Supplementary Material

Inferring the proportion of undetected cholera infections from serological and clinical surveillance in an immunologically naive population

Flavio Finger, Joseph Lemaitre, Stanley Juin, Brendan Jackson, Sebastian Funk, Justin Lessler, Eric Mintz, Patrick Dely, Jacques Boncy, Andrew S Azman

Previous estimates of infection to clinical case ratio

Table S1: previous published estimates of the cholera infection to clinical case ratio

Location and year	Study design & population	Method to determine infection	Biotype	Estimate	Reference
East Pakistan, 1966–1967	Longitudinal study of all available village residents in an endemic area	Serology	El Tor	0 out of 27 infected required treatment 22 out of 27 infections inapparent	(Mccormack, Islam, and Fahimuddin 1969)
East Pakistan, 1968-1969	Case contact follow-up	Bacterial culture & serology	Classical	Infection to case ratio 4:1	(Bart et al. 1970)
			El Tor	Infection to case ratio 36:1	
(unspecified)	Surveys	Bacterial culture	Classical	59% inapparent infections and 15% very mild symptoms but detected only in bacteriological surveys	(Gangarosa 1974)
			El Tor	75% inapparent infections and 18% very mild symptoms but detected only in bacteriological surveys	
Louisiana, US, 1978	Case contact follow-up	Bacterial culture	El Tor	3 out of 6 infections inapparent	(Blake et al. 1980)

Location and year	Study design & population	Method to determine infection	Biotype	Estimate	Reference
Review of published sources	Case contact follow-up	Bacterial culture	Classical	50% of infected were asymptomatic (mean over 5 studies)[1]	(Feachem 1982)
			El Tor	70% of infected were asymptomatic (mean over 4 studies)[1]	
Gulf Coast Oil Rig, US, 1983	Outbreak investigation	Serology	El Tor	1/16 infections inapparent	(Johnston et al. 1983)
Truk, Micronesia, 1982	Cross-sectional study	Serology	El Tor	68% of infections inapparent	(Harris et al. 1986)
Review of published sources	-	-	(not specified)	1/2 to 1/100 infections leads to symptomatic cholera in endemic areas	(Glass and Black 1992)
Bay of Bengal, 1891–1940	Model of historical incidence of clinical cholera	-	Classical	high proportion of inapparent infections	(King et al. 2008)
Sitakunda, Bangladesh, 2022	Longitudinal study	Serology	El Tor	1 symptomatic case per 600 infections 1 medically-attended case per 2,340 infections	(Hegde et al. 2023)

[1] The studies included in this aggregated estimate are not mentioned separately in this table.

Predictors of self-reported cholera diagnosis

We used multivariable logistic regression in a stepwise procedure to identify predictors of self-reported watery diarrhea or cholera diagnosis as model outcomes. First, univariate models with sex, age (in years), location (13 levels, one for each village included in the survey) and vibriocidal titer (log2-transformed) as the independent variable were fit for each of the two dependent variables. Only significant predictors in the univariate models (p-value <0.2) were included in the multivariable analysis. We didn't adjust for household clustering, but verified that this didn't have a significant impact on the standard deviations of the results of the regression.

Age and log2-transformed vibriocidal titer as well as some of the location levels were significant predictors for both outcomes in univariate regression. We thus compared multivariable models with and without location, and including age and vibriocidal titer, using the Akaike Information Criterion. The best model (for each of the two outcomes) included location as fixed effect (13 levels). Results from these final models are presented in Tables S2a and S2b Note that this analysis requires individual level data, which isn't included in the authors data repository to protect privacy of the survey participants. Interested researchers are invited to contact the authors.

A higher log2-transformed vibriocidal titer was associated with self-reported watery diarrhea (Odds Ratio (OR) 1.10 (95% CI 1.07 to 1.14)) when adjusting for age and location. In addition, log2-transformed vibriocidal titer was associated with self-reported cholera diagnosis by a health care worker was (OR 1.12 (95% CI 1.08 to 1.16)) when adjusting for age and location.

Table S2a: Logistic regression analysis of the association of vibriocidal titer with self-reported watery diarrhea

