Estimating seroprevalence with imperfect serological tests: exploiting cutoff-free approaches

Judith A Bouman[1,*], Sebastian Bonhoeffer[1], and Roland R Regoes[1,*]

[1]Institute of Integrative Biology, ETH Zurich, Zurich, Switzerland
[*]Corresponding authors: Judith A. Bouman and Roland R. Regoes,
(judith.bouman@env.ethz.ch, roland.regoes@env.ethz.ch).

May 18, 2020

Abstract

Large-scale serological testing in the population is essential to determine the true extent of the current Coronavirus pandemic. Serological tests measure antibody responses against pathogens and define cutoff levels that dichotomize the quantitative test measures into sero-positives and negatives. With the imperfect tests that are currently available to test for past SARS-CoV-2 infection, the fraction of seropositive individuals in serosurveys is a biased estimator of seroprevalence and is usually corrected post-hoc to account for the sensitivity and specificity. Here we use a likelihood-based inference method — previously called *mixture models* — for the estimation of the seroprevalence that does not require to define cutoffs by integrating the quantitative test measures directly into the statistical inference procedure. We confirm that this likelihood-based method outperforms the methods based on cutoffs and post-hoc corrections leading to less variation in point-estimates of the seroprevalence and its temporal trend. We illustrate how the likelihood-based method can be used to optimize the design of serosurveys with imperfect serological tests. We also provide guidance on the number of control and case sera that are required to quantify the test's ambiguity sufficiently to enable the reliable estimation of the seroprevalence. Lastly, we show how this approach can be used to identify classes of case sera with an unknown distribution of quantitative test measures that have not been used for test validation. An R-package with the likelihood and power analysis functions is provided. Our study advocates to using serological tests without cutoffs, especially if they are used to determine parameters characterizing populations rather than individuals. This approach circumvents some of the shortcomings of cutoff-based methods with post-hoc correction at exactly the low seroprevalence levels and test accuracies that we are currently facing in COVID-19 serosurveys.

Author Summary

As other pathogens, SARS-CoV-2 elicits antibody responses in infected people that can be detected in their blood serum as early as a week after the infection until long after recovery. The presence of SARS-CoV-2 specific antibodies can therefore be used as a marker of past infection, and the prevalence of seropositive people, i.e. people with specific antibodies, is a key measure to determine the extent of the Coronavirus pandemic. The serological tests, however, are usually not perfect, yielding false positive and negative results. Here we exploit an approach that refrains from classifying people as seropositive or negative, but rather compares the antibody level to that of confirmed cases and controls. This approach leads to more reliable seroprevalence estimates, especially for the low prevalence and low test accuracies that we face during the current Coronavirus pandemic. We also show how this approach can be extended to infer the presence of cases that have not been used for validating the test, such as people that underwent a mild or even asymptomatic infection.

1 Introduction

Serological tests can be used to determine the infection history on two levels: the level of single individuals or that of an entire population. If infection leads to immunity, determining the infection history of an individual can be very useful. In the context of the Coronavirus crisis, for example, it has been suggested that seropositive people could go back to work without posing a risk to themselves or others. In Germany and Italy, among other countries, there is even a discussion about issuing certificates of seropositivity [1, 2]. Such a use of serological tests requires very high test quality, in particular a low false positive rate.

Another, to this date, more common use of serological tests are serosurveys to determine the total number of cumulative infections in a population. Serosurveys are usually conducted after an epidemic, leaving enough time for test refinement. In the current situation, however, an estimate of the true extent of the COVID-19 pandemic in different countries is urgently needed to make policy decisions. This pressure may limit the time available to improve the tests.

Currently available serological tests have high specificity and relatively low sensitivity [3, 4], as expected for this type of testing [5]. Studies which evaluated the tests on severe cases report high specificity and sensitivity [6], however, the discriminatory power of the tests is less clear for mild cases and even asymptomatic cases [7].

Less than ideal sensitivity and specificity are known to introduce biases in the estimation of seroprevalence. Post-hoc corrections of the binomial estimator of seroprevalence have been developed and refined over the past decades [8, 9, 10]. Such a correction has recently been used in the estimation of the seroprevalence in the Coronavirus pandemic [11]. A Bayesian version of the correction has been proposed to infer age-stratified seroprevalences in the context of epidemiological transmission models [12]. However, the corrections work well only for high seroprevalence and relatively high specificity.

In this study, we use an alternative approach to deal with imperfect tests when estimating seroprevalence, this approach has been used before for several other pathogens [13, 14, 15, 16, 17, 18, 19]. Rather than relying on dichotomized serological test data, i.e. a positive or negative test result for each individual in the cohort, and correcting post-hoc for the test's sensitivity and specificity, this method uses the quantitative test measures together with data on the distribution of these measures in controls and conformed cases. Such quantitative measures differ depending on the diagnostic techniques used. In ELISA-based serological tests, fluorescence is determined in units of optical density or "arbitrary units". Tests involving neutralization assays usually report the serum titers needed to achieve viral neutralization of a certain level, e.g. NT50. The method works irrespective of the type of assay that is used, and we refer to the assay results generically as the *quantitative test measures*.

We show that this inference framework enables us to estimate seroprevalence and its change over time without bias. We confirm that cutoffs introduced to dichotomize serological tests results lead to strong biases. Most importantly, however, the alternative inference method outperforms the classical Rogan-Gladen corrections, especially for low seroprevalence and low test accuracies.

We demonstrate how the alternative inference method can be used in power analyses of serosurveys. Specifically, we determine the minimal sample size needed to estimate the seroprevalence with less than 25% deviation of its true value for tests of a given accuracy. We also investigate how many control and case sera are required to quantitatively capture the distribution of test measures in these two groups with enough certainty to allow the reliable estimation of the seroprevalence. Moreover, we introduce an extended likelihood that allows the detection of a discrepancy between the serological observations in the epidemiological surveys and the quantitative test measures collected during the validation phase of the of the serological assay. This does not only allow us to identify a missing case distribution of, for instance, mild cases, but also to estimate the relative prevalence of those missing cases.

