# SARS-COV-2 SEROPREVALENCE AND VACCINE CORRELATE OF PROTECTION STANDARDIZATION

Samuel P. Rosin

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biostatistics in the Gillings School of Global Public Health.

Chapel Hill
2023

Approved by:

Michael G. Hudgens

Bonnie E. Shook-Sa

Haibo Zhou

Stephen R. Cole

Jessie K. Edwards

## ABSTRACT

Samuel P. Rosin: SARS-CoV-2 Seroprevalence and Vaccine Correlate of Protection Standardization
(Under the direction of Michael G. Hudgens and Bonnie E. Shook-Sa)

In the COVID-19 pandemic, there was great interest in population seroprevalence estimation of individuals with antibodies against SARS-CoV-2 and in evaluation of antibodies as surrogate markers for vaccine efficacy. In the first paper, methods for estimation of seroprevalence from surveys which can have selection bias and serologic tests which can have measurement error are presented. These challenges are addressed with leveraging of auxiliary data, e.g., population census data, and of laboratory studies of false positive and false negative rates. Direct standardization is used for development of nonparametric and parametric seroprevalence estimators. The estimators are proven consistent and asymptotically normal. Simulation studies demonstrate performance across a variety of selection and misclassification biases scenarios. The proposed methods are applied to SARS-CoV-2 seroprevalence studies in New York City, Belgium, and North Carolina.

Drawing simple comparisons of COVID-19 vaccine trial efficacy estimates is problematic without considering factors affecting trial context and design, including characteristics of a study's population (Rapaka et al., 2022). A meta-analytic paradigm for surrogate endpoint evaluation entails estimating an association between the treatment effects on the surrogate and clinical endpoints, respectively, using data from multiple clinical trials. This approach can estimate the association between vaccine induced anti-SARS-CoV-2 antibodies and vaccine efficacy against symptomatic COVID-19 illnesss. In the second paper, multiple vaccine trials are standardized to a common target population. Meta-analytic causal association parameters, estimators, and the asymptotic distributions of the estimators are considered. A hypothesis test of an implication of a conditional exchangeability assumption is proposed. Simulation studies demonstrate the methods in scenarios motivated by data from several U.S. government Phase 3 SARS-CoV-2 vaccine trials.

When data are fused across data sets, often the random variables are assumed to be independent but not identically distributed, as in the preceding chapters. However, standard estimating equation theory assumes an independent and identically distributed set up. In the third chapter, the consistency and asymptotic normality of estimating equation estimators when data are independent but not identically distributed is considered. Regularity conditions for consistency and asymptotic normality in the non-iid setting are presented and examples for application of the estimating equation theory to data fusion estimators are provided.

To Maddie for her love and support.

## ACKNOWLEDGEMENTS

Sincerest thank yous to my advisors Michael Hudgens and Bonnie Shook-Sa for their persistence, causing my development as a thinker, writer, and researcher. The learning opportunities were priceless. I thank Jess Edwards, Haibo Zhou, and Stephen Cole for their service on my doctoral committee and for their thoughtful suggestions and insights that improved the research.

The research in Chapter 2 was co-authored with Bonnie Shook-Sa, Stephen Cole, and Michael Hudgens and published in *Journal of the Royal Statistical Society, Series A, Statistics in Society*. I thank Harm van Bakel, Juan Manuel Carreno, Frans Cuevas, Florian Krammer, and Viviana Simon for sharing their data and Dirk Dittmer and the ScreenNC team for sharing their data. The research was supported in part by the National Institute of Health (NIH R01AI085073), the UNC Chapel Hill Center for AIDS Research (UNC-CH CFAR P30AI050410 and R01AI157758), and the National Science Foundation Graduate Research Fellowship Program (NSF GRFP).

My studies at UNC were supported by a NIH training grant in Biostatistics and Mental Health Neuroimaging and Genomics (T32 MH106440-01), the NSF GRFP, and in part through Developmental funding from the University of North Carolina at Chapel Hill Center for AIDS Research (CFAR), an NIH funded program P30AI050410. I would like to acknowledge Jason Fine and John Gilmore who provided me with the T-32 and Bonnie Shook-Sa who provided me with the UNC-CFAR GRA. Thank you to my BIOS colleagues and friends and the administrative staff for encouragement and kindness. Last but not least, I thank my family, especially Mom, Dad, and Dina, who have done so much for me.

