Aggregation in schistosomiasis: comparison of the relationships between prevalence and intensity in different endemic areas

H. L. GUYATT[1], T. SMITH[1], B. GRYSEELS[2], C. LENGELER[3], H. MSHINDA[4], S. SIZIYA[5], B. SALANAVE[6], N. MOHOME[7], J. MAKWALA[8], K. P. NGIMBI[8] and M. TANNER[1]

[1] Department of Public Health and Epidemiology, Swiss Tropical Institute, Socinstrasse 57, Basel CH-4002, Switzerland
[2] Department of Parasitology, University of Leiden, P.B. 9605, 2300 RC Leiden, The Netherlands
[3] Tropical Health and Epidemiology Unit, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, UK
[4] Ifakara Centre, PO Box 53, Ifakara, Tanzania
[5] Tropical Diseases Research Center, POB 71 769 Ndola, Zambia
[6] IFORD, B.P. 1556, Yaoundé, Cameroon
[7] OCEAC, B.P. 288, Yaoundé, Cameroon
[8] Département de Démographie, Université de Kinshasa, Zaire

SUMMARY

Distributions of the intensities of helminth infections within their host populations are invariably aggregated. In the case of the intestinal nematodes, the degrees of aggregation have been shown to be species specific, and constant for any given species despite geographical variation in study sites. This species-specific aggregation can be quantified and used as a tool in planning control interventions. One practical application is that the prevalence of infection can be used to predict the prevalence of heavy infection and thus the risks of morbidity. This paper investigates the patterns of aggregation in schistosome egg counts in different endemic areas in Africa (data sets were obtained from Burundi, Cameroon, Tanzania, Zambia and Zaire). The analysis demonstrates that the degree of parasite aggregation, for both *Schistosoma mansoni* and *S. haematobium*, differs amongst the different study localities. This is probably due to area-specific differences in host exposure and immunity. This implies that for these schistosome species, it is not possible to predict egg count distributions or morbidity levels from prevalence data alone.

Key words: *Schistosoma mansoni*, *Schistosoma haematobium*, prevalence, intensity, epidemiology, control.

INTRODUCTION

Patterns of schistosomiasis infection are shaped by two factors: host exposure and host immunity. Both these factors demonstrate marked heterogeneity within any given human population (Butterworth *et al.* 1984; Wilkins *et al.* 1987; Chandiwana & Woolhouse, 1991), implying that the distribution of infection in any one community will also be heterogeneous. Indeed, parasitic helminths are invariably more aggregated than is predicted by a Poisson model for the distribution of parasites among hosts; most hosts have a few parasites and a few hosts have many parasites. These aggregated distributions of parasite intensities have been observed for all the major helminths of humans: the schistosomes, the filarial worms and the three most prevalent intestinal nematodes – *Ascaris lumbricoides*, *Trichuris trichiura* and the hookworms (see Anderson & May, 1991).

The negative binomial distribution has been widely used to describe these aggregated distributions. This is a two-parameter distribution defined by the mean intensity of infection and the parameter k, an inverse measure of the degree of aggregation. Estimates of k have been obtained from cross-sectional studies of single human populations in order to quantify the degree of parasite aggregation. These k values for human helminths tend to be small, indicating severe aggregation, and tend to lie within a restricted range (0.1–1.0) (Anderson & May, 1991).

If data from several surveys with similar values of k are available, the negative binomial model predicts a specific form of relationship between prevalence and mean intensity. From this a global estimate of k can be obtained, even when the complete frequency distributions for the different surveys are not available.

In the cases of *A. lumbricoides* and hookworm, similar relationships between prevalence and intensity of a given parasite are found in numerous different geographical localities. This makes it possible to quantify the aggregation with a single value of k for each species (Guyatt *et al.* 1990;

Lwambo, Bundy & Medley, 1992). This *k* value can then be used to predict frequencies of heavy infections from simple prevalence data, greatly simplifying screening for high-risk areas of morbidity (Guyatt & Bundy, 1991). Estimates of the distribution of the infectious load are also required by models used to predict the implications of selective control measures, the distribution of morbidity within a community and the dynamics of transmission (Anderson & May, 1985; Anderson & Medley, 1985). These models become tractable if a uniform *k* value can be assumed.

In the case of the intestinal nematodes, it is possible to obtain estimates of adult worm burden by anthelmintic expulsion. However, because schistosomes inhabit the blood vessels, adult schistosome worms are only obtainable by autopsy. Therefore, the only community infection data available for the schistosomes is based on egg counts. Schistosome egg counts represent only an indirect measure of worm burdens. Egg count distributions depend on the life-expectancies of the worms, and on immune status, whilst transmission can be highly focal. All these factors might lead to a poor fit of the negative binomial model to faecal egg intensities (De Vlas & Gryseels, 1992; De Vlas *et al.* 1992), and might cause variation in levels of aggregation across geographically separate human populations.

In this study we investigated the extent of aggregation in egg-counts for both *Schistosoma haematobium* and *S. mansoni* using survey data from areas of different endemicity in Africa. In addition to examining the fit of the negative binomial model, we used logistic regression methods to compare the empirical relationships found in different study areas between prevalence of infection, mean intensity of infection and prevalence of heavy infection. We consider the implications of these relationships for the planning of control strategies.

MATERIALS AND METHODS

Study populations

Surveys for *S. haematobium* were carried out in Tanzania, Cameroon and Zambia. In each case, survey teams visited primary schools and sampled children aged between 7 and 17 years. The numbers of schools and total numbers of children sampled are summarized in Table 1.

The Tanzanian surveys took place in the Kilombero district, Morogoro region. This district covers 14000 km² and has approximately 170000 inhabitants in 45 villages. There are 77 primary schools in the district (Lengeler *et al.* 1991).

The data from Cameroon were collected in the Mayo Danay district which covers 5260 km² and has a total population of 310000 persons. The area is part of the Sahelian belt, with a hot and dry climate

and a short rainy season (Red Urine Study Group, 1994).

The Zambian surveys took place in Isoka district in the Northern Province. Isoka district covers 14000 km² with an estimated population of 100000. The rainy season is from November to March (Red Urine Study Group, 1994).

Surveys for *S. mansoni* were carried out in Zaire and in two regions of Burundi (Cohoha Lake and Rusizi Plain).

