ONLINE FIRST ARTICLE
doi.org/10.26398/IJAS.215

Statistica Applicata - Italian Journal of Applied Statistics

Received: 22.10.22/ **Accepted:** 09.08.24
Published online: 30.11.24

SCHISTOSOMIASIS IN UGANDA: WHAT FACTORSAFFECT MAINLY THE SPREAD OF THE INFECTION?

Francesca Bassi, Department of Statistical Sciences, University of Padova, Italy. ORCID: 0000-0002-3257-7029. francesca.bassi@unipd.it. Corresponding author.

Salvatore Ingrassia, Department of Economics and Business, University of Catania, Italy. ORCID: 0000-0003-2052-4226.

Saint Kizito Omala, Department of Statistical Methods and Actuarial Science, Makerere University, Uganda. ORCID: 0000-0003-4073-5565

Chiara Tognon, Department of Statistical Sciences, University of Padova, Italy.

Abstract. Schistosomiasis represents a heavy burden for developing countries. In particular, the infection is quite widespread in Uganda where it reaches high prevalence in many regions. In this paper, we investigate the factors that mainly influence the probability of contracting the infection based on a dataset formed by 24,918 observations. Data was collected between April and June 2017 through a survey on households in some districts of Uganda. Due to the hierarchical structure of the data, the analysis has been carried out with multilevel regression models. Results show that hygienic conditions and the absence of water resources strongly correlate with the infection's spread.

Keywords: NTD (neglected tropical disease), Africa, risk factors, prevalence

1. Introduction

Schistosomiasis is an infectious disease, also known as bilharzia, and it is caused by a species of parasitic worms of the genus *Schistosoma*. It occupies the third place among tropical infections with the most disastrous effects of mortality and morbidity in developing countries. It only precedes malaria helminthiasis infection (Ahmed, 2020). Transmission occurs through contact with water contaminated by excrements that contain parasite eggs. The larvae open once inside the human host, and the female specimens continue the reproduction by depositing new eggs (WHO, 2021a). Parasites survive, on average, between three and 10 years inside the human body, although they can survive for even 40 years. In these cases, if a patient becomes infected during childhood or adolescence, he can carry the infection up to adulthood, even after the disappearance of medical symptoms. Thus, the reported number of people infected could be underestimated. Furthermore, schistosomiasis can cause comorbidity with other infectious diseases, such as hepatitis, HIV and malaria (Ahmed, 2020). Some studies report that genital *Schistosoma* infection increases the risk of contracting HIV by three/four times (Colley et al., 2014).

From a social point of view, the disease can be disabling: in children, it can cause anaemia, growth problems, malnutrition and impaired cognitive capacity; while in adults, it can have straining economic effects owing to its impact on their ability to work. All daily activities that involve contact with unsafe water constitute a risk factor for the transmission of the infection. Schistosomiasis infections occur among the poorest and most rural communities, where the main occupations are agriculture and fishing, activities that involve assiduous contacts, without any hygienic control, with polluted water which carries the parasites. The Special Program for Research and Training in Tropical Diseases of the WHO has started the development of vaccines for the most diffused and most dangerous viruses, including schistosomiasis (Assad and Torrigiani, 1985).

Uganda is one of the countries where schistosomiasis is endemic; there are areas with a high prevalence of infection among the population. An estimated over four million people have become infected with the disease, and 55% of the Ugandan population is at risk (Loewemberg, 2014).

In recent years, the government of Uganda, with support from some other organizations, has tried to improve health conditions. In particular, the Ministry of Health devised a specific plan to control neglected tropical diseases (Borne and NTDD, 2021). The project involves specialists from various fields, including scientists and experts in communication, technicians, entomologists and analysts, who collaborate to identify the most problematic areas and plan immediate intervention.

In this framework, the present paper aims to investigate the main factors that contributed to the spread of the infection in Uganda. To this end, a dataset comprising 24,918 individuals has been analyzed, with data collected through a survey of households in some districts of Uganda. Due to the hierarchical structure of the data, the analysis has been conducted using multilevel regression models.

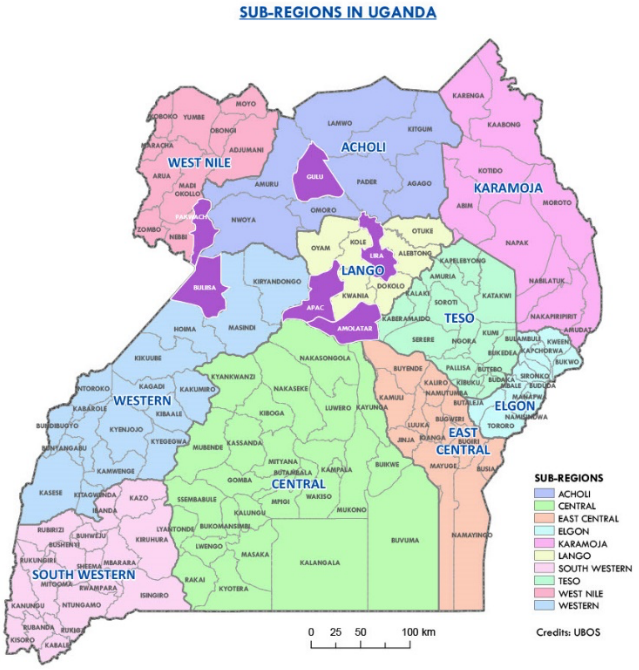
The rest of the paper is organized as follows: In section 2, we provide an overview of the features and geographic distribution of schistosomiasis, also reporting measures which were put in place to limit the occurrence of the infection and outlining the goals set for the future by health authorities. We analyze, in more detail, the case of Uganda, giving an overall picture of the development of schistosomiasis in the country and a brief excursus of the demographic characteristics of the population. In section 3, we describe our dataset, the sampling and collection method, and perform exploratory analyses. In section 4, we introduce multilevel modelling, highlighting the importance of evaluating the inclusion of a hierarchical component in the analyses in the epidemiological field. These models and their properties are described from a statistical point of view. In particular, the multilevel logistic regression model is estimated to identify factors which mostly affect prevalence. In section 5, we report the results of the estimation of the multilevel models. These results show which factors contribute more to increase the risk of infection, considering both within and between groups variability. Section 6 contains a discussion and concluding remarks.

