Quantifying previous SARS-CoV-2 infection through mixture modelling of antibody levels

C. Bottomley, M. Otiende, S. Uyoga, K. Gallagher, E. W. Kagucia, A. O. Etyang, D. Mugo, J. Gitonga, H. Karanja, J. Nyagwange, I. M. O. Adetifa, A. Agweyu, D. J. Nokes, G. M. Warimwe & J. A. G. Scott

As countries decide on vaccination strategies and how to ease movement restrictions, estimating the proportion of the population previously infected with SARS-CoV-2 is important for predicting the future burden of COVID-19. This proportion is usually estimated from serosurvey data in two steps: first the proportion above a threshold antibody level is calculated, then the crude estimate is adjusted using external estimates of sensitivity and specificity. A drawback of this approach is that the PCR-confirmed cases used to estimate the sensitivity of the threshold may not be representative of cases in the wider population—e.g., they may be more recently infected and more severely symptomatic. Mixture modelling offers an alternative approach that does not require external data from PCR-confirmed cases. Here we illustrate the bias in the standard threshold-based approach by comparing both approaches using data from several Kenyan serosurveys. We show that the mixture model analysis produces estimates of previous infection that are often substantially higher than the standard threshold analysis.
Establishing the amount of previous infection with SARS-CoV-2 is key to predicting the future impact of the virus. The evidence to date suggests that reinfection is uncommon, at least in the short term, and associated with mild disease\(^1,2\). Therefore, knowing how many people have previously been infected can help to establish to what extent a population is protected by natural infection.

The proportion previously infected is usually estimated from serological surveys (i.e. data on antibody levels). The conventional analysis of these data involves estimating the proportion above an arbitrary threshold and adjusting for the sensitivity and specificity at that threshold\(^3,4\). However, sensitivity is usually estimated using samples from PCR-positive cases who are symptomatic and have been recently infected. Since these samples typically have higher antibody levels than samples from the general population of previously infected individuals, including those with asymptomatic infection, this may lead to the overestimation of sensitivity and underestimation of the proportion previously infected\(^5\). Bias of this kind, that is bias that arises because sensitivity (or specificity) is estimated in a non-representative sample, is often referred to as spectrum bias. Mixture models offer an alternative approach to the analysis of serological data that does not involve specifying a threshold and is therefore not vulnerable to spectrum bias\(^6,7\).

In this paper, we use data on antibody concentrations—optical density (OD) ratios measured by ELISA—from several Kenyan SARS-CoV-2 serosurveys to compare the standard threshold-based analysis with a mixture modelling approach.

In the mixture model, we assume the observed distribution of antibody levels is a mixture of two unobserved distributions—the distribution in individuals who have experienced previous infection and the distribution in those who have not. The model is therefore characterised by the two component distributions and the proportion in each component. For the uninfected component we specify that log antibody concentrations follow a normal distribution, and for the infected component we specify they follow a skew normal distribution. To fit the model, we fix the variance of the uninfected component at a value estimated from pre-COVID-19 samples, and estimate the remaining parameters using a Markov chain Monte Carlo algorithm. We show that this mixture modelling approach generally produces higher estimates of the proportion previously infected than the standard threshold analysis.

Results

We found that the positive (previously infected) and negative (previously uninfected) distributions estimated from the mixture model did not segregate as clearly as expected based on the distributions observed in pre-COVID-19 and PCR-positive samples (Figs. 1 and 2, Supplementary Table 1). In most surveys, this was because the positive distribution was shifted to the left relative to the distribution in PCR-positive samples, i.e. the mean was lower than in the PCR-positive samples (mean log\(_2\) OD ratios = 3.07). In contrast, the mean of the negative distribution was usually similar to the mean observed in the pre-COVID-19 samples (mean log\(_2\) OD ratios = −0.17). However, there was some variation by region, and in the surveys done in truck drivers the means were higher, while in the surveys done in pregnant women they were lower. In most surveys, the skew parameter of the positive distribution was close to zero, and the scale parameter, which determines the spread of the positive distribution, was similar to the standard deviation in PCR-confirmed cases (SD log\(_2\) OD ratios = 1.32).

