Vibrio Cholerae O1 Transmission in Bangladesh: Insights from a Nationally-Representative Serosurvey Supplement

Andrew S Azman[a], Stephen A Lauer[a], M. Taufiq Rahman Bhuiyan[b], Francisco J Luquero[c,d], Daniel T Leung[e], Sonia Hegde[a], Jason Harris[f,g,h], Kishor Kumar Paul[b], Fatema Khaton[b], Jannatul Ferdous[b], Justin Lessler[a], Henrik Salje[i,a], Firdausi Qadri[b], Emily S Gurley[a,b]

[a]*Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, USA*
[b]*icddr,b, Dhaka, Bangladesh*
[c]*Epicentre, Paris, France*
[d]*Department of International Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, USA*
[e]*Division of Infectious Diseases, University of Utah School of Medicine, Salt Lake City, USA*
[f]*Division of Infectious Diseases, Massachusetts General Hospital, Boston, USA*
[g]*Division of Global Health, Massachusetts General Hospital, Boston, USA*
[h]*Department of Pediatrics, Harvard School of Medicine, Boston, USA*
[i]*Mathematical Modelling of Infectious Diseases Unit, Institut Pasteur, Paris, France*

S1. Cholera seroincidence model and inference

The primary goal of these analyses is to estimate the proportion of the population infected by *Vibrio cholerae* O1 in the previous year, which we refer to as the 'seroincidence rate' and denote as π. We use a previously validated random forest model to classify whether each member of a recent nationally-representative serosurvey was infected in the year before the before the survey. We treat the binary outcome of this model like a diagnostic test, which when summarized at the population-level can be adjusted for sensitivity and specificity. As detailed in the paper, the serosurvey was a two-stage cluster survey, with 70 communities selected with probability proportional to each community's population and at least 10 households (with at least 40 total samples) sampled from each community.

We make three estimates of the seroincidence rate, all of which rely on a Bayesian hierarchical model that is specified below. For the 'survey estimate', we assume that the nationwide estimate of seroincidence is equivalent to the in-sample estimate of seroincidence, since the serosurvey sample was nationally representative. For the 'overall estimate', we extrapolate the survey estimate to both the unsampled populations within the sampled communities as well as to the 97,092 unsampled communities throughout Bangladesh. For the 'spatial estimate', we extend the survey estimate to the rest of the country using a logistic regression model including covariates and a Matern spatial covariance function.

S1.1. Bayesian hierarchical model

We model the number of predicted seropositive people in each serosurvey household, z_h, with a Bayesian hierarchical model similar to that of Makela, Si, and Gelman,[1] augmented to account for the sensitivity and specificity of the random forest model and with a binomial outcome in place of a Bernoulli outcome:

$$z_h \sim \text{Binomial}(n_h, \pi_h\theta_{1|1} + (1-\pi_h)(1-\theta_{0|0})) \tag{1}$$
$$\pi_h = \text{logit}^{-1}(\alpha_{c[h]}) \tag{2}$$
$$\alpha_c \sim \text{Normal}(\alpha_0 + \gamma\log(N_c), \sigma^2) \tag{3}$$
$$\alpha_0, \gamma \sim \text{Normal}(0,1) \tag{4}$$
$$\sigma \sim \text{Normal}^{+}(0,1). \tag{5}$$

The number of predicted seropositive people in a household, z_h, is determined by the number of members sampled in the household n_h and the probability of a household member testing positive. We separate the probability of testing positive into two parts: the true positive rate, calculated as the household seroincidence rate π_h multiplied by the sensitivity $\theta_{1|1}$; and the false positive rate, calculated as the seronegative rate

$(1-\pi_h)$ multiplied by one minus the specificity $(1-\theta_{0|0})$. Each household in community c is assumed to have the same underlying community-level rate α_c. The community-level rate is logit normal with its mean determined by the sum of a country-level intercept, α_0, and linear term to account for the (log) population of each community (N_c with coefficient γ) and a variance σ^2. Since the probability that a community was sampled was proportional to its population, we need to account for any relationship between community population and seroincidence rate to control for confounding. As in Makela, Si, and Gelman, we use a standard normal as a prior on α_0 and γ and a standard normal truncated to be positive as a prior for σ.

Upon estimating π_h, we can make predictions for the number of seroincident individuals in each sampled household, $\hat{y}_h$:

$$\hat{y}_h = \text{Binomial}(n_h, \pi_h) \tag{6}$$

$$\hat{y}_{survey} = \sum_{h=1}^{H} \hat{y}_h \tag{7}$$

$$\pi_{survey,c} = \frac{\sum_{h \in c} \hat{y}_h}{\sum_{h \in c} n_h} \tag{8}$$

$$\pi_{survey} = \frac{\hat{y}_{survey}}{\sum_{h=1}^{H} n_h}. \tag{9}$$

The number of seroincident individuals within a surveyed household is a binomial draw from that household with probability π_h. These draws can be averaged across each community to calculate the community-specific survey estimates $\pi_{survey,c}$ or across all households to calculate the nationwide survey estimate π_{survey}.

We can also make predictions for the number of seroincident individuals in unsampled households within sampled communities $\hat{y}_{unobs}$ and the number of seroincident individuals in unsampled communities $\hat{y}_{unsamp}$ to calculate the overall seroincidence rate $\pi_{overall}$:

$$\hat{y}_{unobs} = \sum_{c=1}^{C_{obs}} \text{Binomial}\left(N_c - \sum_{h \in c} n_h, \text{logit}^{-1}(\alpha_c)\right) \tag{10}$$

$$\hat{y}_{unsamp} = \sum_{c=C_{obs}+1}^{C} \text{Binomial}\left(N_c, \text{logit}^{-1}(\alpha_{unsamp,c})\right) \tag{11}$$

$$\alpha_{unsamp,c} = \text{Normal}\left(\alpha_0 + \gamma \log(N_c), \sigma^2\right) \tag{12}$$

$$\pi_{overall} = \frac{(\hat{y}_{survey} + \hat{y}_{unobs} + \hat{y}_{unsamp})}{\sum_{c=1}^{C} N_c}. \tag{13}$$

We separate sampled and unsampled communities by using the first C_{obs} indices of c to represent sampled communities, while indices $C_{obs}+1, ..., C$ represent unsampled communities. For unobserved households within sampled communities, we assume that the seroincidence rate is the same as that amongst the sampled households, $\text{logit}^{-1}(\alpha_c)$. For unobserved communities, we draw a new seroincidence rate $\text{logit}^{-1}(\alpha_{unsamp,c})$, where $\alpha_{unsamp,c}$ is a draw from a normal distribution with a mean determined by a function of the country-level intercept α_0 and the population of that community N_c and a variance of σ^2. The sum of these estimates divided by the entire population is the overall estimate of the nationwide seroincidence, $\pi_{overall}$.

To fit the Bayesian hierarchical model and make both the survey and overall seroincidence estimates we used a model built using the Stan probabilistic programming language.[2,3]

S1.2. Integrated nested Laplace approximations and the spatial estimate

Our primary estimate of the country-wide seroincidence used in the main analyses is however the spatial estimate. To produce this estimate, we extend the community-specific survey estimates to the entire country using a logistic regression model with a Matern spatial covariance function and covariates with integrated nested Laplace approximations (INLA),[4] as described in the main text. Specificially, we fit an INLA model to each of 1,000 posterior draws of community seroincidence from the Bayesian hierarchical model (described above) and then predict the seroincidence for all 5km by 5km grid-cells across the country. In the end, we generate 1,000 maps of cholera seroincidence rates and we take a population-weighted average to produce

the nationwide spatial estimates. Our primary results are the median spatial estimate and the 95% credible interval.