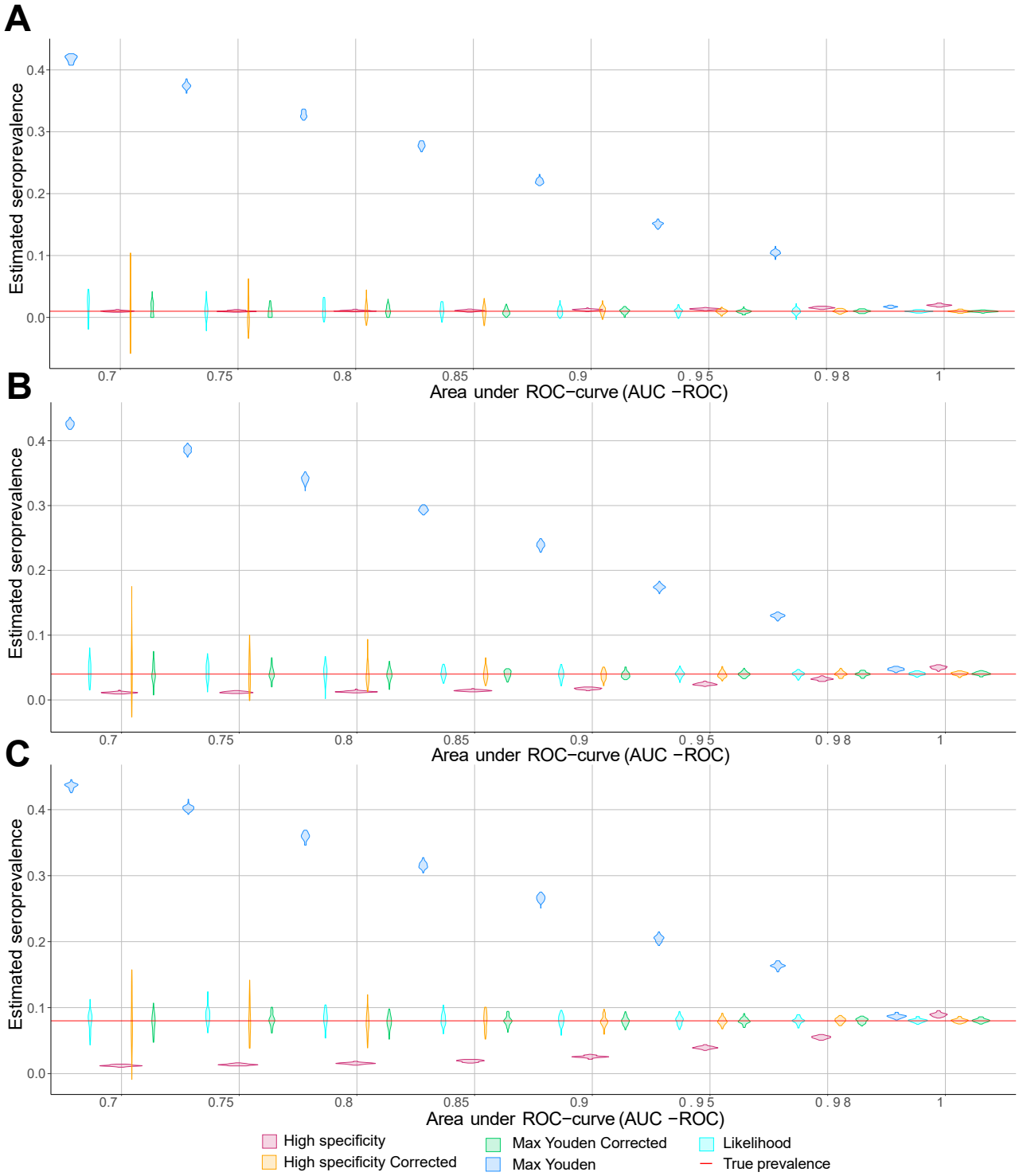

Figure 1: Point-estimates of seroprevalence calculated by the cutoff and likelihood-based methods. The violin plots show the distribution of point-estimates for 50 *in silico* serosurveys conducted with cohorts of 10′000 virtual individuals. The dashed red line shows the true seroprevalence we assumed in the simulated serosurveys.

2 Results

Likelihood-based method that relies on quantitative test measures outperforms corrected cutoff-based methods

To compare the likelihood-based inference method to classical cutoff-based methods without and with the Rogan-Gladen correction, we simulated serosurveys conducted with serological tests of varying accuracies. As a proxy for the accuracy, we used the area under the ROC-curve (AUC-ROC) and varied it from 0.7 to 1. This range is consistent with the AUC-ROC values of many diagnostic tools across disciplines [20]. (The sensitivity and specificity corresponding to the standard cutoffs across the range of AUC-ROC values we consider here are given in Fig S2.) In the simulated serosurveys, we assumed seroprevalences of 1%, 4% and 8% and involved 10′000 virtual individuals. We then derived estimates of the seroprevalence with cutoff-based methods without and with the Rogan-Gladen correction, and the likelihood-based method.

The results of these analyses are shown in Fig 1. Our analysis confirms the strong biases in the estimate of the seroprevalence derived with the two traditional cutoff-based methods without correction. For the 'high specificity'-method, in which the specificity is set to 99% and the sensitivity is calculated according to the assumed test accuracy, the prevalence is underestimated. For the method where the cutoff maximizes the Youden index, seroprevalence is overestimated. Additionally, the dichotomization by a cutoff leads to a reduction in the variation of the point-estimates. This seeming increased reliability of the estimates, however, is an artefact of the dichotomization, and should not be interpreted as an advantage of these methods.

We find that the common Rogan-Gladen correction [8] (see Methods) largely alleviates the biases in the estimation of seroprevalence (see Fig 1). Note that for low prevalence levels or low test accuracy, the corrected point estimate of the prevalence can become negative. This is due to the fact that the number of observed seropositives — being a realization of a stochastic process — can be smaller than the expected number of false positives. For the same reason, the Rogan-Gladen correction give rise to more variable estimates compared to the likelihood-based method for low seroprevalence levels and low test accuracies. In contrast, the likelihood-based method does not inflate the variation of point-estimates and does not result in negative seroprevalence estimates for any true seroprevalence level or test accuracy.

Note, that the variation in point estimates is strongly correlated with the variation of bootstrap estimates (see Fig S3). Thus, the insights into the variation are also informative on the width of the confidence intervals for each method. The variability in the Rogan-Gladen corrected point-estimates is very similar to the variability in the estimates obtained by using the Bayesian correction (see Methods and Fig S4 A). However, the confidence intervals of the Rogan-Gladen corrected point-estimates are larger than the credible intervals obtained for the Bayesian corrections (Fig S4 B).

Cutoff-based methods lead to less reliable estimates of the temporal trends in seroprevalence

In the ongoing Coronavirus pandemic, the seroprevalence is still increasing, and an estimation of its temporal trend is an urgent public health objective. To assess how well the different methods can estimate temporal trends in seroprevalence, we simulated serosurveys during an ongoing epidemic. In particular, we assumed that the seroprevalence increases from 1.5% to 15%.

We found that the uncorrected cutoff-based method both underestimate the temporal trend in seroprevalence (Fig 2). The correction removes the bias in the estimate, but introduces a large variation in the point-estimates of the temporal trend. The estimate for the temporal trend can even be negative for the reason mentioned above. In contrast, the likelihood-based method leads to more reliable point-estimates the temporal trend without inflating their variation.

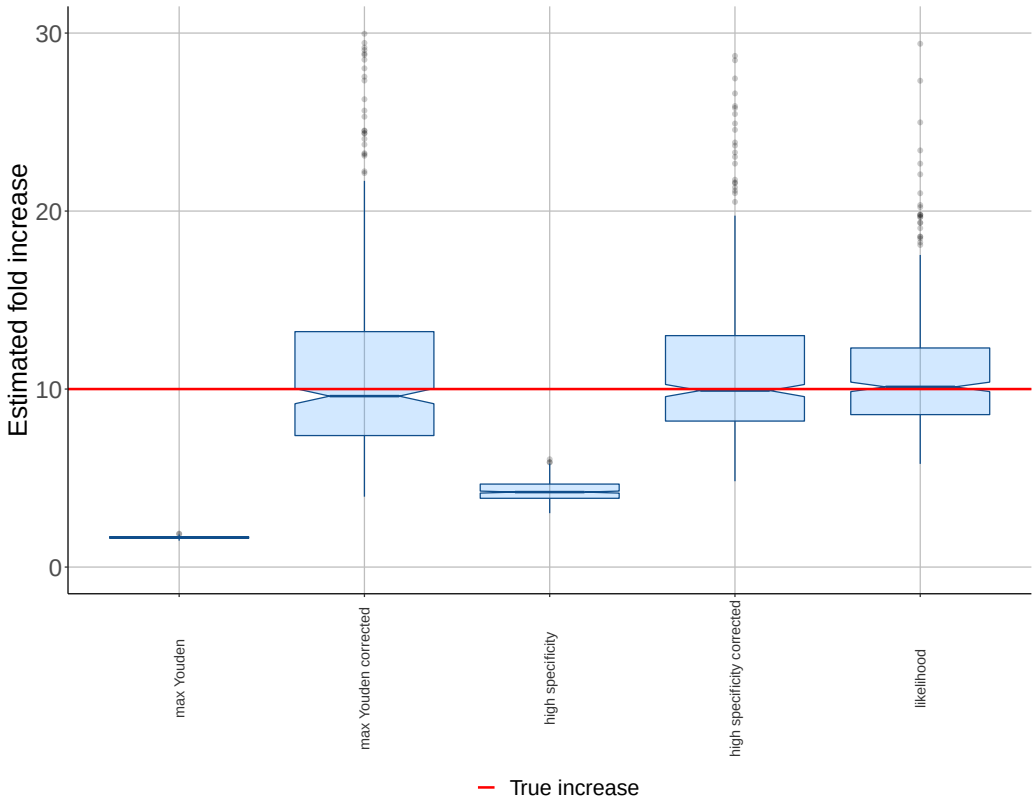

Figure 2: Estimated fold increases in seroprevalence for the cutoff (both corrected and uncorrected) and likelihood-based method. In the simulated serosurveys, we assumed the seroprebvalence to increase from 1.5% to 15%, resulting in a true fold increase of 10. The boxes show the distribution of 50 *in silico* serosurveys for both seroprevalence levels conducted with cohorts of 5000 individuals and a test with an AUC-ROC value of 0.95.

