Biometrics yy, 1–21 DOI: 10.1111/j.1541-0420.2005.00454.x
xx 202x

# Estimating SARS-CoV-2 Seroprevalence

**Samuel Rosin[1*], Stephen R. Cole[2], and Michael G. Hudgens[1**]**

Departments of [1]Biostatistics and [2]Epidemiology

University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, U.S.A.

[*]*email:* srosin@live.unc.edu

[**]*email:* mhudgens@bios.unc.edu

Summary: Diagnostic testing is a longstanding area of biostatistics, and it has perhaps never been of greater importance than during the COVID-19 pandemic. Governments and public health authorities worldwide have relied on seroprevalence studies to provide estimates of the proportions of persons with antibodies to SARS-CoV-2. However, serological assays are prone to both false positives and false negatives, and nonrandom population sampling induces selection bias. While ad hoc estimators that attempt to correct for these biases have been used in serosurveys, usually making an assumption of conditional independence of sampling given covariate data, these estimators' statistical characteristics have not previously been analyzed. In this article we consider a frequentist point estimator that addresses both of these challenges, show that it is consistent and asymptotically normal under regularity conditions, and introduce a consistent variance estimator using estimating equation theory. Simulation studies confirm the performance of the estimators in finite sample settings, and the methods are applied to a serosurvey of asymptomatic residents of North Carolina.

Key words: Data fusion; Diagnostic tests; Estimating equations; SARS-CoV-2; Seroepidemiologic studies; Standardization

This paper has been submitted for consideration for publication in *Biometrics*

## 1. Introduction

Biostatisticians who did not heed the National Academy of Medicine's 2001 call to "do more to improve our ability to prevent, detect, and control microbial threats to health" (Smolinski, Hamburg, and Lederberg, 2003) surely now recognize the importance of infectious diseases as the world faces the COVID-19 pandemic. Many considerations, some of them requiring statistical estimation, have guided governmental and public health policy decisions, including decisions about when to lock down and reopen various sectors of society. Among these factors is the proportion of the population that has antibodies to SARS-CoV-2 (the virus that causes COVID-19 disease), a quantity that has been estimated in dozens of studies in different populations worldwide. Having antibodies means that a person was previously infected with the virus and may have some level of immunity to future infection. However, the surveys used to estimate the prevalence of antibodies in the population ("serosurveys", which estimate seroprevalence), and in fact all medical prevalence studies that use diagnostic tests, usually suffer from at least two sources of error.

First, blood assays used to test for antibodies almost always suffer from false positives and false negatives. This measurement error can bias results very significantly, especially when seroprevalence is low. An example given by Sempos and Tian (2021) is instructive. Assume that true seroprevalence is 1%, and antibody tests are performed using an assay which perfectly identifies true positives as positive, so with 100% sensitivity, and nearly perfectly identifies true negatives as negative, so with 99% specificity. Despite the strength of this assay, approximately 50% of the people identified by the assay as having antibodies will be false positives. Clearly, sensitivity and specificity should be estimated and accounted for in order to have an accurate overall seroprevalence estimate. A traditional method to combat such error is to run validation studies assessing the assay's performance on true positives and true negatives. For SARS-CoV-2, true serology status is measured by gold-standard

methods such as reverse transcription polymerase chain reaction testing (RT-PCR), and/or by testing remnant blood samples that were drawn before major spread of the pathogen in 2019. Sensitivity and specificity are estimated from the validation data, and in turn these estimates are used to adjust the final seroprevalence estimate (Rogan and Gladen, 1978; Messam et al., 2008; Sempos and Tian, 2021).

Second, serosurveys are usually not conducted by simple random sampling, but by convenience sampling, which leads to selection bias. Covariate data and a conditional independence assumption, which is equivalent to assuming the sample arose by stratified random sampling on the covariate data used, are often used afterwards in an attempt to generalize estimates to broader populations of interest (Shook-Sa, Boyce, and Aiello, 2020). Under this unverifiable assumption, a final prevalence is calculated using direct standardization (Miettinen, 1985; Van Belle et al., 2004; see Hernán and Robins, 2020 for use in causal inference). Finally, we note that these methodological challenges are not unique to COVID-19 or serosurveys, but are broadly applicable to medical prevalence studies that use diagnostic tests.

A frequentist point estimator of seroprevalence that combines these two adjustments, using discrete demographic variables for standardization, has been used in recent serosurveys (Havers et al., 2020; Barzin et al., 2020). To our knowledge the statistical properties of this estimator, a Rogan-Gladen with standardization estimator for estimating prevalence under measurement error and selection bias, have not previously been analyzed. In this article we show that, under regularity conditions, the estimator is consistent and asymptotically normal. These recent serosurveys constructed confidence intervals using a two-stage bootstrap, but here we use estimating equations to derive an empirical sandwich variance estimator and show that it is consistent for the true asymptotic variance of the point estimator. In an example of data fusion, an estimating equations framework is used to incorporate all three

datasets - two validation datasets and the prevalence study dataset - in both the theoretical proofs and the derivation of the variance estimator.

The article is organized as follows. Section 2 develops the methods generally for broad application to prevalence estimation problems, and is composed of three subsections. Section 2.1 reviews the problem of prevalence estimation under measurement error, and rederives the variance estimator of the Rogan-Gladen estimator using an estimating equations approach. This approach is developed in order to derive the nonparametric point estimators and consistent variance estimator in the setting of Section 2.2 when, in addition to measurement error, selection bias is also present. Section 2.3 introduces modeling strategies for when sample data do not exist for all covariate combinations. A simulation study in Section 3 demonstrates that the nonparametric point estimator is approximately unbiased and that 95% Wald-type confidence intervals based on the confidence variance estimator give approximately nominal coverage. The methods are applied to a serosurvey of asymptomatic residents of North Carolina in Section 4. Lastly, we conclude with a discussion.

## 2. Methods

### 2.1 *Prevalence estimation under measurement error*

*Problem Setup.* Let $Y$ denote a binary random variable with expectation $\pi = P(Y = 1)$. Our goal is to draw inference about $\pi$. Let $X$ be a mismeasured version of $Y$. Let $\rho = P(X = 1)$ denote the expected value of $X$, $\sigma_e = P(X = 1|Y = 1)$ denote sensitivity, and $\sigma_p = P(X = 0|Y = 0)$ denote specificity. Suppose we observe $n_1$ independent and identically distributed (iid) copies of $X$ from strata of the population where $Y = 1$, and denote these observations by $X_1, \ldots, X_{n_1}$. Suppose we also observe $n_2$ iid copies of $X$ from strata of the population where $Y = 0$; denote these observations by $X_{n_1+1}, \ldots, X_{n_1+n_2}$. Finally, we also observe $n_3$ iid copies of $X$ from the (unstratified) population, denoted by $X_{n_1+n_2+1}, \ldots, X_n$

where $n = n_1 + n_2 + n_3$. Let $\delta$ be an indicator of which study an observation of $X$ is from, with $\delta_i = 1$ if $i \leqslant n_1$, $\delta_i = 2$ if $n_1 < i \leqslant n_1 + n_2$, and $\delta_i = 3$ otherwise. Note that $\sum I(\delta_i = j) = n_j$, $j = 1, 2, 3$, and assume $n_j/n \to c_j \in (0, 1)$ as $n \to \infty$, where here and throughout summations are taken from $i = 1$ to $n$ unless otherwise specified.