S1.3. Random forest predictions

Azman *et al.* fitted a random forest model using age, sex, vibriocidal titers (Ogawa and Inaba), anti-LPS IgG and IgA antibodies, and anti-CTB IgG and IgA antibodies, and blood group to classify the seropositive status of individuals from a longitudinal cohort study in Bangladesh.[5] Since no blood group information was collected in the serosurvey, we fit a new random forest model to the cohort data using all of the remaining covariates. For each observation in the cohort study, we use the proportion of trees that predict that the observation is seropositive as the probability of seropositivity. From these probabilities, we calculate the receiver operating characteristic (ROC) curve for the cohort predictions and calculate the cutoff that maximizes the Youden's J statistic, *i.e.* the sum of the sensitivity and the specificity.[6]

We use the random forest model to predict the seropositivity status of each participant in the serosurvey; the participants whose probability of seropositivity exceed the Youden cutoff are classified as seropositive. The serosurvey predictions are aggregated to the household unit, providing us with z_h in Equation 1.

S1.4. Specificity and sensitivity of the random forest predictions

As with all imperfect tests, population-level (*e.g.*, aggregated) random forest model seroincidence estimates can be corrected for the test's specificity and sensitivity, when known. To estimate the specificity and sensitivity of this random forest model we conducted leave-one-individual-out cross validation (LOOCV) on the original cohort data used in Azman *et al.*, where the seropositive status of the participants was known. For each individual in the cohort, we fit a random forest model to the rest of the cohort, calculate the Youden cutoff, and predict the seropositivity of the left-out individual, which we call LOOCV predictions and denote z^{ℓ}.

To estimate the specificity, $\theta_{0|0}$ in Equation 1, we include the LOOCV predictions in our Bayesian hierarchical model:

$$z_i^{\ell} \mid y_i = 0 \sim Bernoulli(1 - \theta_{0|0}). \tag{14}$$

The LOOCV predictions of the seronegative observations from the cohort study (*i.e.* $y_i = 0$) are Bernoulli random variables with the probability of seropositivity equal to $(1 - \theta_{0|0})$.

The sensitivity of the random forest predictions, $\theta_{1|1}$ in Equation 1, varies across days since infection due to the decay in antibody response over time. After infection with *V. cholerae* O1, most antibodies rise including vibriocidals, one of the most informative markers, which peak around 7-10 days post-infection. As time since infection increases, the antibody profile of an individual, in general, returns to pre-infection levels. The vibriocidal titers decay quickly in the first three months before decaying more slowly over the following three years. This decay is illustrated by the decline in raw sensitivity of the random forest predictions over time (Table S1).

Table S1: The sensitivity of the random forest predictions of seropositivity over days since infection.

Days since infection	Observations	Sensitivity
7-10	311	96.8%
24-41	293	97.3%
76-109	164	72.0%
154-199	137	46.7%
261-274	42	38.1%
353-363	37	32.4%

To account for this decay, we estimate the sensitivity as a time-varying quantity rather than as a static quantity and rewrite the overall sensitivity as a joint probability:

$$\underbrace{\theta_{1|1}}_{\substack{\text{overall}\\ \text{sensitivity}}} = \underbrace{\mathbb{P}(Z = 1 \mid Y = 1, T)}_{\text{time-varying sensitivity}} \underbrace{\mathbb{P}(T \mid Y = 1)}_{\substack{\text{daily probability}\\ \text{of infection}}}, \tag{15}$$

where Z is the result of the test (*i.e.* the random forest model), Y is the true seropositive status of the individual, and T is the time since infection in days. We need to estimate the time-varying sensitivity ($\mathbb{P}(Z = 1 \mid Y = 1, T)$) and the probability of being infected $T = t$ days ago ($\mathbb{P}(T \mid Y = 1)$). Since sensitivity only concerns seropositive individuals (*i.e.* $Y = 1$) and seropositivity is by our definition infection over the past 365 days, T is restricted to be less than or equal to 365 days for all components of $\theta_{1|1}$.

S1.4.1. Time-varying sensitivity

We estimate the time-varying sensitivity of the random forest predictions in the cohort study using a logistic regression model with a cubic polynomial for the log of days since infection, similar to the method used by Leisenring *et al.*:[7]

$$\text{logit}(Z^{ell} \mid Y = 1, T = t) = \beta_0 + \beta_1 \log(t) + \beta_2 \log(t)^2 + \beta_3 \log(t)^3. \tag{16}$$

We assume that the sensitivity of the test depends only on the time since infection, T. The posterior median and 95% credible interval for the sensitivity at each time since infection (from 7 to 365), $\mathbb{P}(Z = 1 \mid Y = 1, T = 7, \ldots, 365)$, is shown Figure SS1.

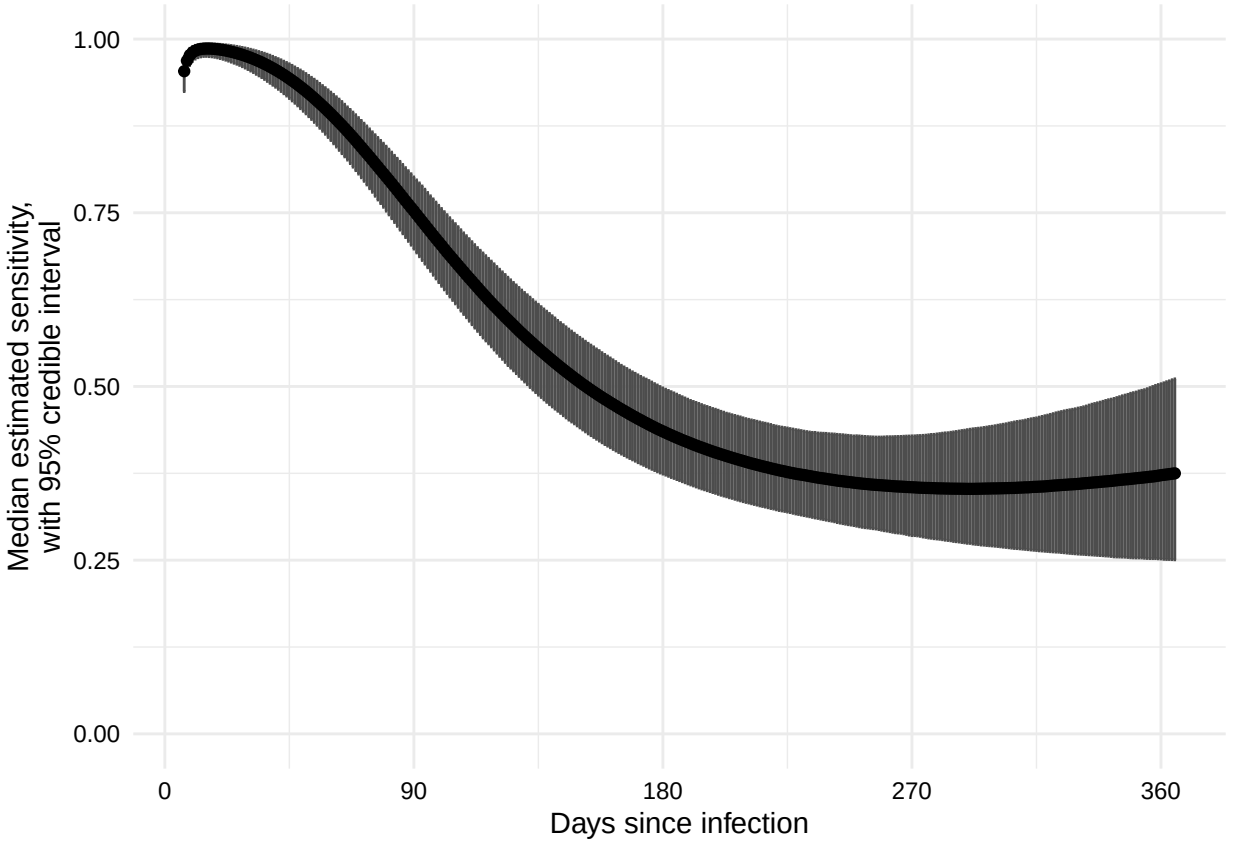


Figure S1: The estimated sensitivity of the random forest model for identifying whether an individual was infected in the last year by the number of days since true infection. The points represent the median estimate from a generalized logistic regression model for sensitivity with cubic polynomial terms for the log of days since infection. The gray error bars represent the 95% credible intervals.

