DISSERTATION

STATISTICAL MODELS FOR COVID-19 INFECTION FATALITY RATES AND DIAGNOSTIC TEST DATA

Submitted by

Sierra Pugh

Department of Statistics

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Summer 2023

Doctoral Committee:

Advisor: Ander Wilson
Co-Advisor: Bailey K. Fosdick

Kayleigh Keller
Mary Meyer
Molly Gutilla

ABSTRACT

STATISTICAL MODELS FOR COVID-19 INFECTION FATALITY RATES AND DIAGNOSTIC TEST DATA

The COVID-19 pandemic has had devastating impacts worldwide. Early in the pandemic, little was known about the emerging disease. To inform policy, it was essential to develop data science tools to inform public health policy and interventions. We developed methods to fill three gaps in the literature. A first key task for scientists at the start of the pandemic was to develop diagnostic tests to classify an individual's disease status as positive or negative and to estimate community prevalence. Researchers rapidly developed diagnostic tests, yet there was a lack of guidance on how to select a cutoff to classify positive and negative test results for COVID-19 antibody tests developed with limited numbers of controls with known disease status. We propose selecting a cutoff using extreme value theory and compared this method to existing methods through a data analysis and simulation study. Second, there lacked a cohesive method for estimating the infection fatality rate (IFR) of COVID-19 that fully accounted for uncertainty in the fatality data, seroprevalence study data, and antibody test characteristics. We developed a Bayesian model to jointly model these data to fully account for the many sources of uncertainty. A third challenge is providing information that can be used to compare seroprevalence and IFR across locations to best allocate resources and target public health interventions. It is particularly important to account for differences in age-distributions when comparing across locations as age is a well-established risk factor for COVID-19 mortality. There is a lack of methods for estimating the seroprevalence and IFR as continuous functions of age, while adequately accounting for uncertainty. We present a Bayesian hierarchical model that jointly estimates seroprevalence and IFR as continuous functions of age, sharing information across locations to improve identifiability. We use this model to estimate seroprevalence and IFR in 26 developing country locations.

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my advisors: Bailey Fosdick and Ander Wilson. Bailey is deeply invested in the success of her students, and I am grateful for her exceptional mentorship, encouragement, and guidance. She introduced me to many interesting research projects and supported me through them. Ander was willing to jump in as a co-advisor in my last semester and has been an incredible advisor. I truly value his feedback and encouragement. This dissertation would not exist without my advisors' constant support and dedication.

Furthermore, I would like to acknowledge and express my deepest appreciation to my previous mentors Mevin Hooten and Matthew Heaton. They provided unwavering support and helped me develop the confidence and skills I needed to complete a PhD.

I am grateful to the many collaborators I have had the opportunity to work with. They provided thought-provoking questions, taught me about their areas of expertise, and enriched my research experience.

I would like to thank my committee members, Kayleigh Keller, Mary Meyer, Molly Gutilla, and preliminary committee member, Julia Sharp, for their valuable feedback and insightful discussions.

I am thankful to my family for supporting, encouraging, and believing in me throughout this journey. Their constant support was and continues to be invaluable.

Lastly, I extend my heartfelt appreciation to my friends and colleagues who offered much-needed encouragement throughout this research endeavor and brought joy along the way.

To everyone who has contributed to my academic and personal growth, directly or indirectly, I extend my sincere gratitude. This dissertation would not have been possible without your support and guidance.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

The coronavirus disease 2019 (COVID-19) pandemic has had devastating impacts worldwide with 6.72 million recorded COVID-19 deaths as of 2023 (Our World in Data, 2022). Cutler and Summers (2020) estimated the cost of the COVID-19 pandemic at over 16 trillion dollars, accounting for lost GDP, premature deaths, long term health impairments, and mental health impairments. Timely and accurate information about the disease was essential to allow data-informed policies and allocation of resources. Perhaps no single piece of information is more essential than understanding who is, or has been, infected with COVID-19. However, early in the pandemic, tests were extremely limited and preferentially given to those showing symptoms (Campbell et al., 2022). Those with asymptomatic or more mild infections were less likely to be recorded, leading to underreporting of COVID-19 cases and reported cases skewed towards more severe cases. Thus, instead of relying on reported cases, antibody tests were used to estimate the proportion of the population that had been infected. As of April 2023, SeroTracker had gathered over 4000 SARS-CoV-2 seroprevalence studies (Arora et al., 2021; SeroTracker, 2023). However, tests for COVID-19 antibodies were rapidly developed with limited lab validation data, leading to uncertainty in the test characteristics. In this dissertation, we develop statistical approaches to improve testing accuracy with antibody tests and improve statistical inference on seroprevalence with antibody tests.

This dissertation is motivated by the COVID-19 pandemic, particularly the data and information needs that arose early in the pandemic. However, the methods developed in this dissertation are broadly applicable to any infectious disease. They are particularly relevant to tests that are rapidly developed for emerging infectious diseases as they address challenges such as limited validation data and heterogeneous reporting of deaths and infections.

1.1 Antibody test and serology study background

1.1.1 Antibody tests

Many types of tests for COVID-19 have been developed. Viral tests are used to identify an active severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infection, whereas antibody tests are used to identify a prior infection. Specifically, antibody tests detect the presence of antibodies against the SARS-CoV-2 virus in the blood, which develop in the weeks following an infection. Different antibody tests are designed to detect different types of antibodies and mimic different structures of the virus (Jacofsky et al., 2020). Thus, the sensitivities and specificities of the tests can vary dramatically, complicating comparisons between different tests and any sharing of information across tests.

Because it takes weeks for the antibodies to develop within an individual that contracted COVID-19 and they persist for months post infection, antibody tests cannot be used to determine who is currently infected or who should quarantine. The advantage of antibody tests is they can retroactively give a measure of previous infections. Antibody tests have been widely used to estimate the proportion of the population that has been infected with SARS-CoV-2 via serology studies (e.g., Garcia-Basteiro et al., 2020; Nisar et al., 2021), as well as to measure the lasting immune response either from a SARS-CoV-2 infection or from a COVID-19 vaccine (e.g., Gallichotte et al., 2021; Ward et al., 2022).

1.1.2 Estimating the number of COVID-19 infections

Reported COVID-19 cases are a dramatic undercount of the total number of COVID-19 cases due to asymptomatic cases and limited testing (National Academies of Sciences, Engineering, and Medicine, 2020). Thus, researchers rely on seroprevalence studies to estimate the number of prior COVID-19 infections. In seroprevalence studies, researchers administer antibody tests to a sample from the population of interest. The proportion of the sample with antibodies against COVID-19 is then used as an estimator for the proportion of the population that has been infected. Seroprevalence studies are not a perfect measure as not everyone who is infected with SARS-CoV-2 develops

antibodies against the virus, i.e., seroconverts (Lipsitch et al., 2020), and the antibody levels can wane over time, i.e., serorevert (Brazeau et al., 2022). Additionally, through vaccination, individuals develop COVID-19 antibodies (Ward et al., 2022). However, in this dissertation, we focus on seroprevalence studies conducted early in the pandemic, before seroreversion and vaccines were large concerns. Vaccines were not available, and there was less time between potential infection and the seroprevalence study to allow for waning antibody levels.

Early seroprevalence studies were, however, complicated by the uncertainty associated with antibody tests. No test is perfectly accurate as they inevitably result in some false positives and false negatives. Thus, prevalence estimates based on seroprevalence studies must account for these error rates. For tests rapidly developed early in the pandemic, lab validation sample sizes were limited, and thus, there was considerable uncertainty associated with the sensitivity and specificity of a test, including the cutoff value needed to achieve a desired specificity for a test. The test characteristic uncertainty translates into uncertainty in the prevalence that must be accounted for (Gelman and Carpenter, 2020).

One of the common approaches for estimating the prevalence from a serology study is the Rogan-Gladen approach (Rogan and Gladen, 1978), which uses the sensitivity, specificity, and positivity rate to estimate the prevalence (e.g., Axfors and Ioannidis, 2022; Levin et al., 2020; Pezzullo et al., 2023). Confidence intervals for the seroprevalence are then created, but many omit the uncertainty in the sensitivity and specificity values themselves. Seeing this gap, Gelman and Carpenter (2020) proposed a Bayesian model for jointly modeling sensitivity, specificity, and prevalence, which fully accounted for the uncertainty from the serology study and lab validation data.

1.1.3 Estimating the infection fatality rate

The infection fatality rate (IFR) for COVID-19 is the proportion of those infected with SARS-CoV-2 who then die from the disease. Thus, it requires a measure of the total number of infections and the total number of deaths. The accuracy of COVID-19 fatality records varied based on re-

sources and death registration systems; however, we lack empirical evidence of their accuracies, so we treat the number of deaths as known. The total number of infections is typically estimated from seroprevalence studies and has uncertainty. This means estimation of IFR has the same sources of uncertainty that serology studies do: sampling variability and test characteristic uncertainty.

A number of previous papers have estimated IFR for specific age bins or as continuous functions of age (e.g., Campbell and Gustafson, 2021; Levin et al., 2022; Perez-Saez et al., 2021). However, many of these papers do not fully account for the test characteristic uncertainty (COVID-19 Forecasting Team, 2022; Levin et al., 2020; O'Driscoll et al., 2021; Pezzullo et al., 2023), and none jointly model the test characteristics, seroprevalence, and IFR in one cohesive model.

1.2 Outline

1.2.1 Estimating antibody test cutoffs

In Chapter 2 we focus on enzyme-linked immunoassay (ELISA) antibody tests, which measure the amount of antibodies in a blood sample that bind to an antigen, a protein meant to mimic a piece of the SARS-CoV-2 virus. These tests result in optical density (OD) measures, which are proportional to the level of antibodies in the sample (Jacofsky et al., 2020; Nasrallah et al., 2021).

Antibody tests require a cutoff for the OD values to classify positive and negative test results. Early in the pandemic, when the proportion of the population infected was low, correctly classifying negative samples was a higher priority in order to have a more accurate test overall and to limit the number of false positive results. For this reason, the Centers for Disease Control and Prevention recommended selecting cutoffs to ensure a high specificity such as 0.995, meaning 99.5% of those without antibodies would test negative (Centers for Disease Control and Prevention, 2020). Common solutions for selecting a cutoff targeting a specific specificity include fitting a normal distribution to the negative controls (or some transformation of them) and estimating the corresponding quantile of this normal distribution, or finding the corresponding empirical percentile of the negative controls (Devanarayan et al., 2017; Jordan and Staack, 2021; Zhang, 2021). Variations on the normal distribution approach include fitting a t-distribution instead (Hoffman and Berger,

2011; Shen et al., 2015) or varying how the normal distribution is fit: mean and standard deviation versus median and mean absolute deviation (Devanarayan et al., 2017).

Because we typically seek to estimate an extreme quantile of the negative controls distribution, we propose using extreme value theory to estimate the cutoff in Chapter 2. We fit a generalized Pareto distribution to the upper tail of the negative controls and estimate the desired quantile using this fitted distribution. Furthermore, out of the existing methods, there lacks a formal comparison showing which cutoff estimation method is preferred when lab validation sample sizes are limited, as was the case early in the COVID-19 pandemic. Thus, we formally compare cutoff estimation methods via a data analysis and simulation study using COVID-19 antibody test data from the first year of the pandemic. We consider seven cutoff estimation methods, including our new method, with varied target specificities and focus on the situation where there is limited validation sample sizes.

1.2.2 Estimating Age-Specific IFR: Binned age case

Case and mortality data are often reported for age ranges, i.e., age bins. The specific age bins used vary across studies, and death reporting procedures vary across locations. In Chapter 3, we present a Bayesian model that jointly models seroprevalence, antibody test characteristics, and age-specific IFR based on the age bins for which the data is available. We account for the challenges of limited serology and test validation sample sizes, misaligned bins between the seroprevalence and fatality data, heterogeneous test characteristics, and uncertainty in the inherent test characteristics. By estimating serology and IFR at the age bin level, we can apply to our model to locations with any granularity level of data. In contrast to existing methods (e.g., COVID-19 Forecasting Team, 2022; Levin et al., 2020), we minimize assumptions by limiting how information was shared across locations and not assuming any particular relationship between IFR and age such as log-linearity. We compare our new method for prevalence estimation to the commonly used Rogan-Gladen method (Rogan and Gladen, 1978), which does not account for test characteristic uncertainty. We demonstrate the improvements to our estimates and their uncertainty when consid-

ering the data sources simultaneously. Using our proposed method we analyzed test characteristic, seroprevalence, and fatality data for 107 locations from 44 developing countries to demonstrate the utility of our tools.

1.2.3 Estimating Age-Specific IFR: Continuous age case

In Chapter 4 we focus on developing country locations with age-specific seroprevalence and fatality data. We estimate location- and age-specific seroprevalence and IFR as continuous functions of age to compare the burden of COVID-19 across locations that had limited data available at varying age bins and population age distributions. To meet this objective, we develop a hierarchical Bayesian model that shares information on the shape of the IFR curve across locations. We further share information across age bins by estimating seroprevalence and IFR as continuous functions of age. The continuous functions naturally allow for misaligned bins between the serology and death data. We estimate and compare age-specific IFR in 26 locations in developing country locations.

1.2.4 Conclusion

In Chapter 5, we summarize the contributions of the prior chapters and explore paths for future work.

Chapter 2

Estimating cutoff values for diagnostic tests to achieve target specificity using extreme value theory

2.1 Introduction

When faced with an emerging infectious disease outbreak, it is imperative to rapidly develop diagnostic tests to determine individual disease status and estimate community prevalence. Both individual- and community-level information is necessary to target public health interventions and deploy medical resources. In addition to designing tests that accurately measure biological samples for evidence of disease (e.g., antibodies), a critical challenge is how to classify quantitative test results as positive or negative. Therefore, a threshold, based on controls with known disease status, must be selected to determine positive and negative test results.

Estimating cutoffs for newly developed tests provides unique challenges. First, tests can show little separation in the distributions for positive and negative controls. The threshold can be chosen to target a particular sensitivity or specificity, but not both. Second, many early tests have a limited number of controls with known disease status. For example, a study found that of 47 coronavirus disease 2019 (COVID-19) antibody tests used in developing countries, the majority had fewer than 200 negative controls and some had as few as 31 (Levin et al., 2022). Thus, estimating the cutoff that will have the desired sensitivity or specificity must be done from limited data.

This raises two important questions. First, what sensitivity or specificity should be targeted? Second, how to best estimate a cutoff value for the target sensitivity or specificity? For emerging diseases, we expect the prevalence to be low. Thus, to optimize the number of tests with the correct result, we should prioritize correctly identifying negative results and consequently have a high specificity (Takahashi et al., 2020; Klumpp-Thomas et al., 2021). For this reason, the Centers for Disease Control and Prevention (CDC) recommended high specificity, such as 0.995, for tests de-

veloped in the early part of the COVID-19 pandemic (Centers for Disease Control and Prevention, 2020). To achieve a target specificity, researchers commonly use the same quantile of the negative controls distribution as a cutoff. Two common approaches to estimating a quantile of the negative controls are to use the empirical quantile or use the quantiles of a parametric distribution, such as normal or lognormal, fitted to the data (Klumpp-Thomas et al., 2021; Devanarayan et al., 2017; Hoffman and Berger, 2011; Zhang et al., 2013). However, these methods have not been specifically evaluated for selecting cutoffs of rapidly developed tests for emerging diseases.

We provide two contributions to the literature. First, we propose a method to estimate a cutoff for a desired target specificity based on extreme value theory. Our proposed approach is to fit a generalized Pareto distribution to the upper tail of the negative control data (Pickands III, 1975). This approach has been broadly used to estimate extreme values of events such as rainfall (Cooley et al., 2007), air pollution exposures (Martín et al., 2022), and stock prices (Kiriliouk et al., 2019), among other applications, but has never been applied to cutoff selection. Second, we compare commonly used methods and the proposed extreme value-based approach, for estimating the cutoffs of emerging disease tests through a simulation study and data application. We compare cutoff estimation methods based on their accuracy in achieving a target specificity, individual tests, and estimating community prevalence. We also compare the impact of target specificities on these outcomes. In our data analysis, we focus on enzyme linked immunosorbent assay (ELISA) antibody test data collected during the first year of the COVID-19 pandemic. However, the methods proposed are general and can be applied to data from any test. In our simulation study, we demonstrated the extreme value method had the least bias for estimating a cutoff for a high target specificity and that lower target specificities are easier to estimate and may perform better when the objective is estimating prevalence.

2.2 Methods

2.2.1 Data

We used two data sources in our analysis. The training dataset contained blood samples from staff at long-term care facilities in Colorado, USA, sampled between June and December of 2020. A total of 226 staff members underwent up to five tests each, resulting in 690 samples. Each sample was tested using three different antibody tests: a neutralization assay test and two different ELISA antibody tests. One ELISA test targeted the spike protein and the other targeted the receptor-binding domain (RBD). The neutralization assay test is considered the "gold standard" in antibody testing, so we used these results to identify positive and negative controls (Bewley et al., 2021; Cohen et al., 2008; Eyal et al., 2005). This resulted in 245 positive controls and 445 negative controls. Additional details are given elsewhere (Gallichotte et al., 2021).

The testing dataset consisted of samples from 186 skilled nursing staff during May 2020. Researchers collected one sample from each staff member and ran multiple antibody tests, including the spike and RBD ELISA tests used in the training dataset, as described elsewhere (Nehring et al., 2023).

For both datasets, we normalized the results of the ELISA tests to account for batch effects. We calculated the positive to negative ratio (P/N) by dividing the average optical density for each sample by the average of the negative controls run on the same plate (Zhang et al., 2013).

2.2.2 Statistical methods to estimate cutoff values

Our objective in determining the cutoff value is to estimate the Q quantile of the negative controls for a target specificity of Q. Let $\boldsymbol{x}$ denote the vector of n negative control test results.

Normal method. The normal method finds the Q quantile of a normal distribution with a mean of $\overline{x}$ and a standard deviation of s_x, where $\overline{x}$ and s_x denote the mean and standard deviation of $\boldsymbol{x}$, respectively (Klumpp-Thomas et al., 2021; Devanarayan et al., 2017; Hoffman and Berger, 2011; Zhang et al., 2013; Jordan and Staack, 2021; Mire-Sluis et al., 2004).

Lognormal method. The lognormal method is the normal method applied to the data after a natural log transformation (Devanarayan et al., 2017; Zhang et al., 2013; Jordan and Staack, 2021). This equates to fitting a lognormal distribution to the raw data and using the Q quantile of that lognormal distribution.

MAD method. The MAD method is a modification of the normal method that replaces the mean with the median, $\widetilde{x}$, and the standard deviation with the scaled mean absolute deviation (MAD), $s_x^{\text{MAD}} = 1.4826 \times \text{median}\{|x_i - \text{median}(\boldsymbol{x})|\}_{i=1}^n$ (Devanarayan et al., 2017; Hoffman and Berger, 2011; Mire-Sluis et al., 2004). This approach is intended to be robust to outliers.

Log MAD method. The log MAD method is the MAD method applied to natural log transformed data (Devanarayan et al., 2017; Hoffman and Berger, 2011).

Empirical method. The empirical method uses the empirical quantile of $\boldsymbol{x}$ as an estimator of the cutoff, avoiding parametric assumptions (Devanarayan et al., 2017; Hoffman and Berger, 2011; Zhang et al., 2013; Jordan and Staack, 2021; Mire-Sluis et al., 2004).

Pareto method using the upper 10% (Pareto 0.9) and upper 5% (Pareto 0.95). The Pareto method, based on extreme value theory, fits a generalized Pareto distribution to the upper tail of $\boldsymbol{x}$. Let u denote some threshold, and $\boldsymbol{y}$ be the values in $\boldsymbol{x}$ that exceed u. Asymptotically, under regularizing conditions, $\boldsymbol{y}$ follows the generalized Pareto distribution as u approaches the upper limit of the distribution for $\boldsymbol{x}$ (Pickands III, 1975; Balkema and De Haan, 1974). The generalized Pareto distribution is

$$G(y; \sigma_u, \xi) = \begin{cases} 1 - \left(1 + \frac{\xi(y-u)}{\sigma_u}\right)^{-\frac{1}{\xi}} & \xi \neq 0 \\ 1 - \exp\left(-\frac{y-u}{\sigma_u}\right) & \xi = 0. \end{cases} \tag{2.1}$$

We make the simplifying assumption that $\xi = 0$, which results in a shifted exponential distribution and has been shown to be preferable for small sample sizes (Rosbjerg et al., 1992). Thus, we only estimate σ_u from the data as u is pre-specified.

We set u to be the k^{th} quantile of $\boldsymbol{x}$ and consider two values of k: 90 and 95 (DuMouchel, 1983; Durán-Rosal et al., 2022). We then fit an exponential distribution to $\boldsymbol{y} - u$. We use maximum likelihood to estimate $\boldsymbol{y} - u \sim \exp(\lambda)$ such that $\widehat{\lambda} = \frac{1}{\overline{y}-u}$ where $\overline{y}$ is the sample mean of $\boldsymbol{y}$.

Since $\boldsymbol{y}$ is assumed to be the upper $(100 - k)\%$ of the data, the upper $Q' = \frac{Q-k/100}{1-k/100}$ quantile of our fitted exponential distribution corresponds to the upper Q quantile of the data overall. Thus, we set the cutoff as

$$C = F^{-1}(Q', \widehat{\lambda}) + u \tag{2.2}$$

$$= -\widehat{\lambda}\log(1 - Q') + u, \tag{2.3}$$

where $F^{-1}(Q', \widehat{\lambda})$ is the inverse CDF of an exponential distribution with a scale parameter of $\widehat{\lambda}$, evaluated at Q'. When $Q = 0.95$ and $k = 95$, the cutoff estimate is equivalent to the empirical method estimate because $Q' = 0$.

Hybrid approaches. We also consider hybrid approaches that provide a data-driven approach to select a cutoff estimation method (Devanarayan et al., 2017; Zhang et al., 2013). We first test for normality using the Shapiro-Wilk test with a significance level of 0.05. If the test fails to reject, we use the normal method. If the test rejects normality, we natural log transform and test for normality again. If the test fails to reject, we use the lognormal method. If the test rejects normality, we use one of three methods: empirical, Pareto 90%, and Pareto 95% (henceforth referred to as hybrid empirical, hybrid Pareto 0.9, and hybrid Pareto 0.95, respectively).

Additional details on the estimation methods are given in Appendix A.1.

2.2.3 Statistical methods to estimate prevalence

To accurately estimate the proportion of the population with antibodies for the disease, the seroprevalence, we account for the sensitivity and specificity of the test via the Rogan-Gladen adjustment (Rogan and Gladen, 1978), modified to disallow any negative estimates. The prevalence

estimator is

$$\widehat{\pi} = \max\left(\frac{\widehat{p} + Q - 1}{Q + \widehat{\text{sens}} - 1}, 0\right), \tag{2.4}$$

where $\widehat{p}$ is the proportion of tests classified as positive in the testing data and $\widehat{\text{sens}}$ denotes the estimated sensitivity of the test: the proportion of the positive controls that correctly tested positive in the training data. We use the target specificity Q as the specificity estimate.

2.2.4 Data analysis

We established cutoffs for both the spike and RBD ELISA tests using two different target specificities: 0.95 and 0.995. For each target specificity and test, we estimated the cutoff using each of the seven methods described above and the three hybrid methods. To estimate the sensitivity for each cutoff, we used the samples with positive neutralization assay results as positive controls.

We then used the cutoffs to classify each observation in the testing dataset as positive or negative. The resulting positivity was used to calculate the Rogan-Gladen adjusted prevalence for each cutoff.

2.2.5 Simulation study

We modeled our simulated data after the training dataset. We fit mixture distributions separately to the positive and negative controls for each test. See Appendix A.2.

We sampled from the fitted mixture distributions to generate data for the simulation study. By sampling from known mixture distributions, we were able to calculate the true quantiles for the population we sampled from, allowing us to assess bias and the root mean squared error (RMSE) of the cutoff value.

We considered eight scenarios in our simulation study. The data was either simulated from the fitted spike P/N ratios distribution (scenario A) or the fitted RBD P/N ratios distribution (scenario B). We varied the training sample size between 50 and 200 controls of each type (positive and negative) and set the target specificity at 0.95 or 0.995. For each simulated training dataset, we

generated a corresponding testing dataset of size 500, with the number of positive and negative controls determined by the prevalence: either 0.05 or 0.3. We generated 10,000 training datasets and testing datasets.

For each training dataset, we estimated the cutoff using all seven methods and the three hybrid methods. Then, we estimated the sensitivity of the cutoff using the proportion of the positive controls in the training dataset that were correctly predicted as positive using that cutoff. We also used each cutoff to classify positive and negative results in the testing dataset. We calculated the Rogan-Gladen adjusted prevalence as previously described. We evaluated the cutoffs in terms of the bias and RMSE. We calculated the accuracy of the predictions for the testing dataset as the proportion of testing dataset observations that were correctly predicted for each cutoff.

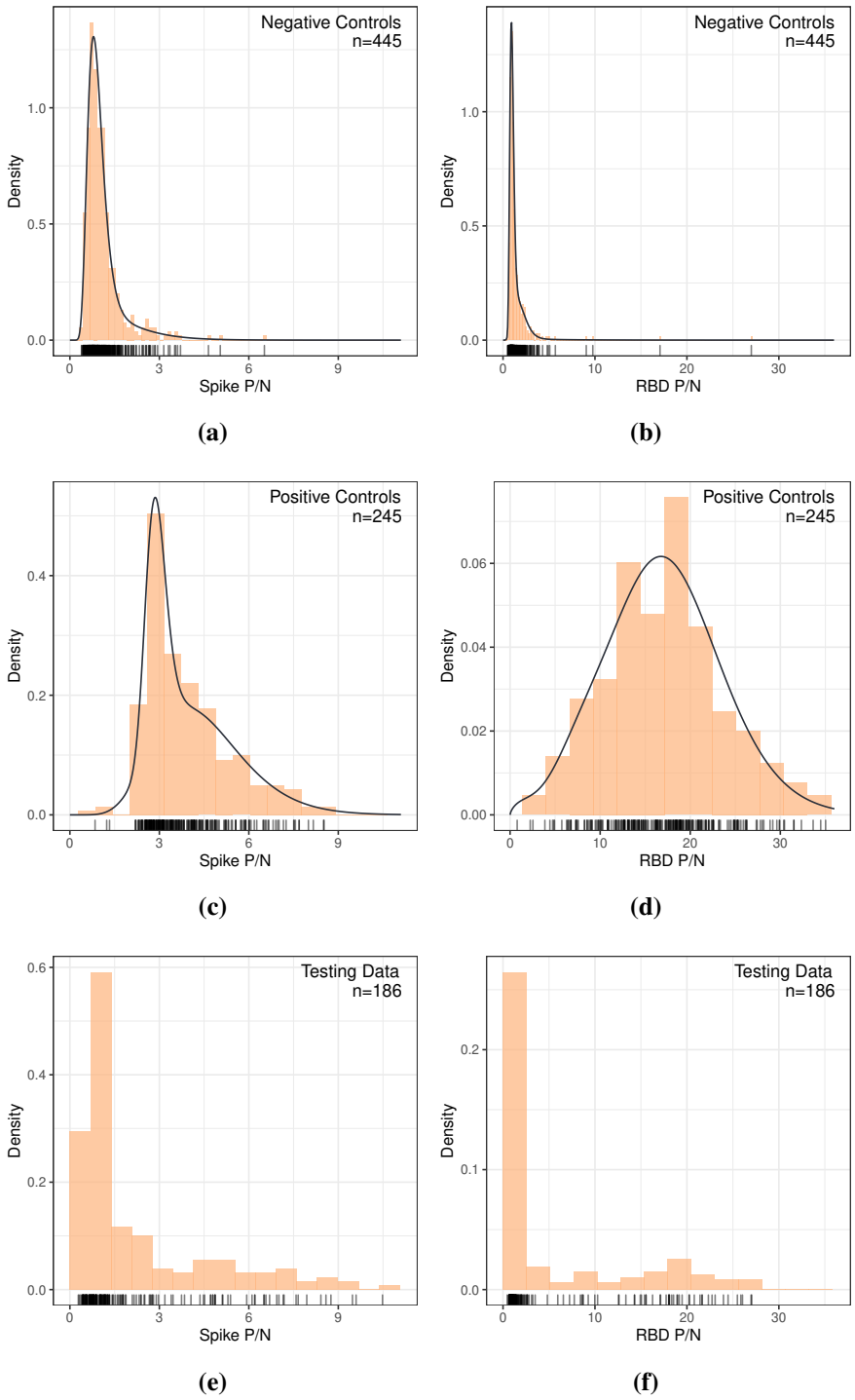


Figure 2.1: (a)-(d) Histogram of the training dataset for each test and control type overlaid with the corresponding mixture distribution from which the data was generated in the simulation study (training data only). The testing data set are in panels (e) and (f). The first column corresponds to the spike test, and the second to the receptor-binding domain (RBD) test. Training data was sampled from staff at long-term care facilities in Colorado, USA between June and December 2020. Testing data collected from skilled nursing staff in Colorado during May 2020.

2.3 Results

2.3.1 Data analysis

Figure 2.1 shows the negative control training data, positive control training data, and testing data for both the spike and RBD tests. The spike test had a smaller range of P/N ratios and less separation between the positive and negative controls. The RBD negative controls had a sparser upper tail, and the positive controls had a more symmetric distribution compared to the spike test.

Spike test

Figure 2.2 shows the training and testing data and the estimated cutoff for each method, target specificity, and test. Appendix A.3 shows the results in numerical form. Overall, the different estimation methods resulted in very different cutoff values. When targeting a specificity of 0.995, the spike test cutoffs ranged from 1.8 to 4.6 compared to a range of 1.5 to 2.5 when targeting a specificity of 0.95. The MAD normal methods consistently estimated the lowest cutoffs, while the empirical and Pareto methods resulted in the highest estimates. Using the hybrid approaches, we rejected normality for the untransformed and natural log transformed data and used the empirical and Pareto estimators.

Because the cutoffs are in the tail of the distribution for the negative controls, there are not many negative control observations between the cutoff values from the different methods (Figure 2.2). Thus, the differences in the cutoffs have minimal impact on the empirical specificities (Table 2.1). The cutoffs had a larger impact on the empirical sensitivity because there were many positive controls in the range of the cutoffs as shown in Figure 2.2. For example, the Pareto 0.9 and the lognormal cutoffs had similar training data empirical specificities, 0.993 versus 0.978, when targeting a specificity of 0.995. However, the empirical sensitivities were substantially different: 0.27 and 0.63, respectively.

The Rogan-Gladen adjusted prevalence estimate for each cutoff method is shown in Table 2.1. The prevalence estimates from cutoffs targeting a specificity of 0.95 ranged from 0.29 to 0.37. Those targeting 0.995 ranged from 0.29 to 0.64. Most prevalence estimates ranged from 0.26

to 0.42 with either target specificity, but the prevalence estimates from the empirical and Pareto cutoffs targeting a specificity of 0.995 were much larger, between 0.61 and 0.64.

RBD test

The estimated cutoffs for the RBD test were also more variable when targeting a specificity of 0.995. The MAD normal cutoffs were the smallest, and the empirical and Pareto cutoffs were similar to each other. We again rejected normality both for the raw and log transformed data, and the hybrid method estimates were equivalent to the empirical and Pareto estimates.

The RBD test showed greater separation in the distributions of the negative controls and positive controls, resulting in higher and more consistent empirical sensitivities, with all sensitivities greater than 0.87 (Table 2.1). The reduced variability in the empirical sensitivity estimates between estimation methods resulted in less variability of the prevalence estimates, compared to the spike tests.

Table 2.1: Rogan-Gladen adjusted prevalence estimate of the testing dataset for each cutoff method, test, and target specificity.

	Spike			RBD		
	Prevalence	Sensitivity	Specificity	Prevalence	Sensitivity	Specificity
Target specificity=0.995						
Empirical	0.64	0.28	0.99	0.26	0.87	0.99
Normal	0.30	0.77	0.97	0.28	0.96	0.99
Log Normal	0.35	0.63	0.98	0.28	0.98	0.98
MAD	0.32	0.99	0.91	0.39	1.00	0.85
Log MAD	0.29	0.93	0.95	0.32	0.99	0.93
Pareto 0.9	0.63	0.27	0.99	0.27	0.91	0.99
Pareto 0.95	0.61	0.31	0.99	0.26	0.87	0.99
Target specificity=0.95						
Empirical	0.29	0.93	0.95	0.30	0.99	0.95
Normal	0.30	0.98	0.94	0.28	0.98	0.98
Log Normal	0.31	0.99	0.93	0.31	0.99	0.94
MAD	0.37	0.99	0.85	0.42	1.00	0.80
Log MAD	0.34	0.99	0.90	0.39	1.00	0.84
Pareto 0.9	0.29	0.96	0.94	0.28	0.99	0.97

Abbreviations: mean absolute deviation, MAD; receptor-binding domain, RBD

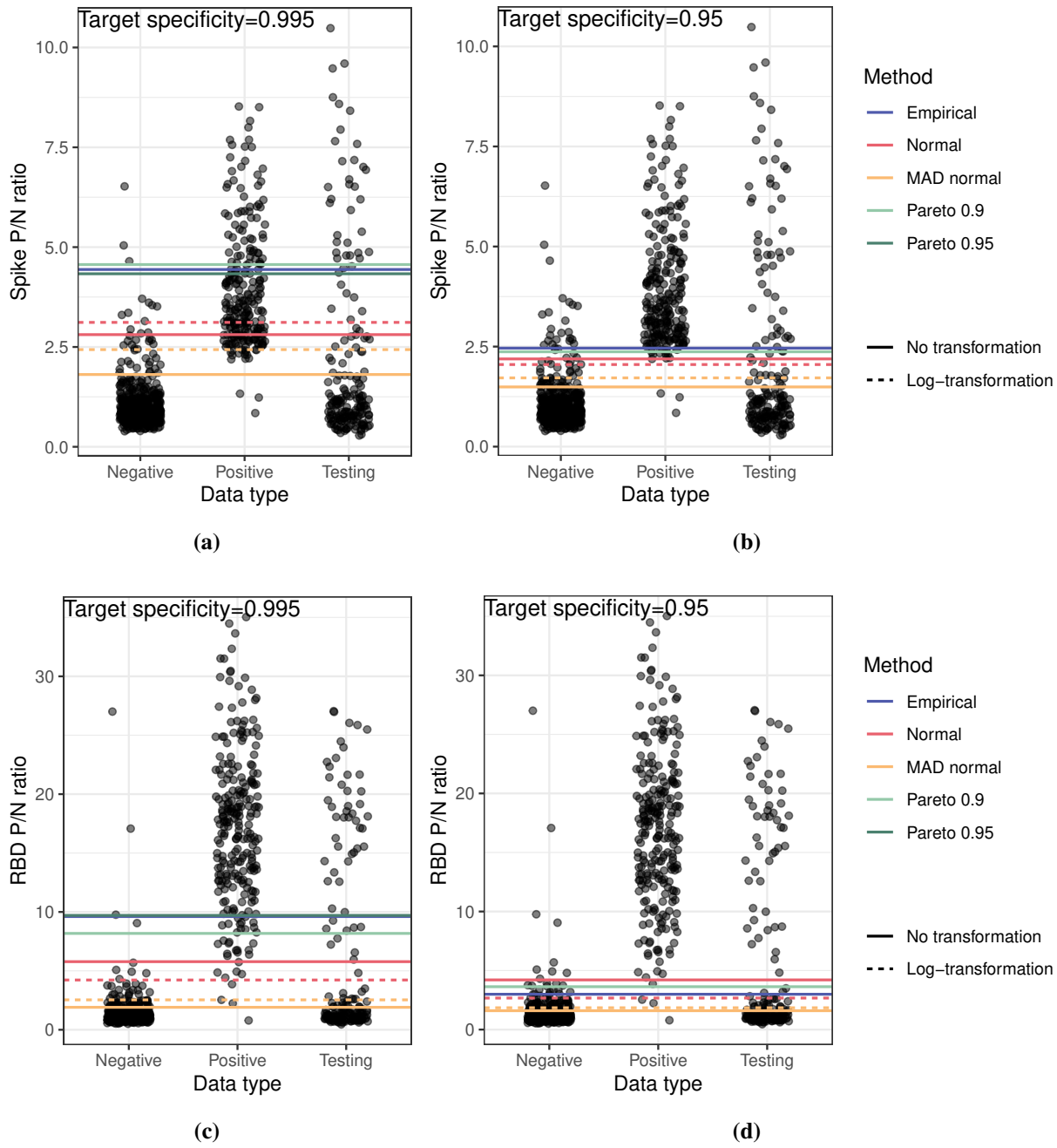


Figure 2.2: P/N ratios for the positive controls, negative controls, and testing data, jittered horizontally. Cutoffs as calculated by each of the seven methods are shown as horizontal lines. The first row shows the spike test cutoffs with (a) a target specificity of 0.995 and (b) a target specificity of 0.95. The second row shows the receptor-binding domain (RBD) test with (c) a specificity of 0.995 and (d) a target specificity of 0.95. Training data was sampled from staff at long-term care facilities in Colorado, USA between June and December 2020. Testing data collected from skilled nursing staff in Colorado during May 2020.