## TABLE OF CONTENTS

## LIST OF TABLES

## LIST OF FIGURES

# CHAPTER 1: LITERATURE REVIEW

## 1.1 COVID-19 pandemic

According to World Health Organization (WHO) estimates, the COVID-19 pandemic has caused more than 14 million deaths. (World Health Organization, 2022). This pandemic is the deadliest since the influenza pandemic of 1918, a catastrophe for global health that killed over 50 million, and perhaps as many as 100 million, people worldwide (Johnson and Mueller, 2002). Over the past half-century, spurred on greatly by reports of AIDS in 1978 and discovery of its viral cause HIV, it has become clear to public health professionals that infectious diseases pose a major threat to humans (Weber et al., 2016; Wilson, 1995). Most new and emerging infectious disease threats are zoonotic, i.e., transmitting from animals, such as HIV, SARS-CoV-2, SARS, and Ebola (Weber et al., 2016). Clinically, COVID-19 disease can be severe (Cevik et al., 2020). The most common symptoms are fever, cough, fatigue, and muscle pain, and the disease can result in acute respiratory distress syndrome and death. (Cevik et al., 2020).

Two important public health challenges during the pandemic have been the study of seroprevalence and the determination of vaccine correlates of protection. Seroprevalence, the proportion of individuals in a population with antibodies to SARS-CoV-2, is a useful quantity for tracking the pandemic's severity and informing public health decisions. This proportion can be combined with further assumptions and data to estimate the attack rate, cumulative incidence rate, and infection fatality rate of SARS-CoV-2, among other quantities of interest (Takahashi et al., 2023; Shioda et al., 2021; Buss et al., 2021; Perez-Saez et al., 2021; Brazeau et al., 2020). Seroprevalence estimates may also be useful in modeling the proportion of people who have reduced susceptibility to SARS-CoV-2 insofar as antibody titers are associated with protection from COVID-19 (Earle et al., 2021; Khoury et al., 2021). Seroprevalence is often estimated by survey research, but such studies are commonly subject to misclassification bias due to false positives and negatives (Bouman et al., 2021) and

selection bias due to the use of non-probability sampling techniques (Accorsi et al., 2021). These challenges are addressed in Chapter 2.

About one year after SARS-CoV-2 was first discovered in humans, the Pfizer-BioNTech and Moderna mRNA-based COVID-19 vaccines were successfully developed and authorized for emergency use by the U.S. Food and Drug Administration (Office of the Commissioner, 2020b,a). The speed of vaccine development was unprecedented, representing a tremendous scientific accomplishment (Krammer, 2020), and these and other vaccines have shown high efficacy against COVID-19 disease (Lipsitch and Kahn, 2021). The Phase 3 trials that supported efficacy required thousands of volunteers in addition to great effort and time. As SARS-CoV-2 mutates and new strains develop, there has been substantial interest in determining an immune biomarker produced by the vaccines that correlates significantly with the efficacy each vaccine demonstrated in the Phase 3 trials; particular attention has been given to the roles of neutralizing and binding antibodies (Krammer, 2021; Khoury et al., 2021; Earle et al., 2021; Gilbert et al., 2022b; Addetia et al., 2020). However, the vaccine efficacy trials also differed in terms of their study populations, timing (and thus which strains of the virus were circulating), and definitions of endpoints, among other factors. As such, it is difficult to compare the point estimates of vaccine efficacy without context (Rapaka et al., 2022). Methods to conduct vaccine correlate of protection analyses while adjusting for selection bias across trials are considered in Chapter 3.

## 1.2 Standardization and the parametric g-formula

In each paper, the generalization of estimates to target populations is considered; in particular, the use of direct standardization (Neison, 1844) and the parametric g-formula (Robins, 1986). These methods adjust for structural biases such as selection bias, confounding, and effect modification (Hernán et al., 2004). In the first and second papers, proportions and causally interpreted parameters are standardized using either direct standardization or the parametric g-formula. Consider measurement of (vectors of) individual outcome data $X$ and covariate data $Z$. Without loss of generality, the covariates can be considered discrete and taking on one of $k$ distinct realized strata $z_1, \ldots, z_k$. The

quantities $\rho_j = \mathbb{E}[X \mid Z = z_j]$ for $j = 1, \ldots, k$ are estimated from the sample data. Presume that the distribution of $Z$ is known in the target population, i.e., $\Pr(Z = z_j)$ is known for $j = 1, \ldots, k$. The standardized prevalence in (or generalized to) the target population is $\rho = \sum_{j=1}^{k} \rho_j \Pr(Z = z_j)$. The procedure applies to continuous covariates $Z$ by integrating over the distribution of $Z$ rather than summing over its discrete levels.

