INTEGRATING PUBLIC HEALTH SURVEILLANCE AND ENVIRONMENTAL DATA TO MODEL PRESENCE OF *HISTOPLASMA* IN THE UNITED STATES

Staci A. Hepler[1], **Kimberly A. Kaufeld**[2], **Kaitlin Benedict**[3], **Mitsuru Toda**[3], **Brendan R. Jackson**[3], **Xiaonan Liu**[1], **David Kline**[4]

[1]Mathematics and Statistics, Wake Forest University, Winston-Salem, NC.

[2]Statistical Sciences Group, Los Alamos National Laboratory, Los Alamos, NM.

[3]Centers for Disease Control and Prevention, Atlanta, GA.

[4]Biostatistics and Data Science, Wake Forest School of Medicine, Winston-Salem, NC.

Abstract

Background: In the United States, the true geographic distribution of the environmental fungus *Histoplasma capsulatum* remains poorly understood but appears to have changed since it was first characterized. Histoplasmosis is caused by inhalation of the fungus and can range in severity from asymptomatic to life-threatening. Due to limited public health surveillance and under detection of infections, it is challenging to directly use reported case data to characterize spatial risk.

Methods: Using monthly and yearly county-level public health surveillance data and various environmental and socioeconomic characteristics, we use a spatio-temporal occupancy model to estimate latent, or unobserved, presence of *H. capsulatum*, accounting for imperfect detection of histoplasmosis cases.

Results: We estimate areas with higher probabilities of the presence of *H. capsulatum* in the East North Central states around the Great Lakes, reflecting a shift of the endemic region to the north from previous estimates. The presence of *H. capsulatum* was strongly associated with higher soil nitrogen levels.

Conclusions: In this investigation, we were able to mitigate challenges related to reporting and illustrate a shift in the endemic region from historical estimates. This work aims to help inform future surveillance needs, clinical awareness, and testing decisions for histoplasmosis.

Keywords

Fungal Disease; Spatio-temporal; occupancy model; Histoplasmosis; Public health surveillance; imperfect detection

Introduction

Histoplasmosis is an infection caused by the inhalation of the environmental fungus *H. capsulatum* that ranges from asymptomatic infection to life-threatening disseminated disease.[1] In the United States, histoplasmosis has traditionally been associated with the areas around the Ohio and Mississippi River Valleys, a finding established by nationwide skin test reactivity surveys conducted during the 1940s and 1950s.[3] However, cases acquired far outside these areas suggest that histoplasmosis is more widespread than originally thought.[4,5,6] Furthermore, public health surveillance data, though subject to underdetection of cases and limited to only a dozen states that require reporting of histoplasmosis, indicate that cases routinely occur in north-central states not previously considered to be endemic.[7,8]

Our current population-level understanding of histoplasmosis is primarily based on passive public health disease surveillance data. Before 2016, there was no standardized national case definition for histoplasmosis, leading to variation in definitions across states (See eAppendix Table 1). Since then, the case definition has been standardized.[8] Despite these advances, surveillance remains limited in geographic scope and in the level of detail of the data collected.[9] Additionally, understanding the spatial risk of histoplasmosis is complicated by challenges with detection.[10] For example, histoplasmosis is likely under detected and frequently misdiagnosed because the signs and symptoms can be similar to those of other common respiratory illnesses.[8,11] A prior study developed a suitability score for *H. capsulatum* based on environmental characteristics, but did not account for under detection of reported histoplasmosis cases.[12] Ignoring under detection can result in biased statistical inference about the presence of *H. capsulatum*.[13]

We developed a spatio-temporal occupancy model to estimate the endemic region for histoplasmosis.[14] The model relates reported histoplasmosis cases to latent, or unobserved, presence of *H. capsulatum* and accounts for imperfect detection in the reported cases. This represents a novel application of a common ecological model to a problem in environmental and infectious disease epidemiology. Our results provide an increased understanding of the areas with the highest probability of the presence of *H. capsulatum*. This information is essential for guiding both healthcare providers' testing decisions and public health prevention strategies.