		Univariate			Multivariate		
	N	**OR[1]**	**95% CI[1]**	**p-value**	**OR[1]**	**95% CI[1]**	**p-value**
sex	2,539						
f		—	—				
m		0.97	0.80, 1.18	0.8			
age	2,536	1.01	1.00, 1.01	<0.001	1.01	1.01, 1.02	<0.001
location	2,539						
Bac Da		—	—		—	—	
Chevre		0.52	0.32, 0.86	0.011	0.52	0.31, 0.88	0.014
Dauphi		0.91	0.57, 1.46	0.7	0.86	0.52, 1.40	0.5
Drouin		1.19	0.79, 1.81	0.4	1.17	0.77, 1.82	0.5
Grand		0.80	0.47, 1.35	0.4	0.77	0.45, 1.32	0.3
Laport		2.01	1.17, 3.47	0.012	2.50	1.42, 4.42	0.002
Latapi		1.02	0.61, 1.69	>0.9	1.02	0.61, 1.72	>0.9
Lemeau		0.83	0.35, 1.83	0.7	0.79	0.32, 1.80	0.6
Petit		0.64	0.32, 1.23	0.2	0.67	0.33, 1.30	0.2
Ponten		0.57	0.36, 0.89	0.014	0.58	0.37, 0.93	0.023
Robine		0.77	0.39, 1.49	0.5	0.77	0.38, 1.50	0.5
Rossig		0.91	0.58, 1.42	0.7	0.85	0.54, 1.35	0.5
Theard		0.53	0.29, 0.96	0.041	0.48	0.25, 0.88	0.021
vibriocidal titer (log2)	2,447	1.09	1.06, 1.12	<0.001	1.10	1.07, 1.14	<0.001

[1] OR = Odds Ratio, CI = Confidence Interval

Table S2b: Logistic regression analysis of the association of vibriocidal titer with self-reported cholera

		Univariate			Multivariate		
	N	**OR[1]**	**95% CI[1]**	**p-value**	**OR[1]**	**95% CI[1]**	**p-value**
sex	2,540						
f		—	—				
m		0.96	0.78, 1.17	0.7			
age	2,537	1.01	1.00, 1.01	0.001	1.01	1.01, 1.02	<0.001
location	2,540						
Bac Da		—	—		—	—	
Chevre		0.54	0.32, 0.92	0.022	0.55	0.32, 0.94	0.028
Dauphi		0.72	0.44, 1.20	0.2	0.71	0.42, 1.21	0.2
Drouin		1.01	0.66, 1.58	>0.9	1.00	0.64, 1.58	>0.9
Grand		0.31	0.15, 0.60	<0.001	0.29	0.14, 0.57	<0.001
Laport		2.20	1.27, 3.85	0.005	2.79	1.57, 5.01	<0.001
Latapi		1.12	0.67, 1.89	0.7	1.14	0.67, 1.95	0.6
Lemeau		0.59	0.21, 1.41	0.3	0.65	0.23, 1.58	0.4
Petit		0.73	0.36, 1.42	0.4	0.77	0.37, 1.52	0.5
Ponten		0.58	0.36, 0.93	0.021	0.59	0.37, 0.97	0.034
Robine		0.89	0.44, 1.72	0.7	0.89	0.44, 1.74	0.7
Rossig		1.02	0.65, 1.62	>0.9	0.96	0.60, 1.54	0.9
Theard		0.61	0.33, 1.12	0.12	0.55	0.29, 1.02	0.064
vibriocidal titer (log2)	2,448	1.10	1.07, 1.14	<0.001	1.12	1.08, 1.16	<0.001

[1] OR = Odds Ratio, CI = Confidence Interval

Alternative methods to infer infection rate

Simpler methods to determine infection by classifying measured titer values according to fixed threshold or a mixture distribution exist. This chapter describes two methods and their application to our data, one relying on a fixed titer threshold to determine infection, and the other on Gaussian mixture models.

Fixed vibriocidal antibody titer thresholds to determine recent infection

We used a previously established vibriocidal antibody titer threshold of 320 to identify people infected with cholera in the previous 200 days (Azman et al. 2019). We accounted for the imperfect sensitivity and specificity of this threshold using the Rogan-Gladen estimator, considering estimates from two different populations, a study among previously uninfected adult North American volunteers challenged with *Vibrio cholerae* O1 and confirmed cholera cases in Bangladesh (Rogan and Gladen 1978). Uncertainty around sensitivity and specificity were propagated through the estimator.

From this approach we estimate that 42.0% (using North American cohort) to 31.2% (95%CI 14.7 to 46.4, Bangladesh cohort) of the population ≥2 years old was infected. While the estimates of the infection rate for 2-4 year olds were slightly lower than those from our main analyses, those for ≥5 were on average 1.6 times lower (Table S2, Figure S1).

Gaussian mixture model to determine recent infection

We also fit a mixture model composed of two normal distributions to the bi-modal distribution of measured $\log_2$-transformed vibriocidal antibody titers of participants. The two normal distributions were interpreted as the baseline antibody titer levels of uninfected and increased titer levels of recently infected participants respectively. We accounted for interval censoring of the titer values and estimated mean and variance of each distribution in addition to the mixing parameter, which we interpreted as the infection rate. Priors for the two normal distributions were Normal(4.5, 5) and Normal(8.5, 5). The prior for the mixture parameter was set to a Beta(1,1) distribution. The model was implemented in Stan using the RStan package (Stan Development Team 2020).

This approach produced similar central estimates than the threshold model for the population aged 5 and above (31.8%; 95% CrI 24.1 to 41.3) and a flat distribution with wide credible interval for 2 to 4 year olds (65.2%; 95% CrI 11.8 to 95.0) (Table S3, Figure S1).