S1.4.2. Daily probability of infection

The estimates of time-varying sensitivity allow us to calculate the overall sensitivity given the time since infection, however we do not know this time for any individual in the serosurvey. We assume that individuals only get infected once in the past year, such that the daily probabilities sum to one across $t = 1, \ldots, 365$. We assume that the risk of infection for each individual was uniformly distributed over the year before sample collection:

$$\mathbb{P}(T = 1 \mid Y = 1) = \mathbb{P}(T = 2 \mid Y = 1) = \cdots = \mathbb{P}(T = 365 \mid Y = 1) = \frac{1}{365}.$$

We set the expected value of any given time since infection to be equal to $\frac{1}{365}$, however the Bayesian hierarchical model allows for variability around each estimate.

Past work on clinical cholera has shown that there is seasonal variation of cholera in Bangladesh, which varies regionally across the country.[8] While we do not expect that incorporating seasonality will greatly affect our overall estimates of seroincidence, if more detailed estimates of seasonality were available across the country, they could be used to refine estimates of these or future analyses.

S1.5. Other estimators and time frames

We fit several alternative models and observe the differences in their seroincidence estimates. We use an 'unadjusted' model, where our random forest estimates are not adjusted by sensitivity and specificity; *i.e.* $\pi_h\theta_{1|1} + (1 - \pi_h)(1 - \theta_{0|0})$ in Equation 1 is replaced by p_h. Previous work showed that a vibriocidal titer (either Inaba and Ogawa) of at least 320 was the best threshold for maximimizing sensitivity and specificity for identifying individuals infected in the previous year; thus we fit a 'vibriocidal' model which used these predictions in place of the random forest predictions for z_h in Equation 1. To see how the seroincidence changed over multiple time frames, we also fit the random forest and vibriocidal models to 100 and 200 days since infection. We use the equivalent of the overall estimate to conduct these additional analyses.

S1.6. Risk factors for seroposivity

We used a series of logistic regression models with a Matern spatial covariance function to explore the association between seropositivity (random forest positive for individuals) and various individual-, household- and community-level covariates. We explored both univariate relationship and multivariate (linear) relationships between the covariates and the binary seropositivity outcome using models with and without different random effects and spatial correlation. The 'full' model, used for the primary analyses in manuscript, included a Matern spatial random field and random effects for both households and communities (assumed to be independent and identically distributed with log-gamma priors). We also estimated the relationship between the covariates and seropositivity with a model including no random effects for household or community and only spatial correlation, and another model including only random effects for household and community without spatial correlation.

S1.7. Results

The three methods produced similar estimates for the nationwide seroincidence rate for the 365-day period preceding the serosurvey (Figure SS2). The median spatial estimate from the INLA 5km by 5km grid-cell maps was 18.7% (95% CI: 8.7-26.8%) This corresponds to a median of 30.4 million (95% CI: 14.1-43.6 million) individuals infected during that time period. The median overall estimate from the Bayesian hierarchical model, $\pi_{overall}$ in Equation 13, was 20.7% (95% CI: 14.1-28.4%). The median in-sample estimate from the same model, π_{survey} in Equation 9, was 20.4% (95% CI: 15.1-26.4%).

Coincidentally, the unadjusted 365-day random forest model estimates are similar to those of the estimates adjusted for sensitivity and specificity, albeit with a narrower credible interval (median: 19.9%, 95% CI: 17.8-22.5%). By comparison, the vibriocidal model yields lower seroincidence rate estimates (median: 12.8%, 95% CI: 6.8-20.1%) than the models based on random forest estimates despite the fact that the proportion of the serosurvey that had vibriocidal titers greater than or equal to 320 is similar to the proportion that was classified as seropositive by the random forest model (19.5% vs. 19.9%). This is due to the vibriocidal estimates having a lower specificity than the random forest models (Figure SS3).

S1.8. Mapping

From the grid-cell estimates, we can make a series of maps to observe the geographic variability of cholera throughout Bangladesh. Maps of the median seroincidence rate and estimated number of annual infections by grid cell are in the main manuscript (Figure 2). To help identify high-risk regions and our confidence in the estimates, we calculated proportion of posterior grid-cell seroincidence estimates with a relative risk

greater than two (Figure SS4). As in the manuscript, we see higher risk in the Bay of Bengal and in pockets in the northwest and north, though less so in the northeast.

Figure SS5 investigates the variance of our estimates As variance scales with the size of the estimate, it is difficult to interpret. The coefficient of variation, the standard deviation is divided by the mean, is another measure often used but it can become very large for places where mean estimates are very small. Instead, we use the width of the logged relative risk credible interval in this map. To do this, we first bound all posterior samples of the relative risk to be between 0.25 and 4 (or -2 and 2 on the log2 scale), which represent reasonable cutoffs for very high and very low risk as only 2.0% of samples across all grid cells are greater than 4 and 16.0% are less than 0.25. Next, we take the log2 difference between the upper and lower bounds of the 95% credible interval for each grid cell. Grid cells with a log2 difference of 4 have an upper bound relative risk that is greater than 4 and a lower bound relative risk less than 0.25, indicating that we are very uncertain of the true risk in that grid cell; 19.8% of the grid cells in our map have a log2 difference of 4 and are displayed in white. Grid cells with a log2 difference of 0 have both upper and lower bounds either above 4 or below 0.25 and would be indicated on the map with maximum opacity if there were any. The colors on the map indicate the posterior median for that grid cell and the opacity indicates the log2 difference between the upper and lower bounds. This map demonstrates how the certainty in our estimates fades with distance from the sampled communities.