METHODS

Data Sources.

We used county-level data on histoplasmosis cases reported to health departments from 12 states (Alabama, Arkansas, Delaware, Illinois, Indiana, Kentucky, Michigan, Minnesota, Mississippi, Nebraska, Pennsylvania, and Wisconsin) during 2011–2014, the most recently available time period (Figure 1). This time period precedes the standardized definition defined in 2016. Monthly data were available from all states except for Delaware and Kentucky, which were reported yearly.[8]

Variables considered to help explain the environmental presence of the fungus included land cover characteristics, nitrogen levels, latitude and longitude, and elevation. To explain

the presence or absence of histoplasmosis diagnoses, we considered cultivated crops, temperature, soil moisture, total population, and socio-economic variables. The land cover characteristics were from the 2010 National Land Cover Database (NLCD), and for this analysis we considered the proportion of each county with cultivated crops, the proportion covered with water, and the proportion that is undeveloped.[15] We obtained soil nitrogen levels from the United States Geological Survey[16] based upon estimated county-level farm and nonfarm nitrogen fertilizer use, in kilograms, from commercial fertilizer sales. Elevation was from the 2010 Global Multi-resolution Terrain Elevation Data (GMTED2010).[17] We obtained temperature data from the National Oceanic and Atmospheric Administration (NOAA) National Centers for Environmental Information (NCEI) and was provided at the locations of the weather stations.[18] We computed daily county-level temperature by averaging the stations within each county. For counties with no weather stations, we averaged the temperature of the adjacent counties. We generated monthly data by averaging the daily values. Surface soil moisture was from the National Integrated Drought Information System (NIDIS) supported by the Climate Prediction Center (CPC).[19] The values estimated by the CPC soil moisture tool as part of the National Weather Service Global Forecast System is calculated using a one-layer hydrologic model[20,21] that calculates soil moisture, evaporation, and runoff using as forcing observed precipitation and temperature. The data provides daily calculated soil moisture values in a 10km x 10km pixel area, which were averaged to the county level by attributing the pixels to the respective county based upon the bounding box and scaled to the monthly level by averaging across the days. We used the American Community Survey 2014 5-year county-level estimates of total population, proportion with private health insurance, and proportion employed in agriculture to characterize county demographics.[22]

Statistical Model.

We used a Bayesian occupancy model (full model specification in eAppendix) to relate the binary indicator of reported histoplasmosis cases to latent presence of *H. capsulatum*.[14] By converting reported histoplasmosis case counts to binary indicators, we mitigate some of the reporting differences across states while retaining the ability to learn about the latent presence of H. capsulatum in the environment. This model accounts for the fact that presence of *H. capsulatum* does not always result in histoplasmosis infection, and even if histoplasmosis is contracted, it may go undetected (i.e. undiagnosed or unreported to public health authorities). Failure to account for the fact that zero cases may be reported even when there is exposure to *H. capsulatum* can result in biased statistical inference.[13] By focusing on presence of *H. capsulatum*, we address a question of interest better suited to the available data that avoids bias due to imperfect detection that would be present in a more traditional count model for rates of disease.

For county i $(i = 1,\dots,943)$ and month t $(t = 1,\dots,48)$, define Y_{it} to be the binary indicator that at least one case of histoplasmosis was reported. Let Z_i be an indicator of presence of *H. capsulatum* in county i. While histoplasmosis cases are reported monthly, we assume presence of *H. capsulatum* is constant over this 4-year period. This assumption is reasonable based on the short time frame under which data from this study were collected relative to known survival rates of *H. capsulatum*.[23] The proposed occupancy model assumes that if