Our supplementary analysis found that our approach to estimating the infection rate led to generally larger estimates than the alternative approaches including the use of a fixed threshold and a mixture model. While commonly used and computationally simpler, neither accounts for the waning of antibody levels in the months post infection. In the absence of well characterized post-infection kinetics or slowly waning antibodies, these approaches may perform well but care should be taken when applying these generally with fast waning antibodies, like vibriocidals (half life ~120 days, (Jones et al. 2022)).

Table S3: Clinical attack rate, attack rate of self reported cholera and watery diarrhoea, and estimates of infection rate with *V. cholerae* O1 according to our main and alternative analyses.

Method	Value [95% CI or CrI][a]		
	All ≥ 2 years	2-4 years	≥5 years
Clinical attack rate estimates			
Clinical attack rate (reported cholera incidence)	18.2%	39.5%	16.4%
Clinical attack rate (self-reported cholera diagnosis)	17.8%	18.5%	17.8%
Clinical attack rate (self-reported watery diarrhoea)	20.6%	25.0%	20.4%
Infection rate			
Infection rate computed using a vibriocidal decay model	52.6% [49.4 - 55.7]	35.5% [24.2 - 51.6]	53.1% [49.4 - 56.4]
Estimate of the infection rate from alternative analyses			
Infection rate computed using a fixed vibriocidal antibody threshold and correcting for sensitivity and specificity computed from a cohort from Bangladesh	31.2% [14.7 - 46.4]	21% [4.1 - 36.2]	31.7% [15.1 - 46.9]
Infection rate computed using a fixed vibriocidal antibody threshold and correcting for sensitivity and specificity computed from a cohort of healthy North-Americans	42.0%[b]	33.9%[b]	42.3%[b]
Infection rate computed using a gaussian mixture model	32.0% [24.5 - 41.3]	65.2% [11.8 - 95.0]	31.8% [24.1 - 41.3]

[a]Credible Intervals are given for Bayesian estimates

[b]Uncertainty cannot be quantified because no confidence intervals on sensitivity and specificity of thresholds are given in the source used