*Estimators and Statistical Properties.* Let $\theta = (\sigma_e, \sigma_p, \rho, \pi)^T$ denote the 4-vector of true parameter values and consider the vector of estimators $\hat{\theta} = (\hat{\sigma}_e, \hat{\sigma}_p, \hat{\rho}, \hat{\pi})^T$ where $\hat{\sigma}_e = \sum I(\delta_i = 1)X_i/n_1$, $\hat{\sigma}_2 = \sum I(\delta_i = 2)(1 - X_i)/n_2$, $\hat{\rho} = \sum I(\delta_i = 3)X_i/n_3$, and $\hat{\pi} = (\hat{\rho} + \hat{\sigma}_p - 1)/(\hat{\sigma}_e + \hat{\sigma}_p - 1)$. (The term $x$-vector is used to denote a vector of dimension $x$.) $\hat{\sigma}_e$, $\hat{\sigma}_p$, and $\hat{\rho}$ are maximum likelihood estimates (MLEs) from their respective binomial distributions while $\hat{\pi}$ is the commonly-used Rogan-Gladen estimator (Rogan and Gladen, 1978) first discussed by Marchevsky in 1974 (Marchevsky 1979). By the invariance of the MLE $\hat{\pi}$ is also an MLE. These estimators are solutions to the 4-vector of estimating equations

$$\sum \psi(X_i; \delta_i, \theta) = \begin{pmatrix} \sum \psi_e(X_i; \delta_i, \theta) \\ \sum \psi_p(X_i; \delta_i, \theta) \\ \sum \psi_\rho(X_i; \delta_i, \theta) \\ \psi_\pi(X_i; \delta_i, \theta) \end{pmatrix} = \begin{pmatrix} \sum I(\delta_i = 1)(X_i - \sigma_e) \\ \sum I(\delta_i = 2)((1 - X_i) - \sigma_p) \\ \sum I(\delta_i = 3)(X_i - \rho) \\ (\rho + \sigma_p - 1) - \pi(\sigma_e + \sigma_p - 1) \end{pmatrix} = 0.$$

Using this estimating equations approach (also known as M-estimation) we show that $\hat{\theta}$ is consistent for $\theta$ and asymptotically normal, and that its variance can be consistently estimated by an empirical sandwich estimator. The notation and approaches used generally follow those described in Boos and Stefanski (2013). Taylor expansion of $n^{-1} \sum \psi(X_i; \delta_i, \hat{\theta})$ around the true parameter value $\theta$ yields

$$0 = n^{-1} \sum \psi(X_i; \delta_i, \hat{\theta}) = n^{-1} \sum \psi(X_i; \delta_i, \theta) + (\hat{\theta} - \theta)^T n^{-1} \sum \dot{\psi}(X_i; \delta_i, \theta) + R_n$$

where $\dot{\psi}(X_i; \delta_i, \theta) = \partial \psi(X_i; \delta_i, \theta)/\partial \theta^T$ is a $4 \times 4$ matrix of partial derivatives and $R_n$ is a remainder term. Rearranging and multiplying by $\sqrt{n}$ yields

$$\sqrt{n}(\hat{\theta} - \theta) = \left[ n^{-1} \sum -\dot{\psi}(X_i; \delta_i, \theta) \right]^{-1} \times \sqrt{n} \left( n^{-1} \sum \psi(X_i; \delta_i, \theta) \right) + \sqrt{n} R_n^* \qquad (1)$$

where $R_n^*$ is also a remainder term. Equation (1) is analyzed as $n \to \infty$ to understand the asymptotic properties of $\hat{\theta}$, with details in Web Appendix A. We conclude that

$$\sqrt{n}(\hat{\theta} - \theta) \to_{d.} \mathcal{N}\left(0, \mathbb{A}(\theta)^{-1}\mathbb{B}(\theta)\mathbb{A}(\theta)^{-T}\right) \text{ and } \sqrt{n}(\hat{\pi} - \pi) \to_{d.} \mathcal{N}\left(0, V_\pi\right)$$

where $\mathbb{A}(\theta)^{-1}\mathbb{B}(\theta)\mathbb{A}(\theta)^{-T}$ is a $4 \times 4$ variance-covariance matrix with lower-right element $V_\pi$,

$$V_\pi = \left(\frac{\pi^2\sigma_e(1-\sigma_e)}{c_1} + \frac{(1-\pi)^2\sigma_p(1-\sigma_p)}{c_2} + \frac{\rho(1-\rho)}{c_3}\right)(\sigma_e + \sigma_p - 1)^{-2}.$$

The proof is similar to that of Boos and Stefanski (2013) Equation 7.10, but because these data are not identically distributed, the Lindeberg-Feller Central Limit Theorem is used in place of the classical (Lindeberg-Lévy) Central Limit Theorem, and the asymptotic assumptions above about each $n_j/n, j = 1, 2, 3$ are necessary.

Since $\sqrt{n}(\hat{\pi}-\pi) \to_{d.} \mathcal{N}(0, V_\pi)$, it follows that $\hat{\pi}$ is approximately distributed $\mathcal{N}(\pi, V_\pi/n)$ as $n$ grows infinitely large (and likewise for the other elements of $\hat{\theta}$). Estimation of $V_\pi$ proceeds by substituting $\hat{c}_j = n_j/n$ for $c_j$, $j = 1, 2, 3$ and using the previously-considered estimators $\hat{\sigma}_e$, $\hat{\sigma}_p$, and $\hat{\pi}$. Denote this variance estimator by $\hat{V}_\pi$. The resulting variance estimator is

$$\hat{V}_\pi/n = \left(\frac{\hat{\pi}^2\hat{\sigma}_e(1-\hat{\sigma}_e)}{n_1} + \frac{(1-\hat{\pi})^2\hat{\sigma}_p(1-\hat{\sigma}_p)}{n_2} + \frac{\hat{\rho}(1-\hat{\rho})}{n_3}\right)(\hat{\sigma}_e + \hat{\sigma}_p - 1)^{-2}, \tag{2}$$

which was derived by Rogan and Gladen (1978) and is consistent for the true variance $V_\pi/n$. The proof of consistency is deferred to Web Appendix A. This variance estimator can be used to construct Wald-type confidence intervals that asymptotically attain nominal coverage probabilities.