Because the mixture model predicted lower antibody levels in previously infected individuals than was observed in the PCR-positive samples, the sensitivity of the threshold (i.e. the proportion of OD ratios > 2 in the positive component) was generally lower than assumed in the standard threshold-based analysis; consequently the mixture model analysis produced higher estimates of the proportion previously infected than the threshold analysis (Fig. 3, Supplementary Fig. 1 and Supplementary Table 2). Across all surveys the mean sensitivity was 65% (cf. 93% sensitivity measured in the validation sample and assumed in the threshold analysis), and the estimated proportion previously infected was on average 1.44-fold higher than in the threshold analysis.

In general, the 95% credible intervals associated with the estimated proportions were wider in the mixture model analysis, with the largest differences occurring when there was strong overlap between the component distributions. For example, the three surveys of truck drivers (in Busia, Magarini and Malaba) produced the widest confidence intervals; they were also the surveys with the greatest overlap between the distributions.

To assess the robustness of our results, we fitted an alternative model where log antibody levels follow a two-component mixture distribution in previously infected individuals and a normal distribution in previously uninfected individuals, as in the original model. The two positive components do not have a clear biological interpretation, nevertheless we might imagine that recent/symptomatic infections predominate in the high-antibody-level component and older/asymptomatic infections predominate in the low-antibody-level component. The predicted distributions and estimates of the proportion positive from this alternative model were similar to those from the original skew normal model suggesting that the results are not sensitive to the distribution assumed for the positive component (Supplementary Fig. 2, Supplementary Fig. 3).

We further tested the mixture model in three simulated data scenarios (Supplementary Table 3). In scenario 1, the test data were generated by combining the PCR-positive and pre-COVID-19 samples. In scenarios 2 and 3 either the positive (scenario 2) or negative component (scenario 3) was simulated from a mixture of two normal distributions and the other component was simulated from a single normal distribution.

In keeping with the results of the previous sensitivity analysis, the mixture model performed well in scenarios 1 and 2: in both cases the estimated proportion positive was close to the expected value (scenario 1: 15% vs 14%; scenario 2: 19% vs 20%). By contrast, in scenario 3—where a mixture was used for the negative component—the mixture model estimate was severely biased (estimated = 36%, expected = 20%). This bias arises because the variance for the negative component is fixed to be equal to the variance estimated from pre-COVID-19 samples (this is one of the constraints used to fit the model). Consequently, variation that should be attributed to the negative component is instead attributed to the positive component thereby overestimating the proportion positive.

The findings from scenario 3 suggest that combining data across different populations may lead to the overestimation of previous infection. To test this hypothesis, we conducted a Kenya-wide analysis by combining the blood donor data and fitting the mixture model to these combined data. Our expectation was that the proportion previously infected would be overestimated in this analysis because of the variation in background antibody levels observed in the region-specific analyses (Supplementary Table 1). Specifically, we hypothesised that the estimate from the analysis of the combined data would be greater than average of the region-specific estimates, after weighting by the number of samples collected in each region. Consistent with this prediction, the estimate from the combined analysis was 44.3% whereas the average of region-specific analyses was only 18.9%. In fact, the
estimate from the combined analysis was greater than any of the region-specific estimates.

Discussion

The mixture model analysis suggests that antibody levels are higher in samples from recent PCR-confirmed cases than in samples from previously infected serosurvey participants. Because of this, threshold-based estimates—which rely on having an accurate sensitivity estimate—underestimate previous infection.

