermore, a significant portion of those combine zero and positive counts (Supplementary Figure 1C). Therefore, the standard binary classification (positive-negative) based on a single KK test has reduced sensitivity (Supplementary Table 1). Overall, our bulk estimates of KK sensitivity are consistent with known results (eg, [4, 15]), but they are not particularly useful at the community level.

Community-Level Analysis

Next, we applied a raw resample method across the SCORE triple test bed. Figure 2*A* illustrates a distribution of 1624 cloud centers, shown as a linked mesh region in the *P-G* plane. Each resample ensemble was generated from 500 random snapshots from a triplet community test *T*. Cloud centers cover a broad swath in the *P-G* plane. It is reasonable to assume they all represent putative "single-slide" snapshots (SCORE or non-SCORE). Figure 2*B* shows a typical normal cloud distribution $D_0(z)$ around its marked center $z_0 = (P_0, G_0) = (0.41, 2.22)$. An arrow drawn from z_0 to the SCORE truth $z_T = (P_T, G_T)$, illustrates the discrepancy between snapshot and the true statistics. In this case, true $P_T \approx 0.85$ exceeds expected resample $P_0 \approx 0.41$, by a factor of 2. Figure 2*C* shows the distribution of resampled *P* values, displayed as means with standard deviations (vertical bars) plotted against true P_T (horizontal axis) across 1624