### 2.2 *Standardization*

*Problem Setup.* Now assume that for the $n_3$ copies of $X$ from the infinite population, a $b$-vector of covariates $\tilde{Z}$ is observed. This infinite population is known as the target population. Assume each covariate is discrete and there are a total of $k \geqslant b$ discrete strata of $\tilde{Z}$, that is, the $i$th covariate has $l_i$ levels, $i = 1, \ldots, b$ and $k = \prod_{i=1}^{b} l_i$. The distributions of $Y$ and $X$ are assumed to be heterogeneous over the strata of $\tilde{Z}$. For notational convenience let $Z$ be

a redefined version of $\tilde{Z}$, where $Z$ is a discrete variable belonging one of $k$ possible strata, $Z \in \{z_1, \ldots, z_k\}$.

Further, we make a conditional independence assumption that the sample is not a simple random sample from the target population but a stratified random sample by $Z$. Assume the stratum proportions in the target population are known and denoted by $\gamma_j \equiv P(Z = z_j)$ where $\gamma_j > 0$ for $j = 1, \ldots, k$ and $\sum_{j=1}^{k} \gamma_j = 1$. The covariates $Z$ are distributed Multinomial with $k$ categories, sample size $n_3$, and an unknown sampling probability vector $(s_1, \ldots, s_k)^T$ where $s_j > 0$, $j = 1, \ldots, k$ and $\sum_{j=1}^{k} s_j = 1$. Note that in the simple random sample case the sampling probabilities are equal to the stratum proportions, $s_j = \gamma_j$, $j = 1, \ldots, k$. The $j$th unknown stratum-conditional mean of $X$ is denoted as $\rho_j = P(X = 1|Z = z_j)$.

Lastly, denote the sample size for the $j$th stratum as $\sum I(\delta_i = 3, Z_i = z_j) = n_{z_j}$. It is possible that for some strata $j$ $n_{z_j} = 0$. This can happen for one of two reasons; first, the sampling scheme excludes them with $s_j = 0$, a situation referred to as deterministic nonpositivity. Alternatively, random nonpositivity can arise if no one who has $Z = z_j$ happened to be sampled (Westreich and Cole, 2010), which may be likely if the sampling probability $s_j$ is nonzero but very small. In this paper we assume there is deterministic positivity such that each stratum has at least some positive probability of inclusion in the sample, $s_j > 0$, $j = 1, \ldots, k$. In this section, if random nonpositivity arises, we use the simple restriction strategy of redefining the target population to only consist of strata $j$ for which $n_{z_j} > 0$ (Westreich and Cole, 2010). A model-based alternative strategy to restriction is introduced in Section 2.3.

*Estimators and Statistical Properties.* Note that $\rho$ is now a weighted average of the stratum-conditional means $\rho_j$ weighted by the known stratum proportions $\gamma_j$, $\rho = \sum_{j=1}^{k} \rho_j \gamma_j$. A simple nonparametric standardization estimator for $\rho$ is $\hat{\rho}_{st} = \sum_{j=1}^{k} \hat{\rho}_j \gamma_j$, where $\hat{\rho}_j = \left(\sum I(Z_i = z_j, \delta_i = 3) X_i\right) / n_{z_j}$ uses each stratum-specific sample prevalence as an estimator,

$j = 1, \ldots, k$. (These observed prevalences are, of course, mismeasured unless $\sigma_e = \sigma_p = 1$.) A final Rogan-Gladen standardization prevalence estimator is then $\hat{\pi}_{st} = (\hat{\rho}_{st} + \hat{\sigma}_p - 1)/(\hat{\sigma}_e + \hat{\sigma}_p - 1)$. Let $\theta = (\sigma_e, \sigma_p, \rho_1, \ldots, \rho_k, \rho, \pi)^T$ denote the $(k+4)$-vector of true parameter values, and notice the estimators just discussed are solutions to a $(k+4)$-vector of estimating equations

$$\sum \psi(X_i, Z_i; \delta_i, \theta) = \left( \sum \psi_e, \sum \psi_p, \sum \boldsymbol{\psi}_{\boldsymbol{\rho}}, \psi_\rho, \psi_\pi \right)^T = 0$$

In the above vector $\sum \psi_e, \sum \psi_p$, and $\psi_\pi$ are identical to the equations from Section 2.1; $\sum \boldsymbol{\psi}_{\boldsymbol{\rho}}$ is a $k$-vector with $j$th element $\sum I(Z_i = z_j, \delta_i = 3)(X_i - \rho_j)$; and $\psi_\rho = \sum_{j=1}^k \rho_j \gamma_j - \rho$. Using similar techniques to those used in Section 2.1, we show that

$$\sqrt{n}(\hat{\theta} - \theta) \rightarrow_d \mathcal{N}\left(0, \mathbb{A}(\theta)^{-1}\mathbb{B}(\theta)\mathbb{A}(\theta)^{-T}\right) \text{ and } \sqrt{n}(\hat{\pi}_{st} - \pi) \rightarrow_d \mathcal{N}\left(0, V_{\pi,st}\right)$$

where $\mathbb{A}(\theta)^{-1}\mathbb{B}(\theta)\mathbb{A}(\theta)^{-T}$ is a $(k+4) \times (k+4)$ variance-covariance matrix with lower-right element

$$V_{\pi,st} = \left( \frac{\pi^2 \sigma_e(1-\sigma_e)}{c_1} + \frac{(1-\pi)^2 \sigma_p(1-\sigma_p)}{c_2} + \sum_{j=1}^k \frac{\gamma_j^2 \rho_j(1-\rho_j)}{c_3 s_j} \right) (\sigma_e + \sigma_p - 1)^{-2}$$

It follows that $\hat{\pi}_{st}$ is approximately distributed $\mathcal{N}(\pi, V_{\pi,st})$ as $n$ grows infinitely large. Variance estimation proceeds by substituting $\hat{c}_i = n_i/n$ for $c_i$, $i = 1, 2, 3$, $\hat{s}_j = n_{z_j}/n_3$ for $s_j$, $j = 1, \ldots, k$, and using the previously-considered estimators $\hat{\sigma}_e, \hat{\sigma}_p, \hat{\rho}_1, \ldots, \hat{\rho}_k, \hat{\rho}_{st}, \hat{\pi}_{st}$. The resulting variance estimator is

$$\hat{V}_{\pi,st}/n = \left( \frac{\hat{\pi}_{st}^2 \hat{\sigma}_e(1-\hat{\sigma}_e)}{n_1} + \frac{(1-\hat{\pi})^2 \hat{\sigma}_p(1-\hat{\sigma}_p)}{n_2} + \sum_{j=1}^k \frac{\gamma_j^2 \hat{\rho}_j(1-\hat{\rho}_j)}{n_{z_j}} \right) (\hat{\sigma}_e + \hat{\sigma}_p - 1)^{-2} \quad (3)$$

and is consistent for the true variance. Recall that the $\gamma_j$s are known and do not need to be estimated. The proofs of asymptotic normality and variance estimator consistency, and the derivation of the variance estimator, are included in Web Appendix B.