The Zairian study took place in the town of Matadi and the administrative zone of Songololo, in the narrow land stretch between the Zaire River and the Angolan republic, west of Kinshasa. The study area covered 9000 km² and had a population of 380000; 180000 of which were in Matadi. The rainy season is from October to May (Red Urine Study Group, 1994).

The Cohoha Lake is situated in the north-east highlands of Burundi. At the time of the surveys, this area was thought to be a new focus of *S. mansoni* (Gryseels *et al.* 1987). The population density in this areas was 150 inhabitants/km². The data comprised a 10% systematic random population sample.

The Rusizi Plain is located in N.W. Burundi. The area covers approximately 1000 km² and has a total population estimated at 140000. The study surveyed a 5% systematic random population sample (Gryseels & Nkulikyinka, 1988).

In order to make the data from the population samples in Burundi comparable with the samples of school children from other areas, the data analysis was performed only on 7 to 17-year-olds within these two studies.

Measurements of prevalence and intensity

The field methodologies for the different study areas were comparable. Each study team also established a quality control system for the processing of stool and urine samples and the reading of egg counts.

For each child in the study egg counts were used to measure the intensity of infection (x). Egg counts for *S. haematobium* (expressed as eggs/10 ml urine) were determined using a single 10 ml urine filtration. Urine filtration was performed with a commercially available low-cost kit using Nucleopore filters (Pleasanton, USA) and purchased through Tropical Health Technology (Doddington, March, Cambridgeshire, UK). Egg intensities of *S. mansoni* were determined using Kato-Katz slides (Katz, Chavez & Pellegrino, 1972). In Zaire, a single 42 mg Kato was used, and in Burundi, duplicate 28 mg Katos were performed for each individual. These intensities were expressed as eggs/g of faeces.

Statistical methods

The parasitological data were analysed using the school or village as the unit of study. For each school

Table 1. *Summary statistics for the study sites: sample sizes, age-distribution of sample and prevalence rates of infection (see Materials and Methods section for references)*

	Country (district)	No. of schools/ villages	Total no. of individuals	Mean age (years)	S.D. of age	Overall prevalence of infection (%)	Range of prevalence rates (%)	Mixed schistosome infection
S. haematobium	Tanzania (Kilombero)	56	4279	11·8	1·6	22·4	1·4–78·4	< 5 %*
	Cameroon (Mayo Danay)	10	747	11·4	1·7	62·5	30·8–87·4	None
	Zambia (Isoka)	10	783	12·4	1·9	32·8	0–85·8	None
S. mansoni	Zaire (Songololo)	60	6142	12·4	2·3	31·0	0–88·2	20 %†
	Burundi (Cohoha Lake)	11	417	11·4	3·0	40·5	0–64·9	None
	Burundi (Rusizi Plain)	36	1626	11·1	3·1	39·2	0–84·0	None

* Of *S. mansoni* infection.
† Of *S. haematobium* infection.

or village, the prevalence of infection $(Pr(x > 0))$, mean intensity of infection, M (arithmetic mean including negatives), and prevalence of heavy infection at different thresholds were determined. The thresholds of heavy infection used in the analysis were mainly based on those initially recommended by WHO (1985): the thresholds chosen for *S. haematobium* were 50, 100 and 200 eggs/10 ml urine and those for *S. mansoni* were 100, 200 and 500 eggs/g of faeces.

Relationships between prevalence and intensity of infection

Prevalence–intensity relationships were investigated by modelling the dependence of $Pr(x > 0)$ on M, comparing different schools/communities within a given study area. The full set of alternative relationships fitted is illustrated using the *S. mansoni* dataset from Songololo (Zaire).

Analysis of helminth infection intensities often starts from a null model of independent distribution of eggs within the population, leading to a Poisson distribution of intensities within any one population and a fixed relationship between $Pr(x > 0)$ and M (Table 2) with $Pr(x > 0)$ binomially distributed about the predicted value of $1 - \exp(-M)$. As a measure of fit of this model to the Zaire data, the deviance (McCullagh & Nelder, 1989) was computed.

The variance of infection intensities within a population is usually greater than that predicted by the Poisson. This is equivalent to the aggregation of infection in some individuals. The negative binomial model arises as a generalization of the Poisson (Elliott, 1977) which allows for this excess variation. The level of this overdispersion is described by the index of aggregation, k. As k approaches infinity, so

the negative binomial distribution converges to a Poisson. For any given negative binomial, k is related to the prevalence and the arithmetic mean intensity (see Table 2). If intensities in a number of populations have different prevalences but a common value of k, this level of aggregation can be estimated by maximum likelihood (Guyatt *et al.* 1990). This model for the inter-population prevalence–intensity relationship was also fitted to the *S. mansoni* data from Songololo.

An alternative null model is that host populations consist partly of a subpopulation not exposed to the parasite, with the exposed subset all infected and intensities in the infected individuals determined by endogenous factors unrelated to exposure. Such a model (the straight line model, see Table 2) predicts that prevalence and mean infection intensity are unrelated. It makes no assumptions about the distribution of intensities among the infected individuals. The maximum likelihood fit to the *S. mansoni* data from Songololo was evaluated again assuming a binomial distribution of $Pr(x > 0)$ about the expected value. Models with further parameters were fitted in order to generalize the straight line null model to allow for effects of variation in exposure. The choice of model was determined by the fit to the data rather than theoretical considerations of how exposure variation might influence distributions of intensities, with the constraint that the theoretical prevalence should be zero at zero intensity. This led to the use of a logistic regression (Dyke & Patterson, 1952; McCullagh & Nelder, 1989) to fit a relationship between $Pr(x > 0)$ and M, as follows:

$$\text{logit}(Pr(x > 0)) = \beta_0 + \beta_1 \ln(M), \tag{1}$$

where $\text{logit}(Pr(x > 0)) = \ln(Pr(x > 0)/(1 - Pr(x > 0)))$. The variation in prevalence about

the fitted values is again assumed to be binomial. This model was fitted using the package EGRET (SERC, 1991). The fit of the line was tested by examining the change in deviance associated with the inclusion of higher order polynomial terms in $\ln(M)$.

This regression line modelled the expected increase in prevalence with increasing mean intensity, but did not correspond to any simple parametric form for the distributions of egg counts within individual populations. If sampling variation were the only source of within-community variability, then the binomial error would account for all the scatter about the regression lines. However, there could be further variation in the patterns of aggregation between schools within a study area, leading to a greater degree of scatter about the regression.

