medRxiv preprint doi: https://doi.org/10.1101/2021.08.20.21262368; this version posted August 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Assessing physical and environmental predictors of bovine *Schistosoma japonicum* infection in rural China

Short title: Predictors of bovine schistosomiasis in rural China

Elise Grover[1], Sara Paull[1], Katerina Kechris[2], Andrea Buchwald[1,3], Katherine James[1], Yang Liu[4], Elizabeth J. Carlton[1*]

[1] Department of Environmental and Occupational Health, Colorado School of Public Health, University of Colorado Anschutz Medical Campus, Aurora, USA

[2] Department of Biostatistics and Informatics, Colorado School of Public Health, University of Colorado Anschutz Medical Campus, Aurora, USA

[3] Center for Vaccine Development and Global Health, University of Maryland School of Medicine

[4] Institute of Parasitic Diseases, Sichuan Center for Disease Control and Prevention, Chengdu, China

***** Corresponding author

Email: elizabeth.carlton@cuanschutz.edu

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2021.08.20.21262368; this version posted August 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Abstract

Background

Bovines have been repeatedly highlighted as a major reservoir for human *Schistosoma japonicum* infection in rural farming villages in China. However, little is known about the individual and environmental risk factors for bovine schistosomiasis infection. The current body of literature on individual-level risk factors features inconsistent, and sometimes contradictory results, and to date, few studies have assessed the broader environmental conditions that predict bovine schistosomiasis.

Methodology/Principal Findings

Using data collected as a part of a longitudinal study in 39 rural villages in Sichuan, China from 2007 to 2016, we aimed to identifying the strongest individual, household and village-level predictors of bovine *S. japonicum* infection. Candidate predictors for this assessment included: 1) physical/biological characteristics of bovines, 2) potential human sources of environmental schistosomes, 3) socio-economic indicators, 4) potential animal reservoirs, and 5) agricultural risk factors. A Random Forests machine learning approach was used to determine which of our candidate predictors serve as the best predictors of bovine schistosomiasis infection in each survey year. Of the five categories of predictors, high-risk agricultural practices and animal reservoirs, specifically, bovine density at the village-level, were repeatedly found to be among the top predictors of bovine *S. japonicum* infection.

Conclusion/Significance

Our findings highlight the potential utility of presumptively treating bovines residing in villages and households that engage in high-risk agricultural practices, or bovines belonging to villages with particularly high levels of bovine ownership. Additionally, village-level predictors were stronger predictors of bovine infection than household-level predictors, suggesting future investigations and interventions may need to apply a broad ecological lens in order to successfully extricate and address environmental sources of ongoing transmission.

medRxiv preprint doi: https://doi.org/10.1101/2021.08.20.21262368; this version posted August 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Author Summary

Schistosomiasis is a burdensome global disease that is frequently transmitted between humans and animals. The parasite that causes schistosomiasis is released into water by snails that become infected via contact with eggs from human or animal feces, allowing other human and animal hosts to become infected when they come in contact with contaminated water. In China, bovines are believed to be the most common animal source of human infections, though little is known about what factors promote bovine infections. Because schistosomiasis is a sanitation-related, water-borne disease transmitted by many animals, we hypothesized that several environmental factors – such as the lack of improved sanitation systems, or participation in agricultural production that is water or fertilizer-intensive – could promote schistosomiasis infection in bovines. Our study investigated this using data collected in 39 villages in a region of China where bovine and human schistosomiasis both occur. We found that several agriculture-related factors and bovine density in the village were predictive of bovine infection status. These findings highlight the importance of assessing environmental sources of disease transmission across large geographic scales, and suggest that preventative treatment of bovines residing in high risk villages may help to control local transmission.

medRxiv preprint doi: https://doi.org/10.1101/2021.08.20.21262368; this version posted August 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Introduction

Schistosomiasis is among the most burdensome helminth infections, with transmission being reported in a total of 78 countries in 2018 and approximately 230 million people in need of preventative treatment (1). Although great strides have been made in the last several decades in the control of schistosomiasis in several countries worldwide (1), pockets of reemergent or persistent transmission within such areas highlight the need for careful consideration of possible local drivers of transmission (2, 3). A poignant example of this is found when looking at the transmission of *Schistosoma japonicum* in China, where despite well-established control programs and great progress towards elimination since the mid-1950s (4), a 2018 national report highlighted that there remains 450 endemic counties where transmission interruption has yet to be achieved (5). *S. japonicum* has been found to be transmitted by at least 40 species of wild and domestic mammals (6), and animal activities near likely transmission sites may be important sources of reemergence and persistence. In China, several domesticated and wild animals have been identified as being capable of carrying and transmitting *S. japonicum*, including bovines, pigs, goats, dogs, cats and rodents (7, 8). Estimates from Jiangxi Province of Eastern China suggest that bovines may be responsible for as much as 75% of human transmission (9). This substantial contribution is thought to be related to the high degree of environmental overlap between humans and bovines during agricultural production, as well as the large amount of fecal output of bovines, which has been estimated to be as high as 100 times that of human fecal production each day (6, 10, 11). Additionally, the high frequency of livestock movement via the livestock trade within mountainous regions of China further highlights the important role that bovines may be playing in the *S. japonicum* transmission cycle in endemic areas (12, 13).

Despite increasing awareness that bovines may be an important driver of human schistosomiasis infection, little is known about what factors are likely to be influencing infection within bovine populations. Studies have highlighted several risk factors associated with other bovine infectious diseases, including individual characteristics like old-age, male sex, a range of breeds and uses (e.g. dairy, beef or agricultural work), group characteristics like herd size and herd density, and environmental characteristics like contact with other animals and the presence/absence of irrigation systems (14-19). While recent assessments of bovine risk of *S. bovis* infection in Eastern Africa have also studied individual-level risk factors such as bovine sex, age, breed and body condition, the results are contradictory (20-26). Reasons for such discrepancies have not been fully elucidated,

medRxiv preprint doi: https://doi.org/10.1101/2021.08.20.21262368; this version posted August 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

though Defersha & Belete (2018) hypothesize that it may be related to variations in management practices for different bovine groups (e.g. separation of sexes or of age groups) and different grazing ranges or grazing patterns allowed on different farms (e.g. smaller grazing area of very young and very old bovines) (23).

Outside of eastern Africa, few studies have set out to characterize predictors of bovine schistosomiasis infection. One study from Malaysia found that low weight, male sex, and older age were all risk factors for *S. spindale* infection in a range of different cattle species, though notably, no water buffalo species were included in this study (27). By comparison, in Southern China, a study conducted primarily among water buffaloes (96.2% water buffalos, 3.8% cattle) found that infection intensity was highest in bovines under the age of two (28). These seemingly contradictory results may potentially be explained by isolation and limited grazing for calves in Malaysia (27), as well as potential genus-related differences in acquired immunity and self-cure rates (29). Of the two main types of bovines found in *S. japonicum* endemic areas, yellow cattle are believed to be more susceptible to infection than water buffalo based on studies assessing worm establishment success in the two genera (30, 31). Nevertheless, He et al. (2001) also point out that water buffaloes may still act as important hosts in marshland areas of China, as they are more likely to spend time in water, and therefore more likely to be involved in the *S. japonicum* transmission cycle (10).

As *S. japonicum* is the primary species responsible for schistosomiasis infection in both humans and bovines in China, and given the considerable role that bovines are posited to have in contributing to human infection risk, studies aimed at assessing potential predictors of *S. japonicum* infection in bovines are of paramount importance. Not only is there a great deal of disagreement in the current body of literature over the key risk factors for bovine schistosomiasis infection, the limited studies to date have almost exclusively focused on physical/individual-level characteristics rather than broader environmental conditions. As such, this study set out to assess potential individual and environmental predictors of bovine *S. japonicum* infection in 2007, 2010 and 2016 at the individual, household and village-levels in a region where schistosomiasis persistence has been demonstrated to exist in both humans and bovine populations.

medRxiv preprint doi: https://doi.org/10.1101/2021.08.20.21262368; this version posted August 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Methods

Village selection

A longitudinal assessment of human and bovine infection was conducted in villages of Sichuan province in 2007, 2010 and 2016. Villages were located in the hilly regions of rural Sichuan and ranged from ~20-150 households and a population of ~50-200 people. Village selection has been described previously (32). Briefly, to identify villages with evidence of *S. japonicum* reemergence, county surveillance records were reviewed from the year that transmission control was achieved in the county through March 2007. Out of eight Sichuan counties where schistosomiasis had been identified despite control efforts (33), three were selected for inclusion based on surveillance record availability and the local control stations' willingness and capacity to collaborate (32). However, due to a 7.9 magnitude earthquake in May 2008 that severely impacted one of the study counties (34), follow-up surveys were conducted in 36 villages in the two remaining counties in 2010. Based on infection rates, 7 of the original 36 villages were surveyed again in 2016, in addition to 3 newly reemerging villages, giving a total of 36 villages included in the analysis in 2007 and 2010, and 10 villages in 2016.

Demographic, household and GPS surveys

A village census was conducted in each collection year and all residents over the age of five were invited to participate in surveys and stool sample screenings for *S. japonicum* infection. In addition, attempts were made to survey all bovines in the village for *S. japonicum* infection. In the summers of 2007, 2010 and 2016, the head of each household was asked to complete a household survey that contained closed-ended questions related to socioeconomic status, domestic and farm animal ownership, sanitation and water access and agricultural practices. Bovine age, type and sex were collected at the time of the bovine infection surveys in 2007 and 2010 (these data were not collected in 2016). Trained staff from the Sichuan Center for Disease Control and Prevention and the county Schistosomiasis Control Stations piloted and conducted all surveys in the local Sichuan dialect.

Ethics statement

This study was approved by the Sichuan Institutional Review Board, the University of California, Berkeley, Committee for the Protection of Human Subjects, and the Colorado Multiple Institutional Review Board. All participants provided written, informed consent. The collection of bovine samples we determined to be exempt

medRxiv preprint doi: https://doi.org/10.1101/2021.08.20.21262368; this version posted August 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

from review by the Animal Care and Use Committee at the University of California, Berkeley and the Institutional Animal Care and Use Committee at the University of Colorado Anschutz.

Infection surveys

Infection surveys were conducted by attempting to test three stool samples on three consecutive days from eligible humans and all bovines in the village. Infection surveys were conducted in November and December of 2007 and 2010, and July 2016. Individual bovines were isolated in a pen or tied up until a stool was produced on three separate days (consecutive, when possible). All stool samples were transported to the central laboratory soon after collection to be examined using the miracidium hatching test, following standard protocols (35). To account for the short survival and rapid hatching of bovine miracidia, the bovine samples were examine for miracidia at one, three and five hours after preparation for at least two minutes each time, whereas human samples were assessed at two, five and eight hours after preparation. One sample from each human was also examined using the Kato Katz thick smear procedure in 2007 and 2010 (36). A bovine was considered positive for *S. japonicum* if any miracidium hatching test was positive. A human was considered positive for *S. japonicum* if any miracidium hatching test or the Kato-Katz test was positive.

For each data collection period, the proportion of bovines in the village that were captured by infection surveys was assessed by comparing the total number of bovines reported in household surveys to the total number of bovines that participated in the infection survey in each village. Between 2007 and 2016, bovine infection status was assessed in 35/36 villages where residents reported owning bovines in 2007, 31/35 villages in 2010, and 8/8 villages in 2016. Details about participation and infection survey completeness are provided in S1 Table.

Predictor selection and definitions

The primary outcome of interest in this analysis was bovine *S. japonicum* infection in 2007, 2010 and 2016. All candidate predictors were defined using either the household surveys or the human and bovine infection surveys and were divided into five categories: 1) biological/physical characteristics; 2) potential human sources of environmental schistosomes; 3) socio-economic indicators; 4) potential animal reservoirs/sources of infection; 5) agricultural risk factors. We included agriculture as its own category because bovines are frequently employed in agricultural work in China (13), and because different crop types and agricultural practices have their own

medRxiv preprint doi: https://doi.org/10.1101/2021.08.20.21262368; this version posted August 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

inherent exposure risks (e.g. planting wet crops like rice may increase the likelihood of contact with snail habitat and exposure to cercariae (37)). Variables identified as predictors with hypothesized similar mechanisms of transmission risk were aggregated where possible (e.g. wet vs. dry crops). Three crop type categories were created: winter crops, summer dry crops and summer wet crops (i.e. rice). Night soil use – that is, the collection of either treated or untreated human and/or animal waste for use as fertilizer – was also included as an agricultural risk factor and divided into three categories: night soil use on winter crops, dry summer crops and wet summer crops.