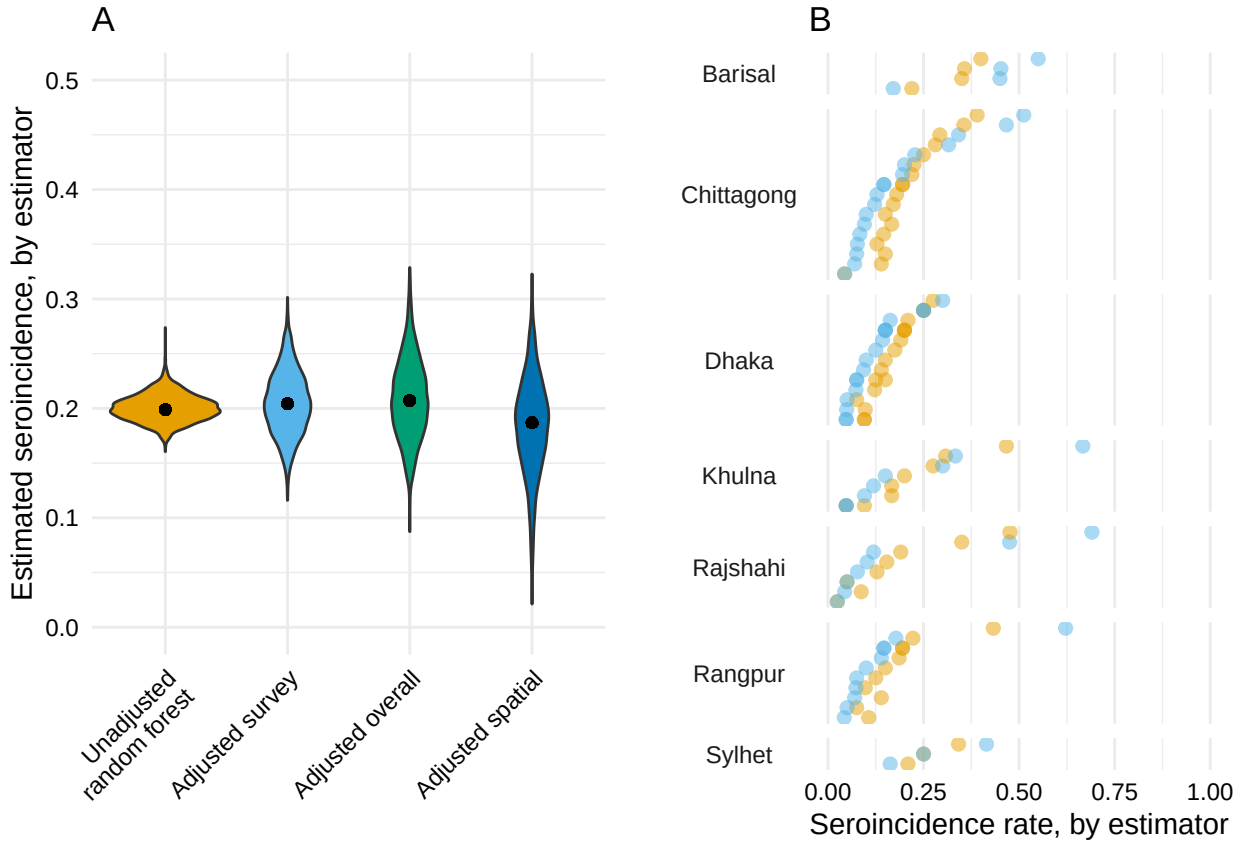
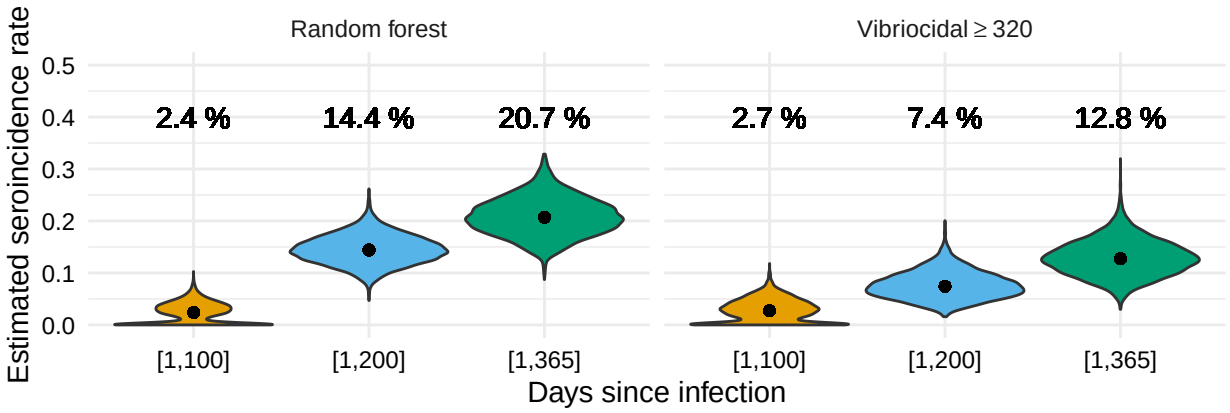


Figure S2: The estimated seroincidence rate distributions by estimator for the whole population (A) and by community (B). The unadjusted random forest estimates (orange) do not account for sensitivity or specificity. The adjusted survey estimates (light blue) are the in-sample estimates from the Bayesian hierarchical model which accounts for the sensitivity and specificity of the random forest estimates. The adjusted overall estimates (green) are from the same model but including predictions for unsampled communities. The adjusted spatial estimates (dark blue) extend the survey estimates to the rest of the country using a logistic regression with a Matern spatial covariance function. Only the unadjusted random forest and adjusted survey estimators make community-specific estimates in (B).



	RF [1,100]	RF [1,200]	RF [1,365]	Vib320 [1,100]	Vib320 [1,200]	Vib320 [1,365]
sensitivity	66.9%	63.8%	54.3%	72.6%	61.3%	53.3%
specificity	93.6%	89.7%	88.9%	80.6%	83.1%	85.0%

Figure S3: The estimated seroincidence rate across varying infection window sizes and estimators. We estimate the seroincidence rate with two different estimators across three window sizes (100, 200, and 365 days). The random forest uses age, sex, and measurements of six antibodies (vibriocidal Inaba, vibriocidal Ogawa, anti-CTB IgG, anti-CTB IgA, anti-LPS IgG and anti-LPS IgA) to classify individuals as seroincident. As a comparison, we use the historical convention where those with either vibriocidal titers greater than or equal to 320 is classified as seroincident. The vibriocidal titer method has lower specificity, which yields lower estimates of seropositivity than those from the random forest model. The estimate of the median seroincidence rate with each estimator is displayed above its distribution. The estimates of the adjusted sensitivity and specificity for each estimator and window size are presented in the table below the figure.

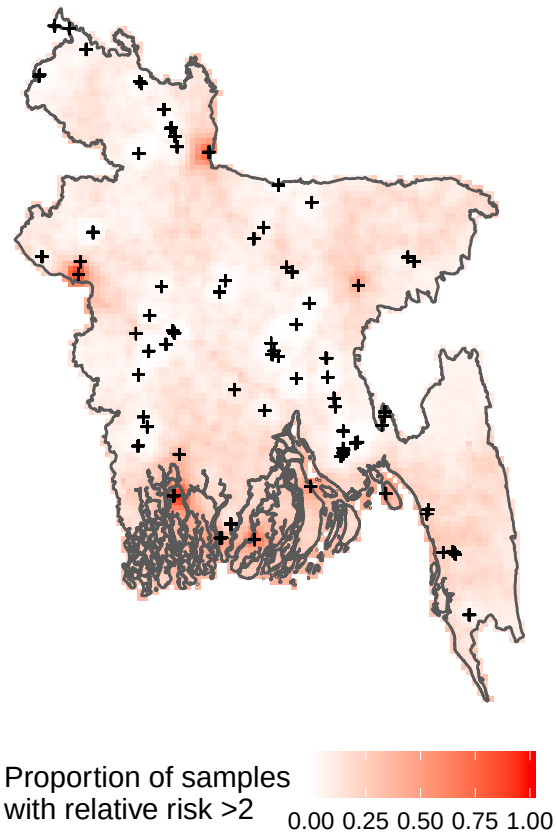


Figure S4: Proportion of posterior samples with relative risk greater than 2 for each 5km x 5km grid cell.