The regression lines and the amount of extra-binomial variation between schools were estimated simultaneously by fitting a logistic-normal model in the package EGRET (SERC, 1991). The logistic-normal model (Pierce & Sands, 1975; SERC, 1991; Breslow & Clayton, 1993) differs from the usual logistic regression model in including an extra error term σ which measures the extra-binomial variation. Assuming I to be a categorical variable taking integer values i, the estimate of the prevalence of infection for school i was:

$$\text{logit}\,(Pr(x > 0\,|\,I = i)) = \beta_0 + \beta_1 \ln(M_i) + \sigma u_i, \qquad (2)$$

where u_i is a perturbation which is different for each school, with the whole set of perturbations normally distributed about 0, with unit standard deviation.

The logistic regression model (2) was fitted to the community-specific data to determine the relationship between prevalence and intensity of infection for each of the endemic areas. Likelihood ratio tests (equivalent to comparisons of the deviance) were used to test whether there were statistically significant differences between countries in the best-fitting values of β_0 and β_1.

Relationship between heavy infection and prevalence of infection

Logistic-normal models were also used to investigate the proportions of individuals ($Pr(x > X)$) exceeding given intensity cutoffs (X) as a function of the overall prevalence.

In order to allow for the structural relationship between prevalence of high intensity and overall prevalence, the models fitted separately for each threshold considered were of the general form:

$$\begin{aligned}\text{logit}\,(Pr(x > X\,|\,x > 0)) \\ = \beta_{2,x} + \beta_{3,x}\,\text{logit}\,(Pr(x > 0)) + \sigma_{1,x}, \quad (3)\end{aligned}$$

where the conditional probability, $Pr(x > X\,|\,x > 0)$, is the prevalence of intensities (x) greater than some cutoff X among the parasite positive individuals in a

school. $\text{Exp}\,(\beta_{2,x})/(1 + \exp(\beta_{2,x}))$ is the estimated proportion of high-intensity infections among all infections. In the limiting case, $X = 0$; $\beta_{2,0} = \infty$; $\beta_{3,0} = 0$. For $X > 0$, the statistical significance of $\beta_{3,x}$ tests the hypothesis that the frequency of high-intensity infections is prevalence dependent.

Estimates of $Pr(x > X\,|\,x > 0)$ were converted to unconditional estimates of $Pr(x > X)$:

$$Pr(x > X) = Pr(x > X\,|\,x > 0)\,Pr(x > 0). \qquad (4)$$

For each threshold, in order to test whether there was heterogeneity between countries in the relationship between high-intensity infection and prevalence, the deviances of models which fitted the different values of $\beta_{2,x}$ and $\beta_{3,x}$ for each country were subtracted from those of equivalent models which fitted single values for each parameter. This gave likelihood ratio chi-square statistics with 2 degrees of freedom.

RESULTS

The overall prevalence of *S. haematobium* and *S. mansoni* in the various studies are shown in Table 1. This table also presents the mean ages of the children studied, which were similar for all the study groups.

The prevalence–intensity relationships estimated by the different models for the Zairian *S. mansoni* dataset are shown in Fig. 1. The deviances for the different models are given in Table 2. The deviance for a model which fits well should be close to the number of residual degrees of freedom. On this criterion, the fits of the Poisson, negative binomial, and straight-line models are all very poor. The logistic model without extra-binomial variation fitted much better. The improvement in deviance associated with the inclusion of the term for extra-binomial variation was also highly significant. The fit of this model was consequently considerably better than that of any of the others, although the deviance remained approximately twice the residual degrees of freedom. However, the fitted prevalence–intensity relationship for the logistic normal model was very similar to that for the ordinary logistic model.

Logistic models including higher-order polynomial terms in $\ln(M)$ made only a small improvement to the fit of the model, indicating that the regression on $\ln(M)$ adequately represents the average dependence of $Pr(x > 0)$ on M.

The prevalence–intensity relationships in the other populations studied are shown in Fig. 2 (*S. haematobium*) and Fig. 3 (*S. mansoni*). All the egg-count distributions are clearly highly aggregated. If the distribution of eggs within a population was Poisson very few uninfected individuals would have been found in any of the communities studied, but the fit of the negative binomial model was invariably much poorer than that of the logistic regressions.
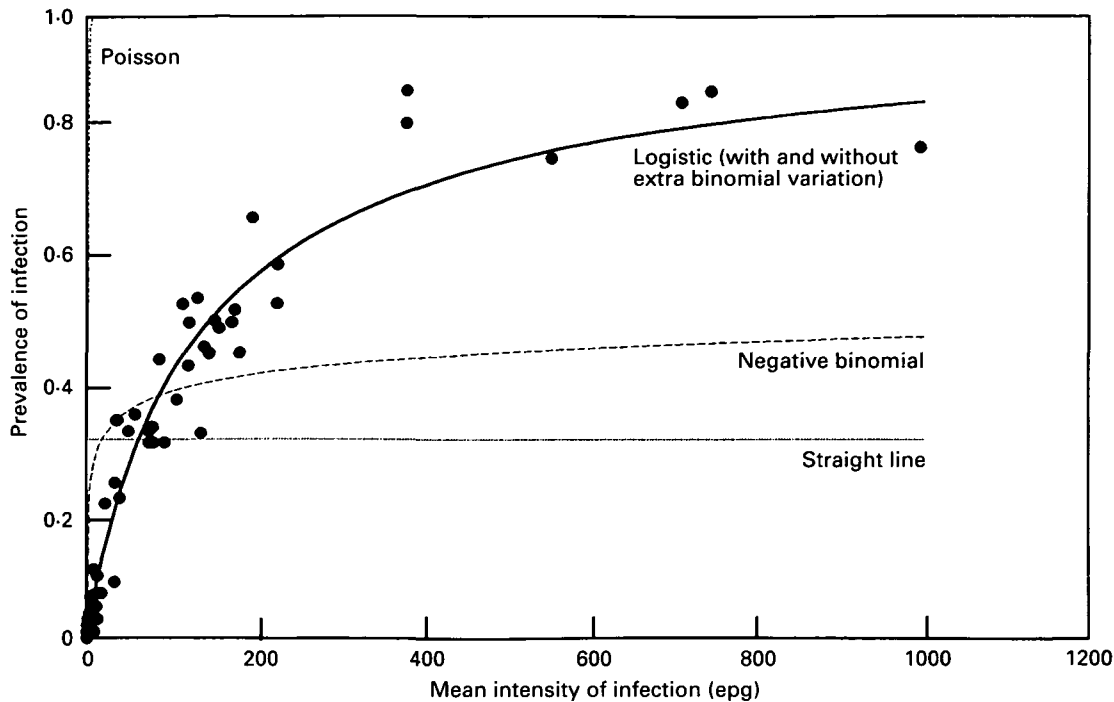
Fig. 1. The observed relationship between prevalence and mean intensity of infection for the *Schistosoma mansoni* data from Songololo (Zaire), and the fitted lines of a Poisson, negative binomial, straight line and logistic model. The details of each model are given in Table 2. The maximum likelihood estimates of the model parameters were as follows: negative binomial $k = 0.070$; straight line $c = 0.320$; $\beta_1 = 0.937$ and $\beta_0 = -4.62$ for the logistic model without extra-binomial variation; $\beta_1 = 0.940$ and $\beta_0 = -4.64$ for the logistic model with extra-binomial variation.