There were minor variations in the household survey content and question formulation across the study period. Namely, some variables were not available in all the study years (e.g. pig ownership was not assessed in 2007). Where possible, continuous/discrete predictors were included over binary measures of a predictor for the household-level predictors. For binary variables, we excluded variables from the analysis of a given collection year if they represented very rare (<10%) or very common conditions (>90%). For continuous variables, variables were excluded when >90% of the observations took a single value. For example, household dog and pig ownership were both excluded in 2016 because >90% of the households owned one or more dogs (a binary variable), while >90% did not own any pigs. A composite household asset score (0-9) was developed for use in this assessment, which included eight household assets assessed in all three collection years (washing machine, television, air conditioner, refrigerator, computer, car, motorcycle), as well as a binary measure indicating that the home was made from either concrete, wood or bricks (vs. adobe).

Because prior work has demonstrated that group-level measures can serve as important predictors of schistosomiasis infection in humans (34), we also generated village-level candidate predictors from the household survey data. Village-level variables represent all households that participated in the household survey from a given village, even if they didn't own bovines. Village-level variables were either the village-average value of continuous household measures, or for binary variables, the proportion of the village population reporting the condition. Notably, the village-level variables excluded all observations from the bovine's own household, and instead used only the data from the other households in the village that participated in the household survey. This allowed for an assessment of how the surrounding village environment impacts individual bovine infection risk, independent of the home environment, whereas the household-level variables aim to unpack the influence of the

medRxiv preprint doi: https://doi.org/10.1101/2021.08.20.21262368; this version posted August 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

unique household environment on bovine infection status. The aforementioned predictor definitions and exclusion criteria led to a total of 31 predictors of bovine infection, which are summarized in Table 1.

Table 1. Summary of predictor variables included in the analysis.

Predictor list	Scale of analysis	Years available	Variable type	Predictor category [a]
Bovine type	Individual	2007, 2010	Binary	Physical/biological
Bovine sex	Individual	2007, 2010	Binary	Physical/biological
Bovine age	Individual	2007, 2010	Continuous	Physical/biological
Number of hatch tests [b]	Individual	All	Discrete	Physical/biological
County of residence	Household	All	Binary	Physical/biological
Number of infected human household members	Household	All	Discrete	Potential human sources
Household has improved sanitation (y/n) [c]	Household	All	Binary	Socio-economic indicators
Household asset score (0-9)	Household	All	Discrete	Socio-economic indicators
Household cat ownership	Household	All	Binary	Animal reservoirs/sources
Household dog ownership	Household	2007, 2010	Binary	Animal reservoirs/sources
Household pig ownership	Household	2010	Discrete	Animal reservoirs/sources
Household bovine ownership	Household	All	Discrete	Animal reservoirs/sources
Household rice area	Household	All	Continuous	Agricultural risk factors
Household dry summer crop area	Household	All	Continuous	Agricultural risk factors
Household winter crop area	Household	All	Continuous	Agricultural risk factors
Household night soil rice: # buckets	Household	All	Discrete	Agricultural risk factors
Household night soil summer dry crop: # buckets	Household	All	Discrete	Agricultural risk factors
Household night soil winter crop: # buckets	Household	All	Discrete	Agricultural risk factors
Village prevalence of human infection	Village	All	Continuous	Potential human sources
Village prevalence of improved sanitation	Village	All	Continuous	Socio-economic indicators
Village mean asset score (0-9)	Village	All	Continuous	Socio-economic indicators
Village prevalence of cat ownership	Village	All	Continuous	Animal reservoirs/sources
Village prevalence of dog ownership	Village	All	Continuous	Animal reservoirs/sources
Village mean number of pigs owned	Village	2010, 2016	Continuous	Animal reservoirs/sources
Village mean number of bovines owned	Village	All	Continuous	Animal reservoirs/sources
Village mean rice area	Village	All	Continuous	Agricultural risk factors
Village mean dry summer crop area	Village	All	Continuous	Agricultural risk factors
Village mean winter crop area	Village	All	Continuous	Agricultural risk factors
Village night soil rice: mean # buckets	Village	All	Continuous	Agricultural risk factors
Village night soil summer dry crop: mean # buckets	Village	All	Continuous	Agricultural risk factors
Village night soil winter crop: mean # buckets	Village	All	Continuous	Agricultural risk factors

[a] Predictors were grouped into five categories relevant to bovine *S. japonicum* infection risk probability: 1) physical/biological characteristics (e.g. old-age); 2) potential human sources of environmental schistosomes (e.g. human *S. japonicum* infection prevalence in the bovine's household); 3) socio-economic indicators (e.g. prevalence of improved sanitation systems in the surround village); 4) potential animal reservoirs/sources of infections (e.g. prevalence of dog ownership in the surrounding village); 5) agricultural risk factors (e.g. a household's total rice crop area).

[b] Because not all bovines produced three stool samples, and examination of a greater number of stool samples can increase the probability of detecting infection, the number of hatch tests used on a given bovine was also included as a predictor in our analyses.

[c] Improved sanitation was defined as access to an improved toilet in the household, including a biogas digester or a three-compartment toilet.

medRxiv preprint doi: https://doi.org/10.1101/2021.08.20.21262368; this version posted August 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Analysis

Across the collection years, 67 bovines with infection data were excluded from this analysis due to lack of household survey data (30/503 bovines in 2007; 36/233 bovines in 2010; and 1/72 in 2016). Infection prevalence was similar among the excluded bovines (11/67, 16.4%) as compared to those included in this analysis (111/741, 15.0%). Among the remaining bovines included, household and village-level variables generally had low levels of missing data (all <20% missing). By contrast, the individual-level bovine data was recorded with less consistency: the variable with the most missing data was bovine sex in 2007 (21.6% missing). Missing values were imputed separately for each collection year for all variables with <25% missing using the rfImpute function from the "randomForest" package in R (38, 39).

Spatial patterns of bovine infection prevalence were inspected using ESRI's ArcGIS ArcMap software release 10.5.1 (40). Categorical versions of each of the individual, household and village-level candidate predictors were generated and compared between *S. japonicum* infected and uninfected bovines to investigate potential changes in predictor distribution patterns by infection status across the study period.

To determine which of our candidate predictors serve as the best predictors of schistosomiasis transmission in 2007, 2010 and 2016, a random forests (RF) machine learning approach was used. For each collection year, 25% of the data was reserved for validation, while the remaining 75% was used for model construction. To address class imbalance in our outcome of interest (bovine *S. japoncium* prevalence of 13.3%, 17.3%, and 19.7% in 2007, 2010 and 2016, respectively), over sampling of the minority class was conducted. For model tuning, 10-fold cross validation was performed using the Caret package in R to help select the optimal maximum node size and the number of variables to try at each branch. Once the optimal value of each of these parameters was determined, a final model was run using 5000 trees per forest (41).

For each collection year, we conducted a total of ten rebalancing and model tuning iterations to assess the degree of stability in our variable importance rankings. The mean decrease in accuracy (MDA) value was used to rank the top ten predictors from each model on a scale of ten to one from most important (highest MDA) to least important (lowest MDA). These variable rankings were then summed across the 10 rebalancing iterations to give a 10-model summary score of variable importance, ranging from 100-1 and the ten highest scoring variables from the ten-model summary score were then reassigned a final ranking of 1-10. Next, using only those predictors ranked first through tenth within each collection year, we performed an additional ten iterations of the

medRxiv preprint doi: https://doi.org/10.1101/2021.08.20.21262368; this version posted August 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

aforementioned balancing and tuning process to create "lean" prediction model summary score of variable importance, thereby reducing excess noise in the variable ranking assessment caused by including a large number of candidate predictors. Because we hypothesized that the inclusion of human infection as a predictor of bovine infection would strongly influence the predictive capacity of our RF models due to a presumed association between bovine and human infection, we also conducted ten iterations of a sensitivity analysis for each collection year that excluded the human infection variables from the assessment. The ability of the full, lean and sensitivity RF models to predict infection status was then assessed using ROC area under the curve and accuracy. In the case of disagreement or a tie when comparing our chosen performance metrics, the sensitivity, kappa and specificity were subsequently compared to select the top performing model for each year.

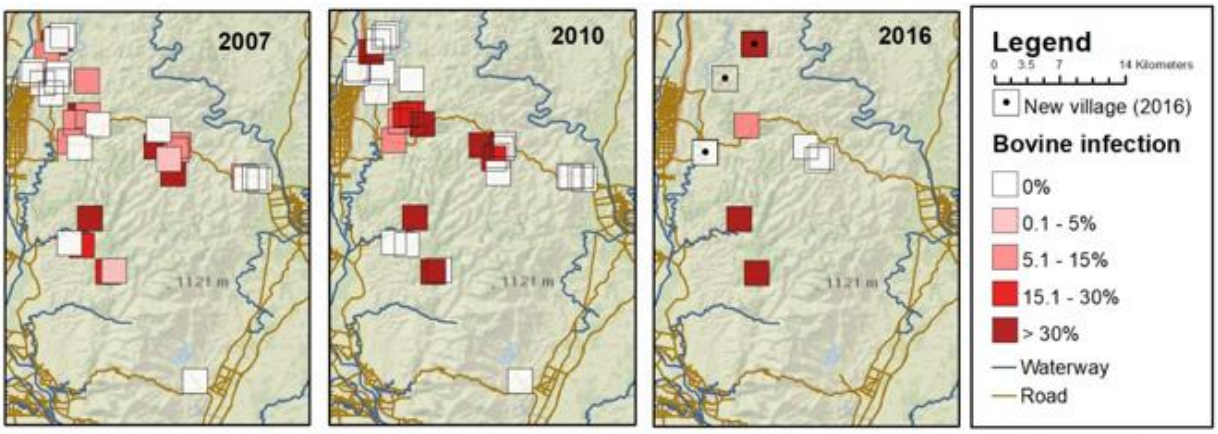
Each of the full model, lean model and sensitivity model summary scores were used to generate heat maps highlighting variable importance scores within each collection year, their change over time, and the frequency with which the different levels of analysis (individual, household or village) were each found to be among the top ten most important predictors. Simple logistic regression analyses were performed to assess the direction of association between the top predictors and bovine infection, dividing continuous variables into tertiles to assess for potential non-linear relationships. The direction of association was recorded for the top predictors within each collection year, using a p-value of <0.2 to indicate weak evidence of a between-group difference. In the case that no difference was indicated between tertile groups at the $p<0.2$ level, the predictors were further divided into quartiles and re-assessed. If still no evidence for a between group difference was identified using quartiles, this point was noted in the results. Density plots by infection status were also examined for a subset of predictors that were found to have a change in the direction of association across the collection years. Stata 15 and R Studio 4.0 were used for all analyses (42, 43).

Results

This analysis included bovines from 37 villages across the study period, with a total of 473 bovines from 35 villages in 2007, 197 bovines from 31 villages in 2010, and 71 bovines from 8 villages in 2016. The overall bovine infection prevalence was 13.3%, 17.3% and 19.2% in 2007, 2010 and 2016, respectively. Figure 1 shows a map of bovine infection distribution by village across two counties.

medRxiv preprint doi: https://doi.org/10.1101/2021.08.20.21262368; this version posted August 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Figure 1. Village-level prevalence of schistosomiasis in bovines in 2007, 2010 and 2016. Unshaded squares indicate study villages where no bovines were tested. Service Layer Credits: National Geographic, ESRI, DeLorme, HERE, UNEP-WCMC, USGS, NASA, METI, NOAA, increment P Corp, and OpenStreetMap Contributors, Geofabrik GmbH, Copyright 2018.



Bovine infection prevalence and individual-level characteristics

Of the individual characteristics assessed in this analysis, none were consistently associated with infection status (Table 2). For example, in 2007, water buffalo were less likely to be infected than cattle (8.1% in water buffalo, 14.2% in cattle), while in 2010, the prevalence of infection was ~17% in both groups. Similarly, in 2007 bovines over 5 years of age were three times more likely to be infected than younger bovines, but in 2010 there wasn't a clear pattern of infection by age. Notably, there were fewer water buffaloes than cattle in both assessment years (18.5% of all bovines were water buffalo in 2007; 15.5% in 2010), bovines were predominantly female (86.3% in 2007; 87.2% in 2010), and ranged widely in age from less than a year to 26 years old.

medRxiv preprint doi: https://doi.org/10.1101/2021.08.20.21262368; this version posted August 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Table 2. Tabulation of individual-level predictors by bovine infection status.