2.3.2 Simulation study

Figures 2.1 (a)-(d) show the distribution functions we generated data from. There was more overlap between the positive and negative cases in the data for scenario A than in scenario B. This is partially a result of the right skew of the positive controls and partially because the tail of the negative controls extends further in scenario A than in scenario B.

Table 2.2: The mean and Monte Carlo standard error in parentheses of the bias and RMSE of the cutoff when targeting a specificity of 0.995. The method(s) with minimal bias and RMSE in each scenario or equivalent after rounding are bolded.

	Scenario A		Scenario B	
	n=50	n=200	n=50	n=200
Bias of cutoff				
Empirical	-0.93 (0.0049)	-0.37 (0.0035)	-3.60 (0.0327)	-1.77 (0.0249)
Normal	-1.74 (0.0027)	-1.70 (0.0014)	-6.26 (0.0158)	-5.71 (0.0112)
Log Normal	-1.34 (0.0030)	-1.36 (0.0015)	-6.72 (0.0070)	-6.79 (0.0032)
MAD	-2.68 (0.0010)	-2.68 (0.0005)	-9.08 (0.0013)	-9.09 (0.0006)
Log MAD	-2.00 (0.0025)	-2.02 (0.0012)	-8.41 (0.0031)	-8.47 (0.0014)
Pareto 0.9	**-0.15 (0.0064)**	**-0.02 (0.0033)**	-3.03 (0.0309)	-2.97 (0.0154)
Pareto 0.95	-0.50 (0.0059)	**-0.02 (0.0035)**	**-2.73 (0.0368)**	**-1.35 (0.0227)**
Hybrid Empirical	-0.93 (0.0048)	-0.38 (0.0036)	-3.59 (0.0327)	-1.77 (0.0249)
Hybrid Pareto 0.9	-0.37 (0.0069)	-0.03 (0.0033)	-3.06 (0.0310)	-2.97 (0.0154)
Hybrid Pareto 0.95	-0.62 (0.0061)	-0.03 (0.0035)	-2.74 (0.0368)	**-1.35 (0.0227)**
RMSE of cutoff				
Empirical	1.35 (0.0039)	0.80 (0.0031)	7.47 (0.0789)	5.28 (0.0360)
Normal	1.82 (0.0024)	1.72 (0.0014)	7.01 (0.0180)	6.14 (0.0090)
Log Normal	1.47 (0.0026)	1.39 (0.0014)	**6.86 (0.0059)**	6.82 (0.0031)
MAD	2.69 (0.0010)	2.68 (0.0005)	9.08 (0.0013)	9.09 (0.0006)
Log MAD	2.07 (0.0022)	2.04 (0.0012)	8.43 (0.0030)	8.48 (0.0014)
Pareto 0.9	**1.28 (0.0059)**	**0.65 (0.0029)**	6.88 (0.0619)	**4.28 (0.0133)**
Pareto 0.95	**1.28 (0.0052)**	0.70 (0.0033)	7.84 (0.0871)	4.74 (0.0294)
Hybrid Empirical	1.34 (0.0037)	0.80 (0.0031)	7.45 (0.0790)	5.28 (0.0360)
Hybrid Pareto 0.9	1.43 (0.0057)	0.66 (0.0030)	6.91 (0.0616)	**4.28 (0.0133)**
Hybrid Pareto 0.95	1.37 (0.0050)	0.71 (0.0033)	7.85 (0.0870)	4.74 (0.0294)

Abbreviations: mean absolute deviation, MAD; root mean squared error, RMSE

Table 2.3: The mean and Monte Carlo standard error in parentheses of the bias and RMSE of the cutoff when targeting a specificity of 0.95. The method(s) with minimal bias and RMSE in each scenario or equivalent after rounding are bolded.

	Scenario A		Scenario B	
	n=50	n=200	n=50	n=200
Bias of cutoff				
Empirical	-0.14 (0.0025)	-0.05 (0.0014)	**-0.02 (0.0041)**	**-0.02 (0.0015)**
Normal	-0.26 (0.0018)	-0.23 (0.0010)	0.65 (0.0105)	1.00 (0.0073)
Log Normal	-0.35 (0.0014)	-0.36 (0.0007)	-0.23 (0.0029)	-0.24 (0.0014)
MAD	-0.92 (0.0007)	-0.92 (0.0004)	-1.29 (0.0009)	-1.29 (0.0004)
Log MAD	-0.68 (0.0012)	-0.68 (0.0006)	-1.03 (0.0015)	-1.06 (0.0007)
Pareto 0.9	**-0.07 (0.0022)**	**-0.03 (0.0012)**	0.65 (0.0077)	0.70 (0.0038)
Hybrid Empirical	-0.17 (0.0024)	-0.05 (0.0014)	-0.03 (0.0041)	**-0.02 (0.0015)**
Hybrid Pareto 0.9	-0.12 (0.0023)	**-0.03 (0.0012)**	0.64 (0.0077)	0.70 (0.0038)
RMSE of cutoff				
Empirical	0.51 (0.0021)	0.29 (0.0012)	0.82 (0.0130)	**0.31 (0.0017)**
Normal	**0.45 (0.0015)**	0.30 (0.0008)	2.20 (0.0341)	1.77 (0.0170)
Log Normal	**0.45 (0.0012)**	0.38 (0.0007)	**0.63 (0.0028)**	0.37 (0.0011)
MAD	0.93 (0.0007)	0.93 (0.0004)	1.30 (0.0009)	1.30 (0.0004)
Log MAD	0.72 (0.0011)	0.69 (0.0006)	1.08 (0.0013)	1.07 (0.0007)
Pareto 0.9	**0.45 (0.0019)**	**0.23 (0.0010)**	1.67 (0.0182)	1.04 (0.0056)
Hybrid Empirical	0.51 (0.0020)	0.29 (0.0012)	0.83 (0.0129)	**0.31 (0.0017)**
Hybrid Pareto 0.9	0.47 (0.0019)	0.24 (0.0010)	1.67 (0.0181)	1.04 (0.0056)

Abbreviations: mean absolute deviation, MAD; root mean squared error, RMSE

Cutoff estimation

Table 2.2 shows the bias and the RMSE of the cutoff for each method when targeting a specificity of 0.995. In the majority of cases, the Pareto methods were superior in terms of bias and RMSE. The only exception was scenario B with a training sample size of 50 where the RMSE was smallest for the lognormal method because the larger bias for this method was offset by the smaller variance.

The cutoff estimates with every method were negatively biased, meaning the cutoff was below the true 0.995 quantile for each method, on average. Thus, the specificity of the estimated cutoff was below the target, on average. The MAD and log MAD methods were the most biased while the Pareto methods were the least biased.

The hybrid methods all had slightly higher RMSE and bias than their corresponding Pareto or empirical methods. Normality and log normality were both rejected for the vast majority of the datasets: 99-100% of datasets with a training sample size of 200 and 59-92% with a training sample size of 50. The results are, therefore, mostly the Pareto and empirical cutoffs but with a small number of poorer performing normal or lognormal cutoffs mixed in.

Table 2.3 shows, when targeting a specificity of 0.95, the magnitude of the bias and RMSE were smaller. The empirical method had the minimal bias under scenario B. The Pareto 0.9 and normal methods had a positive bias for scenario B, compared to the negative bias when targeting a specificity of 0.995.

Prevalence estimation

Table 2.4: The mean and middle 95% (2.5% quantile, 97.5% quantile) of the Rogan-Gladen adjusted prevalence estimates when targeting a specificity of 0.995. The method(s) with the prevalence estimate nearest the truth in each scenario or equivalent after rounding are bolded.

	Scenario A				Scenario B			
	n=50		n=200		n=50		n=200	
Prevalence=0.05								
Empirical	0.07	(0.01, 0.13)	0.06	(0.00, 0.10)	**0.07**	**(0.02, 0.13)**	**0.05**	**(0.03, 0.08)**
Normal	0.09	(0.05, 0.13)	0.09	(0.06, 0.11)	0.08	(0.04, 0.15)	0.06	(0.05, 0.10)
Log Normal	0.08	(0.04, 0.12)	0.08	(0.05, 0.11)	**0.07**	**(0.05, 0.12)**	0.07	(0.05, 0.09)
MAD	0.14	(0.10, 0.20)	0.14	(0.11, 0.17)	0.19	(0.11, 0.26)	0.19	(0.14, 0.23)
Log MAD	0.10	(0.06, 0.15)	0.10	(0.07, 0.13)	0.13	(0.06, 0.22)	0.12	(0.08, 0.17)
Pareto 0.9	**0.06**	**(0.00, 0.14)**	**0.05**	**(0.00, 0.10)**	**0.07**	**(0.02, 0.10)**	**0.05**	**(0.04, 0.07)**
Pareto 0.95	0.07	(0.00, 0.13)	**0.05**	**(0.00, 0.10)**	**0.07**	**(0.01, 0.13)**	**0.05**	**(0.03, 0.07)**
Hybrid Empirical	0.07	(0.01, 0.12)	0.06	(0.00, 0.10)	**0.07**	**(0.02, 0.13)**	**0.05**	**(0.03, 0.08)**
Hybrid Pareto 0.9	0.07	(0.00, 0.14)	**0.05**	**(0.00, 0.10)**	**0.07**	**(0.02, 0.11)**	**0.05**	**(0.04, 0.07)**
Hybrid Pareto 0.95	0.07	(0.00, 0.13)	**0.05**	**(0.00, 0.10)**	0.08	(0.01, 0.13)	**0.05**	**(0.03, 0.07)**
Prevalence=0.30								
Empirical	0.32	(0.19, 0.48)	0.31	(0.21, 0.41)	**0.31**	**(0.24, 0.38)**	**0.30**	**(0.26, 0.34)**
Normal	0.33	(0.25, 0.41)	0.33	(0.28, 0.37)	0.32	(0.28, 0.37)	0.31	(0.29, 0.34)
Log Normal	0.33	(0.23, 0.43)	0.32	(0.26, 0.38)	0.32	(0.29, 0.35)	0.31	(0.30, 0.33)
MAD	0.37	(0.33, 0.42)	0.36	(0.34, 0.39)	0.40	(0.34, 0.46)	0.40	(0.36, 0.44)
Log MAD	0.34	(0.27, 0.40)	0.33	(0.30, 0.36)	0.36	(0.30, 0.43)	0.35	(0.32, 0.39)
Pareto 0.9	**0.31**	**(0.09, 0.63)**	**0.30**	**(0.20, 0.42)**	**0.31**	**(0.23, 0.37)**	**0.30**	**(0.27, 0.33)**
Pareto 0.95	0.32	(0.14, 0.55)	**0.30**	**(0.19, 0.43)**	**0.31**	**(0.22, 0.39)**	**0.30**	**(0.26, 0.34)**
Hybrid Empirical	0.32	(0.20, 0.48)	0.31	(0.21, 0.41)	**0.31**	**(0.24, 0.37)**	**0.30**	**(0.26, 0.34)**
Hybrid Pareto 0.9	**0.31**	**(0.09, 0.60)**	**0.30**	**(0.20, 0.42)**	**0.31**	**(0.23, 0.37)**	**0.30**	**(0.27, 0.33)**
Hybrid Pareto 0.95	0.32	(0.15, 0.54)	**0.30**	**(0.19, 0.43)**	**0.31**	**(0.22, 0.39)**	**0.30**	**(0.26, 0.34)**

Abbreviations: mean absolute deviation, MAD

Table 2.4 shows simulation results for the Rogan-Gladen adjusted prevalence estimates when targeting a specificity of 0.995. The Pareto cutoffs had little bias but had larger variability when targeting a specificity of 0.995. In every case, the average of the prevalence point estimates was closest to the truth using one of the Pareto methods. However, in scenario A the Pareto estimates, especially with a sample size of 50, were more variable than the normal-based methods. Table 2.5 shows the prevalence when targeting a specificity of 0.95. The variability for the Pareto and empirical methods were lower when targeting a lower specificity, and particularly at the smaller sample size. With both target specificities, the MAD and log MAD methods were positively biased,

Table 2.5: The mean and middle 95% (2.5% quantile, 97.5% quantile) of the Rogan-Gladen adjusted prevalence estimates when targeting a specificity of 0.95. The method(s) with the prevalence estimate nearest the truth in each scenario or equivalent after rounding are bolded.

	Scenario A				Scenario B			
	n=50		n=200		n=50		n=200	
Prevalence=0.05								
Empirical	**0.06**	**(0.00, 0.15)**	**0.05**	**(0.01, 0.10)**	0.07	(0.01, 0.15)	**0.05**	**(0.02, 0.10)**
Normal	0.07	(0.01, 0.14)	0.07	(0.03, 0.10)	0.06	(0.00, 0.16)	0.04	(0.00, 0.10)
Log Normal	0.08	(0.03, 0.13)	0.07	(0.04, 0.11)	0.08	(0.02, 0.16)	0.07	(0.03, 0.12)
MAD	0.15	(0.09, 0.25)	0.15	(0.10, 0.20)	0.21	(0.12, 0.29)	0.20	(0.15, 0.26)
Log MAD	0.11	(0.05, 0.20)	0.11	(0.07, 0.15)	0.16	(0.06, 0.26)	0.16	(0.11, 0.22)
Pareto 0.9	**0.06**	**(0.00, 0.13)**	**0.05**	**(0.02, 0.09)**	**0.05**	**(0.00, 0.14)**	0.03	(0.01, 0.08)
Hybrid Empirical	**0.06**	**(0.00, 0.14)**	**0.05**	**(0.01, 0.10)**	0.07	(0.01, 0.15)	**0.05**	**(0.02, 0.10)**
Hybrid Pareto 0.9	**0.06**	**(0.00, 0.13)**	**0.05**	**(0.02, 0.09)**	**0.05**	**(0.00, 0.14)**	0.03	(0.01, 0.08)
Prevalence=0.30								
Empirical	**0.31**	**(0.23, 0.38)**	**0.30**	**(0.26, 0.34)**	0.31	(0.27, 0.37)	**0.30**	**(0.28, 0.34)**
Normal	0.32	(0.26, 0.37)	0.31	(0.28, 0.34)	0.31	(0.25, 0.38)	0.29	(0.26, 0.34)
Log Normal	0.32	(0.27, 0.37)	0.32	(0.29, 0.35)	0.32	(0.27, 0.39)	0.31	(0.28, 0.35)
MAD	0.38	(0.32, 0.45)	0.37	(0.34, 0.41)	0.41	(0.35, 0.48)	0.41	(0.37, 0.46)
Log MAD	0.35	(0.30, 0.41)	0.34	(0.31, 0.38)	0.38	(0.31, 0.45)	0.38	(0.34, 0.43)
Pareto 0.9	**0.31**	**(0.23, 0.36)**	**0.30**	**(0.26, 0.33)**	**0.30**	**(0.26, 0.36)**	0.29	(0.26, 0.32)
Hybrid Empirical	**0.31**	**(0.23, 0.37)**	**0.30**	**(0.26, 0.34)**	0.31	(0.27, 0.38)	**0.30**	**(0.28, 0.34)**
Hybrid Pareto 0.9	**0.31**	**(0.23, 0.37)**	**0.30**	**(0.26, 0.33)**	**0.30**	**(0.26, 0.37)**	0.29	(0.26, 0.32)

Abbreviations: mean absolute deviation, MAD

while the other methods had a smaller bias, generally positive. The hybrid method estimates were again similar to the corresponding empirical and Pareto estimates.

Test accuracy

We consider the accuracy of the cutoff estimation methods for classifying individuals as positive or negative in the testing data. Tables 2.6 and 2.7 show the proportion of testing set observations correctly classified with a target specificity of 0.995 and 0.95, respectively. The MAD methods' cutoffs were negatively biased, leading to a lower specificity and decreased accuracy in low prevalence scenarios. The Pareto methods had the highest accuracy (or equivalent to the highest accuracy) when prevalence was 0.05.

When the prevalence was higher at 0.3 and using the lower target specificity, the Pareto method was most accurate in scenario B. All but the MAD methods performed similarly for scenario A. With the higher target specificity, the MAD cutoffs had highest accuracy for scenario A, and the lognormal method was most accurate for scenario B.

Table 2.6: The mean and middle 95% (2.5% quantile, 97.5% quantile) of the accuracy of the test as measured by the proportion of testing dataset observations correctly predicted when targeting a specificity of 0.995. The method(s) with highest accuracy in each scenario or equivalent after rounding are bolded.

	Scenario A				Scenario B			
	n=50		n=200		n=50		n=200	
Prevalence=0.05								
Empirical	**0.96**	**(0.93, 0.97)**	**0.96**	**(0.95, 0.97)**	0.97	(0.93, 0.99)	0.98	(0.96, 0.99)
Normal	0.95	(0.91, 0.97)	**0.96**	**(0.93, 0.97)**	0.97	(0.90, 0.99)	0.98	(0.95, 0.99)
Log Normal	**0.96**	**(0.93, 0.97)**	**0.96**	**(0.94, 0.97)**	0.97	(0.93, 0.99)	0.98	(0.96, 0.99)
MAD	0.90	(0.84, 0.95)	0.91	(0.87, 0.94)	0.86	(0.78, 0.93)	0.86	(0.81, 0.90)
Log MAD	0.94	(0.89, 0.97)	0.95	(0.92, 0.97)	0.92	(0.83, 0.98)	0.93	(0.87, 0.97)
Pareto 0.9	**0.96**	**(0.94, 0.97)**	**0.96**	**(0.95, 0.97)**	**0.98**	**(0.95, 0.99)**	**0.99**	**(0.97, 1.00)**
Pareto 0.95	**0.96**	**(0.93, 0.97)**	**0.96**	**(0.95, 0.97)**	**0.98**	**(0.94, 0.99)**	0.98	(0.96, 0.99)
Hybrid Empirical	**0.96**	**(0.93, 0.97)**	**0.96**	**(0.95, 0.97)**	0.97	(0.93, 0.99)	0.98	(0.96, 0.99)
Hybrid Pareto 0.9	**0.96**	**(0.93, 0.97)**	**0.96**	**(0.95, 0.97)**	**0.98**	**(0.94, 0.99)**	**0.99**	**(0.97, 1.00)**
Hybrid Pareto 0.95	**0.96**	**(0.93, 0.97)**	**0.96**	**(0.95, 0.97)**	**0.98**	**(0.94, 0.99)**	0.98	(0.96, 0.99)
Prevalence=0.30								
Empirical	0.85	(0.73, 0.95)	0.81	(0.73, 0.89)	0.95	(0.74, 0.99)	0.94	(0.77, 0.99)
Normal	0.91	(0.81, 0.96)	0.91	(0.84, 0.96)	0.96	(0.90, 0.99)	0.97	(0.94, 0.99)
Log Normal	0.88	(0.78, 0.96)	0.87	(0.81, 0.94)	**0.97**	**(0.94, 0.99)**	**0.98**	**(0.96, 0.99)**
MAD	**0.93**	**(0.88, 0.96)**	0.93	(0.90, 0.95)	0.89	(0.84, 0.95)	0.90	(0.86, 0.93)
Log MAD	0.92	(0.82, 0.96)	**0.94**	**(0.88, 0.96)**	0.94	(0.87, 0.98)	0.94	(0.90, 0.98)
Pareto 0.9	0.81	(0.71, 0.95)	0.79	(0.73, 0.86)	0.95	(0.74, 0.99)	0.96	(0.87, 0.99)
Pareto 0.95	0.82	(0.71, 0.95)	0.79	(0.73, 0.86)	0.94	(0.72, 0.99)	0.94	(0.78, 0.99)
Hybrid Empirical	0.85	(0.73, 0.95)	0.81	(0.73, 0.89)	0.95	(0.74, 0.99)	0.94	(0.77, 0.99)
Hybrid Pareto 0.9	0.82	(0.71, 0.95)	0.79	(0.73, 0.86)	0.95	(0.74, 0.99)	0.96	(0.87, 0.99)
Hybrid Pareto 0.95	0.83	(0.71, 0.95)	0.79	(0.73, 0.86)	0.94	(0.72, 0.99)	0.94	(0.78, 0.99)

Abbreviations: mean absolute deviation, MAD

Table 2.7: The mean and middle 95% (2.5% quantile, 97.5% quantile) of the accuracy of the test as measured by the proportion of testing dataset observations correctly predicted when targeting a specificity of 0.95. The method(s) with highest accuracy in each scenario or equivalent after rounding are bolded.

	Scenario A				Scenario B			
	n=50		n=200		n=50		n=200	
Prevalence=0.05								
Empirical	0.93	(0.86, 0.97)	0.94	(0.91, 0.97)	0.94	(0.86, 0.99)	0.95	(0.91, 0.98)
Normal	0.93	(0.87, 0.97)	0.94	(0.90, 0.96)	0.94	(0.85, 0.99)	0.96	(0.90, 0.99)
Log Normal	0.93	(0.87, 0.96)	0.93	(0.90, 0.96)	0.93	(0.84, 0.98)	0.93	(0.89, 0.97)
MAD	0.85	(0.76, 0.92)	0.86	(0.81, 0.90)	0.80	(0.72, 0.88)	0.81	(0.76, 0.85)
Log MAD	0.89	(0.81, 0.95)	0.90	(0.86, 0.94)	0.85	(0.75, 0.94)	0.85	(0.79, 0.90)
Pareto 0.9	**0.94**	**(0.88, 0.97)**	**0.95**	**(0.91, 0.97)**	**0.95**	**(0.87, 0.99)**	**0.97**	**(0.93, 0.99)**
Hybrid Empirical	0.93	(0.86, 0.97)	0.94	(0.91, 0.97)	0.94	(0.85, 0.99)	0.95	(0.91, 0.98)
Hybrid Pareto 0.9	**0.94**	**(0.88, 0.97)**	**0.95**	**(0.91, 0.97)**	**0.95**	**(0.86, 0.99)**	**0.97**	**(0.93, 0.99)**
Prevalence=0.30								
Empirical	0.93	(0.84, 0.96)	**0.94**	**(0.88, 0.96)**	0.95	(0.89, 0.99)	0.96	(0.93, 0.98)
Normal	0.93	(0.88, 0.96)	**0.94**	**(0.92, 0.96)**	0.95	(0.88, 0.99)	**0.97**	**(0.92, 0.99)**
Log Normal	**0.94**	**(0.90, 0.96)**	**0.94**	**(0.92, 0.96)**	0.94	(0.88, 0.98)	0.95	(0.91, 0.98)
MAD	0.89	(0.82, 0.94)	0.89	(0.86, 0.93)	0.86	(0.79, 0.91)	0.86	(0.82, 0.89)
Log MAD	0.92	(0.86, 0.96)	0.92	(0.89, 0.95)	0.89	(0.82, 0.95)	0.89	(0.84, 0.93)
Pareto 0.9	0.93	(0.85, 0.96)	**0.94**	**(0.90, 0.96)**	**0.96**	**(0.90, 0.99)**	**0.97**	**(0.94, 0.99)**
Hybrid Empirical	0.93	(0.84, 0.96)	**0.94**	**(0.88, 0.96)**	0.95	(0.89, 0.99)	0.96	(0.93, 0.98)
Hybrid Pareto 0.9	0.93	(0.85, 0.96)	**0.94**	**(0.90, 0.96)**	**0.96**	**(0.90, 0.99)**	**0.97**	**(0.94, 0.99)**

Abbreviations: mean absolute deviation, MAD

2.4 Discussion

It is imperative to rapidly develop and deploy prognostic tests for emerging infectious diseases that can be used to classify individuals and estimate prevalence in a community. A common challenge for tests is determining a cutoff value to separate positive and negative cases as there is often overlap in the results between the positive and negative cases. This is especially challenging with early tests for emerging diseases for which there is limited training data with validated positive and negative controls. Common approaches to estimating cutoff values are using the quantile of a parametric distribution fit to the negative control test data or using the empirical quantile of the negative control test data. Yet, there is little guidance on how to select a cutoff to separate positive and negative results, especially for small data sets. Here, we proposed using methods from extreme value theory, specifically using the generalized Pareto distribution to estimate the upper tail of the negative control training data and its quantiles, to estimate a cutoff value to achieve a target specificity. We compared the proposed approach and common alternatives in a simulation study.

Our simulation demonstrated that when targeting a very high specificity, 0.995 as recommended by the CDC early in the COVID-19 pandemic (Centers for Disease Control and Prevention, 2020), the Pareto methods proposed had lower bias and RMSE for estimating a cutoff value. When targeting a lower target specificity of 0.95, the empirical method consistently performed well. Methods that relied on parametric distributions (e.g., normal, lognormal, MAD normal and MAD lognormal) generally had large bias and RMSE. As a result, the Pareto methods for the higher target specificity and the empirical methods for the lower target specificity, estimated cutoffs closest to the target specificity of any methods considered.

Estimating a cutoff that results in the desired target specificity is essential for both estimating prevalence and classifying individuals. The methods showing minimal bias in cutoff estimation also had the least biased prevalence estimates.

In addition to the cutoff estimation method, a second factor that influences cutoff estimation accuracy is the target specificity. We compared the recommended target specificity of 0.995 to a target specificity of 0.95 and found the desired target specificity varied according to the goal

of the analysis as well as the prevalence of the population. In the low prevalence, as setting we might expect for an emerging disease, using a higher target cutoff of 0.995, as compared to the more moderate 0.95, resulted in better accuracy for classifying individuals as positive or negative (Tables 2.6 and 2.7). When prevalence is low, specificity affects the accuracy more than sensitivity. However, in a higher-prevalence setting, the need for a large specificity is balanced with the need for a large sensitivity as there are more true positives that could potentially be misclassified. Thus, with higher prevalence, accuracy was overall higher when targeting a specificity of 0.95 instead of 0.995. The slightly lowered target specificity allowed for much higher sensitivities. When targeting a specificity of 0.995, the median sensitivities ranged from 0.31 to 0.99 across the various methods and simulation scenarios. When targeting a specificity of 0.95, the median sensitivities ranged significantly less, from 0.94 to $>$0.99 (Appendix A.4).

We also found the target specificity impacted uncertainty in prevalence estimation. The average prevalence estimate was near the true value for the empirical and Pareto methods with either target specificity, but the variability of the prevalence estimate was generally lower for these methods when targeting a specificity of 0.95. The standard error was over two times larger for the Pareto methods when using a target specificity of 0.995 compared to 0.95 for the smaller training sample size of 50: the scenario where the Pareto methods was the least biased for prevalence estimates. The empirical standard error was also larger when using a target specificity of 0.995, but to a lesser degree for scenario A. Therefore, if the goal is estimating the population prevalence, we recommend a lower target specificity as this reduces the variability of the cutoffs and in turn the variability of the prevalence estimates.

The results of our data analysis of two COVID-19 antibody tests are consistent with the results of the simulation study. The Pareto and empirical methods, which showed minimal negative bias in the simulation study, also tended to have the highest cutoff estimates in the data analysis. The MAD methods showed considerable negative bias in the simulation study and had the smallest estimates in the data analysis. Additionally, like the simulation study, the prevalence estimates showed more variability when targeting a specificity of 0.995 rather than 0.95.

The performance of the cutoff estimators and the resulting accuracy at the individual level and prevalence estimators at the community-levels will vary depending on the shape of the distributions of positive and negative results and the separation between those two distributions. The shape of the distribution impacts how accurately the target specificity can be estimated for the methods using parametric assumptions. The separation of the distributions impacts accuracy, sensitivity, and prevalence estimates. If the distributions show considerable overlap, the accuracy is lowered, and a cutoff cannot be selected that results in both a highly sensitive and highly specific test. We only generated data from two possible distributions and two possible sample sizes, so the results of our simulation study should be limited to this context.

The normal methods rely on stricter distributional assumptions than the Pareto or empirical. In our simulation, the assumptions of the normal methods were not met. Even in the hybrid methods when the normal and lognormal methods were only used if we failed to reject normality, the bias and RMSE were still larger than if we only used the empirical or Pareto methods. In contrast, the empirical method makes no distributional assumptions. The Pareto methods rely on general regularity conditions about the negative controls distribution (Pickands III, 1975; Balkema and De Haan, 1974).

Because an emerging disease has potential cross-reactivity and few true positives expected, we focus on methods for establishing cutoffs that target a high specificity (Takahashi et al., 2020). However, in other applications, approaches that consider both the sensitivity and specificity, as well as the relative costs of false positive and false negative results and the prevalence, may be preferred (Felder and Mayrhofer, 2022; Greiner et al., 2000; Hajian-Tilaki, 2018; Linnet and Brandt, 1986). When only estimating prevalence, some methods forgo establishing a cutoff and instead fit a mixture model (Bottomley et al., 2021; Bouman et al., 2021; Hitchings et al., 2023; Schaarschmidt et al., 2015; Vink et al., 2015) or a latent class model (Kostoulas et al., 2021; Laurin et al., 2019; Symons et al., 2021) to the continuous test results.

Based on our simulation and data analysis, we recommend using the Pareto methods or the empirical method to estimate the cutoff when developing tests, depending on the target specificity.

When targeting a high specificity such as 0.995, we demonstrated the Pareto method was least biased in estimating the cutoff with the desired specificity, as well as showed minimal bias in estimating the population prevalence. When targeting a lower specificity, the empirical method performed consistently well. The commonly used normal methods showed a larger negative bias than either of these methods, and the MAD normal methods showed a considerable bias in our simulation.

The choice of target specificity of the cutoff should account for the goals of the test. If the goal is to have accurate tests at the individual level, a high target specificity of 0.995 was ideal in our simulation when in the low prevalence case we would expect of an emerging disease. However, if the goal is instead to use the antibody test to estimate the population prevalence, a lower target specificity may be preferred to reduce variability in the cutoff and thus in the prevalence estimate.

Chapter 3

Data integration via Bayesian modeling for estimation of COVID-19 infection fatality rates

3.1 Introduction

COVID-19 severely disrupted daily life globally in early 2020 with at least 643.89 million infections and 6.64 million deaths worldwide as of December 2022 (Jha et al., 2022; Our World in Data, 2022). However, availability of location-specific data on the number of COVID-19 infections and fatalities is inconsistent. Some locations have age-specific records of both infections and fatalities, while others report these for varying numbers of age bins or lack such data entirely. We developed a statistical model to formally assess the location-specific burden of COVID-19 in terms of infection rates and fatality rates. We perform data integration to use all the available data sources and get age bin-level estimates rather than population estimates, leveraging all granularity provided by the data. The proposed model is an advance over existing methods because it fully accounts for uncertainty in infection and fatality rates. Using this model, we estimate seroprevalence and infection fatality rates for 107 locations in developing countries as well as the associated uncertainty, as detailed in Levin et al. (2022). In this chapter we provide the technical details of our model. We compare our model for infection and fatality rates to previous methods, highlighting the importance of fully accounting for uncertainty in the infections and the fatalities data. Our results can be used to facilitate data-driven, age-specific decisions regarding distribution of vaccines, treatments, and international aid, as well as interventions such as social distancing measures.

We use the proportion of infections that result in a death to compare locations. This can be estimated as the number of COVID-19 deaths divided by the reported number of COVID-19 infections–the case fatality rate (Alimohamadi et al., 2021). However, the number of reported cases is likely a drastic under-count of the total number of COVID-19 infections (National Academies

of Sciences, Engineering, and Medicine, 2020). At the start of the pandemic tests were not available (Vandenberg et al., 2021). Once tests were developed, asymptomatic cases were still likely to go unreported as did many symptomatic cases (Lau et al., 2021). Thus, instead of relying on the reported number of cases, we estimate the number of infections to infer the infection fatality rate (IFR): the number of COVID-19 deaths divided by the estimated number of COVID-19 infections.

In our analysis, we use serology survey data to estimate the number of COVID-19 infections. In serology studies, a sample of the population is selected, and each study participant is tested for COVID-19 antibodies. People who have been infected with SARS-CoV-2 build antibodies against the virus (i.e., seroconvert) and should test positive. The proportion of the serology study sample that tests positive for COVID-19 antibodies is then used to estimate the proportion of the population that has antibodies–the seroprevalence, which is used as an estimate for the proportion of the population that has experienced a COVID-19 infection.

In estimating the seroprevalence from a serology study, it is essential to account for the false positive and false negative rates of the antibody tests. As an example, consider a test with 100% sensitivity (i.e., the probability an individual with COVID-19 antibodies tests positive is one) in a location with a low seroprevalence of 5%, as we might expect early in a pandemic (Table 3.1). If the test had perfect specificity (i.e., the probability an individual without COVID-19 antibodies tests negative is one), we would expect none of our positive test results to be false positives, and the positivity rate would be an unbiased estimator for the seroprevalence. However, if the specificity was lowered to just 0.95, then the expected false discovery rate jumps to 49%, meaning we would expect almost half of the cases we identify as positive to be false positives. Thus, our estimated seroprevalence should be about half of the positivity rate.

Table 3.1: The expected percent of positive test results that are false positives for various specificity values, assuming perfect sensitivity and a 0.05 prevalence.

Specificity	False discovery rate
1.000	0%
0.975	32%
0.950	49%

Because the COVID-19 antibody tests used in serology studies during the first year of the pandemic were rapidly developed with limited lab validation, it is also essential to account for the *uncertainty* of the false positive and false negative rates (Gelman and Carpenter, 2020; Larremore et al., 2022). Early test developers evaluated the sensitivity and specificity for each assay by testing positive and negative controls. They ran their COVID-19 antibody tests on blood samples from individuals either known to have been infected with SARS-CoV-2 (positive controls) or those known not to have been infected (negative controls), and the number of correct test results were recorded. Sometimes many controls were used (1000 or more), while in other cases very few were used (around 30). For the tests evaluated with few controls, the confidence interval for the test characteristics can be wide enough that the end points result in drastically different seroprevalence estimates. In our dataset, the Luminex S protein trimer IgG assay had the fewest negative controls with just 31, and the 95% confidence interval for the specificity is (0.888, 1). Extending our prior example, assume a sensitivity of 100%, a true prevalence of 0.05, and assume that of 100 individuals tested, 10 had a positive test result. The false discovery rate is

$$\text{FDR} = \frac{P(\text{false positive})}{P(\text{test positive})} = \frac{(1-\pi)(1-\text{spec})}{(1-\pi)(1-\text{spec}) + \pi\text{sens}}, \tag{3.1}$$

for a sensitivity of sens, specificity of spec, and true prevalence of π. First considering a specificity of 0.888, we would expect 68% of positive tests to be false positives, i.e., about 7 of those that were tested. The prevalence could then be estimated as the 3 true positives over the 100 total for an estimated prevalence of 0.03. If the specificity were instead 1, we would expect all of the positive tests to be true positives, resulting in a prevalence estimate of 0.1.

Many of the previous approaches for estimating IFR using seroprevalence studies have accounted for the test characteristics in a two-stepped approach. First, they estimate the seroprevalence and uncertainty, accounting for the test characteristics, then they use the seroprevalence estimates to estimate IFR. One of the common approaches is to use the Rogan-Gladen adjusted confidence interval to estimate seroprevalence (Rogan and Gladen, 1978), which relies on the sensitivity and specificity point estimates (e.g., Axfors and Ioannidis (2022); Levin et al. (2020)).

However, this approach does not account for any additional uncertainty due to the test characteristics being unknown. To account for the test characteristic uncertainty, some such as Perez-Saez et al. (2021) and Tunheim et al. (2022) followed Gelman and Carpenter (2020) to simultaneously estimate the sensitivity, specificity, and seroprevalence via a Bayesian model to fully account for the test characteristics as well as their uncertainty. They then used these results and associated uncertainty to estimate the IFR. By separating estimation into two steps, any information from the fatality data is not accounted for in estimating the seroprevalence and test characteristics.

When estimating the IFR from many previous studies, obtaining sufficient lab validation and serology study data for each study can be challenging, so some previous meta-analyses, such as Campbell and Gustafson (2021) and Pezzullo et al. (2023), relied on the seroprevalence estimates and confidence intervals reported in the previous studies. Pezzullo et al. (2023) directly utilized the reported confidence intervals that accounted for test characteristics, while Campbell and Gustafson (2021) inverted confidence intervals of the seroprevalence as reported in the serology studies to back out the correlated sample size and number testing positive. They then used these in a Bayesian model to estimate the IFR. However, the serology studies included in the meta-analyses used different methods to account for test characteristics, with some accounting for test characteristic uncertainty and others not. This means the IFR estimates inconsistently accounted for test characteristic uncertainty.