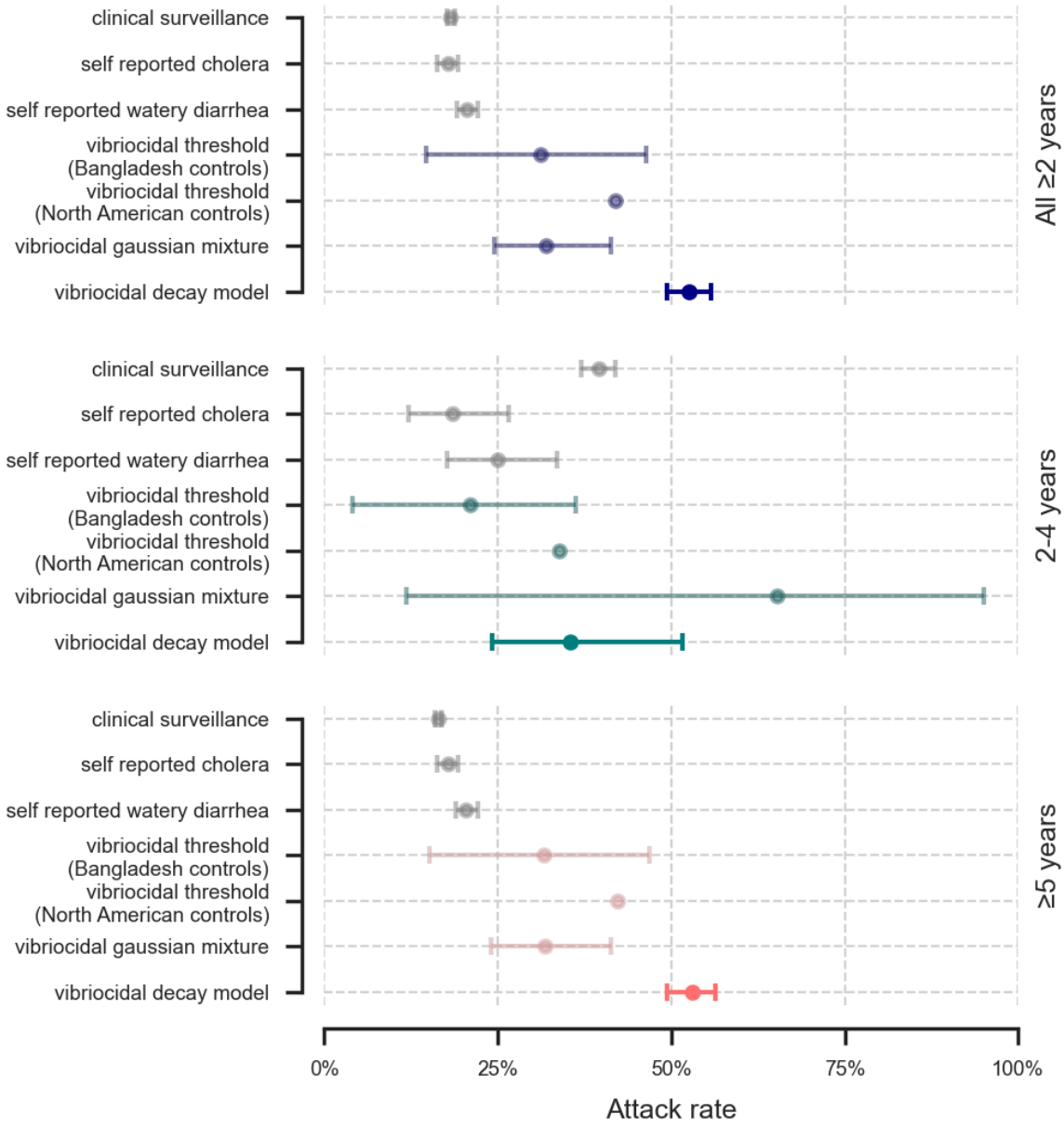

Figure S1: Estimates of clinical attack rate (grey) and infection rates (colored) for age groups 2 years and above, 2-4 years and 5 years and above.

Bayesian vibriocidal titer decay model

Parameter values and fit

The vibriocidal titer decay, parametrized by the individual baseline titer ϖ^i, the individual titer rise λ^i and the titer decay rate r, didn't change significantly during the model inference, and the posterior values are similar to the priors from Jones et al. (2022). Samples from the posterior distribution are shown in Figure S2. There is a significant overlap between baseline titer of the uninfected population and titer values of the infected population, even when infected titer values are at their maximum. This overlap explains the sigmoid curve and a part of the uncertainty on our parameter identification. In addition, the figure shows, where the participants who reported a watery diarrhea onset date lie on the decay curve. Most individuals have a vibriocidal titer compatible with our decay model, though there is a proportion of people who reported a cholera diagnosis who have a titer lower than expected.

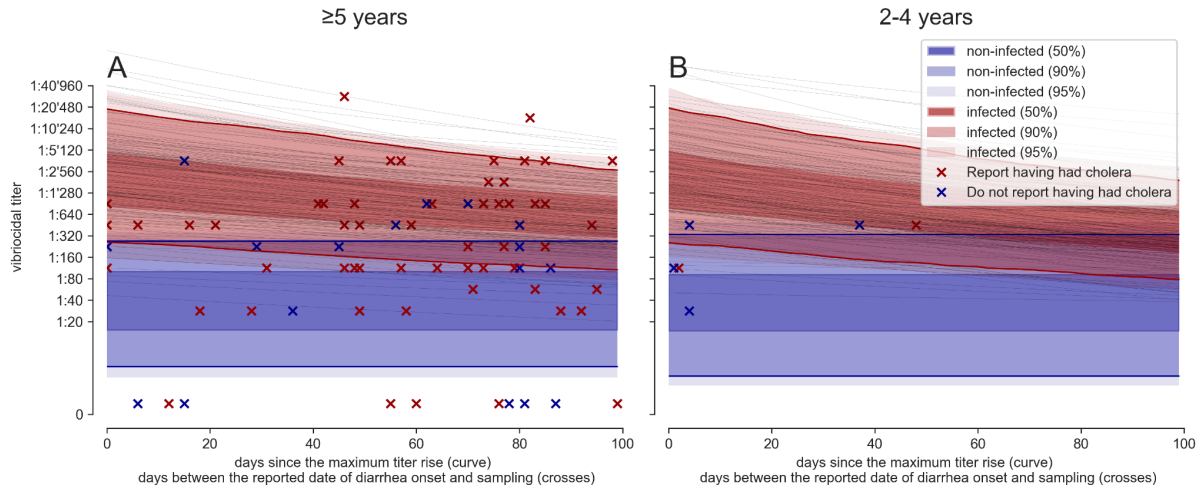

Figure S2: Visualization of the posterior of the vibriocidal titer decay model for individuals 5 and older (A) and those 2-4 years old (B). The decay of infected individuals (red shade, with 100 samples in gray lines) is shown from the maximum titer rise, 11 days after infection. Uninfected individual titer values remain constant at their baseline level (blue shade). The crosses show the position of survey participants who reported watery diarrhea onset date, as the delay between onset and the survey on the x-axis, relative to their measured vibriocidal titer on the y-axis. Those reporting to have been diagnosed with cholera by a health care worker are marked in red.

Posterior predictive checks

Posterior predictive checks compare the observed data with replicated data from the fitted decay model. The discrepancies between posterior predictive checks and posterior values may come from a inadequacy in the model definition (Figure S3). All observations fall within the range of posterior predictive values, but for the lower titer values, the model

underestimates the median count. One possible explanation would be that the discrete latent parameters indicating infection are not inferred correctly, a common pitfall in Bayesian inference. However, given that the posterior of the Bernoulli draws of the infection indicator δ^i are correctly distributed with respect to its inferred parameter value, the infection rate, iAR, we ruled out this possibility. It is thus most likely that this bias is due to our model specification and the difficulty to reconcile observed data with the provided decay parameters and the case time series.

