Adaptation of the Rao-Wu rescaling bootstrap for seroprevalence estimation in complex survey studies

Nishma Vias

Senior Honors Thesis
Biostatistics
University of North Carolina at Chapel Hill

April 15, 2021

Approved:

_____
Dr. Bonnie Shook-Sa, Thesis Advisor

_____
Dr. Jane Monaco, Committee Member

_____
Dr. Ross Boyce, Committee Member

Adaptation of the Rao-Wu rescaling bootstrap for seroprevalence estimation in complex survey studies

Abstract

The coronavirus 2019 (COVID-19) pandemic has highlighted a need for accurate seroprevalence estimators in order to monitor virus transmission and mortality rates. Furthermore, seroprevalence data is used to inform public health policy. Adjustments for diagnostic test sensitivity and specificity in seroprevalence estimation are complicated by concerns regarding proper confidence interval (CI) coverage and width given that the proportion of a population considered seropositive may be relatively small. As such, these methods are not widely implemented at present, particularly for complex survey studies. This paper presents a two-stage, non-parametric bootstrap method, adapted from the Rao-Wu rescaling bootstrap, for adjusted CI construction. Simulation of a stratified, multi-stage cluster sample was conducted to assess the performance of the proposed method as measured by empirical bias, 95% CI coverage, and CI width. Across 500 simulations, the mean empirical bias was 0.0008 (range, -0.0411–0.0348), the CI empirical coverage was 95.6%, and CI width was 0.0487. Further application of the proposed method to preliminary study data from a Chatham County, North Carolina seroprevalence study demonstrated directionally appropriate adjustments for the diagnostic test sensitivity and specificity. The proposed method, paired with the Rogan-Gladen estimator for true prevalence, will allow for sensitivity-specificity adjusted seroprevalence estimation in complex survey designs while minimizing the CI width and coverage issues that are common for small proportions.

Background

COVID-19, the viral disease caused by the novel SARS-CoV-2, was declared a pandemic by the World Health Organization (WHO) on March 11, 2020 following alarming levels of global spread and severity.[1] Following the regulatory approval of seroprevalence assays for SARS-CoV-2 antibodies, seroprevalence studies have emerged as a key means for surveillance.[2] Accurate seroprevalence estimates are necessary for modeling virus transmission and determining mortality rates, which in turn guide public health policies surrounding COVID-19.[2] The demand for rapid implementation and data turnaround from these studies, however, has revealed several key issues related to data reporting and generalizability of results.

The quality of seroprevalence studies conducted to date has varied greatly. While many studies report clear, detailed explanations of the methods used in analysis, some published reports of studies provide incomplete information about the study design.[3] Missing information makes it difficult to assess the quality of the estimates from such studies. A number of seroprevalence studies to date have been based on convenience samples recruited from clinics, social media platforms, and shopping centers.[4–6] Convenience samples are appealing because they are relatively inexpensive and can be acquired quickly. Selection bias, however, is often inherent in convenience samples due to a wide range of factors that may influence participation, which may limit generalizability of results to the total population.[2] Probability-based survey sampling techniques are based on random selection within a population while allowing quantification of a participant's chance of selection.[2] As a result, well-designed probability samples allow for design-based representative estimates of the target population.

Other issues arise from the nature of the diagnostic tools used in seroprevalence studies. The primary aim of a seroprevalence study is to estimate the proportion of a population that is carrying antibodies for a specific infection. An individual's antibody status is determined using a diagnostic test, typically a chemiluminescent immunoassay (CLIA) or enzyme-linked immunosorbent assay (ELISA). While antibodies tend to be highly specific in nature, no diagnostic test is perfect, which introduces an associated measurement error.[7] Disregarding a test's sensitivity, or the probability of a positive result when applied to a known positive, and specificity, or the probability of a negative result when applied to a known negative, can lead to biased estimates.[8] Seroprevalence estimates should be adjusted to account for sensitivity and specificity of the diagnostic test.

Rogan and Gladen[9] have proposed a prevalence estimator for known sensitivity and specificity. Sensitivity and specificity, however, are often not known quantities but also experimentally-determined estimates.[7] Biggs et al.[10] notes serological assay error as an important study limitation. While some Bayesian and bootstrap techniques have been developed to adjust for the uncertainty in these measures, they have not been widely implemented. Furthermore, there is a lack of sensitivity-specificity adjustment methods developed specifically for complex survey designs.