The parametric g-formula, introduced by Robins (1986), generalized standardization to the time-varying exposure setting in causal inference; for the purposes of this thesis, consider the g-formula in the time-fixed setting. Let $Y^a$ be the potential outcome if, possibly counter to fact, exposure (or treatment) were $A = a$. Standard assumptions for causal inference must be made to draw causal conclusions with the g-formula. These assumptions are conditional exchangeability $Y^a \perp\!\!\!\perp A \mid Z$, causal consistency $Y = Y^a I(A = a)$, positivity $f_{A|Z}(a \mid Z) > 0$ with probability 1 where $f_{A|Z}(a \mid Z)$ is the conditional probability density function of $A \mid Z$, no interference between units (Hudgens and Halloran, 2008), and no measurement error. If these assumptions hold, then the causal quantities of interest can be identified from the data: $\mathbb{E}[Y^a] = \int_Z \mathbb{E}[Y \mid A = a, Z = Z]\, dF_Z(z)$ where $F_Z$ is the distribution function of $Z$.

In the causal inference setting, the parametric g-formula requires the fitting of an outcome model for $Y$ conditional on exposure $A$ and covariates $Z$. Causal effects can also be consistently estimated using inverse probability weighting, which involves fitting a propensity model to estimate $\mathbb{E}[A \mid Z = Z]$, the conditional probability of being exposed given covariate values. Outcome and propensity models can be combined in augmented inverse probability weighted estimators, which feature the 'double robustness' property. The double robustness property assures the estimator is consistent and asymptotically normal under regularity conditions if at least one of, but not necessarily both of, the outcome or propensity models is correctly specified (Daniel, 2018).

### 1.3 Seroprevalence estimation

In Chapter 2, methods for estimating seroprevalence of SARS-CoV-2 are considered. Consider a seroprevalence study as arising from random sampling of a target population. Estimating

seroprevalence can then be seen as the general problem of estimating the prevalence $\pi \in [0, 1]$ of any population characteristic with a measurement instrument (assay, diagnostic test, or survey) that may potentially have error. Such measurement error can be applied to diagnostic testing (Gastwirth, 1987), survey research (Dean and Pagano, 2015; Hemenway, 1997), etc. The error can have significant effects on prevalence estimates.

Notationally, there exists some unobserved random variable $Y$ at the individual (or unit) level which is distributed Bernoulli with expectation $\pi$. Two frequently-used measures of a test's efficiency are its sensitivity $\sigma_e = \Pr(X = 1 \mid Y = 1)$, the probability that someone who truly has the condition tests positive, and specificity $\sigma_p = \Pr(X = 0 \mid Y = 0)$, the probability that someone who truly does not have the condition tests negative (Wilson and Jungner, 1968). The observed random variable is the test result $X \sim \text{Bernoulli}(\rho)$ where $\rho = \pi\sigma_e + (1 - \pi)(1 - \sigma_p)$ by the law of total probability. Sensitivity and specificity suffice for the population health aims of Chapter 2, but other quantities such as the positive predictive value $\Pr(Y = 1 \mid X = 1)$ and negative predictive value $\Pr(Y = 0 \mid X = 0)$ often have clinical relevance.

In many cases, a main study and two validation studies are conducted. (1) Measure sensitivity by sampling $n_1$ individuals from strata of the population having the characteristic, where 'having the characteristic' is measured by a gold standard; (2) Measure specificity from a sample of $n_2$ individuals from strata of the population not having the characteristic; (3) Sample $n_3$ individuals from the general population of interest. The samples are drawn from populations large enough such that the probability of an individual being selected in more than one sample is negligible and thus ignored.

### 1.3.1 Frequentist methods

An empirical estimator that plugs in sample proportions for their population parameters is a simple prevalence estimator. Denote the sample proportions of positives in the three sample as $\hat{\sigma}_e$, $1 - \hat{\sigma}_p$, and $\hat{\rho}$. The plug-in estimator $\hat{\pi}_{RG} = \{\hat{\rho} - (1 - \hat{\sigma}_p)\}/\{\hat{\sigma}_e - (1 - \hat{\sigma}_p)\}$ results from algebraic manipulation of the definition for $\rho$. The estimator $\hat{\pi}_{RG}$ is usually attributed to Rogan and Gladen

(1978) (see also Gart and Buck, 1966; Marchevsky, 1979) and is widely used in biomedical and epidemiologic applications (Messam et al., 2008; Greiner and Gardner, 2000; Manuel et al., 2010; Maxim et al., 2014; Hens et al., 2012; Shoukri, 2003). Rogan and Gladen (1978) considered an estimator of the variance of $\hat{\pi}_{RG}$ derived from the delta method.