There are two potential explanations for the higher antibody levels in recent PCR-confirmed cases. First, PCR-confirmed cases are more likely to be symptomatic, and symptoms such as cough, fever, and severe disease (hospitalisation) are positively associated with higher antibody levels. Second, PCR testing is usually done soon after infection when antibody levels are high. In the PCR-positive samples used in our study, for example, the median time between symptom onset and blood sample collection was 21 days, which may have been ideal to capture the peak antibody response. By contrast, many survey participants will have been infected several months earlier and their antibody levels may have waned in the meantime. Antibody waning has been reported in a number of studies, and those that have specifically accounted for waning by assuming a constant rate of seroreversion—rather than by accounting for spectrum bias more generally as we have done—have predicted a significant impact on seroprevalence estimates. In general, waning is greatest for anti-nucleocapsid antibodies, but it can also be significant for anti-spike protein antibodies. For example, in one study involving milder cases of infection the half-life of anti-spike antibodies was estimated to be just 73 days.

The mixture model results suggest variability in the background levels of anti-spike IgG between different populations. In addition to variation by region, we observed higher IgG levels in truck drivers and lower levels in pregnant women. This variation will bias the standard threshold analysis. For example, in a population with low background IgG levels, as observed in pregnant women, the specificity estimate will be too low, and the sensitivity estimate too high. The variation will also bias mixture model analyses if data are combined across different populations, as exemplified by the Kenya-wide analysis of blood donor data.

The reasons for the variation in baseline IgG levels are unclear. It could simply reflect temporal variation in the laboratory procedures, though the negative control should guard against this bias. Alternatively, it could be related to differences between populations in exposure to infection and possibly also infective dose. Several studies of pre-COVID-19 antibody levels have reported variation between populations, with antibody levels generally being higher in African populations than non-African populations. Furthermore, anti-SARS-CoV-2 antibodies are known to cross-react with antibodies against other coronaviruses, and possibly also antibodies against dengue and malaria, though the latter finding was not confirmed in a more recent study. For pregnant women, it is possible that low antibody levels are a feature of immune environment in pregnancy.

The major limitation of the mixture modelling approach is that it is sensitive to the variance assumed for the uninfected population. Ideally the variance estimate should come from the population being surveyed, but in practice it will often be necessary to use an estimate from a different population. For example, we used pre-COVID-19 samples from blood donors in...
Fig. 2 Mixture distributions fitted to anti-spike IgG antibody data collected in serological surveys of Kenyan blood donors, antenatal care (ANC) attendees, healthcare workers (HCW), and truck drivers. The red distribution represents predicted responses in individuals previously infected with SARS-CoV-2 and the blue distribution represents predicted responses in previously uninfected individuals.

Fig. 3 Previous SARS-CoV-2 infection in Kenyan blood donors, antenatal care attendees, health care workers and truck drivers estimated via threshold analysis and mixture modelling. Estimates of the proportion previously infected are shown with 95% credible intervals.
In Nairobi and their contacts (14 pre-symptomatic, 55 symptomatic, 92 asymptomatic). The median time between the PCR test and blood sample collection was 21 days (IQR: 15, 34). The validation was based on a threshold OD ratio of 2, and the proportion previously infected, is a function of the proportion previously infected, \( p \), and the sensitivity and specificity of the threshold.

\[
y = \text{Binomial}(p_{\text{th}}, N)
\]
\[
x = \text{Binomial}(\text{se}, N_x)
\]
\[
z = \text{Binomial}(\text{sp}, N_y)
\]
\[
p_{\text{th}} = \text{se} \times p + (1 - \text{sp}) \times (1 - p).
\]

**Priors:**
- \( p \sim \text{Uniform}(0, 1) \)
- \( \text{se} \sim \text{Uniform}(0, 1) \)
- \( \text{sp} \sim \text{Uniform}(0, 1) \)

**Mixture model.** We fitted a two-component mixture model where individuals are classified according to whether or not they have experienced SARS-CoV-2 infection. We assumed that log2 OD ratios follow a skew normal among previously infected individuals (parameters: location = \( \xi \), scale = \( \omega \) and skew = \( \alpha \)) and a normal distribution among previously uninfected individuals (parameters: mean = \( \theta \), standard deviation = \( \nu \)).

\[
\mathcal{N}(\theta, \nu) + \text{SkewNormal}(\xi, \omega, \alpha),
\]

where \( p \) = proportion previously infected.