	2007			2010			2016		
	N Positive	Total tested	% Positive	N Positive	Total tested	% Positive	N Positive	Total tested	% Positive
Total	**63**	**473**	**13.32%**	**34**	**197**	**17.26%**	**14**	**71**	**19.72%**
County									
1	15	191	7.85%	17	71	23.94%	3	21	14.29%
2	48	282	17.02%	17	126	13.49%	11	50	22.00%
Bovine sex									
Female	50	320	15.63%	29	163	17.79%	ND [a]	ND [a]	
Male	9	51	17.65%	4	24	16.67%	ND [a]	ND [a]	
Missing	4	102	3.92%	1	10	10.00%			
Bovine age [b]									
<2	3	43	6.98%	6	35	17.14%	ND [a]	ND [a]	
2 to 4	26	220	11.82%	9	59	15.25%	ND [a]	ND [a]	
5+	28	137	20.44%	18	93	19.35%	ND [a]	ND [a]	
Missing	6	73	8.22%	1	10	10.00%			
Bovine type									
Water buffalo	7	86	8.14%	5	29	17.24%	ND [a]	ND [a]	
Cattle	54	380	14.21%	28	158	17.72%	ND [a]	ND [a]	
Missing	2	7	28.57%	1	10	10.00%			
Hatch tests (N)									
1	3	59	5.08%	0	29	0.00%	1	10	10.00%
2	8	80	10.00%	6	41	14.63%	2	9	22.22%
3	52	334	15.57%	28	127	22.05%	11	52	21.15%

[a] No data (ND). Data was not collected on bovine sex, age or type in 2016.

[b] Although age is broken into categories to facilitate a comparison of the distributions between 2007 and 2010, it was assessed as a continuous variable in our RF models.

Bovine infection prevalence and household-level characteristics

Bovine infection prevalence was highest among bovines in households where one or more humans were infected, and in households that did not own pigs (Table 3). Relationships between bovine infection and the rest of the household predictors were inconsistent across years. For example, access to improved sanitation and infection status by sanitation group shifted across our study period, rising from 22.8% of households reporting improved sanitation in 2007 and roughly equal infection prevalence in the two sanitation groups, to 52.1% with improved sanitation by 2016 and a higher probability of infection in the households with unimproved sanitation (23.5%) compared to the households with improved sanitation (16.2%). Across the study period, there was a steady increase in the prevalence of households reporting planting rice (69.1% in 2007; 71.5% in 2010; 81.7% in 2016), other summer crops (77.2% in 2007; 98% in 2010; 100% in 2016) and winter crops (97.7% in 2007; 99% in 2010; 100% in 2016). For rice crops in 2007 and 2010, the prevalence of bovine infection increased as the area of rice crop planted increased, whereas in 2016, this pattern did not hold. Other noteworthy changes in agricultural production across the study period includes a decrease in night soil use on rice and winter crops: the proportion of

medRxiv preprint doi: https://doi.org/10.1101/2021.08.20.21262368; this version posted August 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

households reporting any night soil use on rice crops dropped from 35.6% to 11.7% between 2007 and 2016, and for winter crops it dropped from 58.3% to 12.6%. By contrast, the proportion of night soil users for summer crops remained relatively constant across the study period (52.4% in 2007; 53.5% in 2010; 50.7% in 2016).

Table 3. Household predictors by bovine infection status.

		2007			2010			2016		
		# Positive	# tested	% Positive	# Positive	# tested	% Positive	# Positive	# tested	% Positive
# of infected humans in the household	0	38	366	10.38%	19	154	12.34%	12	61	19.67%
	1+	19	84	22.62%	13	29	44.83%	2	10	20.00%
	Missing	6	23	26.09%	2	14	14.29%	0	0	--
Household asset score (0-9)	0 - 2	22	184	11.96%	4	24	16.67%	2	10	20.00%
	2 - 3	31	211	14.69%	15	89	16.85%	3	14	21.43%
	4+	10	78	12.82%	15	84	17.86%	9	47	19.15%
Improved sanitation	No	47	365	12.88%	27	129	20.93%	8	34	23.53%
	Yes	16	108	14.81%	7	68	10.29%	6	37	16.22%
Household owns dogs [a]	No	6	67	8.96%	3	29	10.34%	2	7	28.57%
	Yes	57	404	14.11%	31	168	18.45%	12	64 [a]	18.75%
	Missing	0	2	0.00%	0	0	--	0	0	--
Household owns cats	No	17	155	10.97%	12	69	17.39%	5	23	21.74%
	Yes	46	312	14.74%	22	128	17.19%	9	48	18.75%
	Missing	0	6	0.00%	0	0	--	0	0	--
# Pigs owned by household [a, b]	0	ND [b]	ND [b]		27	116	23.28%	14	64 [a]	21.88%
	1+	ND [b]	ND [b]		7	81	8.64%	0	7	0.00%
# Other bovines owned by household	0	49	334	14.67%	20	126	15.87%	8	38	21.05%
	1+	14	139	10.07%	14	71	19.72%	6	33	18.18%
Total rice crop area (mu)	0	16	146	10.96%	5	56	8.93%	5	13	38.46%
	<2	24	195	12.31%	16	84	19.05%	4	31	12.90%
	2+	23	132	17.42%	13	57	22.81%	5	27	18.52%
Total dry summer crop area (mu)	0	8	108	7.41%	0	4	0.00%	0	0	--
	<3	39	242	16.12%	18	99	18.18%	2	26	7.69%
	3+	16	123	13.01%	16	94	17.02%	12	45	26.67%
Total winter crop area (mu)	<2	14	142	9.86%	9	52	17.31%	7	29	24.14%
	2-3	33	223	14.80%	18	97	18.56%	7	29	24.14%
	4+	16	108	14.81%	7	48	14.58%	0	13	0.00%
# buckets night soil on rice	0	40	304	13.16%	29	168	17.26%	8	53	15.09%
	1+	22	168	13.10%	5	29	17.24%	1	7	14.29%
	Missing	1	1	100.0%	0	0	--	5	11	45.45%

medRxiv preprint doi: https://doi.org/10.1101/2021.08.20.21262368; this version posted August 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

# buckets night soil on dry summer crops	0	31	248	12.50%	24	105	22.86%	6	36	16.67%
	1-25	6	53	11.32%	5	32	15.63%	5	22	22.73%
	>25	26	172	15.12%	5	60	8.33%	3	13	23.08%
# buckets night soil household winter crops	0	27	197	13.71%	31	137	22.63%	9	62	14.52%
	1-26	5	79	6.33%	0	16	0.00%	1	3	33.33%
	>26	31	197	15.74%	3	44	6.82%	4	6	66.67%

[a] Variables with >90% of observations taking on a single value were excluded from the RF assessment. This exclusion criteria applied twice in 2016: >90% of the included households reported owning at least one dog, and >90% of households owned zero pigs.

[b] ND. Data was not collected on pig ownership in 2007.

Bovine infection prevalence and village-level characteristics

Bovine infection prevalence was highest in villages with high levels of bovine ownership (Table 4). For the remaining village-level predictors however, the infection patterns were inconsistent. For example, bovine infection prevalence was highest in 2007 and 2010 for bovines residing in villages where a high percentage of the human population was infected, whereas in 2016, that pattern did not hold. In 2007, infection prevalence incrementally decreases as the percent of households in the village that own dogs increases, but infection prevalence was higher among bovines residing in villages with higher dog ownership in 2010 and 2016. Similarly, in 2007, the prevalence of bovine infection is highest in villages where more night soil is used on rice crops, dry summer crops and winter crops, but in 2010, bovine infection prevalence decreases as the surrounding village's night soil use increases. Notably, the average amount of night soil being applied to crops dropped across the study period for all crop types.

Table 4. Village-level predictors by bovine infection status.

		2007			**2010**			**2016**		
		# Positive	**# tested**	**% Positive**	**# Positive**	**# tested**	**% Positive**	**# Positive**	**# tested**	**% Positive**
Human infection prevalence (%)	0-2.53%	14	147	9.52%	0	76	0.00%	4	27	14.81%
	2.54- 11.11%	13	158	8.23%	11	57	19.30%	8	29	27.59%
	≥11.12%	36	168	21.43%	23	64	35.94%	2	15	13.33%
Mean household asset score (0-9)	<2	39	225	17.33%	0	0	0.00%	0	0	0.00%
	2-3	24	248	9.68	26	143	18.18%	2	23	8.70%
	≥4	0	0	0.00%	8	54	14.81%	12	48	25.00%

medRxiv preprint doi: https://doi.org/10.1101/2021.08.20.21262368; this version posted August 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

% households with an improved toilet	<10%	14	180	7.78%	3	24	12.50%	0	0	0.00%
	10-50%	38	215	17.67%	29	133	21.80%	6	24	25.00%
	≥50%	11	78	14.10%	2	40	5.00%	8	47	17.02%
% of households that own at least one dog	<70%	17	98	17.35%	2	51	3.92%	7	50	14.00%
	70-85%	34	217	15.67%	18	94	19.15%	7	21	33.33%
	≥85%	12	158	7.59%	14	52	26.92%	0	0	0.00%
% of households that own at least one cat	<45%	12	109	11.01%	12	71	16.90%	4	36	11.11%
	45-60%	21	124	16.94%	14	73	19.18%	5	13	38.46%
	≥60%	30	240	12.50%	8	53	15.09%	5	22	22.73%
Mean number of pigs owned	0	ND [a]	ND [a]		1	21	4.76%	11	20	55.00%
	0.01 - 1	ND [a]	ND [a]		26	119	21.85%	1	45	2.22%
	≥ 1	ND [a]	ND [a]		7	57	12.28%	2	6	33.33%
Mean number of bovines owned	< 0.5	5	116	4.31%	4	59	6.78%	10	57	17.54%
	0.5 – 1	40	300	13.33%	19	99	19.19%	4	14	28.57%
	≥ 1	18	57	31.58%	11	39	28.21%	0	0	0.00%
Mean area of rice planted (mu)	<0.75	14	143	9.79%	9	90	10.00%	7	18	38.89%
	0.75-1.5	19	183	10.38%	12	48	25.00%	5	41	12.20%
	≥1.5	30	147	20.41%	13	59	22.03%	2	12	16.67%
Mean area of dry summer crops planted (mu)	<1	11	184	5.98%	1	14	7.14%	0	3	0.00%
	1 -2.5	44	195	22.56%	23	110	20.91%	2	26	7.69%
	≥2.5	8	94	8.51%	10	73	13.70%	12	42	28.57%
Mean area of winter crops plants (mu)	<2	21	149	14.09%	17	77	22.08%	12	53	22.64%
	2-2.75	18	178	10.11%	11	66	16.67%	0	2	0.00%
	≥2.75	24	146	16.44%	6	54	11.11%	2	16	12.50%
Mean # buckets of night soil used on rice crops	<1	3	58	5.17%	29	146	19.86%	8	41	19.51%
	1-9.9	21	153	13.73%	4	38	10.53%	6	27	22.22%
	≥10	39	262	14.89%	1	13	7.69%	0	3	0.00%
Mean # buckets of night soil used on dry summer crops	<10	13	190	6.84%	24	99	24.24%	1	15	6.67%
	10-30	18	109	16.51%	7	47	14.89%	13	47	27.66%
	≥30	32	174	18.39%	3	51	5.88%	0	9	0.00%
Mean # buckets of night soil used on winter crops	<10	1	26	3.85%	24	121	19.83%	14	71	100.0%
	10-30	19	218	8.72%	7	41	17.07%	0	0	--
	≥30	43	229	18.78%	3	35	8.57%	0	0	--

[a] ND. Data was not collected on pig ownership in 2007.

medRxiv preprint doi: https://doi.org/10.1101/2021.08.20.21262368; this version posted August 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Predictors of bovine infection

The full models, lean models and sensitivity models within a given collection year all resulted in relatively stable rankings, while more variability in predictor rankings is seen when comparing across collection years (Figure 2). Within each model type for a given year, the ten iterations of re-balancing and tuning led to some variation in the MDA scores across the top ten predictors, with the top five more consistent in their high rankings. This is particularly prominent in the 2007 and 2010 models, whereas 2016 showed more variation overall. With few exceptions (4/60 model iterations), variables that scored in the top five within any of the ten iterations of either the full or sensitivity models were among the top ten predictors using the 10-model summary scores.