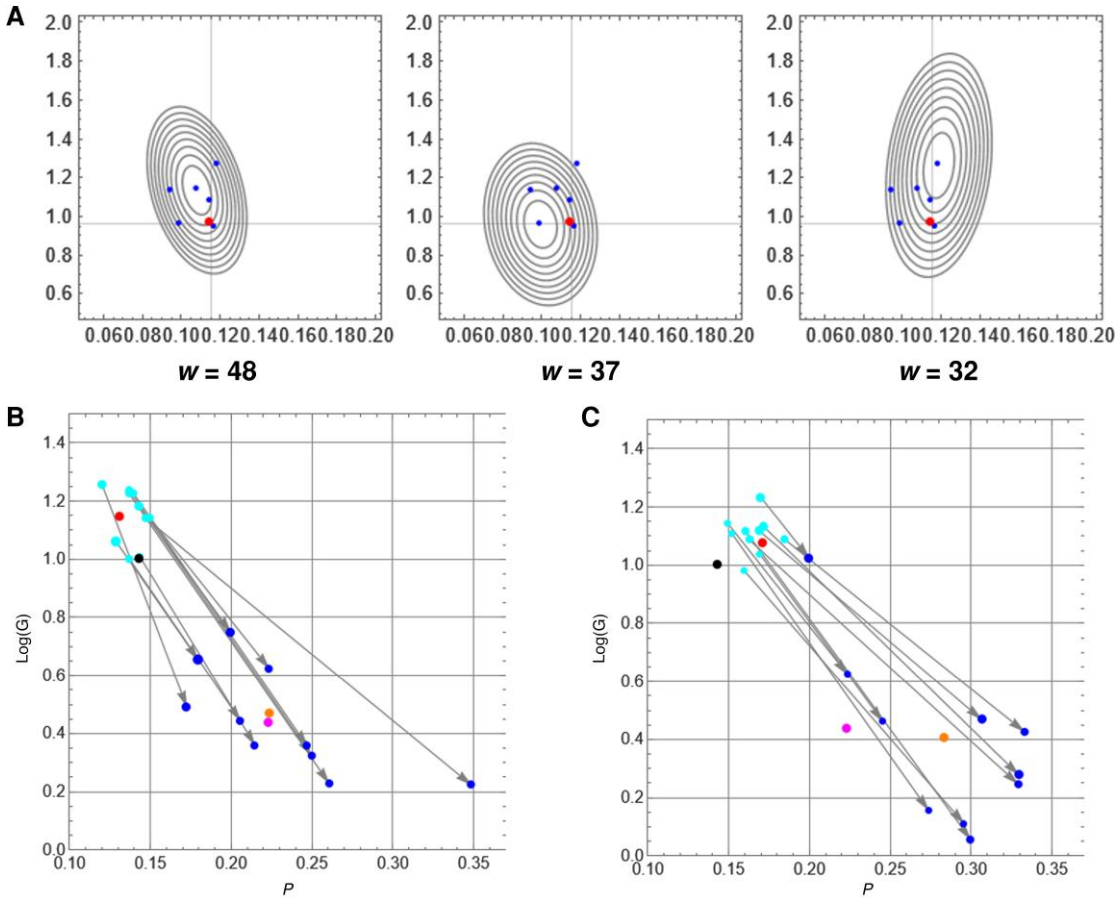

Figure 3. *A*, Local Schistosomiasis Consortium for Operational Research and Evaluation (SCORE) cluster $\{z_j\}$ (*blue centers*) of reference point z_0 (*red*), and 3 selected clouds with likelihood weights $\{w_j\}$. *B, C,* Cluster-based reconstruction of *PG* truth (*P* for prevalence and *G* for geometric mean) for a red reference z_0, drawn from a SCORE community test (*black cloud center*): local cluster (*cyan*), is shifted to SCORE "truth" (*blue*), and its weighted mean (*orange*) serves to estimate the true *PG* for the red reference. Estimated *PG* is compared with SCORE truth (*magenta*).

community tests. We observe a consistent downward bias of resampled *P* values, relative to truth, and wide dispersal within vertical (error) bars. Our goal is to restore the truth from "observation" using the SCORE cluster method.

Local SCORE Clusters and Synthetic Composite Communities

We use a test bed of 1624 SCORE clouds, centered at $z_j = (P_j, G_j)$, and their normal distributions $\{D_j(z)\}$. Given a single-slide raw data *T*, and its reference *PG* values $z_0 = (P_0, G_0)$, we select a local SCORE cluster of z_0, based on likelihood weights $\{w_j = D_j(z_0): j = 1, \ldots, m\}$, as explained above. In most applications, the 10 highest choices are used ($m = 10$). Figure 3*A* shows highlighted local clusters on the "SCORE center mesh" for 3 selected reference points. Figure 3*A* illustrates a local SCORE cluster of 6 for a red reference point z_0, with 3 "highest likelihood" cloud distributions D_j, along with their weights (w).

A local SCORE cluster of raw data *T* serves to generate a composite SCORE-like community, made of a triplet test bed $\{T_j: j = 1, \ldots, m\}$ contributing in proportion to its likelihood weights, $\{w_j\}$. Many relevant statistics can be extracted from such SCORE-like composites. For instance, a composite *PG* cloud of z_0 could be generated from weighted resampled cluster clouds, and its center z_C estimated via a weighted sum of the constituent center, $z_C = \sum_i z_i w_i$. Other *T* statistics could be inferred from its SCORE cluster composite, including true *PG* estimates, graded prevalence (eg, World Health Organization [WHO] light-moderate-heavy [LMH] prevalence), and so on.

True Prevalence -Intensity Reconstruction

Given a single-slide community test *T* with reference point $z_0 = (P_0, G_0)$, we select its local SCORE cluster and link each constituent cluster center z_i to its SCORE truth, Z_i, weighted via the rescaled likelihood value, $w_i \propto D_i(z_0)$, $\sum_i w_i = 1$. The proposed cluster estimate of true *PG* is given by a weighted mean:

$$Z_C = (P_C, G_C) = \sum\nolimits_i w_i Z_i \qquad (2)$$

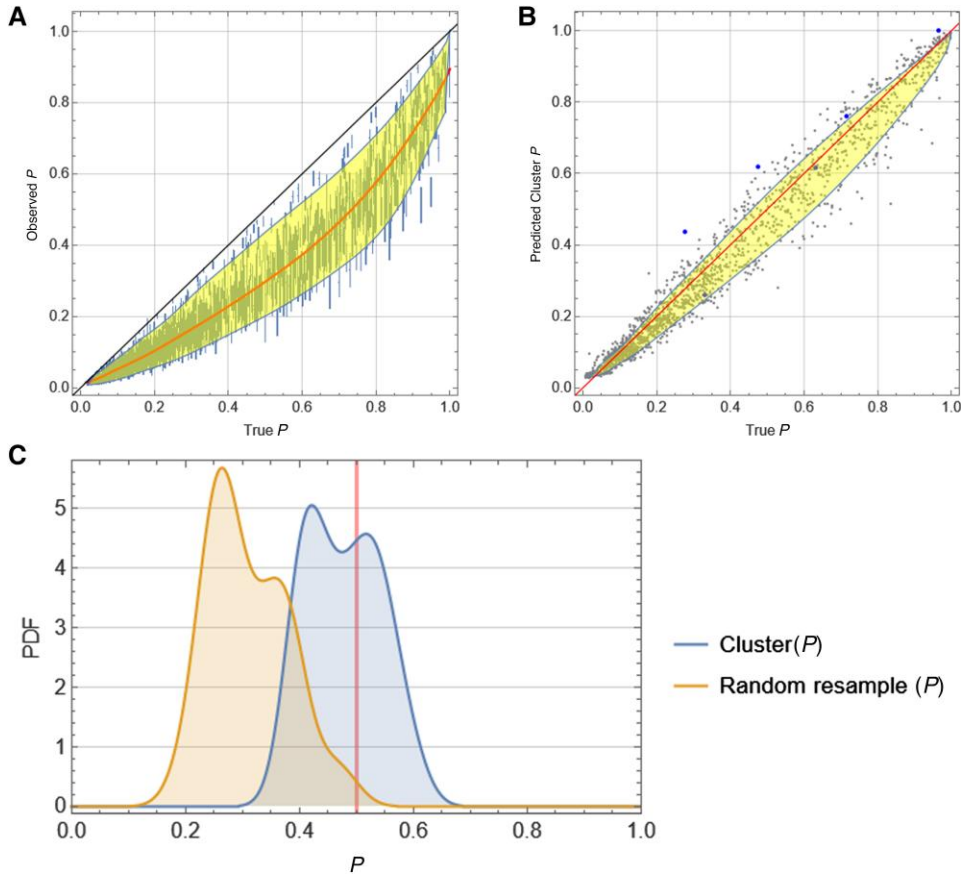

Figure 4. *A*, Raw resample errors bars of Figure 2*B* augmented with quantile marginal curves as functions of true P_T. *B*, Scatterplot of cluster selection scheme (P_T, P_C) for combined data set made of 1624 Schistosomiasis Consortium for Operational Research and Evaluation (SCORE) values (*gray dots*) and additional communities (*blue dots*), augmented with quantile marginal curves. *C*, Cross-sectional snapshot distributions of *A* and *B* at a fixed $P_T = .5$: raw resample (*yellow*) versus cluster (*blue*). Abbreviation: PDF, probability distribution function.

We illustrate the procedure for specific test data in Figure 3*B* and 3*C*. Two reference points z_0 (red) are extracted from resamples of 2 SCORE communities. In both cases, local SCORE cluster centers (cyan) are shifted toward their true *PG* values (blue), whose weighted mean (orange) gives cluster truth, Z_C. As both reference points z_0 come from SCORE resamples (cloud centers marked in black), we can compare cluster truth (orange) with the "SCORE truth" (magenta). In Figure 3*B*, 2 estimates come fairly close, so SCORE cluster gives a good approximation of the truth. In Figure 3*C*, 2 estimates are further apart (so cluster Z_C overestimates Z_T), which is partly owing to the position of the selected reference (red) relative to the SCORE center (black). We also note that the "red resample" (Figure 3*B*) has higher likelihood than (Figure 3*C*).

Model Validation

To validate our model, we applied it to SCORE communities and an additional collection of non-SCORE communities with multiple-test data. In each case, a single-test random snapshot (T) was drawn from multiday scores, and the clustering method was applied to the reference values z_0. Cluster outcomes, particularly estimated true P_C was compared with the source truth P_T. In both cases, SCORE and non-SCORE, the truth is available via multiday averaging. In general, we should not expect a perfect match between P_T and P_C, as shown in the previous section. So, our validation scheme aims to assess statistical robustness of cluster method across the entire community span.