Improving on these methods, we present a Bayesian model which performs data integration to simultaneously estimate the test characteristics, seroprevalence, and IFR in one cohesive model, fully accounting for the uncertainty in the test characteristics. We highlight the benefits of our cohesive model by comparing our results to the commonly used Rogan-Gladen adjustment (Rogan and Gladen, 1978) and the Rogan-Gladen based IFR estimate. We estimate IFR and seroprevalence for each age bin the data was reported at, which can harmonize across data reported at any data granularity and has the additional benefit of avoiding imposing assumptions about how age bins relate to each other. We simultaneously model data, regardless fatality data availability, estimating seroprevalence for all locations and IFR for locations with fatality data.

3.2 Data

We analyze the data published in Levin et al. (2022), which consists of 107 non-overlapping locations from 44 developing countries, with developing countries defined as low- and middle-income countries as designated by the International Monetary Fund (International Monetary Fund, 2021). These locations had varying amounts of data as shown in Table 3.2: 63 had COVID-19 fatality data as well as seroprevalence study data, while the remaining 44 locations only reported seroprevalence study data, with no corresponding fatality data available. Of the studies with fatality data, 28 had age-specific fatality data reported for up to eight age bins, and 35 only reported the total number of deaths across all ages. Nine of the seroprevalence only locations did not report age-specific data, while the other 35 reported data for between two and ten age bins. In addition to the seroprevalence and fatality data, we modeled the lab validation data for each of the tests used in the seroprevalence studies.

Table 3.2: Number of locations with and without death data for each country as well as the range of the number of seroprevalence age bins and death bins for each location within a country.

Region	Country	# of locations	# with death data	# without death data	# sero bins	# death bins
Latin America	Argentina	3	2	1	1–7	4–7
Latin America	Bolivia	1	1	–	1	1
Latin America	Brazil	9	9	–	1–8	1–8
Latin America	Chile	1	1	–	5	5
Latin America	Colombia	10	10	–	5–8	5–7
Latin America	Dominican Republic	1	–	1	1	–
Latin America	Ecuador	1	1	–	5	5
Latin America	Mexico	1	1	–	4	1
Latin America	Paraguay	1	1	–	6	6
Latin America	Peru	4	4	–	1–8	1–8
Africa	Cameroon	1	–	1	5	–
Africa	Dem. Republic of Congo	1	–	1	3	–
Africa	Ethiopia	3	2	1	1–4	1
Africa	Kenya	1	1	–	7	5
Africa	Libya	1	–	1	1	–
Africa	Mozambique	12	1	11	4	1
Africa	Nigeria	1	–	1	7	–
Africa	Senegal	1	1	–	6	1
Africa	South Africa	5	2	3	1–6	1
Africa	South Sudan	1	–	1	8	–
Africa	Zambia	1	1	–	1	1

Africa	Zimbabwe	1	–	1	9	–
Europe	Albania	1	–	1	1	–
Europe	Bosnia and Herzegovina	1	1	–	5	1
Europe	Georgia	1	–	1	5	–
Europe	Hungary	1	1	–	3	3
Europe	Poland	1	1	–	8	1
Europe	Russia	1	1	–	4	1
Middle East	Iran	1	1	–	4	1
Middle East	Iraq	1	–	1	1	–
Middle East	Jordan	1	1	–	8	4
Middle East	Oman	1	1	–	4	1
Middle East	Palestine	2	–	2	8	–
Middle East	United Arab Emirates	1	–	1	8	–
Middle East	Yemen	1	–	1	6	–
East Asia	China	3	2	1	5	1–4
East Asia	Laos	1	–	1	4	–
East Asia	Malaysia	1	–	1	2	–
East Asia	Mongolia	1	–	1	10	–
East Asia	Thailand	1	–	1	3	–
South Asia	Bangladesh	1	–	1	8	–
South Asia	India	20	13	7	1–8	1–8
South Asia	Nepal	1	1	–	10	1
South Asia	Pakistan	3	2	1	1–4	1

3.2.1 Inclusion criteria

Levin et al. (2022) searched for COVID-19 seroprevalence studies publicly disseminated by December 17th, 2021. Each study was then evaluated in terms of sampling bias and only studies deemed to have representative samples were included in the analysis.

The analysis was restricted to seroprevalence studies conducted before April 2021. By only considering studies within the first year of the pandemic, we can reduce the impact of seroreversion, people losing antibodies against COVID-19 over time and therefore no longer testing positive with antibody tests. Additionally, vaccines were not widely distributed before March of 2021, so we can reasonably assume that individuals in these studies only developed COVID-19 antibodies as a result of a prior infection and not from a vaccine.

The timing of serology studies was also considered in relation to a “wave" of COVID-19. If the number infections was dramatically increasing in a condensed period of time (defined as the number of reported COVID-19 deaths increasing by a factor of at least three between the

midpoint of the study and four weeks past the midpoint), it is nearly impossible to match the number of COVID-19 fatalities to the cumulative infections due to the delay between infection, seroconversion, a potentially fatal outcome, and reporting of deaths. Therefore, we opted to use only studies that were not conducted during an accelerating outbreak.

3.2.2 Data structure

Seroprevalence data

The seroprevalence study data consisted of the number of study participants tested and the number of the participants that then tested positive on the COVID-19 antibody test. These counts were either reported for specific age bins (e.g., $(0, 20], (20, 40], (40, 60], \ldots$) or for the study as a whole, representing the average for all ages. The number of participants in an age bin ranged from 8 to 12,897, with a median sample size of 371. The seroprevalence studies in our analysis were conducted between April of 2020 and March of 2021 with 80% of the studies conducted between mid-June and December of 2020.

Death data

For each study with fatality data, researchers collected the cumulative number of COVID-19 deaths up to fourteen days past the midpoint of the corresponding serology study. The fourteen-day lag was introduced to account for the timing between infection and a potentially fatal outcome. For data gathered from official reports, rather than case data, an additional fourteen days was added to account for the lag in death reporting (see Levin et al. (2022)). Because of the heterogeneity in study date, population of the study area, and COVID-19 risk by age, the number of COVID-19 deaths varied dramatically between locations, as well as across age bins within a location. The cumulative number of deaths in an age bin ranged from 0 to 151,910.

Test characteristic data

We analyzed lab validation data for each test assay used in the seroprevalence studies. In total, 47 assays were used with 13 assays used in multiple studies. The number of positive and negative

controls tested for each assay is shown in Figure 3.1. The number of positive controls tested ranged from 29 to 1000 and the number of negative controls tested ranged from 31 to 5991. Twelve assays had less than 100 negative controls, two of which were used in four or more seroprevalence studies. For most assays, more negative controls were used than positive controls.

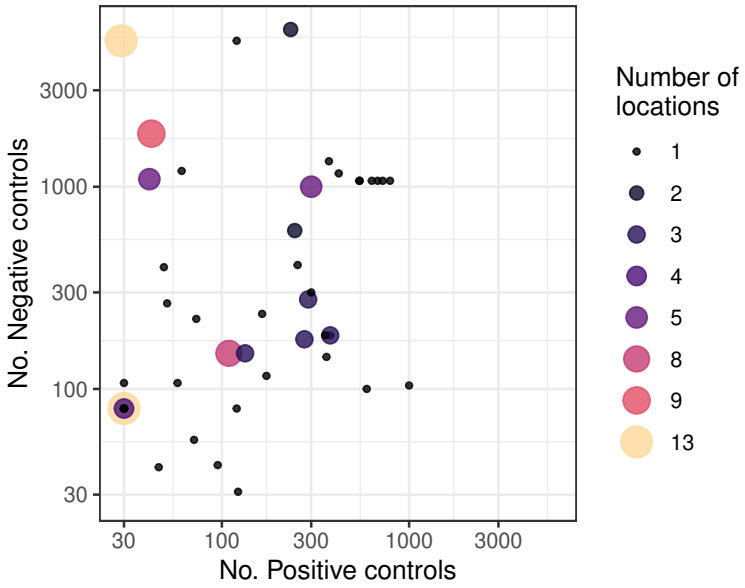


Figure 3.1: Number of positive controls and number of negative controls for each test assay used. The size and color of the point indicates the number of locations using each assay.

Population age distribution data

Population data was collected for various age bins for each study location, primarily from websites such as populationstat.com, worldpopulationreview.com, and populationpyramid.net as well as from census data, when available. When age bins were wider than five years, we combined the location-specific data with national population data from similar sources to estimate the population for five-year age intervals. Details are given in Appendix B.1.

3.3 Methods

We present a cohesive Bayesian model, which simultaneously models fatality, serology, and test assay lab data to allow for age-specific and location-specific seroprevalence and IFR inference. In Section 3.3.1 we present a model for the seroprevalence and test characteristic, which applies to all locations. Then, in Section 3.3.2 we present a model for the fatality data, which only applies to locations with fatality data. Thus, the model simultaneously estimates seroprevalence and test characteristics for all locations, while also estimating the IFR for those with fatality data. We discuss handling of misalignment between seroprevalence and fatality age bins in Section 3.3.3 and conclude by discussing priors is Section 3.3.4.

3.3.1 Modeling seroprevalence data

We model the number of study participants that test positive in age bin A and location ℓ, $R^{\star}_{\ell,A}$, as following a binomial distribution

$$R^{\star}_{\ell,A} \sim \text{Bin}(n_{\ell,A}, p_{\ell,A}), \tag{3.2}$$

where $n_{\ell,A}$ denotes the serology study sample size. Therefore, we assume tests to be independent, Bernoulli trials, with the probability an individual tests positive equal to $p_{\ell,A}$. For locations with multiple age bins, A represents the interval containing the ages of a particular bin. For locations without age-specific data, $A = [0, 100+)$ represents all ages.

Test characteristics

Because the seroprevalence tests used are imperfect, the test positivity rate differs from the seroprevalence rate. The test positivity rate includes individuals with correct positive results and false positive results, whereas the seroprevalence rate includes individuals with correct positive results and false negative results. Thus, the test positivity rate, $p_{\ell,A}$, can be expressed as a function of the test characteristics and the seroprevalence. Let $\pi_{\ell,A}$ be the seroprevalence for age bin A at

location ℓ, one of our primary parameters of interest. Then

$$p_{\ell,A} = \text{sens}_{t_\ell}\pi_{\ell,A} + (1 - \text{spec}_{t_\ell})(1 - \pi_{\ell,A}), \tag{3.3}$$

where t_ℓ denotes the test assay used at location ℓ, sens_{t_ℓ} is the sensitivity, and spec_{t_ℓ} is the specificity of the test. Thus, $\text{sens}_{t_\ell}\pi_{\ell,A}$ represents the probability of a true positive: the probability someone tests positive given they are seropositive times the probability they are seropositive. Similarly, the probability of a false positive is represented by $(1 - \text{spec}_{t_\ell})(1 - \pi_{\ell,A})$.

To account for the uncertainty in the test characteristics, we treat the sensitivity and specificity as unknown parameters. We model the lab validation data with binomial distributions

$$x_{\text{sens},t} \sim \text{Bin}(n_{\text{sens},t}, \text{sens}_t) \tag{3.4}$$

$$x_{\text{spec},t} \sim \text{Bin}(n_{\text{spec},t}, \text{spec}_t), \tag{3.5}$$

where $n_{\text{sens},t}$ and $n_{\text{spec},t}$ denote the number of positive and negative controls tested, respectively, while $x_{\text{sens},t}$ and $x_{\text{spec},t}$ denote the number of controls that had the correct test result.

3.3.2 Modeling fatality data

A primary objective is to estimate the IFR for the location with fatality data. However, we only observe data for the death rate, which can be defined as the proportion of the population, for a particular location and age bin, that died of COVID-19. The IFR can be expressed as the proportion that experience a COVID-19 infection times the probability an individual dies of COVID-19 given they were infected, or equivalently, the seroprevalence, $\pi_{\ell,A}$, times the infection fatality rate, $\text{IFR}_{\ell,A}$. We can convert this death rate to the expected number of deaths by multiplying by the population, $N_{\ell,A}$.

Let $D^\star_{\ell,A}$ represent the number of COVID-19 fatalities for location ℓ and age bin A. We model the number of deaths with a Poisson distribution with the mean equal to the expected number of

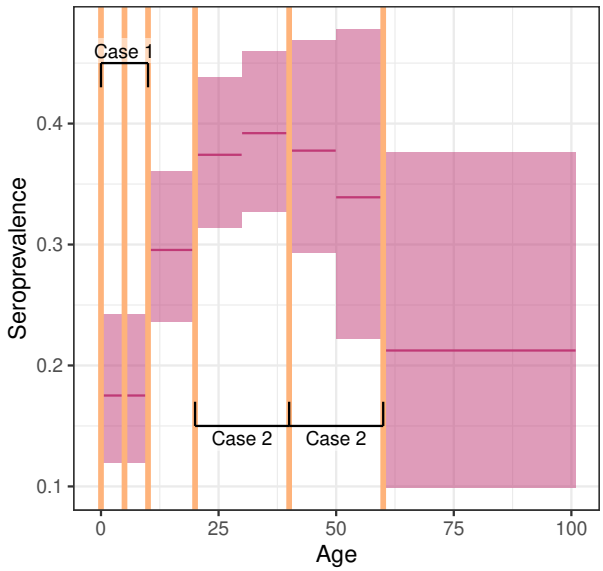


Figure 3.2: The death age bin cutoffs (vertical lines) compared to the seroprevalence study age bins (pink boxes) for Nairobi County, Kenya. The pink horizontal lines give the Rogan-Gladen seroprevalence estimate and the boxes give the 95% confidence interval.

deaths:

$$D^{\star}_{\ell,A} \sim \text{Poisson}(N_{\ell,A} \times \pi_{\ell,A} \times \text{IFR}_{\ell,A}). \tag{3.6}$$

We chose to use a Poisson distribution, similar to O'Driscoll et al. (2021), because a COVID-19 death is a relatively rare event.

The mean in (3.6) utilizes estimates of $\pi_{\ell,A}$ and $\text{IFR}_{\ell,A}$ at the same age bin A. However, the seroprevalence study age bins did not always match the age bins the deaths were reported at. In the following section, Section 3.3.3, we discuss how we handled this issue.

3.3.3 Aligning age bins

The mismatches between seroprevalence study age bins and fatality data age bins at a given location can be classified into one of four categories.

(1) **Fatality data age bins are nested within seroprevalence study age bins:** If we use the seroprevalence estimate from the larger age bin and match it to the fatality data for a subset of

that bin, we assume the average seroprevalence is a sufficient estimate for the smaller subset of ages. This assumption is particularly bad for cases like Nairobi County, Kenya where the seroprevalence estimates vary greatly across age bins (see Figure 3.2). The seroprevalence estimate from the 0-9 age bin of 0.18 is likely too low for the 5-9 age bin and too high for the 0-4 age bin.

Solution: We aggregated the death age bins to match the seroprevalence study age bin.

(2) **Seroprevalence study age bins are nested within fatality data age bins:** Similar to case (1), we do not want to assume the death rate for the entire age bin is representative of the death rate for a subset of the age bin. As an example, consider the 20-39 and 40-59 death data age bins in Figure 3.2 where the serology data is available for 10-year bins.

Solution: We aggregated the information across serology bins to create an average seroprevalence for the fatality age bin. Let B represent the death age bin. We took a weighted average of the seroprevalence estimates for the age bins contained in B with weighting proportional to the population in each serology age bin.

(3) **Seroprevalence study age bins are not nested within the death age bins but are within two years of the fatality data age bins:** Our dataset contained a number of locations where the bounds were within two years of aligning, particularly around ages 18-20. For example, consider the youngest age bins in Chennai, India. The serology study data was available for age bins 0-19 and 20-29, whereas the fatality data was available for age bins 0-18, 19-24, and 25-29. The death age bins are almost nested within the serology age bins except the discrepancy in the cutoff at age 18 versus 19.

Solution: The seroprevalence age limits were adjusted to match those of the death data so the bins aligned or they fell into Case 1 or Case 2.

(4) **Seroprevalence study age bins are not nested and are more than two years away from the fatality data age bins:** In some cases, the only common age bin bounds were at ages 0 and 100.

Solution: The seroprevalence or fatality data age bins were pooled until the location fell into one of the three cases above.

Using the rule set introduced here, this process can be automated. Thus, the model can be applied to new data sets without the need for manually aligning the age bins for each location.

3.3.4 Priors

Identifiability

Our interest is in estimating the seroprevalence rate and the infection fatality rate, but we only observe data on the positivity rate, death rate, and lab validation data. Considering first the infection fatality rate, there are infinitely many combinations of $\pi_{\ell,A}$ and $\text{IFR}_{\ell,A}$ that give the same death rate ($\pi_{\ell,A} \times \text{IFR}_{\ell,A}$) in (3.6). While there are data to inform estimation of test characteristics, seroprevalence, and IFR, we found that the death data had a tendency to dominate the model. The number of data points used to estimate the death rate, the population $N_{\ell,A}$, is generally very large relative to the seroprevalence study sample sizes or the lab validation sample size.

Further, in the case of seroprevalence, we observe data on the test positivity, but not the seroprevalence itself. There are infinitely many combinations of the triplet sens_{t_ℓ}, spec_{t_ℓ}, and $\pi_{\ell,A}$ that could give the same positivity, $p_{\ell,A}$, in (3.3). The lab validation data can inform the sens_{t_ℓ} and spec_{t_ℓ} estimates. However, particularly in cases of small lab validation sample sizes, the lab validation data can be overwhelmed.

To improve identifiability and leverage known serology test development processes, we utilize a combination of weakly informative and strongly informative priors. The details of these priors are described in the following subsections.

Seroprevalence priors

We use weakly informative priors for seroprevalence. For each age bin A and location ℓ we model prevalence with a beta distribution:

$$\pi_{\ell,A} \sim \text{Beta}(2, 6). \tag{3.7}$$

The resulting distribution had a mode at 0.167 and 94% of the mass below 0.5. All the seroprevalence studies were concluded before March of 2021, so we do not expect seroprevalence to exceed 0.5. Given the wide spread of COVID-19, we also do not expect seroprevalence to be zero, so we chose a prior with a nonzero mode. Note, these priors do not allow any pooling of information across locations. We use independent priors for each $\pi_{\ell,A}$ to avoid making any assumptions about how the seroprevalence is related within a location or across locations.

3.3.5 Test characteristic priors

We use more informative priors for the test characteristics, similar to Gelman and Carpenter (2020). Seroprevalence tests are designed to prioritize specificity, with the Centers for Disease Control and Prevention (2020) recommending a specificity of 99.5% when developing COVID-19 antibody tests. Therefore, we use a stronger prior for specificity favoring larger values, and an informative but not as strong prior on the test sensitivity

$$\text{sens}_t \sim \text{Beta}(10, 1), \tag{3.8}$$

$$\text{spec}_t \sim \text{Beta}(50, 1). \tag{3.9}$$

The sensitivity and specificity are assumed to be the same across studies using the same test assay, but the sensitivity or specificity of one assay is independent of the sensitivity or specificity of another assay *a priori*.

IFR prior

We developed a mildly informative IFR prior. We know the IFR generally increases with age (Levin et al., 2020; Pezzullo et al., 2023; Starke et al., 2021). Therefore, we want a prior that allows for larger IFR values when the age bin contains older ages. In addition, age bins containing only older ages should have a larger IFR compared to age bins that contain a wider range of ages. In general, we expect more variability in the IFR for smaller age bins, and less variability for wider age bins. Therefore, we define our prior for IFR as a function of the age bin limits, $L_{\ell,A}$ and $U_{\ell,A}$ corresponding to the lower age and upper age cutoffs, respectively, for age bin A and location ℓ. When the upper age is open ended (e.g., 85+), we set $U_{\ell,A} = 100$.

After consulting with COVID-19 epidemiologists and other experts and looking at studies of the COVID-19 IFR in high-income countries (Levin et al., 2020), we set independent prior distributions for IFR accordingly:

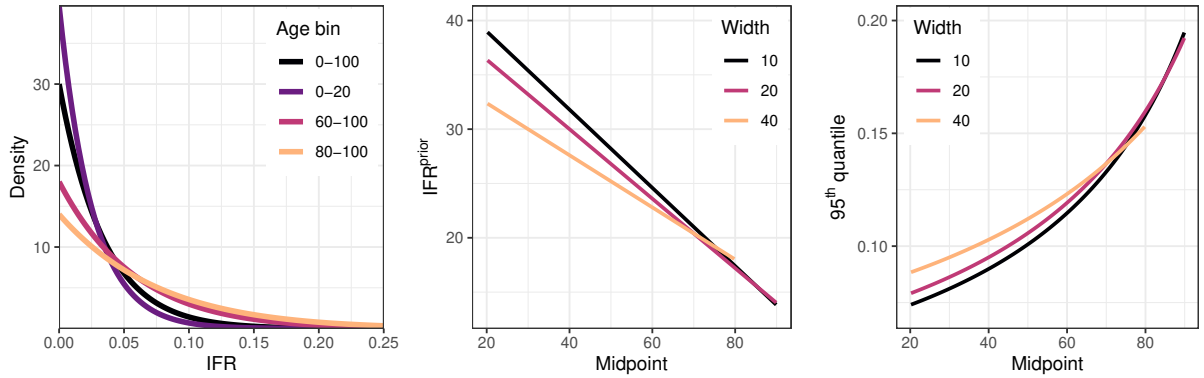
$$\text{IFR}_{\ell,A} \sim \text{Beta}(1, \text{IFR}_{\ell,A}^{\text{prior}}) \tag{3.10}$$

$$\text{IFR}_{\ell,A}^{\text{prior}} = 30 - 20\left[\frac{U_{\ell,A} - 50}{50}\left(1 - \frac{U_{\ell,A} - L_{\ell,A}}{100}\right)\right]. \tag{3.11}$$

If the age bin represents the entire population ($L_{\ell,A} = 0$, $U_{\ell,A} = 100$), the prior simplifies to a $\text{Beta}(1, 30)$. This is mildly informative, but still appropriately vague, with over 85% of the mass below 0.07 (see panel (a) of Figure 3.3). If the upper bound of the age bin is above 50, $\text{IFR}_{\ell,A}^{\text{prior}}$ will be less than 30, resulting in a flatter prior which allows for larger IFR values. If the upper bound is below 50, $\text{IFR}_{\ell,A}^{\text{prior}}$ will be greater than 30, giving a prior that is more concentrated near zero. The amount that is added or subtracted from 30 then depends on the width of the age bin, $U_{\ell,A} - L_{\ell,A}$, with more dramatic adjustments being made for narrower age bins, which are likely more variable. These patterns are visualized in panel (a) of Figure 3.3 with example prior distributions for four age bins.

Panel (b) illustrates the $\text{IFR}_{\ell,A}^{\text{prior}}$ value as a function of the width and midpoint age of the age bin. As the midpoint of the age bin increases, $\text{IFR}_{\ell,A}^{\text{prior}}$ decreases and the slope is steeper for narrower

age bins. Specifically, $\text{IFR}^{\text{prior}}$ for the 40-width bin is smaller than that of the 10-width bin for midpoints less than age 75, and smaller than that of the 20-width bin for midpoints less than 70. Because a wider age bin would incorporate older individuals, who we expect to have a higher IFR, we believe the IFR would be larger for this wider age bin. After around age 70 $\text{IFR}^{\text{prior}}$ is larger for the age bins with a width of 40, resulting in a tighter prior around zero. This is because the wider age bin would incorporate more younger individuals and should therefore have a smaller IFR. This is further illustrated by panel (c). The 95^{th} quantile for the 40-year width bin is larger below about age 70 and smaller above age 70 for.



(a) **(b)** **(c)**

Figure 3.3: Panel (a) shows the prior distribution for example age bins. Panel (b) shows $\text{IFR}^{\text{prior}}$ values for age bins centered at the midpoint value and with a total width indicated by the color. Note, $U_{\ell,A}$ has an upper bound of 100, so the width 40 bin can have a maximum midpoint of 80. Panel (c) shows the 95^{th} quantile of the prior distribution for various widths and midpoints.

3.4 Estimation

The joint posterior distribution of $(\{\pi_{\ell,A}\}, \{\text{IFR}_{\ell,A}\}, \{\text{sens}_t\}, \{\text{spec}_t\})$ given the seroprevalence, lab validation, and fatality data $(\{n_{\ell,A}\}, \{R^{\star}_{\ell,A}\}, \{x_{\text{sens},t}\}, \{x_{\text{spec},t}\}, \{n_{\text{sens},t}\}, \{n_{\text{spec},t}\}, \{D^{\star}_{\ell,A}\})$ is not available in a closed form. We obtained 10,000 posterior parameter samples from three Markov chains using version 3.32.2 of Stan (Stan Development Team, 2022). The first 5000 iterations were discarded as burn-in. This resulted in an effective sample size of at least 1,200 for each parameter. The traceplots and $\widehat{R}$ suggested convergence. We therefore leverage a Markov

chain Monte Carlo algorithm to sample from the posterior distribution of the parameters given the data.

3.5 Comparison of methods

We compare our Bayesian estimates of the test characteristics, seroprevalence, and IFR to the Rogan-Gladen (RG) based estimates commonly used in practice (Axfors and Ioannidis, 2022; Levin et al., 2020; Pezzullo et al., 2023). For test characteristics, we consider the raw estimates $\widehat{\text{sens}}_t = x_{\text{sens},t}/n_{\text{sens},t}$ and $\widehat{\text{spec}}_t = x_{\text{spec},t}/n_{\text{spec},t}$ for sensitivity and specificity, respectively. The adjusted RG seroprevalence estimate is then

$$\widehat{\pi}_{\ell,A} = \frac{p_{\ell,A} + \widehat{\text{spec}}_{t_\ell} - 1}{\widehat{\text{spec}}_{t_\ell} + \widehat{\text{sens}}_{t_\ell} - 1}. \tag{3.12}$$

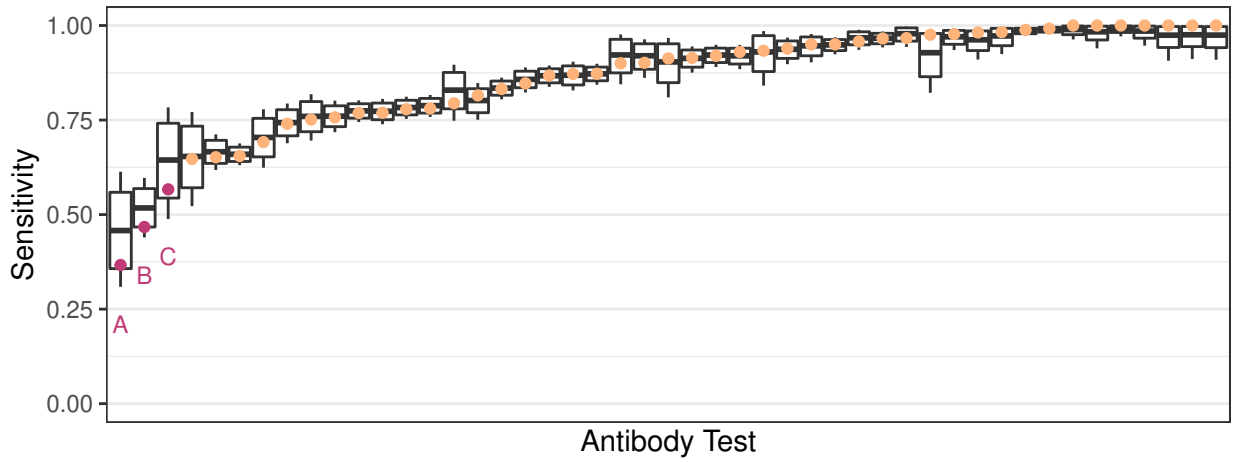
To create confidence intervals for the RG seroprevalence, we use the `epiR` package (Stevenson and Sergeant, 2023) to create Blaker's confidence intervals (Blaker, 2000) that account for the adjustments made to the positivity rate, but do not account for the uncertainty in $\widehat{\text{sens}}$ and $\widehat{\text{spec}}$.

The naive estimate for IFR, based on the RG prevalence estimate, is then

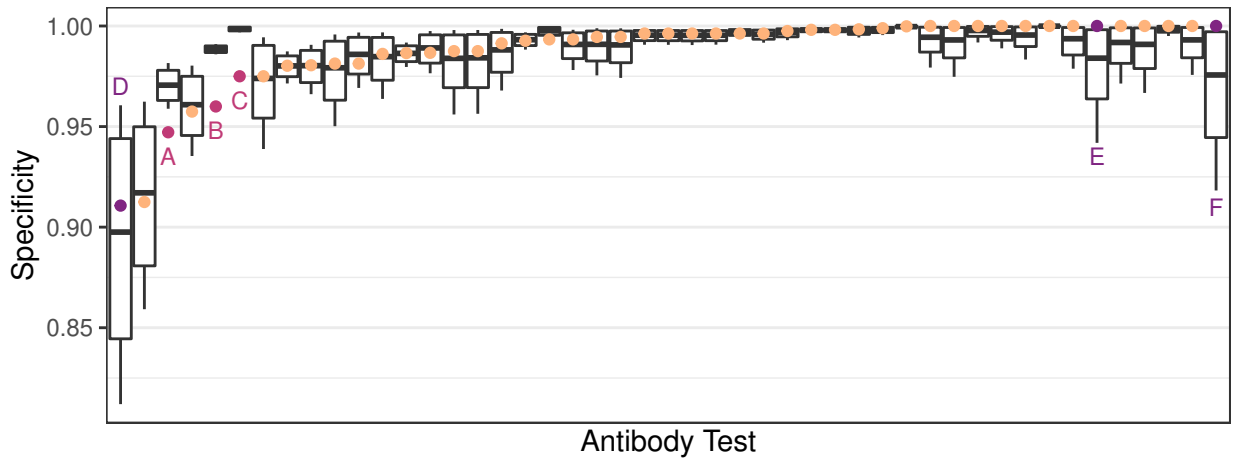
$$\widehat{\text{IFR}}_{\ell,A} = \frac{D^\star_{\ell,A}/N_{\ell,A}}{\widehat{\pi}_{\ell,A}}. \tag{3.13}$$

Following Levin et al. (2020), we calculated naive confidence intervals for the IFR by replacing $\widehat{\pi}_{\ell,A}$ with either the upper bound or the lower bound of the corresponding confidence interval for $\widehat{\pi}_{\ell,A}$ in (3.13).

Across these 107 locations, we estimate seroprevalence for 471 age bins and estimate IFR for 182 age bins, where the serology study sample, test assay accuracy, amount of lab validation data, and population vary.



(a)



(b)

Figure 3.4: Boxplots show the posterior distribution of the (a) sensitivity and (b) specificity for each test assay. The whiskers indicate the 95% credible interval, the box indicates the 80% credible interval, and the center line is at the posterior mean. The pink and yellow dots indicate the raw estimates (number of correctly identified controls/number of controls tested). Raw sensitivity and specificity estimates of interest are emphasized and labeled in pink and purple.

3.5.1 Test characteristics

Figure 3.4 shows the posterior distributions for the test characteristics. In most cases, the raw estimate is within the range of the posterior distribution, with a few exceptions.

First, we consider the sensitivity estimates (Figure 3.4(a)). The average difference between the posterior mean and the raw estimate was 0.013, with a standard deviation of 0.014. The posterior mean was over 0.05 greater than the raw estimate for the three assays with the smallest raw sensi-

tivities (highlighted in pink). Two of the assays (A and C) had only 30 positive controls, while the third (B) had 122 positive controls. These unusually low estimates with limited sample sizes were pulled slightly upwards by the prior on sensitivity.

For the specificity estimates (Figure 3.4(b)), the mean difference between raw estimates and the posterior means was 0.007, with a standard deviation of 0.009. We focus on the tests where the difference between estimates was greater than 0.01. There were a few notable differences between the raw estimate and the posterior mean (highlighted in pink) that demonstrate the benefits of simultaneously estimating the test characteristics and seroprevalence, rather than estimating these separately. To illustrate, we focus on the Qingdao Hightop Biotech IgM/IgG Duo assay (labeled B). The raw specificity estimate was 0.96. This means we expect a minimum of 4% of the tests to be positive. However, multiple locations using this assay had positivity rates far below 4%. Tete, Mozambique had positivity rates as low as 0.3% and 0.9% in age bins with sample sizes of 623 and 336, respectively. By simultaneously estimating seroprevalence and the test characteristics, the model identified that given the positivity rates, it was likely the specificity was higher than the raw estimate. The other assays with raised specificity estimates were also used in locations with lower than expected positivity rates, assuming the specificity was equal to the raw estimate. This benefit of joint modeling had also been noted by Larremore et al. (2022).

Additionally, three tests (highlighted in purple) had posterior means for specificity that were between 0.01 and 0.02 smaller than the raw estimates. Both E and F had small numbers of negative controls (42 and 31, respectively) and raw specificities of 1.0, so the lowered estimates reflect the uncertainty about the specificities. For D, the uncertainty in the specificity estimate was much larger than the change in the point estimate as shown by Figure 3.4 (b). The lower raw specificity and smaller sample size (56) led to additional uncertainty for this test.

3.5.2 Seroprevalence point estimate

Figure 3.5 shows the RG estimates compared to the posterior means for seroprevalence. For 90% of the bins, the posterior mean seroprevalence was within 0.035 of the RG estimate. The 46

bins with a difference in estimates greater than 0.035 fall into at least one of three cases. Case 1: there are 13 RG estimates more than 0.035 from the posterior mean that are negative. This occurs when the Bayesian model does not allow for negative estimates, and thus, its estimates are larger. Case 2: there are 32 age bins for which the posterior mean sensitivity or specificity differed substantially from the raw estimates as explained in Section 3.5.1. These resulted in changes to the seroprevalence estimate. Case 2 contains case 1 as all the age bins with negative RG estimates also had a change in the associated sensitivity or specificity estimate. Finally, the 14 remaining bins with large differences between the RG estimate and the posterior mean all had small serology sample sizes, ranging from 8 to 68. When seroprevalence was low (RG less than 0.05), the large uncertainty allowed for higher posterior means as the Bayesian estimates cannot go below zero, leading to right skewed posterior distributions. When seroprevalence was higher (greater than 0.25), the posterior means were smaller the RG estimates, as weakly suggested by the prior.

Figure 3.6 (a) shows an example location where the RG and posterior mean estimate notably differ. In the 80+ age bin in Karnataka, India, 5 out of 9 study participants tested positive for COVID-19 antibodies. With raw sensitivity and specificity estimates of 0.85 and 1.0, respectively, the RG seroprevalence estimate is 0.66. Due to the small serology study sample size, our prior for seroprevalence was rather influential for this age bin. The posterior mean was pulled down to 0.39, aligning with the prior mode of 0.167.

3.5.3 Seroprevalence intervals

Figure 3.7 shows the width of the RG 95% confidence intervals compared to the 95% credible interval for seroprevalence. Panel (a) shows that the seroprevalence sample size directly relates to the width of the confidence intervals and credible intervals. The age bins with a confidence interval width greater than 0.3 had a serology study sample size between 8 and 68. In these cases where the serology study provided limited information, the RG confidence intervals were wider than the credible intervals. This reflects the slight shrinkage in the parameter posterior that results from our prior when serology sample sizes are low. The difference in confidence interval and credible

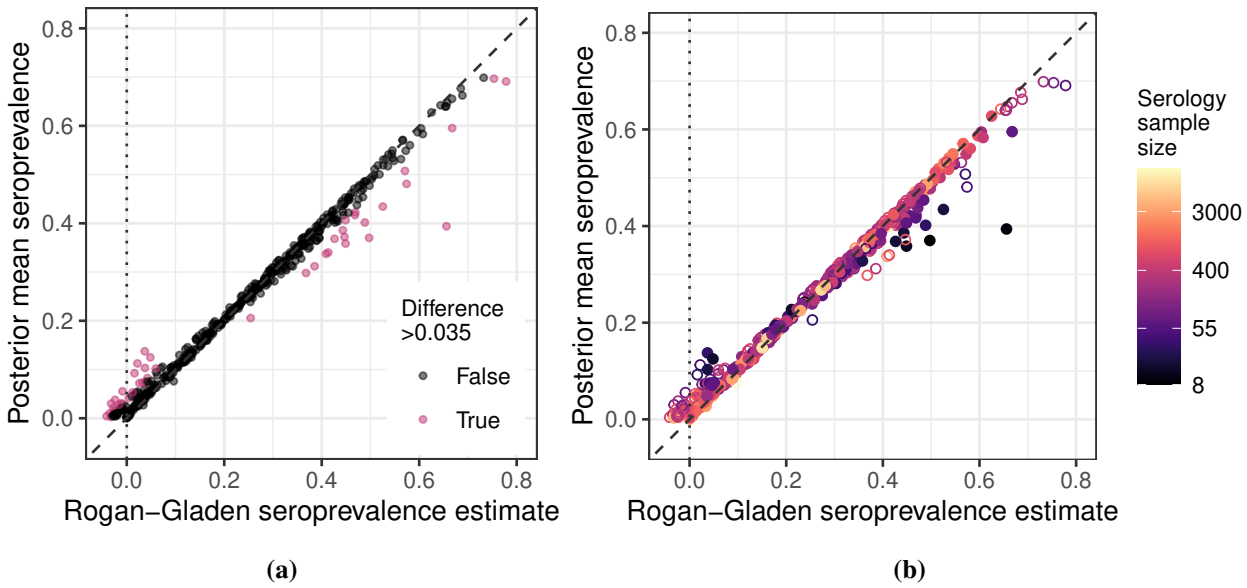


Figure 3.5: The Rogan-Gladen estimates of seroprevalence compared to the posterior mean of the seroprevalence. The points are colored by (a) an indicator for which estimate differ by more than 0.035 or (b) the serology study sample size. In (b), open circles indicate points where, compared to the raw estimate, the posterior mean of the sensitivity differed from the raw estimate by more than 0.05 or the posterior mean of the specificity differed from the raw estimate by more than 0.01. The dotted line shows $x = 0$.

interval widths is especially apparent for bins with RG intervals widths greater than 0.6. The upper two age bins in Karnataka (Figure 3.6) correspond with two of the black points in Figure 3.7(a) with serology study sample sizes of 29 and 9.