2. Demographic and hygienic framework

Uganda is located in the Central-Eastern Africa. To begin with, we illustrate the demographic and hygienic context in Uganda that influences the spread of the infection of schistosomiasis. The most common type is the *Schistosoma mansoni*, which causes infections at the intestinal level; it is estimated that more than four million people have become infected with the disease and that 55% of the Ugandan population is at risk (Loewemberg, 2014). Across the country, the prevalence in the population is 22%, according to a PMA (2020). The prevalence of the infection varies a lot, and there are communities where positivity reaches 92%, while there are areas where it is almost absent (Exum et al., 2019). The problem mainly concerns rural areas, but it can also affect urban areas, such as the district of Kampala, Uganda's capital city, where the prevalence is 10%. The lethality of the infection is 1.8 over 100,000 inhabitants, according to WHO (2019).

The main reasons for such a high prevalence are lifestyle, habits and the lack of health and hygiene rules. Another problem that contributes to potentially infecting the water is the management of the sewerage system for the disposal of excrement. According to a report by the African Development Fund, in 2017, only 7% of the country was covered by a sewer network (African Development Fund, 2017). The most used system, especially in rural areas but also in the suburbs of large cities, is that of the latrines, which are very often without covers. Frequently, the territorial conformation and the absence of carriage roads make emptying these latrines difficult, making them unusable and, above all, creating problems when the rains carry around the faecal material in excess. For this reason, latrines are very often emptied illegally in the streams of water along the roadsides, favouring the proliferation of bacteria and parasites.

Poor health conditions are also associated with a lack of basic hygienic standard practices, such as washing hands with soap after going to the bathroom; in 2014, only 29% of the population followed this habit (Loewenberg, 2014).



Credits: UBOS

Figure 1. Districts selected for the survey (in purple)

2.1 Demographic and family overview

To better understand the problem, we present here a brief overview of Uganda's main demographic and family characteristics. According to the data collected through the Uganda Demographic and Health Survey (Uganda Bureau of Statistics, 2021), the population amounts to 41 million, and 27% of them live in urban centres. The territory is divided into five regions and 15 sub-regions, each divided into districts and parishes (Figure 1). Buganda (Central 1-Central 2) is the most inhabited region, with nearly 10 million residents (23% of the Ugandan population). Regions also differ by household size; in Teso, a rural area, for example, the average number of household members is 5.9. In Kampala, it is lower, equal to 3.4. The population is very young (54% under 18 and 4% over 60), and life expectancy is 63 years (United Nations, Department of Economic and Social Affairs, Population Division, 2019). In some areas (South-East and Kampala), the literacy rate is over 80%; in the North, only slightly over 30%. An important portion of the population does not have access to health facilities in case of illness, especially in rural areas, either due to the distance, the costs, or because they do not consider it necessary. As for the work situation, most of it is related to the primary sector: 68% of the population is engaged in agriculture, fishing and forestry. The data from the survey reported median monthly earnings for an employed person of UGX 200,000 (about 50 euros); 30% of households live below the poverty line, which means less than US$ 1.77 per person per month. The housing situation is an important indicator for assessing the conditions of life of the population. The fact that there are certain in-household services and features has a major impact on the health and well-being of individuals. Fundamental to guaranteeing basic sanitary conditions is the state and the typology of the toilets. The form that is commonly used is the latrine (83% of cases), while 7% of households do not have any available toilet. The toilets equipped with a drain, which guarantees greater cleanliness and hygiene, are present in only 3% of homes. In particular, in the Northeast region of Karamoja, 70% of the population lives in houses without any form of toilet. Personal cleaning services are also not guaranteed in most of the country: 82% of households do not have tools for washing hands, and only 9% of cases are there running water for washing. As for the availability of drinking water, access to clean and safe sources is guaranteed to 79% of the population. In general, the sources of water are at a distance of less than three km from the house, and only 8% have to travel a longer distance. In this context, it is evident, especially in certain areas, that there are favourable conditions for a high incidence of the infection of schistosomiasis.

As stated above, schistosomiasis prevalence in the country is greater than 22%; however, due to the many socio-economic and cultural differences among districts just described, prevalence varies greatly across the territory; moreover, we expect differences also in the factors – and their magnitude - affecting the risk of infection.

3. Explorative data analysis

The data analyzed in this paper were collected between April and June 2017, with a survey of households in some districts of Uganda, some located in the North-central area of the country and the other two, Buliisa and Pakwach, as indicated in Figure 1.