We proceed by selecting a reference point z_0 for each sampled community and generating its local SCORE cluster $\{z_i : j = 1, \ldots, m\}$, along with the "source truth" $\{Z_j\}$ and likelihood weights $\{w_j\}$. The resulting cluster $Z_C = (P_C, G_C)$, Equation (2), is compared with the source truth $Z_T = (P_T, G_T)$. The scatterplot (P_T, P_C) is shown in Figure 4*B*; it includes 1624 SCORE test bed (gray points) and 7 non-SCORE communities (blue). Figure 4*B* can be contrasted to the "raw data resampling" of Figure 4*A*. The latter (raw resampling) exhibits a systematic downward bias (gross underestimation of truth) and high variability. The former (cluster selection) is closer to diagonal truth with narrower dispersal.

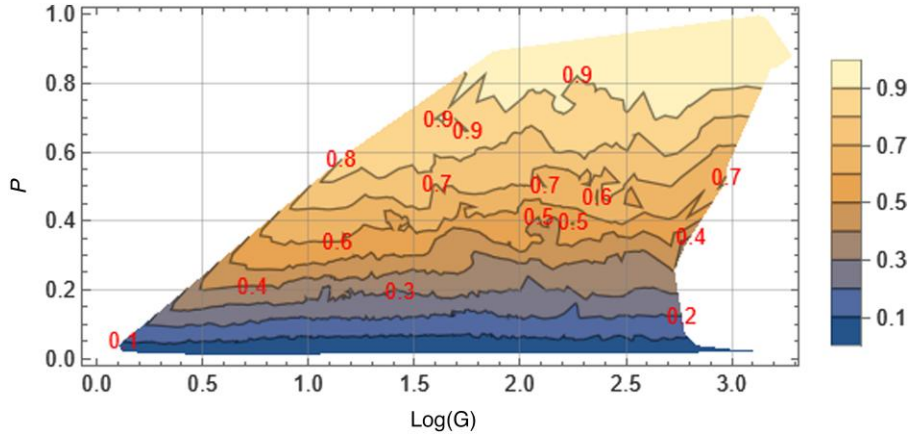

Figure 5. Schistosomiasis Consortium for Operational Research and Evaluation (SCORE) version of topography of the *P-G* plane described by de Vlas et al (in their "pocket chart" [7]); red isocontours mark fixed values of true $P_T = f(G, P)$, as a function of (*P*,*G*).

Figure 4*C* illustrates this gap by cross-sectional (vertical) slices of 2 distributions (Figure 4*A* and 4*B*) at a fixed value $P_T = 0.5$.

In both cases (Figure 4*A* and 4*B*), we approximated a discrete data point (mean + uncertainty) by smooth functions (blue curves) that mark 5%–95% quantile ranges of vertical scatters in different P_T prevalence bands. Those functions could serve as crude markers of predicted "true ranges" for single-slide community snapshot. However, wide marginal ranges, particularly Figure 4*A*, have little predictive value for prevalence assessment. To illustrate this point, we take a single-slide (observed) prevalence $P = 0.4$; Figure 4*A* would predict the true range ($0.48 < P_T < 0.78$)—well above observation with wide margins of uncertainty, while Figure 4*B* would narrow the range ($0.35 < P_T < 0.5$)—closer to the predicted cluster truth. Overall, the blue margin curves of (Figure 4*B* and 4*C*) rely on prevalence estimates alone and give widely uncertain predictions. The key advantage of cluster selection method is the greater accuracy of true prediction and reduced uncertainty, achieved via combined *P*-*G* statistics.

P-G Estimates of Truth

An ideal tool for program managers would be a simple "function" or "numeric code" that would take a raw data input—for example, observed (*P*, *G*) values—and predict true prevalence $P_T = f(P, G)$, within error margins. A version of such function (a "pocket chart") was proposed by de Vlas et al [7], using a statistical model of community EPG test that was based on negative binomial distributions. It assumed a hypothetical worm burden stratification of host communities and an egg release process by host strata, both described by suitable negative binomial distributions. In this model, true prevalence P_T corresponds to the "positive worm burden," while the estimated "EPG prevalence" (test observation), $P < P_T$, comprises "positive egg release" fractions of all infected strata. The main result is formulated in terms of function $P_T = f(P, G)$. Our SCORE cluster method yields an empirical version of a pocket chart (Figure 5). It resembles qualitatively the pocket chart of de Vlas et al [7], but the SCORE topography is more rugged. Our "SCORE cluster" method, however, goes beyond the topographic chart of Figure 5. It provides a simple and efficient computational tool (SCORE calculator), that can take any raw input and estimate its true *P*-*G* statistics within uncertainty margins.