Agricultural variables were most frequently ranked in the top ten across all years. Specifically, the household area of winter crops planted, the mean area of rice planted in the surrounding village, and the mean amount of night soil applied to dry summer crops in the surrounding village were all ranked in the top ten for all collection years and the full, lean and sensitivity analyses. Additionally, the total household area of summer crops planted, the village mean area of winter crops and the mean number of bovines owned by the surrounding village were also all among the top ten predictors in at least one of the three model types used for 2007, 2010 and 2016. Of those predictors that ranked in the top ten in at least one collection year, four were scaled to the village-level, and two were assessed at the household-level. Because the full list of predictors changed slightly across the collection years, a supplemental analysis was conducted in which only predictors that were available in all three collection years were included in the RF models. This analysis demonstrated that 1) intra-year rankings and extra-year patterns did not change substantially, and 2) agricultural variables remained the most prominent predictor category when comparing across the entire study period. See S1 Figure for details of the supplemental analysis.

Figure 2. Variable importance rankings and direction of association for candidate predictors of bovine *S. japonicum* infection in 2007, 2010 and 2016. Variable importance rankings are based on a composite of mean decrease in accuracy scores for 10 random forest (RF) models for each model type (full, lean and sensitivity) and collection year. The direction of association was determined through logistic regression, using tertile categories for continuous variables to assess evidence for non-linearity. A p-value of <0.2 was used to indicate evidence of a between-group difference, and, when a between group difference was found, the direction of association is indicated. See S2 Table for detailed logistic regression results.

medRxiv preprint doi: https://doi.org/10.1101/2021.08.20.21262368; this version posted August 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Predictor categories	2007			2010			2016		
	Full	Lean	Sens.	Full	Lean	Sens.	Full	Lean	Sens.
Physical/biological characteristics									
Bovine type							ND	ND	ND
Bovine sex							ND	ND	ND
Bovine age	3 ↑	1 ↑	2 ↑				ND	ND	ND
Number of hatch tests				2 ↑	2 ↑	1 ↑			
County of residence									
Potential human sources of schistosomes									
Human infection prevalence in the village	5 ↑	3 ↑	Excl.	1 ↑	1↑	Excl.			Excl.
Number of infected human household members			Excl.	6 ↑	7 ↑	Excl.			Excl.
Socio-economic indicators									
Village prevalence of improved sanitation						9 ∩			
Household has improved sanitation (y/n)									
Village mean asset score (0-9)			10 ↓				8 ↑	5 ↑	8 ↑
Household asset score (0-9)						10 ø			
Potential animal reservoirs/sources of infection									
Village prevalence of cat ownership							9* ↑	10↑	10 ↑
Household cat ownership									
Village prevalence of dog ownership							9* ↑	9 ↑	
Household dog ownership							Excl.	Excl.	Excl.
Village mean number of pigs owned	ND	ND	ND				3 ↓	7 ↓	4 ↓
Household pig ownership	ND	ND	ND	7 ↓	8 ↓	6 ↓	Excl.	Excl.	Excl.
Village mean number of bovines owned	8 ↑	8 ↑	6 ↑			8 ↑	2 ↑	1 ↑	1 ↑
Household bovine ownership									
Agricultural risk factors									
Village mean rice area	6 ↑	5 ↑	3 ↑	5 ↑	5 ↑	3* ↑	5 ↓	8 ↓	5 ↓
Household rice area				8 ↑	10↑	7 ↑			
Village mean dry summer crop area	2 ↑	7 ↑	4 ↑				6 ↑	6 ↑	7 ↑
Household dry summer crop area	4 ↑	6 ↑	5 ↑	4 ↓	4 ↓	3* ↓			9 ↑
Village mean winter crop area	1 ↑	2 ↑	1 ↑	10↓	6 ↓		1 ↓	2 ↓	2 ↓
Household winter crop area	7 ↑	4 ↑	7 ↑	9 ø	9 ø	5 ø	7 ↓	4 ↓	6 ↓
Village night soil rice: mean # buckets									
Household night soil rice: # buckets									
Village night soil summer dry crop: mean # buckets	9 ↑	10↑	8 ↑	3 ↓	3 ↓	2 ↓	4 ↑	3 ↑	3 ↑
HH night soil summer dry crop: # buckets									
Village night soil winter crop: mean # buckets	10* U	9 U	9 U						
Household night soil winter crop: # buckets	10* U								

Symbol key:

- ↑ Positive association
- ↓ Negative association
- ∩ Non-linear association (Rise-Fall)
- U Non-linear association (Fall-Rise)
- ND Data not collected
- Excl. Variable excluded from model
- * Tied for importance rank
- ø No evidence for a significant between - group difference found

Color key: Variable importance rankings (1st-10th) by scale

Individual	1st – 2nd	3rd – 4th	5th – 6th	7th – 8th	9th – 10th
Household	1st – 2nd	3rd – 4th	5th – 6th	7th – 8th	9th – 10th
Village	1st – 2nd	3rd – 4th	5th – 6th	7th – 8th	9th – 10th

= Not ranked 1-10 for the given model/year

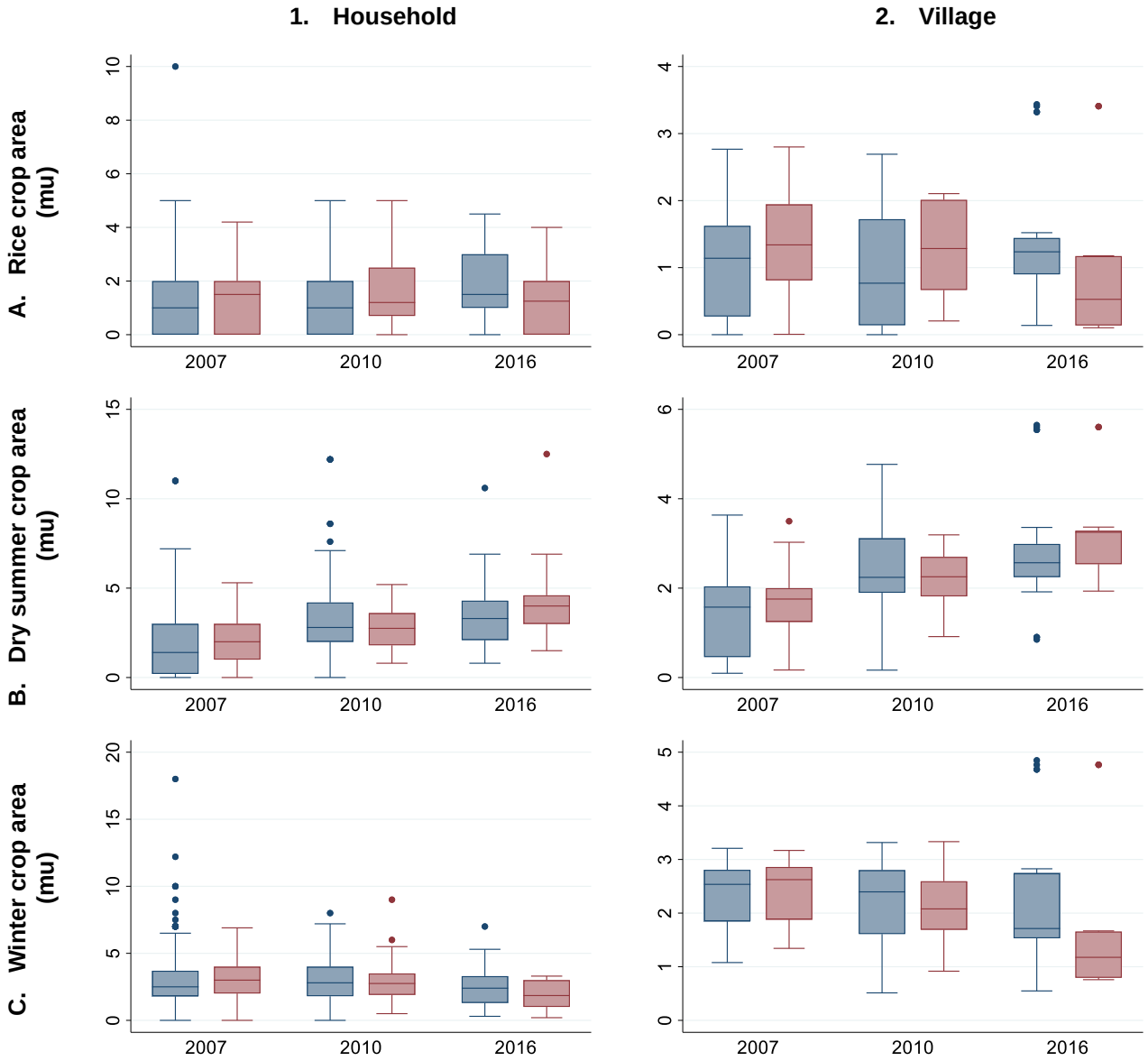
medRxiv preprint doi: https://doi.org/10.1101/2021.08.20.21262368; this version posted August 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Despite the inter-year agreement for several of the agricultural variables' high importance rankings, the direction of association between the top agricultural predictors and bovine infection was not consistent across the three collection years. For example, the logistic regression assessments suggest that the direction of association with bovine infection flips from positive to negative for village rice crop area (2007 & 2010 = ↑; 2016 = ↓), and winter crop area (2007 = ↑; 2010 & 2016 = ↓), while for household summer crop area and village night soil use on summer crops, the direction of association flips from positive to negative to positive (2007 = ↑; 2010 = ↓; 2016 = ↑). Notably, in 2007 increases in all the key agricultural predictors were associated with an increase in bovine infection risk, apart from night soil use on winter crops. By contrast, in 2010 and 2016 our models indicate a mixture of positive and negative associations across the key agricultural predictors, and in one instance (household winter crop area in 2010), no evidence of a relationship was found.

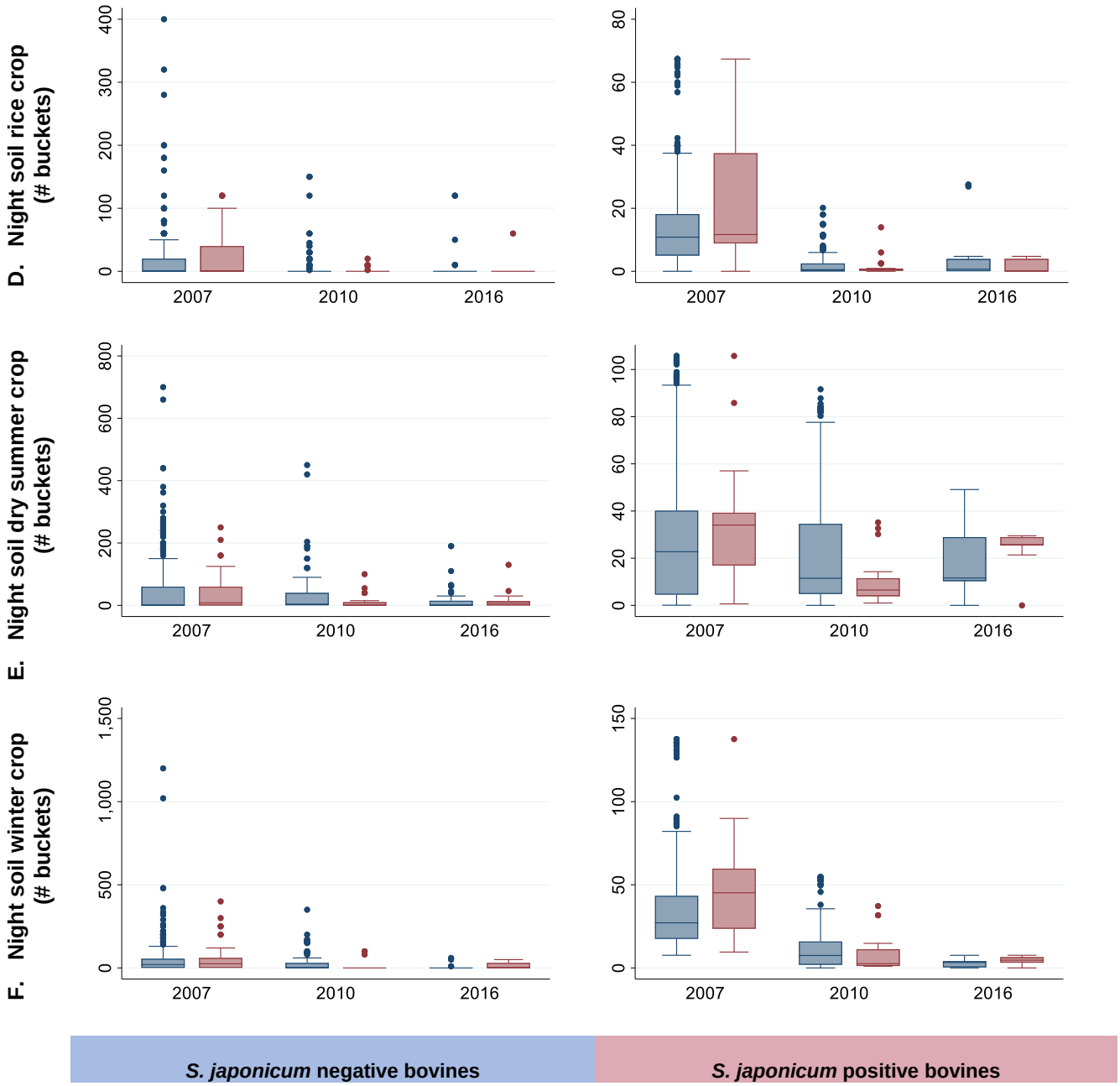
As mentioned above, the proportion of households planting rice, dry summer crops and winter crops, and the proportion of households reporting night soil use on rice and winter crops (but not dry summer crops) all shifted over the study period. Figure 3 depicts these shifting patterns over time, illustrating changes in the distribution of different agricultural practices by bovine infection status between 2007 and 2016. Despite the previously noted rise in the prevalence of households farming rice, dry summer crops and winter crops over the study period, panels A-C of Figure 3 show that only dry summer crop farming saw a notable increase in the total and mean area of crop being planted by households and villages between 2007 and 2016. On the other hand, a general decrease in the overall range and mean number of buckets of night soil being applied to rice, winter crops and, to a lesser extent, dry summer crops, can be observed when comparing between 2007 and 2016 (Figure 3, panels D-F).