The data was collected by the Ugandan Bureau of Statistics (UBOS). A multistage cluster sampling design was used to select households (Turner, 1996), which is specific for situations where resources are scarce, and there is no sampling frame of the target population (Milligan et al., 2004). This procedure solves some of the problems of the classic Expanded Program on Immunization (EPI) sampling method used by the WHO for surveys in developing countries. The EPI method is based on information available from the most recent census. Areas are selected with probabilities proportional to the number of inhabitants; this last information is, however, often inaccurate in regions where the rate of population growth, even in short periods, is very high. Households in the sampled areas are selected by the interviewer using a non-random sampling procedure (Brogan et al., 1994). In the multistage cluster sampling method, sampling areas are determined with probability proportional to the population counted in the most recent census; the difference consists in how households are selected. Using maps, sampled areas are divided into smaller segments, containing more or less the same number of houses. A sample of segments is randomly selected, and all households are included in cluster sampling.

With reference to this specific survey, 20 areas corresponding to the parishes (municipalities) within the districts were selected with probability proportional to the population. Every parish was then divided into segments with around 30 households. Interviews were administered to each head of the selected households; questions regarding schistosomiasis with reference to all household members were posed, and information on household members and characteristics of households was also collected. The achieved sample comprised 24,918 individuals from a total of 3,966 households.

Table 1. Variables collected in the survey: brief description

Name	**Description**
Second-level variables	
District	There are seven districts (6, 14, 16, 20, 21, 22, 24)
Parish	Each district contains 20 parishes
Members	Number of family components
Watersource	There are 8 categories: pipe at residence, public water tap, deep well, hand pump well, open well, river/stream, lake, spring.
Timesource	Time to reach water: less than 30 minutes, 31-60 minutes, 61-90 minutes, more than 60 minutes
Wateryear	Binary variable indicating if water is available all year or not.
Bicycle	Binary variable indicating if the family owns a bicycle.
Radio	Binary variable indicating if the family owns a radio.
Penpres	Binary variable indicating if the family owns a pen for cows.
Toilet	Type of toilet present in household: no, shared, family use only, public, other
Garbage	Binary variable indicating if there is garbage near the house.
Feces	Binary variable indicating if there are excrements near the house.
First-level variables	
Age	There are 8 categories in years: <5, 5-9, 10-14, 15-19, 20-29, 30-39, 40-49, >49.
Sex	2 categories: male, female.
Marital	5 categories: single, married, widowed, divorced, separated.
Education	5 categories: no education, primary level, secondary education, diploma, university.
Nourished	Binary variable indicating if the person is malnourished.
Extended-sprain	Binary variable indicating if the person has abdominal sprain.
Abdominal-pain	Binary variable indicating if the person has abdominal pain.
Rash	Binary variable indicating if the person has skin rash.
Canread	Binary variable indicating if the person can read.
Relation	Relation with head of household

Table 1 describes the variables used for the analyses; a few others were discarded because of too many missing data. The questionnaires administered in the district of Buliisa (number 20) presented very large percentages of missingness, therefore, we eliminated the corresponding records from the dataset. A few other records collected in other districts were eliminated either for missing data in many variables or for irrelevance, for example, data on district 14, where only five interviews were conducted. In cases of a percentage lower than 1 of missingness, we proceeded with imputation[1]. We ended up with 21,029 observations on individuals nested in 3,280 households, distributed in five districts as reported in Table 2, which also shows the prevalence of schistosomiasis in the districts and the overall population: we can observe great heterogeneity across the territories (confirmed by a Chi-squared test with p-value lower than 0,001).

Table 2. Individuals and household distribution in districts and prevalence of schistosomiasis

District	Individuals		Households		Prevalence
	n	%	*n*	%	%
6 *Pakwach*	4,050	18.39	666	20.30	33.83
16 *Gulu*	3,868	20.87	574	17.50	8.79
21 *Apac*	4,389	20.47	667	20.34	15.13
22 *Lira*	4,305	21.00	682	20.79	10.59
24 *Amolatar*	4,417	19.26	691	21.07	13.79
Total	21,029	100	3.280	100	16.36

Table 3 reports the prevalence of the disease by households' and members' characteristics; only statistically significance associations (Chi-squared test) are reported. We can see that the occurrence of the illness is positively associated with many characteristics of the individual and the household. Females, are more likely to get infected. The high heterogeneity observed for prevalence in the five districts may be due to the fact that families have different ways of life and different household's facilities in the different areas of the country. In Table 3, we also register the prevalence in the case of symptoms that are considered typical of schistosomiasis. The status of these symptoms is positively correlated to the prevalence (confirmed by a Chi-squared test; *p*-values are reported in Table A.1 together with Cramer V test).