Beyond Prevalence-Intensity

Prevalence serves as a key measure of community infection widely used in control programs (see, eg, WHO road map strategies [18]). Another important statistic proposed by the WHO is the graded prevalence based on EPG counts: $0 < \text{EPG} < 100$ (light), $100 < \text{EPG} < 400$ (moderate), and $400 < \text{EPG}$(heavy). We call them {*L*, *M*, *H*}, with the latter (*H*) serving as a proxy of schistosomiasis morbidity (heavy infections are often correlated with chronic conditions). Thus, WHO control strategies rely not only on the combined prevalence, $P = L + M + H$, but also include *H*, as a specific target.

The SCORE cluster method allows one to assess different community statistics, including LMH. As noted above, a single-day community test would give specific LMH values, but a multiday test would generate an LMH ensemble via resampling, alongside the "true LMH," based on average EPG scores Equation (1). As above, one should not expect resampled LMH to match the truth. Figure 6*A* illustrates LMH discrepancies for a specific community test. Here resampled *L*; and *M* underestimates the truth, while *H* exceeds it.

The cluster selection for LMH proceeds as above; starting with a (*P*,*G*) reference point z_0 (raw test data), we generate its local SCORE cluster $\{z_i, w_i{:}i = 1, \ldots, m\}$, along with the cluster's triplet truth—a collection of multiday tests $\{T_1; \ldots; T_m\}$. Each T_i has its true LMH (L_i, M_i, H_i), weighted by a *PG* likelihood value w_i of z_i. The resulting cluster estimate