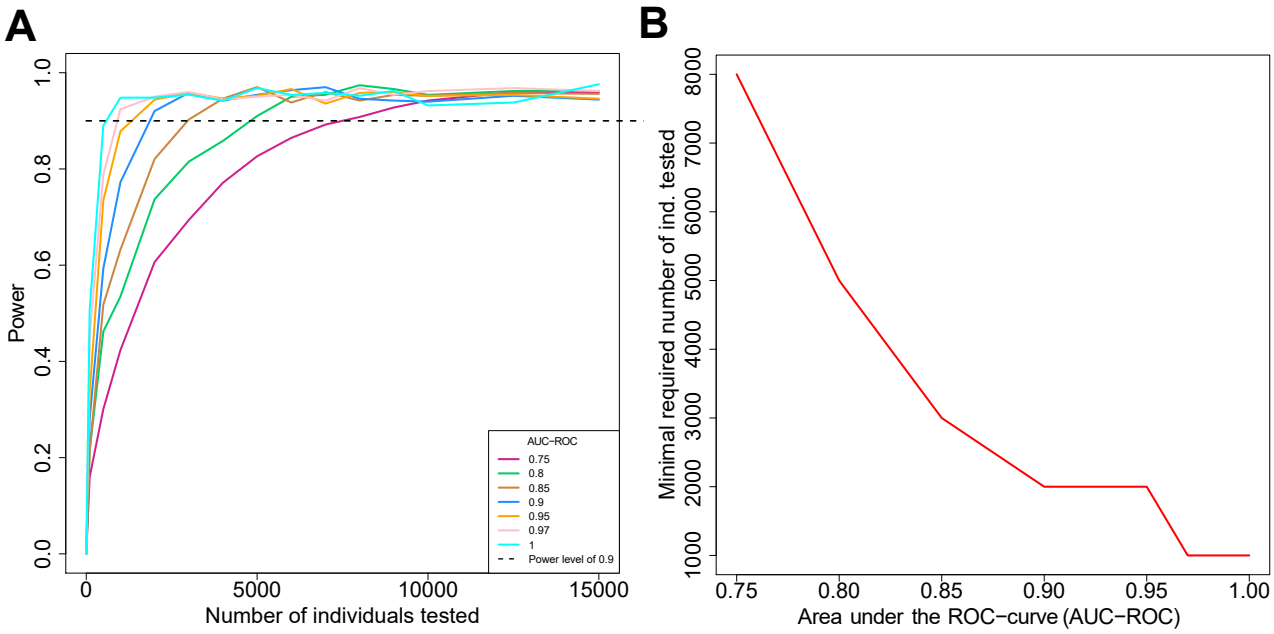

Figure 3: Statistical power of the likelihood-based method. (A) Statistical power versus sample size for varying levels of test accuracy (AUC-ROC). The power is calculated as the fraction of simulated serosurveys that resulted in a seroprevalence estimate that is within 25% of the true seroprevalence and for which the true seroprevalence level lies within 2 standard deviations of the estimated value. Each point in the graph represents the result of 500 *in silico* serosurveys. (B) The minimal number of virtual individuals necessary to obtain a statistical power of 0.9 over a range of AUC-ROC values.

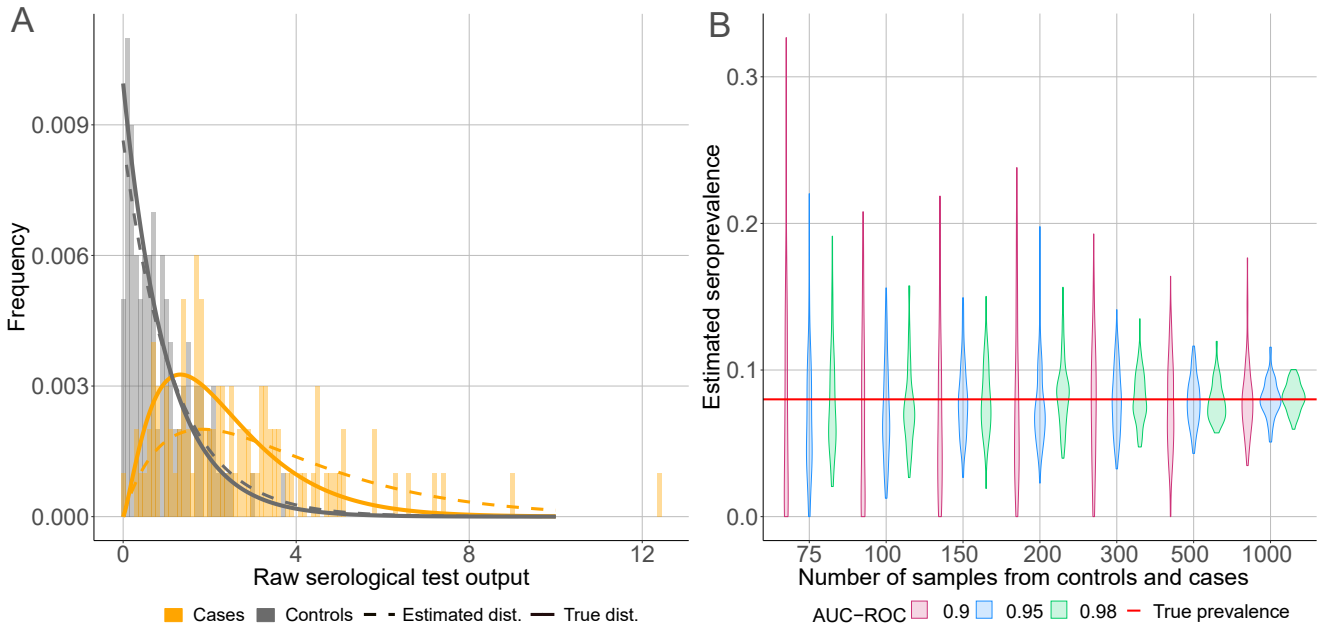

Figure 4: Point estimates of seroprevalence for various numbers of control and case sera used to calibrate the serological test. (A) An example of the result of testing the sera of 50 controls and cases to calibrate their distributions of quantitative test measures. (B) Each violin shows the distribution of the estimated seroprevalence of 50 *in silico* serosurveys conducted with cohorts of 10′000 virtual individuals. The dashed red line shows the true seroprevalence we assumed in the simulated serosurveys (8%).

Low specificity/sensitivity can be compensated by enrolling more people

To understand how the test accuracy affects the statistical power of serosurveys for the likelihood-based method, we simulated many serosurveys with tests characterized by varying AUC-ROC values and evaluated their success. A serosurvey was defined to be successful if the estimate of the seroprevalence was sufficiently close to the true seroprevalence (see Methods). The statistical power is defined as the fraction of successful *in silico* serosurveys.

We find that low accuracy of the serological test can be compensated by higher sample sizes (Fig 3A). For example, to achieve a statistical power of 0.9, 1000 individuals need to be enrolled into the serosurvey for a high-accuracy test (AUC-ROC=0.98). In contrast, for a lower accuracy test (AUC-ROC=0.8) 5000 are required (Fig 3B).

The number of case and control sera used for validation of serological tests impacts the reliability of the seroprevalence estimate

The likelihood-based method relies directly on the quantitative test measures and their distribution for control and case sera. Therefore, the reliability of the methods are expected to depend on the precision with which these distributions have been determined. The precision depends on the number of control and case sera used in test validation.

To assess the required number of cases and controls that should be sampled to obtain precise test measure distributions, and hence reliable estimates of the seroprevalence, we simulated the validation of serological tests with various numbers of control and case sera (see Methods). In brief, test measures for control and case sera are drawn from "true" distributions. Subsequently the test measures are used to reconstruct the distribution of control and case test measures. This reconstruction has its own uncertainty, which should be

included in the estimation of seroprevalence.

We find that we require at least 150 control and case sera to obtain sufficiently reliable seroprevalence estimates (Fig 4). Clearly, the number of required samples decreases when the accuracy of the test, measured by AUC-ROC, increases. For a true prevalence of 8%, as is used in Fig 4, and a AUC-ROC value of 0.9, a high number of samples is needed to obtain a small confidence band around the estimate. However, for a AUC-ROC value of 0.95, sampling 150 cases and controls would be sufficient. In this analysis, we have focused on the point estimates and their variation as a proxy for the reliability of seroprevalence estimation. As we mentioned above, the variation in point-estimates is strongly correlated with the variation in bootstrap estimates of a single point-estimate (see Fig S3). Thus, our results on the variation of the point estimates apply equally to the confidence intervals.