Table 3. Prevalence by individuals and households' characteristics

Characteristic	Category	Prevalence %	Characteristic	Category	Prevalence %
Sex	Male	15.76	Radio	Radio	12.68
	Women	16.91		No radio	21.27
Canread	Able to read	12.19	Bicycle	Bicycle	13.87
	Unable to read	18.94		No bicycle	20.42
Age	Age <5	16.29	Water source	Deep well	8.67
	Age 5-9	16.57		Hand-pump well	18.05
	Age 10-14	15.43		Lake	7.78
	Age 15-19	18.04		Open well	13.17
	Age 20-29	13.68		Pipe	4.44
	Age 30-39	15.21		Public water tap	8.15
	Age 40-49	17.35		River	36.90
	Age >49	21.72		Spring	10.59
Education	No education	20.85	Water in a year	Water all year	14.27
	Primary education	12.70		Water not all year	23.18
	Secondary education	9.84	Time to water source	minutes to water <31	15.56
	Certificate/Diploma	10.51		minutes to water 31-60	15.39

[1] For some variables it was possible to make imputations of missing values with the following rules for each specific variable:
"Totmembers": has been imputed with the number of the members corresponding to the same "idnumber";
"Marital": for all those under the age of 16 it was assumed that they were not married;
For the other variables, we studied the distribution of the missing values within families. If data was missing for one variable for all the components, a decision was made to eliminate the observations relating to the entire household; otherwise, if the absence of information was only associated with a family member, the household was retained in the analysis.
Other missing data was treated during estimation as missing at random.

	University degree and above	9.76		minutes to water >60	21.73
Waste management	Garbage near house	23.58	Rash	Rash	36.27
	No garbage near house	8.08		No rash	13.09
Faeces	Excrements near house	32.36	Abdominal-pain	Abdominal pain	28.54
	No excrement near house	6.86		No abdominal pain	11.38
Toilet	No toilet	18.87	Extended-spleen	Extended sleen	31.51
	Shared toilet	15.00		No extended spleen	12.98
	Household toilet	10.84	Nourished	Malnourished	23.07
	Public toilet	9.00		Nourished	12.92
	Other	25.32			

4. Multilevel modelling in epidemiology

Multilevel models have always been applied, particularly in economic and social studies, where we often deal with hierarchical data (Guo and Zhao, 2000). Using the multilevel approach, it is possible to take into account the fact that observations might not be independent because they are nested in higher-level units. Traditional methods of statistical inference assume that observations are independent, but this is not necessarily true in hierarchical structures as in the study sample. Multilevel modelling accounts for eventual correlation among first-level units.

In epidemiological studies, the multilevel approach has historically been used less, as the focus of the investigations is usually on individual risk factors that influence the occurrence of a disease or on aggregated data (Weinmayr et al., 2017). In this case, the idea is that the individual's aspects alone are capable of explaining the causes of a disease (Diez Roux and Aiello, 2005), and it is not allowed to simultaneously consider individual and group effects on the risk of contracting the disease. More recently, the interest in multilevel analysis has also increased in epidemiological studies in order to include the possibility of variability both within and between groups.

An example in which the importance of the multilevel component is evident is the study on the propensity to smoke in adolescents, which may depend on how widespread smoking is in the group of peers over individual characteristics (Diez Roux and Aiello, 2005). Another example is that of HIV or other sexually transmitted diseases, where the transmission rate can be determined by social norms and behavioural habits of the group to which individuals belong, not only by individual behaviour (Diez Roux and Aiello, 2005).

In the epidemiological context, there is a considerable number of diseases that have a highly variable prevalence in different geographical areas. In these cases, it is important to consider the effects of contextual factors on individual health outcomes (Weinmayr et al., 2017). This approach also makes it possible to plan interventions both at the individual and community level while not considering the subdivision into groups, and there is an actual risk of drawing incorrect conclusions. Taking into account the hierarchical structure of the data considers that observations are not independent (Guo and Zhao, 2000); assuming independence, in this case, would result in biased estimates. Multilevel models permit estimation of the impact of covariates at the different levels; moreover, total variance can be decomposed in order to assess how much variability in the data is due to within and between groups factors.

3.1. Multilevel regression model

Equation (1) describes a multilevel regression model,

$$y_{ij} = \beta_{0j} + \beta_{1j} x_{ij} + \varepsilon_{ij} \qquad (1)$$

y_{ij} is a dependent variable, observed on unit i (first level), belonging to group j (second level), $i=1,..,n_j, j=1,...,J$; x_{ij} is a first-level covariate, and ε_{ij} is a random error with distribution $N(0,\sigma^2)$. β_{0j} and β_{1j} are, respectively, the random intercept and the random slope.

In the case of the random intercept model, we can calculate the intra-class correlation coefficient (ICC), which is a measure of the proportion of total variance explained by between groups variability:

$$ICC = \frac{\tau_0^2}{\sigma^2 + \tau_0^2}.$$

A low value of the ICC indicates that between groups variability is low; therefore, it is not necessary to estimate hierarchical models. The minimum value of ICC necessary to perform multilevel modelling depends on sample size and other data characteristics (Musca et al., 2011).

In epidemiology, we frequently deal with a binary response variable to indicate the presence or absence of a disease. Indicating with p_{ij}=P(y_{ij}=1), a random intercept multilevel logistic regression model is defined by equation (2) (Merlo et al., 2016):

$$logit(p_{ij}) = \gamma_0 + \sum_{h=1}^{r} \gamma_h x_{hij} + U_{0j} \qquad (2)$$

where x_h, with $h=1,..,r$, are first-level covariates; γ_0 is the mean of p_{ij} in the logistic scale and, in epidemiology, it can be seen as prevalence (Snijders and Bosker, 1999). In this case, the ICC has the following form:

$$ICC = \frac{\tau_0^2}{\frac{\pi^2}{3} + \tau_0^2}.$$

In the case of the multilevel logistic regression model, an alternative measure of between groups variability is the Median Odds Ratio (MOR):

$$MOR = \exp\left(\sqrt{2\tau_0^2}\phi^{-1}(0.75)\right) = \exp\left(\sqrt{2\tau_0^2} \cdot 0.6745\right) \cong \exp\left(\sqrt{\tau_0^2} \cdot 0.95\right).$$

where ϕ^{-1} is the 75th percentile of the standard Normal distribution. This indicator, such as the ICC, can be used to compare alternative models. It assumes values ≥1; being equal to 1 in the case of no between groups variability.