Table 2. Models for the relationship between prevalence and mean intensity: fit of the models to the *Schistosoma mansoni* data from Songololo (Zaire)

Model		Deviance	Residual degrees of freedom
Poisson	$Pr(x > 0) = 1 - \exp(-M)$	256478·9	58
Negative binomial	$Pr(x > 0) = 1 - (1 + M/k)^{-k}$	797·6	57
Straight line	$Pr(x > 0) = \text{constant}$	2237·5	57
Logistic*	$Pr(x > 0) = aM^b/(1 + aM^b)$	145·3	56
Logistic*†	$Pr(x > 0) = aM^b/(1 + aM^b)$	111·6	55

* $a = \exp(\beta_0)$; $b = \beta_1$.
† Including extra binomial variation term.

Comparison of study localities demonstrates that the prevalence–intensity relationships differ between areas. For *S. haematobium* (Fig. 2), the most aggregated distributions were observed for the Kilombero district (Tanzania), since the prevalence there was lower at high intensities. The least aggregated distributions are observed in the Mayo Danay district (Cameroon), where much higher prevalences of infection are observed at comparable intensities. The logistic regression analysis indicated that the differences between the intercepts for the three different countries were statistically significant (likelihood ratio (LR) chi-square 53·3, 2 D.F., $P < 0.001$), but the best fitting gradients were very similar. The regression lines illustrated are therefore from a regression model which fitted different values of the intercept (β_0), but the same slopes (β_1) for the three countries (see Fig. 2).

It is apparent from Fig. 2 that there is also scatter about the regression lines, particularly in the Tanzanian dataset. The estimate of σ for *S. haematobium* was much higher than its standard error, indicating that there is highly significant extra-binomial variation about the regression lines. Therefore, not only do the three countries differ in the degree of aggregation, but the different schools within these countries also show significantly different levels of aggregation.
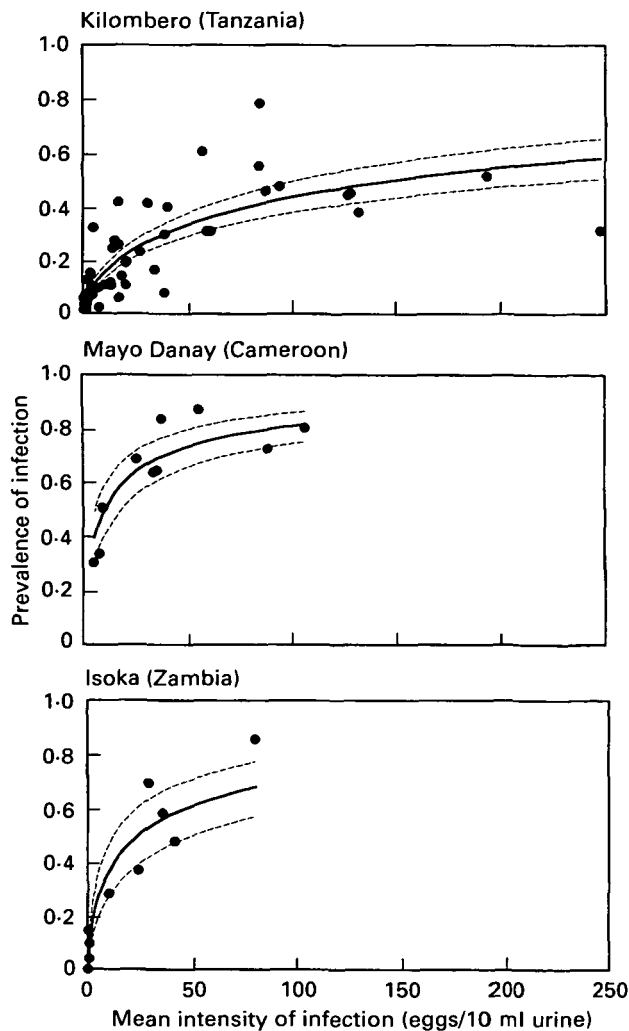
Fig. 2. The relationship between prevalence and mean intensity of infection for *Schistosoma haematobium* in endemic areas in the Kilombero district (Tanzania), Mayo Danay district (Cameroon) and the Isoka district (Zambia). The fitted lines represent the best fit of the logistic regression model (see text) with $\beta_1 = 0.63$ and $\beta_0 = -3.14$ for Kilombero, -1.44 for Mayo Danay and -2.00 for Isoka. The error term σ was 0.544 (s.e. $= 0.59$). The dashed lines represent the 95% C.I. for the fitted line.
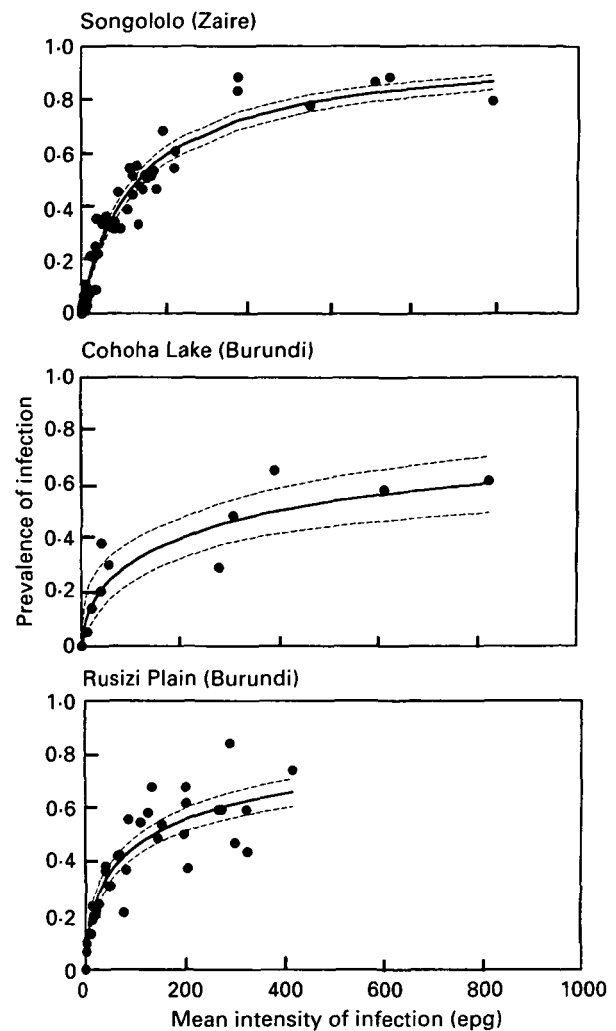
Fig. 3. The relationship between prevalence and mean intensity of infection for *Schistosoma mansoni* in the endemic areas of Songololo district (Zaire), Cohoha Lake (Burundi) and Rusizi Plain (Burundi). The fitted lines represent the best fit of the logistic regression model (see text) with the following parameter values: $\beta_1 = 0.94$ and $\beta_0 = -4.6$ (Songololo), $\beta_1 = 0.57$, $\beta_0 = -3.44$ (Cohoha Lake), $\beta_1 = 0.60$, $\beta_0 = -2.96$ (Rusizi Plain). The error term σ was 0.321 (s.e. $= 0.056$) for Songololo, 0.242 (s.e. $= 0.150$) for Cohoha Lake and 0.258 (s.e. $= 0.082$) for Rusizi Plain. The dashed lines represent the 95% C.I. for the fitted line.