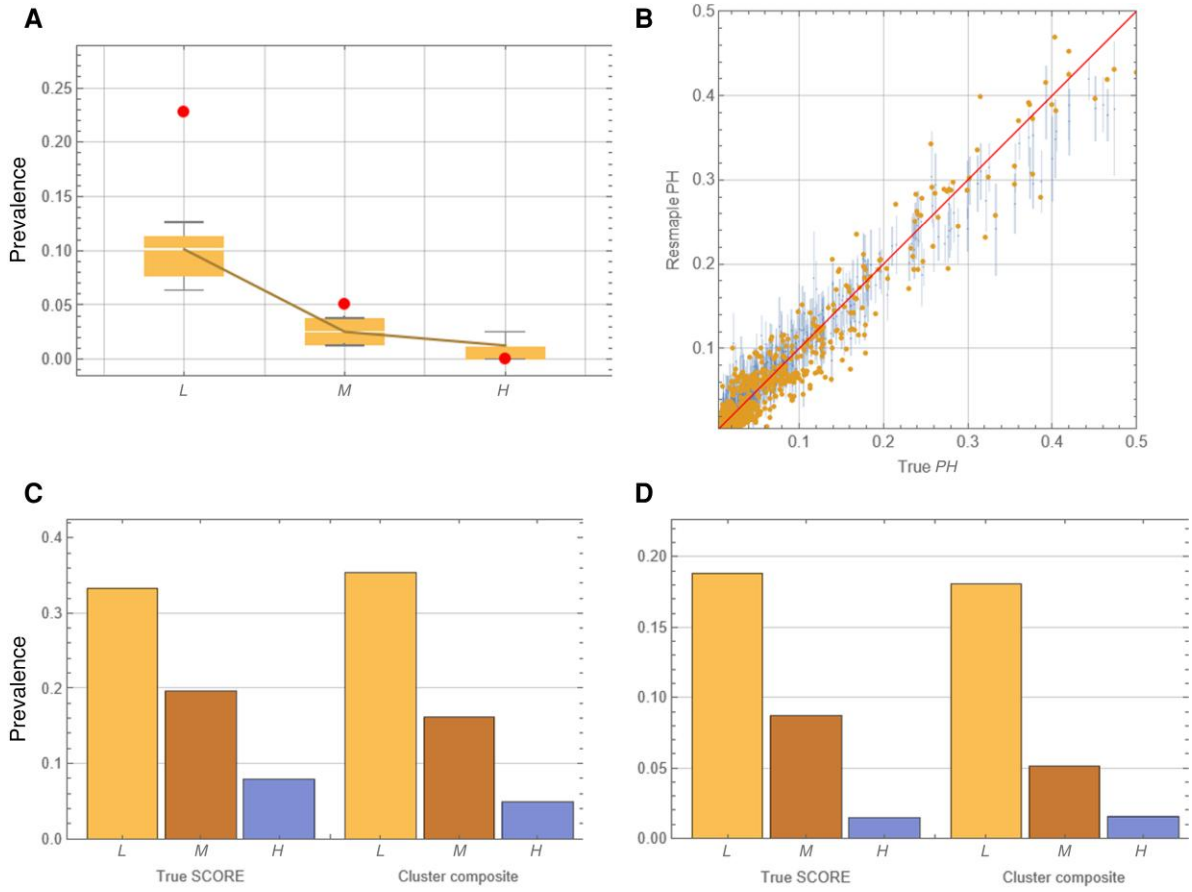

Figure 6. Estimates of graded light-moderate-heavy (LMH) prevalence. *A,* LMH ensemble distribution (*box-whiskers*) compared with true LMH (*red dots*) for a specific community test. *B,* Heavy prevalence *H* (*PH*) for 1620 community tests: bars correspond to raw resample estimates, and yellow dots are cluster-composite estimates of *H*. Both are plotted against the true Schistosomiasis Consortium for Operational Research and Evaluation (SCORE) *H*. *C, D,* LMH prevalence values (true vs cluster composite) for 2 selected community tests.

of LMH is given by the weighted sum

$$(L_C, M_C, H_C) = \sum_i w_i (L_i, M_i, H_i) \qquad (3)$$

similar to Equation (2). Indeed, each cluster test T_i contributes a w_i share of the composite SCORE ensemble.

To assess the validity of cluster estimate (Figure 3) across all SCORE communities, we compared predicted LMH_C with SCORE truth LMH_T, in particular, heavy prevalence H. The scatterplot of Figure 6*B* shows points (H_T, H_C)—yellow dots, with gray bars marking raw resample estimates of H. As above, we do not expect a perfect match, but overall cluster selection exhibits higher accuracy and lower variability (yellow dots vs gray shaded area).

DISCUSSION

Conventional approaches to schistosomiasis surveillance and control rely on single-slide KK diagnostics. KK diagnostics exhibit high variability on an individual (day-to-day) basis, and the resulting community assessment can be grossly underestimated relative to a putative multislide truth. To fill in the missing "observation-truth" gap, we propose to use an extensive data set of multislide community tests of the SCORE project.

Our analysis combined conventional measures of community infection for multiday tests (see, eg, [4, 6, 8–10]), with alternative statistics derived from data resampling. We consistently use prevalence-intensity statistics, the latter measured by the geometric mean (G) of positive counts. While prevalence P was the key target, we found that P alone was insufficient for accurate assessment, but combined (P-G) provided essential tools for analysis; indeed, the entire scheme was set up and carried out in the P-G (prevalence-intensity) plane (cf [7]).

The key application of our method is leveraging the SCORE data set for any single-slide community test data T (SCORE or non-SCORE) to estimate its (unknown) true statistics (e.g. prevalence, intensity). This is accomplished by identifying a cluster of SCORE communities similar to T, measured by the likelihood of generating T statistics, via resampling. Once the SCORE cluster is identified, one can create a "virtual SCORE

replica" of community snapshot T and infer its statistics, including unknown true (*P*-*G*), along with the uncertainty (error) margins.

The effect of (*PG*) statistics combined with the cluster selection model is illustrated in Figure 4. Figure 4*A*, derived from raw data resampling (expected outcomes of single-slide KK screening), shows broad and consistently biased error margins (well below true *P*). Figure 4*B*, derived by the SCORE cluster scheme, comes much closer to the truth, with reduced error margins.