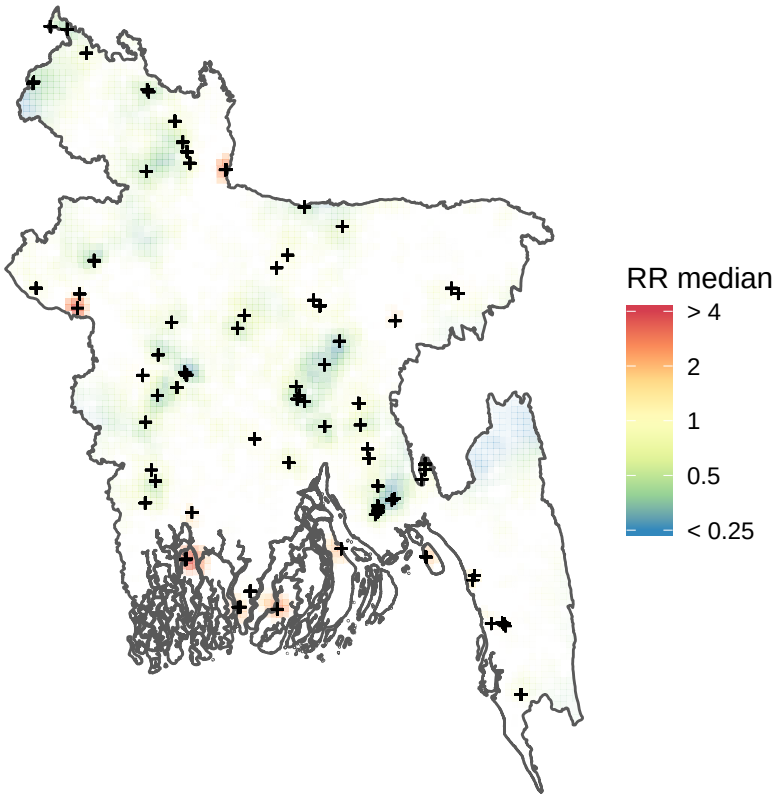


Figure S5: The median relative risk estimate for each 5km x 5km grid cell with the opacity determined by the width of the 95% credible interval. White grid cells have 95% credible intervals where the lower bound is less than 0.25 and the upper bound is greater than 4. Places with narrower credible intervals have greater opacity. The most opaque cells are those where both the upper and lower bound are either less than 0.25 or above 4.

S1.9. Leave-one-out cross validation

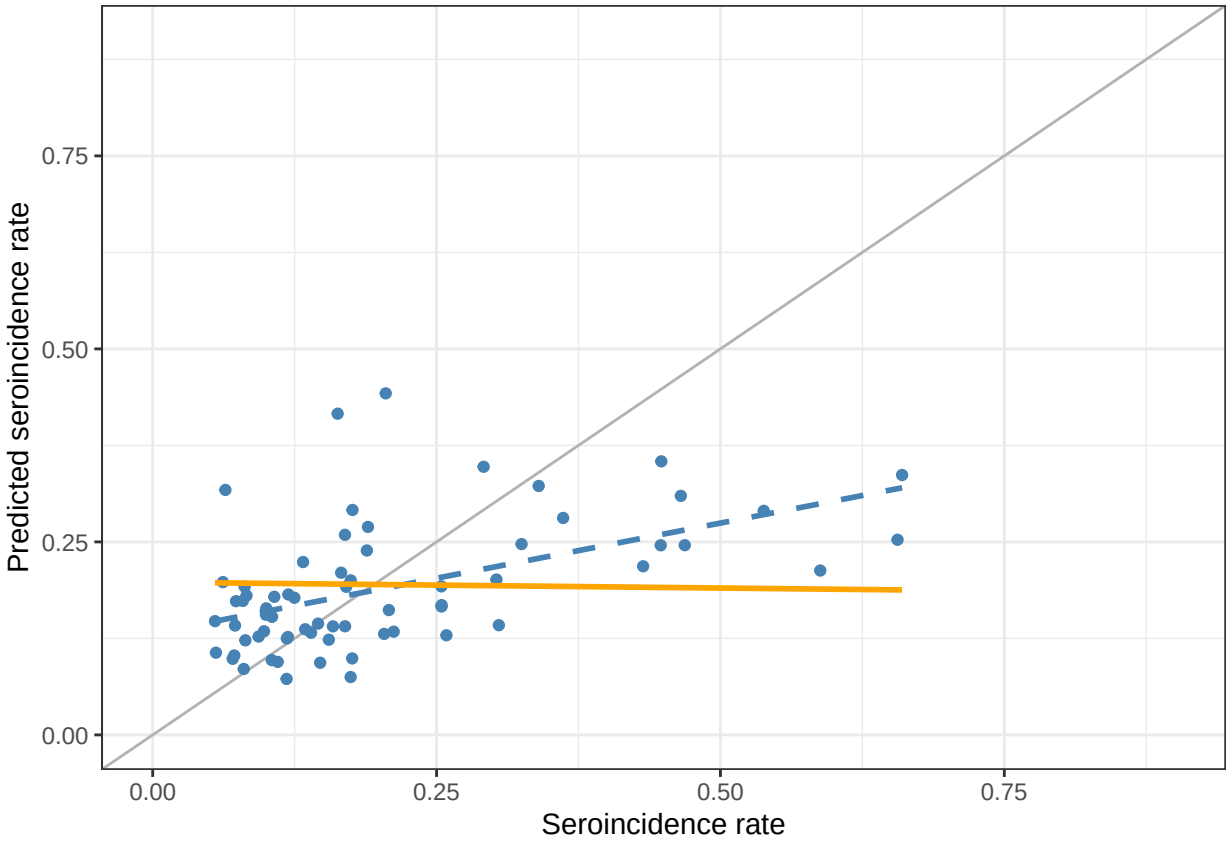


Figure S6: The results from leave-one-out cross validation. Results from cross-validation where each grid cell was held out of the INLA model, one at a time, and the posterior predictive mean for that location was estimated (y-axis). The blue dashed line illustrates the best fitting linear model prediction. The solid orange line is the best fitting linear model prediction from a naive model that predicts the average of the mean of the other sampled grid cells.