5. Numerical results

In order to study which factors have the greatest effects in determining the probability of contracting the infection, we estimated logistic regression models. The classic logistic model (no multilevel) is used as the baseline to evaluate the improvement introduced by a multilevel component in the analysis. Different multilevel regression models were tested, selecting the best in terms of the lowest value of the ICC and MOR. For now, we focus on the results of the analysis, while the numerical details of fit are given in the Appendix.

In the ordinary logistic regression model (that is, no multilevel) for evaluating the probability of contracting the infection, we considered the following predictors: symptoms and individual and household characteristics; a few variables, like extended spleen and bicycle ownership, were eliminated due to multicollinearity issues.

The results showed that living in different districts has a significant effect. For example, in reference to district 6, the other districts exhibited a decrease in the probability of becoming ill. Household factors that have a statistically significant effect on the probability of contracting the illness are the type of water source, specifically when it is a river or stream; the walking time necessary to reach the water source: the longer the time, the higher the risk; the absence of a toilet in the house; the presence of garbage and excrements near the house; and the household size.

Among the individual level effects, we observe that age increases the probability of having the disease, as well as the fact that the individual is unable to read. Quite surprisingly, possession of symptoms has no significant effect.

We now proceed to evaluate whether the insertion of a second-level component, i.e., considering the hierarchical structure of our data, improves the model. As already stated, districts in Uganda differ for many characteristics that affect individuals and households, and the prevalence of schistosomiasis shows a great variability across the country. The ordinary logistic regression model assumes that individuals who make up the sample belong to the same population. Instead, it is reasonable to assume that there are differences between districts that generate a correlation structure in the variables collected on household located in the same area. The multilevel logistic regression model estimates the effects of covariates on prevalence, taking into account this correlation structure.

To this end, we estimated, as a starting point, a logistic random intercept model without covariates (unconditional) with the probability of contracting schistosomiasis as the dependent variable. This model takes into account the idea that individuals are nested within households and therefore, observations on members of the same family may be correlated. The values of the ICC and the MOR indicate significant association within the groups; therefore the multilevel approach is appropriate for analyzing these data (see Table A.2). Further, we insert covariates that might affect prevalence and improve model fit and having an effect on the variability explained by the model.

Table 4. Conditional logistic random intercept model for the probability of contracting schistosomiasis: estimation results

Variable	Estimate	Standard error	p-value	Odds-ratio
Constant	-4.63	0.43	0.00*	
District 6 ref.				
16	-2.59	0.26	0.00*	0.07
21	-1.40	0.25	0.00*	0.25
22	-1.83	0.26	0.00*	0.16
24	-1.22	0.24	0.00*	0.29
Timesource <31 minutes ref.				
31-60 minutes	0.41	0.14	0.00*	0.21
>60 minutes	1.98	0.18	0.00*	0.52
Watersource hand-pump well ref.				
Deep well	-1.15	0.14	0.02*	0.32
Lake	-1.45	0.18	0.00*	0.23
Open well	-0.23	0.19	0.23	
Pipe at residence	-0.49	0.69	0.48	
Public water tap	0.12	0.38	0.74	
River/stream	0.31	0.27	0.26	
Spring	-0.11	0.25	0.44	
Wateryear	0.04	0.15	0.80	
Toiliet family ref.				
No	0.75	1.47	0.61	
Shared	0.58	1.40	0.68	
Public	-0.04	0.62	0.95	
Other	-0.78	0.17	0.00*	0.46
Radio	0.26	0.16	0.10	
Garbage	0.39	0.15	0.01*	1.47
Feces	2.19	0.17	0.00*	8.97
Members	0.23	0.03	0.00*	
F_age	1.03	0.01	0.00*	
F_canread	-1.87	0.30	0.00*	
Nourished	-0.03	0.13	0.81	
Rash	-0.06	0.13	0.63	
Abdominal-pain	0.08	0.07	0.31	
Canread	-0.17	0.10	0.09	
Education 1 ref.				
2	0.00	0.09	0.99	
3	-0.13	0.17	0.44	
4	0.33	0.28	0.23	
5	-014	0.41	0.72	
Marital 1 ref				
2	-0.09	0.12	0.70	
3	-0.07	0.17	0.93	

4	-0.12	0.19	0.78	
5	-0.08	0.56	0.69	
Sex male ref.	0.06	0.07	0.32	
Relation 1 ref.				
2	-0.05	0.12	0.70	
3	-0.02	0.17	0.93	
4	-0.05	0.19	0.78	
5	0.22	0.56	0.69	
Age in years	1.00	0.00	0.99	
τ^2_0	6.78	0.45	0.00*	
Model fit	ICC=0.6732	MOR=11.45	AIC=12,503	

* statistically significant at 5%.

Table 4 contains the estimates of the best fitting logistic random intercept model with covariates (conditional); covariates have been selected on the basis of the reduction of the ICC and MOC; alternative models have been compared with the AIC index.