Our work highlights the role of combined *P*-*G* statistics for accurate community assessment, which could be relevant to other helminth infections. It also raises a challenging problem of developing *P*-*G*–based control guidelines that would extend the current WHO strategies based on prevalence alone [18]. Indeed, worm burden and the resulting egg release could vary widely within and between host communities. The intensity variable *G* complements prevalence *P*, as shown by SCORE data analysis (Figure 2). So, communities with near-identical values for *P* could exhibit vastly different *G* values (higher burden). They differ by EPG count distributions or graded prevalence levels, such as WHO LMH prevalence. Supplementary Figure 3 demonstrates the effect of increased intensity *G* within a narrow prevalence band ($P \approx 0.2$). Higher intensity (and the associated worm burden) could affect other features in such communities, for instance their potential response to MDA control (see, eg, [8]). So control strategies that rely on prevalence alone could expect a multitude of different outcomes.

Our work suggests that schistosomiasis control guidelines should account for *P*-*G* statistics to provide a more robust framework for disease control and elimination. However, the problem of extending WHO control guidelines would require not only a more detailed analysis of SCORE-like large-scale data sets (MDA response patterns) but also extensive numeric exploration of dynamic transmission models for SCORE-like communities (see, eg, [8, 9]).

Limitations and Extensions

The current version of cluster selection scheme uses 1624 SCORE community tests, augmented with an additional 5-country data set. It can be applied to any single-slide raw test data (reference point $z_0 = (P_0, G_0)$) that falls within or near the lamina-shaped region of Figure 2*A*). Reference points outside this range may not produce statistically significant clusters for estimating truth. We do not know whether the SCORE data set covers all possible transmission environments and infection patterns, but it looks sufficiently representative in terms of prevalence values.

The proposed cluster method can be extended beyond the SCORE test bed. Indeed, any multiday test community data can be added to augment the SCORE pool, like the 5-country data set. Such extensions could contribute to improved prediction and reduce uncertainty. Our modeling and analysis have focused on diagnostic test uncertainties alone. Another significant source of data uncertainty comes from statistical sampling, including target population groups and geographic regions. Future work will combine the methods and tools of both approaches to advance the goals of accurate assessment and control predictions. Going beyond diagnostic assessment, our approach can be combined with dynamic transmission models ([8–10, 19]) for reanalysis of SCORE MDA-progress patterns, persistent hot spots, and efficient control strategies.

Data and Computer Resources

The basic data source for our modeling and analysis is the publicly available SCORE data set (www.clinepidb.org); it also contains an additional file called "SCORE Five Country CCA Evaluation Cross-sectional." The computer codes and procedures were developed and run on the Wolfram Mathematica platform. Based on those, we deployed an open-source SCORE calculator. The code can take any raw KK test input (EPG counts) for single or multiple communities, in any data format, and will output their true statistics (*PG* and LMH) within error margins. This easy-to-use tool requires minimal experience in Mathematica and is available at a GitHub link (https://github.com/mln27/SCORE-Calculator).

Supplementary Data

Supplementary materials are available at *Open Forum Infectious Diseases* online. Consisting of data provided by the authors to benefit the reader, the posted materials are not copyedited and are the sole responsibility of the authors, so questions or comments should be addressed to the corresponding author.

Acknowledgments

We thank our Schistosomiasis Consortium for Operational Research and Evaluation (SCORE) colleagues for providing SCORE data sets and sharing their analysis and insights. We received much useful feedback from N. Lo (University of California at San Francisco), and V. Mallikarjun (Case Western Reserve University) has contributed to computer development.

Author contributions. D. G. and M. L. N. M. designed the project and interpreted the results. D. G. performed the analysis and computer implementation. All authors drafted the manuscript and reviewed and approved the final version.

Potential conflicts of interest. All authors: No reported conflicts.