To make it easier to interpret the model parameters, we reparameterised the model in terms of the difference, \( \delta \), between the means of the two distributions:

\[
\xi = \theta + \delta - \omega \sqrt{2/\pi} \frac{\alpha}{\sqrt{1 + \alpha^2}}.
\]

Since mixture models can be difficult to fit when there is overlap of the component distributions, we used several constraints to facilitate parameter estimation. First, we fixed the standard deviation in the uninfected, \( \nu \), to be equal to the standard deviation in the pre-COVID-19 samples. Second, we used an informative prior for \( \delta \) to constrain the difference in means—the prior puts 5% probability on the difference exceeding the difference between the mean in symptomatic PCR-positive cases (mean log2 OD ratios = 3.43) and the mean in pre-COVID-19 samples (mean log2 OD ratios = –0.17). The prior for \( \delta \) was also designed to avoid label switching by ensuring \( \delta > 0 \). Finally, we used an informative prior for \( \alpha \) to rule out strong skew in either direction.

**Priors:**
- \( \theta \sim \text{Normal}(0, 10) \)
- \( \delta \sim \text{Normal}^+(0, 1.83) \)
- \( \ln \omega \sim \text{Normal}(0, 10) \)
- \( a \sim \text{Normal}(0, 1) \)
- \( p \sim \text{Uniform}(0, 1) \).

In addition to estimating the model parameters, we estimated the sensitivity and specificity at various threshold values. The sensitivity corresponds to the proportion above the threshold in the skew normal distribution and the specificity corresponds to the proportion below the threshold in the normal distribution. Both quantities were estimated using the sample of parameter values drawn from the posterior distribution.

**An alternative specification of the mixture model.** As a sensitivity analysis, we fitted an alternative model where we assumed that the distribution among previously infected individuals follows a mixture distribution with mixing parameter \( q \).

\[
(1 - p) \times \text{Normal}(\theta_1, \nu_1) + pq \times \text{Normal}(\theta_2, \nu_2) + (1 - q) \times \text{Normal}(\theta_3, \nu_3).
\]
As with the skew-normal model, the model was reparameterised in terms of the difference in mean between the positive and negative component, that is we defined \( \theta_1 = \alpha q_1 + (1 - q) \theta_2 - \delta \), where \( \delta \) represents the difference between the means. The priors for \( \nu_i (i = 2, 3) \) were chosen to ensure that these standard deviations are of similar magnitude to the standard deviation observed in PCR-positive individuals (SD log. OD ratios = 1.32 = exp(0.28)) and we used the constraint \( \theta_1 > \theta_2 \) to avoid the problem of label switching and ensure the identifiability of these parameters.

Priors:
\[
\begin{align*}
\delta & \sim \text{Normal}(0, 1.83) \\
\theta_1 & \sim \text{Normal}(0, 10) \quad i = 2, 3 \\
\ln \nu_i & \sim \text{Normal}(0.28, 0.2) \quad i = 2, 3 \\
p & \sim \text{Uniform}(0, 1) \\
q & \sim \text{Uniform}(0, 1).
\end{align*}
\]

Reporting summary: Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability
Source data are provided in the supplementary materials (xlsx file). Source data are provided with this paper.

Code availability
Stan code for fitting the Bayesian models is provided in Supplementary Notes 1–3 and R/Stan code for all analyses, including the generation of tables and figures, is available at: https://github.com/christian-bottomley/mixture_model_sarscov2.