When the RG widths were less than 0.2, the credible interval width tended to align with the RG width or were wider. In the cases where the credible intervals width was greater than the confidence interval width by at least 0.02, the number of positive and negative controls was small (see Figure 3.7(b)). The limited number of controls increased the uncertainty about the sensitivity and specificity, which in turn increased uncertainty about seroprevalence in the Bayesian estimates. The RG intervals treat sensitivity and specificity as known, and do not account for this added uncertainty.

Figure 3.6(b) shows the confidence intervals and credible intervals for Asunción and Central Department, Paraguay. The test sensitivity and specificity were established with 30 positive controls and 80 negative controls, resulting in non-negligible uncertainty in the test characteristics. Notably, the 95% credible intervals for sensitivity and specificity are $(0.84, 0.99)$ and $(0.86, 0.96)$,

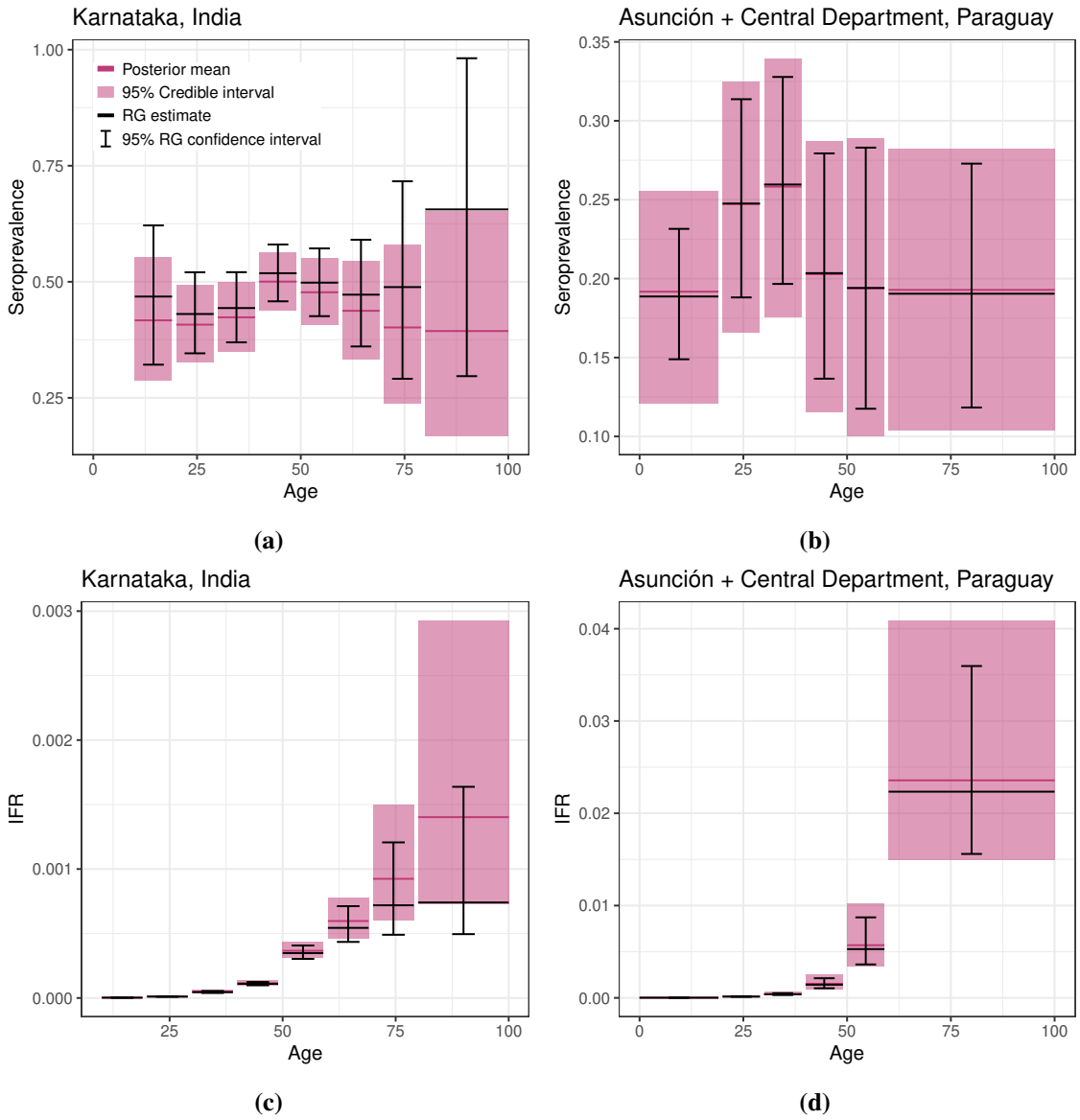


Figure 3.6: Bayesian posterior means and 95% credible intervals for seroprevalence compared to Rogan-Gladen point estimates and 95% confidence intervals.

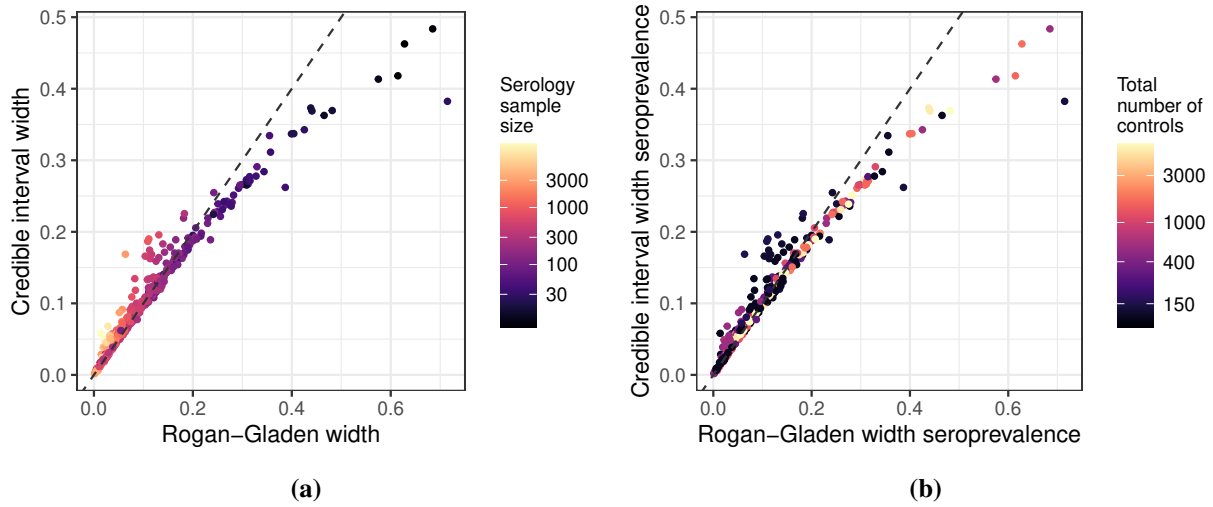


Figure 3.7: Width of the 95% confidence interval for the Rogan-Gladen seroprevalence estimate compared to the width of the 95% credible interval for seroprevalence. The points are colored by (a) the serology sample size and (b) the total number of positive and negative controls combined.

respectively, and the serology sample sizes ranged from 151 to 603 across the age bins. In general, the RG confidence intervals, which only account for serology study sampling variability and not test characteristic uncertainty are distinctly narrower.

3.5.4 IFR

Figure 3.8(a) shows the RG-based and Bayesian IFR point estimates are largely similar. Of the 11 age bins with a difference in IFR estimate greater than 0.0015, all had a corresponding difference in the seroprevalence estimates. Either the RG seroprevalence estimate was larger and the IFR estimate was smaller compared to the Bayesian estimates, or the reverse was true. Note, the RG-based IFR estimate is not useful if the RG estimate is negative or 0. In our dataset, none of the locations containing fatality data had a negative RG estimate, so the RG estimate can be applied to all age bins with fatality data.

Figure 3.8(b) compares the widths of the 95% intervals. The naive confidence interval for IFR assumes there is no variability in the number of deaths, whereas the Bayesian model assumes that deaths follow a Poisson distribution. As noted by Campbell and Gustafson (2021), the difference between the observed number of deaths and the expected number of deaths is an additional source

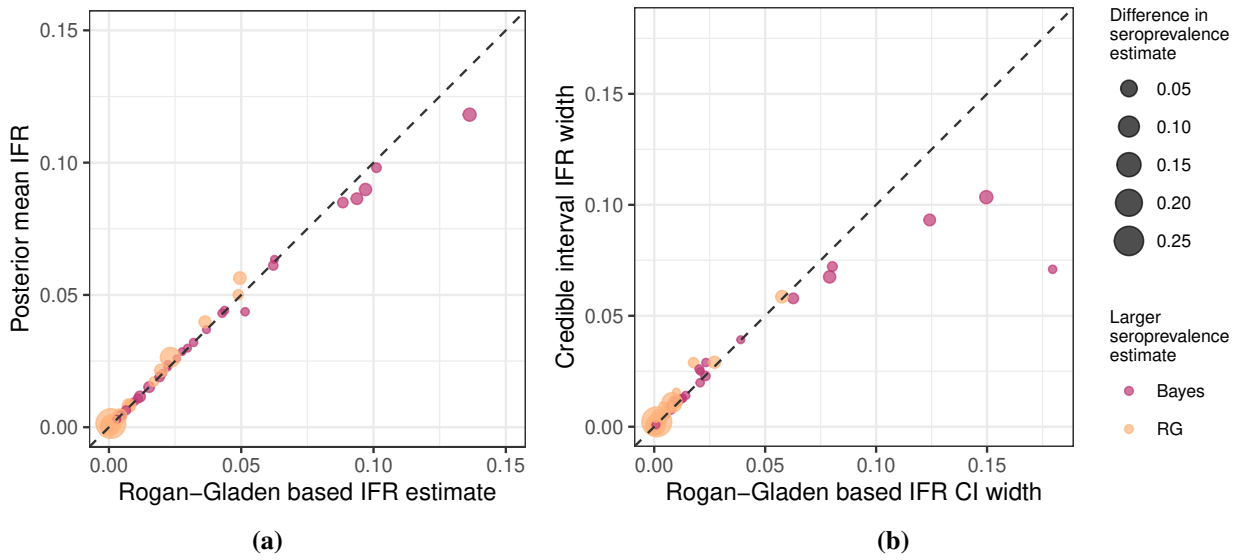


Figure 3.8: (a) The RG-based IFR estimate compared to the posterior mean for IFR and (b) the width of the 95% RG-based IFR confidence interval compared to the width of the 95% IFR credible interval. Points are colored according to which method gave a larger seroprevalence estimate, and the size indicates the magnitude of the difference between the seroprevalence estimates of the two methods. One outlying age bin with an RG based IFR estimate of 0.63 and confidence interval width over 12 was removed.

of uncertainty that should be accounted for. Additionally, the Bayesian model accounts for test characteristic uncertainty that the RG estimates do not. Most of the credible intervals (91%) are wider than the confidence intervals, accounting for these added sources of uncertainty.

Eleven age bins had a confidence interval width greater than the credible interval width. In each case, the Bayesian seroprevalence estimate was larger than the RG estimate. Thinking of the IFR as a ratio of the death rate divided by the seroprevalence, the impact of small changes in the seroprevalence estimate on the IFR estimate increases as seroprevalence decreases. Thus, the uncertainty in IFR will naturally be larger for the smaller seroprevalence estimates. For example, in the 60+ age bin in Cuiabá, Brazil, the 95% RG confidence interval for seroprevalence and the 95% credible interval for seroprevalence had similar widths: 0.111 and 0.113, respectively. However, the RG seroprevalence estimate was 0.016 smaller than the posterior mean of 0.160, which translated into the RG-based IFR confidence interval being wider than the credible interval with a width of 0.079 for the RG-based interval and a width of 0.068 for the credible interval. If

we add 0.016 to the seroprevalence confidence interval bounds, the new RG-based IFR confidence interval would be narrower than the credible interval for IFR with a width of 0.062.

Omitted from the Figure 3.8 are age bins with zero deaths. If zero deaths were observed for a particular location and age bin, then the RG-based IFR estimate is zero and the interval is exactly zero. Additionally, one location had a negative lower bound for the RG estimate, disallowing computation of a confidence interval for IFR.

3.6 Inferences for Developing Countries

The posterior mean seroprevalence for locations with age-specific serology studies is shown in Figures 3.9-3.11. Seroprevalence does not appear to change much across the age bins for most locations, with a few exceptions such as West Bank, Palestine or Nairobi Country, Kenya. In each of these locations the seroprevalence is lower for the oldest age bin, suggesting the older, more vulnerable individuals were protected from infection to some degree. Some regions such as East Asia or Africa have generally lower seroprevalence levels, but we cannot make direct comparisons between locations because the serology studies were conducted at different times. For example, the serology studies in China were conducted in April 2020 while the serology studies in India were conducted between June 2020 and January 2021. The seroprevalence for the three locations in China is low as these studies were conducted early in the pandemic. The seroprevalence is generally higher in India where studies occurred later in the pandemic.

Figure 3.12 shows the age-specific IFR estimates for the corresponding death and serology bins. IFR is generally monotonically increasing with a few minor exceptions in places like Nairobi County, Kenya or Villavicencio, Colombia. Karnataka, India shows lower IFR estimates for all ages compared to the other countries and the high-income country benchmark. However, in India, only 24% of deaths are well-certified, meaning only 24% of deaths are attributed to a specific, well-defined cause (Fullman et al., 2017), suggesting COVID-19 deaths may be under-reported. If COVID-19 deaths are under-reported, the true IFR would be above our estimate, and closer to the high-income country benchmark. In general, locations show similar IFR estimates.

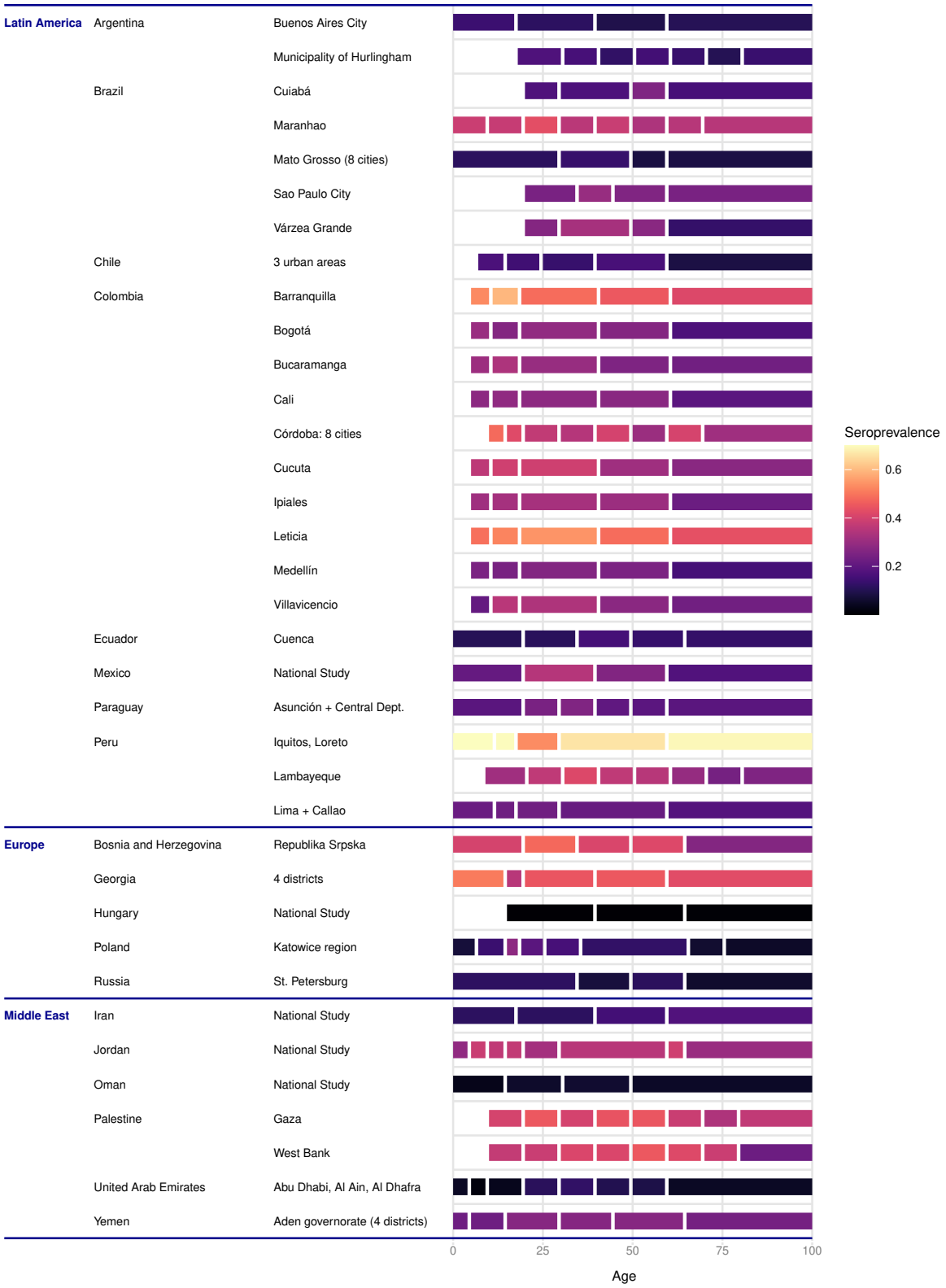


Figure 3.9: Posterior mean seroprevalence for locations with age-specific seroprevalence data.

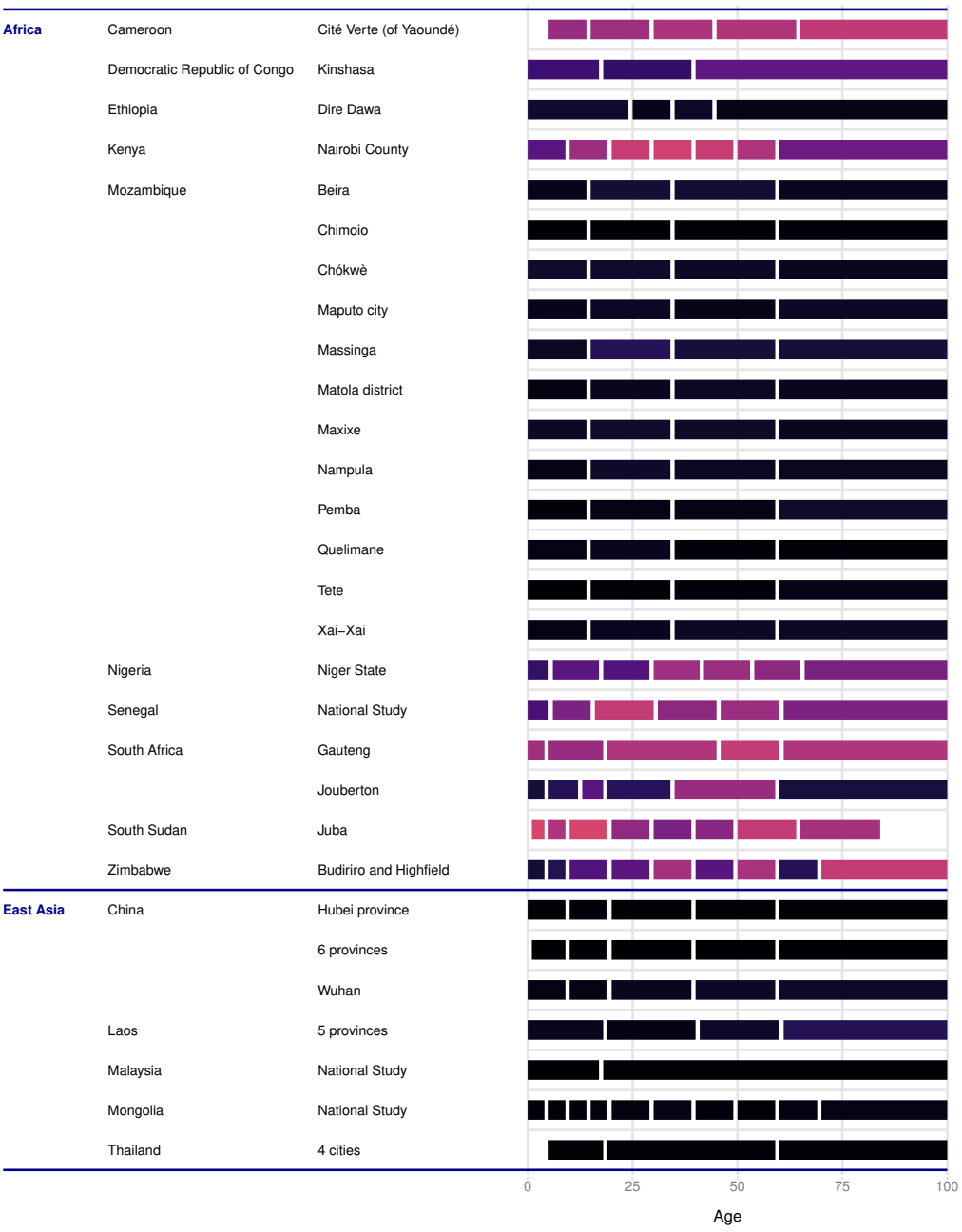


Figure 3.10: Posterior mean seroprevalence for locations with age-specific seroprevalence data.

Figure 3.11: Posterior mean seroprevalence for locations with age-specific seroprevalence data.

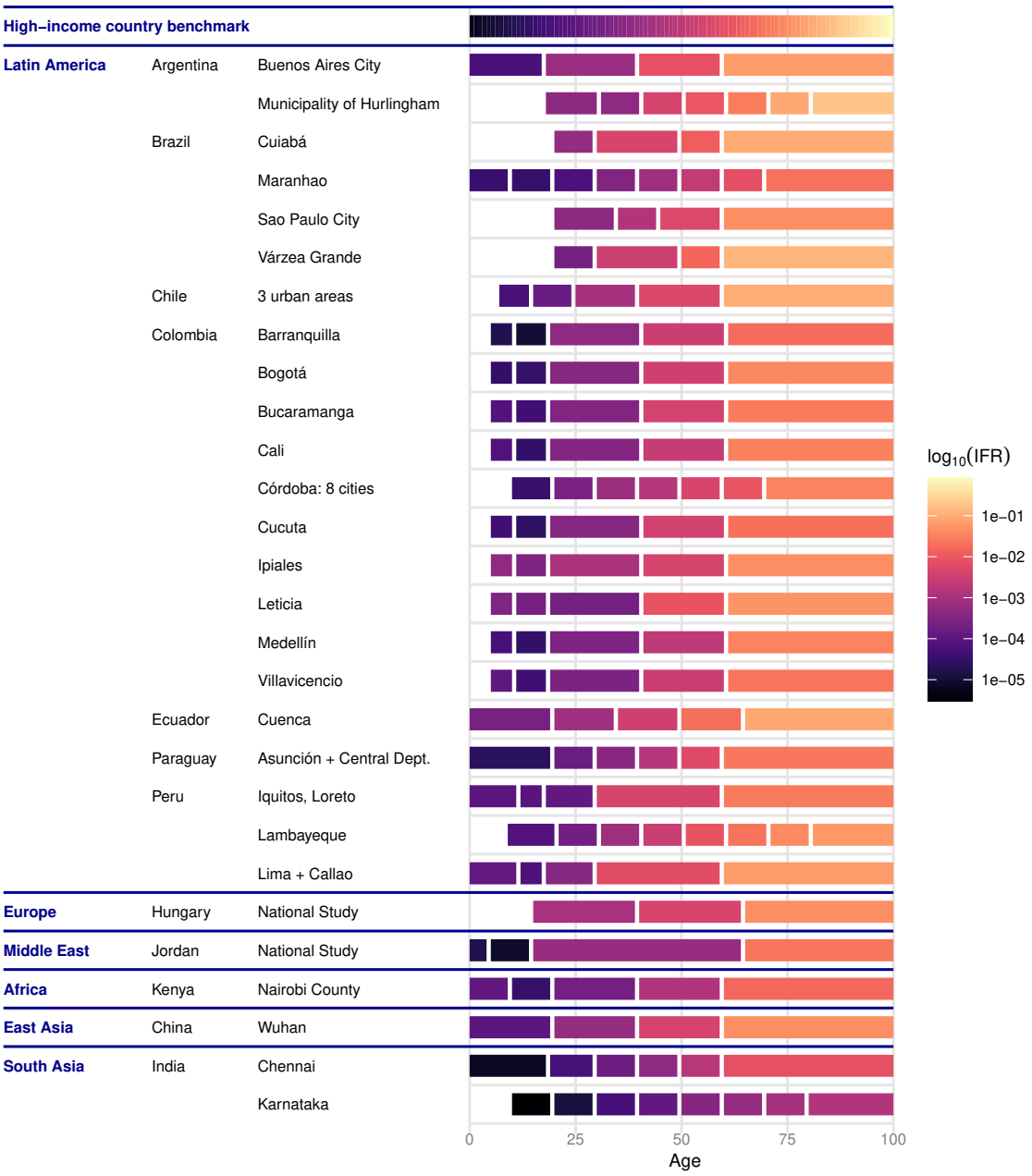


Figure 3.12: Posterior mean IFR for locations with age-specific fatality data. These are compared to the high-income country benchmark from Levin et al. (2020).

3.7 Discussion

The Bayesian model introduced here achieves our goal of enabling simultaneous inference for age-specific and location-specific seroprevalence, test characteristic, and IFR (when applicable) for each of our developing-country locations. By estimating the sensitivity and specificity for each assay, our seroprevalence and IFR estimates incorporate the full uncertainty in the test characteristics. Our choice of priors allows data-driven estimates while incorporating prior knowledge, allowing for estimation of a flexible model with limited data. Our model makes minimal assumptions about the relationship between age and seroprevalence, or age and IFR by estimating these at the bin level. The bin-level model also allows for modeling locations with any level of data whether that be just seroprevalence data, or seroprevalence and fatality data, and either age-specific or population level data. While our model was developed to model the IFR of COVID-19 in developing countries, the flexibility and lack of assumptions make it applicable to modeling IFRs in general for diseases using seroprevalence studies.

Comparing our model to the Rogan-Gladen based estimates we found our model improved both the point estimates and the uncertainty quantification of these estimates. Our model constrained seroprevalence to be between zero and one, unlike the Rogan-Gladen estimates, and it allowed for IFR estimates with uncertainty even when zero deaths occurred. By simultaneously modeling test characteristics, seroprevalence, and IFR, the uncertainty of the test characteristics was fully propagated into the seroprevalence and IFR estimates. The IFR credible intervals tended to be wider than the confidence intervals after accounting for variability in the number of deaths, uncertainty in the seroprevalence, and uncertainty in the test characteristics. Additionally, the simultaneous modeling meant all of our estimates accounted for all the data sources, collectively, rather than considering each data source separately. For example, the specificity estimates reflected both the lab validation data and the serology study data in our model.

We found seroprevalence estimates were generally consistent across age bins within a location, suggesting developing countries were not able to shelter older individuals in most of our study locations. However, the average seroprevalence varied dramatically across locations as studies

were performed at different times during the first year of the pandemic. IFR overall increased with age, and most of the locations had similar estimates for commensurate age bins.

Chapter 4

Hierarchical Bayesian modeling of age-specific COVID-19 infection fatality rates in developing countries

4.1 Introduction

The 2019 coronavirus disease (COVID-19) pandemic has had devastating impacts worldwide, with at least 6.59 million confirmed COVID-19 deaths as of November 2022 (Jha et al., 2022; Our World in Data, 2022). However, comparing the burden across locations with varying age distributions is specifically difficult because COVID-19 infections and fatalities vary substantially by age (Levin et al., 2020; Pezzullo et al., 2023; Starke et al., 2021). Estimating age-specific metrics of the COVID-19 burden is, therefore, necessary to make meaningful comparisons across locations and best allocate scarce resources with age-specific policies. Between location comparisons are further complicated by small sample sizes and uncertainty about testing characteristics of newly developed COVID-19 tests, such as sensitivity and specificity, both of which result in additional uncertainty about location-specific estimates. Policy makers turned to these studies to make real-time decisions, and therefore, it was essential to quantify and communicate uncertainty in the estimates. In this study, we focus on estimating age-specific metrics of the COVID-19 burden, along with corresponding uncertainty bounds, and comparing the burden across developing country locations with different age distributions.

There are several possible measures of disease burden. In this paper, we focus on estimating prevalence and the infection fatality rate (IFR)–the proportion of individuals infected with severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) that subsequently die from the disease. Compared to the more commonly used and easier to estimate case fatality rate (CFR)–defined as the ratio of the number of deaths to the number of reported cases–IFR depends on a measure of the

number of prior infections, including unreported infections. A key benefit of IFR is it not biased by preferential testing and reporting.

Early in the COVID-19 pandemic, tests for active infection of COVID-19 were limited and essentially unavailable in many areas. Thus, reported cases were likely a substantial under count of infections (National Academies of Sciences, Engineering, and Medicine, 2020) and researchers instead turned to serology studies to estimate the number of infections. In a serology study, a sample of the population is tested with an antibody test to identify those who have had a COVID-19 infection and seroconverted, i.e., built antibodies against the virus. The proportion of individuals with SARS-CoV-2 antibodies–the seroprevalence–is then used as a proxy for the proportion of the population that has been infected. Antibody tests used early in the pandemic were developed quickly using limited validation data to estimate test characteristics such as specificity and sensitivity. As a result, estimates of prevalence from serology studies were plagued by uncertainty due to often small samples from the population, inaccuracies from the antibody tests, and further general uncertainty in the antibody test's characteristics. Accounting for these avenues of uncertainty was paramount in reporting uncertainty in community prevalence (Gelman and Carpenter, 2020).

Discrepancies between the effects of COVID-19 for the young and old was quickly recognized. Levin et al. (2020) estimated that variation in age distributions between countries explained approximately 90% of the variation in population-level IFRs. Specifically, they found the IFR for high-income countries to be about 0.01% at age 25 and about 15% at age 85. Thus, age-specific IFR estimates are required to make meaningful comparisons of the disease burden across locations and inform age-based policies that mitigate disease transmission (Malani et al., 2022). This relies upon information on age-specific infections and age-specific fatalities. Most studies of COVID-19 IFR reported estimates for coarse age bins (e.g., Perez-Saez et al., 2021; O'Driscoll et al., 2021; Pezzullo et al., 2023). A limited number of studies have reported estimates of COVID-19 IFR as a continuous function of age (Levin et al., 2020, 2022; COVID-19 Forecasting Team, 2022), but all of these studies rely on multi-step modeling approaches and lack location-specific curves.

The majority of studies quantifying the burden COVID-19 have focused on high income countries, with large serology studies and highly granular death data. The data availability and quality in developing counties posses additional challenges for estimating age-specific rates and between country comparison. In developing countries, serology studies typically had smaller sample sizes and reported data on age inconsistently, with some locations providing individual data and other reporting only coarse age bins. As a result, existing methods that rely on detailed infection and death data are not suitable.

In this study, we focus on estimating infection and fatality rates for 26 developing country locations, where serology studies were reported for coarse age bins with often low sample sizes. We propose a Bayesian hierarchical modeling approach to jointly estimate continuous age-specific and location-specific seroprevalence and IFR curves that account for the many aforementioned challenges: the many sources of uncertainty, binned nature of seroprevalence and death data, and the inconsistent and coarse data resolution at which the developing country data are reported. By jointly modeling seroprevalence, antibody test assay characteristics, and the IFR, we fully propagate uncertainty to all model estimates. Our hierarchical modeling framework shares information across age bins within a location, as well as across locations to improve estimation in cases of limited data. While prior works have addressed some aspects of this problem, we present a holistic solution that addresses all objectives simultaneously. Using our proposed model, we estimate continuous seroprevalence curves for all locations, which allow for age-specific comparisons between our developing country study locations and high-income country estimates.

4.2 Data

In this study we revisit the serology and death data from developing countries described in Levin et al. (2022). The rich data set contains seroprevalence study data for a total of 107 locations from 44 countries. Of the 107 locations, 63 had corresponding COVID-19 fatality data. Lab validation data was assimilated from antibody test manufacturers and the population age distribu-

tion was gathered for each location. The data sources vary in terms of information provided, age aggregation, antibody test and test characteristics, and quality.

Because our aim is to estimate seroprevalence and IFR as flexible, continuous functions of age, we restricted our analysis to studies with age-specific seroprevalence and death data. Making inference about the shape of a curve from only a few data points is extremely challenging, and prior analysis found IFR showed heterogeneity across locations that was not well explained by observed covariates, such as healthcare capacity or GDP, limiting the potential of borrowing information in a data-informed manner (Levin et al., 2022). Therefore, we restrict our attention to studies with at least four serology age bins and at least five death age bins to allow inference on the structure of IFR with age, resulting in a total of 26 study locations for this analysis. Twenty-two locations are from Latin America, two are from South Asia, and there is one study each from Africa and the Middle East (see Figure 4.1). Each study focused on a specific region within the country, with the exception of Jordan, which is a national study. For example, there are nine studies within Colombia, each confined to a different, non-overlapping region.

4.2.1 Serology data

As described in Levin et al. (2022), we collected and evaluated serology studies from government reports, published papers, and preprints. Studies using convenience samples like blood donors, volunteers, or residual sera were excluded, as well as those with imbalanced gender ratios, inaccurate test assays, insufficient reported data, and studies occurring during accelerating outbreaks. All studies were concluded before March 2021–before vaccines were readily available in developing countries. Therefore, a positive antibody test due to a COVID-19 vaccine, rather than a prior infection, is extremely unlikely in the data.