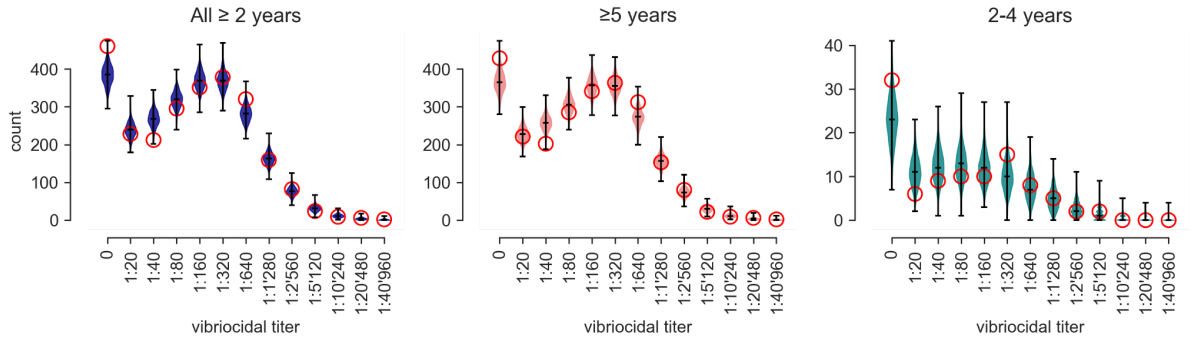

Figure S3: Posterior predictive checks of the three bayesian decay models developed for each age group. The red circle represents the observed count in each titer bin, and the violin plots the posterior predictive values.

MCMC diagnostic checks

The $\hat{R}$ convergence diagnostic, also known as the Gelman-Rubin diagnostic compares between- and within-chain estimates for model parameters to assess chain mixing (Vehtari et al. 2021). If chains have not well mixed, R-hat is larger than 1 and it is recommended not to use samples if the $\hat{R} > 1.05$. In Figure S4, we present the $\hat{R}$ for all parameters of the 5 and older and 2-4 years old models. All parameters in all models have an $\hat{R} < 1.05$.

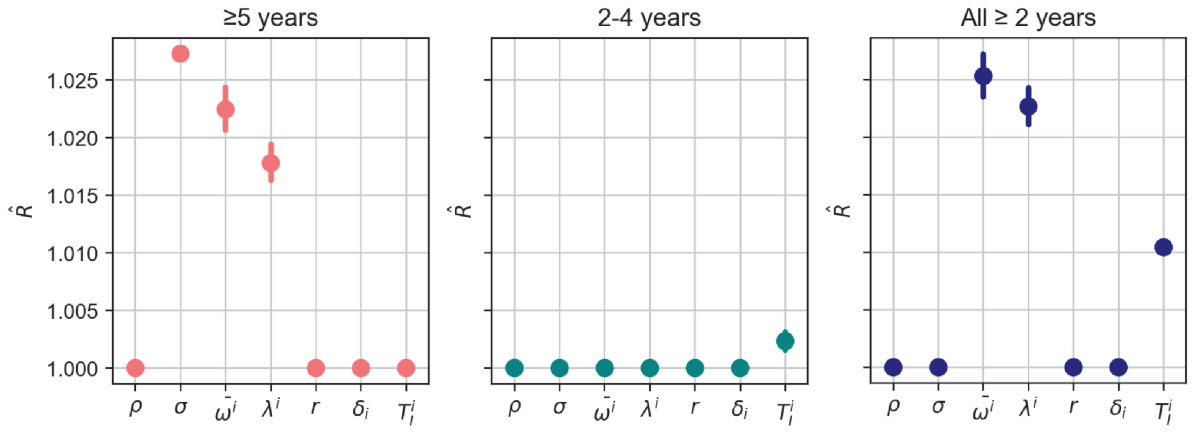

Figure S4: $\hat{R}$ convergence diagnostic for our model parameters. For parameters with individual values (subscript i), we report the mean (dot) and 95% interval (bars).

Posterior

In Figure S5 the prior and posterior of all our model parameters are presented. For the parameters with individual values (subscript i), we show all values across all individuals.