Received: 23 April 2021; Accepted: 17 September 2021; Published online: 26 October 2021

References
1. Abu-Raddad, L. J. et al. Assessment of the risk of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) reinfection in an intense reexposure setting. Clin. Infect. Dis. ciaa1846, https://doi.org/10.1093/cia/ciaa1846 (2020).
2. Lumley, S. F. et al. Antibody status and incidence of SARS-CoV-2 infection in health care workers. N. Engl. J. Med. 384, 533–540 (2021).
3. Stringhini, S. et al. Seroprevalence of anti-SARS-CoV-2 IgG antibodies in Geneva, Switzerland (SEROCoV-POP): a population-based study. Lancet 396, 313–319 (2020).
4. Uyoga, S. et al. Seroprevalence of anti-SARS-CoV-2 IgG antibodies in Kenyan blood donors. Science 371, 79–82 (2021).
5. Takahashi, S., Greenhouse, B. & Rodriguez-Barraquer, I. Are seroprevalence estimates for severe acute respiratory syndrome coronavirus 2 biased? J. Infect. Dis. 222, 1772–1775 (2020).
6. Bouman, J. A., Riosa, J., Bonhof, S. & Regoes, R. R. Estimating the cumulative incidence of SARS-CoV-2 with imperfect serological tests: exploiting cutoff-free approaches. PLoS Comput. Biol. 17, e1008728 (2021).
7. Gay, N. J. Analysis of serological surveys using mixture models: application to a survey of parvovirus B19. Stat. Med. 15, 1567–1573 (1996).
8. Lumley, S. F. et al. The duration, dynamics and determinants of SARS-CoV-2 antibody responses in individual healthcare workers. Clin. Infect. Dis. ciaa004, https://doi.org/10.1093/cia/ciaa004 (2020).
9. Gaskell, K. M. et al. Extremely high SARS-CoV-2 seroprevalence in a strictly-Orthodox Jewish community in the UK. medRxiv https://doi.org/10.1101/2021.02.01.21250893 (2021).
10. Buss, L. F. et al. Three-quarters attack rate of SARS-CoV-2 in the Brazilian Amazon during a largely unmitigated epidemic. Science 371, 288–292 (2021).
11. Chia, W. N. et al. Dynamics of SARS-CoV-2 neutralising antibody responses and duration of immunity: a longitudinal study. Lancet Microbe 2, e240–e249 (2021).
12. Peluso, M. J. et al. SARS-CoV-2 antibody magnitude and detectability are driven by disease severity, timing, and assay. medRxiv https://doi.org/10.1101/2021.03.03.212251639 (2021).
13. Muecksch, F. et al. Longitudinal serological analysis and neutralizing antibody levels in coronavirus disease 2019 convalescent patients. J. Infect. Dis. 223, 389–398 (2021).
14. Figueiredo-Campos, P. et al. Seroprevalence of anti-SARS-CoV-2 antibodies in COVID-19 patients and healthy volunteers up to 6 months post disease onset. Eur. J. Immunol. 50, 2025–2040 (2020).
15. Chen, S., Flegg, J. A., White, L. J. & Aguas, R. Levels of SARS-CoV-2 population exposure are considerably higher than suggested by seroprevalence surveys. medRxiv https://doi.org/10.1101/2021.01.08.21249432 (2021).
16. Brand, S. P. C. et al. COVID-19 transmission dynamics underlying epidemic waves in Kenya. medRxiv https://doi.org/10.1101/2021.06.17.21259100 (2021).
17. Ibarroondo, F. J. et al. Rapid decay of anti–SARS-CoV-2 antibodies in persons with mild COVID-19. N. Engl. J. Med. 383, 1085–1087 (2020).
18. Tso, F. Y. et al. High prevalence of pre-existing serological cross-reactivity against severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2) in sub-Saharan Africa. Int. J. Infect. Dis. 102, 577–583 (2021).
19. Emmerich, P. et al. Limited specificity of commercially available SARS-CoV-2 IgG ELISAs in serum samples of African origin. Trop. Med. Int. Health. 26, 621–631 (2021).