Finally, estimating seroprevalence is challenging because at the beginning of the pandemic, overall prevalence of SARS-CoV-2 antibodies tended to be low in most populations, leading to estimates close to 0.[8] In general, confidence limits are more informative than single point estimates because they help quantify the uncertainty associated with the estimate. Several concerns arise when constructing confidence limits for small proportions[11]:

Coverage. (1 - α)% coverage is ideal for a (1 - α)% confidence interval (CI). It is consistent with current practice to say that for a 95% CI, approximately 95% of similarly computed CIs are expected to contain the parameter, or true population value. For example, a 95% CI with true coverage of 99% is considered too conservative, while one that covers 90% is considered anti-conservative.

CI Width. Provided that the ideal coverage is met, a CI should be narrow in order to be as informative as possible.

Aberrations. Anomalies, including CIs of width zero and confidence limits less than 0 or greater than 1 for a proportion may be observed. In particular, the standard Wald CI, which is based on a normal approximation, can undercover the true proportion.[8]

A common approach for small proportions is to use the logit transformation of the proportion to compute a CI since it ensures that the CI will fall in the [0,1] range.[12] There are also alternatives to the standard Wald-typed methods, which all use the continuous normal distribution to approximate the binomial distribution. The Clopper-Pearson CI, also known as the Exact CI, is based on an exact binomial distribution with a lower bound of 0 and upper bound of 1. It is considered to produce nominal coverage and tighter confidence intervals,[11] performing exceedingly well for small proportions when intracluster correlation is high.[13] However, it can be unnecessarily conservative.[11] The Wilson CI cannot produce a negative lower limit and is especially recommended for small proportions, although it may not be conservative enough.[11,13]

An alternative method for constructing CIs is the use of bootstrap techniques. The Rao-Wu rescaling bootstrap does not rely on any parametric assumptions, including large sample normality, to construct confidence intervals.[14] Instead, repeated replicate sampling is used to construct a distribution of estimates that captures uncertainty. The CIs are derived from this distribution. As a result, the Rao-Wu bootstrap can be specifically applied where small sample size and/or fundamentally irregular distributions lead to asymmetric sampling distributions.[14] Havers et al.[4] utilized a bootstrap method to account for sensitivity-specificity adjustments and construct CIs for seroprevalence estimates based on a convenience sample. The Rao-Wu bootstrap is promising for use with complex survey designs due to its ability to accommodate complex sample design features, including sampling weights, stratification, and clustering; and its ability to handle small sample sizes.[15]

The aim of this paper is to develop, evaluate, and apply a bootstrap method for seroprevalence CI construction that (1) can be used for complex survey designs, (2) adjusts for sensitivity-specificity of diagnostic tests, and (3) achieves the nominal level of confidence interval coverage, even for small proportions.

Methods

Assume that the true population seroprevalence (p) will be estimated based on the results of a diagnostic test conducted on a sample of participants from a stratified, multi-stage cluster sample. In this design, primary sampling units (PSUs), or clusters of observational units, are stratified. A random sample of PSUs is then selected, and units are subsequently randomly selected within sampled PSUs. Stratified designs allow the sampling frame to be divided into mutually exclusive and exhaustive categories, such as high- or low-risk groups. Sampling may then be used to over-select from certain strata as necessary. For example, oversampling could be used to give high-risk individuals a greater probability of selection compared to low-risk individuals. Sample weights are then applied to adjust the seroprevalence estimate appropriately for the sampling strategy used. Further, clustered sampling is often used to reduce costs and improved logistic feasibility of in-person data collection.

We derive point estimates and confidence intervals in three stages to account for increasing complexity in the sensitivity and specificity adjustments.

Stage 1: No measurement error. First, we assume that sensitivity and specificity of the diagnostic test were both 100%, i.e., that there was no error in the diagnostic test. We calculate seroprevalence population estimates ($\hat{p}$) using the ratio estimator common in survey sampling (Equation 1).[14 (p. 160)] This method utilizes the binary classification from the assay y_i ($y_i = 1$ indicates that individual i tested positive for antibodies, $y_i = 0$ otherwise) and sample weights (w_i) from each individual i in the observed sample of n individuals.

$$\hat{p} = \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i} \qquad \text{(Equation 1)}$$

We consider two possible confidence intervals in Stage 1. The Wald CI is presented in Equation 2.

$$\hat{p} \pm Z_{1-\frac{\alpha}{2}} \widehat{SE}(\hat{p}) \qquad \text{(Equation 2)}$$

Where α is the desired significance level and $\widehat{SE}(\hat{p})$ is the estimated standard error of $\hat{p}$.