How can we incorporate the precision, with which the test measure distributions have been determined, into a confidence interval of the seroprevalence? We suggest a two-step bootstrap procedure. First, we resample the test measures of control and case sera, and use the resampled dataset to obtain bootstrap estimates of the parameters of the test measure distributions. Second, we resample the test measures of the serosurvey, and, using the bootstrap estimate of the control and case distributions of test measures, obtain bootstrap estimates of the seroprevalence. This procedure makes sure that the uncertainties arising from the lack of test accuracy as well as the lack of precision in the test measure distributions are appropriately accounted for in the estimation of the seroprevalence.

Alternatively, a one-step Bayesian procedure can be used that estimates the parameters of the control and case sera distributions as well as the seroprevalence at once by combining the various likelihoods for these processes. We provide an R-function for both the two-step and the one-step procedure we propose in the supplementary R-package.

Detecting discrepancies between test validation and serosurvey data

The reliance of the likelihood-based method on quantitative test data, allows us to investigate inconsistencies between the populations used for validating the test on the one hand, and the population recruited into the serosurvey on the other. For example, assume that the serological test has been validated only with sera from individuals who underwent a severe infection, in addition to sera from controls. Can we adopt the likelihood framework to detect the presence of individuals who underwent a mild infection with a distribution of test measures different from severe cases? Could we even estimate their seroprevalence?

To investigate this question, we set up a simulation assuming three distributions of test measures, corresponding to severe and mild cases and controls. We further assume that the total seroprevalence is 10%, of which 14% were severe infections, consistent with observations at the beginning of the COVID-19 epidemic in China [21]. We then simulate a serosurvey with 10'000 individuals that results in quantitative test measures randomly drawn from these three distribution according to their seroprevalences.

We then analyse the simulated serosurvey data assuming a likelihood based on quantitative test measure of only severe cases and controls. This model misspecification reflects a scenario, in which a large fraction of the infected population has been neglected for test validation. As expected, this analysis underestimates the true seroprevalence (see Fig 5 A). The misspecified likelihood, however, is valuable as a null model against which more complex likelihood can be tested.

To test for the presence of mild cases in the serosurvey, we analyse the simulated serosurvey data with a likelihood that extends the misspecified, two-distribution likelihood by an additional unknown distribution of test measures in mild cases. A likelihood ratio test can then be used to identify the existence of this population of mild cases, even though it was ignored during test validation. This procedure is generic and does not require that the ignored cases be mild. It can detect any type of cases as long as their test measure distribution differs from those included in the test validation.

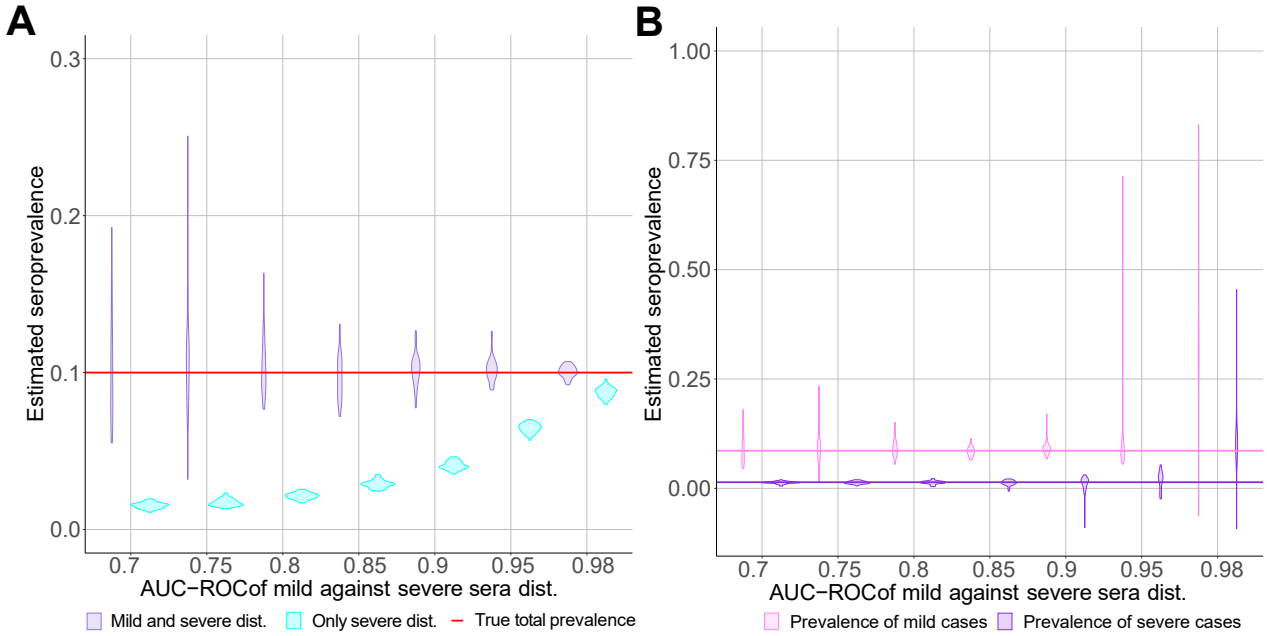

Figure 5: Estimates of the seroprevalence in a population with uninfected, as well as mildly and severely infected individuals. The x-axes represent the AUC-ROC value between the mild and severe case distribution. Each violin represents the result of 50 simulated serosurveys with 10′000 individuals per serosurvey. (A) Cyan violins show estimates of the total seroprevalence based on a likelihood that ignores mild cases, whereas purple violins show estimates with a likelihood including mild cases. The true total prevalence of severe and mild infections is 0.1. (B) The estimated seroprevalence of the mild (light purple) and the severe (dark purple) cases. The true prevalence of mild infections is 0.086 and the prevalence of severe cases is 0.014.

Lastly, the fit of the three-distribution likelihood yields estimates for the seroprevalences of mild and severe cases, and the mean and variance of the test measure distribution of mild cases. This analysis — without knowledge of the third distribution — succeeds in estimating the total seroprevalence without bias (Fig 5 A). It also yields estimates for the seroprevalences for mild and severe cases separately (Fig 5 B). The identifiability of the individual seroprevalences depends strongly on how distinct the test measure distributions of mild and severe cases and the controls are.

3 Discussion

In this study, we analyze a likelihood-based method proposed before for various other viruses that allows the unbiased estimation of seroprevalence in a population [13, 14, 15, 16, 17, 18, 19]. The method relies on the quantitative serological test measurements before they have been dichotomized, i.e. categorized as "positive" or "negative" using a cutoff.

We confirm the well-known fact that such a dichotomization leads to strong biases in the estimate of seroprevalence and its time trends, when estimates are not corrected for the sensitivity and specificity of the serological test [8]. Even though the commonly used Rogan-Gladen and Bayesian corrections alleviate the biases, the estimates of the seroprevalence are generally less reliable than they could be. Estimates with the Rogan-Gladen correction can even be negative for low test accuracies and for the low seroprevalences that we expect in the current Coronavirus pandemic.

We show that the likelihood-based method corrects the biases in a way that leads to less variable estimates. It is also applicable through the entire range of test accuracies and seroprevalences, and should therefore be of value specifically for COVID-19 serosurveys.

As inputs, the likelihood function accepts any common type of quantitative test measures of specific antibody levels, such as optical densities or "arbitrary units" in ELISAs or neutralization titers that inhibit viral replication to any degree specified (e.g NT50 or NT90) in neutralization assays. It is not essential that these measures are linearly correlated with the level of antibodies. It is important, however, that the isotype, units and scales are the same as those for the test measure distributions of control and case sera. While the likelihood-based method has been developed to increase the reliability of seroprevalence estimates especially for tests with low accuracy, it also works for perfect tests. Also in this case, a definition of a cutoff discriminating test measures into seropositives and negatives is not necessary for estimating the seroprevalence with the likelihood-based method.

Paired with the simulation of serosurveys, the likelihood-based method can provide guidance on the design of serosurveys. While we find that a low test accuracy can be compensated by higher sample sizes in a serosurvey, the increase in required sample size can be large (see Fig 3). Thus, investing in test refinement should be carefully weighed against the additional overhead of a larger study. The likelihood-based method may open up avenues for serosurveys in low and middle income countries in cases when highly accurate, and therefore expensive serological tests are not available.