References

1. Lamberton PH, Kabatereine NB, Oguttu DW, Fenwick A, Webster JP. Sensitivity and specificity of multiple Kato-Katz thick smears and a circulating cathodic antigen test for *Schistosoma mansoni* diagnosis pre- and post-repeated-praziquantel treatment. PLoS Negl Trop Dis **2014**; 8:e3139.
2. Hubbard A, Liang S, Maszle D, Qiu D, Gu X, Spear R. Estimating the distribution of worm burden and egg excretion of *Schistosoma japonicum* by risk group in Sichuan Province, China. Parasitology **2002**; 125:221–31.
3. Gryseels B, De Vlas SJ. Worm burdens in schistosome infections. Parasitol Today **1996**; 12:115–9.
4. Bärenbold O, Garba A, Colley DG, et al. Estimating true prevalence of *Schistosoma mansoni* from population summary measures based on the Kato-Katz diagnostic technique. PLOS Negl Trop Dis **2021**; 15(4):e0009310.
5. de Vlas SJ, Gryseels B. Underestimation of *Schistosoma mansoni* prevalences. Parasitol Today **1992**; 8:274–7.

6. Kokaliaris C, Garba A, Matuska M, et al. Effect of preventive chemotherapy with praziquantel on schistosomiasis among school-aged children in sub-Saharan Africa: a spatiotemporal modelling study. Lancet Infect Dis **2022**; 22:136–49.
7. de Vlas SJ, Engels D, Rabello AL, et al. Validation of a chart to estimate true *Schistosoma mansoni* prevalences from simple egg counts. Parasitology **1997**; 114(pt 2):113–21.
8. Li EY, Gurarie D, Lo NC, Zhu X, King CH. Improving public health control of schistosomiasis with a modified WHO strategy: a model-based comparison study. Lancet Glob Health **2019**; 7:e1414–22.
9. Lo NC, Gurarie D, Yoon N, et al. Impact and cost-effectiveness of snail control to achieve disease control targets for schistosomiasis. Proc Natl Acad Sci U S A **2018**; 115:E584–91.
10. Gurarie D, King CH, Yoon N, Li E. Refined stratified-worm-burden models that incorporate specific biological features of human and snail hosts provide better estimates of *Schistosoma* diagnosis, transmission, and control. Parasit Vectors **2016**; 9:1–19.
11. Colley DG, Jacobson JA, Binder S. Schistosomiasis Consortium for Operational Research and Evaluation (SCORE): its foundations, development, and evolution. Am J Trop Med Hyg **2020**; 103(suppl 1):5–13.
12. Colley DG, Fleming FM, Matendechero SH, et al. Contributions of the Schistosomiasis Consortium for Operational Research and Evaluation (SCORE) to schistosomiasis control and elimination: key findings and messages for future goals, thresholds, and operational research. Am J Trop Med Hyg **2020**; 103(suppl 1):125–34.
13. Bergquist NR. Schistosomiasis Consortium for Operational Research and Evaluation: mission accomplished. Am J Trop Med Hyg **2020**; 103(suppl 1):1.
14. Kittur N, Binder S, Campbell CH, et al. Defining persistent hotspots: areas that fail to decrease meaningfully in prevalence after multiple years of mass drug administration with praziquantel for control of schistosomiasis. Am J Trop Med Hyg **2017**; 97:1810–7.
15. Bärenbold O, Raso G, Coulibaly JT, N'Goran EK, Utzinger J, Vounatsou P. Estimating sensitivity of the Kato-Katz technique for the diagnosis of *Schistosoma mansoni* and hookworm in relation to infection intensity. PLOS Negl Trop Dis **2017**; 11:e0005953.
16. Secor WE, Wiegand RE, Montgomery SP, Karanja DMS, Odiere MR. Comparison of school-based and community-wide mass drug administration for schistosomiasis control in an area of western Kenya with high initial *Schistosoma mansoni* infection prevalence: a cluster randomized trial. Am J Trop Med Hyg **2020**; 102: 318–27.
17. Kittur N, King CH, Campbell CH, et al. 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**; 101:617–27.
18. World Health Organization. WHO guideline on control and elimination of human schistosomiasis. Geneva: World Health Organization; **2022**.
19. Gurarie D, Yoon N, Li E, et al. modelling control of *Schistosoma haematobium* infection: predictions of the long-term impact of mass drug administration in Africa. Parasit Vectors **2015**; 8:1–14.