medRxiv preprint doi: https://doi.org/10.1101/2021.08.20.21262368; this version posted August 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Figure 3. Changes in agricultural practices and the relationship between bovine infection and agricultural predictors over time. For each of the agricultural predictors included in this analysis, boxplots are used to represent the distribution of uninfected (blue), infected (red), for household-level (left) and village-level variables (right) in 2007, 2010 and 2016.



medRxiv preprint doi: https://doi.org/10.1101/2021.08.20.21262368; this version posted August 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .



In addition to the agricultural variables, there are also some other notable predictors that stand out in one or more collection year. Village bovine ownership is among the top ten predictors in at least one RF analysis from each year, all of which indicate that an increase in bovine ownership in the surrounding village corresponds with an increase in bovine infection risk. Human infection prevalence in the surrounding village was among the top five predictors of bovine infection in 2007 and 2010, and the number of infected humans within the household was among the top ten predictors in 2010. For both the household and village human infection predictors, an increase

medRxiv preprint doi: https://doi.org/10.1101/2021.08.20.21262368; this version posted August 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

in human infections was associated with an increase in bovine infections. When the human infection predictors were removed for the sensitivity analysis, the rankings of the remaining predictors did not shift substantially in any collection year. Of the physical/biological characteristics assessed, bovine age was among the top predictors in 2007, with the logistic regression results suggesting that a bovine's infection risk increased with age. In all of the 2010 analyses, the number of hatch tests was an important predictor of bovine infection, a feature not shared by the 2007 and 2016 analyses. This may be related to the relatively high proportion of bovines that had less than three hatch test results in 2010 (35.5%), as compared to 2007 (29.4%) and 2016 (26.8%).

Of the three different analyses performed (full, lean and sensitivity) for each collection year, the full models (i.e. those that included the full list of predictors available in a given year) tended to perform the best, as is highlighted in Table 5. Overall our models had high accuracy values, with the top performing models producing a maximum accuracy of 0.864 (95% CI: 0.79 – 0.92) in 2007, 0.816 (95% CI: 0.68 – 0.91) in 2010, and 1.0 (0.81 – 1.0), in 2016. However, due to class imbalance in our reserved test datasets (see the no information rate (NIR) in Table 5), the Kappa value is a useful performance metric for our models, as this takes class imbalance into account. According to the benchmarks laid out by Landis and Koch (1977), the Kappa statistics from our 2007 analyses suggest a "Fair" level of agreement (0.21 – 0.40) between our best RF models and the true known values in 2007. For 2010, the highest Kappa statistic came from the full predictor analysis, with a Kappa of 0.463, indicating a "Moderate" level of agreement (0.41 – 0.60) between the prediction model and the reserved test dataset (44). In 2016, both the full and sensitivity models achieved perfect prediction (Kappa = 1) for the test dataset in at least one of the ten model iterations, whereas the Kappa statistic for the top performing lean model was 0.853, or "Almost Perfect", according to Landis & Koch (44).

medRxiv preprint doi: https://doi.org/10.1101/2021.08.20.21262368; this version posted August 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Table 5. Comparison of model performance metrics for the top performing model from the full, lean and sensitivity analyses in 2007, 2010 and 2016. The top performing model was defined as the one with the highest accuracy for each analysis type (full, lean & sensitivity) and collection year (2007, 2010, 2016). In the case of a tie for the highest accuracy value, the sensitivity, kappa and specificity were subsequently compared to select the top performing model for each analysis type and year.

Performance metrics	**2007 Models**			**2010 Models**			**2016 Models**		
	Full	**Lean**	**Sens.**	**Full**	**Lean**	**Sens.**	**Full**	**Lean**	**Sens.**
Accuracy	0.864	0.864	0.856	0.816	0.816	0.816	1	0.944	1
95% CI	0.79 - 0.92	0.79 - 0.92	0.78 - 0.91	0.68 - 0.91	0.68 - 0.91	0.68 - 0.91	0.81 - 1	0.73 - 1	0.81 - 1
No info rate (NIR)	0.864	0.864	0.864	0.837	0.837	0.837	0.778	0.778	0.778
p-value (Acc > NIR)	0.566	0.566	0.667	0.729	0.779	0.729	0.011	0.067	0.011
Kappa [α]	0.313	0.313	0.246	0.463	0.360	0.416	1	0.853	1
Sensitivity	0.313	0.313	0.25	0.75	0.5	0.505	1	1	1
Specificity	0.951	0.951	0.951	0.829	0.878	0.625	1	0.929	1
Pos pred Value	0.5	0.5	0.444	0.462	0.444	0.854	1	0.8	1
Neg Pred Value	0.898	0.898	0.890	0.944	0.9	0.455	1	1	1

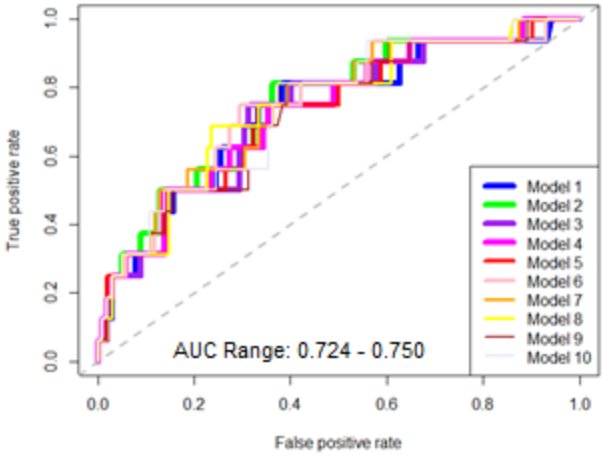
[α] Due to the high degree of imbalance between the outcome classes across the study period, the Kappa values is a useful metric for our models, as it helps to correct bias that results when rewarding the prediction of the majority class. The benchmark values outlined by Landis & Koch (1977) are useful here for determining the relative strength of the predictive models: <0.00 = Poor; 0.00 – 0.20= Slight; 0.21 – 0.40 = Fair; 0.41 – 0.60 = Moderate; 0.61 – 0.81 = Substantial; 0.81 – 1.0 = Almost Perfect.

While there was some variation in model performance across the ten iterations of RF models for each analysis year, overall the models were relatively stable. For the ten iterations of full analyses conducted for each collection year, the AUC ranged from 0.724 – 0.75 in 2007, 0.816 – 0.819 in 2010, and 0.982 – 1.0 in 2016. Figure 5 illustrates the ROC curve and corresponding best and worst AUC for each of the ten RF models of the full predictor list analyses.

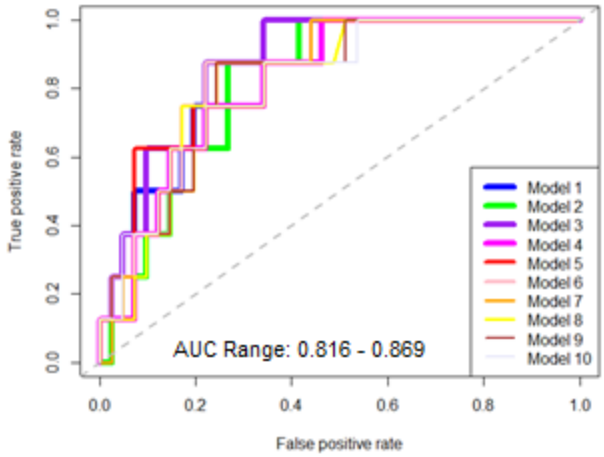
Figure 5. Receiver operator curves for each of the ten full RF model iterations conducted for 2007, 2010 (A) The ten RF models ROC curves for 2007 are shown in the top panel. The AUC in the 2007 full models ranged from 0.724 – 0.75. (B) The ten RF models ROC curves for 2010 are shown in the middle panel. The AUC in the 2010 full models ranged from 0.816 – 0.869. (C) The ten RF models ROC curves for 2016 are shown in the bottom panel. The AUC in the 2016 full models ranged from 0.982 – 1.0.

medRxiv preprint doi: https://doi.org/10.1101/2021.08.20.21262368; this version posted August 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

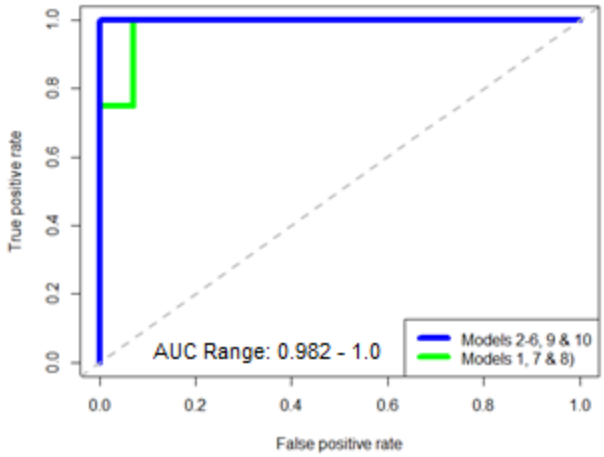
A) 2007 Full Models



B) 2010 Full Models



C) 2016 Full Models



medRxiv preprint doi: https://doi.org/10.1101/2021.08.20.21262368; this version posted August 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Discussion

Of the five categories that were assessed as potential predictors of bovine infection in this study (physical/biological characteristics, human infection-related, socio-economic, potential animal reservoirs and agricultural factors), agricultural factors were important predictors of bovine *S. japonicum* infection in all collection years. Night soil use on summer crops, the village-level area of rice crops, and both the household and village-level areas of summer and winter crops were each ranked among the top five predictors for one or more collection years in our RF models. Interestingly, for 2007, all of the ranked agricultural variables except one were associated with an increase in bovine infection risk in our logistic regression assessments, whereas in 2010 and 2016, these agricultural factors were found to be variably positively and negatively associated with infection risk. This finding may be related to changing norms and interventions that have taken hold in recent years as a result of increasing awareness of the potential risks posed by both bovines as a reservoir of schistosomiasis, and specific agricultural practices. For example, across our study period, we saw a steady increase in the prevalence of households planting rice (69.1% in 2007; 71.5% in 2010; 81.7% in 2016), and a simultaneous decrease in the prevalence of households applying any night soil to their rice crops (35.6% in 2007, 14.7% in 2010; 11.7% in 2016). These shifting norms in rice production and night soil use likely resulted in a decrease in the overall concentration of night soil on rice crops within our study villages, which in turn, may help to explain why the village-level rice crop area shifts from having a positive association with bovine infection in 2007 and 2010, to a negative association by 2016.