With respect to the traditional logistic model, the multilevel specification has a better fit, showing a lower AIC value. Statistically significant estimated coefficients have the same sign in the two models, but different magnitude. Some second-level variables revealed statistically insignificance when considering the hierarchical structure of the data (all levels of the variable describing the type of toilet, except "others", and if water is available all year). The effects of the other second-level variables show larger estimates in the multilevel model. We inserted two new second-level variables, measuring the average age of household members (*F_age*) and the percentage of household members who can read (*F_canread*); these variables are correlated with the occurrence of schistosomiasis. In the multilevel approach, in general, first-level variables result appears less important than in the traditional logistic model.

Table 4 shows also the estimated odds-ratio for the statistically significant levels of categorical variables. These odds-ratios allow to understand better the effect of each covariate and its levels on the risk of schistosomiasis. The probability of contracting schistosomiasis increases by 47% if there is the presence of garbage near the house, and it is almost nine times higher in presence of feces. Risk of disease increases also with the distance to the water source, being 50% greater when the required time is between 30 and 60 minutes, and almost three times bigger for distances that require more than one-hour journey. Getting water from a deep well or a lake decreases the risk of infection with respect to the use of a hand pump. People leaving in districts different from Pakwach (number 6) have a lower risk of contracting the illness. From the statistically significant continuous variables, we see that one additional family member increases the risk by 25%; one year change in household members' average age increases the risk by 3%; a higher proportion of household members who can read decreases the risk.

Finally, although the multilevel logistic regression model improves the fit to the data explaining part of within-groups correlation, still the ICC and the MOR have high values, indicating that it is necessary to specify an even better model to adequately identify the factors affecting schistosomiasis. In particular, the results obtained with the logistic multilevel regression model indicate that this is a sort of household disease since many characteristics of the household and the its location appear as important. Moreover, it is very plausible that occurrence of the disease tends to happen within members of the same family. In our dataset, on average, in families where at least one member suffers from schistosomiasis, 50% of household members are ill. For these reasons, we decided to explore which conditions may cause the fact that at least one family member is infected. The following analyses, then, focus on the probability of observing at least one infection in the household. With this scope, we considered observations on 3,280 families, who become our first-level units. Information at individual level is aggregated with reference to the corresponding household. Second-level units are the 96 parishes or municipalities in which the families live.

Table 5 lists estimation results of a logistics multilevel regression model for the probability of observing at least one member of the family affected by schistosomiasis; only significant estimates are reported. We created some new variables summarizing characteristics of parishes: average number of family components (*P_members*), proportion of families who need to move for at least one hour to reach water (*P_timesorce60*), proportion of families with garbage and excrements near the house (*P_garbage*, *P_feces*), average age of family members (*P_age*). With reference to these second-level variables, we inserted in the model also the contextual effects; i.e., the deviations from the group mean for each family (*P_timesorce60_c*, *P_members_c*, *P_garbace_c*, *P_feces_c*) as suggested by Feaster et al. (2011).

Table 5. Conditional logistic random intercept model for the probability of observing at least one infected member in the family: estimation results

Variable	Estimate	Standard error	p-value	Odds-ratio
Constant	-4.91	1.26	0.00^*	
P_members	0.23	0.02	0.00^*	1.77
P_timesource_60	2.39	0.84	0.00^*	10-94
P_garbage	1.65	0.59	0.00^*	5.20
District 6 ref.				
16	-2.10	0.43	0.00^*	0.12
21	-1.08	0.39	0.01^*	0.34
22	-1.14	0.42	0.00^*	0.32
24	-0.66	0.40	0.10	
Watersource hand-pump well ref.				
Deep well	-0.78	0.36	0.03^*	0.46
Lake	-0.92	0.42	0.03^*	0.40
Open well	-0.46	0.18	0.01^*	0.63
Pipe at residence	0.34	0.53	0.51	
Public water tap	0.09	0.35	0.80	
River/stream	-0.10	0.31	0.74	
Spring	-0.21	0.23	0.37	
Toiliet family ref.				
No	0.50	1.11	0.65	
Shared	0.52	0.99	0.60	
Public	0.05	0.45	0.91	
Other	-0.65	0.14	0.00^*	0.52
Feces	0.59	0.13	0.00^*	1.81
F_age	0.02	0.01	0.03^*	1.02
P_timesource_60_c	0.45	0.13	0.00^*	1.56
P_members_C	0.23	0.02	0.00^*	1.25
τ^2_0	1.00	0.20	0.00^*	
Model fit	ICC=0.2333	MOR=2.60	AIC=3,433	

* statistically significant at 5%.

The risk of observing at least one infected person in the family decreases by 88%, 66% and 68% respectively, living in districts 16, 21, 22, with respect to district 6. Families who take water from lakes or wells show lower risks with respect to families using hand pumps; also the type of toilet access, specifically respondents that has a toilet used only a household, diminishes the risk. The risk of at least one infection in the family, on the other hand, increases with the average age of family members and excrements near the house. With reference to the parish where the family lives, factors that positively affect the risk are the average number of members in each household, the proportion of families who have a travel time greater than one hour to reach the source of water and the proportion of families in the parish who have garbage around the house. Estimation results, in this case, show that factors affecting schistosomiasis in families are related both to the characteristics of the household and of the parish where the family lives. Statistically significant second-level variables indicate that there is non-negligible heterogeneity between parishes.