For each location, researchers reported results for between four and eight age bins. The total number of participants tested and the number of participants that tested positive were then recorded for each bin. Figure 4.1 shows the age bins for each study location, as well as the number of participants (i.e., sample size) in each bin. Most locations have wider age bins and often smaller

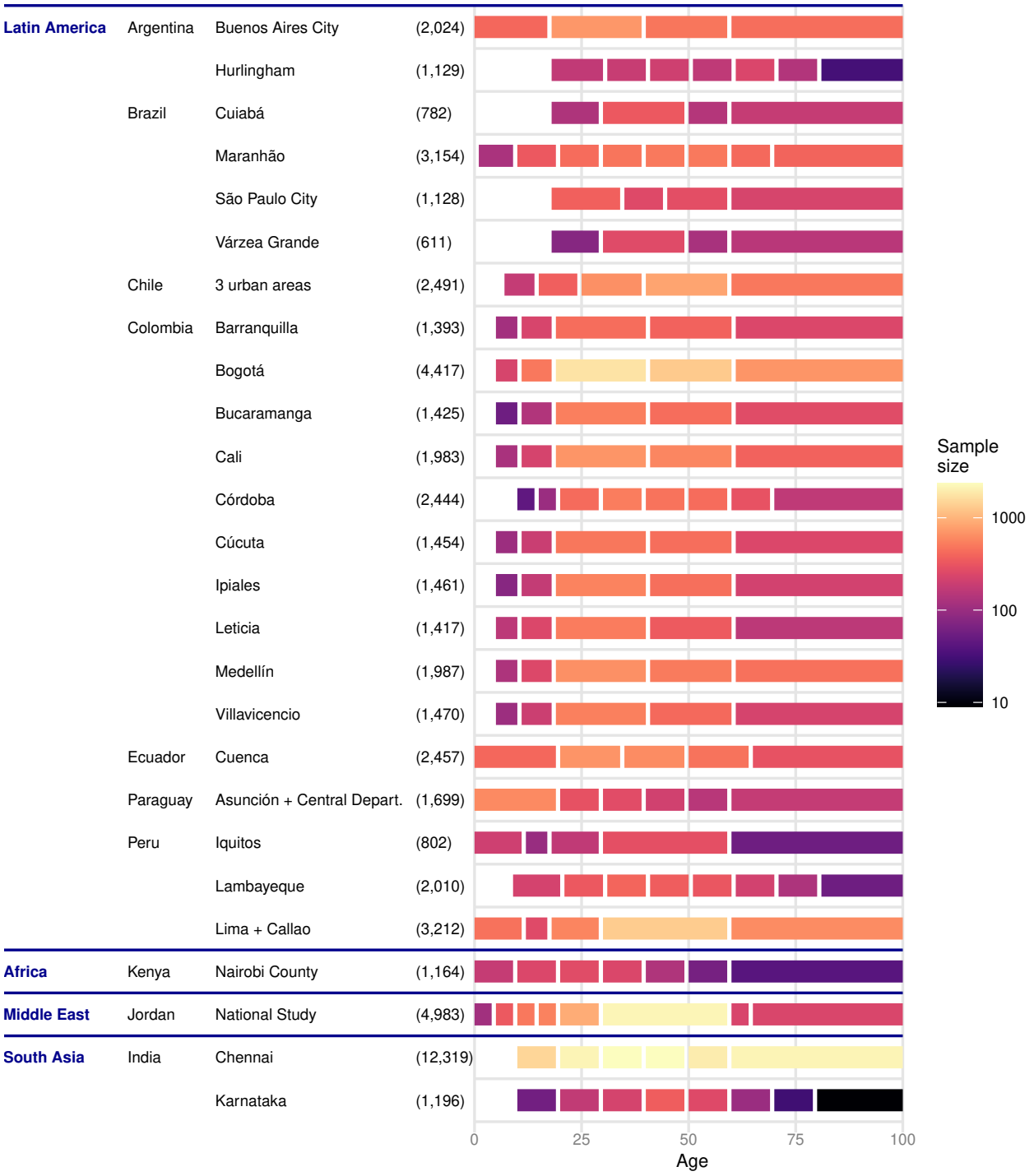


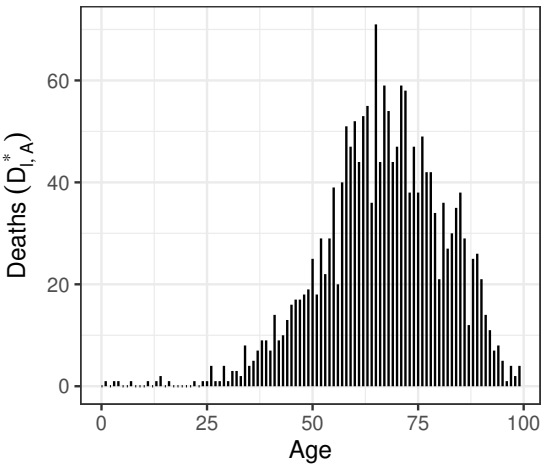
Figure 4.1: Number of participants in serology study for each age bin and location ($n_{\ell,A^R_{\ell,b}}$). Total sample size shown in parentheses.

sample sizes for the oldest individuals as seen in, for example, Hurlingham, Argentina and Lambayeque, Peru. Karnataka, India had the smallest average bin sample size with between 9 and 353

samples for each age bin, whereas Chennai, India had the largest average bin sample size with between 1473 and 2353 samples for each bin. Panel (a) of Figure 4.2 shows an example of the seroprevalence data for Lambayeque, Peru.

Age bin $\left(A^R_{\ell,b}\right)$	Sample size $\left(n_{\ell,A^R_{\ell,b}}\right)$	Number positive $\left(R^*_{\ell,A^R_{\ell,b}}\right)$
9-20	220	59
21-30	329	102
31-40	394	139
41-50	353	109
51-60	315	94
61-70	215	54
71-80	129	26
81+	55	12

(a)



(b)

Figure 4.2: (a) Seroprevalence data collected between 6/24/2020 7/10/2020 and (b) corresponding cumulative death data $(D^\star_{\ell,A^D_{\ell,b}})$ for Lambayeque, Peru.

4.2.2 Test characteristic data

For each test used in a seroprevalence study, we directly model the test assay lab validation data, namely the number of positive/negative controls tested in the validation study and the number that test positive in the validation study, to estimate the test sensitivity and specificity. In total, thirteen tests were used in our sample of seroprevalence studies, with six of the test assays used in multiple studies (see Figure 4.3). Between 29 and 632 positive controls and between 42 and 5272 negative controls were used to validate each test assay. For instance, the Coretest COVID-19 IgM/IgG Antibody Test used in Lambayeque, Peru was developed using 73 positive controls, 58 of which tested positive, giving a crude sensitivity estimate of 0.795, and 222 negative controls, which all correctly tested negative, giving a crude specificity estimate of 1.

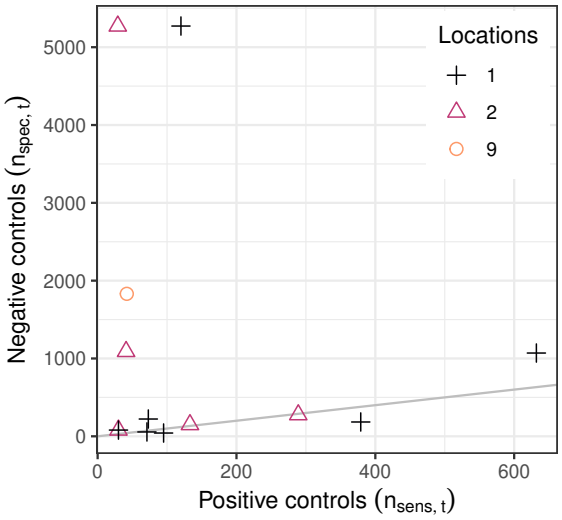


Figure 4.3: Number of positive controls ($n_{\text{sens},t}$) and number of negative controls ($n_{\text{spec},t}$) for each test assay. The color and shape of the point indicate the number of locations using the assay, either 1, 2, or 9. The line $x = y$ is shown in grey.

4.2.3 Death data

We modeled aggregate data on the number of confirmed and suspected COVID-19 fatalities associated with each serology study location using individual case data, when available, and public health reports otherwise. To account for the delay in time between infection and fatality, we collected the cumulative number of deaths up to fourteen days after the midpoint date of the associated seroprevalence study when using individual case data and the cumulative number of deaths up to twenty-eight days after the midpoint date of the study when using public health reports due to the lag in death reporting.

Twenty study locations had individual case data available, from which we can determine the number of deaths for each one-year age bin (e.g., $[0, 1)$, $[1, 2)$, etc.). For the remaining six locations, death data was available for between six and eighteen age bins. Lambayeque, Peru, for example, has individual case data summarized in one-year age bins (Figure 4.2 panel (b)). Fifteen ages had zero deaths, with the most deaths, namely 71, occurring at age 65.

4.2.4 Age distribution data

The population distribution data, collected from census data and websites such as worldometers.info, worldpopulationreview.com, and populationpyramid.net, contained the number of people in each study area at a resolution of between seven to twenty age bins. These distributions were further refined to a 5-year age bin resolution using national age distribution data (see Appendix C.1). Acknowledging that age is a fundamentally continuous measure, our model operates on a density function of age. For application to this developing country data set, the 5-year binned age distribution data was further converted to a continuous function of age by fitting a locally weighted linear regression line. The resulting smooth function for the age distribution was then scaled to integrate to one to create a density. While a simpler approach would have been to assume the population is uniform within an age bin, this assumption is clearly violated for older age groups where the risk of death is largest and most variable. These facts motivated our more complex approach, which is detailed further in Appendix C.1.

4.3 Methods

We propose a joint model for seroprevalence and IFR using a Bayesian hierarchical framework. Our approach includes two submodels to combine the many data sources shown in Figure 4.4. First, we use a logistic regression model to model the serology study data as a function of age. To account for uncertainty in the test characteristics, we use binomial models for the positive and negative controls to model sensitivity and specificity. Second, we model the fatality data using a Poisson regression as a function of age and estimated age-specific seroprevalence. Together, these submodels allow for location- and age-specific seroprevalence and IFR estimates that fully account for the uncertainty of each data source.

4.3.1 Modeling seroprevalence

We seek to model seroprevalence as a continuous function of age from the serology study data that is only reported for age bins. We assume the observed data is informative about the

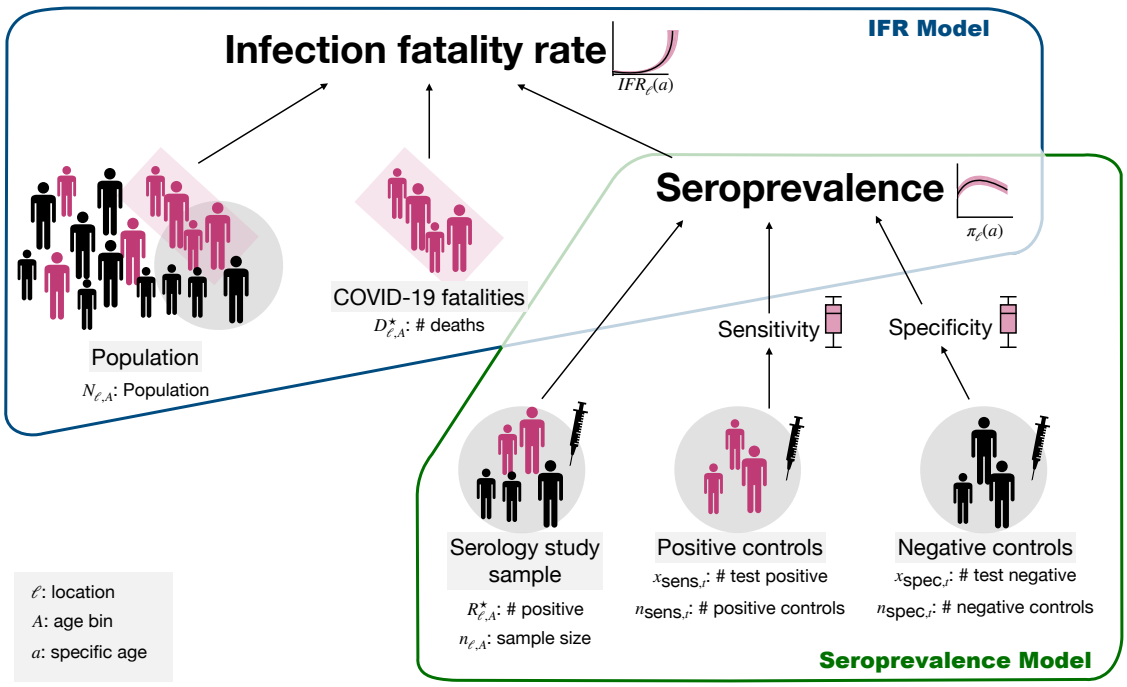


Figure 4.4: The relationship between our data sources and the seroprevalence and IFR functions of interest. Note, this represents a single location and a single age bin within this location.

(population weighted) average seroprevalence within the respective age bin. Specifically, we define this average seroprevalence as the integral of the continuous seroprevalence function with respect to the population's age distribution over the bin. In the rest of this section, we describe our model for the continuous seroprevalence function and the test validation data.

Modeling observed serology data

Let $\pi_\ell(a)$ represent the unknown seroprevalence rate for location ℓ and age a, one of our primary parameters of interest. We define $b \in \{1, \ldots, B^R_\ell\}$ as indices for the serology age bins at location ℓ. Let $A^R_{\ell,b}$ denote an age bin of the serology study at location ℓ, $R^\star_{\ell,A^R_{\ell,b}}$ denote the number of individuals who tested positive in age bin $A^R_{\ell,b}$ in study ℓ, and $n_{\ell,A^R_{\ell,b}}$ denote the number of individuals tested in age bin $A^R_{\ell,b}$ at location ℓ.

We model the number of people who test positive as coming from a binomial distribution

$$R^\star_{\ell,A^R_{\ell,b}} \sim \text{Binomial}(n_{\ell,A^R_{\ell,b}}, p_{\ell,A^R_{\ell,b}}), \tag{4.1}$$

where the average test positivity in that age bin is represented by $p_{\ell,A^R_{\ell,b}}$. The test positivity represents the proportion of individuals' tests we expect to be positive, while the seroprevalence represents the proportion of individuals with COVID-19 antibodies (i.e., are seropositive). Utilizing a binomial distribution assumes the tests are independent of each other, which aligns with the study inclusion criteria of Levin et al. (2022) that required samples to be representative.

We model the average positivity for age bin $A^R_{\ell,b}$ at location ℓ, $p_{\ell,A^R_{\ell,b}}$, as the integral of a continuous age-specific positivity function, $p_\ell(a)$, which is integrated with respect to the location ℓ's population age distribution. Mathematically this can be expressed

$$p_{\ell,A^R_{\ell,b}} = \int_{A^R_{\ell,b}} p_\ell(a) \frac{f_\ell(a)}{\int_{A^R_{\ell,b}} f_\ell(x)dx} da, \tag{4.2}$$

where $f_\ell(a)$ denotes the population age density at location ℓ evaluated at age a. The age distribution is normalized by the total population in age bin $A^R_{\ell,b}$, so $p_{\ell,A^R_{\ell,b}}$ is a weighted average positivity rate, weighted by the relative population within the interval. Details for defining $f_\ell(a)$ are given in Appendix C.1.

Since serology tests are not perfectly accurate, false positive results and false negative results are expected, and the frequency of these depends on the true seroprevalence and characteristics of the test (Gelman and Carpenter, 2020). For the test assay t_ℓ used at location ℓ, the relationship between the test positivity rate and true seroprevalence is

$$p_\ell(a) = \pi_\ell(a)\text{sens}_{t_\ell} + (1 - \pi_\ell(a))(1 - \text{spec}_{t_\ell}), \tag{4.3}$$

where sens_{t_ℓ} and spec_{t_ℓ} represent the sensitivity and specificity of the test, respectively. The first term in $p_\ell(a)$ equals the proportion of the population correctly identified as seropositive and the second term is the proportion of the population incorrectly identified as seropositive.

Seroprevalence as a continuous function of age

We model seroprevalence for location ℓ and age a, $\pi_\ell(a)$, as a linear function of covariates on the logit scale:

$$\text{logit}\left(\pi_\ell(a)\right) = \gamma_{\ell,0} + \boldsymbol{z}'_{\ell,a}\boldsymbol{\gamma}_\ell, \tag{4.4}$$

where $\gamma_{\ell,0}$ is an intercept, $\boldsymbol{z}_{\ell,a}$ is a p-dimensional vector of covariates, and $\boldsymbol{\gamma}_\ell = (\gamma_{\ell,1}, \ldots, \gamma_{\ell,p})'$ is a p-dimensional vector of coefficients. Because seroprevalence is a proportion, we use the logit link to constrain $\pi_\ell(a)$ to be between zero and one. We are specifically interested in seroprevalence as a function of age and, therefore, specify $\boldsymbol{z}'_{\ell,a}$ to be a natural spline of age. Yet, our framework is general and can accommodate any covariates of interests. For example, population density in a location or an indicator for whether the location underwent government delegated stay-at-home orders could be added as covariates. These may be informative predictors if seroprevalence is higher in denser populations or mobility restrictions prevented transmission leading to lower seroprevalence. In our application, we lacked rich covariate information on all locations, so $\boldsymbol{z}'_{\ell,a}$ was specified to be a natural spline of age (details given in Section 4.3.3).

Modeling test validation data

At the onset of the pandemic, serological test assays were developed quickly with limited controls. Therefore, we consider sensitivity and specificity to be unknown parameters and explicitly model the lab validation data as done in Gelman and Carpenter (2020), Stringhini et al. (2020), and Larremore et al. (2022). Let $n_{\text{sens},t}$ denote the number of positive controls and $n_{\text{spec},t}$ denote the number of negative controls for test assay t. Further, let $x_{\text{sens},t}$ denote the number of positive controls that tested positive and $x_{\text{spec},t}$ denote the number of negative controls that tested negative. Naive point estimates of the test sensitivity and specificity are then $x_{\text{sens},t}/n_{\text{sens},t}$ and $x_{\text{spec},t}/n_{\text{spec},t}$, respectively. Rather than taking the test characteristics as known, we estimate spec_t and sens_t

along with their uncertainty from the lab validation data. The model is

$$x_{\text{sens},t} \sim \text{Binomial}(n_{\text{sens},t}, \text{sens}_t), \tag{4.5}$$

$$x_{\text{spec},t} \sim \text{Binomial}(n_{\text{spec},t}, \text{spec}_t). \tag{4.6}$$

We assume each control sample tested is a Bernoulli trial, where positive controls test positive with probability equal to the sensitivity of the assay, and negative controls test negative with probability equal to the specificity of the assay.

When a test is used at multiple study locations, the sensitivity and specificity are assumed to be the same in each study. However, we do not assume any relationship between the sensitivity and specificity across tests. One reason for this is that antibody tests are designed to target different types of antibodies, such as IgG versus IgM, as well as rely on binding different regions of SARS-CoV-2 (Jacofsky et al., 2020). Therefore, while we pool information across locations that use the same test, we do not pool information about the sensitivity and specificity across tests.

Seroprevalence, sensitivity, and specificity jointly determine the positivity rate. With small sample sizes, sensitivity and specificity are weakly identified. We have prior information on the target test characteristics for antibody tests, so we use informative priors to improve identifiability. The priors selected for the developing countries data set specifically are discussed in Section 4.3.3.

4.3.2 Modeling IFR

The model for IFR is similar to the model for seroprevalence as we model IFR as a function of age and other covariates. We link the continuous age-specific IFR parameter to the data by calculating the average IFR at the observed age bins via integration.

Modeling observed death counts

Let $A^D_{\ell,b}$ denote an age group for which deaths caused by COVID-19 are recorded, $b \in \{1, \ldots, B^D_\ell\}$. We model the $D^\star_{\ell,A^D_{\ell,b}}$ deaths at location ℓ and age bin $A^D_{\ell,b}$ as following a Poisson distribution:

$$D^\star_{\ell,A^D_{\ell,b}} \sim \text{Poisson}(N_{\ell,A^D_{\ell,b}} \Lambda_{\ell,A^D_{\ell,b}}), \tag{4.7}$$

where $N_{\ell,A^D_{\ell,b}}$ denotes the population at location ℓ in age group $A^D_{\ell,b}$, and $\Lambda_{\ell,A^D_{\ell,b}}$ represents the proportion of individuals in location ℓ and age group $A^D_{\ell,b}$ expected to die from the disease. Thus, the product $N_{\ell,A^D_{\ell,b}} \Lambda_{\ell,A^D_{\ell,b}}$ represents the number of individuals in that location and age bin expected to die from the disease. Similar to O'Driscoll et al. (2021), we chose to model deaths with a Poisson distribution rather than a binomial distribution because COVID-19 deaths are a relatively rare event. Note, deaths are modeled as a census of the population, assuming deaths are accurately reported.

The death rate for age a is given by $\pi_\ell(a) \times \text{IFR}_\ell(a)$: the probability of infection times the probability of death given infection at age a. Similar to the average positivity in (4.2), we define the population weighted average death rate for age bin $A^D_{\ell,b}$, $\Lambda_{\ell,A^D_{\ell,b}}$, as

$$\Lambda_{\ell,A^D_{\ell,b}} = \int_{A^D_{\ell,b}} \pi_\ell(a) \times \text{IFR}_\ell(a) \frac{f_\ell(a)}{\int_{A^D_{\ell,b}} f_\ell(x)dx} da. \tag{4.8}$$

There are infinitely many possible combinations of seroprevalence and IFR that can result in the same $\Lambda_{\ell,A^D_{\ell,b}}$, so we recommend an informative prior for at least one of these when the serology data is limited. In our analysis, we apply informative priors to the parameters in the serology model as described in Section 4.3.3.

IFR as a continuous function of age

Let $\text{IFR}_\ell(a)$ be the IFR at location ℓ and age a. We define $\log(\text{IFR}_\ell(a))$ as a function of a natural spline of age, $\boldsymbol{x}_{\ell,a} \in \mathbb{R}^q$ and associated location-specific coefficients $\boldsymbol{\beta}_\ell = (\beta_{\ell,1}, \ldots, \beta_{\ell,q})'$

$$\log\left(\text{IFR}_\ell(a)\right) = \beta_{\ell,0} + \boldsymbol{x}'_{\ell,a}\boldsymbol{\beta}_\ell, \tag{4.9}$$

where $\beta_{\ell,0}$ represents the study location-specific intercept. This framework is general and can incorporate additional covariates.

Since we expect the IFR to be generally similar across study locations, we use hierarchical priors to pool information about $\boldsymbol{\beta}_\ell$ across locations. We further expect overall IFR levels to be more similar for locations within the same country compared to study locations in different countries, so we model the IFR intercept with a common country effect. This is particularly relevant for developing countries as the death registration systems can differ dramatically between countries (Karanikolos et al., 2020). Thus, we define priors on the coefficients as

$$\beta_{\ell,0} \sim \mathcal{N}(\beta_{\text{global},0} + \beta_{\text{country},c_\ell}, \sigma_0^2), \tag{4.10}$$

$$\beta_{\ell,i} \sim \mathcal{N}(\beta_{\text{global},i}, \sigma_i^2), \qquad \text{for i } \in \{1, \ldots, q\}. \tag{4.11}$$

The intercept ($\beta_{\ell,0}$) is informed by an overall global intercept parameter ($\beta_{\text{global},0}$) as well as a country specific intercept ($\beta_{\text{country},c_\ell}$). The other covariates, which control the shape of the IFR function, pool information at the global level using $\beta_{\text{global},i}$, but not at the country level.

4.3.3 Covariate and prior distribution selection

The methods have been discussed rather generally up until this point as they are applicable to any disease setting. However, we now focus on modeling choices made to account for the specific nuances of our application: COVID-19 in developing countries from June 2020 to March 2021. Specifically, we discuss the covariates and the prior distributions in this section.

Covariates

For modeling IFR as a function of age, we used natural cubic splines with boundary knots at 0 and 80, and internal knots at 10 and 60 as the covariates, $\boldsymbol{x}_{\ell,a}$. Thus, log IFR is modeled as cubic functions of age between 0 and 80 and is constrained to be log-linear above age 80. The knots were selected based on expert opinion and prior literature (Cai et al., 2021; COVID-19 Forecasting Team, 2022). Similarly, we used natural cubic splines for the seroprevalence covariates, $\boldsymbol{z}_{\ell,a}$, with an internal knot at 60 and the boundary knots at 10 and 80. We decreased the number of internal knots in the natural spline for the serology model because each location in our data set contains only three to eight age bins, with over half the observations having five bins.

Prior specification

Estimation of our model is straightforward in a Bayesian context. Similar to Gelman and Carpenter (2020), we set independent, informative beta priors for the sensitivity and specificity of each of the $T = 13$ tests:

$$\text{sens}_t \sim \text{Beta}(10, 1), \qquad \text{for } t \in \{1, \ldots, 13\}, \tag{4.12}$$

$$\text{spec}_t \sim \text{Beta}(50, 1), \qquad \text{for } t \in \{1, \ldots, 13\}. \tag{4.13}$$

Seroprevalence assays are generally designed to ensure a high specificity. For example, the Centers for Disease Control and Prevention recommended a specificity of 0.995 for COVID-19 antibody tests (Centers for Disease Control and Prevention, 2020). The prior specified for specificity has a 0.9 probability the specificity is greater than 0.95. Because tests are designed to prioritize specificity, the sensitivity is typically more variable, reflected by our more dispersed prior with 0.9 probability the sensitivity is greater than 0.79.

We utilize informative priors on the coefficients for the seroprevalence function at each location based on prior studies. Specifically, we specify independent normal distributions centering the

non-intercept coefficients about zero:

$$\gamma_{\ell,0} \sim \mathcal{N}(-1, 1.5), \tag{4.14}$$

$$\gamma_{\ell,j} \sim \mathcal{N}(0, 0.05), \qquad \text{for } j \in \{1, 2\}. \tag{4.15}$$

Here $\gamma_{\ell,0}$ denotes the intercept at location ℓ and $\gamma_{\ell,j}$ for $j > 0$ are the coefficients associated with the natural spline, controlling the shape of the seroprevalence function. Because seroprevalence is modeled as a logit-linear function of the covariates, the prior on the intercept is a right skewed distribution with a median of about 0.27 and a 95$^{\text{th}}$ quantile of about 0.81. The more informative priors on the spline coefficients ($j > 0$) have a prior mean of zero, corresponding to a seroprevalence function that does not vary with age, which has been noted in the literature. Levin et al. (2022) found the ratio of seroprevalence for those age 60+ compared to 40-59 was not significantly different from one in most locations. Esteve et al. (2020) also found countries where multigenerational households are common, such as developing countries (Ruggles and Heggeness, 2008), may not be able to shield their older population from COVID-19.

We use weakly-informative priors for the age spline coefficient parameters associated with IFR:

$$\begin{aligned}
\beta_{\text{global},i} &\sim \mathcal{N}(0, 5), && \text{for } i \in \{0, \ldots, 3\}, \\
\beta_{\text{country},c_\ell} &\sim \mathcal{N}(0, \sigma_{\text{country}}), && \\
\sigma_i &\sim \text{half-normal}(0, 2), && \text{for } i \in \{0, \ldots, 3\}, \\
\sigma_{\text{country}} &\sim \text{half-normal}(0, 2). &&
\end{aligned} \tag{4.16}$$

Note, IFR is modeled on the log scale, so these priors are relatively non-informative. For example, with a prior probability of 0.8, the global intercept, $\beta_{\text{global},0}$ is between -6.4 and 6.4, corresponding to a multiplicative effect between 0.002 and 606.5.

4.3.4 Estimation and Inference

Inference was based on the joint posterior distribution of the IFR parameters $(\{\beta_{\ell,i}\}, \{\beta_{\text{global},i}\}, \{\beta_{\text{country},c}\}, \{\sigma_i\}, \sigma_{\text{country}})$ and serology parameters $(\{\gamma_{\ell,j}\}, \{\text{sens}_t\}, \{\text{spec}_t\})$ given the death, serology, and antibody test validation data $(\{n_{\ell,A^R_{\ell,b}}\}, \{R^\star_{\ell,A^R_{\ell,b}}\}, \{D^\star_{\ell,A^D_{\ell,b}}\}, \{x_{\text{sens},t}\}, \{x_{\text{spec},t}\}, \{n_{\text{sens},t}\}, \{n_{\text{spec},t}\})$. The joint posterior is not available in closed form, so we obtain posterior draws through a Markov chain Monte Carlo (MCMC) algorithm. We approximated the integrals in (4.2) and (4.8), which calculate the average positivity and death rates for the age bins, using trapezoidal Reimann sums scaling the mesh appropriately to increase computation speed, while maintaining sufficient accuracy. The MCMC algorithm was implemented in version 2.21.5 of RStan (Stan Development Team, 2022).

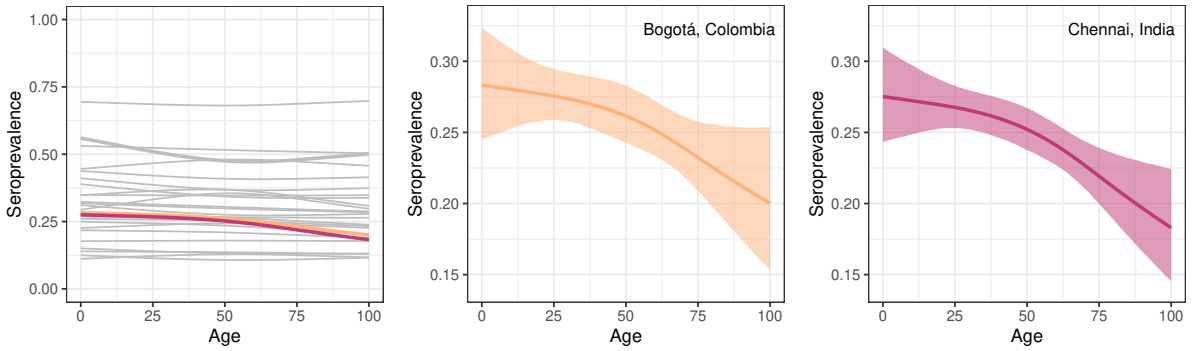
We use the $\gamma_{\ell,j}$ and $\beta_{\ell,j}$ posterior draws to calculate the location-specific IFR and seroprevalence functions of age as in (4.9) and (4.4). These allow for age-specific comparisons, rather than age bin level comparisons. We summarized the parameters of interest in terms of posterior means and 95% credible intervals. Note, we do not interpret $\beta_{\text{global},i}$ as a "global average" in the geographic sense since the study locations are not a representative sample of the world as a whole. Similarly, $\beta_{\text{country},c_\ell}$ is not interpreted as the adjustment for an entire country since the study locations are not a representative sample of locations within a country.

4.4 Analysis of the developing countries data

To sample from the posterior distribution, we ran three chains with a burn-in of 2500 iterations and the subsequent 3000 samples were retained from each chain. Code is available at https://github.com/pughs/covid-ifr. Convergence was assessed via traceplots, effective sample size, and $\widehat{R}$ (Gelman et al., 1995). Each parameter had an effective sample size greater than 1000 and $\widehat{R}$ within 0.0042 of one (see Appendix C.2).

Figure 4.5 shows the posterior mean seroprevalence curves for each study location. While most curves do not show age-specific trends, a few show clear deviations, notably, Bogotá, Colombia and Chennai, India. The 95% credible interval of the seroprevalence function did not contain

a flat trend for both of these locations. Both showed a decreased seroprevalence for the oldest individuals. While most locations showed roughly flat seroprevalence, the average seroprevalence varies substantially across locations. This is unsurprising as we expect heterogeneity due to factors such as when the seroprevalence study was conducted and when waves of COVID-19 arose in each location.

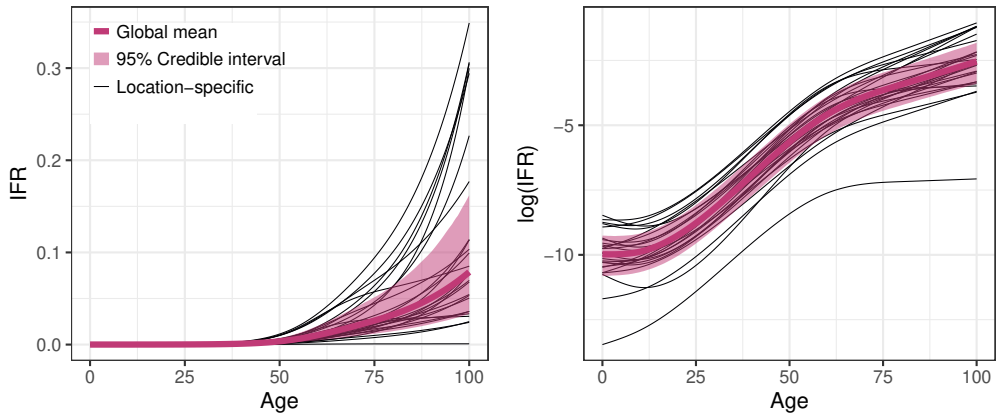


(a) **(b)** **(c)**

Figure 4.5: (a) Posterior mean seroprevalence curve for each location, colored to emphasize those locations where seroprevalence varies by age. Panels (b) and (c) show the posterior mean and 95% credible interval for the locations highlighted in (a). All studies were conducted between June 2020 and March 2021.

Figure 4.6 shows the global IFR curve as well as the posterior mean IFR curve for each location, where we again use the term “global" in the modeling rather than geographic sense. Viewing IFR on the log-scale (Figure 4.6(b)), we can see the locations generally follow a s-curve with IFR increasing less quickly or even decreasing with age for children and older individuals. The individual locations show variation in this trend in terms of their intercepts and shapes. IFR posterior mean increased with age for children in some locations but decreased with age in others, although not significantly. Similarly, the IFR flattened out more or less for older individuals depending on the location. These results highlight the need for a model that can estimate location-specific, non-linear IFR curves.

Figure 4.7 focuses on the seroprevalence and IFR functions for Cuiabá, Brazil and Chennai, India. The posterior mean and 95% credible intervals are compared to naive estimates at the age bin level. For seroprevalence, the Rogan-Gladen estimate is shown, which adjusts the raw prevalence



(a) **(b)**

Figure 4.6: Posterior mean of the global IFR curve with 95% credible interval for the mean. Panel (a) shows the IFR on its original scale and panel (b) shows the log-scale. The posterior means of location-specific IFR curves are shown in black.

by estimates of the test sensitivity and specificity (Rogan and Gladen, 1978). For IFR, the naive estimate represents the death rate divided by the Rogan-Gladen seroprevalence estimate, assuming the death rate and seroprevalence are constant within an age bin.

We can see the strength of modeling the seroprevalence and IFR as smooth functions of age rather than modeling each age bin separately. We do not expect seroprevalence or IFR to largely jump across consecutive age bins as is observed in the naive estimates, but rather expect some smooth underlying function. For example, the 50-59 serology age bin in Cuiabá had a smaller sample size relative to the surrounding bins and showed an unusual jump in seroprevalence. Rather than fitting a seroprevalence curve that goes through the naive seroprevalence estimate, the model pools information across adjacent age bins. However, the seroprevalence curve can overcome the mean zero prior when the data suggests it, as is shown for Chennai, India (see panel (b) of Figure 4.7). Additionally, by sharing information across age bins, the uncertainty of our seroprevalence function is smaller than the uncertainty of the naive point estimates in most cases.

Figure 4.7 again shows the benefits of estimating IFR as a continuous function of age. In Cuiabá, the death data is aggregated into less than 10 bins. By estimating a smooth, underlying function of IFR, we are able to estimate IFR for each specific age, not just averages for the age bins the data is available at. In Chennai, the IFR function passes through roughly the center of the naive

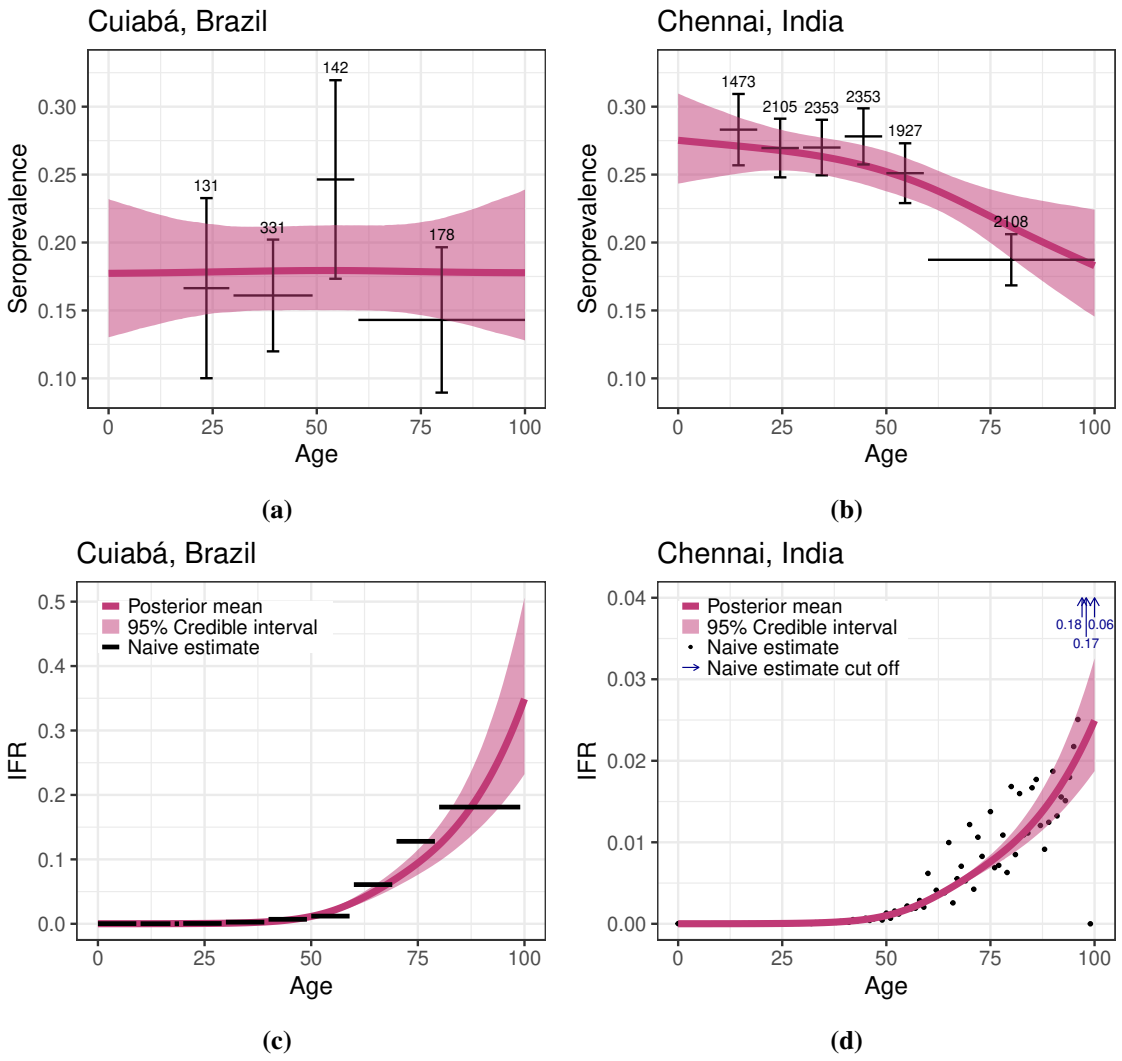


Figure 4.7: Panels (a) and (b) show the posterior mean and 95% credible interval for (a) Cuiabá, Brazil's and (b) Chennai, India's seroprevalence curves, annotated with the seroprevalence study sample size for each bin. The Rogan-Gladen estimators and approximate 95% confidence intervals are shown as error bars. Panels (c) and (d) show the posterior mean and 95% credible interval for the IFR curves. Naive estimates for the IFR are shown as points when single year age bins are available for the death data and as black lines when the death data is binned.

estimates, estimating a smooth function given the extremely noisy single year naive estimates (see Figure 4.7(d)). Seroprevalence and IFR functions for each of the 26 locations in our analysis are included in Appendices C.4 and C.5.