### 2.3 *Modeling Mismeasured Stratum-Specific Prevalences*

*Problem Setup.* Standardization required estimating each mismeasured stratum-specific prevalence $\rho_j = P(X = 1 | Z = z_j)$. The restriction strategy led to each $n_{z_j}$ being a positive nonzero integer, which was necessary to compute the nonparametric estimator

$\hat{\rho}_j = (\sum I(Z_i = z_j, \delta_i = 3)X_i)/n_{z_j}$. In this section we fit a logistic regression to estimate $\rho_j$ in cases where random nonpositivity arises and $n_{z_j} = 0$ for certain strata $j \in \{1, \ldots, k\}$.

The logistic model is $\text{logit}(\rho_j) = \beta^T h(z_j)$, where $\beta$ is a $p$-vector of regression coefficients with intercept $\beta_1$. $h(z)$ is a user-specified $p$-vector function of the covariates $z$ that may include indicator variables and interaction terms. Let $\text{supp}(z)$ be the covariate support of $z$ in the sample, that is, $\text{supp}(z) = \{Z_j : n_{z_j} > 0\}$ with dimension $\dim(\text{supp}(z)) = \sum_{j=1}^{k} I(n_{z_j} > 0)$. The model assumes $p \leqslant \dim(\text{supp}(z)) \leqslant k$, with $\dim(\text{supp}(z)) = k$ only when there is no non-positivity (and the estimators of Section 2.2 can be used with no restriction needed). The $j$th element of $h(z)$ is denoted $h_j(z)$, with $h_1(z)$ set equal to one to correspond to the intercept.

*Estimators and Statistical Properties.* Each stratum-specific mismeasured prevalence $\rho_j$ is no longer a parameter, but a function of the estimated regression coefficients $\hat{\beta}$ and that stratum's covariates, denoted $\rho_j(\hat{\beta}, z_j) = \text{logit}^{-1}(\hat{\beta}^T h(z_j))$. The final model-based Rogan-Gladen model-based standardization estimator is then $\hat{\pi}_{mst} = (\hat{\rho}_{mst} + \hat{\sigma}_p - 1)/(\hat{\sigma}_e + \hat{\sigma}_p - 1)$, where $\hat{\rho}_{mst} = \sum_{j=1}^{k} \hat{\rho}_j(\hat{\beta}, z_j)\gamma_j$. An estimating equations approach can again be used to analyze the estimator, where the $k$ estimating equations for $\rho_1, \ldots, \rho_k$ from Section 2.2 are replaced with $p+1$ estimating equations for $\beta_1, \ldots, \beta_p$, corresponding to the score equations from logistic regression. Let $\theta = (\sigma_e, \sigma_p, \beta_1, \ldots, \beta_p, \rho, \pi)^T$ denote the $(p+4)$-vector of true parameter values. The estimators are then solutions to the $(p+4)$-vector of estimating equations:

$$\sum \psi(X_i, Z_i; \delta_i, \theta) = \left(\sum \psi_e, \sum \psi_p, \sum \psi_\beta, \psi_\rho, \psi_\pi\right)^T = 0$$

In the above vector $\sum \psi_e, \sum \psi_p$, and $\psi_\pi$ are identical to the equations from Section 2.2; $\psi_\rho = \sum_{j=1}^{k} \text{logit}^{-1}(\beta^T h(z_j))\gamma_j - \rho$; and $\sum \psi_\beta$ is a $p$-vector with $j$th element $\sum \psi_{\beta_j} = \sum I(\delta_i = 3)(X_i - \text{logit}^{-1}\{\beta^T h(Z_i)\})h_j(Z_i)$. Using similar techniques to those in Sections 2.1 and 2.2,

we show that

$$\sqrt{n}(\hat{\theta}-\theta) \rightarrow_{d.} \mathcal{N}\left(0, \mathbb{A}(\theta)^{-1}\mathbb{B}(\theta)\mathbb{A}(\theta)^{-T}\right) \text{ and } \sqrt{n}(\hat{\pi}_{mst}-\pi) \rightarrow_{d.} \mathcal{N}\left(0, V_{\pi,mst}\right)$$

where $\mathbb{A}(\theta)^{-1}\mathbb{B}(\theta)\mathbb{A}(\theta)^{-T}$ is a $(p+4)\times(p+4)$ variance-covariance matrix with lower-right element $V_{\pi,mst}$. The proof of asymptotic normality and form of $V_{\pi,mst}$ are in Web Appendix C. It follows that $\hat{\pi}_{mst}$ is approximately distributed $\mathcal{N}(\pi, V_{\pi,mst})$ as $n$ grows infinitely large. Variance estimation involves substituting $\hat{c}_j = n_j/n$ for $c_j$, $j=1,2,3$, using $\hat{\sigma}_e, \hat{\sigma}_p, \hat{\rho}_{mst}, \hat{\pi}_{mst}$, and empirical estimators for other components of $V_{\pi,mst}$. The resulting variance estimator is denoted $\hat{V}_{\pi,mst}/n$ and is also expressed in Web Appendix C.

## 3. Simulations

Simulation studies were conducted in order to assess (a) the mean empirical bias, $\hat{\pi}-\pi$, of each estimator; (b) the comparison between the mean asymptotic standard error $\sqrt{\hat{V}_\pi/n}$ (ASE) and mean empirical standard error (ESE), computed as the standard error of $\hat{\pi}$; and (c) if 95% Wald-type confidence intervals based on the ASE attain approximately nominal coverage. Three general settings were evaluated: (i) The setting of Section 2.1, where there is measurement error but no selection bias; (ii) The selection bias setting of Section 2.2 with a single covariate $Z$ taking one of two levels; (iii) A more complex selection bias setting with a higher-dimensional covariate vector $Z$.

Simulations for setting (i) used the following steps in each of $n_{sim} = 2500$ iterations:

(1) Data $X_1, \ldots, X_{n_1}$ are generated with true sensitivity $\sigma_e$ and sample size $n_1$. Data $X_{n_1+1}, \ldots, X_{n_2}$ are generated with true specificity $\sigma_p$ and sample size $n_2$.