S1.10. Risk Factors for seroposivity

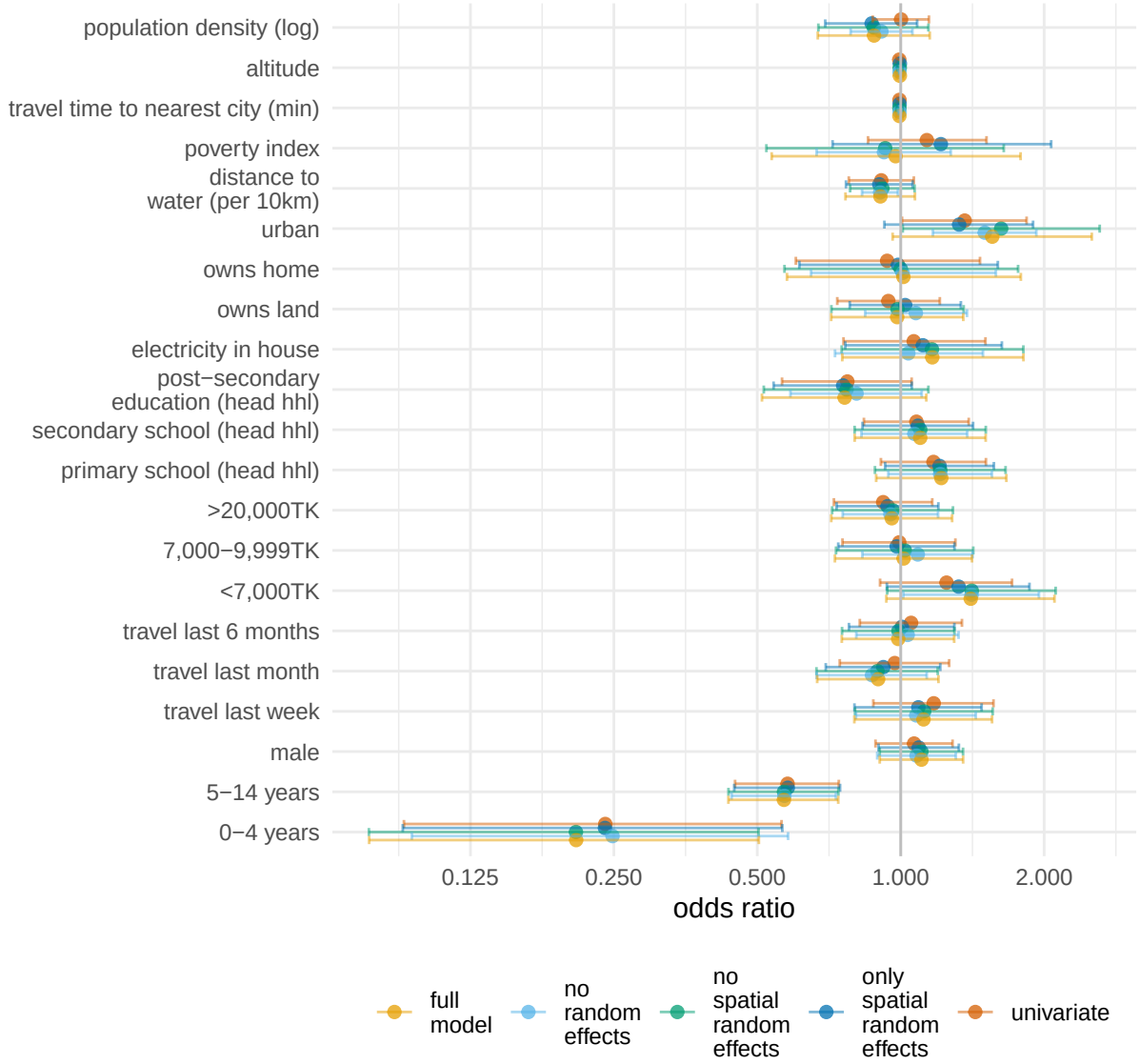


Figure S7: Estimates of odds ratios for seropositivity from different models.

S2. Additional descriptive analyses

In this section we present additional descriptive analyses to illustrate the distributions of each of the antibody levels in different ways and characteristics of the cohort.

Figure S8: Population pyramid of survey participants. Dots illustrate the expected proportion of each age-sex category according to the 2012 Bangladesh census.

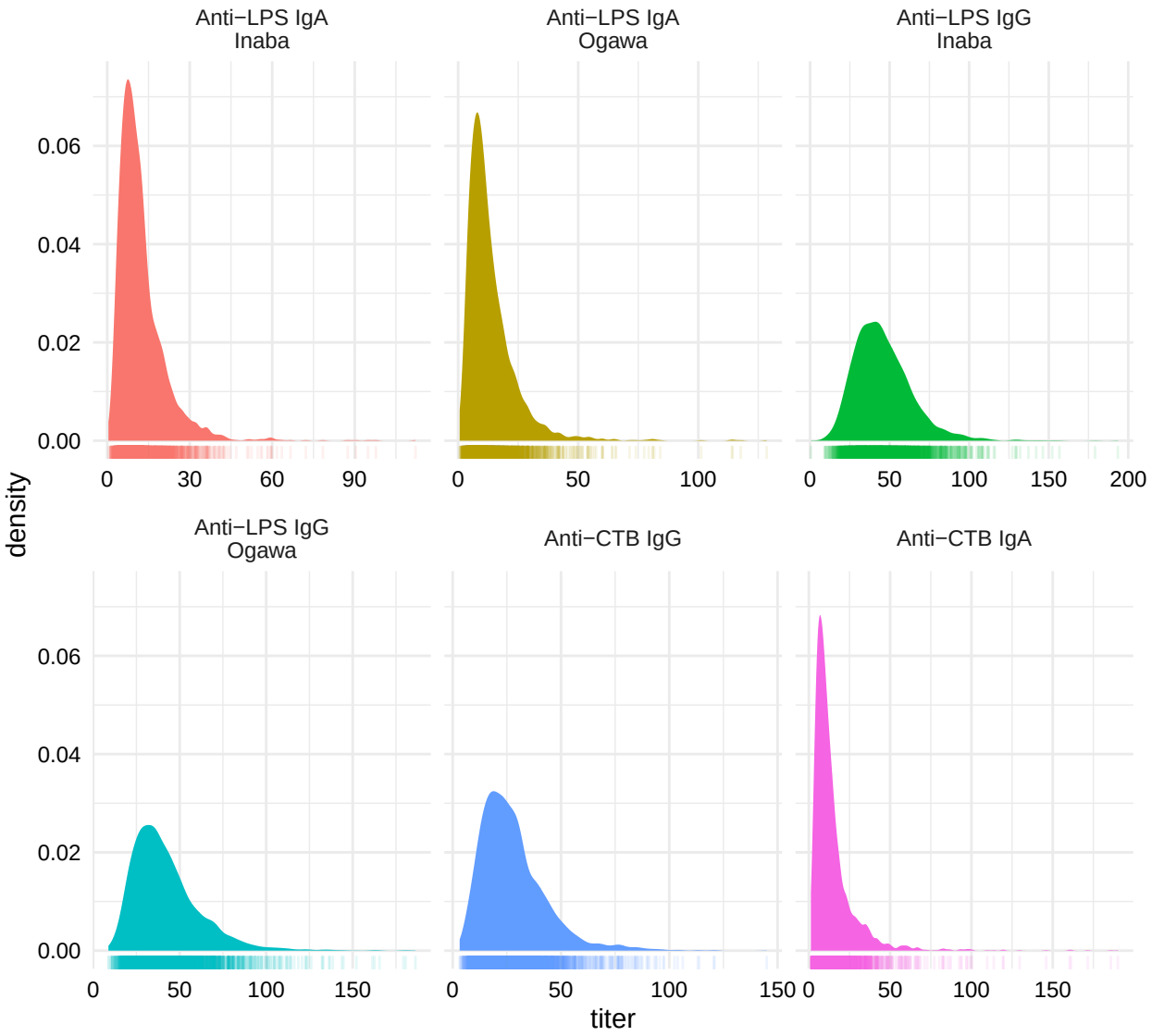


Figure S9: Distributions (smoothed) of antibodies measured by ELISA. Smoothed densities estimated using ggplot with default parameters (geom_density) with locations of data points shown in the rug plot below.