6. Conclusions

In this paper, we study factors affecting the diffusion of schistosomiasis in some districts of Uganda, taking into account the hierarchical nature of the phenomenon: infected individuals belong to households, which are clustered in parishes. As recognized by the reviewed literature (Diez Roux and Aiello, 2005), in epidemiological studies, it is important to consider the multilevel structure of the data in order to obtain reliable estimates. Observations on members of the same household are correlated, as well, as observations on households living in the same parish; not considering this correlation in model specification, gives rise to biased estimates.

Estimating logistic random intercept models with our data allows us to correctly evaluate the effects on contracting schistosomiasis of all variable-levels: characteristics of individuals, families and parishes; moreover, total variance can be appropriately decomposed across within and between-groups parts.

Statistical analyses of our data identified significant associations between the prevalence of schistosomiasis and individual characteristics; however, the effects of individual traits disappeared in multilevel estimation in favour of family characteristics, especially relating to the hygienic conditions in and around the house. This result is consistent with other studies from the reference literature that identify poor sanitation as one of the main risk factors for infection (see, for example, Colley et al., 2014). Another family-related important risk factor emerging from our analyses is the contact with polluted water, as also indicated by WHO (2021), especially those families who have a long journey to reach the water source are more exposed to schistosomiasis. These results highlight the need to strengthen public information campaigns as both WHO and the Ugandan Ministry of Health are preoccupied with the prevention of the occurrence of schistosomiasis (Borne and NTDD, 2021; WHO, 2021b). These campaigns should be especially targeted for families with the highest illiteracy rates, who are also those most at risk of infection. Another significant and positive effect on the risk of infection is due to the mean age of family members, which can be explained by referring to the results in the literature on the longevity of the infection in the human body (Colley et al., 2014). Usually, schistosomiasis is contracted at a young age, and then it remains in the human body for an important part of adult life. In endemic areas, such as in the case of districts of Uganda that we analyzed, the most widespread form of schistosomiasis has a chronic characteristic due to repeated contact with infected larvae. In larger families, the risk of getting sick increases.

When estimating the logistics random intercept model for the probability of at least one infection in the family, other important aspects emerged regarding the area where the family lives: there is evidence of high heterogeneity between the different municipalities or parishes. Again, hygienic conditions in the area are strictly linked to the risk of infection. In particular, the risk is higher in areas where there is a greater proportion of households with garbage near the house or more distant from water sources. Areas with larger families also show a higher risk of schistosomiasis in the household. An important result emerging from all models is the difference between the district of Pakwach (number 6) and the others. Pakwach district has a different geographical location, being in the region of the West Nile, near the White Nile, while the others are in the central belt of the country. In Pakwach, fragility and backwardness conditions regarding developing countries emerge more sharply (UBOS, 2021). In fact, in this area, we find the highest percentages of houses with faeces or garbage around, illiteracy rates, and malnourished individuals.

For what concerns future developments of this research, it might be of some interest to estimate random slope models. Our multilevel logistic regression models assumes that the models describing prevalence in the districts have all the same slope; we allow for randomness only for the intercept. Incorporating random variation also in the slope might increase model fit.

Referring to the preventive treatments that are carried out in the areas where the disease is endemic, in a future study, it could be interesting to use this information, if available, to analyze the impact of the treatment on the probability of contracting schistosomiasis on the individual and the whole family unit.

Declarations

Ethical Approval

Not applicable

Competing interests

Not applicable

Authors' contributions

The authors equally contributed to the analyses and the manuscript.

Funding

Not applicable

Availability of data and materials

Not applicable

References

African Development Bank Group (2017), *Program: Kampala Sanitation Program-Phase 1 Country: Uganda*, https://projectsportal.afdb.org/dataportal/VProject/show/P-UG-E00-008, accessed 14/01/22.

Ahmed, S.H. (2020), Schistosomiasis (Bilharzia), *Medscape*, https://emedicine.medscape.com/article/228392-overview, accessed 14/01/22.

Assaad, F., Torrigiani, G. (1985), Who's vaccine development programme. *European Journal of Epidemiology*, Vol.1, pp. 1–4. https://doi.org/10.1007/BF00162305

Borne V. and Neglected Tropical Diseases Division, Ministry of Health, Uganda (2021), *Sustainability Plan for Neglected Tropical Diseases Control Program 2020–2025*, https://www.health.go.ug/cause/sustainability-plan-for-neglected-tropical-diseases-control-program-2020-2025/, accessed 14/01/22.

Brogan D, Flagg EW, Deming M, Waldman R. (1994), Increasing the accuracy of the Expanded Programme on Immunization's cluster survey design. *Annals of Epidemiology*, Vol. 4, pp. 302–311.

Casulli A (2021), New global targets for NTDs in the WHO roadmap 2021–2030, *PLOS Neglected Tropical Diseases*, 15(5): e0009373, https://doi.org/10.1371/journal.pntd.000937.

Cohen J. (2016), Unfilled Vials, *Science*, Vol. 351(6268), pp. 16-19

Colley D.G., Bustinduy A.L., Secor W.E. and King C.H. (2014), Human schistosomiasis, *Lancet*, Vol. 383 (9936), pp. 2253–2264.

Diez Roux A.V. and Aiello A.E. (2005), Multilevel Analysis of Infectious Diseases, *Journal of Infectious Diseases*, Vol. 191 Suppl 1, pp. S25–S33.