Because the likelihood-based method relies directly on the quantitative test measures, it provides better insights into how the precision, with which the test accuracy has been determined, affects the estimates of seroprevalence. We show that, for the expected current seroprevalence of the Coronavirus pandemic and an optimistically high test accuracy of AUC-ROC= 0.95, approximately 150 control and case sera are required to adequately determine the distribution of their test measures that ensure for sufficiently reliable estimates of a seroprevalence. Again, the effort put into test refinement should be carefully weighed against the effort put into mapping the test accuracy thoroughly. Additionally, we provide a procedure to calculate a confidence interval of the estimated seroprevalence that appropriately accounts for the uncertainties arising from the lack of test accuracy as well as the lack of precision in the test measure distributions.

Similar to the cutoff-based methods, it has been shown the likelihood-based method can be easily extended by categorical or continuous covariates of the seroprevalence, such as sex or age [13, 14, 15, 16, 17, 18, 19]. Also temporal changes in seroprevalence — as predicted, for example, by epidemiological models — can

be incorporated into this framework. The likelihood-based method provides a more natural approach to integrating multiple test measures, such as IgA and IgG levels. In cutoff-based methods this may require complex cutoff functions that depend on the multiple measures and will complicate the determination of sensitivity and specificity.

The likelihood-based method relies on unbiased data from cases. Unbiased means this cohort should contain individuals who have undergone severe *and* mild infections, and the proportions of the infections with different severities should recapitulate the proportions in the population. We have illustrated that an over-representation of severe cases compared to mild cases in the validation of the serological test leads to underestimation of the total seroprevalence (Fig 5). Thus, it is preferable to include cases detected by contact tracing rather than by more biased detection channels, such as hospitalisation. This requirement, however, applies generally to the adequate determination of test sensitivity and specificity that are necessary to determine cutoffs or to apply methods that use post-hoc corrections.

Unlike the cutoff-based methods, however, a method based on the quantitative test measures allows us to infer the presence of cases that were not used for the validation of the test. To this end, the likelihood needs to be extended by an unknown distribution of test measures. In addition to being able to identify a discrepancy of test validation versus serosurvey data, this extended likelihood also allows for the estimation of the seroprevalence of the subpopulation with the unknown distribution (Fig 5). This issue is crucial for a reliable estimation of the true extent of the current COVID-19 pandemic and still unknown. The likelihood can be easily extended to include more than two types of cases. We recommend to always fit a likelihood with an additional unknown distribution and to test its fit against a null likelihood that includes all known case distributions. This guards against strong biases arising from neglecting a large fraction with distinct test measure distributions.

This work as well as other similar studies that investigate the effect of test accuracy on estimating seroprevalence [12], highlight the necessity of integrating the development of serological tests with the design of the serosurveys, in which they will be applied. This is especially relevant during the current Coronavirus pandemic as the serological tests are developed at the same time as the serosurveys are being conducted.

4 Methods

The likelihood The likelihood-based approach to inference relies directly on the quantitative measure obtained from serological tests [14, 15, 16, 17]. It estimates the seroprevalence (π) by maximizing the likelihood of observing those quantitative test measures, given their distribution for case and control sera (see Fig 6A). The likelihood for the data is shown in Eq 1.

$$ll(U) = \sum_{i=1}^{i=n} \log \left(p(\sigma_i = 1|\pi)\, p(U_i|\sigma_i = 1) + p(\sigma_i = 0|\pi)\, p(U_i|\sigma_i = 0) \right) \tag{1}$$

Here, U is a vector with the quantitative test measures of all n tests, σ is a binary vector of length n with their underlying true serological status (1 for infected and 0 for not infected). The probabilities $p(U_i|\sigma_i = 1)$ and $p(U_i|\sigma_i = 0)$ capture the distributions of quantitative test measures of control and case sera, and $p(\sigma_i = 1|\pi)$ and $p(\sigma_i = 0|\pi)$ denote the probability of sampling individuals who have been truly infected or not. The units of U can be anything that is commonly used in serological tests, such as optical densities obtained from ELISAs or neutralization titers obtained in neutralization assays.

Cutoff-based methods Most commonly, the quantitative test measures are dichotomized into seropositives and negatives using a cutoff. There are many ways to estimate a cutoff value for a test [22, 23]. One strategy is to set the cutoff such that the test is highly specific (99%) (see Fig 6C). This is equivalent to minimizing the number of false positives. This will go at the cost of the sensitivity and will lead to more false negatives. This method is referred to as the 'high specificity' method.

Another method that is often used to determine the cutoff is to maximize the Youden index (Youden

index = sensitivity + specificity - 1 [24]) (see Fig 6C). Graphically, this is equivalent to maximizing the distance between the diagonal and the receiver-operator characteristic (ROC) curve (see Fig 6D). This method is referred to as the 'max Youden' method.

The standard Rogan-Gladen correction of the binomial seroprevalence estimate q is[8]:

$$\pi = \frac{q + s - 1}{r + s - 1}. \tag{2}$$

Here, r is the sensitivity and s the specificity of the test, and π is the corrected estimator of seroprevalence. Effectively, this correction adds the expected number of false negative and subtracts the expected number of false positives from the number of observed positives. The correction works best when the expected number of false positives is smaller than the observed number of positives. We apply this correction to the maximal Youden and the high specificity method.

An alternative way to correct for the sensitivity and specificity is described by Larremore et al. (2020). They use a Bayesian approach that samples from a posterior distribution that includes the estimates of the sensitivity and specificity [12]. We also apply this correction to both the maximal Youden and the high specificity method.

Simulations We simulate serosurveys by assuming a given prevalence and distributions of the quantitative test measures for control and case sera. For each virtual individual enrolled in these *in silico* serosurveys, we conduct a Bernoulli trial with the probability set to the true seroprevalence. In a second step, we simulate the serological test of each individual. To this end, we draw a random quantitative test measure from the distribution of either case or control sera — depending on whether the individual is truly seropositive or not.

To simulate serological tests with various accuracy, we use several distributions of quantitative test measures for the case sera. We assume the distribution of quantitative test measures for the control sera to be Γ-distributed with a shape and scale parameter of 1, which results in a mean of 1, and the distribution for case sera to be Γ-distributed with varying means and a rate parameter of 1. The scale parameter of the distribution for case sera determines the amount of overlap between the quantitative test measure of cases with controls, and thus modifies accuracy (Fig 6).

The results of these simulated serosurveys are then analyzed with the likelihood- and the various cutoff-based methods.

Power analysis To estimate the statistical power of a serosurvey, we conducted 500 serosurveys for each given test accuracy and sample size. Here, the power is determined as the proportion of simulations for which the seroprevalence is estimated successfully. We define an estimate to be successful when it deviates less than 25% from the true seroprevalence and the true seroprevalence is within 2 standard deviations of the estimate.

Inferring the distribution of quantitative test measures for control and case sera To capture the uncertainty in the distribution of quantitative test measures, we simulated experiments that aim to determine them. We assume true distributions as specified above. We then draw quantitative test measures for a given number of controls and cases from these true distributions. From these *in silico* data, we estimate the parameters of the distribution for control and case sera. Hereby, we assume that distribution of the measures for control and case sera are Γ-distributed, and we estimate the shape and scale parameter of the distribution of the case sera and the scale parameter of the distribution of the control sera. These estimated distributions are then used to analyse the simulated data.