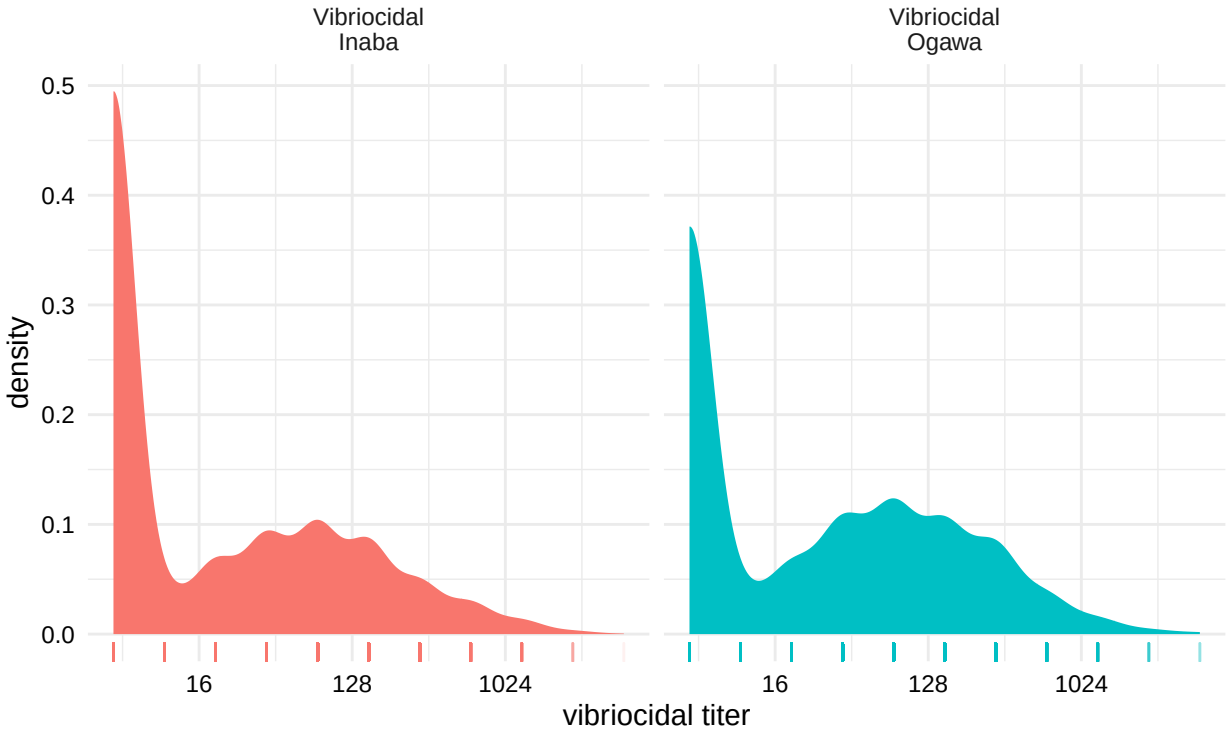


Figure S10: Distributions (smoothed) of vibriocidal antibodies. Smoothed densities estimated using ggplot with default parameters (geom_density) with locations of data points shown in the rug plot below.

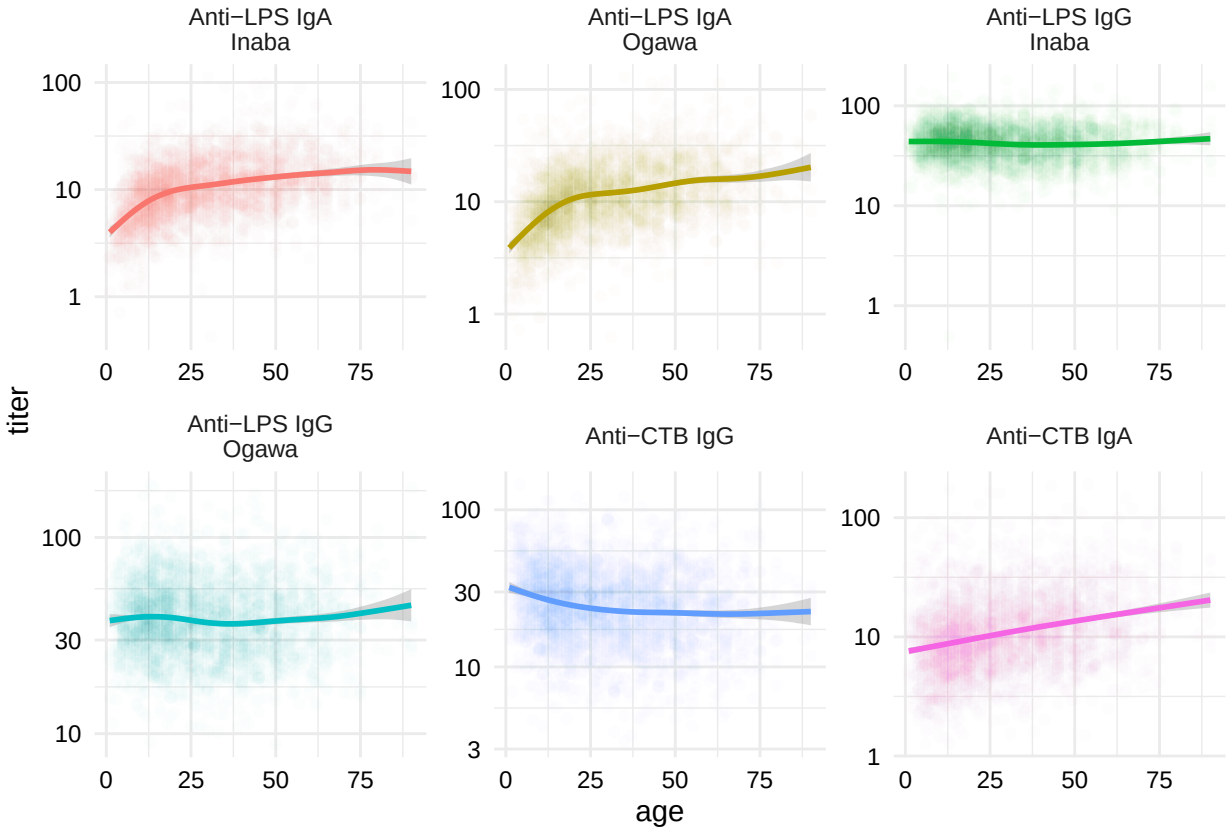


Figure S11: Distributions of ELISA antibody titers by age. Dots represent individual datapoints and lines represent the fit of a generalized additive model using a cubic spline.