Note that $\widehat{SE}(\hat{p})$ is estimated using design-consistent estimation methods in software such as SAS 'survey' procedures, the R 'survey' package, or SUDAAN. To produce a CI using the

logit of $\hat{p}$, $\hat{L}$, we first apply the logit transformation as shown in Equation 3.[12] We then calculate a Wald CI for $\hat{L}$. The upper and lower limits of the CI, $\hat{L}_{upper}$ and $\hat{L}_{lower}$, respectively, are then transformed back to the domain of p using an inverse-logit transformation (Equations 4 and 5).

$$\hat{L} = \log\left(\hat{p}/(1-\hat{p})\right) \quad \text{(Equation 3)}$$

$$LL: 1/(1+\exp(-\hat{L}_{lower})) \quad \text{(Equation 4)}$$

$$UL: 1/(1+\exp(-\hat{L}_{upper})) \quad \text{(Equation 5)}$$

Stage 2: Adjustment for known sensitivity and specificity. We next assume that the sensitivity and/or specificity of the diagnostic test is less than 100%, but that they are fixed, known quantities. For this scenario, we use the estimator proposed by Rogan and Gladen[9] to estimate the population seroprevalence (Equation 6) and its estimated standard error (Equation 7). Confidence intervals in this stage are computed using the formulas from Stage 1, substituting $\widehat{SE}(\hat{p})$ with $\widehat{SE}(\hat{p}')$.

$$\hat{p}' = \frac{\hat{p}+spec-1}{sens+spec-1} \quad \text{(Equation 6)}$$

$$\widehat{SE}(\hat{p}') = \frac{\widehat{SE}(\hat{p})}{sens+spec-1} \quad \text{(Equation 7)}$$

Where spec is the known specificity and sens is the known sensitivity of the diagnostic test.

Stage 3: Adjustment for estimated sensitivity and specificity. In the final stage, values for sensitivity and specificity of the diagnostic test are considered to be laboratory estimates with associated variability. As in Stage 2, we use the Rogan-Gladen estimator ($\hat{p}'$) to calculate seroprevalence point estimates (Equation 6). However, treating the estimated sensitivity and specificity as known when estimating variance, as in Equation 7, is anti-conservative. To create confidence intervals around our estimates that appropriately account for the estimated sensitivity and specificity, we use a two-stage non-parametric bootstrap, based off the Rao-Wu rescaling bootstrap described by Heeringa et al.[14] (p. 107-108)

We draw *B* replicate samples from the primary sample by selecting $m_h = n_h - 1$ PSUs from each stratum h, with replacement, where n_h is the number of PSUs that were sampled in stratum h. Within each replicate sample b, bootstrap weights ($w_{hi}^{(b)}$) are calculated for each individual i in stratum h, based on the number of times their PSU was selected within b (Equation 8).

$$w_{hi}^{(b)} = w_i \frac{n_h}{(n_h-1)} * r^{(b)} \quad \text{(Equation 8)}$$

Where $r^{(b)}$ is the number of times the PSUassociated with unit i is selected in replicate b.

Subsequently, we calculate weighted population estimates ($\hat{p}^b$) for the proportion of individuals with positive serology tests, for each replicate sample b.

Next, we account for variation in the sensitivity and specificity estimates. Assume that sensitivity and specificity were estimated based on lab tests of samples of sizes t_1 and t_2, respectively. We take random draws from two independent binomial distributions to estimate sensitivity and specificity for each bootstrap sample. That is, we let $X_1 \sim Binomial(t_1, sens)$ and $X_2 \sim Binomial(t_2, spec)$ and estimate sensitivity ($sens^b = x_1/t_1$) and specificity ($spec^b = x_2/t_2$). We then estimate seroprevalence for each bootstrap sample using the Rogan-Gladen estimator ($\hat{p}'^b$), where sensitivity and specificity are the bootstrap-specific values ($sens^b$ and $spec^b$). This process accounts for variation both in selection of the sample and in estimation of the sensitivity and specificity of the diagnostic test.