20. Ng, K. W. et al. Preexisting and de novo humoral immunity to SARS-CoV-2 in humans. Science 370, 1339–1343 (2020).
21. Lustig, Y. et al. Potential antigenic cross-reactivity between severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and dengue viruses. Clin. Infect. Dis. ciaa1207, https://doi.org/10.1093/cid/ciaa1207 (2020).
22. Yadoulenton, A. et al. Limited specificity of serologic tests for SARS-CoV-2 antibody detection. Benin. Emerg. Infect. Dis. 27, 233–237 (2021).
23. Nkuba Ndoye, A. et al. Challenges in interpreting SARS-CoV-2 serological results in African countries. Lancet Glob. Health. 9, e588–e589 (2021).
24. Sherer, M. L. et al. Dysregulated immunity in SARS-CoV-2 infected pregnant women. medRxiv https://doi.org/10.1101/2020.11.13.20233173 (2020).
25. Adetifa, I. M. O. et al. Temporal trends of SARS-CoV-2 seroprevalence during the first wave of the COVID-19 epidemic in Kenya. Nat. Commun. 12, 3966 (2021).
26. Eityang, A. O. et al. Seroprevalence of antibodies to SARS-CoV-2 among health care workers in Kenya. Clin. Infect. Dis. ciab346, https://doi.org/10.1093/cid/ciab346 (2021).
27. Kagucia, E. W. et al. Seroprevalence of anti-SARS-CoV-2 IgG antibodies among truck drivers and assistants in Kenya. medRxiv https://doi.org/10.1101/2021.02.02.21215294 (2021).
28. Lucinde, R. et al. Sero-surveillance for IgG to SARS-CoV-2 at antenatal care clinics in two Kenyan referral hospitals. medRxiv https://doi.org/10.1101/2021.02.05.21250735 (2021).
29. Amanat, F. et al. A serological assay to detect SARS-CoV-2 seroversion in humans. Nat. Med. 26, 1033–1036 (2020).
30. Mattiuzzo, G., Bentley, E. M., Hassall, M. & Routley, S. Establishment of the WHO international standard and reference panel for anti-SARS-CoV-2 antibody. https://www.who.int/publications/m/item/WHO-BS-2020.2403 (2020).
31. Carpenter, B. et al. Stan: a probabilistic programming language. J. Stat. Softw. 76, 1–32 (2017).
32. R Core Team. R: a language and environment for statistical computing. (2021).

Acknowledgements
We thank the serosurvey participants for their contribution to this research. This project was funded by the Wellcome Trust (grant numbers 220911/Z/20/Z, 203077/Z/16/Z), the Bill and Melinda Gates Foundation (INV-017547) and by the Foreign Commonwealth and Development Office (FCDO) through the East Africa Research Fund (EARE/ITT/039). J. Anthony G. Scott is funded by a Wellcome Trust Senior Research Fellowship (214320). Christian Bottomley is funded through an award that is jointly funded by the UK Medical Research Council (MRC) and the UK Foreign, Commonwealth and Development Office (FCDO) under the MRC/FCDO Concordat agreement and is also part of the EDCTP2 programme supported by the European Union (MR/B010161/1). For the purpose of open access, the authors have applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission. This work is published with the permission of the Director of the Kenya Medical Research Institute.

Author contributions
Conceptualization: A.S. and C.B. Methodology: C.B., M.O. and A.S. Investigation: S.U., K.G., E.K., A.E., D.M., J.G., H.K., J.N., I.A., A.A. and G.W. Writing—original draft: C.B. Writing—review & editing: C.B., M.O., K.G., E.K., A.E., I.A., A.A., D.N. and A.S.

Competing interests
The authors declare no competing interests.