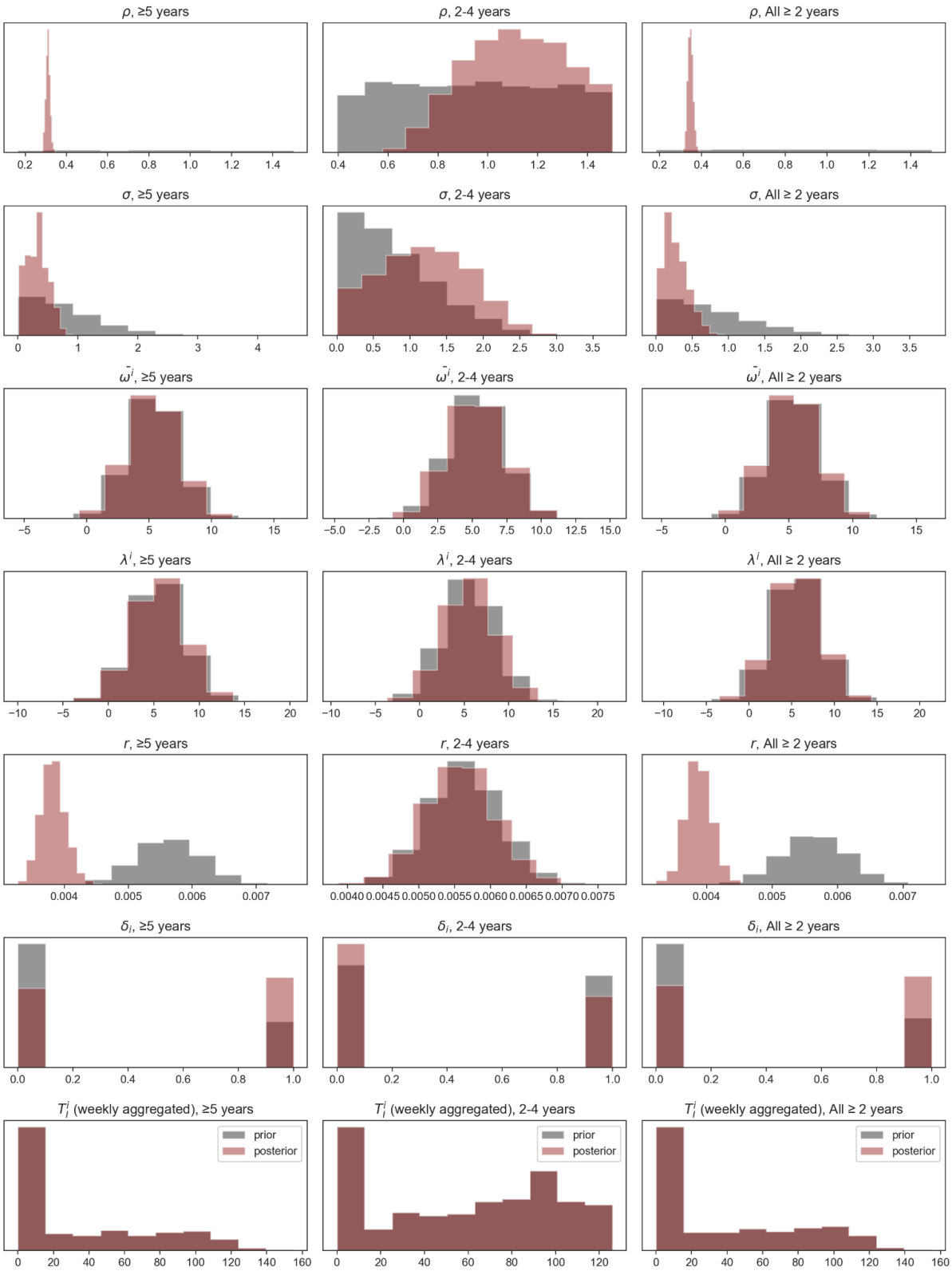

Figure S5: Prior (grey) and posterior (red) distributions of our model parameters. The day of infection (T_I^i) has been aggregated by week to improve clarity.

Sensitivity analysis on the decay rates of antibody

The antibody decay rate parameter r is derived from prior studies and informed by a single, relatively narrow prior common to all age groups (Table 1). A hypothetical explanation for the observed difference in infection rate and infection-to-reported-case ratio between age groups could be a true decay rate varying by age. We performed a sensitivity analysis, assessing the influence of the decay rate on the infection-to-case ratio, specifically whether different decay rates could explain the differences in infection rate between age groups.

We fit our models for 2-4 years olds and 5 years and older with a prior on the decay rate multiplied by 1/10, 1/2, 5, and 9. As shown in Figure S6, the differences in decay rate influence our estimates, changing the density and modes of the infection to the reported case ratio and of the infection rate. For example in 2-4 years olds, the mean of the infection to the reported case ratio varied from 0.8 (95%CrI 0.7-1.1) with a multiplier of 1/10 to 1.3 (95%CrI 0.7-2.2) with a multiplier of 9. This variation, however, remains small compared to the variation between age groups. The distinction between the two age groups remains clear even when assuming a much faster or slower decay rate for children. The different infection rates and infection to reported case ratios for the two age groups are thus not an artifact of our choice of the decay rate prior. Similarly, when varying the decay rates in the population 5 years and older, the variations of our main outcomes, the infection rate and the number of infections per reported case, remain small, except for the case when the decay rate would be several times higher than in our main model.