Assessments conducted in China early in the new millennium repeatedly highlighted bovines as a key source of environmental contamination and as the main animal reservoir of *S. japonicum* in the country (9, 28, 45). Beginning in 2004, a new government-led approach to eliminating schistosomiasis transmission in China was adopted, which – in conjunction with infrastructure improvements in rural areas and several new schistosomiasis elimination interventions – featured replacing bovines with machinery in agricultural production (46). Thus, the negative associations that were found intermittently between bovine infection and some of our agricultural variables in 2010 and 2016 may be linked to the added precautions that were being adopted when bovines were being used for agriculture, or because bovines were being reallocated for other purposes (e.g. beef production) as machinery became the norm for large crop areas or those deemed high risk (e.g. wet rice crops). Increasing

medRxiv preprint doi: https://doi.org/10.1101/2021.08.20.21262368; this version posted August 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

recognition of the potential risks posed by night soil use during our study period (32) may have also contributed to some decreases in environmental contamination as a result of decreases in night soil applications and/or the more careful treatment of night soil prior to field applications. Indeed, a downward trend in the range of reported night soil use (total and mean number of buckets) on crops can be observed in Figure 4 (parts D-F), though notably, we do not see any substantial shift in the overall proportion of households that reported any night soil use on summer crops over the years (52.4% in 2007; 53.5% in 2010; 50.7% in 2016) (Table 3). The continued prominence of applying some amount of night soil to summer crops, paired with the steady increase in the total area of summer crops being planted by villagers over the study period (see Figure 4, part B) may help to explain why night soil use on summer crops returns to being positively association with bovine infection status in 2016.

Bovine ownership in the surrounding village was in the top ten predictors of RF models and bovine density in a village was positively associated with bovine infection in our regression models in all collection years. These findings align well with the existing literature that points to bovines as the most important reservoir of *S. japonicum* infection in China (9, 45), and suggests that being in close proximity to higher densities of bovine hosts may correspond with increasing infection risk, as has been found for other bovine pathogens (47, 48). However, it is worth noting that household-level bovine ownership was not among the top predictors in any of our RF models, highlighting that the larger-scale lens (i.e. village-level analysis scale) may be particularly important to future investigations and control strategies. Likewise, recent informal interviews with locals from our study sites have revealed that bovines are infrequently kept near the home, as allowing bovines to graze (and defecate) freely is an economical and efficient way of raising bovines, further illustrating that the household scale may not always be broad enough to capture larger scale trends. Instead, villagers opt to bring their bovines to the mountains to graze during the day, which subsequently presents more opportunities for contact between bovines from different households, and may ultimately result in more widespread environmental contamination (e.g. bovine feces washed into nearby irrigation ditches after precipitation).

In the developmental stages of this analysis, we hypothesized that human infection prevalence and the number of infected people in the household would be among the top predictors of bovine infection status, given the known link between human schistosomiasis and bovine reservoirs (e.g. 45). It was therefore somewhat surprising to find that household-level human infection was only ranked as important from RF models in 2007, and village human infection prevalence was only ranked as important in 2007 and 2010. One potential explanation for

medRxiv preprint doi: https://doi.org/10.1101/2021.08.20.21262368; this version posted August 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

the apparent drop in the importance of human infection status as a predictor of bovine infection could be related to the aforementioned bovine-removal phenomenon, in which bovines are increasingly being removed from the village area and brought to alternative mountain locations for grazing, resulting in less frequent contact between bovines and humans, but more opportunities for contact with other bovines. In fact, the drop in the important rankings of human infection status in 2016 coincides with a jump in the variable importance rankings for village-level bovine ownership (6th – 8th in 2007 and 2010; 1st -2nd in 2016), providing further support of the theory that bovines may be becoming increasing important reservoirs of continued schistosomiasis infection. On the other hand, an altogether different explanation for the differences in the 2016 rankings compared to 2007 and 2010 is that the 2016 data collection simple didn't have a large enough sample size to allow for the detection of a true relationship between relatively rare events.

As such, one limitation of this assessment was the relatively small sample sizes, particularly in 2016 (N=71), though to a lesser extent, 2010 (N=197) and 2007 (N=473), given the correspondingly large number of predictors that were included in the full predictor models (N=29, N=31, N=26, in 2007, 2010 and 2016 respectively). While RF models are generally acknowledged as being able to handle assessments of high dimensional data even with relatively small sample sizes (49), it remains that small samples sizes can still give rise to the aforementioned issue of non-detection of rare events. Another limitation to this assessment is that RF models tends to favor continuous predictors over categorical measures, as they allow for a wider range of potential split points for classifying observations. For this reason, it is not particularly surprising that age was the only predictor from the individual/physical characteristics predictor group that was ranked among the top ten predictors, as the remaining individual characteristics were binary measures. Another notable limitation of the variable importance rankings used in RF models is that they become less reliable when predictors are highly correlated with one another (50). This may be particularly important to the rankings ascribed to the agricultural variables, as correlation between the area of the different crop types planted and the amount of night soil used on each crop tended to be high across all collection years, with the highest predictor correlations found in the 2016 collection year (See S2 – S4 Figures for correlation matrices). This is notable, as a higher degree of instability in the variable importance rankings was also found for 2016 as compared to 2010 or 2007, suggesting predictor correlation may be responsible. We therefore recommend that the variable rankings presented from this analyses be interpreted more holistically (e.g. agricultural variables are strong predictors of bovine infection), and advise

medRxiv preprint doi: https://doi.org/10.1101/2021.08.20.21262368; this version posted August 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

caution when comparing unique variable ranking values against one another (e.g. rice crop area is less important than winter crop area).

Our main interests in this assessment were to 1) identify the best predictors of bovine *S. japonicum* infection within rural farming communities in Sichuan China, and 2) to ascertain whether there are broader trends in bovine infection distribution across individual, household or village-levels scales or over time. Our RF assessments have highlighted several key patterns that were repeated across multiple collection years and multiple iterations of three different models. Agricultural factors and high levels of bovine ownership at the village-level were repeatedly found to be among the top predictors of bovine *S. japonicum* infection, highlighting the potential utility of presumptively treating bovines belonging to villages with particularly high levels of bovine ownership, or those who engage in high-risk agricultural practices such as planting rice. Additionally, village-level predictors tended to be better predictors of bovine infection than household-level predictors, suggesting that interventions may need to take a multipronged approach to address broader ecological sources of ongoing transmission.

medRxiv preprint doi: https://doi.org/10.1101/2021.08.20.21262368; this version posted August 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Competing interests: The authors have declared that no competing interests exist.

Data Availability: Due to the inclusion of potentially identifying information (e.g. socio-economic indicators and human infection status), access to study data must be requested through the Carlton Lab group to ensure the protection of research subjects and compliance with all ethical guidelines. Please contact Elizabeth Carlton at elizabeth.carlton@cuanschutz.edu for more information.

Funding: This research was supported by grants from the National Institute of Allergy and Infectious Diseases: R01AI134673 (EJC, PI), R21AI115288 (EJC, PI) and R01AI068854 (Robert Spear, PI). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author Contributions: Conceptualization: EC, YL, SP, AB; Data Curation: DL, EG; Formal Analysis: EG, SP, KK, AB, EC; Funding Acquisition: EC, YL, SP; Investigation: YL, DL, EC; Methodology: EG, SP, EC, AB, KK; Project Administration: YL, DL; Resources: EC, YL, DL; Validation: EG, SP, EC, KK; Writing – Original Draft Preparation: EG, SP, EC, KJ; Writing – Review & Editing: EG, SP, KK, YL, DL, EC, AB, KJ.

Acknowledgments: We are grateful for the support and efforts of the field research team members from the Institute of Parasitic Diseases and the county anti-schistosomiasis control stations for their efforts in collecting the data presented here.

medRxiv preprint doi: https://doi.org/10.1101/2021.08.20.21262368; this version posted August 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

References

1. World Health Organization. Schistosomiasis: Key Facts 2020 [Available from: https://www.who.int/news-room/fact-sheets/detail/schistosomiasis.
2. Kittur N, King CH, Campbell CH, Kinung'hi S, Mwinzi PNM, Karanja DMS, 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. The American journal of tropical medicine and hygiene. 2019;101(3):617-27.
3. Song LG, Wu XY, Sacko M, Wu ZD. History of schistosomiasis epidemiology, current status, and challenges in China: on the road to schistosomiasis elimination. Parasitology research. 2016;115(11):4071-81.
4. Xu J, Steinman P, Maybe D, Zhou XN, Lv S, Li SZ, et al. Evolution of the National Schistosomiasis Control Programmes in The People's Republic of China. Adv Parasitol. 2016;92:1-38.
5. Zhang LJ, Xu ZM, Guo JY, Dai SM, Dang H, Lü S, et al. [Endemic status of schistosomiasis in People's Republic of China in 2018]. Zhongguo xue xi chong bing fang zhi za zhi = Chinese journal of schistosomiasis control. 2019;31(6):576-82.
6. Gray DJ, Williams GM, Li Y, McManus DP. Transmission Dynamics of Schistosoma japonicum in the Lakes and Marshlands of China. PloS one. 2009;3(12):e4058.
7. Li H, Dong GD, Liu JM, Gao JX, Shi YJ, Zhang YG, et al. Elimination of schistosomiasis japonica from formerly endemic areas in mountainous regions of southern China using a praziquantel regimen. Veterinary parasitology. 2015;208(3-4):254-8.
8. Van Dorssen CF, Gordon CA, Li Y, Williams GM, Wang Y, Luo Z, et al. Rodents, goats and dogs – their potential roles in the transmission of schistosomiasis in China. Parasitology. 2017;144(12):1633-42.
9. GUO J, LI Y, GRAY D, NING A, HU G, CHEN H, et al. A DRUG-BASED INTERVENTION STUDY ON THE IMPORTANCE OF BUFFALOES FOR HUMAN SCHISTOSOMA JAPONICUM INFECTION AROUND POYANG LAKE, PEOPLE'S REPUBLIC OF CHINA. The American Journal of Tropical Medicine and Hygiene. 2006;74(2):335-41.
10. He YX, Salafsky B, Ramaswamy K. Host--parasite relationships of Schistosoma japonicum in mammalian hosts. Trends Parasitol. 2001;17(7):320-4.
11. Ross AG, Sleigh AC, Li Y, Davis GM, Williams GM, Jiang Z, et al. Schistosomiasis in the People's Republic of China: prospects and challenges for the 21st century. Clinical microbiology reviews. 2001;14(2):270-95.
12. Zhou YB, Liang S, Jiang QW. Factors impacting on progress towards elimination of transmission of schistosomiasis japonica in China. Parasites & vectors. 2012;5:275.
13. Zheng J, Guo JG, Wang XF, Zhu HQ. Relationship of the livestock trade to schistosomiasis transmission in mountainous area. Zhongguo ji sheng chong xue yu ji sheng chong bing za zhi = Chinese journal of parasitology & parasitic diseases. 2000;18(3):146-8.
14. Soomro A. A Study on Prevalence and Risk Factors of Brucellosis in Cattle and Buffaloes in District Hyderabad, Pakistan. Journal of Animal Health and Production. 2014;2:33-7.
15. Mugizi DR, Boqvist S, Nasinyama GW, Waiswa C, Ikwap K, Rock K, et al. Prevalence of and factors associated with Brucella sero-positivity in cattle in urban and peri-urban Gulu and Soroti towns of Uganda. J Vet Med Sci. 2015;77(5):557-64.
16. Skuce RA, Allen AR, McDowell SWJ. Herd-Level Risk Factors for Bovine Tuberculosis: A Literature Review. Veterinary Medicine International. 2012;2012:621210.
17. Nzalawahe J, Kassuku AA, Stothard JR, Coles GC, Eisler MC. Trematode infections in cattle in Arumeru District, Tanzania are associated with irrigation. Parasites & vectors. 2014;7:107.

medRxiv preprint doi: https://doi.org/10.1101/2021.08.20.21262368; this version posted August 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

18. Deka RP, Magnusson U, Grace D, Lindahl J. Bovine brucellosis: prevalence, risk factors, economic cost and control options with particular reference to India-a review. Infection Ecology & Epidemiology. 2018;8(1):1556548.

19. Tempia S, Salman M, Keefe T, Morley P, Freier J, DeMartini J, et al. A sero-survey of rinderpest in nomadic pastoral systems in central and southern Somalia from 2002 to 2003, using a spatially integrated random sampling approach. Revue scientifique et technique. 2010;29(3):497.

20. Chanie M, Dejen B, Fentahun T. Prevalence of cattle schistosomiasis and associated risk factors in Fogera cattle, south Gondar zone, Amhara national regional state, Ethiopia. Journal of Advanced Veterinary Research. 2012;2:153-6.