For *S. mansoni* (Fig. 3), the most aggregated parasite population is observed in Cohoha Lake (Burundi), followed by Rusizi Plain (Burundi), with Songololo district (Zaire) exhibiting the lowest parasite aggregation. There were again statistically significant differences between the intercepts for the different studies (LR chi-square $= 15.1$, $P < 0.001$). Unlike the situation for *S. haematobium*, the differences between the gradients were statistically significant (LR chi-square $= 26.0$, $P < 0.001$). Different gradients (and different values of σ) were therefore estimated for the three different countries. Although the scatter around the regression lines is much greater in the datasets from Burundi (which had smaller sample sizes) than in the data from Songololo district in Zaire, the amount of extra-binomial

variation, as measured by the parameter σ, was similar in all three datasets. This again indicated overall highly significant variation between communities in the level of aggregation.

Differences in the distribution of infection will also be manifested in differences in the relationships between prevalence of heavy infection and prevalence of any infection. The observed and fitted relationships between prevalence of heavy infection and overall prevalence of infection for *S. haematobium* and *S. mansoni* are illustrated in Figs 4 and 5 respectively. There is a direct correlation between the two prevalences. This relationship varies between countries and according to the heavy infection
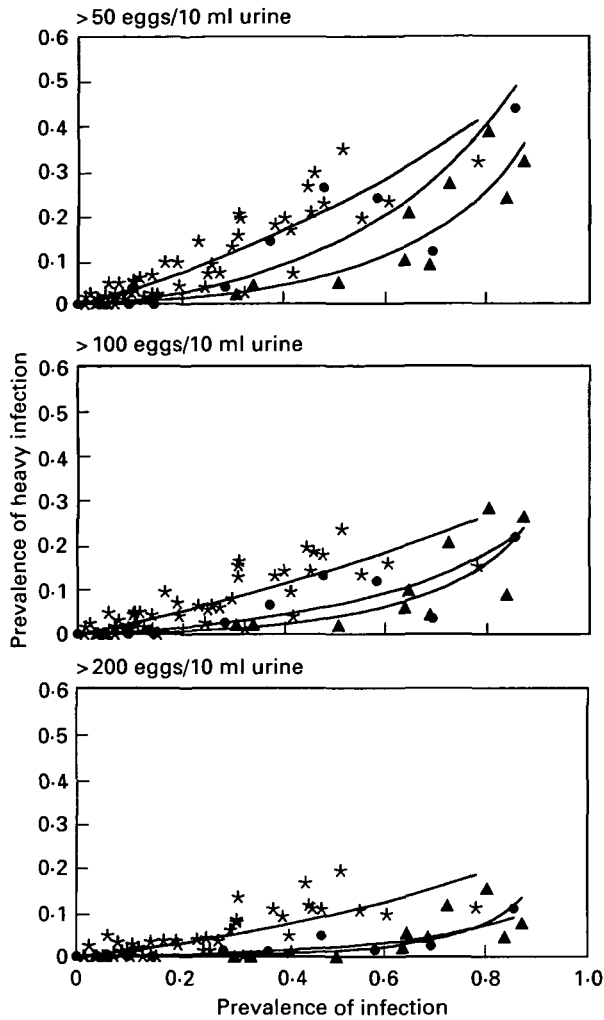
Fig. 4. The relationship between prevalence of heavy infection and prevalence of infection for *Schistosoma haematobium* in the endemic areas of Kilombero (Tanzania) (∗), Mayo Danay (Cameroon) (▲) and Isoka (Zambia) (●). The fitted lines represent the best fit of the logistic regression model (see text) with the parameter values in Table 3.
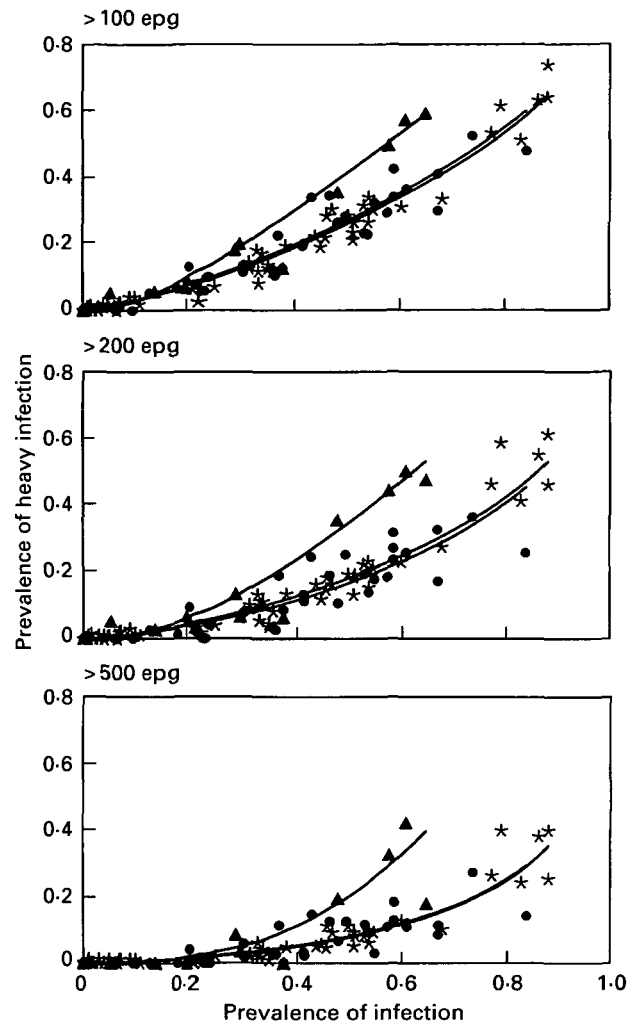
Fig. 5. The relationship between prevalence of heavy infection and prevalence of infection for *Schistosoma mansoni* in endemic areas in Songololo (Zaire) (∗), Cohoha Lake (Burundi) (▲) and Rusizi Plain (Burundi) (●). The fitted lines represent the best fit of the logistic regression model (see text) with the parameter values in Table 4.