Figure 4.8 provides estimates and credible intervals for IFR at age sixty at all locations. Because we model IFR as a continuous function of age, we can estimate IFR for this specific age rather than being restricted to discussing the average IFR over the age bin at which the data was

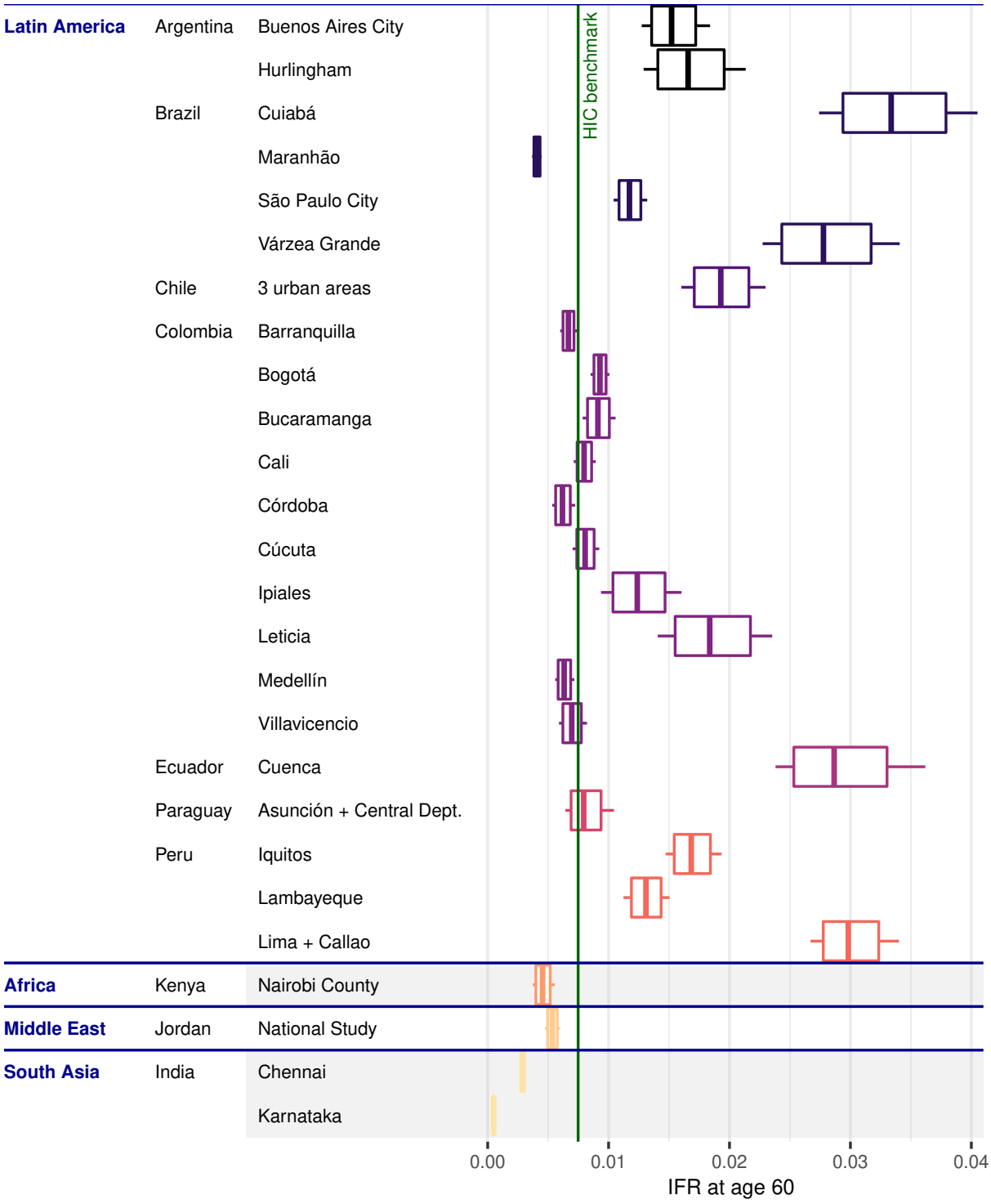


Figure 4.8: IFR at age 60 for each location. Whiskers indicate 95% credible intervals, boxes indicate 80% credible intervals, and the center line indicates the posterior median of the posterior distribution for age-60 IFR. The high-income countries (HIC) benchmark from Levin et al. (2020) is shown as a vertical line. Locations with a grey background have less than 50% of deaths well certified (Fullman et al., 2017).

reported. These location-specific estimates are compared to the estimated IFR of high-income countries at age 60 given in Levin et al. (2020). Note, this high-income country benchmark was deemed an appropriate comparator for the studies in this data set by Levin et al. (2022) in terms of the timing and the inclusion criteria of the seroprevalence studies and seroreversion/death reporting adjustments made. We can see most locations have an IFR near or above the high-income countries estimate at age 60. Of the eight study locations below the benchmark, three have less than 50% of all deaths well certified, meaning less than 50% of deaths in that country are registered to a specific, well-defined cause (Fullman et al., 2017). These three are Nairobi County, Kenya, Chennai, India, and Karnataka, India. Combining this result with the findings in Knutson et al. (2022), which found COVID-19 deaths are likely to be undercounted in many developing countries, suggests the estimated IFR may be below the high-income country benchmark because of an under-reporting of COVID-19 deaths, rather than an actual decrease in the risk of COVID-19 at these locations. For example, da Silva et al. (2020) found the increased number of natural cause deaths for those over the age of 60 during the pandemic compared to before was around two times the number of reported COVID-19 deaths in Maranhão, Brazil. If the increased number of natural cause deaths could be attributed to unreported COVID-19 deaths, the IFR estimate would approximately double and exceed the high-income country benchmark.

If we view the number of reported deaths in our study locations as minimums, with death reports either accurate or an underreport, then our IFR estimates can be interpreted as lower-bounds for the true IFR. In this case, COVID-19 has been catastrophic for some developing country locations like Cuiabá, Brazil and Lima, Peru that have an IFR at least three times that of high-income countries for those aged sixty.

Figure 4.9 shows the effect of the age distribution on the population IFR. In Buenos Aires City, Argentina and Cali, Colombia the population IFR was significantly higher when based on their own age distributions compared to the standardized age distribution. This is because the average age was smaller for the standardized distribution than that for their specific locations. In the case of Buenos Aires City, the population IFR estimate was considerably closer to the other locations' when controlling for the age distribution. We see the opposite at locations like Maranhão, Nairobi County, and the Jordan national study, where the average age was smaller for the specific locations than the standardized distribution. The standardized distribution estimates were closer to the other locations for these locations as well. By estimating seroprevalence and IFR as functions of age, we are able to separate out the effects of the seroprevalence, IFR, and age distribution on the population level estimate.

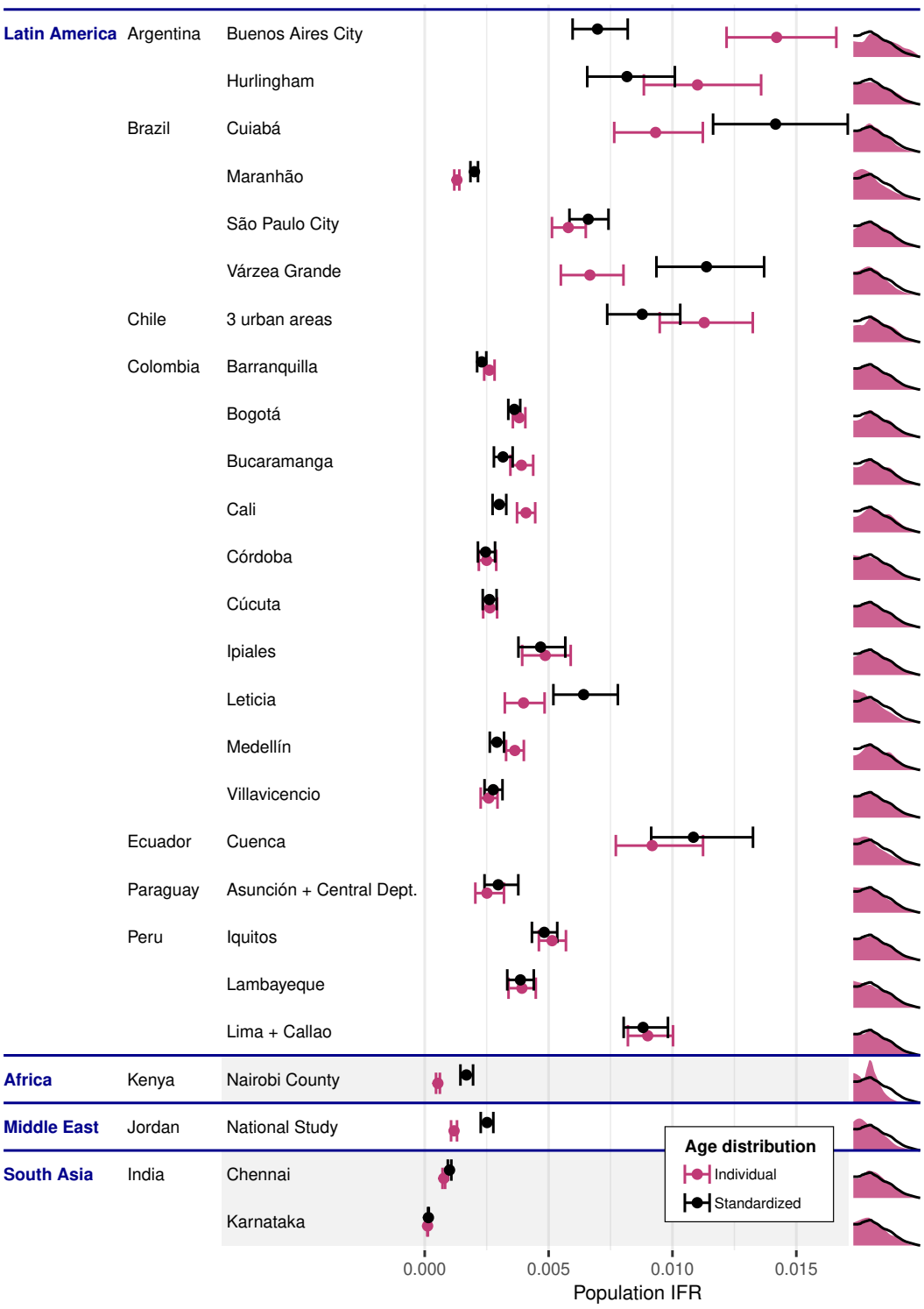


Figure 4.9: The posterior mean population IFR (the point) with a 95% credible interval. Estimates are based on the location-specific age distribution ("Individual") or based on the median age distribution across our study locations ("Standardized"). Locations with a grey background have less than 50% of deaths well certified (Fullman et al., 2017). The age distribution for each location (filled) compared to the standardized age distribution (the line) is shown on the right.

4.5 Discussion

In this work, we introduced an adaptable Bayesian hierarchical model that enables estimation of location- and age-specific IFR and seroprevalence curves, which accurately reflect uncertainty due to limited data and uncertainty in the underlying test characteristics. The Bayesian framework presents a natural framework for propagating test characteristic uncertainty and allowed us to incorporate prior knowledge to improve identifiability.

Apply this model to our developing countries dataset, we found seroprevalence was not dependent on age for most locations, while IFR increased non-linearly with age. The IFR curves were diverse, emphasizing the need for location-specific estimates. IFR for those aged 60 in our study locations was near or above the age 60 high-income country benchmark for most locations. Finally, we showed the importance of considering the location's age distribution by comparing the population IFR using the location-specific age distribution to the population IFR using a standardized age distribution.

While our model was tailored to the developing countries COVID-19 data set, the model is applicable to estimating IFR for any disease using seroprevalence studies and reported deaths. Specifically, this methodology may be particularly useful for modeling novel emerging pathogens. During the initial surge of COVID-19, one of the primary issues noted in public communications was the challenge of incorporating uncertainty into estimates (see, e.g., Dean (2020); Foad (2021)). Our model may be useful for estimating IFR for future novel pathogens, when it is essential to maintain sufficient model flexibility, such as avoiding assumptions about the relationship of mortality with age, before much is known about the disease.

Our model is flexible and can meet the unique challenges and opportunities of future applications. For example, additional covariates beyond age could be added to the vector of covariates for serology, $\boldsymbol{z}_{\ell,a}$, or deaths, $\boldsymbol{x}_{\ell,a}$. While we chose not to pool information across the seroprevalence coefficients in our application, information could be pooled following the hierarchical modeling framework presented for the IFR coefficients.

Chapter 5

Conclusion

The COVID-19 pandemic quickly and dramatically changed life for people worldwide. At the start of the pandemic, it was impossible to track all the COVID-19 cases, especially as viral tests were not initially available. To estimate the number of infections and consequently the risk of COVID-19, antibody tests were rapidly developed. These tests had diverse and unknown test characteristic, and the studies employing these tests had varying sample sizes and age-specific granularity. Thus, we required methods to make inference and policy decisions based on limited and disparate datasets.

In this dissertation we present methods to fill three gaps in the literature. First, there was a lack of guidance on how to establish cutoffs for antibody tests when limited samples with known disease status were available. Second, there was no cohesive model to estimate test characteristics, seroprevalence, and IFR to account for the many sources of uncertainty. Third, building off the previous gap, there were no cohesive methods that treated seroprevalence and IFR as continuous functions of age while accounting for the three data sources.

In Chapter 2 we proposed using extreme value theory to select a cutoff for antibody tests. We compared the extreme value method to common, existing methods for selecting cutoffs through a simulation study and data analysis. We also evaluated how the choice of target specificity impacts test characteristics. We demonstrated that the extreme value-based method was superior for a high target specificity, while the empirical quantile approach was superior for a lower target specificity. We also found that the common normal distribution-based methods could have considerable bias. A lower target specificity was ideal for estimating seroprevalence, while the higher target specificity was ideal for overall test accuracy when the prevalence was low. These findings can be applied to selecting cutoffs for any diagnostic test with limited lab validation data.

In Chapter 3 we developed a cohesive model to estimate seroprevalence and IFR while fully accounting for test characteristic uncertainty. Our model was able to synthesize all available data,

regardless of whether age-specific or fatality data were available. Applying our model to a COVID-19 developing countries dataset, we found the IFR behaved similar across countries with IFR increasing with age. Seroprevalence was relatively independent of age, so developing countries were not able to shelter older individuals as well as high-income countries. We compared the results of our model to common Rogan-Gladen-based estimates (Rogan and Gladen, 1978). We found the sources of uncertainty omitted by the Rogan-Gladen based intervals were non-negligible, with the credible intervals for IFR wider than the corresponding confidence intervals for the vast majority of age bins. We further demonstrated the improvements in the estimates that come from modeling the data sources simultaneously rather than independently.

In Chapter 4 we extended the model from Chapter 3 to estimate serology and IFR as continuous functions of age, while still jointly estimating test characteristics, seroprevalence, and IFR. This allowed for comparison of the seroprevalence or IFR for specific ages, rather than estimates at the age bin level. We verified seroprevalence was roughly independent of age when considering it as a continuous function. We additionally found evidence that IFR was not log-linear, which provides further motivation for using the nonparametric approach that we developed in this thesis. This model could be applied to future emerging diseases as it is flexible enough to accommodate mismatched age bins in the data and does not make strong assumptions about the relationship between seroprevalence or IFR and age.

5.1 Impact

The methods in this dissertation are motivated by the testing and data challenges faced in the early stages of the COVID-19 pandemic, but they are broadly applicable. Seroprevalence studies utilizing antibody tests are used for many diseases including HIV (Brookmeyer and Gail, 1988; Sakarovitch et al., 2007; Stengaard et al., 2021), H1N1 (Wong et al., 2013; Zimmer et al., 2010), and MERS-CoV (Degnah et al., 2020; Ryu et al., 2019). The methods we have developed can be applied to similar emerging diseases in the future. The challenges of diverse test characteristics, limited sample sizes, and variable data availability are most severe for the early stages of emerging

infectious diseases, but the methods we present in this dissertation can be applied to diseases at any stage.

5.2 Future work

There are many possible directions for future work. We established positive and negative controls in Chapter 2 using a neutralization assay test. We could explore the impact of uncertainty is the lab validation controls, due to imperfect sensitivity and specificity, on the cutoff methods. In Chapters 3 and 4 we could explore combining multiple sources of death data. In our analysis, we used the confirmed and suspected cases. We could explore leveraging both the confirmed and suspected cases and the excessive mortality: the difference between the total number of deaths recorded for the time period of interest and the total number of deaths recorded for a commensurate time period prior to 2020. Future work could also include modeling the timing between infection, fatality, and death reporting to estimate the true number of deaths corresponding to the seroprevalence study time period (Jersakova et al., 2022; O'Driscoll et al., 2021; Perez-Saez et al., 2021; Seaman et al., 2022).

Bibliography

Alimohamadi, Y., Tola, H. H., Abbasi-Ghahramanloo, A., Janani, M., and Sepandi, M. (2021). Case fatality rate of COVID-19: A systematic review and meta-analysis. *Journal of Preventive Medicine and Hygiene*, 62(2):E311.

Arora, R. K., Joseph, A., Van Wyk, J., Rocco, S., Atmaja, A., May, E., Yan, T., Bobrovitz, N., Chevrier, J., Cheng, M. P., et al. (2021). Serotracker: A global SARS-CoV-2 seroprevalence dashboard. *The Lancet Infectious Diseases*, 21(4):e75–e76.

Axfors, C. and Ioannidis, J. (2022). Infection fatality rate of COVID-19 in community-dwelling elderly populations. *European Journal of Epidemiology*, pages 1–15.

Balkema, A. A. and De Haan, L. (1974). Residual life time at great age. *The Annals of Probability*, 2(5):792–804.

Beirlant, J., Goegebeur, Y., Segers, J., and Teugels, J. L. (2006). *Statistics of extremes: Theory and applications*. John Wiley & Sons.

Bewley, K. R., Coombes, N. S., Gagnon, L., McInroy, L., Baker, N., Shaik, I., St-Jean, J. R., St-Amant, N., Buttigieg, K. R., Humphries, H. E., et al. (2021). Quantification of SARS-CoV-2 neutralizing antibody by wild-type plaque reduction neutralization, microneutralization and pseudotyped virus neutralization assays. *Nature Protocols*, 16(6):3114–3140.

Blaker, H. (2000). Confidence curves and improved exact confidence intervals for discrete distributions. *Canadian Journal of Statistics*, 28(4):783–798.

Blostein, M. and Miljkovic, T. (2019). On modeling left-truncated loss data using mixtures of distributions. *Insurance: Mathematics and Economics*, 85:35–46.

Bottomley, C., Otiende, M., Uyoga, S., Gallagher, K., Kagucia, E., Etyang, A., Mugo, D., Gitonga, J., Karanja, H., Nyagwange, J., et al. (2021). Quantifying previous SARS-CoV-2 infection through mixture modelling of antibody levels. *Nature Communications*, 12(1):6196.

Bouman, J. A., Riou, J., Bonhoeffer, S., and Regoes, R. R. (2021). Estimating the cumulative incidence of SARS-CoV-2 with imperfect serological tests: Exploiting cutoff-free approaches. *PLoS Computational Biology*, 17(2):e1008728.

Brazeau, N. F., Verity, R., Jenks, S., Fu, H., Whittaker, C., Winskill, P., Dorigatti, I., Walker, P. G., Riley, S., Schnekenberg, R. P., et al. (2022). Estimating the COVID-19 infection fatality ratio accounting for seroreversion using statistical modelling. *Communications Medicine*, 2(1):1–13.

Brookmeyer, R. and Gail, M. H. (1988). A method for obtaining short-term projections and lower bounds on the size of the AIDS epidemic. *Journal of the American Statistical Association*, 83(402):301–308.

Cai, R., Novosad, P., Tandel, V., Asher, S., and Malani, A. (2021). Representative estimates of COVID-19 infection fatality rates from four locations in India: Cross-sectional study. *BMJ Open*, 11(10):e050920.

Campbell, H., de Valpine, P., Maxwell, L., de Jong, V. M. T., Debray, T. P. A., Jaenisch, T., and Gustafson, P. (2022). Bayesian adjustment for preferential testing in estimating infection fatality rates, as motivated by the COVID-19 pandemic. *The Annals of Applied Statistics*, 16(1):436 – 459.

Campbell, H. and Gustafson, P. (2021). Inferring the COVID-19 infection fatality rate in the community-dwelling population: A simple Bayesian evidence synthesis of seroprevalence study data and imprecise mortality data. *Epidemiology & Infection*, 149:1 – 14.

Castellanos, M. E. and Cabras, S. (2007). A default Bayesian procedure for the generalized Pareto distribution. *Journal of Statistical Planning and Inference*, 137(2):473–483.

Centers for Disease Control and Prevention (2020). Interim guidelines for COVID-19 antibody testing. https://www.cdc.gov/coronavirus/2019-ncov/lab/resources/antibody-tests-guidelines.html. Published May 23, 2020. Updated August 1, 2020. Accessed January 7, 2021.

Cohen, B., Doblas, D., and Andrews, N. (2008). Comparison of plaque reduction neutralisation test (PRNT) and measles virus-specific IgG ELISA for assessing immunogenicity of measles vaccination. *Vaccine*, 26(50):6392–6397.

Cooley, D., Nychka, D., and Naveau, P. (2007). Bayesian spatial modeling of extreme precipitation return levels. *Journal of the American Statistical Association*, 102(479):824–840.

COVID-19 Forecasting Team (2022). Variation in the COVID-19 infection-fatality ratio by age, time, and geography during the pre-vaccine era: A systematic analysis. *The Lancet*, 399(10334):1469–1488.

Cutler, D. M. and Summers, L. H. (2020). The COVID-19 pandemic and the $16 trillion virus. *JAMA*, 324(15):1495–1496.

da Silva, A. A. M., Lima-Neto, L. G., de Maria Pedrozo e Silva de Azevedo, C., da Costa, L. M. M., Martins ça, M. L. B., Barros Filho, A. K. D., Bastos Wittlin, B., de Souza, B. F., de Oliveira, B. L. C. A., de Carvalho, C. A., et al. (2020). Population-based seroprevalence of SARS-CoV-2 and the herd immunity threshold in Maranhão. *Revista de Saude Publica*, 54.

Dean, J. (2020). Experts: Acknowledge uncertainty in COVID communication. https://news.cornell.edu/stories/2020/09/experts-acknowledge-uncertainty-covid-communication. Accessed: 2023-1-04.

Degnah, A. A., Al-Amri, S. S., Hassan, A. M., Almasoud, A. S., Mousa, M., Almahboub, S. A., Alhabbab, R. Y., Mirza, A. A., Hindawi, S. I., Alharbi, N. K., et al. (2020). Seroprevalence of MERS-CoV in healthy adults in western Saudi Arabia, 2011–2016. *Journal of Infection and Public Health*, 13(5):697–703.

Devanarayan, V., Smith, W. C., Brunelle, R. L., Seger, M. E., Krug, K., and Bowsher, R. R. (2017). Recommendations for systematic statistical computation of immunogenicity cut points. *The AAPS Journal*, 19(5):1487–1498.

Drees, H., Ferreira, A., and De Haan, L. (2004). On maximum likelihood estimation of the extreme value index. *Annals of Applied Probability*, pages 1179–1201.

DuMouchel, W. H. (1983). Estimating the stable index α in order to measure tail thickness: A critique. *The Annals of Statistics*, 11(4):1019–1031.

Durán-Rosal, A. M., Carbonero, M., Gutiérrez, P. A., and Hervás-Martínez, C. (2022). A mixed distribution to fix the threshold for Peak-Over-Threshold wave height estimation. *Scientific Reports*, 12(1):17327.

Esteve, A., Permanyer, I., Boertien, D., and Vaupel, J. W. (2020). National age and coresidence patterns shape COVID-19 vulnerability. *Proceedings of the National Academy of Sciences*, 117(28):16118–16120.

Eyal, O., Olshevsky, U., Lustig, S., Paran, N., Halevy, M., Schneider, P., Zomber, G., and Fuchs, P. (2005). Development of a tissue-culture-based enzyme-immunoassay method for the quantitation of anti-vaccinia-neutralizing antibodies in human sera. *Journal of Virological Methods*, 130(1-2):15–21.

Felder, S. and Mayrhofer, T. (2022). The optimal cutoff of a diagnostic test. In *Medical Decision Making: A Health Economic Primer*, pages 173–192. Springer.

Foad, C. (2021). Embracing pandemic uncertainty in science, society and policy. https://royalsociety.org/blog/2021/07/ embracing-pandemic-uncertainty-in-science-society-and-policy/. Accessed: 2023-1-04.

Fullman, N., Barber, R. M., Abajobir, A. A., Abate, K. H., Abbafati, C., Abbas, K. M., Abd-Allah, F., Abdulkader, R. S., Abdulle, A. M., Abera, S. F., et al. (2017). Measuring progress and projecting attainment on the basis of past trends of the health-related Sustainable Development Goals in 188 countries: An analysis from the Global Burden of Disease Study 2016. *The Lancet*, 390(10100):1423–1459.

Gallichotte, E. N., Nehring, M., Young, M. C., Pugh, S., Sexton, N. R., Fitzmeyer, E., Quicke, K. M., Richardson, M., Pabilonia, K. L., Ehrhart, N., et al. (2021). Durable antibody responses in staff at two long-term care facilities, during and post SARS-CoV-2 outbreaks. *Microbiology Spectrum*, 9(1):e00224–21.

Garcia-Basteiro, A. L., Moncunill, G., Tortajada, M., Vidal, M., Guinovart, C., Jimenez, A., Santano, R., Sanz, S., Méndez, S., Llupià, A., et al. (2020). Seroprevalence of antibodies against SARS-CoV-2 among health care workers in a large Spanish reference hospital. *Nature Communications*, 11(1):3500.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995). *Bayesian Data Analysis*. Chapman and Hall/CRC.

Gelman, A. and Carpenter, B. (2020). Bayesian analysis of tests with unknown specificity and sensitivity. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 69(5):1269–1283.

Greiner, M., Pfeiffer, D., and Smith, R. D. (2000). Principles and practical application of the receiver-operating characteristic analysis for diagnostic tests. *Preventive Veterinary Medicine*, 45(1-2):23–41.

Hajian-Tilaki, K. (2018). The choice of methods in determining the optimal cut-off value for quantitative diagnostic test evaluation. *Statistical Methods in Medical Research*, 27(8):2374–2383.

Hitchings, M. D. T., Patel, E. U., Khan, R., Srikrishnan, A. K., Anderson, M., Kumar, K. S., Wesolowski, A. P., Iqbal, S. H., Rodgers, M. A., Mehta, S. H., Cloherty, G., Cummings, D. A. T., and Solomon, S. S. (2023). A mixture model to estimate SARS-CoV-2 seroprevalence in Chennai, India. *American Journal of Epidemiology*.

Hoffman, D. and Berger, M. (2011). Statistical considerations for calculation of immunogenicity screening assay cut points. *Journal of Immunological Methods*, 373(1-2):200–208.

International Monetary Fund (2021). World economic outlook database. https://www.imf.org/en/Publications/WEO/weo-database/2021/April. Accessed: 2022-09-16.

Jacofsky, D., Jacofsky, E. M., and Jacofsky, M. (2020). Understanding antibody testing for COVID-19. *The Journal of Arthroplasty*, 35(7):S74–S81.

Jersakova, R., Lomax, J., Hetherington, J., Lehmann, B., Nicholson, G., Briers, M., and Holmes, C. (2022). Bayesian imputation of COVID-19 positive test counts for nowcasting under reporting lag. *Journal of the Royal Statistical Society. Series C, Applied Statistics*.

Jha, P., Brown, P. E., and Ansumana, R. (2022). Counting the global COVID-19 dead. *The Lancet*, 399(10339):1937–1938.

Jordan, G. and Staack, R. F. (2021). An alternative data transformation approach for ADA cut point determination: Why not use a Weibull transformation? *The AAPS Journal*, 23(5):97.

Karanikolos, M., McKee, M., et al. (2020). How comparable is COVID-19 mortality across countries? *Eurohealth*, 26(2):45–50.

Kiriliouk, A., Rootzén, H., Segers, J., and Wadsworth, J. L. (2019). Peaks over thresholds modeling with multivariate generalized Pareto distributions. *Technometrics*, 61(1):123–135.

Klumpp-Thomas, C., Kalish, H., Drew, M., Hunsberger, S., Snead, K., Fay, M. P., Mehalko, J., Shunmugavel, A., Wall, V., Frank, P., et al. (2021). Standardization of ELISA protocols for serosurveys of the SARS-CoV-2 pandemic using clinical and at-home blood sampling. *Nature Communications*, 12(1):113.

Knutson, V., Aleshin-Guendel, S., Karlinsky, A., Msemburi, W., and Wakefield, J. (2022). Estimating global and country-specific excess mortality during the COVID-19 pandemic. *arXiv preprint arXiv:2205.09081*.

Kostoulas, P., Eusebi, P., and Hartnack, S. (2021). Diagnostic accuracy estimates for COVID-19 real-time polymerase chain reaction and lateral flow immunoassay tests with Bayesian latent-class models. *American Journal of Epidemiology*, 190(8):1689–1695.

Larremore, D. B., Fosdick, B. K., Zhang, S., and Grad, Y. H. (2022). Optimizing prevalence estimates for a novel pathogen by reducing uncertainty in test characteristics. *Epidemics*, 41:100634.

Lau, H., Khosrawipour, T., Kocbach, P., Ichii, H., Bania, J., and Khosrawipour, V. (2021). Evaluating the massive underreporting and undertesting of COVID-19 cases in multiple global epicenters. *Pulmonology*, 27(2):110–115.

Laurin, E., Morrison, D., Gardner, I. A., Siah, A., Powell, J. F., and Kamaitis, M. (2019). Bayesian latent class analysis of ELISA and RT-rPCR diagnostic accuracy for subclinical Renibacterium salmoninarum infection in Atlantic salmon (Salmo salar) broodstock. *Journal of Fish Diseases*, 42(2):303–313.

Levin, A. T., Hanage, W. P., Owusu-Boaitey, N., Cochran, K. B., Walsh, S. P., and Meyerowitz-Katz, G. (2020). Assessing the age specificity of infection fatality rates for COVID-19: Systematic review, meta-analysis, and public policy implications. *European Journal of Epidemiology*, 35(12):1123–1138.

Levin, A. T., Owusu-Boaitey, N., Pugh, S., Fosdick, B. K., Zwi, A. B., Malani, A., Soman, S., Besançon, L., Kashnitsky, I., Ganesh, S., et al. (2022). Assessing the burden of COVID-19 in developing countries: Systematic review, meta-analysis and public policy implications. *BMJ Global Health*, 7(5):e008477.

Linnet, K. and Brandt, E. (1986). Assessing diagnostic tests once an optimal cutoff point has been selected. *Clinical Chemistry*, 32(7):1341–1346.

Lipsitch, M., Grad, Y. H., Sette, A., and Crotty, S. (2020). Cross-reactive memory T cells and herd immunity to SARS-CoV-2. *Nature Reviews Immunology*, 20(11):709–713.

Malani, A., Soman, S., Ramachandran, S., Chen, A., and Lakdawalla, D. N. (2022). Vaccine allocation priorities using disease surveillance and economic data. Technical report, National Bureau of Economic Research.

Martín, J., Parra, M. I., Pizarro, M. M., and Sanjuán, E. L. (2022). Baseline methods for the parameter estimation of the generalized Pareto distribution. *Entropy*, 24(2):178.

Mire-Sluis, A. R., Barrett, Y. C., Devanarayan, V., Koren, E., Liu, H., Maia, M., Parish, T., Scott, G., Shankar, G., Shores, E., et al. (2004). Recommendations for the design and optimization of immunoassays used in the detection of host antibodies against biotechnology products. *Journal of Immunological Methods*, 289(1-2):1–16.

Nasrallah, G. K., Dargham, S. R., Shurrab, F., Al-Sadeq, D. W., Al-Jighefee, H., Chemaitelly, H., Al Kanaani, Z., Al Khal, A., Al Kuwari, E., Coyle, P., et al. (2021). Analytic comparison between three high-throughput commercial SARS-CoV-2 antibody assays reveals minor discrepancies in a high-incidence population. *Scientific Reports*, 11(1):11837.

National Academies of Sciences, Engineering, and Medicine (2020). *Evaluating data types: A guide for decision makers using data to understand the extent and spread of COVID-19*. The National Academies Press, Washington, DC.

Nehring, M., Pugh, S., Dihle, T., Gallichotte, E., Nett, T., Weber, E., Mayo, C., Lynn, L., Ebel, G., Fosdick, B. K., et al. (2023). Laboratory-based SARS-CoV-2 receptor-binding domain serologic assays perform with equivalent sensitivity and specificity to commercial FDA-EUA approved tests. *Viruses*, 15(1):106.

Nisar, M. I., Ansari, N., Khalid, F., Amin, M., Shahbaz, H., Hotwani, A., Rehman, N., Pugh, S., Mehmood, U., Rizvi, A., et al. (2021). Serial population-based serosurveys for COVID-19 in two neighbourhoods of Karachi, Pakistan. *International Journal of Infectious Diseases*, 106:176–182.

Our World in Data (2022). COVID-19 data explorer. https://ourworldindata.org/explorers/coronavirus-data-explorer. Accessed: 2023-6-1.

O'Driscoll, M., Ribeiro Dos Santos, G., Wang, L., Cummings, D. A., Azman, A. S., Paireau, J., Fontanet, A., Cauchemez, S., and Salje, H. (2021). Age-specific mortality and immunity patterns of SARS-CoV-2. *Nature*, 590(7844):140–145.

Perez-Saez, J., Lauer, S. A., Kaiser, L., Regard, S., Delaporte, E., Guessous, I., Stringhini, S., Azman, A. S., Alioucha, D., Arm-Vernez, I., et al. (2021). Serology-informed estimates of SARS-CoV-2 infection fatality risk in Geneva, Switzerland. *The Lancet Infectious Diseases*, 21(4):e69–e70.

Pezzullo, A. M., Axfors, C., Contopoulos-Ioannidis, D. G., Apostolatos, A., and Ioannidis, J. P. (2023). Age-stratified infection fatality rate of COVID-19 in the non-elderly population. *Environmental Research*, 216:114655.

Pickands III, J. (1975). Statistical inference using extreme order statistics. *The Annals of Statistics*, pages 119–131.

Rogan, W. J. and Gladen, B. (1978). Estimating prevalence from the results of a screening test. *American Journal of Epidemiology*, 107(1):71–76.

Rosbjerg, D., Madsen, H., and Rasmussen, P. F. (1992). Prediction in partial duration series with generalized Pareto-distributed exceedances. *Water Resources Research*, 28(11):3001–3010.

Ruggles, S. and Heggeness, M. (2008). Intergenerational coresidence in developing countries. *Population and Development Review*, 34(2):253–281.

Ryu, B., Cho, S.-I., Oh, M.-d., Lee, J.-K., Lee, J., Hwang, Y.-O., Yang, J.-S., Kim, S. S., and Bang, J. H. (2019). Seroprevalence of Middle East respiratory syndrome coronavirus (MERS-CoV) in public health workers responding to a MERS outbreak in Seoul, Republic of Korea, in 2015. *Western Pacific Surveillance and Response Journal: WPSAR*, 10(2):46.