21. Kebede A, Dugassa J, Haile G, Wakjira BM. Prevalence of bovine of schistosomosis in and around Nekemte, East Wollega zone, Western Ethiopia. Journal of Veterinary Medicine and Animal Health. 2018;10:123-7.

22. Gebremeskel AK, Simeneh ST, Mekuria SA. Prevalence and Associated Risk Factors of Bovine Schistosomiasis in Northwestern Ethiopia. World. 2017;7(1):01-4.

23. Defersha T, Belete B. The Neglected Infectious Disease, Bovine Schistosomiasis: Prevalence and Associated Risk Factors for its Occurrence among Cattle in the North Gulf of Lake Tana, Northwest Ethiopia. J Vet Med Health. 2018;2:112.

24. Yihunie A, Urga B, Alebie G. Prevalence and risk factors of bovine schistosomiasis in Northwestern Ethiopia. BMC veterinary research. 2019;15(1):12.

25. Tsega M, Derso S. Prevalence of bovine schistosomiasis and its associated risk factor in and around Debre Tabor town, north west of Ethiopia. Europ J Biol Sci. 2015;7:108-13.

26. Lulie B, Guadu T. Bovine schistosomiasis: A threat in public health perspective in Bahir Dar town, northwest Ethiopia. Acta Parasitologica Globalis. 2014;5(1):1-6.

27. Tan TK, Low VL, Lee SC, Panchadcharam C, Kho KL, Koh FX, et al. Detection of Schistosoma spindale ova and associated risk factors among Malaysian cattle through coprological survey. Japanese Journal of Veterinary Research. 2015;63(2):63-71.

28. Gray DJ, Williams GM, Li Y, Chen H, Li RS, Forsyth SJ, et al. A cluster-randomized bovine intervention trial against Schistosoma japonicum in the People's Republic of China: design and baseline results. The American journal of tropical medicine and hygiene. 2007;77(5):866-74.

29. Li YS, McManus DP, Lin DD, Williams GM, Harn DA, Ross AG, et al. The Schistosoma japonicum self-cure phenomenon in water buffaloes: potential impact on the control and elimination of schistosomiasis in China. International journal for parasitology. 2014;44(3-4):167-71.

30. Xu S, Shi F, Shen W, Lin J, Wang Y, Lin B, et al. Vaccination of bovines against schistosomiasis japonica with cryopreserved-irradiated and freeze-thaw schistosomula. Veterinary parasitology. 1993;47(1-2):37-50.

31. He Y, Xu S, Shi F, Shen W, Hsü S, Hsü H. Comparative studies on the infection and maturation of schistosoma japonicum in cattle and buffaloes. Current Zoology. 1992;38(3):266-71.

32. Carlton EJ, Bates MN, Zhong B, Seto EYW, Spear RC. Evaluation of Mammalian and Intermediate Host Surveillance Methods for Detecting Schistosomiasis Reemergence in Southwest China. PLoS Negl Trop Dis. 2011;5(3).

33. Liang S, Yang C, Zhong B, Qiu D. Re-emerging schistosomiasis in hilly and mountainous areas of Sichuan, China2006. 139-44 p.

34. Carlton EJ, Liu Y, Zhong B, Hubbard A, Spear RC. Associations between Schistosomiasis and the Use of Human Waste as an Agricultural Fertilizer in China. PLoS Negl Trop Dis. 2015;9(1).

35. Control DoD. Textbook for Schistosomiasis Control. Shanghai Shanghai Publishing House for Science and Technology. 2000:72-6.

medRxiv preprint doi: https://doi.org/10.1101/2021.08.20.21262368; this version posted August 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

36. Katz N, Chaves A, Pellegrino J. A simple device for quantitative stool thick-smear technique in Schistosomiasis mansoni. Rev Inst Med Trop Sao Paulo. 1972;14(6):397-400.

37. Gordon CA, Kurscheid J, Williams GM, Clements ACA, Li Y, Zhou XN, et al. Asian Schistosomiasis: Current Status and Prospects for Control Leading to Elimination. Trop Med Infect Dis. 2019;4(1).

38. Breiman LC, A.; Liaw, A.; Wiener, M. . Breiman and Cutler's Random Forests for Classification and Regression. https://cran.r-project.org/web/packages/randomForest/randomForest.pdf; 2018.

39. RColorBrewer S, Liaw MA. Package ‘randomForest’.

40. Environmental Systems Research Institute (ESRI). ArcGIS Desktop Release. 10.5.1 ed. Redlands, CA. 2017

41. Breiman L, Cutler A. Manual–setting up, using, and understanding random forests V4. 0. 2003. URL https://www stat berkeley edu/~ breiman/Using_random_forests_v4 0 pdf. 2011.

42. StataCorp. Stata Statistical Software: Release 15. College Station, TX: StataCorp LP; 2015.

43. RStudio Team. RStudio: Integrated Development for R. RStudio, PBC, Boston, MA, 2020. 2020.

44. Landis JR, Koch GG. The Measurement of Observer Agreement for Categorical Data. Biometrics. 1977;33(1):159-74.

45. Gray DJ, Williams GM, Li Y, Chen H, Forsyth SJ, Li RS, et al. A cluster-randomised intervention trial against Schistosoma japonicum in the Peoples' Republic of China: bovine and human transmission. PloS one. 2009;4(6):e5900.

46. Liu Y, Zhong B, Wu Z-S, Liang S, Qiu D-C, Ma X. Interruption of schistosomiasis transmission in mountainous and hilly regions with an integrated strategy: a longitudinal case study in Sichuan, China. Infectious Diseases of Poverty. 2017;6(1):79.

47. Spencer SE, Besser TE, Cobbold RN, French NP. 'Super' or just 'above average'? Supershedders and the transmission of Escherichia coli O157:H7 among feedlot cattle. J R Soc Interface. 2015;12(110):0446.

48. Meadows AJ, Mundt CC, Keeling MJ, Tildesley MJ. Disentangling the influence of livestock vs. farm density on livestock disease epidemics. Ecosphere. 2018;9(7):e02294.

49. Biau G, Scornet E. A random forest guided tour. TEST. 2016;25(2):197-227.

50. Strobl C, Boulesteix A-L, Kneib T, Augustin T, Zeileis A. Conditional variable importance for random forests. BMC Bioinformatics. 2008;9(1):307.

medRxiv preprint doi: https://doi.org/10.1101/2021.08.20.21262368; this version posted August 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Supporting Information

S1 Table. Completeness of bovine infection and household surveys. Some differences between the number of bovines reported by households and the number of bovines tested may have arisen due to the lag time between the household surveys, which were completed in June/July in 2007 and 2010 and the infection surveys, which were conducted in November and December in 2007 and 2010. Both the household surveys and infection surveys were conducted during June/July of 2016.

	2007			**2010**			**2016**		
	N	Total possible	%	N	Total possible	%	N	Total possible	%
Number of villages where 1+ household reported owning bovines	36	36	100%	35	36	97.2%	8	10	80%
Number of villages where 1+ bovine tested	35	36	97.2%	31	35	88.6%	8	8	100%
Number of villages where:									
≥80% of bovines tested	18	35	51.4%	13	31	41.9%	4	8	50%
40 – 80 % tested	14	35	40.0%	14	31	45.2%	4	8	50%
<40% tested	3	35	8.6%	4	31	12.9%	0	8	0%
Total number of bovines owned, as reported on household surveys	675	--	--	371	--	--	95	--	--
Total number of bovines tested	503	675	74.5%	233	371	62.8%	72	95	58.6%
Hatch tests									
1	68	503	13.5%	34	233	14.6%	10	72	8.6%
2	91	503	18.1%	49	233	21.0%	9	72	20.7%
3	344	503	68.4%	150	233	64.4%	53	72	73.6%
Total number of bovines with household survey data	473	503	94.0%	197	233	84.5%	71	72	98.6%

medRxiv preprint doi: https://doi.org/10.1101/2021.08.20.21262368; this version posted August 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

S2 Table. Simple logistic regression analyses to determine the direction of association between bovine infection status and each predictor, by collection year. Tertiles (and sometimes quartiles) by year were used in simple logistic regression analyses to help investigate potential non-linearity. Results highlighted in gray indicate that the predictor was one of the top ten predictors in one or more RF analyses for a given collection year.

	2007				2010				2016			
	Tertile	**Point estimate**	**SE**	**P-value**	**Tertile**	**Point estimate**	**SE**	**P-value**	**Tertile**	**Point estimate**	**SE**	**P-value**
Individual characteristics												
Bovine age												
	≤ 3	ref			≤ 3	Ref						
	3.1 - 5	0.58	0.34	0.094*	3.1 - 6	0.10	0.46	0.824				
	≥ 5.1	0.92	0.36	0.011*	≥ 6.1	0.01	0.47	0.988				
# of hatch tests												
	1	Ref			1	p.f.p.	p.f.p	p.f.p.*	1	Ref		
	2	0.73	0.70	0.297	2	Ref			2	0.94	1.32	0.476
	3	1.24	0.61	0.043	3	0.50	0.49	0.308	3	0.88	1.11	0.426
Human sources												
Human infection: household count	0	Ref			0	Ref			0	Ref		
	1+	0.93	0.31	0.003	1+	1.75	0.45	<0.001*	1+	0.02	0.85	0.981
Human infection: village prevalence	≤ 2.6%	Ref			0%	p.f.p.	p.f.p	p.f.p.*	≤ 2.53%	Ref		
	2.7 - 11.9%	-0.24	0.40	0.558	0.1 – 10.8%	Ref			2.54 – 4.49%	0.99	0.71	0.164
	≥ 12%	1.06	0.33	0.001*	≥ 10.9%	1.11	0.43	0.010*	≥ 4.5%	-0.10	0.82	0.907
Socio-economic												
Vil: Imp. sanitation	≤ 7.2%	Ref			≤ 25%	Ref			≤ 48%	Ref		
	7.3 – 23.9%	1.29	0.35	<0.001	26 – 42%	0.87	0.43	0.046*	49 – 53%	p.f.p.	p.f.p	p.f.p.
	≥ 24%	0.36	0.41	0.379	≥ 43%	-0.99	0.62	0.112*	≥ 54%	0.61	0.65	0.346
Vil. asset score (0-9)	≤ 1.654	Ref			≤ 2.77	Ref			≤ 4.03	Ref		
	1.655 – 2.5	-1.09	0.35	0.002*	2.78 – 3.79	0.38	0.44	0.384	4.04 - 4.59	1.40	0.86	0.104*
	≥ 2.51	-1.15	0.34	<0.001*	≥ 3.8	-0.30	0.50	0.549	≥ 4.6	1.23	0.90	0.170*
HH asset score (0-9)					≤ 2	Ref						
	≤ 1	Ref			3	0.32	0.51	0.531	≤ 3	Ref		
	2 - 3	0.24	0.30	0.427	4	0.04	0.59	0.951	4	-1.06	0.89	0.234
	≥ 4	0.08	0.41	0.845	≥ 5	0.41	0.53	0.440	≥ 5	0.51	0.68	0.453
Animal reservoir												
Vil.: cat ownership	≤ 47.9%	Ref			≤ 44.9%	Ref			≤ 43.9%	Ref		
	48 – 64.9%	-0.61	0.37	0.095	45 – 55.9%	0.94	0.47	0.044	44 – 58.9%	2.81	1.11	0.012*
	≥ 65%	0.31	0.31	0.324	≥ 56%	0.13	0.53	0.813	≥ 59%	2.07	1.14	0.069*
Vil.: dog ownership	≤ 75%	Ref			≤ 71.9%	Ref			≤ 64%	Ref		
	76 - 85.2%	-0.43	0.31	0.166	72 - 80%	2.40	0.77	0.002	64.1 – 74.9%	0.24	0.82	0.769
	≥ 85.2%	-0.95	0.36	0.009	≥ 81%	2.38	0.77	0.002	≥ 75%	1.25	0.77	0.104*
HH: pigs owned					0	Ref						
					1	-0.42	0.67	0.534				
					>1	-1.50	0.56	0.008*				
Vil. mean pigs owned					≤ 0.49	Ref			≤ 0.14	Ref.		
					.5 – 0.84	0.34	0.43	0.436	0.21 – 0.43	-2.05	1.09	0.060*
					≥ 0.85	-0.38	0.52	0.470	≥ 0.44	-1.30	0.83	0.119*
Vil.: mean bovines owned	≤ 0.65	Ref			≤ 0.55	Ref			≤ 0.30	Ref		
	0.66 – 0.89	1.19	0.38	0.002*	0.56 – 0.74	-0.10	0.51	0.839	0.30 – 0.44	0.63	0.96	0.513
	≥ 0.89	1.16	0.40	0.003*	≥ 0.75	0.77	0.45	0.086*	≥ 0.441	2.24	0.86	0.009*