$H.$ $capsulatum$ is absent ($Z_i = 0$), then there are no reported cases of histoplasmosis in that county throughout the study period ($Y_{it} = 0$ for all $t = 1, \ldots, T$). However, if $H.$ $capsulatum$ is present ($Z_i = 1$), then there may or may not be any reported cases for any month within the study period ($Y_{it} = 0$ or $Y_{it} = 1$ for any given month t, where $t = 1, \ldots, 48$). To incorporate all data available, we let $Y_i^{(k)}$ be the indicator that there was at least one reported case of histoplasmosis in county i during year $k = 1,2,3,4$, and we set $Y_i^{(k)}$ equal to 1 if $Y_{it} = 1$ for any month within year k and 0 otherwise. This hierarchical model allows us to simultaneously model all counties regardless of reporting frequency, and we borrow strength from the counties with monthly observations to infer monthly outcomes for the remaining counties.

Define $\psi_i \equiv P(Z_i = 1)$ as the probability of $H.$ $capsulatum$ presence in county i and $p_{it} = P(Y_{it} = 1 | Z_i = 1)$ the detection probability, or the probability of there being at least one reported case of histoplasmosis in county i during month t given that $H.$ $capsulatum$ is present. We assume the following Bayesian occupancy model:

$$Y_i^{(k)} = 1 - \prod_{t = 12k - 11}^{12k} (1 - Y_{it})$$

$$Y_{it} \mid Z_i, \ p_{it} \sim Bernoulli(Z_i p_{it})$$

$$Z_i \mid \psi_i \sim Bernoulli(\psi_i).$$

For the counties where monthly data are available, Y_{it} is observed. However, for the counties with yearly data, Y_{it} is a latent binary random variable for each t. We assume a probit link for each probability, ψ_i and p_{it}, and relate these quantities to environmental variables and spatial or temporal random effects.[24,25] The probit link was chosen for computational considerations (see eAppendix Section 1.2 for more details). More specifically, we assume the probability of the presence of $H.$ $capsulatum$ relates to environmental covariates and a spatial random effect that accounts for our belief that $H.$ $capsulatum$ is more likely to be present in a county if it is present in neighboring counties. That is,

$$\psi_i = \Phi(X_i \alpha + \eta_i)$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution, X_i is a vector of standardized covariates related to exposure risk, α is a vector of regression coefficients, and η_i is a spatial random effect (see eAppendix for detailed specification). We assume the probability of observing at least one diagnosed case of histoplasmosis, given $H.$ $capsulatum$ is present relates to land-use and socio-environmental covariates, as well as population size since more populous counties are more likely to have a detected case. The model for the detection probability includes a random effect to account for temporal autocorrelation within a county. That is,

$$p_{it} = \Phi(\boldsymbol{W}_{it}\boldsymbol{\beta} + v_{it}),$$

where $\boldsymbol{W}_{it}$ is a vector that contains standardized covariates related to detectability, and $\boldsymbol{\beta}$ is a vector of regression coefficients. Note that $\boldsymbol{W}_{it}$ is specified to assume state-specific intercepts so that the intercepts represent a statewide average detection rate. We include the temporal random effect v_{it} to account for the notion that detectability within county i during year t is likely related to detectability for county i during year $t-1$. Thus, we assume for $t=1$ $v_{it} \sim N(0, \tau^2)$ and for $t = 2, ..., T$, $v_{it} \sim N(\rho v_{i,t-1}, \tau^2)$. The temporal random effects are assumed to be independent across space. For identifiability of the temporal random effect, we enforce a mean-zero centering constraint when $t=1$.

Our model is fit within the Bayesian framework. Details on the prior distributions and the Markov chain Monte Carlo (MCMC) algorithm used to simulate from the posterior distribution are in eAppendix Section 1.2. We initially included a larger set of explanatory variables and performed reversible jump MCMC to determine which variables had a relatively high posterior probability of inclusion in the model.[26] See eAppendix Section 2.1 for details. The final model was then fit using the variables listed in Tables 1 and 2.