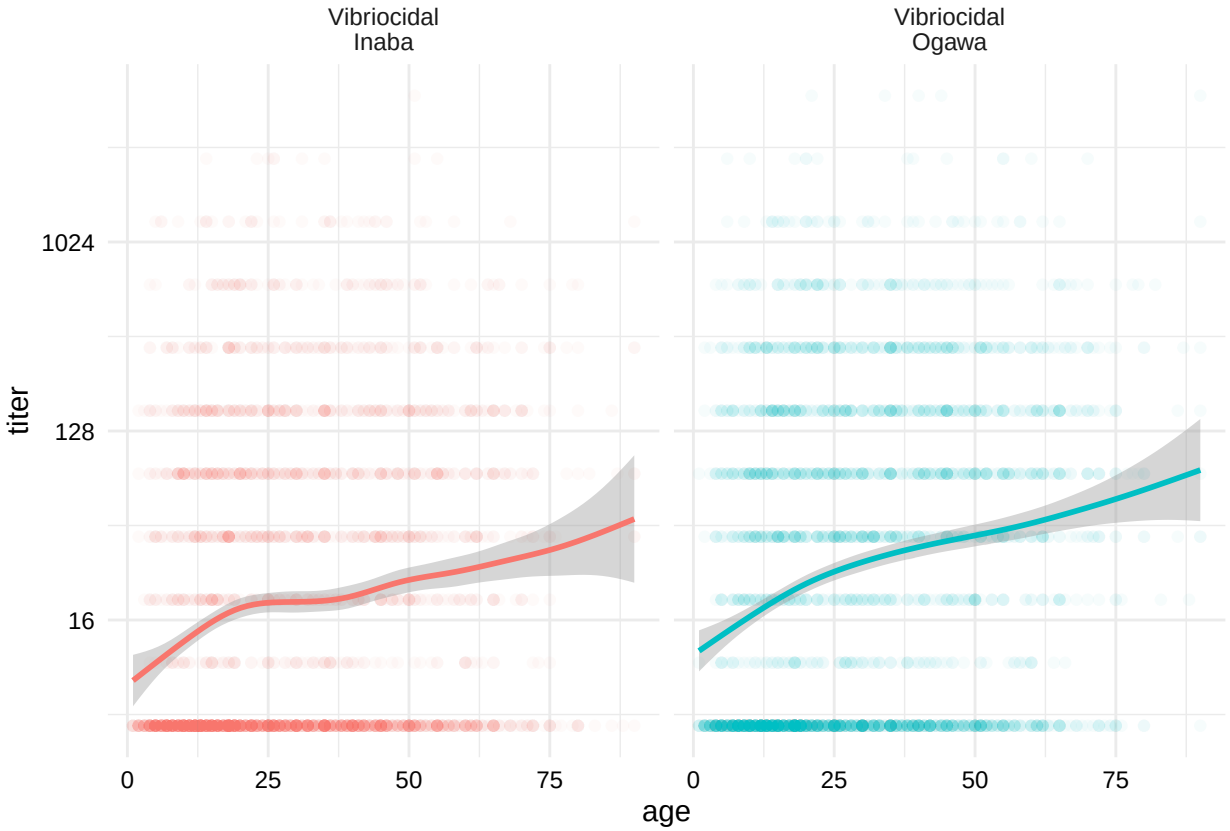


Figure S12: Distributions of vibriocidal antibody titers by age. Dots represent individual datapoints and lines represent the fit of a generalized additive model using a cubic spline.

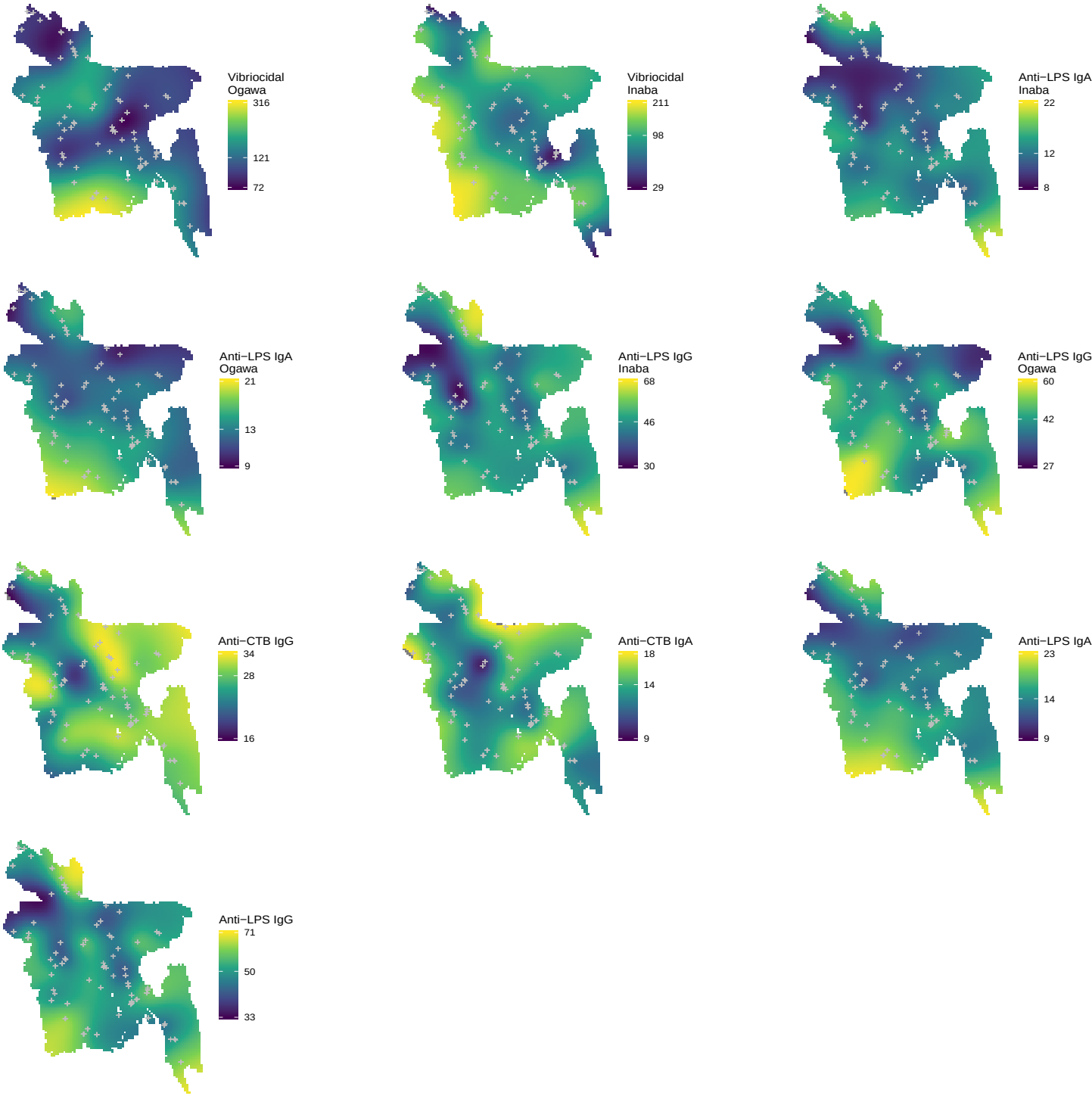


Figure S13: Smoothed maps of the antibody levels for each biomarker based on generalized additive models (GAMs) including a thinplate spline for geographic coordinates and age. Predictions are made for individuals of age 25 to reflect that of adults.

References

[1] Makela S, Si Y, Gelman A. Bayesian inference under cluster sampling with probability proportional to size. Statistics in Medicine 2018;37:3849–68. doi:10.1002/sim.7892.

[2] Carpenter B, Gelman A, Hoffman MD, Lee D, Goodrich B, Betancourt M, et al. Stan: A Probabilistic Programming Language. Journal of Statistical Software 2017;76:1–32. doi:10.18637/jss.v076.i01.

[3] Stan Development Team. RStan: The R interface to Stan 2018.

[4] Rue H, Martino S, Chopin N. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 2009;71:319–92. doi:10.1111/j.1467-9868.2008.00700.x.

[5] Azman AS, Lessler J, Luquero FJ, Bhuiyan TR, Khan AI, Chowdhury F, et al. Estimating cholera incidence with cross-sectional serology. Science Translational Medicine 2019;11:eaau6242. doi:10.1126/scitranslmed.aau6242.

[6] Youden WJ. Index for rating diagnostic tests. Cancer 1950;3:32–5. doi:10.1002/1097-0142(1950)3:1<32::AID-CNCR2820030106>3.0.CO;2-3.

[7] Leisenring W, Pepe MS, Longton G. A Marginal Regression Modelling Framework for Evaluating Medical Diagnostic Tests. Statistics in Medicine 1997;16:1263–81. doi:10.1002/(SICI)1097-0258(19970615)16:11<1263::AID-SIM550>3.0.CO;2-M.

[8] Das SK, Begum D, Ahmed S, Ferdous F, Farzana FD, Chisti MJ, et al. Geographical diversity in seasonality of major diarrhoeal pathogens in Bangladesh observed between 2010 and 2012. Epidemiology & Infection 2014;142:2530–41. doi:10.1017/S095026881400017X.