medRxiv preprint doi: https://doi.org/10.1101/2021.08.20.21262368; this version posted August 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Agriculture												
Vil. mean rice area												
	≤ 0.84	Ref			≤ 0.61	Ref			≤ 0.95	Ref		
	0.84 – 1.47	-0.33	0.37	0.380	0.62 – 1.29	2.33	0.77	0.003*	0.96 – 1.34	-0.97	0.70	0.166*
	≥ 1.47	0.55	0.32	0.081*	≥ 1.30	2.35	0.77	0.002*	≥ 1.35	-1.61	0.86	0.061*
HH rice area												
	≤ 0.5	Ref			≤ 0.5	Ref			≤ 1.2	Ref		
	0.6 – 1.5	0.36	0.35	0.301	0.6 – 1.5	0.57	0.49	0.253	1.2 – 2.3	-0.49	0.67	0.463
	≥ 1.6	0.54	0.34	0.114	≥ 1.6	0.76	0.48	0.116*	≥ 2.4	-1.25	0.87	0.149
Vil sum. crop area												
	≤ 0.73	Ref			≤ 2.03	Ref			≤ 2.412	Ref		
	0.74 – 1.96	1.51	0.39	<0.001*	2.03– 2.76	0.02	0.44	0.966	2.413 – 2.98	0.88	0.92	0.336
	≥ 1.97	0.90	0.42	0.031*	≥ 2.77	-0.52	0.49	0.285	≥ 2.99	1.81	0.86	0.034*
HH sum. crop area												
	≤ 1	Ref			≤ 2.2	Ref			≤ 2.8	Ref		
	1.1 – 2.2	0.69	0.34	0.041*	2.3 – 3.8	-0.09	0.43	0.830	2.81 – 4.1	1.70	0.86	0.046*
	≥ 2.3	0.40	0.33	0.230	≥ 3.9	-0.69	0.50	0.167*	≥ 4.2	0.84	0.92	0.362
Vil. win. crop area*												
	≤ 1.869	Ref										
	1.87 – 2.55	0.01	0.40	0.981	≤ 1.869	Ref			≤ 1.58	Ref		
	2.56 – 2.81	0.01	0.40	0.981	1.87 – 2.58	-0.50	0.45	0.269	1.59 – 1.89	-0.86	0.70	0.219
	≥2.82	0.48	0.37	0.193*	≥ 2.59	-0.60	0.46	0.193*	≥ 1.9	-1.55	0.86	0.071*
HH win. crop area*												
	1.79	Ref			≤ 1.9	Ref			≤ 1.2	Ref		
	1.8 –2.5	-0.03	0.41	0.944	2 – 2.8	-0.07	0.53	0.895	1.3 – 2.1	-0.62	0.81	0.442
	2.6 – 3.7	0.54	0.38	0.152*	2.9 – 3.7	0.36	0.51	0.617	2.2 – 3.2	-0.76	0.80	0.340
	≥ 3.8	0.32	0.39	0.413	≥ 3.8	-0.23	0.54	0.678	≥ 3.3	-1.17	0.90	0.193*
Vil. night soil sum.												
	≤ 7.3	Ref			≤ 6.39	Ref			≤ 11.2	Ref		
	7.4 – 37	1.18	0.38	0.002*	6.4 – 14.4	-0.12	0.41	0.766	11.3 – 27.9	2.48	1.11	0.025*
	≥ 38	1.06	0.38	0.005*	≥ 14.5	-1.80	0.66	0.006*	≥ 28	1.95	1.14	0.087*
Vil. night soil winter												
	≤ 23.6	Ref			≤ 2.589	Ref			≤ 2.9	Ref		
	23.7 – 42.6	-0.69	0.44	0.120*	2.59 – 11.2	-0.82	0.44	0.066	3 – 4.2	0.41	0.87	0.642
	≥ 42.7	1.04	0.32	0.001*	≥ 11.3	-1.39	0.54	0.010	≥ 4.3	1.79	0.76	0.019
HH night soil winter*												
	0	0.90	0.56	0.105*								
	1-20	Ref										
	21 – 59	1.11	0.59	0.060*	0	Ref			0	Ref		
	≥ 60	0.96	0.58	0.097*	>0	-1.71	0.63	0.006	>0	2.00	0.76	0.009

*In the case that the tertiles did not show evidence of any moderate difference (p<0.2) between one or more groups, quartiles were tried. When still no difference was found between groups, this was noted in results Table 4.

p.f.p = perfect failure predicted.

medRxiv preprint doi: https://doi.org/10.1101/2021.08.20.21262368; this version posted August 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

S1 Figure. Supplemental analysis assessing changes over time. Two additional RF model iterations were run for each collection year that only included those predictors that were available in all three of the collection years. The top ten predictors for these two iterations were given a score of 1-10, and the summed scores were used to determine the variable ranking 1st – 10th for each collection year, as well as a final variable ranking "all year score" that summed the rankings across all six iterations (two per collection year) conducted.

	2007	2010	2016	All year score
Physical characteristics				
Number of hatch tests		2		9
County of residence				
Infection				
Human infection: household count		9		
Human infection: village prevalence	5	1		4*
Socio-economic indicators				
Village prevalence of improved sanitation			10	
Household has improved sanitation (y/n)				
Village mean asset score (0-9)	10		4	10
Household asset score (0-9)				
Animal ownership				
Village prevalence of cat ownership			9	
Household cat ownership				
Village prevalence of dog ownership			6	
Household bovine ownership				
Village prevalence of bovine ownership	8		1*	8
Agriculture				
Village mean rice area	6	5	7*	6
Household rice area		6		
Village mean dry summer crop area	2	10	7*	7
Household dry summer crop area	4	4		4*
Village mean winter crop area	1	8	3	2
Household winter crop area	3	7	5	3
Village night soil rice: mean # buckets				
Household night soil rice: # buckets				
Village night soil summer dry crop: mean # buckets	7	3	1*	1
Household night soil summer dry crop: # buckets				
Village night soil winter crop: mean # buckets				
Household night soil winter crop: # buckets	9			

* = Tied for importance rank
= Not ranked 1-10 for the given model/year

Color key: Variable importance rankings (1st-10th) by scale

Individual	1st – 2nd	3rd – 4th	5th – 6th	7th – 8th	9th – 10th
Household	1st – 2nd	3rd – 4th	5th – 6th	7th – 8th	9th – 10th
Village	1st – 2nd	3rd – 4th	5th – 6th	7th – 8th	9th – 10th

S2 Figure. Correlation matrix for 2007 predictors. A correlation matrix for predictors included in the 2007 RF models is provided to highlight those predictors whose relative variable ranking positions may be less reliable due to correlation with other influential predictors. Only predictors with a correlation coefficient of < -0.499 or > 0.499 are included. The 2007 correlation matrix demonstrates that there are some strongly correlated predictors, particularly in the agricultural predictor category, that may be impacting their relative importance rankings.

	County	Hatch tests	HH asset	HH rice crop area	HH sum. crop area	HH win. crop area	HH NS sum. crop	Vil. asset	Vil cat own	Vil. bov. own	Vil hum. Inf. prev.	Vil rice area	Vil. sum. crop area	Vil. win. crop area	Vil. NS sum. crop	Vil. NS win. crop	Vil NS Rice crop
County	1.0																
Hatch tests	0.553	1.0															
HH asset			1.0														
HH rice area				1.0													
HH sum. crop area	0.578				1.0												
HH win. crop area					0.559	1.0											
HH NS sum. Crop					0.594		1.0										
Vil Asset	-0.728		0.576		-0.501			1.0									
Vil Cat Own								-0.565	1.0								
Vil Bov. Own								-0.520		1.0							
Vil Hum. Inf. Prev.											1.0						
Vil rice area	-0.664			0.691	-0.575			0.544				1.0					
Vil sum. crop area	0.799				-0.699			-0.632				-0.696	1.0				
Wil win. crop area													0.648	1.0			
Vil NS sum. crop	0.674				0.644			-0.626				-0.599	0.814	0.712	1.0		
Vil NS win. Crop													0.547			1.0	
Vil NS rice crop											0.598						1.0

Color Key:

Moderate positive correlation	0.50 – 0.599	Moderate negative correlation	-0.599– -0.50
Strong positive correlation	0.60 – 0.799	Strong negative correlation	-0.799 – -0.60
Very strong positive correlation	≥ 0.80	Very strong negative correlation	≤ -0.80

S3 Figure. Correlation matrix for 2010 predictors. A correlation matrix for predictors included in the 2010 RF models is provided to highlight those predictors whose relative variable ranking positions may be less reliable due to correlation with other influential predictors. Only predictors with a correlation coefficient of < -0.499 or > 0.499 are included. The 2010 correlation matrix demonstrates that there are just a few strongly correlated predictors in the agricultural predictor category. As well as the socio-economic indicator category that may be impacting relative importance rankings.

	County	HH rice crop area	HH sum. crop area	HH win. crop area	HH NS sum. crop	HH NS win. crop	Vil. asset	Vil cat own	Vil. bov. own	Vil pigs	Vil rice area	Vil. sum. crop area	Vil. win. crop area	Vil. NS sum. crop	Vil. NS win. crop
County	1.0														
HH rice area		1.0													
HH sum. crop area			1.0												
HH win. crop area			0.575	1.0											
HH NS sum. Crop					1.0										
HH NS win. Crop					0.565	1.0									
Vil Asset	-0.771						1.0								
Vil Cat Own	0.589						-0.533	1.0							
Vil Bov. Own	0.556						-0.525		1.0						
Vil pigs										1.0					
Vil rice area		0.553									1.0				
Vil sum. crop area	0.576						-0.569	0.535				1.0			
Vil win. crop area								0.789		0.515		0.675	1.0		
Vil NS sum. crop												0578		1.0	
Vil NS win. crop						0.509								0.878	1.0

Color Key:

Moderate positive correlation	0.50 – 0.599	Moderate negative correlation	-0.599– -0.50
Strong positive correlation	0.60 – 0.799	Strong negative correlation	-0.799 – -0.60
Very strong positive correlation	≥ 0.80	Very strong negative correlation	≤ -0.80

S4 Figure. Correlation matrix for 2016 predictors. A correlation matrix for predictors included in the 2016 RF models is provided to highlight those predictors whose relative variable ranking positions may be less reliable due to correlation with other influential predictors. Only predictors with a correlation coefficient of < -0.499 or > 0.499 are included. The 2016 correlation matrix demonstrates that there are several strongly correlated predictors across the different predictor categories that may be impacting relative importance rankings for the 2016 RF models.

	County	Hatch tests	HH hum. inf. prev	HH rice crop area	HH win. crop area	HH NS sum. crop	HH NS rice crop	Vil hum inf prev	Vil. asset	Vil toilet imp	Vil. dog own	Vil cat own	Vil pigs own	Vil rice crop area	Vil. sum. crop area	Vil. win. crop area	Vil. NS sum. crop	Vil. NS win. crop	Vil. NS rice crop
County	1.0																		
Hatch tests	0.615	1.0																	
HH hum. Inf.	-0.625		1.0																
HH rice area				1.0															
HH win. crop area				0.806	1.0														
HH NS sum. crop						1.0													
HH NS rice crop						0.834	1.0												
Vil. hum. inf prev.	-0.839	-0.641	0.661					1.0											
Vil asset									1.0										
Vil toilet Imp.										1.0									
Vil. dog own											1.0								
Vil cat Own											0.800	1.0							
Vil pigs	-0.502		0.611					0.838	-0.563				1.0						
Vil rice area								0.669	-0.635			0.522	0.858	1.0					
Vil sum. crop area								0.501	-0.612			0.721	0.801	0.678	1.0				
Vil win. crop area								0.590	-0.653			0.640	0.820	0.955	0.774	1.0			
Vil NS sum. crop										-0.645							1.0		
Vil NS win. crop	0.649							-0.546		-0.814							0.578	1.0	
Vil. NS rice crop																	0.581		1.0

Color Key:

Moderate positive correlation	0.50 – 0.599	Moderate negative correlation	-0.599– -0.50
Strong positive correlation	0.60 – 0.799	Strong negative correlation	-0.799 – -0.60
Very strong positive correlation	≥ 0.80	Very strong negative correlation	≤ -0.80

750