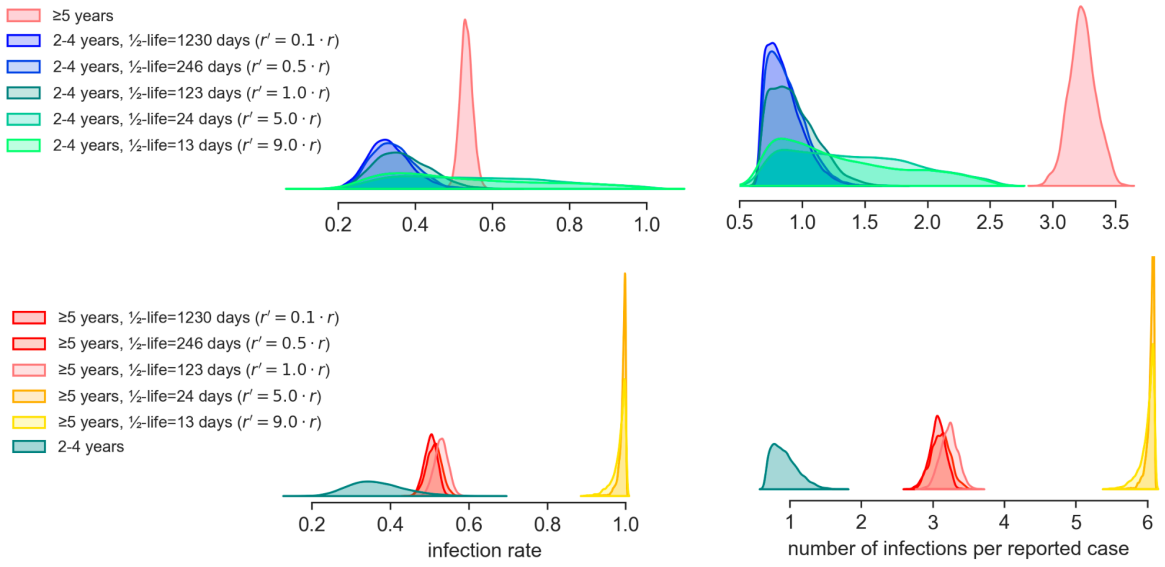

Fig S6: infection to reported case ratio and infection rate derived from our main model for individuals 5 years and older, and from a model with different priors on the decay rate parameter for the individuals 2-4 years old (top). and the same quantities derived from our main model for 2-4 years olds, and from a model with different priors on the decay rate parameter for the individuals 5 years and older (bottom).

Sensitivity analysis on the baseline titer

The baseline titer and titer rise are obtained from a study in Bangladesh (Jones et al. 2022). It is likely that this population is regularly exposed to cholera, different from the population of Haiti in 2010. We perform a sensitivity analysis on the baseline titer to evaluate the influence of these priors on our main results. We decrease the prior of the (log-transformed) baseline titer by ¼ and ½ of its standard deviation (3.96 as per Table 1). Given that the baseline titer $\bar{\omega^i}$ and the titer rise λ^i are sampled from a multivariate distribution with negative covariance, these lower baseline titers are compensated by a higher titer rise. As shown in Figure S7, lower priors on the baseline titers significantly increase the inferred infection rate and thus the number of infections per reported case. For example in 2-4 years old, the mean of the infection to the reported case ratio increased to 1.3 (95%CrI 0.9-1.7) with a decrease of the prior of the baseline titer by ¼ of its standard deviation, and to 1.8 (95%CrI 1.3-2.2) with ½ of its standard deviation, and to respectively 4.2 (95%CrI 4.0-4.4) and 5.5 (95%CrI 5.3-5.7) for individuals 5 year and older. The difference between age groups is still present if the prior value remains the same for both groups.

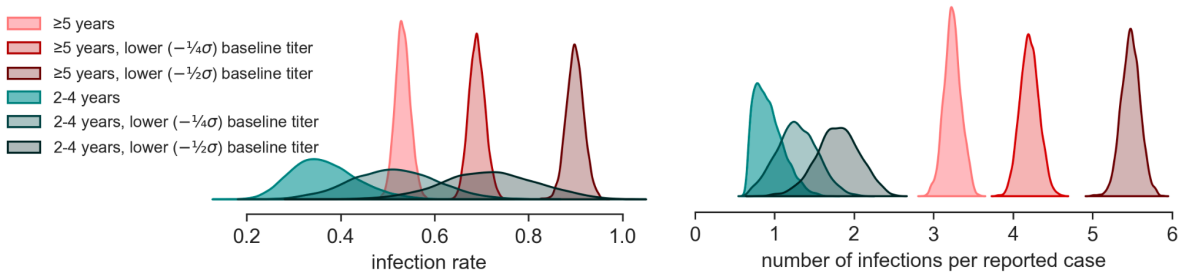

Fig S7: infection to reported case ratio and infection attack rate derived from our main model and models with the prior of the log-transformed baseline titer decreased by ¼ and ½ of its standard deviation (darker shades).