Sakarovitch, C., Alioum, A., Ekouevi, D. K., Msellati, P., Leroy, V., and Dabis, F. (2007). Estimating incidence of HIV infection in childbearing age African women using serial prevalence data from antenatal clinics. *Statistics in Medicine*, 26(2):320–335.

Schaarschmidt, F., Hofmann, M., Jaki, T., Grün, B., and Hothorn, L. A. (2015). Statistical approaches for the determination of cut points in anti-drug antibody bioassays. *Journal of Immunological Methods*, 418:84–100.

Seaman, S. R., Samartsidis, P., Kall, M., and De Angelis, D. (2022). Nowcasting COVID-19 deaths in England by age and region. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 71(5):1266–1281.

SeroTracker (2023). Serotracker. https://serotracker.com/. Accessed: 2023-04-21.

Shen, M., Dong, X., and Tsong, Y. (2015). Statistical evaluation of several methods for cut-point determination of immunogenicity screening assay. *Journal of Biopharmaceutical Statistics*, 25(2):269–279.

Smith, R. L. (1985). Maximum likelihood estimation in a class of nonregular cases. *Biometrika*, 72(1):67–90.

Stan Development Team (2022). RStan: The R interface to Stan. R package version 2.21.5.

Starke, K. R., Reissig, D., Petereit-Haack, G., Schmauder, S., Nienhaus, A., and Seidler, A. (2021). The isolated effect of age on the risk of COVID-19 severe outcomes: A systematic review with meta-analysis. *BMJ Global Health*, 6(12):e006434.

Stengaard, A. R., Combs, L., Supervie, V., Croxford, S., Desai, S., Sullivan, A. K., Jakobsen, S. F., Santos, Q., Simões, D., Casabona, J., et al. (2021). HIV seroprevalence in five key populations in Europe: A systematic literature review, 2009 to 2019. *Eurosurveillance*, 26(47):2100044.

Stevenson, M. and Sergeant, E. (2023). *epiR: Tools for the Analysis of Epidemiological Data*. R package version 2.0.60.

Stringhini, S., Wisniak, A., Piumatti, G., Azman, A. S., Lauer, S. A., Baysson, H., De Ridder, D., Petrovic, D., Schrempft, S., Marcus, K., et al. (2020). Seroprevalence of anti-SARS-CoV-2 IgG antibodies in Geneva, Switzerland (SEROCoV-POP): A population-based study. *The Lancet*, 396(10247):313–319.

Symons, R., Beath, K., Dangis, A., Lefever, S., Smismans, A., De Bruecker, Y., and Frans, J. (2021). A statistical framework to estimate diagnostic test performance for COVID-19. *Clinical Radiology*, 76(1):75–e1.

Takahashi, S., Greenhouse, B., and Rodríguez-Barraquer, I. (2020). Are seroprevalence estimates for severe acute respiratory syndrome coronavirus 2 biased? *The Journal of Infectious Diseases*, 222(11):1772–1775.

Tunheim, G., Rø, G. Ø. I., Tran, T., Kran, A.-M. B., Andersen, J. T., Vaage, E. B., Kolderup, A., Vaage, J. T., Lund-Johansen, F., and Hungnes, O. (2022). Trends in seroprevalence of SARS-CoV-2 and infection fatality rate in the Norwegian population through the first year of the COVID-19 pandemic. *Influenza and Other Respiratory Viruses*, 16(2):204–212.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Vandenberg, O., Martiny, D., Rochas, O., van Belkum, A., and Kozlakidis, Z. (2021). Considerations for diagnostic COVID-19 tests. *Nature Reviews Microbiology*, 19(3):171–183.

Vink, M. A., van de Kassteele, J., Wallinga, J., Teunis, P. F., and Bogaards, J. A. (2015). Estimating seroprevalence of human papillomavirus type 16 using a mixture model with smoothed age-dependent mixing proportions. *Epidemiology*, 26(1):8–16.

Ward, H., Whitaker, M., Flower, B., Tang, S. N., Atchison, C., Darzi, A., Donnelly, C. A., Cann, A., Diggle, P. J., Ashby, D., et al. (2022). Population antibody responses following COVID-19 vaccination in 212,102 individuals. *Nature Communications*, 13(1):907.

Wong, J. Y., Wu, P., Nishiura, H., Goldstein, E., Lau, E. H., Yang, L., Chuang, S., Tsang, T., Peiris, J. M., Wu, J. T., and Cowling, B. J. (2013). Infection fatality risk of the pandemic A(H1N1)2009 virus in Hong Kong. *American Journal of Epidemiology*, 177(8):834–840.

Zhang, J. (2021). Hospital avoidance and unintended deaths during the COVID-19 pandemic. *American Journal of Health Economics*, 7(4):405–426.

Zhang, L., Zhang, J. J., Kubiak, R. J., and Yang, H. (2013). Statistical methods and tool for cut point analysis in immunogenicity assays. *Journal of Immunological Methods*, 389(1-2):79–87.

Zimmer, S. M., Crevar, C. J., Carter, D. M., Stark, J. H., Giles, B. M., Zimmerman, R. K., Ostroff, S. M., Lee, B. Y., Burke, D. S., and Ross, T. M. (2010). Seroprevalence following the second wave of Pandemic 2009 H1N1 influenza in Pittsburgh, PA, USA. *PLoS One*, 5(7):e11601.

Appendix A

Supplemental Material for Chapter 2

A.1 Cutoff estimation methods

A.1.1 Empirical quantile estimation

To calculate the Q empirical quantile from a sample of size n, we calculate a weighted average of the two order statistics surrounding the desired quantile. Specifically, $(1-\gamma)x_{(i)}+\gamma x_{(i+1)}$ where $x_{(i)}$ denotes the i^{th} order statistic, $i = \lfloor (n-1)Q + 1 \rfloor$, $\gamma = (n-1)Q + 1 - \lfloor (n-1)Q + 1 \rfloor$, and $\lfloor \cdot \rfloor$ is the floor function.

A.1.2 Hybrid methods

Figure A.1 shows the flow chart for the hybrid methods.

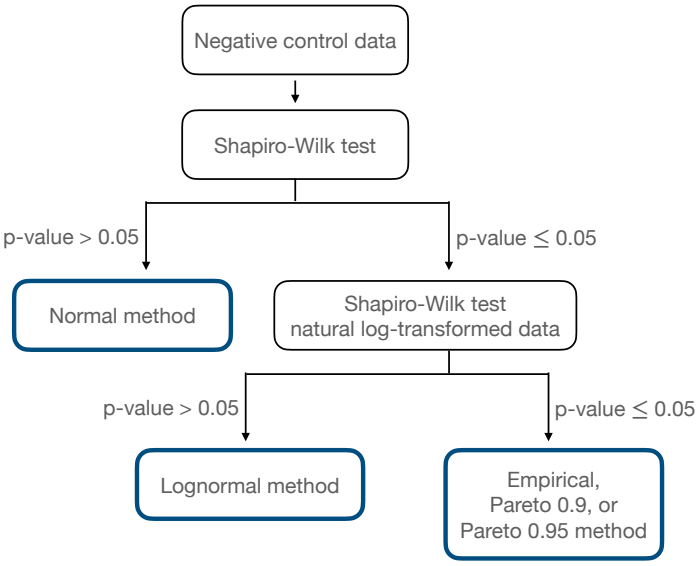


Figure A.1: Flow chart indicating how the cutoff is estimated for the hybrid methods.

A.2 Mixture distribution for the simulation study

For each test (spike or RBD) and control type (positive or negative), we fit mixture distributions of the form

$$g(x) = \sum_{i=1}^{K} \pi_i f_i(x) \tag{A.1}$$

where K is the number of components, π_i gives the weight of each component, $f_i(x)$ is the probability density function of each component evaluated at x, and $g(x)$ is the resulting mixture distribution evaluated at x. We considered gamma, Weibull, and lognormal distributions and either two or three components. All possible combinations of these distribution were fit using the `ltmix` package in R for each number of components Blostein and Miljkovic (2019). We selected the best model for each in terms of BIC and visual inspection. The resulting mixture distributions are given in Table A.1.

Table A.1: The mixture distribution fit to each test and control type. The mixture probabilities are given by π_i.

Test	Control type	π_1	Distribution 1	π_2	Distribution 2	π_3	Distribution 3
Spike	Positive	0.65	gamma(8.42, 0.53)	0.35	lognormal(1.06, 0.12)		
	Negative	0.13	lognormal(0.75, 0.42)	0.87	lognormal(-0.13, 0.32)		
RBD	Positive	0.36	gamma(5.87, 2.32)	0.59	gamma(13.33, 1.48)	0.05	gamma(1.67, 4.71)
	Negative	0.27	gamma(6.62, 0.29)	0.69	lognormal(-0.05, 0.22)	0.03	lognormal(1.49, 0.88)

Abbreviations: receptor-binding domain, RBD

A.3 Data analysis cutoffs

Table A.2 gives the results from Figure 2 in the main text in numerical form. Specifically, it gives the cutoffs for each test, method, and target specificity.

Table A.2: Estimated cutoff for each estimation method on each training data source.

	Empirical	Normal	Log Normal	MAD	Log MAD	Pareto 0.9	Pareto 0.95
Target specificity=0.95							
Spike	2.5	2.2	2.1	1.5	1.7	2.4	2.5
RBD	3.0	4.2	2.7	1.6	1.8	3.6	3.0
Target specificity=0.995							
Spike	4.4	2.8	3.1	1.8	2.4	4.6	4.3
RBD	9.6	5.8	4.2	1.9	2.5	8.2	9.7

Abbreviations: mean absolute deviation, MAD; receptor-binding domain, RBD

A.4 Sensitivity estimates

Table A.3: The median and middle 95% (2.5% quantile, 97.5% quantile) of the sensitivity. The method with the largest sensitivity in each scenario is bolded.

	Scenario A				Scenario B			
	n=50		n=200		n=50		n=200	
Target specificity=0.995								
Empirical	0.51	(0.11, 0.98)	0.38	(0.13, 0.69)	0.97	(0.12, 0.99)	0.91	(0.23, 0.99)
Normal	0.84	(0.39, 0.99)	0.80	(0.52, 0.96)	0.98	(0.71, 0.99)	0.97	(0.81, 0.99)
Log Normal	0.64	(0.30, 0.97)	0.62	(0.43, 0.88)	0.98	(0.92, 0.99)	0.98	(0.96, 0.99)
MAD	**0.99**	**(0.96, 1.00)**	**0.99**	**(0.98, 0.99)**	**0.99**	**(0.99, 1.00)**	**0.99**	**(0.99, 1.00)**
Log MAD	0.94	(0.44, 0.99)	0.93	(0.69, 0.98)	0.99	(0.98, 1.00)	0.99	(0.99, 0.99)
Pareto 0.9	0.35	(0.03, 0.96)	0.31	(0.12, 0.55)	0.96	(0.12, 0.99)	0.93	(0.56, 0.98)
Pareto 0.95	0.42	(0.05, 0.97)	0.31	(0.10, 0.57)	0.96	(0.05, 0.99)	0.90	(0.25, 0.98)
Hybrid Empirical	0.52	(0.11, 0.97)	0.38	(0.13, 0.70)	0.97	(0.12, 0.99)	0.91	(0.23, 0.99)
Hybrid Pareto 0.9	0.40	(0.03, 0.97)	0.31	(0.12, 0.56)	0.96	(0.12, 0.99)	0.93	(0.56, 0.98)
Hybrid Pareto 0.95	0.45	(0.05, 0.97)	0.31	(0.10, 0.58)	0.96	(0.05, 0.99)	0.90	(0.25, 0.98)
Target specificity=0.95								
Empirical	0.97	(0.52, 1.00)	0.95	(0.69, 0.99)	0.99	(0.97, 0.99)	0.99	(0.99, 0.99)
Normal	0.98	(0.69, 1.00)	0.97	(0.88, 0.99)	0.99	(0.88, 0.99)	0.99	(0.92, 0.99)
Log Normal	0.98	(0.83, 1.00)	0.98	(0.95, 0.99)	0.99	(0.98, 0.99)	0.99	(0.99, 0.99)
MAD	**1.00**	**(0.99, 1.00)**	**1.00**	**(0.99, 1.00)**	**1.00**	**(0.99, 1.00)**	**1.00**	**(1.00, 1.00)**
Log MAD	0.99	(0.96, 1.00)	0.99	(0.98, 1.00)	0.99	(0.99, 1.00)	0.99	(0.99, 1.00)
Pareto 0.9	0.96	(0.54, 1.00)	0.94	(0.74, 0.99)	0.99	(0.92, 0.99)	0.99	(0.97, 0.99)
Hybrid Empirical	0.97	(0.52, 1.00)	0.95	(0.69, 0.99)	0.99	(0.97, 0.99)	0.99	(0.99, 0.99)
Hybrid Pareto 0.9	0.97	(0.54, 1.00)	0.94	(0.74, 0.99)	0.99	(0.92, 0.99)	0.99	(0.97, 0.99)

Abbreviations: mean absolute deviation, MAD

A.5 Properties of estimators

A.5.1 Empirical quantile

If the cumulative distribution function (CDF) of the negative controls, F, is differentiable with a positive derivative at $F^{-1}(Q)$, the empirical quantile is asymptotically normally distributed (van der Vaart, 1998). Let $\widehat{F}_n^{-1}(\cdot)$ denote the empirical quantile based on a random sample of n observations as defined in Appendix A.1.1 and $F^{-1}(\cdot)$ denote the inverse CDF. Then asymptotically,

$$\sqrt{n}\left(\widehat{F}_n^{-1}(Q) - F^{-1}(Q)\right) \sim \mathcal{N}\left(0, \frac{Q(1-Q)}{f^2\left(F^{-1}(Q)\right)}\right) \tag{A.2}$$

where $f^2(\cdot)$ is the squared derivative of F.

However, we focus on the limited sample size case, which poses unique challenges. For example, our estimate will be a value contained in the range of the sample, but with a sample size of 50 we would only expect 22% of samples to contain the 99.5% quantile. Thus, 78% of the empirical quantile estimates for 0.995 must be below the true quantile. This induces bias in the empirical estimator for the small sample sizes (see Tables 2.2 and 2.3). Additionally, in our limited sample size scenario, the empirical quantile can have considerable variability. Thus, we consider the Pareto methods as a parametric alternative.

A.5.2 Extreme value theory

Typically, the peaks over threshold approach in extreme value theory is used in the scenario where we seek to extrapolate. For example, Castellanos and Cabras (2007) used 35 years of flood data to predict the 25-, 50-, and 100-year events, i.e., the largest flood we would expect to see in 25, 50, or 100 years, respectively. Thus, we make parametric assumptions to extrapolate to unobserved years.

Let F be the CDF for a random variable X that is in the domain of attraction of an extreme value distribution, and let $Y = X - u$ for threshold u. Then the distribution of the exceedances of

the threshold follow a generalized Pareto distribution (GPD)

$$P(Y > y|Y > 0) \approx \begin{cases} 1 - (1 + \frac{\xi y}{\sigma})^{-1/\xi} & \text{if } \xi \neq 0, \\ 1 - \exp(-\frac{y}{\sigma}) & \text{if } \xi = 0, \end{cases} \tag{A.3}$$

as u approaches the upper limit of the support of X (Beirlant et al., 2006).

A feature of typical extreme value theory applications is that in such applications we seek to estimate increasingly extreme events that are contained in an increasingly extreme section of the tail of the distribution. Therefore, as the sample size increases, the threshold u can also increase while staying below the extreme event being estimated. In our application, the quantile being estimates is fixed and increasing the threshold u will ultimately result in the situation where u is greater than the target specificity. For example, consider a target specificity of 0.995. If the threshold u exceeds the 0.995 quantile, then we cannot use the exceedances to estimate the 0.995 quantile. Thus, the concept of u approaching the upper limit of the support does not apply for our application. However, due to our limited sample sizes, we aim to make parametric assumptions to minimize variability, and the GPD (generalized Pareto distribution) is a reasonable distribution to assume for the exceedances. In our scenario with the least data, where we have a training sample size of 50 and fit a GPD to the upper 5% of the data, we must estimate the parameters of the GPD with only three data points. Thus, with limited data, we further make the assumption that the shape parameter, ξ, is zero, corresponding to a shifted exponential distribution, to reduce the number of parameters that need to be estimated.

Next we consider estimation of the parameters of the generalized Pareto distribution. Smith (1985) showed that for $\xi > -1/2$, the standard regularity conditions for asymptotic properties of maximum likelihood estimates hold. Thus, assuming the model is correct, the maximum likelihood estimates are consistent and have a convergence rate of $\sqrt{n_u}$ where n_u is the number of exceedances above the threshold u. These properties also hold for the exponential distribution, corresponding to our estimator which fixes $\xi = 0$ and estimates only the scale parameter. As our estimator of the Q quantile is a linear function of the scale parameter, the estimator is also consistent for the

target quantile by the continuous mapping theorem. Our modeling approach would fall within this framework, where we assume a fixed threshold and increasing sample size above the threshold. However, the properties discussed here assume that the exceedances exactly follow a GPD, or more specifically the exponential distribution in our case, while the GPD is in practice just the limiting distribution of the exceedances.

Instead of assuming the exceedances follow a GPD, some such as Drees et al. (2004) have studied the estimation of the GPD parameters for general underlying distribution. They estimate the parameters as the threshold u approaches the upper limit of the distribution for any distribution F that is in the domain of attraction of the extreme value distribution. Drees et al. (2004) showed that when the threshold u is based on upper order statistics and $\xi > -1/2$, the maximum likelihood estimates are asymptotically normal with convergence rate $\sqrt{n_u}$. However, such work assumes the sample size, n, approaches infinity and the number of exceedances, n_u, approaches infinity such that $\lim_{n\to\infty} n/n_u = 0$. As before, this type of asymptotics does not apply to our setting because the fixed quantile we seek to estimate would eventually fall below the threshold.

In summary, we employ the GPD because it is a sensible model for the tail, justified by results from extreme value theory. At the same time, our small sample size setting is quite different from the typical extremes setting, and its asymptotic results are less relevant here. When the threshold u is fixed, there is no convergence of the tail to a GPD, so we would not expect consistency of our estimator in general. If the sample size is large enough, the empirical estimator performs well. For the limited sample sizes of our application, our simulation study showed the parametric alternative outperforms the nonparametric empirical estimator for the higher target specificity.

Appendix B

Supplemental Material for Chapter 3

B.1 Combining bins using the population age distribution

Let $f_\ell(a)$ denote the number of individuals of age a at location ℓ for $a \in \{0, 1, \ldots, 84+\}$. Note, if population age structure is only available in 5-year age bins, then define

$$f_\ell(a) = \sum_{b \in \{0,5,\ldots,80\}} \frac{f_\ell([b, b+5))}{5} \mathbb{I}_{[b,b+5)}(a) \tag{B.1}$$

where $f_\ell([b, b+5))$ is the proportion of the population ages $[b, b+5)$.

In cases where the location specific age structure is only available in larger bins, but the national age structure is available in 5-year age bins, we leverage the national age structure to inform the location specific age structure as follows. Let A denote an interval the location specific age structure is available for (e.g., $[0, 18)$). If $f(A)$ is the proportion of the population at location ℓ with an age in A and $f_n(a)$ is the proportion of the population aged a at the national level, then we estimate $f_\ell(a)$, the proportion at location ℓ that is age a, as

$$f_\ell(a) = f(A) \frac{f_n(a)}{\sum_{b \in A \cap \mathbb{N}} f_n(b)}. \tag{B.2}$$

Essentially, we rescale $f_n(a)$ such that the total mass in A matches the observed total mass in A at location ℓ, $f(A)$. Since we model seroprevalence as constant past age 85, we let $f_\ell(85)$ represent the proportion of the population aged 85 or older, rather than just the proportion aged 85.

Appendix C

Supplemental Material for Chapter 4

C.1 Approximating the age distribution

The exact age distribution is not available for any particular location. Instead, we have the age distribution split into age bins that do not necessarily correspond to those of the serology study or death data. Let $\mathcal{A}$ be the set of such disjoint age bins corresponding to the age distribution data. Let $\widehat{f}_\ell(A)$ denote the the observed proportion of the population in age bin A, for $A \in \mathcal{A}$. Then we define the step function

$$\widetilde{f}_\ell(a) = \sum_{A \in \mathcal{A}} \frac{\widehat{f}_\ell(A)}{|A|} \mathbb{I}(a \in A) \tag{C.1}$$

where $|A|$ gives the length of interval A.

When the age bins for a population were larger than five, we further refined the step function using the national age distribution for that location, $\widehat{f}_{n_\ell}(A)$, which was expanded using (C.1) to $\widetilde{f}_{n_\ell}(a)$. We then define

$$\widetilde{f}_\ell(a) = \sum_{A \in \mathcal{A}} \widehat{f}_\ell(A) \frac{\widetilde{f}_{n_\ell}(a)}{\sum_{b \in A \cap \mathbb{N}} \widetilde{f}_{n_\ell}(b)} \mathbb{I}(a \in A) \tag{C.2}$$

where $\mathbb{N}$ denotes the natural numbers. Thus, the national age distribution is rescaled so the proportion of the population in A equals $\widehat{f}_\ell(A)$, the proportion for location ℓ.

Finally, we used locally weighted smoothing (LOESS) to smooth the population age distribution, rather than using the step function defined in (C.2). We used the points from the step function up to age 85, as this was the limit of our data in some locations. We added a data point at age 100 with a value of 0 as individuals over the age of 100 are very rare in developing countries. If any predictions were less than zero, we subtracted the minimum values from all the predictions to give all predictions a positive value. This made minimal changes as the negative values were always

negligible. Finally, we rescaled the predictions to sum to one. These resulting predictions, shown in Figure C.1, are what we defined to be $f_\ell(a)$.

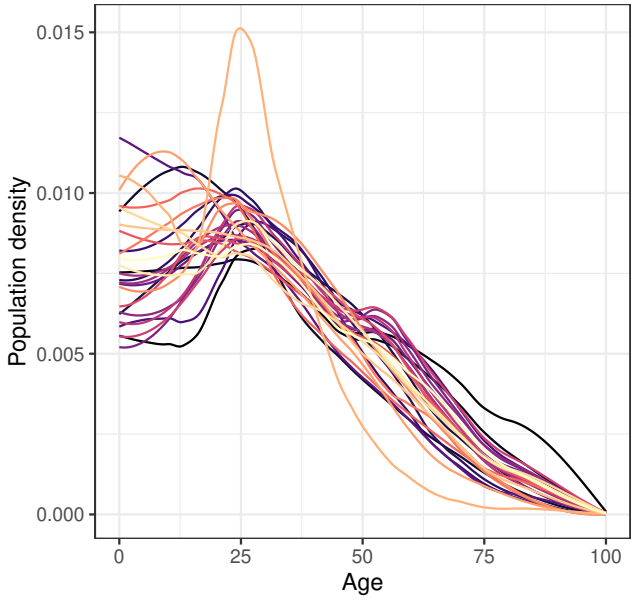


Figure C.1: Population density for each location, $f_\ell(a)$.

C.2 Convergence diagnostics

The range of $\widehat{R}$ and effective sample size are shown for groupings of parameters in Table C.1 (Gelman et al., 1995). All effective samples sizes were above 1000 and $\widehat{R}$ were within 0.0042 of one for each parameter. The traceplots for the three parameters with the smallest effective samples sizes are shown in Figure C.2. Traceplots for the other parameters are similar, suggesting convergence. Sampling via RStan took 129 minutes using a standard laptop with a 2.7 GHz quad-core processor and 16 GB of RAM.

Table C.1: The range of $\widehat{R}$ and the effective sample size (ESS) for each grouping of parameters.

	$\widehat{R}$	ESS
$\beta_{\ell,i}$	$(0.9998, 1.0014)$	$(5673, 11765)$
$\beta_{\text{global},i}$	$(1.0002, 1.0042)$	$(1338, 3634)$
$\beta_{\text{country},c_\ell}$	$(1.0007, 1.0037)$	$(1419, 2493)$
σ_i	$(1.0001, 1.0002)$	$(2936, 3517)$
σ_{country}	$(1.0005, 1.0005)$	$(2952, 2952)$
$\gamma_{\ell,j}$	$(0.9998, 1.0017)$	$(5109, 15224)$
sens_t	$(1.0000, 1.0014)$	$(2929, 11443)$
spec_t	$(0.9998, 1.0006)$	$(5869, 14531)$

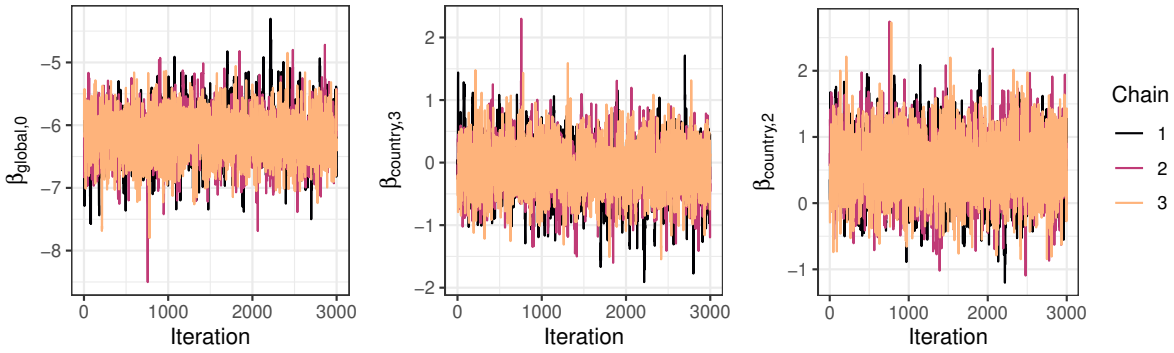


Figure C.2: Traceplots of (a) $\beta_{\text{global},0}$, (b) $\beta_{\text{country},3}$, and (c) $\beta_{\text{country},2}$, colored by chain.

C.3 Seroprevalence curves by study date

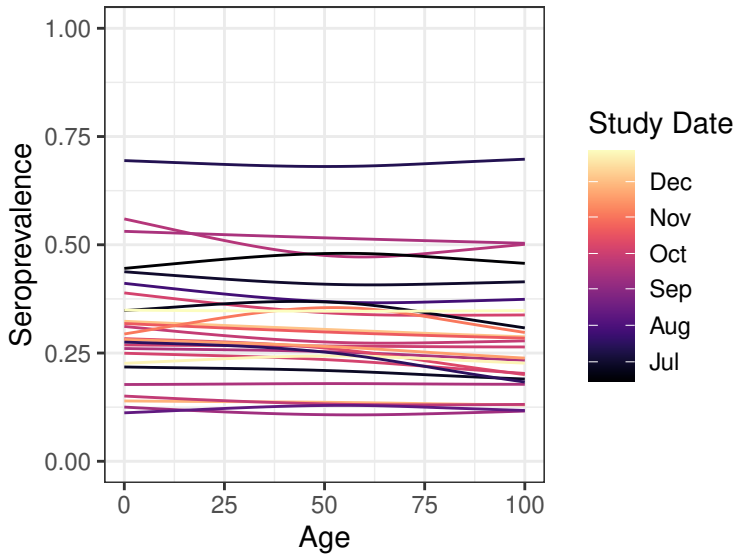
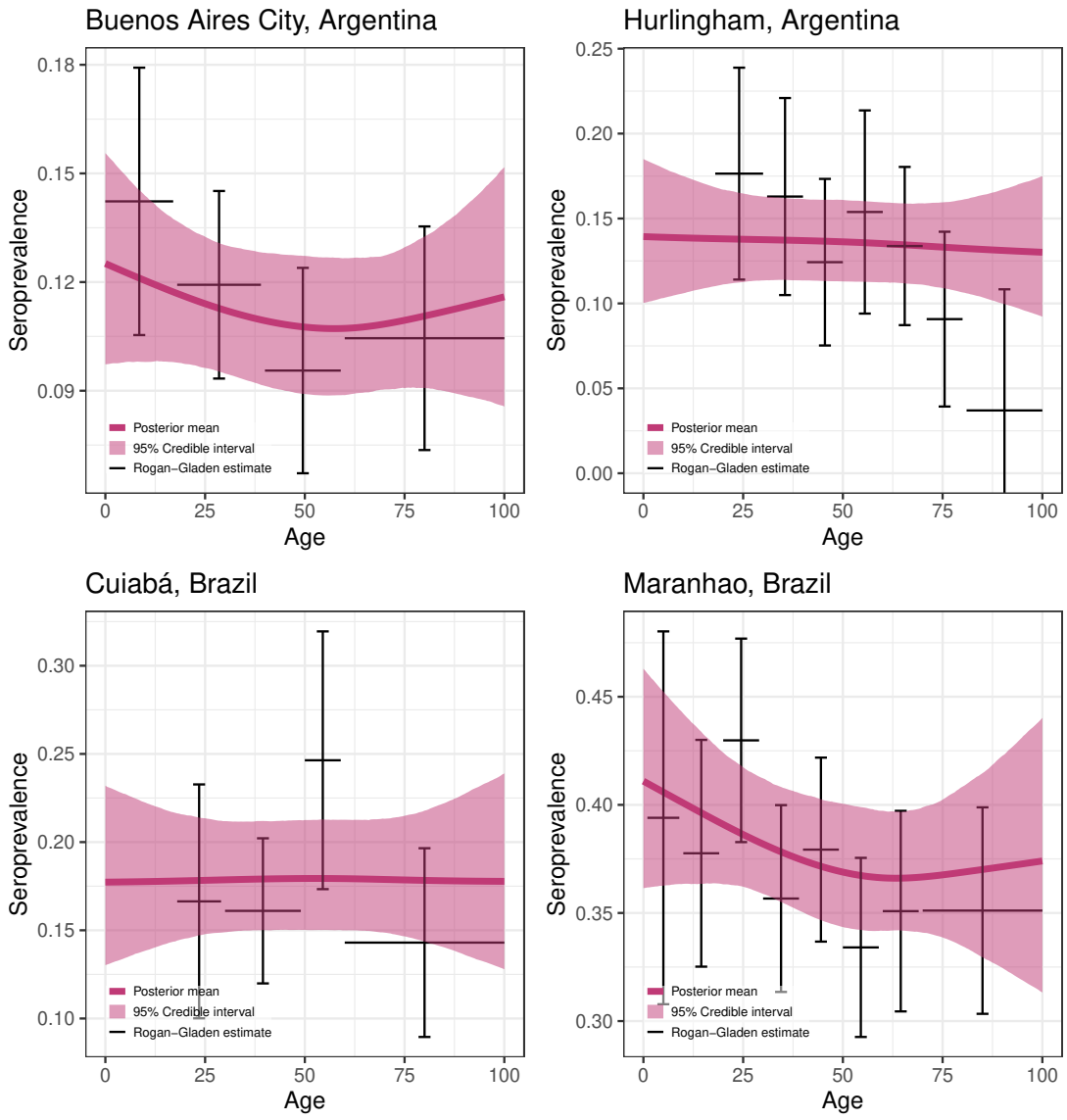
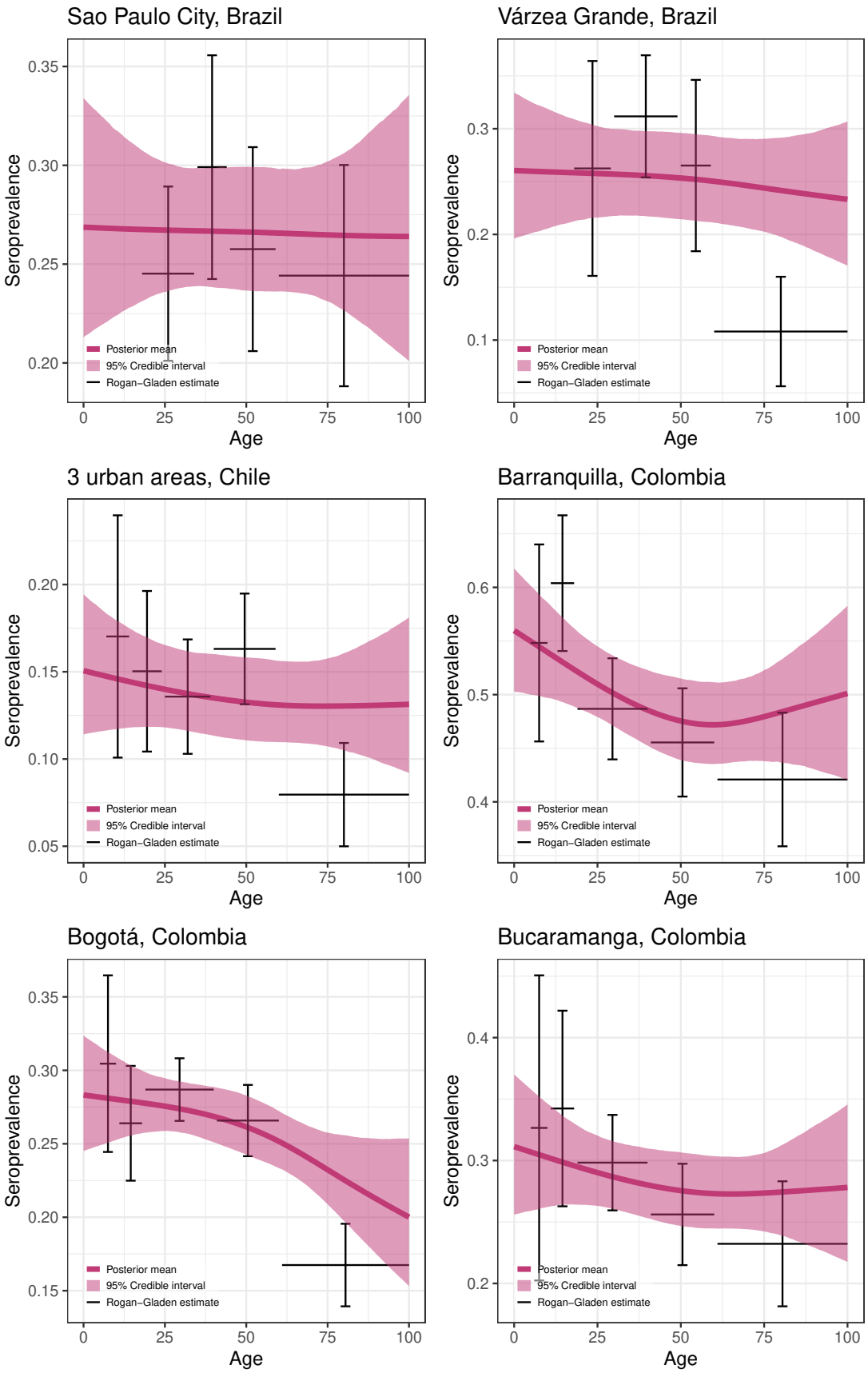


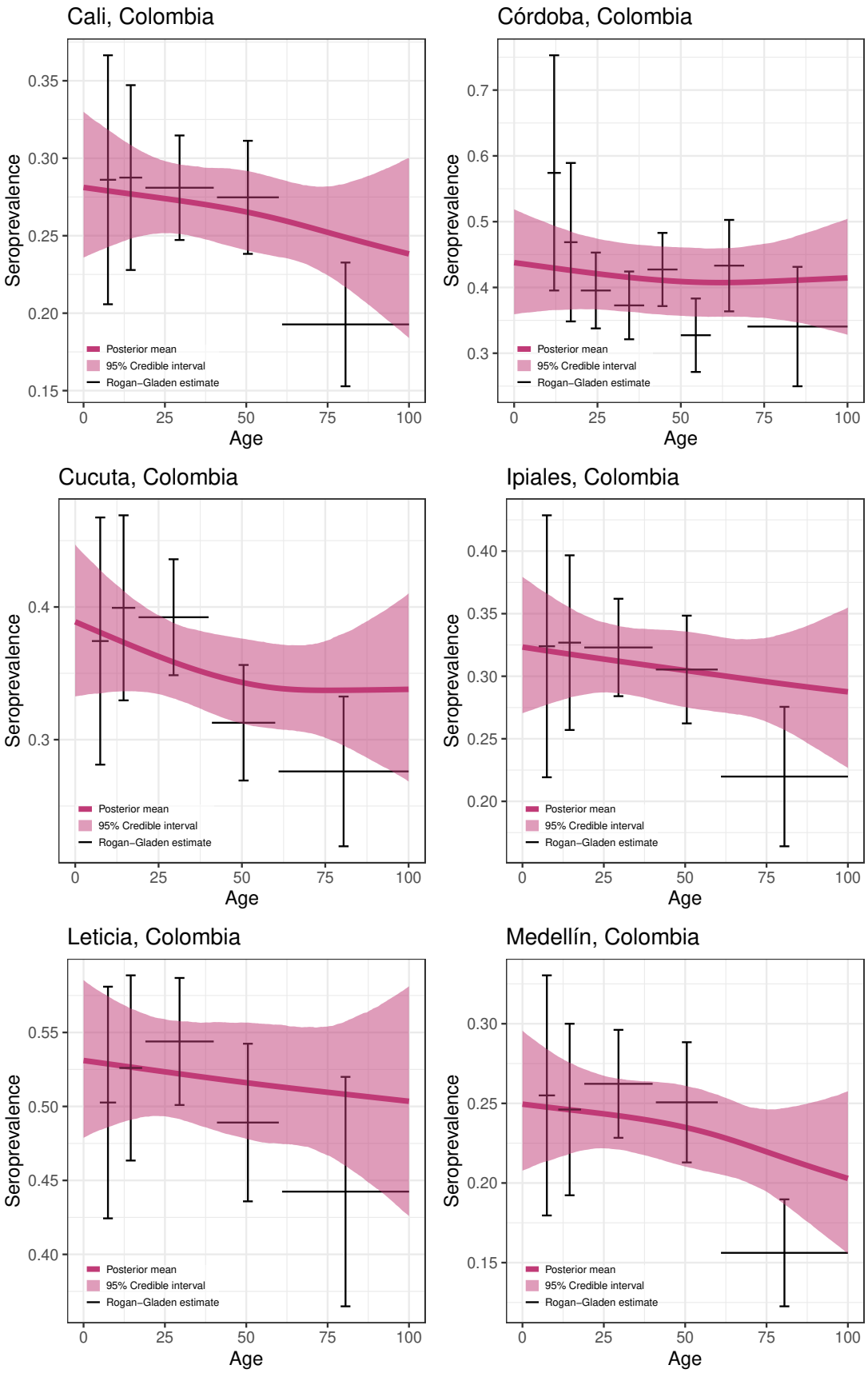
Figure C.3: Posterior mean seroprevalence curve for each study location, colored by the start date of each study.