threshold used. The higher the degree of aggregation, the higher the prevalence of heavy infection at any given overall prevalence of infection. As the value of the threshold increases, the differences between the study areas decrease. In the case of *S. haematobium*, for example, at a threshold of > 200 eggs/10 ml, the predicted relationships between heavy infection and infection are almost identical for the Mayo Danay district (Cameroon) and the Isoka district (Zambia) (see Fig. 4).

In *S. haematobium* there are statistically significant differences between countries in the relationship between prevalence of heavy infection and prevalence of infection. For each cutoff, there was a highly significant difference between the estimated value of β_2 for the different studies, indicating that the difference between the countries was not due to chance (see Table 3). A different regression was again fitted for each country for illustrative purposes

(Fig. 4). In all three countries, there was a marked relationship between the proportion of infected individuals with high intensities, and the overall prevalence. Figure 4 illustrates the fitted curves, plotting the fitted values of the prevalence of heavy infection against $Pr(x > 0)$ separately for each country.

Equivalent analysis was undertaken for *S. mansoni* (see Table 4 and Fig. 5). As for *S. haematobium*, there was a highly significant difference between the values of β_2 for the different *S. mansoni* studies (see Table 4). There were similar levels of aggregation in the Songololo district (Zaire) and the Rusizi Plain (Burundi) datasets, but the Cohoha Lake (Burundi) dataset was clearly different from these others, with a much higher level of aggregation (see Fig. 5).

On an operational level, the above analysis permits prediction of the prevalence of heavy infection at any

Table 3. *Parameter estimates for models of probability of high-intensity infection as a function of study area, and prevalence of Schistosoma haematobium*

| | Intensity cutoff (X) in eggs/10 ml | | | | | |
| | 50 | | 100 | | 200 | |
	β_2	β_3	β_2	β_3	β_2	β_3
Kilombero district (Tanzania)	$-0{\cdot}228$	$0{\cdot}259$	$-0{\cdot}877$	$0{\cdot}136$	$-1{\cdot}45$	$0{\cdot}224$
Mayo Danay district (Cameroon)	$-1{\cdot}80$	$0{\cdot}749$	$-2{\cdot}48$	$0{\cdot}779$	$-3{\cdot}70$	$1{\cdot}03$
Isoka district (Zambia)	$-0{\cdot}962$	$0{\cdot}693$	$-1{\cdot}89$	$0{\cdot}467$	$-3{\cdot}16$	$0{\cdot}556$
Likelihood ratio statistics*	$6{\cdot}5$ ($0{\cdot}04$)	$4{\cdot}7$ ($0{\cdot}10$)	$12{\cdot}0$ ($0{\cdot}003$)	$4{\cdot}8$ ($0{\cdot}09$)	$16{\cdot}7$ ($<0{\cdot}001$)	$4{\cdot}0$ ($0{\cdot}13$)

* Likelihood ratio chi-squared statistics (2 D.F.) testing the statistical significance of differences in parameter values between countries. Values in parentheses are P-values.

Table 4. *Parameter estimates for models of probability of high-intensity infection as a function of study area, and prevalence of Schistosoma mansoni*

| | Intensity cutoff (X) in eggs/g | | | | | |
| | 100 | | 200 | | 500 | |
	β_2	β_3	β_2	β_3	β_2	β_3
Songololo district (Zaire)	$0{\cdot}048$	$0{\cdot}466$	$0{\cdot}598$	$0{\cdot}499$	$1{\cdot}73$	$0{\cdot}653$
Cohoha Lake district (Burundi)	$1{\cdot}53$	$1{\cdot}15$	$0{\cdot}767$	$1{\cdot}15$	$-0{\cdot}427$	$1{\cdot}40$
Rusizi Plain district (Burundi)	$0{\cdot}134$	$0{\cdot}461$	$-0{\cdot}735$	$0{\cdot}534$	$-1{\cdot}68$	$0{\cdot}631$
Likelihood ratio statistics*	$31{\cdot}7$ ($<0{\cdot}001$)	$5{\cdot}2$ ($0{\cdot}076$)	$28{\cdot}3$ ($<0{\cdot}001$)	$4{\cdot}2$ ($0{\cdot}124$)	$21{\cdot}2$ ($<0{\cdot}001$)	$2{\cdot}9$ ($0{\cdot}23$)

* Likelihood ratio chi-squared statistics (2 D.F.) testing the statistical significance of differences in parameter values between countries. Values in parentheses are P-values.