(2) The sample from the population $X_{n_2+1}, \ldots, X_n$ is generated based on $Y, \sigma_e$, $\sigma_p$, and $\pi$.

(3) $\hat{\pi}$ is computed from the observed data $(X_i, \delta_i)_{i=1}^n$ using the Rogan-Gladen approach detailed in Section 2.1. A 95% Wald-type confidence interval for $\pi$ is constructed as $\left[\hat{\pi} \pm 1.96\sqrt{\hat{V}_\pi}\right]$

Simulations for setting (ii) used the following steps in each of $n_{sim} = 2500$ iterations:

(1) Validation data are generated as in step 1 from setting (i).

(2) One discrete covariate with two levels is defined as $Z \in \{z_1, z_2\}$. Known stratum proportions were $\gamma_1 = \gamma_2 = .5$. The distribution of $Z$ in the sample is assumed to be $Z \sim \text{Multi}_2(n_3, (.2, .8))$.

(3) The conditional distribution of the true outcome given covariates is assumed to be $P(Y = 1|Z = z_1) = .075$ and $P(Y = 1|Z = z_2) = .025$, so true prevalence was $\pi = \sum_{j=1}^{2} P(Y = 1|Z = z_j)\gamma_j = .05$ .

(4) The sample from the population $X_{n_2+1}, \ldots, X_n$ is generated based on $Y, Z, \sigma_e$, and $\sigma_p$.

(5) $\hat{\pi}_{st}$ is computed from the observed data $(X_i, Z_i, \delta_i)_{i=1}^{n}$ using the estimating equations approach detailed in Section 2.2. A 95% Wald-type confidence interval for $\pi$ is constructed as $\left[\hat{\pi}_{st} \pm 1.96\sqrt{\hat{V}_{\pi,st}}\right]$. The Rogan-Gladen estimator $\hat{\pi}$ and a confidence interval based on $\hat{V}_{\pi}$ are computed for comparison.

Simulations for setting (iii) used the following settings in each of $n_{sim} = 500$ iterations. Two different settings are considered in Step 2, leading to settings (iii)(a) and (iii)(b).

(1) Data $X_1, \ldots, X_{n_1}$ are generated with true sensitivity $\sigma_e$ and sample size $n_1$. Data $X_{n_1+1}, \ldots, X_{n_2}$ are generated with true specificity $\sigma_p$ and sample size $n_2$.

(2) Three covariates are defined as $Z_1 \in \{z_{10}, z_{11}\}$, $Z_2 \in \{z_{20}, z_{21}, z_{22}, z_{23}\}$, and $Z_3 \in \{z_{30}, z_{31}, z_{32}, z_{33}, z_{34}\}$. There were thus $k = 2 \times 4 \times 5 = 40$ strata with known stratum proportions $\gamma_j, j = 1, \ldots, 40$. The distribution of $Z$ in the sample is assumed to be multinomial with 40 categories, sample size $n_3$, and known probability vector $(s_1, \ldots, s_{40})^T$. The $\gamma_j$s and $s_j$s for settings (iii)(a) and (iii)(b), respectively, are enumerated in Web Tables 1 and 2. Setting (iii)(a) featured some very small sampling probabilities $s_j$ like 0.00001 for small strata (low $\gamma_j$), while setting (iii)(b), perhaps less realistically, featured very small $s_j$s for some strata of relatively large proportions.

(3) The conditional distribution of the true outcome given covariates, $Y|Z$, is also assumed to be binomial with expectation according to a logistic model for the stratum-specific true prevalence:

$$P(Y=1|Z) = \{1+\exp(-\beta_0 - \beta_1 I(Z_1 = z_{11}) - \beta_2 I(Z_2 = z_{20}) - \beta_3 I(Z_2 = z_{21}) - \beta_4 I(Z_3 = z_{30}) - \beta_5 I(Z_3 = z_{31})\}^{-1}$$

The parameters $\beta_1 = -1$, $\beta_2 = -.6$, $\beta_3 = .8$, $\beta_4 = .6$, $\beta_5 = .4$ were set to reflect differential prevalences by stratum. The intercept $\beta_0$ was set to $-2.6375$ so that prevalence was approximately $\pi = \sum_{j=1}^{k} P(Y=1|Z=z_j)\gamma_j \approx 0.05$.

(4) The sample from the population $X_{n_2+1}, \ldots, X_n$ is generated based on $Z, Y|Z, \sigma_e$, and $\sigma_p$.

(5) $\hat{\pi}_{mst}$ is computed from the observed data $(X_i, Z_i, \delta_i)_{i=1}^n$ using the estimating equations approach with a correctly-specificed outcome regression model detailed in Section 2.3. A 95% Wald-type confidence interval for $\pi$ is constructed as $\left[\hat{\pi}_{mst} \pm 1.96\sqrt{\hat{V}_{\pi,mst}}\right]$. The estimators $\hat{\pi}_{st}$ and $\hat{\pi}$ and their respective confidence intervals are computed for comparison.

Simulation settings (i) and (ii) use the parameter values of $\sigma_e = \sigma_p = .98, n_1 = n_2 = 1000, n_3 = 20000$. Simulation setting (iii) used the same settings, except $n_3$ was set equal to 2000 to induce greater amounts of random nonpositivity, leading to an average of 10/40 strata randomly having no observed data ($n_{z_j} = 0$). Results are displayed in Table 1. Setting (iii)(a) **Use higher values of $n_{sim}$ when settings of simulations are finalized.**

[Table 1 about here.]

The Rogan-Gladen estimator $\hat{\pi}$ works well in setting (i), where there is measurement error but no selection bias, with a very low mean bias and empirical 95% confidence interval coverage of 94.0%. However, when selection bias is introduced in setting (ii), the average bias is now -.015, meaning that the method significantly overestimates a true prevalence of 5% to be on average 6.5%, and CI coverage is 6.4%. The nonparametric Rogan-Gladen with

standardization estimator $\hat{\pi}_{st}$ has good bias and coverage properties, with near-nominal CI coverage of 94.1%.

Though random nonpositivity is introduced in setting (iii), the model-based standardization estimator $\hat{\pi}_{mst}$ did not perform better than the nonparametric standardization estimator $\hat{\pi}_{st}$. In setting (iii)(a) bias was low for both estimators, but the average bias for $\hat{\pi}_{st}$ was -.0003, much smaller in magnitude than the corresponding measure of -.0011 for $\hat{\pi}_{mst}$. Coverage was closer to nominal for $\hat{\pi}_{mst}$ at 95.6% compared to 90.2% for $\hat{\pi}_{st}$. However, the ASE was farther away from the ASE for $\hat{\pi}_{mst}$ than for $\hat{\pi}_{st}$, so the coverage results could be an artifact of the simulation settings. In setting (iii)b $\hat{\pi}_{mst}$ displayed more bias than even $\hat{\pi}$, and its ASE was much smaller than the ESE. Surprisingly, its coverage was closer to nominal than that of $\hat{\pi}_{st}$, but this could again be an artifact of the combination between $\hat{\pi}_{mst}$ overestimating the true prevalence and $\hat{V}_{\pi,mst}$ underestimating the true variance.