This process results in B estimates of $\hat{p}'^b$, one for each iteration of the bootstrap. We then use the distribution of all B estimates to calculate confidence interval endpoints at the $\frac{\alpha}{2}$ and $1 - \frac{\alpha}{2}$ percentiles.

Simulation Study

A simulation study was conducted to assess the performance, measured by empirical bias, 95% CI coverage, and CI width, of the estimators described in Stage 3 of the Methods section. The simulation was modeled after the Chatham County COVID-19 Cohort (C4) study, which is based on a stratified, 2-stage cluster design aiming to recruit Chatham County residents from 300 households across 141 sampled census blocks, which are the PSUs.[16] Additional details about the C4 study are provided in the Application section below.

The following simulation was conducted in SAS Studio 3.8 with 500 iterations.

Generate population data. First, the true infection status and diagnostic test status were generated for all individuals in the target population, as follows. Census data were used to estimate the number of households in Chatham County within each PSU. As in the C4 study, PSUs were stratified into three income strata. The base probability of infection for households in each income stratum 1 through 3 was assumed to be 0.15, 0.10, or 0.05, respectively. A cluster effect was generated by a random uniform variable with range ±0.5%. The true sensitivity value for the iteration was defined as $S_1/145$, where $S_1 \sim Binomial(145, 0.897)$ given that 0.897 was the lab-reported sensitivity for the serology assay used in the C4 study, based on 145 known positive samples. The true specificity value for the iteration was similarly derived ($spec^{(lab)} = 0.993, n = 274$). Within each household in the population, a true infection status was generated for the sampled individual based on the combined base risk and

cluster effect; the true value of p was thus determined for the population. Additionally, a diagnostic test status was generated for each individual based on the sensitivity and specificity for the iteration. Across the 500 simulated samples, the mean value of p was 0.0907 with a range of 0.0857 to 0.0957.

Determine sample estimate. A sample of PSUs was selected from each stratum (n_1 = 51, n_2 = 51, n_3 = 60) with probability of selection proportional to the number of households in each PSU. Up to 6 households were randomly selected without replacement from each PSU, as in the study design. The Rogen-Gladen corrected sample estimate $\hat{p}'$ was calculated from the ratio estimate for the mean diagnostic test status in the sample, $\hat{p}$.

Construct confidence interval. 1000 bootstrap samples were selected with replacement from the sample with $m_h = n_h - 1$ (i.e., with m_1 = 50, m_2 = 50, m_3 = 59) and $\hat{p}'^b$ was calculated for each bootstrap sample as described in the methods above. The CI was defined as the 2.5 and 97.5 quantiles of the distribution of 1000 $\hat{p}'^b$ estimates (corresponding to a 95% CI).

The simulation study resulted in 500 values of $\hat{p}'$ and corresponding confidence intervals. Empirical bias ($\hat{p}' - p$) remained low across the 500 iterations, with a mean of 0.0008 and range from -0.0411 to 0.0348. The distribution of empirical bias, shown in Figure 1, is centered close to zero. The CI width, determined by the mean width from all iterations, was 0.0487. Empirical coverage, measured as the proportion of CIs that contained the corresponding true value p, was 95.6%, just above the target value of 95%.

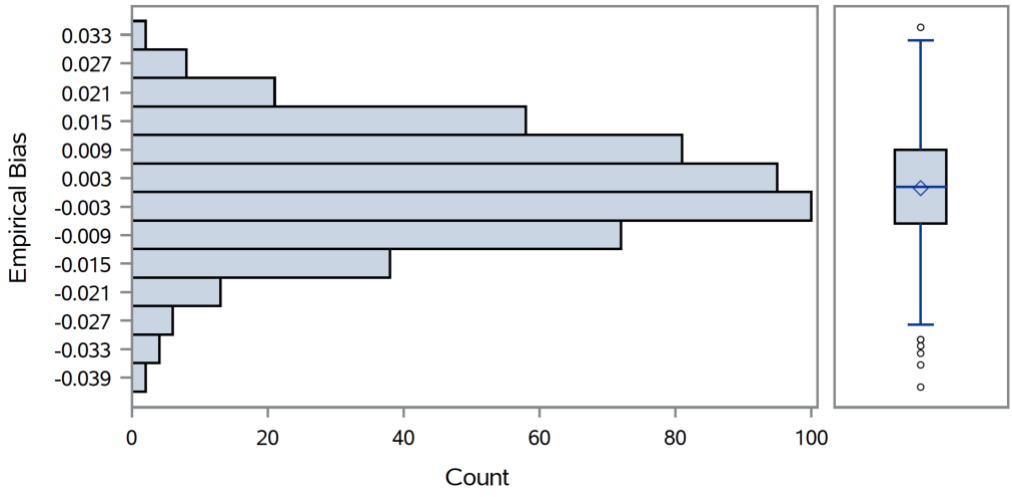

Figure 1. Histogram and boxplot for empirical bias (500 simulations; mean, 0.0008; range, -0.0411–0.0348).