References

Azman, Andrew S., Justin Lessler, Francisco J. Luquero, Taufiqur Rahman Bhuiyan, Ashraful Islam Khan, Fahima Chowdhury, Alamgir Kabir, et al. 2019. “Estimating Cholera Incidence with Cross-Sectional Serology.” *Science Translational Medicine* 11 (480): eaau6242. https://doi.org/10.1126/scitranslmed.aau6242.

Bart, Kenneth J., Zahidul Huq, Moslemuddin Khan, Wiley H. Mosley, Nuruzzaman, and A. K. M. Golam Kibriya. 1970. “Seroepidemiologic Studies during a Simultaneous Epidemic of Infection with El Tor Ogawa and Classical Inaba Vibrio Cholerae.” *Journal of Infectious Diseases* 121 (May): S17–24. https://doi.org/10.1093/infdis/121.Supplement.S17.

Blake, P. A., D. T. Allegra, J. D. Snyder, T. J. Barrett, L. McFarland, C. T. Caraway, J. C. Feeley, et al. 1980. “Cholera--a Possible Endemic Focus in the United States.” *The New England Journal of Medicine* 302 (6): 305–9. https://doi.org/10.1056/NEJM198002073020601.

Feachem, R. G. 1982. “Environmental Aspects of Cholera Epidemiology. III. Transmission and Control.” *Tropical Diseases Bulletin* 79 (1): 1–47.

https://www.cabdirect.org/cabdirect/abstract/19822700907.
Gangarosa, E. J. 1974. “The Epidemiologic Basis of Cholera Control.” *Bulletin of the Pan American Health Organization (PAHO);8(3),1974*. https://iris.paho.org/handle/10665.2/27038.
Glass, Roger I., and Robert E. Black. 1992. “The Epidemiology of Cholera.” In *Cholera*, edited by Dhiman Barua and William B. Greenough, 129–54. Boston, MA: Springer US. https://doi.org/10.1007/978-1-4757-9688-9_7.
Harris, Jeffrey R., Scott D. Holmberg, Robert D. R. Parker, Dale E. Kay, Timothy J. Barrett, Charles R. Young, Myron M. Levine, and Paul A. Blake. 1986. “Impact of Epidemic Cholera in a Previously Uninfected Island Population: Evaluation of a New Seroepidemiologic Method.” *American Journal of Epidemiology* 123 (3): 424–30. https://doi.org/10.1093/oxfordjournals.aje.a114257.
Hegde, Sonia, Ashraf Islam Khan, Javier Perez-Saez, Ishtiakul Islam Khan, Juan Dent Hulse, Md Taufiqul Islam, Zahid Hasan Khan, et al. 2023. “Estimating the Gap between Clinical Cholera and True Community Infections: Findings from an Integrated Surveillance Study in an Endemic Region of Bangladesh.” medRxiv. https://doi.org/10.1101/2023.07.18.23292836.
Johnston, Jeffrey M., Deborah L. Martin, James Perdue, Louise M. McFarland, Charles T. Caraway, Edwin C. Lippy, and Paul A. Blake. 1983. “Cholera on a Gulf Coast Oil Rig.” *New England Journal of Medicine* 309 (9): 523–26. https://doi.org/10.1056/NEJM198309013090903.
Jones, Forrest K., Taufiqur R. Bhuiyan, Rachel E. Muise, Ashraful I. Khan, Damien M. Slater, Kian Robert Hutt Vater, Fahima Chowdhury, et al. 2022. “Identifying Recent Cholera Infections Using a Multiplex Bead Serological Assay.” Edited by Barry R. Bloom. *mBio* 13 (6): e01900-22. https://doi.org/10.1128/mbio.01900-22.
King, Aaron A., Edward L. Ionides, Mercedes Pascual, and Menno J. Bouma. 2008. “Inapparent Infections and Cholera Dynamics.” *Nature* 454 (7206): 877–80. https://doi.org/10.1038/nature07084.
Mccormack, William M., Md Shafiqul Islam, and Md Fahimuddin. 1969. “A Community Study of Inapparent Cholera Infections.” *American Journal of Epidemiology* 89 (6): 658–64. http://aje.oxfordjournals.org/content/89/6/658.
Rogan, Walter J., and Beth Gladen. 1978. “Estimating Prevalence from the Results of a Screening Test.” *American Journal of Epidemiology* 107 (1): 71–76. https://doi.org/10.1093/oxfordjournals.aje.a112510.
Stan Development Team. 2020. *RStan: The R Interface to Stan*. http://mc-stan.org/.
Vehtari, Aki, Andrew Gelman, Daniel Simpson, Bob Carpenter, and Paul-Christian Bürkner. 2021. “Rank-Normalization, Folding, and Localization: An Improved Rˆ for Assessing Convergence of MCMC (with Discussion).” *Bayesian Analysis* 16 (2): 667–718. https://doi.org/10.1214/20-BA1221.