**I think it could be worth investigating improvements to the methodology of the model-based estimator $\hat{\pi}_{mst}$. One idea that comes to mind is to only use the model-based estimator for strata where there is nonpositivity, while the nonparametric estimator is used for all strata where data are observed. What are your thoughts?**

## 4. Application to a Seroprevalence Study

The methods developed in Section 2 were applied to a serosurvey, ScreenNC, which tested a convenience sample of asymptomatic patients in North Carolina for antibodies to SARS-CoV-2 (Barzin et al., 2020). ScreenNC assessed $n_3$ = 2,973 patients age 20+ who were asymptomatic for COVID-19 and seeking unrelated medical care at nine outpatient clinical sites and two emergency room sites associated with the University of North Carolina (UNC) Health Network. Blood draws were taken between April 28 to June 19, 2020, several weeks after the first COVID-19 case was reported in NC (March 2) and stay-at-home orders were

commenced (March 30). The Abbott Architect SARS-CoV-2 IgG assay was used to detect the presence of antibodies. The UNC lab that conducted the assays also conducted validation studies that assessed sensitivity by testing patients confirmed to be positive for SARS-CoV-2 with RT-PCR ($n_1 = 40$), and assessed specificity by testing remnant serum samples, mainly before January 1st, 2020, assumed to be negative for SARS-CoV-2 ($n_2 = 277$).

The Rogan-Gladen with standardization estimator of Section 2.2 was used to analyze ScreenNC, though care needs to be taken regarding the applicability of assumptions. The ScreenNC sample was compared to the population of patients accessing the UNC Health Network during the same timeframe ($n$ = 21,901) and there was differential sampling by the covariates of age group (7 strata), race (5 strata), and sex (2 strata) (Barzin et al., 2020). We implicitly assumed that the sample was like a stratified random sample by these three covariates only. While this assumption may be reasonable, the relatively small size of this UNC target population may violate the infinite superpopulation assumption. We thus also standardized the results to the 2019 NC population ($n$ = 7,873,971) using Census Bureau data from the American Community Survey (U.S. Census Bureau, 2019), though the stratified random sampling assumption is less reasonable for this target population because not all NC residents are in the UNC Health Network. **Add Table 1 from Barzin et al., plus include ACS demographics, to show how biased the sampling is.**

Of the $70 = 7 \times 5 \times 2$ strata in the UNC target population, there were three strata for which there was no sample data, so restriction was used to deal with this non-positivity; that is, a new target population was derived that included only the 67 strata with sample data. The 2019 ACS did not have data on people who refused to give race or who had unknown race, so sample patients in that racial stratum were recategorized into Other race, leading to $56 = 7 \times 4 \times 2$ total strata. Of these, two strata did not have sample data, so the target population was restricted to the 54 relevant strata. The logistic regression methods of Section

2.3 were also used to account for the non-positivity, with the main effects model

$$\begin{aligned}\text{logit}(\rho_j) = \beta_1 &+ \beta_2 SexMale_j + \beta_3 RaceBlack_j + \beta_4 RaceOther_j + \beta_5 RaceUnknown_j \\ &+ \beta_6 RaceWhite_j + \beta_7 Age30to39_j + \beta_8 Age40to49_j + \beta 9 Age50to59_j + \beta_{10} Age60to69_j \\ &+ \beta_{11} Age70to79_j + \beta_{12} Age80Plus_j\end{aligned}$$

fit for the UNC target population, where each variable is an indicator variable. Interaction effects were not considered due to the small number of positive test results, which could lead to model overfitting. The same model was fit for the NC census target population, but excluding the unknown race variable.

The uncorrected sample percentage of positive tests was $\hat{\rho} = n_3^{-1} \sum I(\delta_i = 3) X_i = 24/2973 = 0.81\%$. Using the Rogan-Gladen methodology from Section 2.1, with no standardization, led to a seroprevalence point estimate of -0.28% (95% CI -1.56%, 1.00%). Since the Rogan-Gladen approach can yield estimates outside of the true parameter range of $(0, 1)$, the final estimates should be truncated to be inside that range for reporting purposes (Hilden, 1979). The nonparametric Rogan-Gladen with standardization results using $\hat{\pi}_{st}$ and $\hat{V}_{\pi,st}$ were -0.23% (-1.52%, 1.06%) for the UNC target population and -0.20% (-1.50%, 1.10%) for the NC census target population. Lastly, the model-based Rogan-Gladen with standardization results using $\hat{\pi}_{mst}$ and $\hat{V}_{\pi,mst}$ were -0.20% (-4.47%, 4.07%) for the UNC target population and -0.19% (-4.46%, 4.08% ) for the NC census target population.

## 5. Discussion

Our contributions were the demonstration of beneficial large-sample properties of the Rogan-Gladen with standardization estimator and derivation of a variance estimator for use in confidence interval construction. Simulation studies confirmed that the variance estimator performed well in finite samples. While correct use of the two-stage bootstrap for variance estimation requires not just computational power but statistical programming expertise,

our variance estimator is a simple formula that could be implemented in basic spreadsheet software such as Microsoft Excel. The use of estimating equations was also shown to be useful for finding efficient estimators when combining multiple datasets. We are aware of only a small number of papers that address this approach (Lawless, Kalbfleisch, and Prentice, 1999; Hirose, 2007; Lee and Hirose, 2010) and it may warrant further investigation.

The simulation studies demonstrated that the Rogan-Gladen with standardization estimator can be necessary in situations of selection bias, like settings (ii) and (iii). The traditional Rogan-Gladen estimator leads to relatively large amounts of bias in these situations, and its 95% confidence intervals cover the true prevalence less than 10% of the time. However, the proposed estimators display good performance in terms of both bias and sampling. The point estimator is approximately unbiased and its confidence interval attains approximately nominal empirical coverage in setting (ii). **Address setting (iii) when complete**.