Results

We see strong and positive associations between the presence of *H. capsulatum* and log farm nitrogen soil content, log non-farm nitrogen soil content, and the percent of the county covered by water (Table 1). We see a negative association with log elevation, indicating a lower probability of presence in areas of higher elevation. We estimated the presence of *H. capsulatum* for the 12 states for which we had histoplasmosis case data and for the states in the immediate surrounding area. We observe the highest estimated probability of the presence of *H. capsulatum* in the northern part of the study region consisting of the East North Central United States. We also observe high estimated probabilities along the Atlantic coastal plain. The probability increases moving northward and eastward through the study region (Figure 2A). As expected, we see higher standard errors in counties that are further from those in the observed data (Figure 2B).

The state-specific estimated intercepts and credible intervals for the detection probability are shown in Figure 2(C). This figure shows that detecting cases of histoplasmosis varies from state to state, and the highest average estimated probability of detection was in Minnesota. Figure 2(D) shows the estimated detection probability for each county averaged over the 48 months, showing a large amount of spatial heterogeneity in histoplasmosis detection. Maps of the county-level detection probabilities for each of the 48 months can be found in the eAppendix.

Table 2 summarizes the posterior distributions for the detection probability regression coefficients. We estimate that more populated areas are more likely to have reported cases of histoplasmosis given the presence of *H. capsulatum*. We observed a negative relationship with the percentage of residents with private health insurance, and we see a moderately positive relationship with having a larger land area covered by cultivated crops.

Discussion

We used a Bayesian occupancy model to relate monthly reported cases of histoplasmosis to the latent presence of *H. capsulatum*, while accounting for differences in case reporting frequency. Histoplasmosis is only reportable in a small number of states and often goes undetected. Thus, analyzing the case data without accounting for imperfect detection of histoplasmosis will likely underestimate the true presence and yield biased statistical inference. In addition, we quantified the relationship between geographical variation in the presence of *H. capsulatum* and environmental characteristics. By incorporating information about environmental covariates and accounting for spatial dependence, we were able to predict the probability of presence for states neighboring the 12 states for which we had histoplasmosis case data. We found that the northern Midwestern United States and the eastern Atlantic coastal region had the highest estimated probability of the presence of *H. capsulatum*, suggesting that the spatial risk of histoplasmosis has changed from what was originally described in the 1950s, which focused counties inside a triangle with points in central Ohio, southern Iowa, and northern Louisiana.[3] This finding is consistent with previous work suggesting an expansion of the endemic areas.[8,12] Specifically, the high probability of histoplasmosis detection in Minnesota is consistent with previous work and may reflect a strong surveillance system and a broad case definition at the time the data were collected[8].

Although occupancy models are commonly used in ecological applications, they are less common in public health. However, imperfect detection is likely present in many public health outcomes due to gaps in reporting systems and undiagnosed or misdiagnosed cases. Failure to explicitly account for imperfect detection can result in biased inference since the geographic variability in disease presence and detectability are otherwise confounded. We have illustrated how this modeling framework can be applied to public health data to quantify spatial variability and identify geographic regions of increased risk.

Our finding of a positive association of histoplasmosis with soil nitrogen content is not surprising. *H. capsulatum* is well-known to thrive in soil enriched with high levels of nitrogen and phosphate.[2,10] Future work could incorporate data that quantifies the geographical distribution of birds and bats (e.g. ebird[27] and North American Bat monitoring program[28]). The association we observed between expected presence of *H. capsulatum* with cultivated crops also supports previous work showing that a high proportion of people with histoplasmosis live in rural areas, and that people with jobs in the agriculture industry may be at increased risk.[9] Similarly, historical evidence from skin test studies shows higher rates of histoplasmosis among people who lived on farms.[29]

One limitation of our approach is that we transformed the reported case data to a binary outcome and thus cannot describe rates of infection. Future work could model raw case counts to inform potential exposure risk. The reporting differences between states coupled with the large proportion of infections that go undetected make this challenging as the reported counts do not accurately reflect the true underlying incidence of histoplasmosis in a given county. Another limitation is the model's inability to predict results for the entire United States, as we did not have evidence to determine whether or not the relationships

with the included environmental variables hold outside of our study region. Last, future studies could also investigate host-related covariates. For example, persons aged 65 and older and those with immunosuppression are more likely to have severe histoplasmosis and are thus more likely to be diagnosed. By incorporating demographic information or information on underlying health status, the model's detection process could be improved.