Incorporating distinct distributions for mild and severe case sera Eq 3 represents a likelihood that allows for the possibility that there are various types of case sera with distinguishable distristributions (See Fig S1) [19]. For the sake of simplicity, we assume that there are two types of case sera: those from individuals

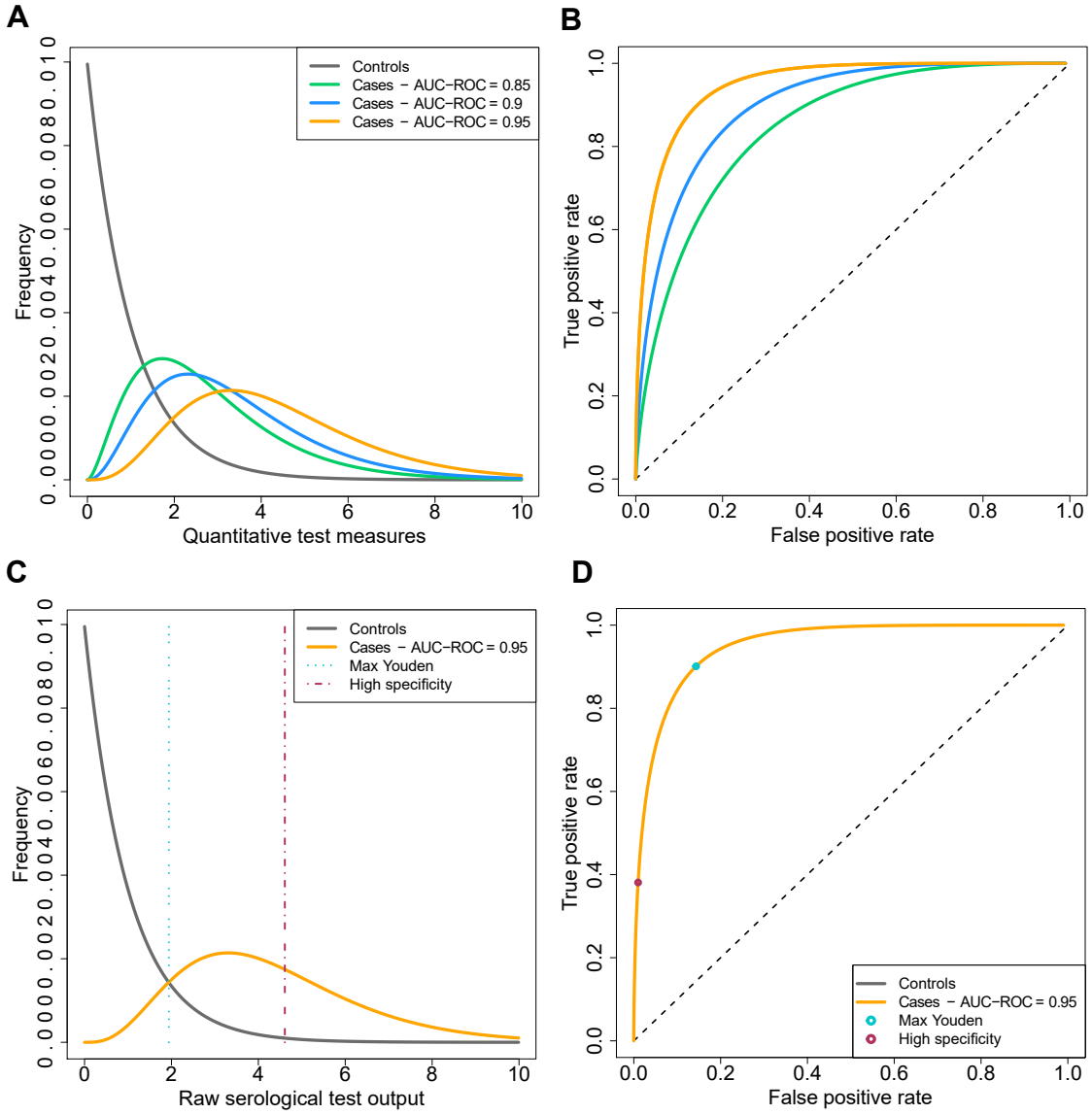

Figure 6: Conceptual diagram of the distribution of the quantitative test measures for control and case sera. (A) Hypothetical probability density of quantitative test measures of control sera and three case sera. (B) ROC-curves corresponding to the distribution of quantitative test measures of the control sera and each of the possible distribution for the case sera. (C) Visualization of the 'maximal Youden' and 'high specificity' cutoffs. (D) Visualization of these cutoffs in the ROC curves.

who underwent a severe infection (severe cases) and those from individuals who underwent a mild infection (mild cases).

$$ll(U) = \sum_{i=1}^{i=n} \log \left(p(\sigma_i = 2|\pi_s, \pi_m)\ p(U_i|\sigma_i = 2) + p(\sigma_i = 1|\pi_m)\ p(U_i|\sigma_i = 1) + p(\sigma_i = 0|\pi_s, \pi_m))\ p(U_i|\sigma_i = 0) \right) \tag{3}$$

Compared to Eq 1, σ_i is now a categorical value that represents whether an individuals is seronegative (0), seropositive due to a mild infection (1) or seropositive due to a severe infection (2). π_{mild} is the seroprevalence of mild infections and π_s the seroprevalence of severe infections. This likelihood is formally equivalent to the one presented in van Boven et al [19], without the age-structure. We used the extended likelihood to estimate the total seroprevalence, the seroprevalence of mild cases and the shape parameter of the distribution of quantitative test measures of the mild case sera.

The identifiability and confidence of the estimates of these three parameters depends on the amount of overlap between the distribution of the mild and severe case sera and the overlap between those and the control case sera. In our simulations, we hold the area under the ROC-curve between the control and severe case sera constant, at a level of 0.95, while we vary the area under the ROC curve between the mild and severe case sera. We set the mean of the distribution of mild cases always to a lower value than that of the severe cases (see Fig S1). Both these means are larger than 1, which is the mean of the distribution of test measures for control sera. We use a scale parameter of 1 for both distributions.

Implementation The likelihood function for estimating seroprevalence as well as the simulations were implemented in the R language for statistical computing [25]. The Bayesian correction is executed with the 'metrop' function in the 'mcmc' package in R [26], with the number of batches equal to 10000 and the length of each batch equal to 100. An R-package containing the code is provided as supplementary material.

5 Acknowledgements

We gratefully acknowledge Claudia Igler, Julien Riou, Christian Althaus and Michiel van Boven for discussions and comments on the manuscript.

References

[1] V. Hackenbroch, "Große Antikörperstudie soll Immunität der Deutschen gegen Covid-19 feststellen," Mar 2020.

[2] J. Horowitz, "In Italy, going back to work may depend on having the right antibodies," Apr 2020.

[3] J. H. C. for Health Security, "Global progress on covid-19 serology-based testing," Apr 2020.

[4] D. Lin, L. Liu, M. Zhang, Y. Hu, Q. Yang, J. Guo, Y. Dai, Y. Xu, Y. Cai, X. Chen, *et al.*, "Evaluations of serological test in the diagnosis of 2019 novel coronavirus (sars-cov-2) infections during the covid-19 outbreak," *medRxiv*, 2020.

[5] A. H. L. Bruning, M. M. G. Leeflang, J. M. B. W. Vos, R. Spijker, M. D. de Jong, K. C. Wolthers, and D. Pajkrt, "Rapid Tests for Influenza, Respiratory Syncytial Virus, and Other Respiratory Viruses: A Systematic Review and Meta-analysis," *Clinical Infectious Diseases*, vol. 65, pp. 1026–1032, 05 2017.

[6] R. Lassaunière, A. Frische, Z. B. Harboe, A. C. Nielsen, A. Fomsgaard, K. A. Krogfelt, and C. S. Jørgensen, "Evaluation of nine commercial sars-cov-2 immunoassays," *medRxiv*, 2020.

[7] N. M. Okba, M. A. Muller, W. Li, C. Wang, C. H. GeurtsvanKessel, V. M. Corman, M. M. Lamers, R. S. Sikkema, E. de Bruin, F. D. Chandler, Y. Yazdanpanah, Q. Le Hingrat, D. Descamps, N. Houhou-Fidouh, C. B. E. M. Reusken, B.-J. Bosch, C. Drosten, M. P. Koopmans, and B. L. Haagmans, "Sars-cov-2 specific antibody responses in covid-19 patients," *medRxiv*, 2020.

[8] W. J. Rogan and B. Gladen, "Estimating prevalence from the results of a screening test," *American journal of epidemiology*, vol. 107, no. 1, pp. 71–76, 1978.

[9] S. Greenland, "Basic methods for sensitivity analysis of biases," *International journal of epidemiology*, vol. 25, no. 6, pp. 1107–1116, 1996.

[10] P. J. Diggle, "Estimating prevalence using an imperfect test," *Epidemiology Research International*, vol. 2011, 2011.

[11] E. Bendavid, B. Mulaney, N. Sood, S. Shah, E. Ling, R. Bromley-Dulfano, C. Lai, Z. Weissberg, R. Saavedra, J. Tedrow, D. Tversky, A. Bogan, T. Kupiec, D. Eichner, R. Gupta, J. Ioannidis, and J. Bhattacharya, "Covid-19 antibody seroprevalence in santa clara county, california," *medRxiv*, 2020.