Exum N.G., Kibira S.P.S., Ssenyonga R., Nobili J., Shannon A.K., et al. (2019), The prevalence of schistosomiasis in Uganda: A nationally representative population estimate to inform control programs and water and sanitation interventions, *PLoS Neglected Tropical Diseases*, Vol. 13(8), e0007617, https://doi.org/10.1371/journal.pntd.0007617.

Feaster D., Brincks A., Robbins M. and Szapocznik J. (2011), Multilevel models to identify contextual effects on individual group member outcomes: a family example, *Family Process*, Vol. 50(2), pp. 167–183.

Guo G. and Zhao H. (2000), Multilevel Modeling for Binary Data, *Annual Review of Sociology*, Vol. 26, pp. 441-462.

Lee V. E. and Bryk A. S. (1989), A multilevel model of the social distribution of high school achievement, *Sociology of Education*, Vol. 62(3), pp.172–192.

Loewenberg S. (2014), Uganda's struggle with schistosomiasis, *Lancet*, Vol. 383(9930), pp. 1707–1708.

Merlo J., Wagner P., Ghith N., and Leckie G. (2016), An Original Stepwise Multilevel Logistic Regression Analysis of Discriminatory Accuracy: The Case of Neighbourhoods and Health, *PLoS One*, Vol. 11(4), e0153778, https://doi.org/10.1371/journal.pone.0153778-

Milligan P., Njie Al. and Bennett S. (2004), Comparison of two cluster sampling methods for health surveys in developing countries, *International Journal of Epidemiology*, Vol. 33, pp. 469–476.

Musca S., Kamiejski R., Nugier A., Méot A., Er-rafiy A., Brauer M. (),Data with Hierarchical Structure: Impact of Intraclass Correlation and Sample Size on Type-I Error, *Frontiers in Psychology*, Vol. 2, Doi: 10.3389/fpsyg.2011.00074.

PMA (2020) https://www.pmadata.org/sites/default/files/data_product_results/PMA2020-Uganda-R1-Sch-brief.pd (accessed 4 August 2022).

Ross R. (1916), An application of the theory of probabilities to the study of a priori pathometry: part I, *Proceedings of the Royal Society Series A*, Vol. 92, pp. 204–30

Snijders T. and Bosker R. (1999), *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling (1 ed.)*, SAGE Publications, New York.

Uganda Bureau of Statistics (UBOS) (2021), *Uganda National Household Survey 2019/2020*, Kampala, Uganda.

United Nations, Department of Economic and Social Affairs, Population Division (2019), *World Population Prospects 2019*, Online Edition, Rev. 1, https://population.un.org/wpp/, accessed 14/01/22.

van der Werf M.J., de Vlas S.J., Brooker S., Looman C.W., Nagelkerke N.J., Habbema J.D. et al. (2003), Quantification of clinical morbidity associated with schistosome infection in sub-Saharan Africa., *Acta tropica*, Vol. 86(2-3), pp. 125–139

Weinmayr G.,Dreyhaupt J., Jaensch A., Forastiere F. and Strachan D.P. (2017), Multilevel regression modelling to investigate variation in disease prevalence across locations, *International Journal of Epidemiology*, Vol. 46(1), pp. 336–347.

World Health Organization (2019), *Global Health Estimates 2019: Deaths by Cause, Age, Sex, by Country and by Region, 2000−2019*, Geneva, https://www.who.int/data/gho/data/themes/mortality-and-global-health-estimates/ghe-leading-causes-of-death, accessed 14/01/22.

World Health Organization, The Global Health Observatory (2020), *Schistosomiasis: Status of Endemic Countries*, www.who.int/data/gho/data/themes/topics/schistosomiasis, accessed 14/01/22.

World Health Organization (2021a), *Schistosomiasis*, www.who.int/news-room/fact-sheets/detail/schistosomiasis, accessed 14/01/22.

World Health Organization (2021b), *Ending the Neglect to Attain the Sustainable Development Goals: a Road Map for Neglected Tropical Diseases 2021–2030. Overview*, Geneva, https://www.who.int/publications/i/item/9789240010352, accessed 14/01/22.

Appendix

Table A.1. p-values of the Chi-square and Cramer V test for association between each variable and prevalence

	Chi-square	Cramer V
Variable	p-value	
District	<0.001	<0.001
Parish	<0.001	<0.001
Members	<0.001	<0.001
Watersource	<0.001	<0.001
Timesource	<0.001	<0.001
Wateryear	<0.001	<0.001
Bicycle	<0.001	<0.001
Radio	<0.001	<0.001
Penpres	<0.001	0.000
Toilet	<0.001	<0.001
Garbage	<0.001	<0.001
Feces	<0.001	0.000
Age	0.011	0.012
Sex	0.025	0.087
Marital	0.115	0.880
Education	<0.001	<0.001
Nourished	<0.001	<0.001
Extended-sprain	<0.001	<0.001
Abdominal-pain	<0.001	<0.001
Rash	<0.001	<0.001
Canread	<0.001	<0.001
Relation	<0.001	<0.001

Table A.2. Unconditional logistic random intercept model: estimation results

	Estimate	Standard error	p-value
Constant γ_0	-4.01	0.12	0.00^*
τ^2_0	13.01	0.85	0.00^*
Model fit	ICC=0.7989	MOR=11.98	AIC=13,092