In the real data application to a serosurvey of asymptotic patients in North Carolina, nonparametric standardization estimates were very similar to estimates made without standardization. This may be due to the small number of positive case, or due to ScreenNC not having dramatically biased sampling on the covariates of race, sex, and age. **Add 1-2 sentences on outcome regression results. Any other ideas on why the results were so similar in the application?**

The present statistical approach has several limitations, so we discuss three alternative frameworks which statisticians could consider. First, serologic tests yield quantitative test measures that are nearly always reduced to a variable such as $Y$ taking on values of positive or negative, but such a dichotomization loses statistical information. Some likelihood-based and mixture-model methods have been proposed (van Boven et al., 2017; Bouman et al., 2020) and such approaches warrant attention. Second, while the approach of Section 2.1 has most frequently been used (Messam et al., 2008) and is computationally simple, Bayesian methods

have also been used for prevalence estimation (Gelman and Carpenter, 2020; Larremore et al., 2020a). A downside to these methods is that they require the analyst to specify prior distributions on parameters of interest such as $\sigma_e, \sigma_p$, and $\pi$. On the other hand, this property does allow for parameter support to be constrained to $[0, 1]$ and for flexible variance estimation, so Bayesian analogues to the methods in Sections 2.2 and 2.3 could be a useful accompaniment to this work. Third, as is made clear from the application, the assumption of an infinite target population may sometimes be unreasonable. A more realistic, though technically demanding, frequentist approach would be to assume a finite population and make use of finite population sampling theory and central limit theorems (Thompson, 1997). Such an assumption would allow for estimating the marginal probability of selection into the sample and the conditional probability of selection given covariates, and thus could enable the use of inverse probability of sampling weights (Buchanan et al., 2018; Lesko et al., 2017) or inverse odds of sampling weights (Westreich et al., 2017).

There are several more natural extensions to this work. The Rogan-Gladen estimator is both biased and can fall outside $[0, 1]$, so we investigated bias-corrected estimators that subtract the bias from the biased estimator, as well as the use of the adjusted maximum likelihood method of Rahme and Joseph (1998). While we were not able to improve on the Rogan-Gladen estimator, and our tentative studies are omitted, improvements may be possible, including those that target estimation when the true values of parameters $\sigma_e$, $\sigma_p$, and $\pi$ are near their boundaries of 0 and 1. Regarding Sections 2.2 and 2.3, our assumption that $n_{z_j} > 0$ for all $j$ seemed to preclude the use of continuous covariates in standardization, but other methods that would allow for such covariates may exist. Further, though logistic regression is a tractable choice of outcome model for asymptotic derivations, nonparametric classification alternatives such as tree-based methods could be considered.

A truer sense of the biostatistical opportunities, and challenges, associated with estimating

prevalence for SARS-CoV-2 requires looking beyond the problem as stated. While standardization methods can yield more accurate results than ignoring selection bias, recent perspectives have pointed out that post-hoc statistical adjustments cannot be as effective as designing more representative studies using probability-based sampling (Shook-Sa et al., 2020; Boyce, Shook-Sa, and Aiello, 2021). Another challenge for serosurveys, and antibody testing generally, is that assay sensitivity is often assessed using samples from hospitalized patients who were sicker, and so may have higher antibody titers, than the average person infected with SARS-CoV-2. When these validation data are used to determine the cutoff for assay positivity, it induces "spectrum bias" which can lead to a higher false negative rate than the validation studies would suggest (Takahashi, Greenhouse, and Rodríguez-Barraquer, 2020). Lastly, there are several interesting open problems in seroprevalence. For instance, two areas that may warrant attention are power calculations for both the validation and main study sample sizes (see Larremore et al., 2020b for a Bayesian approach) and deriving estimators for studies where positive patients are retested. Perhaps in years to come the Rogan-Gladen with standardization estimator could be one method among a suite of complementary techniques used in seroprevalence studies.

## Acknowledgements

**Acknowledgements go here**

## References

Barzin, A., Schmitz, J. L., Rosin, S., et al. (2020). SARS-CoV-2 Seroprevalence among a Southern U.S. Population Indicates Limited Asymptomatic Spread under Physical Distancing Measures. *MBio*, **11** (5), e02426-20.

Boos, D. D., & Stefanski, L. A. (2013). M-Estimation (Estimating Equations). In D. D. Boos & L. A. Stefanski, *Essential Statistical Inference* (2013, pp. 297–337). Springer New York.

Bouman, J. A., Riou, J., Bonhoeffer, S., & Regoes, R. R. (2020). Estimating cumulative incidence of SARS-CoV-2 with imperfect serological tests: Exploiting cutoff-free approaches Preprint. medRXiv. https://doi.org/10.1101/2020.04.29.068999.

Boyce, R., Shook-Sa, B., & Aiello, A. (2020). A tale of two studies: Study design and our understanding of SARS-CoV-2 seroprevalence. Accepted Manuscript. *Clinical Infectious Diseases.* https://doi.org/10.1093/cid/ciaa1868.

Buchanan, A. L., Hudgens, M. G., Cole, S. R., Mollan, K. R., Sax, P. E., Daar, E. S., Adimora, A. A., Eron, J. J., & Mugavero, M. J. (2018). Generalizing Evidence from Randomized Trials using Inverse Probability of Sampling Weights. *Journal of the Royal Statistical Society. Series A, (Statistics in Society)*, **181** (4), 1193–1209. https://doi.org/10.1111/rssa.12357

Gelman, A., & Carpenter, B. (2020). Bayesian analysis of tests with unknown specificity and sensitivity. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **69** (5), 1269–1283. https://doi.org/10.1111/rssc.12435.

Havers, F. P., Reed, C., Lim, T., et al. (2020). Seroprevalence of Antibodies to SARS-CoV-2 in 10 Sites in the United States, March 23-May 12, 2020. *JAMA Internal Medicine* **180** (12). https://doi.org/10.1001/jamainternmed.2020.4130.

Hernán M. A., Robins, J. M. (2020). *Causal Inference: What If.* Chapman & Hall/CRC.

Hilden, J. (1979). A further comment on "Estimating prevalence from the results of a screening test." *American Journal of Epidemiology*, **109** (6), 721–722. https://doi.org/10.1093/oxfordjournals.aje.a112737.

Hirose, Y. (2007). *M*-Estimators in Semi-Parametric Multi-Sample Models. Unpublished manuscript. http://www.mcs.vuw.ac.nz/research/publications/reports/mscs/mscs08-05.pdf.

Larremore, D. B., Fosdick, B. K., Bubar, K. M., Zhang, S., Kissler, S. M., Metcalf, C. J.

E., Buckee, C., & Grad, Y. (2020). Estimating SARS-CoV-2 seroprevalence and epidemiological parameters with uncertainty from serological surveys. Preprint. medRXiv. https://doi.org/10.1101/2020.04.15.20067066.

Larremore, D. B., Fosdick, B. K., Zhang, S., & Grad, Y. H. (2020b). Jointly modeling prevalence, sensitivity and specificity for optimal sample allocation. Preprint. bioRXiv. https://doi.org/10.1101/2020.05.23.112649.

Lawless, J. F., Kalbfleisch, J. D., & Wild, C. J. (1999). Semiparametric methods for response-selective and missing data problems in regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **61** (2), 413–438. https://doi.org/10.1111/1467-9868.00185.