Table 5. *The predicted prevalence of heavy Schistosoma haematobium infection at a given prevalence of infection*

| Prevalence of infection (%) | Heavy infection threshold (X) (eggs/10 ml urine) | Predicted prevalence of heavy infection (%) | | |
		Kilombero district (Tanzania)	Mayo Danay district (Cameroon)	Isoka district (Zambia)
50	50	22·2	7·1	13·8
	100	14·7	3·9	6·6
	200	9·5	1·2	2·0
25	50	10·3	2·6	5·4
	100	7·0	1·4	2·7
	200	4·3	0·4	0·8

given prevalence of infection, and thus identification of 'high-risk' areas based solely on the proportion of individuals infected. WHO (1985) recommended that in children 7–14 years of age, $>50\%$ prevalence of infection is high risk for schistosomiasis, 25–50% is moderate risk and $<25\%$ is low risk.

Using the fitted relationships between heavy infection and prevalence of infection, it was possible to predict what these threshold prevalences of infection are equivalent to in terms of the prevalence of heavy infection in the different areas for both *S. haematobium* (Table 5) and *S. mansoni* (Table 6). It can be seen that there are marked differences in the minimum prevalence of heavy infection at each of these prevalences of infection. For instance, a high-risk area corresponding to an infection prevalence of 50% could reflect between 1 and 22% heavy infection for *S. haematobium* and between 8 and

Table 6. *The predicted prevalence of heavy* Schistosoma mansoni *infection at a given prevalence of infection*

Prevalence of infection (%)	Heavy infection threshold (X) (eggs/g)	Predicted prevalence of heavy infection (%)		
		Songololo district (Zaire)	Cohoha Lake district (Burundi)	Rusizi Plain district (Burundi)
50	100	25·6	41·1	26·7
	200	17·7	34·1	16·2
	500	7·6	19·7	7·9
25	100	11·4	18·2	12·0
	200	7·6	13·9	6·8
	500	2·9	6·3	3·0

41 % for *S. mansoni*, depending on the threshold intensity chosen. If, for instance, the WHO definitions of heavy infection are used (> 50 eggs/10 ml for *S. haematobium* and > 100 eggs/g for *S. mansoni*) (WHO, 1985), then the range of minimum prevalences of heavy infection is reduced to between 7 and 22 % for *S. haematobium* and 26 and 41 % for *S. mansoni*. As the value of the threshold increases, the differences between these countries are reduced. The scatter around these predicted curves, caused by differences in aggregation between communities within a study area, suggests that in practice this range may be higher.

DISCUSSION

Frequently, the only data available from epidemiological surveys of helminths are prevalences of infection. This has encouraged attempts to predict intensity distributions from simple prevalence data, since morbidity risks are related to intensities (Guyatt *et al.* 1990; Guyatt & Bundy, 1991; Lwambo, Medley & Bundy, 1992). The distribution of schistosome intensities within a host population can only be investigated indirectly using egg counts. Jordan (1972) presented data on the relationship between prevalence and intensity of *S. mansoni* infection from studies in Tanzania, Brazil and St Lucia. The pattern suggested a direct relationship between prevalence and intensity, which was the same for the different study sites.

Data from the present study found the general pattern of an increase in prevalence with mean intensity at low intensities, but less relationship at high endemicities. The *k* values were very low, indicating extreme aggregation, but the prevalence–intensity relationships across populations did not fit the negative binomial with a constant *k*.

Logistic regressions provided a much better fit to this relationship but these were of little value in predicting the risk of high-intensity infection from prevalence data, because different endemic areas have different prevalences corresponding to given mean intensities. There was also more variation in

prevalence between schools within countries than would be expected if the risk of infection was related in a consistent way to the average intensity in the community.

What causes these differences in egg count distributions? It is unlikely that these differences are artefactual and are the result of differences in the ability of technicians to detect infected individuals. If technicians from one study were missing a large number of infected individuals, then the data would suggest an apparently higher degree of parasite aggregation within this area. However, it is improbable that a technical error could be responsible for all the observed differences between studies. Each team had established a quality control scheme at each level of the parasitological investigation. For example, the two studies in Burundi used the same technicians, whereas the relationship between prevalence and intensity in these two areas differed markedly.

The differences in parasite aggregation probably reflect geographical variation in current and past patterns of transmission. One aspect which would need to be pursued is the whole question of immunity. Information is now accumulating on the determinants of immunity under different situations (Hagan *et al.* 1991), and it is likely that differences in the immunity to infection in different areas could partly account for differences in observed aggregation. For example in Burundi, Cohoha Lake exhibited a much higher degree of parasite aggregation than the Rusizi Plain. At the time of data collection the Cohoha Lake region was a relatively new focus for *S. mansoni*, hence it is possible that the levelling effect of acquired immunity was less, resulting in more heterogeneity in infection intensities. Since experience of infection is believed to be a major determinant of differences in immunity, variations in current exposure both between and within sites provide an alternative explanation. The study area in the Cohoha Lake region had a more focal transmission than that in the Rusizi Plain. A higher focality of transmission is likely in itself to result in higher degrees of aggregation.

There is clearly a need for standardized procedures in defining transmission sites with an aim to determining the transmission pattern within an area, and thus predicting the distribution of schistosomiasis intensities. The permanence of breeding sites, age at first infection (an indicator of the rate of transmission), seasonality of infection rate, the bionomics of the vectors, and behavioural and socio-economic factors related to the human host could all influence focality of transmission. By incorporating such factors into a model it might be possible to predict the level of aggregation and the prevalence of heavy infection at any given prevalence of infection. Such a standardized procedure would be an extension of the methodologies proposed by Ali & Byslov (1986; based on Dixon (1985)) and Tanner (1989). It is also apparent that such an approach would greatly benefit from a much clearer understanding of how the geographical limits of a transmission area should be defined.

In the last 10 years, the prevalence of infection has been used as an indicator for the prevalence of heavy infection (WHO, 1985). However, as this study has shown, the relationship between these variables is area-specific, suggesting that such general standards are not appropriate for schistosomiasis. It is encouraging that the prevalence thresholds for high risk of schistosomiasis (WHO, 1985) have not been maintained in the recent WHO recommendations (WHO, 1993), and this paper provides a justification for this.