In sum, the simulation study demonstrates low empirical bias and appropriate CI coverage for the methods proposed in Stage 3 when sensitivity and specificity are estimated within the context of a complex sample design.

Application

The Chatham County COVID-19 Cohort (C4) study is a longitudinal, prospective, population-based study aimed at estimating the prevalence of SARS-CoV-2 infection in central North Carolina. Study participants are Chatham County residents age 18 or older who were selected through a stratified, 2-stage cluster design targeting a total enrollment of 300 participants across 141 randomly selected census blocks. Census blocks were stratified across 3 income brackets and selected using probability proportional to size with replacement (PPS-WR) sampling, with the number of occupied households in each block serving as the measure of size. Individuals from approximately 6 households were sampled per census block. Households in census blocks with more concentrated Hispanic/Latino and/or Black/African American populations (based on 2010 Census data) were oversampled to facilitate seroprevalence estimation by race/ethnicity.[16]

For the seroprevalence study, serum samples are being collected from participants once a month via venous blood draws during in-person clinic visits or using a Tasso serum self-collection device at home. Serum samples are tested with ELISA, using the recombinant spike protein antigen to detect total SARS-CoV-2 Ig in plasma[17] (sensitivity = 0.897, specificity = 0.993). This assay detects seropositivity due to either prior SARS-CoV-2 infection or vaccination.

The proposed method described in Stage 3 of the Methods section was applied (B = 1000) to preliminary study data collected through March 9, 2021. The most recent serology result for each of the 118 participants recruited to date from 74 census blocks was analyzed using base (unadjusted) sampling weights. The resulting unadjusted and sensitivity-specificity adjusted seroprevalence estimates, along with their corresponding CIs, are shown in Table 1. Both estimates were noticeably greater than initially expected due to the beginning of vaccination amongst the cohort. The upwards adjustment of the estimate and associated confidence interval were consistent with each other. Without using the Rogan-Gladen estimator and the bootstrap CI, seroprevalence would be underestimated due to the assay's very high specificity but lower sensitivity, which has the tendency to produce more false negative results than false positive results. The bootstrap CI width was greater than the unadjusted CI width. The wide CI widths could be due to the preliminary data still representing only about half of the target sample size.

Table 1. Adjusted and unadjusted seroprevalance estimates and CI widths based on C4 study data collected through March 9, 2021 (n = 118).

	Seroprevalence Estimate (%)	CI Width (%)
Adjusted – Bootstrap CI	32.0 (20.8, 44.9)	24.1
Unadjusted	29.1 (19.0, 39.2)	20.2

Summary

Conditions of the COVID-19 pandemic have emphasized the need for accurate seroprevalence estimators to monitor virus transmission and mortality rates. Furthermore, these estimators should account for measurement error introduced by diagnostic tests and be compatible with probability-based survey sampling study designs, which are often used to improve feasibility of data collection and generalizability of results to the target population. The Rogan-Gladen estimator, paired with a CI generated through the proposed two-stage non-parametric bootstrap method, will allow for seroprevalence estimation in complex survey designs while adjusting for sensitivity-specificity of diagnostic tests as well as minimizing CI width and coverage issues that are common for small proportions. Simulations demonstrated low empirical bias (mean, 0.0008; range, -0.0411–0.0348) and nominal CI coverage (95.6%) when diagnostic test sensitivity and specificity are estimated. Furthermore, the methods accounted for the diagnostic test's lower sensitivity and higher specificity when applied to preliminary data from the C4 study.

There are limitations associated with the proposed methods. The bootstrap technique is computationally intensive, especially when a large number of bootstrap samples is desired. Additionally, the methods require knowledge of diagnostic test sensitivity-specificity estimates as well as the related validation study sample sizes. Despite these limitations, the proposed methods provide a promising approach towards confidence interval construction for seroprevalence estimates from complex sample designs.