Lee, A., & Hirose, Y. (2010). Semi-parametric efficiency bounds for regression models under response-selective sampling: The profile likelihood approach. *Annals of the Institute of Statistical Mathematics*, **62** (6), 1023–1052. https://doi.org/10.1007/s10463-008-0205-1.

Lesko, C. R., Buchanan, A. L., Westreich, D., Edwards, J. K., Hudgens, M. G., & Cole, S. R. (2017). Generalizing study results: A potential outcomes perspective. *Epidemiology*, **28** (4), 553–561. https://doi.org/10.1097/EDE.0000000000000664.

Marchevsky, N. (1979). Re: "Estimating prevalence from the results of a screening test." *American Journal of Epidemiology*, **109** (6), 720–721. https://doi.org/10.1093/oxfordjournals.aje.a112736.

Messam, L. L. McV., Branscum, A. J., Collins, M. T., & Gardner, I. A. (2008). Frequentist and Bayesian approaches to prevalence estimation using examples from Johne's disease. *Animal Health Research Reviews*, **9** (1), 1–23. https://doi.org/10.1017/S1466252307001314.

Miettinen, O.S. (1985). *Theoretical Epidemiology: Principles of Occurrence Research in*

*Medicine*. John Wiley & Sons. (pp. 266-271).

Rahme, E., & Joseph, L. (1998). Estimating the prevalence of a rare disease: Adjusted maximum likelihood. *Journal of the Royal Statistical Society: Series D (The Statistician)*, **47** (1), 149–158. https://doi.org/10.1111/1467-9884.00120.

Rogan, W. J., & Gladen, B. (1978). Estimating Prevalence from the Results of a Screening Test. *American Journal of Epidemiology*, **107** (1).

Sempos, C. T., & Tian, L. (2021). Adjusting Coronavirus Prevalence Estimates for Laboratory Test Kit Error. *American Journal of Epidemiology*, **190** (1), 109–115. https://doi.org/10.1093/aje/kwaa174.

Shook-Sa, B. E., Boyce, R. M., & Aiello, A. E. (2020). Estimation Without Representation: Early Severe Acute Respiratory Syndrome Coronavirus 2 Seroprevalence Studies and the Path Forward. *The Journal of Infectious Diseases*, **222** (7), 1086-1089. https://doi.org/10.1093/infdis/jiaa429.

Smolinski, M. S., Hamburg, M. A., Lederberg, J., eds. (2003). *Microbial threats to health: emergence, detection, and response*. Washington, D.C.: National Academy Press.

Takahashi, S., Greenhouse, B., & Rodriguez-Barraquer, I. (2020). Are SARS-CoV-2 Seroprevalence Estimates Biased? *The Journal of Infectious Diseases*, **222** (11), 1772-1775. https://doi.org/10.1093/infdis/jiaa523.

Thompson, M. (1997). *Theory of sample surveys* (Vol. 74). CRC Press.

U.S. Census Bureau (2019). American Community Survey, 2019 American Community Survey 1-Year Estimates, Public Use Microdata Sample 2019. Table generated by Samuel Rosin using data.census.gov. Retrieved from https://data.census.gov/mdat/#/, 2 February 2020.

Van Belle, G., Fisher, L. D., Heagerty, P. J., & Lumley, T. (2004). *Biostatistics: a Methodology for the Health Sciences* (2nd ed., pp. 640-648). John Wiley & Sons. https://doi.org/

10.1002/0471602396

van Boven, M., van de Kassteele, J., Korndewal, M. J., van Dorp, C. H., Kretzschmar, M., van der Klis, F., de Melker, H. E., Vossen, A. C., & van Baarle, D. (2017). Infectious reactivation of cytomegalovirus explaining age- and sex-specific patterns of seroprevalence. *PLoS Computational Biology*, **13** (9). https://doi.org/10.1371/journal.pcbi.1005719

Westreich, D., & Cole, S. R. (2010). Invited Commentary: Positivity in Practice. *American Journal of Epidemiology*, **171** (6), 674–677. https://doi.org/10.1093/aje/kwp436.

Westreich, D., Edwards, J. K., Lesko, C. R., Stuart, E., & Cole, S. R. (2017). Transportability of Trial Results Using Inverse Odds of Sampling Weights. *American Journal of Epidemiology*, **186** (8), 1010–1014. https://doi.org/10.1093/aje/kwx164.

## Supporting Information

Web Appendices A-C, referenced in Sections 2.1, 2.2, and 2.3 respectively, and Web Table 1 referenced in Section 3, are available with this paper at the Biometrics website on Wiley Online Library. R code based on the simulations in Section 3 and data analysis in Section 4 is included on Wiley Online Library and is also available at [**add code to GitHub and link**].

*Received xx* 202x. *Revised yy* 202y. *Accepted zz* 202z.

**Table 1**

*Simulation results from Section 3. $\hat{\pi}$ denotes the Rogan-Gladen estimator of Section 2.1, $\hat{\pi}_{st}$ denotes the Rogan-Gladen with standardization estimator of Section 2.2, and $\hat{\pi}_{mst}$ denotes the model-based Rogan-Gladen with standardization estimator of Section 2.3. For each estimator $\hat{\pi}$, bias is the average of $\hat{\pi} - \pi$, ESE is the empirical standard error of $\hat{\pi}$, ASE is the mean asymptotic standard error, and EC is empirical coverage. Settings (i) and (ii) are based on 2500 simulations while setting (iii) is based on 500 simulations.*

| Setting | Estimator | Bias | ESE | ASE | EC |
|---|---|---|---|---|---|
| (i) | $\hat{\pi}$ | $5.4 \times 10^{-5}$ | .005 | .005 | 94.0% |
| (ii) | $\hat{\pi}$ | $-.015$ | .005 | .005 | 6.4% |
| (ii) | $\hat{\pi}_{st}$ | $-6.1 \times 10^{-5}$ | .005 | .005 | 94.1% |
| (iii)(a) | $\hat{\pi}$ | $-.022$ | .007 | .007 | 8.2% |
| (iii)(a) | $\hat{\pi}_{st}$ | $-2.6 \times 10^{-4}$ | .0148 | .0132 | 90.2% |
| (iii)(a) | $\hat{\pi}_{mst}$ | $-1.1 \times 10^{-3}$ | .0136 | .0156 | 95.6% |
| (iii)(b) | $\hat{\pi}$ | $-.012$ | .007 | .007 | 62.0% |
| (iii)(b) | $\hat{\pi}_{st}$ | .004 | .025 | .028 | 82.6% |
| (iii)(b) | $\hat{\pi}_{mst}$ | .025 | .066 | .024 | 90.0% |