In conclusion, this analysis has shown that it is not possible to work with pre-set standards when planning schistosomiasis control interventions probably because of area-specific differences in exposure and immunity. It clearly follows that cost-effective targeting of interventions would greatly benefit from a better understanding of the interaction between exposure, immunity and infection intensities indicative of disease.

REFERENCES

ALI, M. & BYSLOV, J. (1986). Simplified data collection and analysis in a *Schistosoma mansoni* endemic area. *Tropical Medicine and Parasitology* **37**, 215–19.

ANDERSON, R. M. & MAY, R. M. (1985). Helminth infections of humans: mathematical models, population dynamics and control. *Advances in Parasitology* **24**, 1–101.

ANDERSON, R. M. & MAY, R. M. (1991). *Infectious Diseases of Humans: Dynamics and Control.* Oxford: Oxford University Press.

ANDERSON, R. M. & MEDLEY, G. F. (1985). Community control of helminth infections of man by mass and selective chemotherapy. *Parasitology* **90**, 629–60.

BRESLOW, N. E. & CLAYTON, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* **88**, 9–25.

BUTTERWORTH, A. E., DALTON, P. R., DUNNE, D. W., MUGAMBI, M., OUMA, J. H., RICHARDSON, B. A., ARAP SIONGOK, T. K. & STURROCK, R. F. (1984). Immunity after treatment of human schistosomiasis mansoni. I. Study design, pretreatment observations and the results of treatment. *Transactions of the Royal Society of Tropical Medicine and Hygiene* **78**, 108–23.

CHANDIWANA, S. K. & WOOLHOUSE, M. E. J. (1991). Heterogeneities in water contact patterns and the epidemiology of *Schistosoma haematobium*. *Parasitology* **103**, 363–70.

DE VLAS, S. J. & GRYSEELS, B. (1992). Underestimation of *Schistosoma mansoni* prevalences. *Parasitology Today* **8**, 274–7.

DE VLAS, S. J., GRYSEELS, B., VAN OORTMARSSEN, G. J., POLDERMAN, A. M. & HABBEMA, J. D. F. (1992). A model for variations in single and repeated egg counts in *Schistosoma mansoni* infections. *Parasitology* **104**, 451–60.

DIXON, H. (1985). Statistical methods applicable to schistosomiasis control programmes. *WHO/SCHISTO/85.81, WHO/ESM/85.1.*

DYKE, G. V. & PATTERSON, H. D. (1952). Analysis of factorial arrangements when the data are proportions. *Biometrics* **8**, 1–12.

ELLIOTT, J. M. (1977). *Some Methods for the Statistical Analysis of Samples of Benthic Invertebrates.* Freshwater Biological Association, Scientific Publication No. 25.

GRYSEELS, B. & NKULIKYINKA, L. (1988). The distribution of *Schistosoma mansoni* in the Rusizi Plain. *Annals of Tropical Medicine and Parasitology* **82**, 581–90.

GRYSEELS, B., NKULIKYINKA, K., KABAHIZI, E. & MAREGEYA, E. (1987). A new focus of *Schistosoma mansoni* in the highlands of Burundi. *Annales de la Société Belge de Médecine tropicale* **67**, 247–8.

GUYATT, H. L. & BUNDY, D. A. P. (1991). Estimating prevalence of community morbidity due to intestinal helminths: prevalence of infection as an indicator of the prevalence of disease. *Transactions of the Royal Society of Tropical Medicine and Hygiene* **85**, 778–82.

GUYATT, H. L., BUNDY, D. A. P., MEDLEY, G. F. & GRENFELL, B. T. (1990). The relationship between the frequency distribution of *Ascaris lumbricoides* and the prevalence and intensity of infection in human communities. *Parasitology* **101**, 139–43.

HAGAN, P., BLUMENTHAL, U. J., DUNN, D., SIMPSON, A. J. G. & WILKINS, H. A. (1991). Human IgE, IgG4 and resistance to reinfection with *Schistosoma haematobium*. *Nature, London* **349**, 243–5.

JORDAN, P. (1972). Epidemiology and control of schistosomiasis. *British Medical Bulletin* **28**, 55–9.

KATZ, N., CHAVEZ, A. & PELLEGRINO, J. (1972). A simple device for quantitative stool thick-smear technique in schistosomiasis mansoni. *Revista do Instituto de Medicina tropical de São Paulo* **14**, 397–400.

LENGELER, C., DE SAVIGNY, D., MSHINDA, H., MAYOMBANA, C., TAYARI, S., HATZ, C., DEGREMONT, A. & TANNER, M. (1991). Community-based questionnaires and health statistics as tools for the cost-efficient identification of communities at risk of urinary schistosomiasis. *International Journal of Epidemiology* **20**, 796–807.

LWAMBO, N. J. S., BUNDY, D. A. P. & MEDLEY, G. F. (1992). A new approach to morbidity risk assessment in hookworm endemic communities. *Epidemiology and Infection* **108**, 469–81.

McCULLAGH, P. & NELDER, J. A. (1989). *Generalized Linear Models*. 2nd Edn. *Monographs on Statistics and Applied Probability*, vol. 37. London: Chapman & Hall.

PIERCE, D. A. & SANDS, B. R. (1975). *Extra-Bernoulli Variation in Binary Data*. Technical report no. 46, Department of Statistics, Oregon State University.

RED URINE STUDY GROUP (1994). Identification of high-risk communities for urinary schistosomiasis: a multi-country study. *Social and Economic Research Project Reports. WHO/TDR/SER-Report Series, Geneva.* (In the Press).

SERC (STATISTICS AND EPIDEMIOLOGY RESEARCH CORPORATION) (1991). *EGRET: User's Manual.* Seattle: Statistics and Epidemiology Research Corporation.

TANNER, M. (1989). Evaluation of public health impact of schistosomiasis. *Tropical Medicine and Parasitology* **40**, 143–8.

WILKINS, H. A., BLUMENTHAL, U. J., HAGAN, P., HAYES, R. J. & TULLOCH, S. (1987). Resistance to reinfection after treatment of urinary schistosomiasis. *Transactions of the Royal Society of Tropical Medicine and Hygiene* **81**, 29–35.

WORLD HEALTH ORGANIZATION (1985). The control of schistosomiasis – report of an expert committee. *WHO Technical Report Series* **379**.

WORLD HEALTH ORGANIZATION (1993). The control of schistosomiasis. Second report of the WHO Expert Committee. *WHO Technical Report Series* **830**.