In conclusion, we have generated a map that quantifies the geographical distribution of endemic regions for histoplasmosis in a relatively large region of the United States, providing a better understanding of the areas where the presence of *H. capsulatum* is likely. A more comprehensive characterization of its spatial distribution is important for guiding clinical decision making and public health surveillance and prevention activities.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

References

1. Kauffman CA Histoplasmosis: a clinical and laboratory update. Clinical microbiology reviews. 2007;20(1):115–132. [PubMed: 17223625]

2. Deepe GS Jr Outbreaks of histoplasmosis: the spores set sail. PLoS pathogens. 2018;14:e1007213. [PubMed: 30212569]

3. Manos NE, Ferebee SH, Kerschbaum WF Geographic variation in the prevalence of histoplasmin sensitivity. Chest. 1956;29:649–668.

4. Benedict K, Thompson III GR, Deresinski S, Chiller T Mycotic infections acquired outside areas of known endemicity, United States. Emerging infectious diseases. 2015; 21:1935–1941. [PubMed: 26485441]

5. Ashraf N, Kubat RC, Poplin V, Adenis AA, Denning DW, Wright L, McCotter O, Schwartz IS, Jackson BR, Chiller T, Bahr NC Re-drawing the maps for endemic Mycoses. Mycopathologia. 2020; 185(5):843–865. [PubMed: 32040709]

6. Lockhart SR, Toda M, Benedict K, Caceres DH, Litvintseva AP, Endemic and other dimorphic Mycoses in the Americas. Journal of Fundi (Basel). 2021; 7(2): 151.

7. https://www.cdc.gov/fungal/fungal-disease-reporting-table.html

8. Armstrong PA, Jackson BR, Haselow D, et al. Multistate epidemiology of histoplasmosis, United States, 2011–2014. Emerging infectious diseases. 2018;24(3):425–431. [PubMed: 29460731]

9. Benedict K, McCracken S, Signs K, Ireland M, Amburgey V, Serrano JA, … & Jackson BR Enhanced surveillance for histoplasmosis—9 states, 2018–2019. Open Forum Infectious Diseases. Volume 7, Issue 9, September 2020, ofaa343. US: Oxford University Press. [PubMed: 32964064]

10. Benedict K, Mody RK Epidemiology of histoplasmosis outbreaks, United States, 1938–2013. Emerging Infectious Diseases. 2016;22(3):370–378. [PubMed: 26890817]

11. Benedict K, Kobayashi M, Garg S, Chiller T, & Jackson BR Symptoms in Blastomycosis, Coccidioidomycosis, and Histoplasmosis Versus Other Respiratory Illnesses in

Commercially Insured Adult Outpatients—United States, 2016–2017. Clinical Infectious Diseases. 2020;ciaa15154.

12. Maiga AW, Deppen S, Scaffidi BK, Baddley J, Aldrich MC, Dittus RS, Grogan EL Mapping *Histoplasma capsulatum* exposure, United States. Emerging Infectious Diseases. 2018;24(10):1835–1839. [PubMed: 30226187]

13. Gu W, Swihart RK Absent or undetected? Effects of non-detection of species occurrence on wildlife-habitat models. Biological Conservation. 2004;116(2):195–203.

14. Royle JA, Dorazio RM Hierarchical modeling and inference in ecology: the analysis of data from populations, metapopulations and communities. Elsevier 2008

15. Multi-Resolution Land Characteristics (MRLC) Consortium National Land Cover Database (NLCD). http://mrlc.gov/data.

16. Brakebill JW and Gronberg JM, 2017, County-Level Estimates of Nitrogen and Phosphorus from Commercial Fertilizer for the Conterminous United States, 1987–2012: U.S. Geological Survey data release, h10.5066/F7H41PKX.