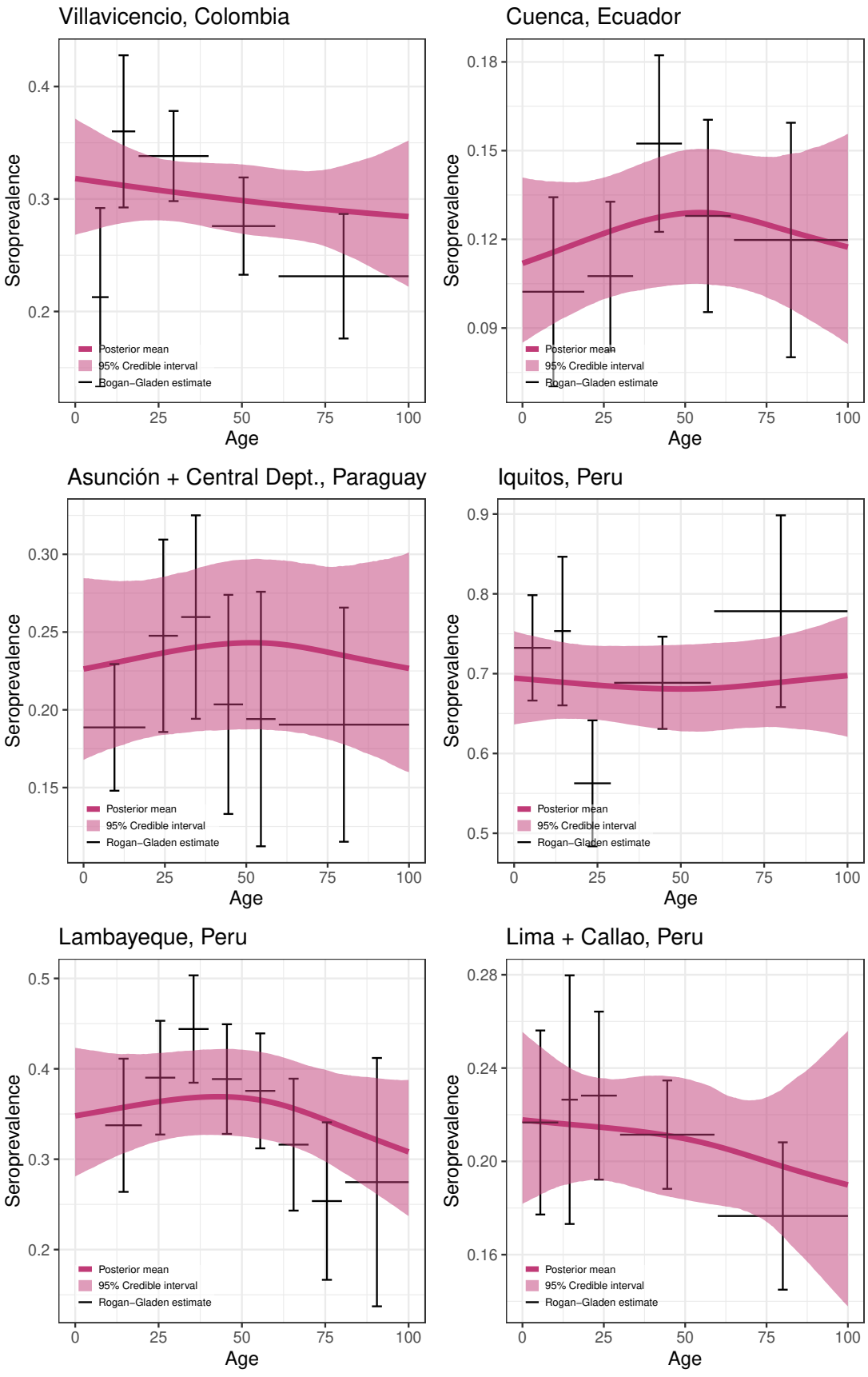
C.4 Seroprevalence curves for each location

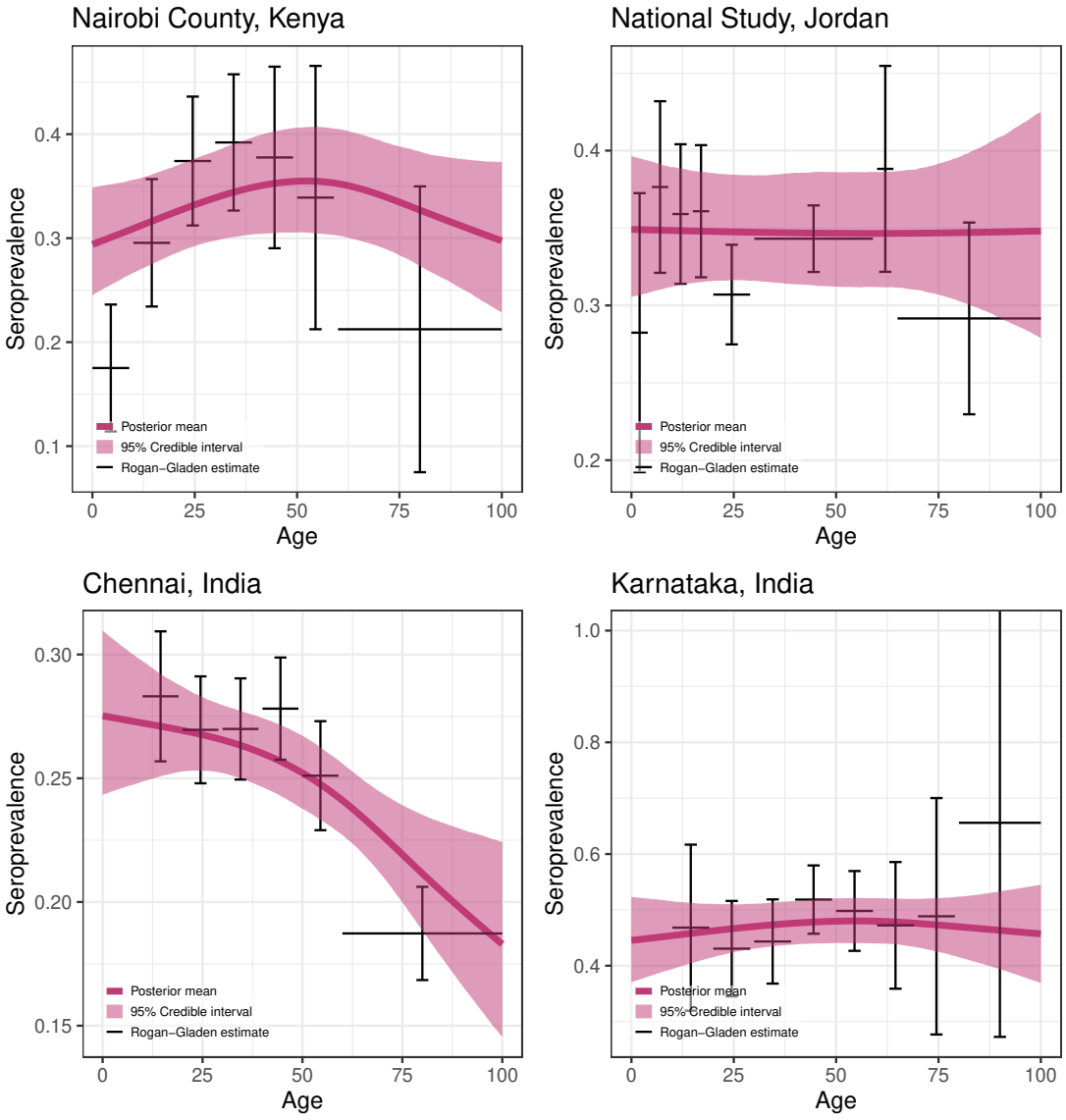
Each of the following plots shows the posterior mean seroprevalence curve with a 95% credible interval. Also shown is the Rogan-Gladen estimate with an approximate confidence interval that treats sensitivity and specificity as known.




Sao Paulo City, Brazil
Seroprevalence
Age
Posterior mean
95% Credible interval
Rogan–Gladen estimate
Várzea Grande, Brazil
3 urban areas, Chile
Barranquilla, Colombia
Bogotá, Colombia
Bucaramanga, Colombia

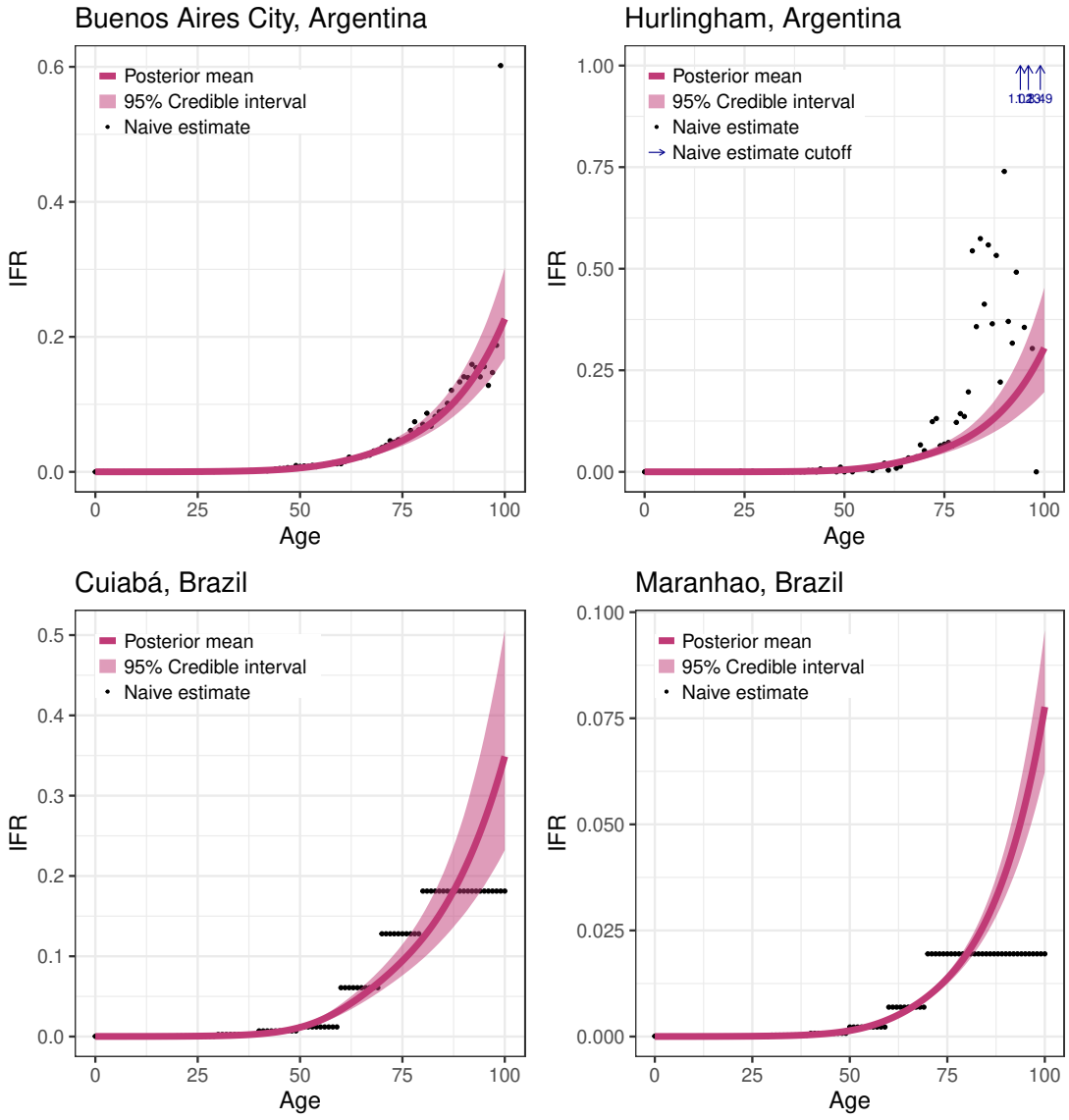

Cali, Colombia
Córdoba, Colombia
Cucuta, Colombia
Ipiales, Colombia
Leticia, Colombia
Medellín, Colombia
Seroprevalence
Age
Posterior mean
95% Credible interval
Rogan–Gladen estimate

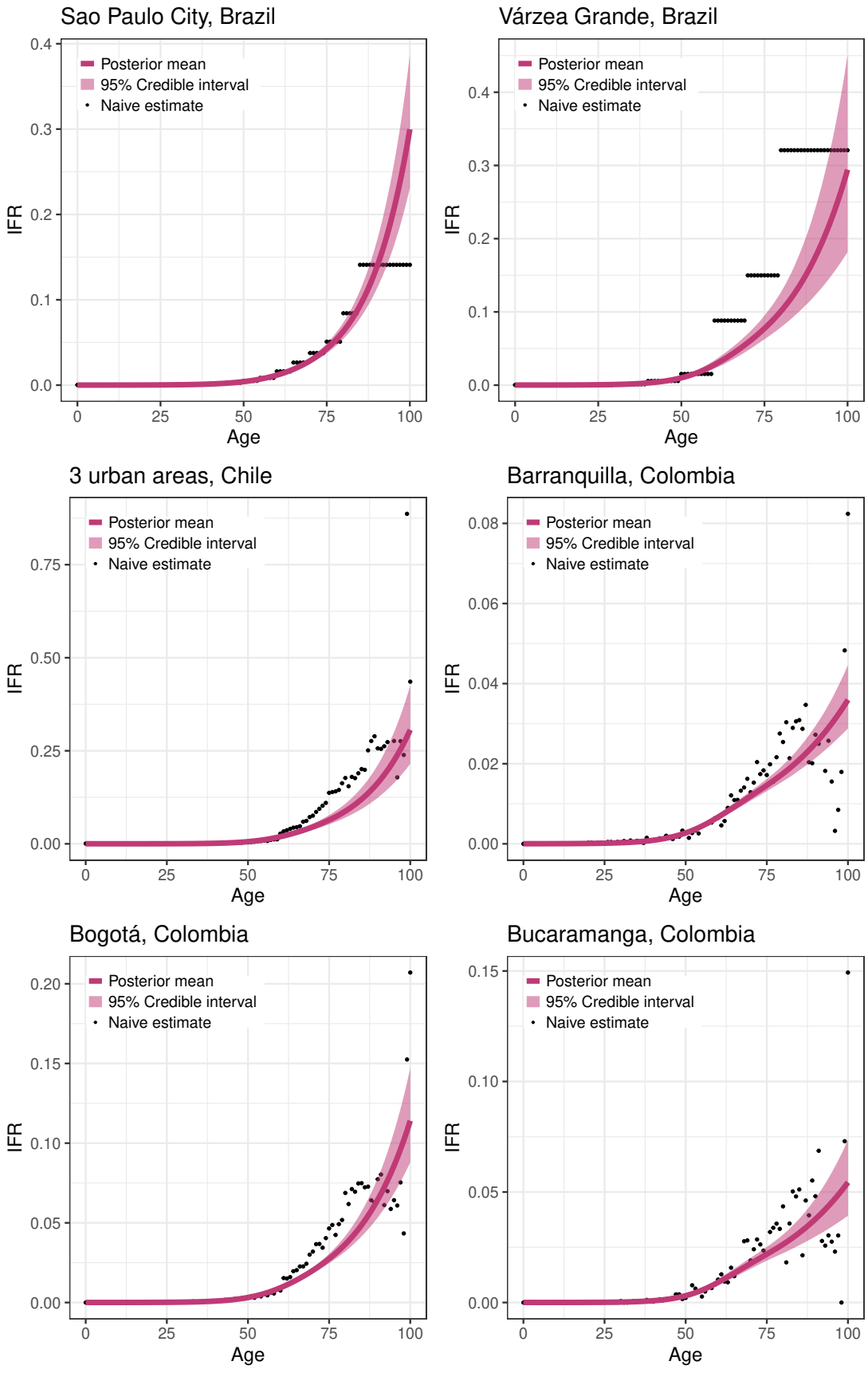

Villavicencio, Colombia
Cuenca, Ecuador
Asunción + Central Dept., Paraguay
Iquitos, Peru
Lambayeque, Peru
Lima + Callao, Peru
Seroprevalence
Age
Posterior mean
95% Credible interval
Rogan–Gladen estimate

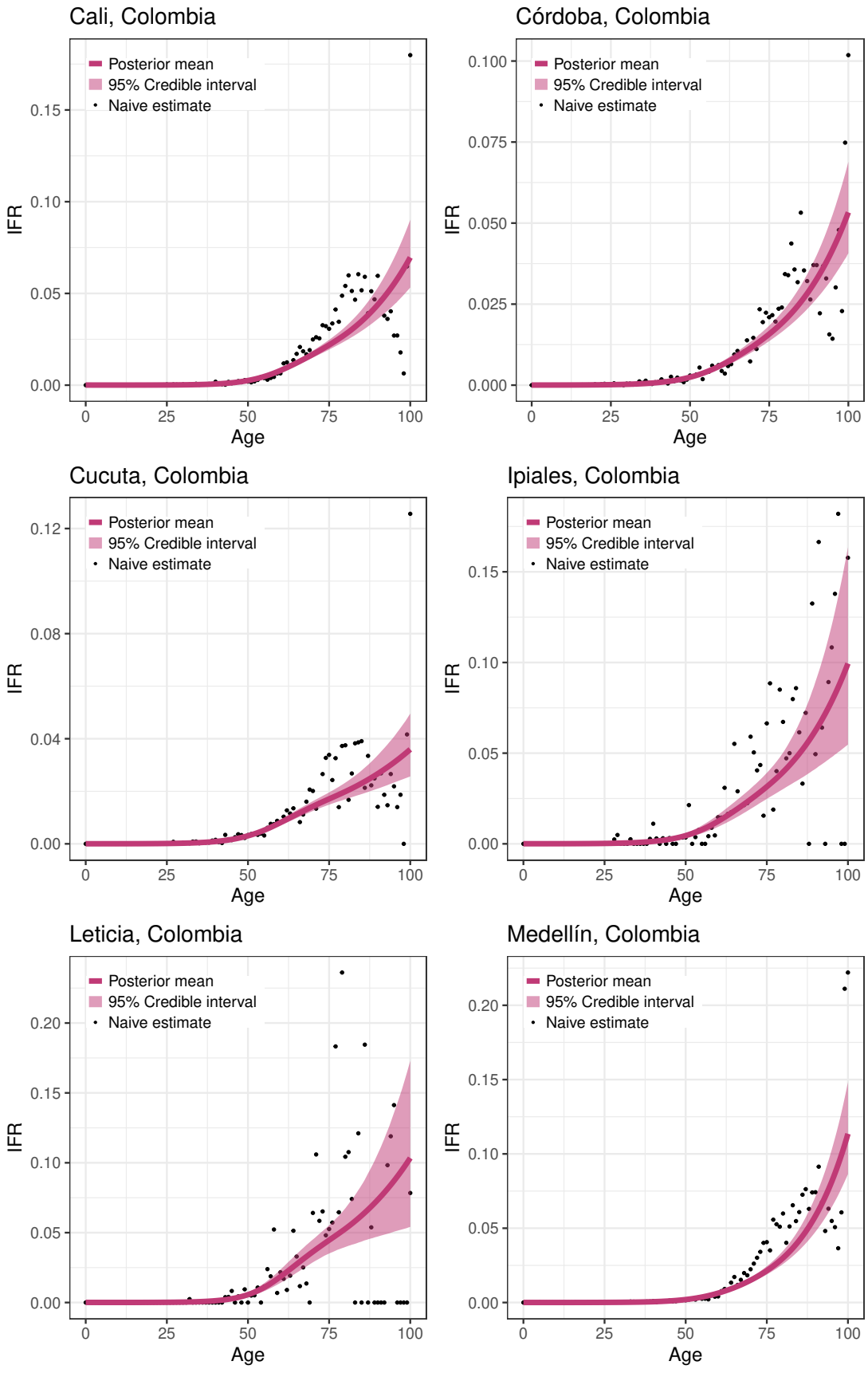

Nairobi County, Kenya
Seroprevalence
Age
Posterior mean
95% Credible interval
Rogan–Gladen estimate
National Study, Jordan
Seroprevalence
Age
Posterior mean
95% Credible interval
Rogan–Gladen estimate
Chennai, India
Seroprevalence
Age
Posterior mean
95% Credible interval
Rogan–Gladen estimate
Karnataka, India
Seroprevalence
Age
Posterior mean
95% Credible interval
Rogan–Gladen estimate

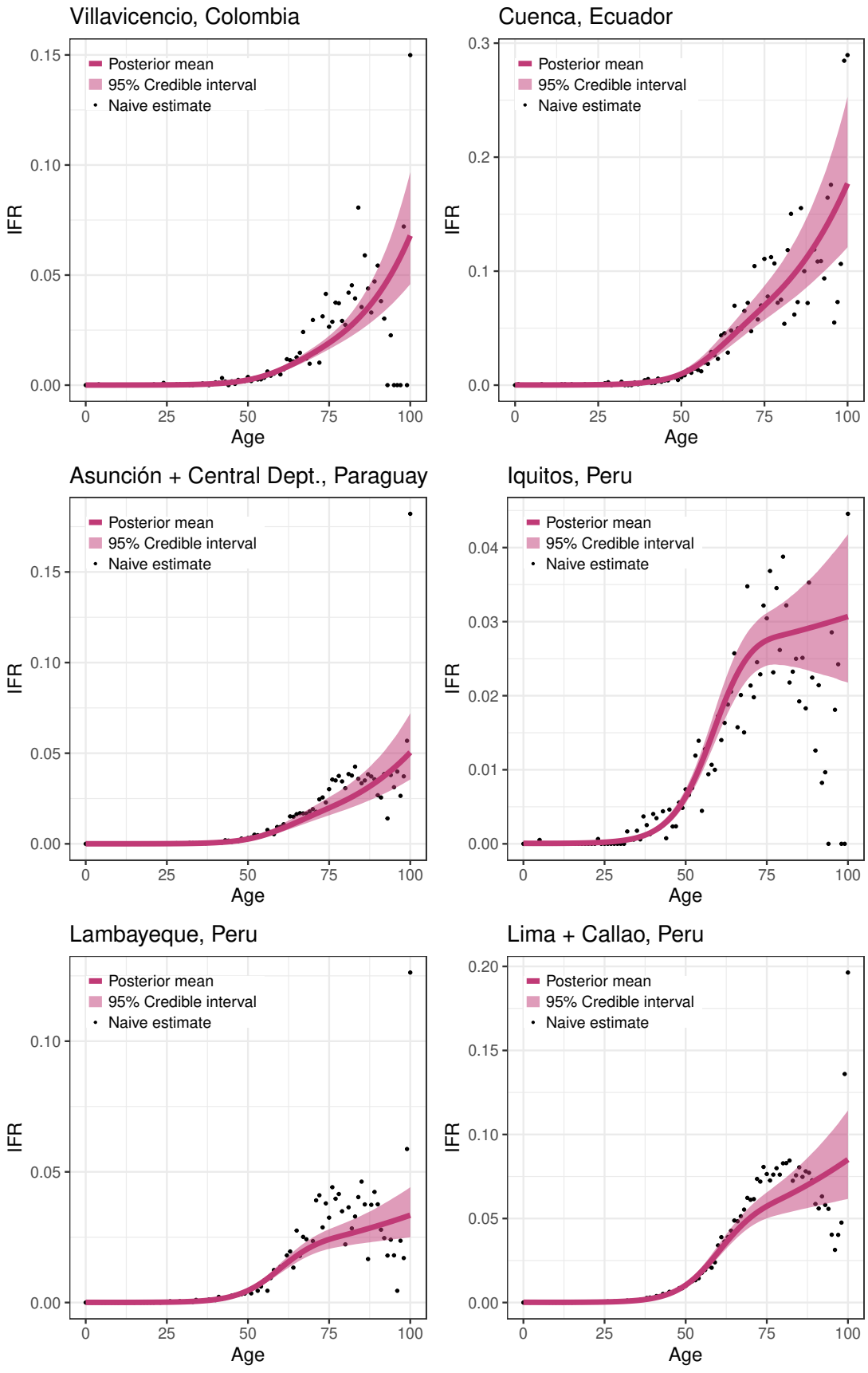
C.5 IFR curves for each location

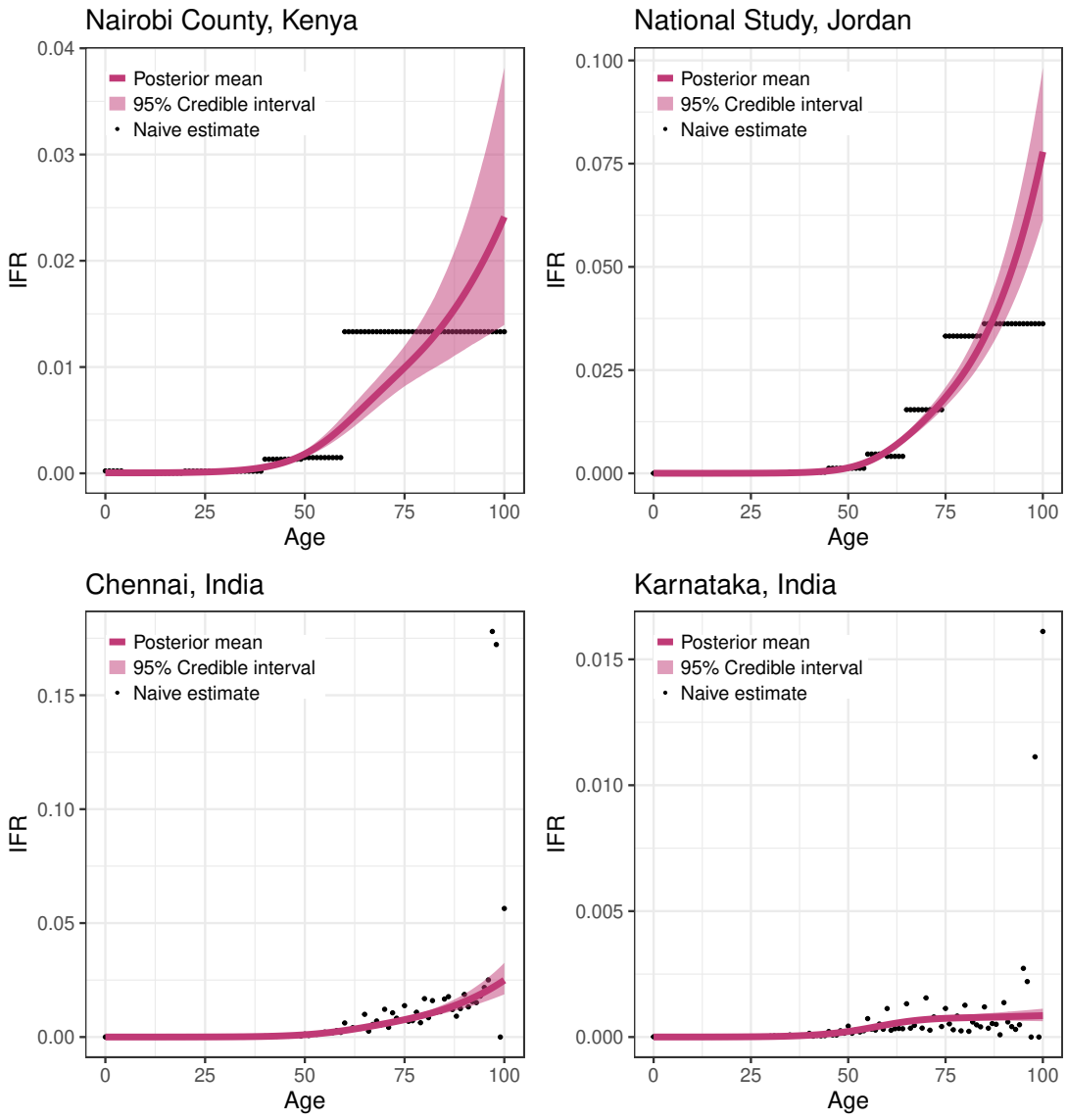
Each of the following plots shows the posterior mean IFR curve with a 95% credible interval. Naive estimates for the IFR are calculated as the empirical death rate divided by the Rogan-Gladen estimator for seroprevalence. Note, when data is only available at the bin level, death and seroprevalence rates are assumed uniform within the age bin for the naive estimate.




Sao Paulo City, Brazil
Várzea Grande, Brazil
3 urban areas, Chile
Barranquilla, Colombia
Bogotá, Colombia
Bucaramanga, Colombia
Posterior mean
95% Credible interval
Naive estimate
IFR
Age


Cali, Colombia
Córdoba, Colombia
Cucuta, Colombia
Ipiales, Colombia
Leticia, Colombia
Medellín, Colombia
Posterior mean
95% Credible interval
Naive estimate
IFR
Age


Villavicencio, Colombia
Cuenca, Ecuador
Asunción + Central Dept., Paraguay
Iquitos, Peru
Lambayeque, Peru
Lima + Callao, Peru
Posterior mean
95% Credible interval
Naive estimate
IFR
Age


Nairobi County, Kenya
National Study, Jordan
Chennai, India
Karnataka, India
Posterior mean
95% Credible interval
Naive estimate
IFR
Age
0.04
0.03
0.02
0.01
0.00
0.100
0.075
0.050
0.025
0.000
0.15
0.10
0.05
0.015
0.010
0.005
0
25
50
75
100

C.6 Data and parameter notation

Data	
$R^{\star}_{\ell,A^R_{\ell,b}}$	Number of individuals in age bin $A^R_{\ell,b}$ for location ℓ that tested positive for COVID-19 antibodies
$n_{\ell,A^R_{\ell,b}}$	Sample size of the serology study in age bin $A^R_{\ell,b}$ for location ℓ
$D^{\star}_{\ell,A^D_{\ell,b}}$	Number of COVID-19 deaths in age bin $A^D_{\ell,b}$ for location ℓ
$N_{\ell,A^D_{\ell,b}}$	Population in age bin $A^D_{\ell,b}$ for location ℓ
$x_{\text{sens},t}$	Number of positive controls for test t that correctly tested positive
$n_{\text{sens},t}$	Number of positive controls for test t
$x_{\text{spec},t}$	Number of negative controls for test t that correctly tested negative
$n_{\text{spec},t}$	Number of negative controls for test t
$f_\ell(a)$	Population density at age a for location ℓ
$\boldsymbol{z}'_{\ell,a}$	Location and age-specific covariates for seroprevalence
$\boldsymbol{x}'_{\ell,a}$	Location and age-specific covariates for IFR
Parameters	
$p_{\ell,A^R_{\ell,b}}$	Probability an individual at location ℓ in age bin $A^R_{\ell,b}$ tests positive for COVID-19 antibodies
$p_\ell(a)$	Probability an age a individual in location ℓ tests positive for COVID-19 antibodies
$\pi_\ell(a)$	Seroprevalence at age a for location ℓ
$\gamma_{\ell,0}$	Intercept of the seroprevalence curve, on the logit scale, for location ℓ
$\boldsymbol{\gamma}_\ell = (\gamma_{\ell,1}, \ldots, \gamma_{\ell,p})'$	Coefficients of the seroprevalence curve at location ℓ
sens_t	Sensitivity for test t
spec_t	Specificity for test t
$\Lambda_{\ell,A^D_{\ell,b}}$	Probability of a COVID-19 death at location ℓ in age bin $A^D_{\ell,b}$
$\text{IFR}_\ell(a)$	Infection fatality rate at age a for location ℓ

$\beta_{\ell,0}$	Intercept of the IFR curve, on the log scale, for location ℓ
$\boldsymbol{\beta}_\ell = (\beta_{\ell,1}, \ldots, \beta_{\ell,q})'$	Coefficients of the IFR curve at location ℓ
$\beta_{\text{global},0}$	Global intercept for IFR
σ_0^2	Variance of location-specific intercepts for IFR
$\beta_{\text{country},c_\ell}$	Country specific effect on the IFR intercept for the country, c, that contains location ℓ
$\sigma^2_{\text{country}}$	Variance of the country effects for the IFR intercepts
$\beta_{\text{global},i}$	i^{th} global coefficient for IFR
σ_i^2	Variance of the i^{th} location-specific coefficients for IFR

C.7 Model summary

$$
\begin{aligned}
R^{\star}_{\ell,A^R_{\ell,b}} &\sim \text{Binomial}(n_{\ell,A^R_{\ell,b}}, p_{\ell,A^R_{\ell,b}}) \\
p_{\ell,A^R_{\ell,b}} &= \int_{A^R_{\ell,b}} p_\ell(a) \frac{f_\ell(a)}{\int_{A^R_{\ell,b}} f_\ell(x)dx} da \\
p_\ell(a) &= \pi_\ell(a)\text{sens}_{t_\ell} + (1-\pi_\ell(a))(1-\text{spec}_{t_\ell}) \\
\text{logit}\left(\pi_\ell(a)\right) &= \gamma_{\ell,0} + \boldsymbol{z}'_{\ell,a}\boldsymbol{\gamma}_\ell \\
\gamma_{\ell,0} &\sim \mathcal{N}(-1, 1.5) \\
\gamma_{\ell,j} &\sim \mathcal{N}(0, 0.05) && \text{for } j \in \{1,2\} \\
x_{\text{sens},t} &\sim \text{Binomial}(n_{\text{sens},t}, \text{sens}_t), \\
x_{\text{spec},t} &\sim \text{Binomial}(n_{\text{spec},t}, \text{spec}_t) \\
\text{sens}_t &\sim \text{Beta}(10, 1) && \text{for } t \in \{1,\ldots,13\} \\
\text{spec}_t &\sim \text{Beta}(50, 1) && \text{for } t \in \{1,\ldots,13\}
\end{aligned}
$$

$$
\begin{aligned}
D^{\star}_{\ell,A^D_{\ell,b}} &\sim \text{Poisson}(N_{\ell,A^D_{\ell,b}}\Lambda_{\ell,A^D_{\ell,b}}) \\
\Lambda_{\ell,A^D_{\ell,b}} &= \int_{A^D_{\ell,b}} \pi_\ell(a) \times \text{IFR}_\ell(a) \frac{f_\ell(a)}{\int_{A^D_{\ell,b}} f_\ell(x)dx} da \\
\log\left(\text{IFR}_\ell(a)\right) &= \beta_{\ell,0} + \boldsymbol{x}'_{\ell,a}\boldsymbol{\beta}_\ell \\
\beta_{\ell,0} &\sim \mathcal{N}(\beta_{\text{global},0} + \beta_{\text{country},c_\ell}, \sigma_0^2), \\
\beta_{\ell,i} &\sim \mathcal{N}(\beta_{\text{global},i}\sigma_i^2), && \text{for i } \in \{1,\ldots,q\} \\
\beta_{\text{country},c_\ell} &\sim \mathcal{N}(0, \sigma_{\text{country}}) \\
\sigma_{\text{country}} &\sim \text{half-normal}(0, 2) \\
\beta_{\text{global},i} &\sim \mathcal{N}(0, 5) && \text{for } i \in \{0,\ldots,3\} \\
\sigma_i &\sim \text{half-normal}(0, 2) && \text{for } i \in \{0,\ldots,3\}
\end{aligned}
$$