[12] D. B. Larremore, B. K. Fosdick, K. M. Bubar, S. Zhang, S. Kissler, C. J. E. Metcalf, C. Buckee, and Y. Grad, "Estimating sars-cov-2 seroprevalence and epidemiological parameters with uncertainty from serological surveys," 2020.

[13] A. Steens, S. Waaijenborg, P. F. Teunis, J. H. Reimerink, A. Meijer, M. van der Lubben, M. Koopmans, M. A. van der Sande, J. Wallinga, and M. van Boven, "Age-dependent patterns of infection and severity explaining the low impact of 2009 influenza a (h1n1): evidence from serial serologic surveys in the netherlands," *American journal of epidemiology*, vol. 174, no. 11, pp. 1307–1315, 2011.

[14] P. Teunis, B. Schimmer, D. Notermans, A. Leenders, P. Wever, M. Kretzschmar, and P. Schneeberger, "Time-course of antibody responses against coxiella burnetii following acute q fever," *Epidemiology & Infection*, vol. 141, no. 1, pp. 62–73, 2013.

[15] D. te Beest, E. De Bruin, S. Imholz, J. Wallinga, P. Teunis, M. Koopmans, and M. Van Boven, "Discrimination of influenza infection (a/2009 h1n1) from prior exposure by antibody protein microarray analysis," *PloS one*, vol. 9, no. 11, 2014.

[16] D. E. Te Beest, P. J. Birrell, J. Wallinga, D. De Angelis, and M. van Boven, "Joint modelling of serological and hospitalization data reveals that high levels of pre-existing immunity and school holidays shaped the influenza a pandemic of 2009 in the netherlands," *Journal of The Royal Society Interface*, vol. 12, no. 103, p. 20141244, 2015.

[17] M. A. Vink, J. van de Kassteele, J. Wallinga, P. F. Teunis, and J. A. Bogaards, "Estimating seroprevalence of human papillomavirus type 16 using a mixture model with smoothed age-dependent mixing proportions," *Epidemiology*, vol. 26, no. 1, pp. 8–16, 2015.

[18] M. A. Vink, J. Berkhof, J. van de Kassteele, M. van Boven, and J. A. Bogaards, "A bivariate mixture model for natural antibody levels to human papillomavirus types 16 and 18: baseline estimates for monitoring the herd effects of immunization," *PloS one*, vol. 11, no. 8, 2016.

[19] M. van Boven, J. van de Kassteele, M. J. Korndewal, C. H. van Dorp, M. Kretzschmar, F. van der Klis, H. E. de Melker, A. C. Vossen, and D. van Baarle, "Infectious reactivation of cytomegalovirus explaining age-and sex-specific patterns of seroprevalence," *PLoS computational biology*, vol. 13, no. 9, p. e1005719, 2017.

[20] J. A. Swets, "Measuring the accuracy of diagnostic systems," *Science*, vol. 240, no. 4857, pp. 1285–1293, 1988.

[21] The Joint Mission headed by B. Aylward and W. Liang, "Report of the who-china joint mission on coronavirus disease 2019 (covid-19)," Feb 2020.

[22] X.-H. Zhou, D. K. McClish, and N. A. Obuchowski, *Statistical methods in diagnostic medicine*, vol. 569. John Wiley & Sons, 2009.

[23] M. S. Pepe *et al.*, *The statistical evaluation of medical tests for classification and prediction*. Medicine, 2003.

[24] W. J. Youden, "Index for rating diagnostic tests," *Cancer*, vol. 3, no. 1, pp. 32–35, 1950.

[25] R Core Team, *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2018.

[26] C. J. Geyer and L. T. Johnson, *mcmc: Markov Chain Monte Carlo*, 2020. R package version 0.9-7.

Supplementary Material

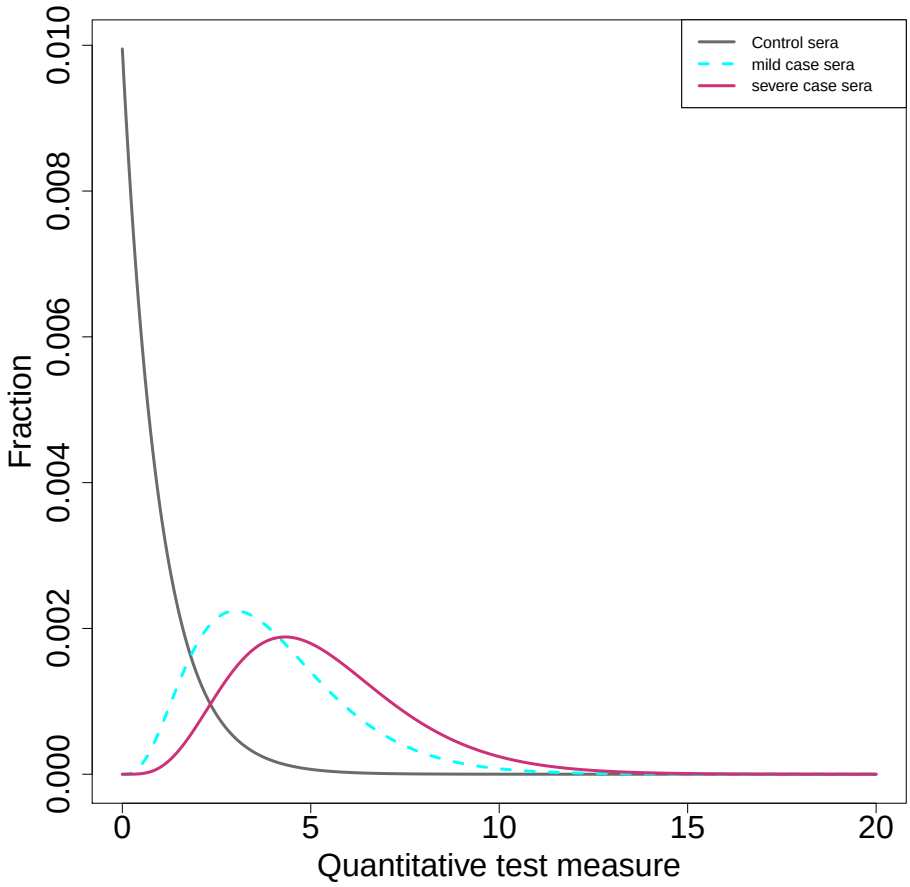

Figure S1: Hypothetical probability density of quantitative test measures of control sera, severe case sera and a potentially distinguishable distribution for the mild case sera.

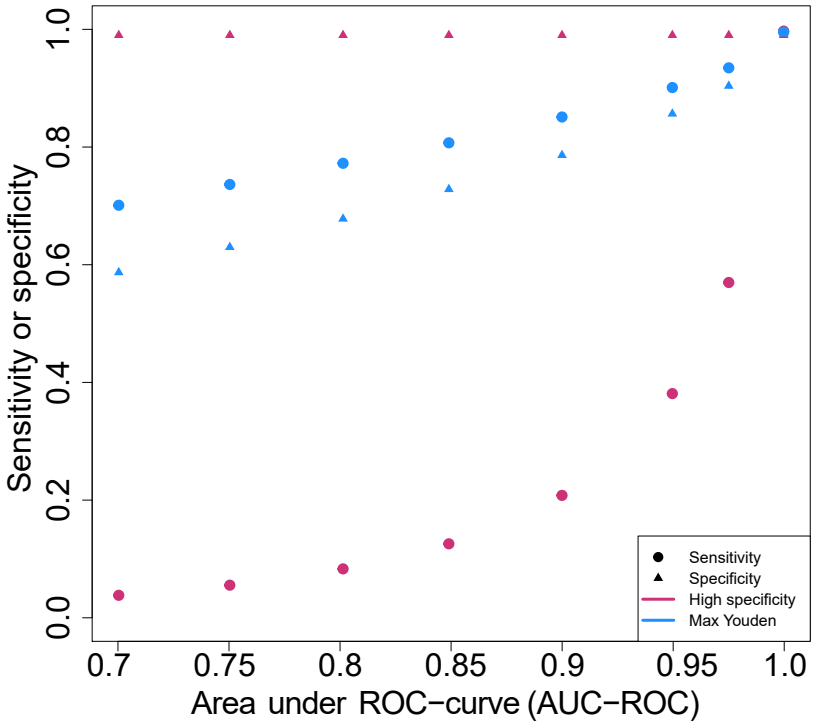

Figure S2: Sensitivity and specificity corresponding to the standard cutoffs across the range of AUC-ROC values we consider in this study.