17. U.S. Geological Survey Global Multi-resolution Terrain Elevation Data 2010 (GMTED2010). Accessed from: http://usgs.gov/centers/eros/data-tools.

18. National Centers for Environmental Information (NCEI), National Oceanic and Atmospheric Administration (NOAA). Accessed from http://ncdc.noaa.gov/data-access.

19. National Integrated Drought Information System (NIDIS), National Oceanic and Atmospheric Administration (NOAA). Accessed from http://drought.gov/topics/soil-moisture.

20. Huang J, van den Dool HM, Georgarakos KP. Analysis of model-calculated soil moisture over the United States (1931–1993) and applications to long-range temperature forecasts. Journal of Climate. 1996 Jun;9(6):1350–62

21. Van den Dool Huug, Huang Jin, and Fan Yun. Performance and analysis of the constructed analogue method applied to US soil moisture over 1981–2001. Journal of Geophysical Research: Atmospheres 2003; 108.D16.

22. U.S. Census Bureau; American Community Survey, 2014 American Community Survey 5-Year Estimates. Accessed from http://data.census.gov.

23. Mahvi TA Factors governing the epidemiology of Histoplasma capsulatum in soil. Mycopathologia et mycologia applicata. 1970;41:167–176. [PubMed: 5535367]

24. Johnson DS, Conn PB, Hooten MB, Ray JC, Pond BA Spatial occupancy models for large data sets. Ecology. 2013;94:801–808.

25. Hepler SA, Erhardt RJ A spatiotemporal model for multivariate occupancy data. Environmetrics. 2021;32(2):e2657.

26. Robert C, Casella G Monte Carlo Statistical Methods. Springer Science & Business Media 2013.

27. eBird, The Cornell Lab or Ornithology, http://ebird.org.

28. The North American Bat Monitoring Program, U.S. Geological Survey, https://sciencebase.usgs.gov/nabat/.

29. Edwards LB, Acquaviva FA, Livesay VT, et al. An atlas of sensitivity to tuberculin, PPD-B, and histoplasmin in the United States. Am Rev Respir Dis. 1969; 99:Suppl:1–132. [PubMed: 4973586]
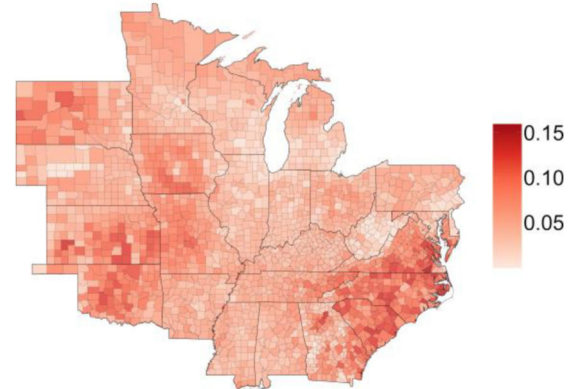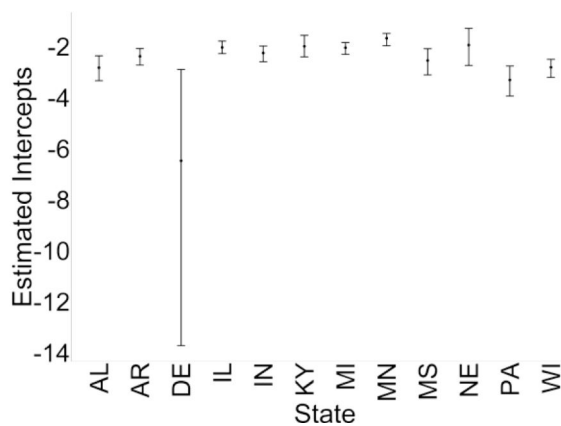
FIGURE 1.
Map of the 12 states included in our analysis. Counties shaded grey had at least one case of histoplasmosis reported to public health authorities during the four-year study period.