References

1. Archived: WHO Timeline - COVID-19. World Health Organization. Accessed October 12, 2020. https://www.who.int/news/item/27-04-2020-who-timeline---covid-19

2. Shook-Sa BE, Boyce RM, Aiello AE. Estimation Without Representation: Early Severe Acute Respiratory Syndrome Coronavirus 2 Seroprevalence Studies and the Path Forward. *J Infect Dis*. 2020;222(7). doi:10.1093/infdis/jiaa429

3. Garcia-Basteiro AL, Moncunill G, Tortajada M, et al. Seroprevalence of antibodies against SARS-CoV-2 among health care workers in a large Spanish reference hospital. *Nat Commun*. 2020;11(1):3500. doi:10.1038/s41467-020-17318-x

4. Havers FP, Reed C, Lim T, et al. Seroprevalence of Antibodies to SARS-CoV-2 in 10 Sites in the United States, March 23-May 12, 2020. *JAMA Intern Med*. 2020;180(12):1576-1586. doi:10.1001/jamainternmed.2020.4130

5. Bendavid E, Mulaney B, Sood N, et al. COVID-19 antibody seroprevalence in Santa Clara County, California. *Int J Epidemiol*. 2021;(dyab010). doi:10.1093/ije/dyab010

6. Sood N, Simon P, Ebner P, et al. Seroprevalence of SARS-CoV-2–Specific Antibodies Among Adults in Los Angeles County, California, on April 10-11, 2020. *JAMA*. 2020;323(23):2425-2427. doi:10.1001/jama.2020.8279

7. EUA Authorized Serology Test Performance. U.S. Food & Drug Administration. Published January 8, 2021. Accessed September 24, 2020. https://www.fda.gov/medical-devices/coronavirus-disease-2019-covid-19-emergency-use-authorizations-medical-devices/eua-authorized-serology-test-performance

8. Shook-Sa BE. Statistical Considerations in the Design and Analysis of SARS-CoV-2 Prevalence Studies. Published online July 31, 2020. Accessed September 2, 2020. https://gillingscovid19.unc.edu/whitepaper/statistical-considerations-design-analysis-sars-cov-2-prevalence-studies

9. Rogan WJ, Gladen B. Estimating Prevalence from the Results of a Screening Test. *Am J Epidemiol*. 1978;107(1):71-76. doi:10.1093/oxfordjournals.aje.a112510

10. Biggs HM, Harris JB, Breakwell L, et al. Estimated Community Seroprevalence of SARS-CoV-2 Antibodies — Two Georgia Counties, April 28–May 3, 2020. *Morb Mortal Wkly Rep*. 2020;69(29):965-970. doi:10.15585/mmwr.mm6929e2

11. Tobi H, van den Berg PB, de Jong-van den Berg LT. Small proportions: what to report for confidence intervals? *Pharmacoepidemiol Drug Saf*. 2005;14(4):239-247. doi:https://doi.org/10.1002/pds.1081

12. Gray A, Haslett S, Kuzmicich G. Confidence Intervals for Proportions Estimated from Complex Sample Designs. *J Official Stat*. 2004;20(4):705-723.

13. Dean N, Pagano M. Evaluating Confidence Interval Methods for Binomial Proportions in Clustered Surveys. *J Surv Stat Methodol*. 2015;3(4):484-503. doi:10.1093/jssam/smv024

14. Heeringa SG, West BT, Berglund PA. *Applied Survey Data Analysis*. 2nd ed. Chapman and Hall/CRC Press; 2017.

15. Girard C. The Rao-Wu Rescaling Bootstrap: From theory to practice. In: *Proceedings of the Federal Committee on Statistical Methodology Research Conference*. Federal Committee on Statistical Methodology; 2009.

16. Law EA, Miller EM, Shook-Sa BE, et al. SARS-CoV-2 Infection in Central North Carolina: Protocol for a Population-Based Longitudinal Cohort Study. Manuscript in preparation, April 2021.

17. Premkumar L, Segovia-Chumbez B, Jadi R, et al. The receptor binding domain of the viral spike protein is an immunodominant and highly specific target of antibodies in SARS-CoV-2 patients. *Sci Immunol*. 2020;5(48). doi:10.1126/sciimmunol.abc8413