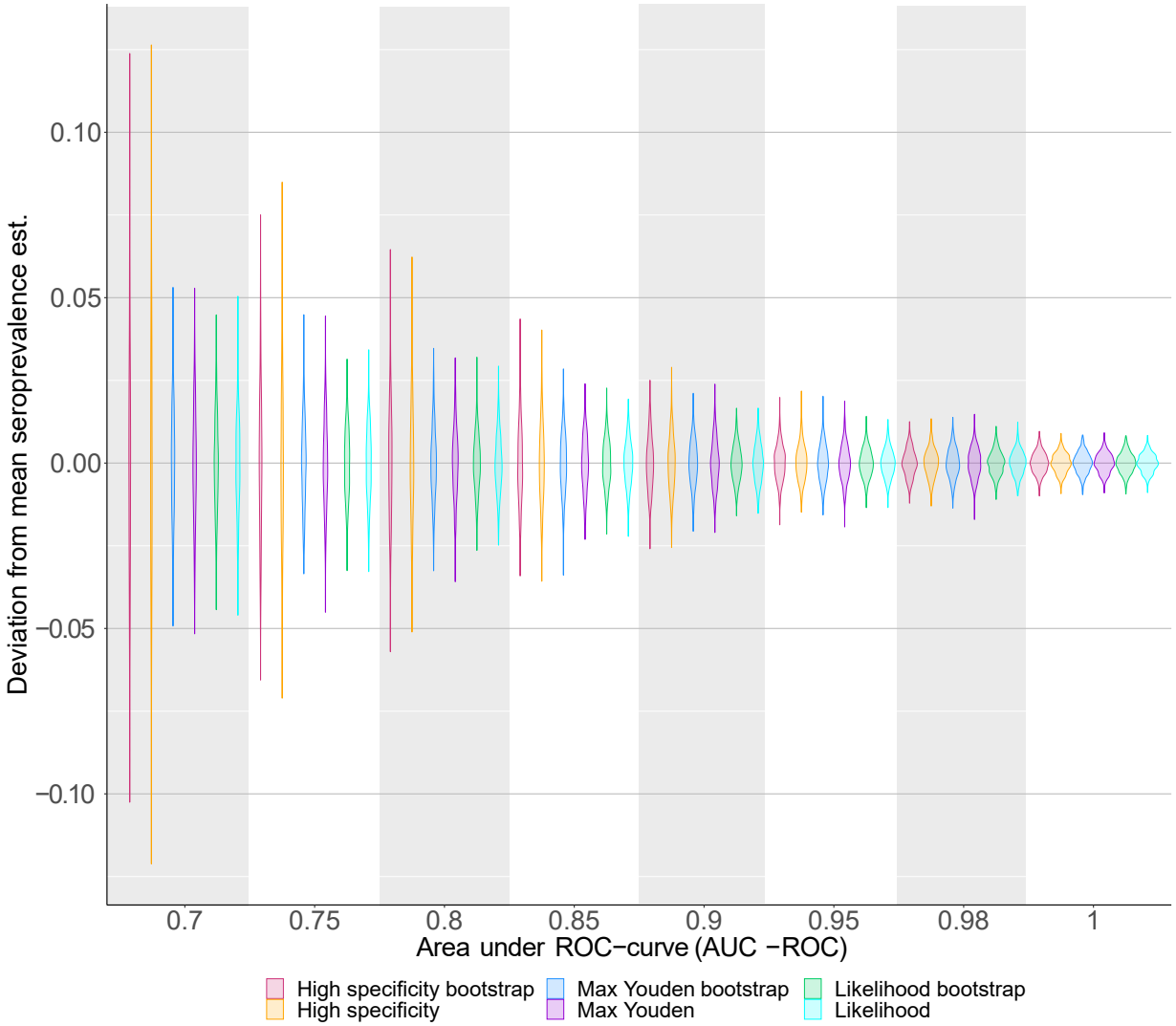

Figure S3: The deviation from the mean seroprevalence estimate for bootstrapping or re-simulating the serosurvey for both cutoff-based and the likelihood-based method. Each violin represents 1000 bootstrap or re-simulating events of a serosurvey with 10′000 individuals, where the true prevalence was equal to 8%.

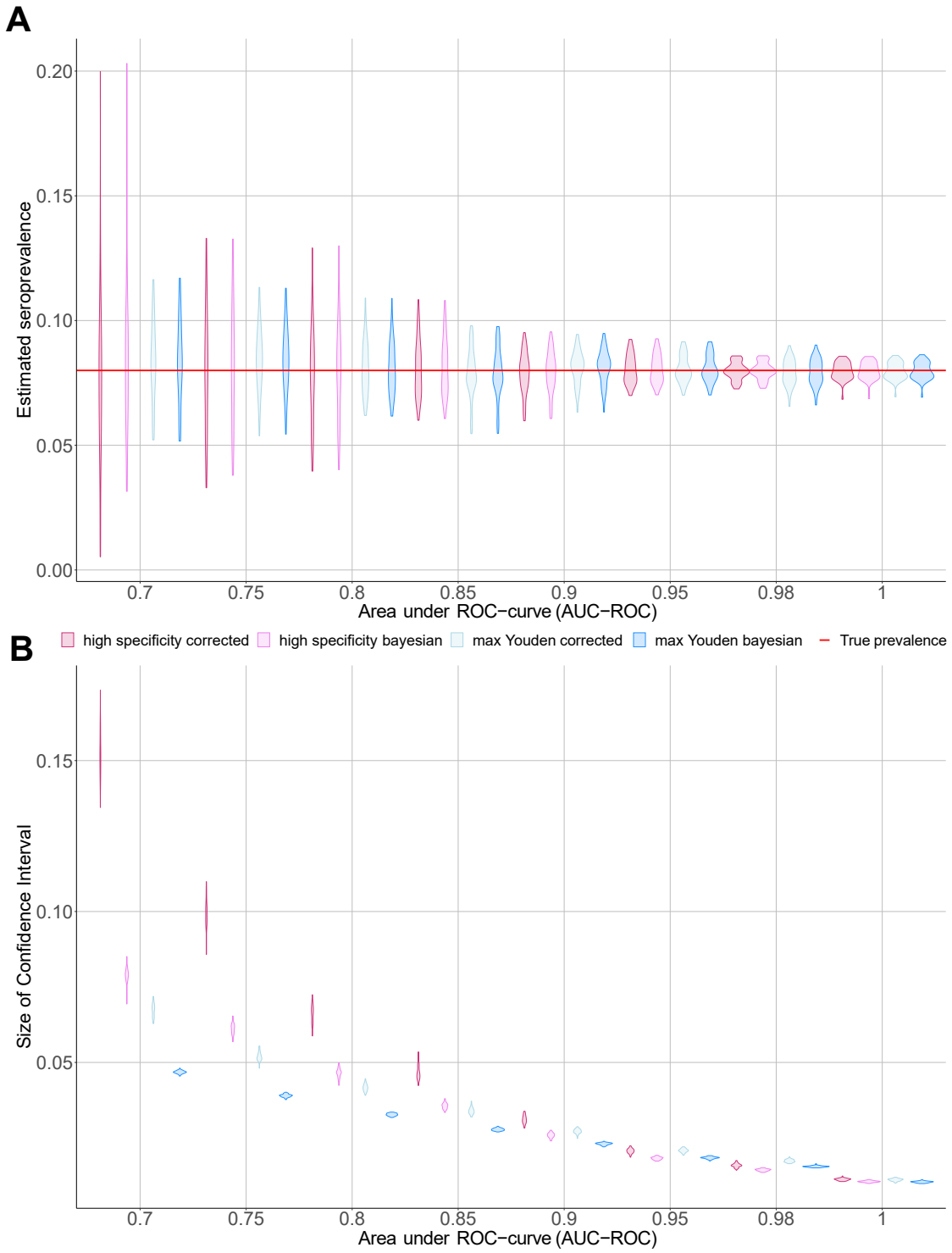

Figure S4: Comparison of cutoff-based methods with the Bayesian and the Rogan-Gladen correction. (A) Estimates of the seroprevalence versus AUC-ROC values. We are plotting the median of the posterior distribution, for the Bayesian correction, and point-estimates for the methods with Rogan-Gladen correction. (B) Size of the confidence or high probability density intervals of the seroprevalence estimates versus AUC-ROC values.