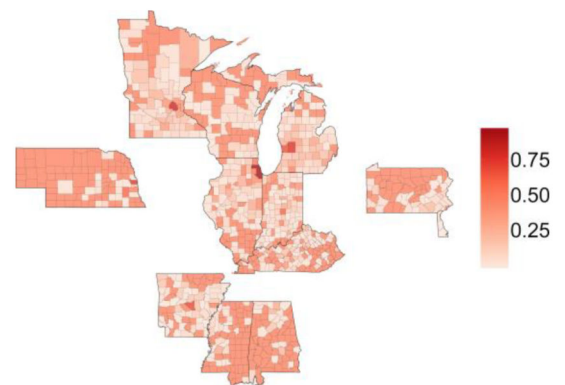(A) Probability of the presence of *H. capsulatum*
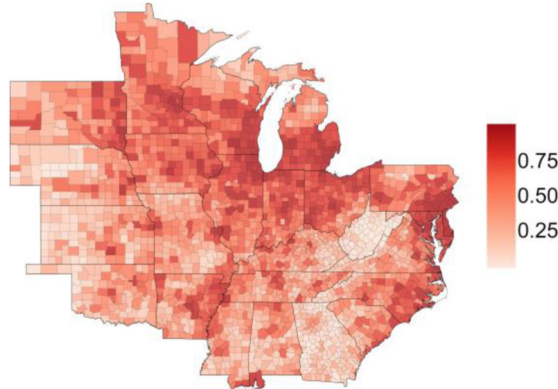
(B) Standard error of estimated posterior probability of presence

(C) Estimates and 95% credible intervals of the state-specific detection posterior probability intercepts

(D) Average detection posterior probability

Figure 2.

(A) Map of the estimated posterior probability of presence of *H. capsulatum*. (B) Standard errors of the estimated posterior probability. (C) Estimates and credible intervals of the state-specific intercepts for the probability of detecting a case of histoplasmosis, given *H. capsulatum* is present. (D) County-level estimated detection posterior probability averaged over the 48 months. Note that the low estimate and wide variability for the average rate of histoplasmosis detection in Delaware shown in (C) is likely due to the fact that Delaware only has three counties with just two reported diagnosed cases throughout the study period.

TABLE 1.

The posterior mean and 95% credible intervals for the covariate effects of the probability of presence of *H. capsulatum*. The last column is the posterior probability of the regression coefficient being positive.

Variable	Estimate	CI	P(>0|data)
Intercept	−0.0375	(−0.1373, 0.0667)	0.2480
Cultivated Crops	0.1566	(−0.0460, 0.3453)	0.9520
Log Farm Nitrogen	0.2272	(0.0442, 0.4217)	0.9880
Log Non-Farm Nitrogen	0.6436	(0.5228, 0.7785)	1.0000
Log Elevation	−0.1922	(−0.3991, −0.0135)	0.0200
Latitude	0.1865	(0.0520, 0.3404)	0.9940
Longitude	0.1898	(0.0692, 0.3300)	1.0000
Land Cover -Water	0.1050	(0.0072, 0.2051)	0.9840

TABLE 2.

The posterior mean and 95% credible intervals of the covariate effects for diagnosed cases of histoplasmosis given presence of *H. capsulatum*. The last column is the posterior probability of the regression coefficient being positive.

Variable	Estimate	CI	P(>0|data)
Cultivated Crops	0.0829	(−0.0620, 0.2321)	0.8660
Soil Moisture	0.0055	(−0.0880, 0.1159)	0.5440
Log Population	0.7292	(0.5525, 0.9288)	1.0000
Log Percent Non-White	0.0040	(−0.1552, 0.1659)	0.5040
Agriculture	0.0290	(−0.1422, 0.1926)	0.6520
Private Insurance	−0.1247	(−0.2771, 0.0126)	0.0380