The simplicity of the Rogan-Gladen estimator makes it accessible for related work; for instance, Lang and Reiczigel (2014) propose improvements to variance estimation using the 'add two pseudo-positives and -negatives' idea of Agresti and Coull (1998). That said, a potential drawback to the Rogan-Gladen estimator is that it can, in rare cases, fall outside of the range $[0, 1]$. This occurs in the unlucky circumstances that $\hat{\rho} < 1 - \hat{\sigma}_p$, $\hat{\sigma}_e < 1 - \hat{\sigma}_p$, or $\hat{\rho} > \hat{\sigma}_e$ (Hilden, 1979). The estimator would be negative if $\sigma_e < 1 - \sigma_p$, but testing 'positive' under such a process would mean that one is more likely to be truly negative than positive, so some authors have considered such a process to not be a "test" (Rogan and Gladen, 1978). The variance estimator is also justified by large-sample approximation, which may not be appropriate in finite samples. Thus some statisticians have advocated for Bayesian approaches. Frequentist developments arising from the COVID-19 pandemic are discussed in Chapter 2.

### 1.3.2 Bayesian methods

Bayesian methods for estimating prevalence account for uncertainty somewhat more naturally through the use of prior distributions, but require the analyst to select the prior and also typically demand more computational resources or time. A typical Bayesian approach is a beta-binomial model, where $X \mid \{\rho, \sigma_e, \sigma_p\} \sim \text{Bin}(n_3, \rho)$ and beta prior distributions are placed on $\rho, \sigma_e$, and $\sigma_p$ (Messam et al., 2008; Gelman and Carpenter, 2020). A Bayesian framework may be particularly helpful for epidemiologic and infectious disease modeling. Seroprevalence estimates can parameterize models for key quantities such as the infection fatality rate, and the Bayesian approach allows uncertainty in both the seroprevalence estimate and the underlying sensitivity and specificity estimates to flow through these calculations (Larremore et al., 2021).

### 1.3.3 Selection bias in seroprevalence studies

Non-probability sampling can lead to biased estimates of population quantities. Perhaps the most infamous example from survey research is that of the *Literary Digest* straw poll to forecast the 1936 presidential election. The poll incorrectly predicted that 60% of the vote would go to Alf Landon over incumbent Franklin Roosevelt because of a combination of non-response and sampling biases (Bryson, 1976). These biases and other survey sampling biases can be recast as examples of selection bias. Specifically, observed and unobserved factors may affect both selection (whether individuals enter the survey) and responses (Winship and Mare, 1992).

Ideally, seroprevalence studies would be performed using random sampling. However, non-probability samples such as convenience samples are often used. Recruitment may happen more quickly and for less cost with convenience samples, but without appropriate adjustment they lead to estimates that often suffer from selection bias (Shook-Sa et al., 2020). Researchers have recommended that surveys follow best practices of using representative sampling frames (e.g., address-based sampling) (Shook-Sa et al., 2020); when this is infeasible or fails to occur, post-hoc adjustment of results is possible if covariates predictive of both survey participation and the outcome are collected (Groves, 2006; Accorsi et al., 2021).

Standardized seroprevalence estimates for SARS-CoV-2 have been made by combining the misclassification bias adjustment of Rogan and Gladen (1978) with direct standardization (Havers et al., 2020; Barzin et al., 2020; Cai et al., 2022). This convenience sample based estimator makes the assumption (either explicitly or implicitly) that each person in a covariate stratum is equally likely to be in the sample, where the covariates defining the covariate strata are chosen by the subject matter experts and analysts (Elliott and Valliant, 2017). The large-sample properties of this nonparametric estimator, and a parametric variation on this estimator, are examined in detail in Chapter 2.

## 1.4 Correlates of protection and surrogate endpoints

In Chapter 3, problems arising in studies of correlates of protection for SARS-CoV-2 are considered. The U.S. Food and Drug Administration typically requires that a large-scale, randomized Phase 3 trial with thousands of participants be conducted to demonstrate efficacy. Such trials are time-consuming, expensive, and demand great effort. Moreover, the different viral strains of the novel coronavirus may necessitate updates to the vaccine each year. Influenza vaccines are updated annually based on the predominant strains in circulation (Cox, 2013). However, conducting expensive Phase 3 trials annually for COVID vaccines may be untenable.

Enter the correlate of protection (CoP). A CoP is an "immune response that is responsible for and statistically interrelated with protection" (Plotkin, 2010), where protection may be vaccine efficacy or another effect measure. The influenza vaccine is a successful example, where a hemagglutinin inhibition antibody titer of 1:40 is a CoP that is used for annual licensure of the updated vaccines (Cox, 2013). Plotkin (2020) reviews CoPs (also known as Correlated Immune Markers) against 20 diseases. These CoPs range from IgG antibodies against the protective antigen to anthrax, to pertussis toxin antibodies, to T cell immune responses against zoster (the disease also known as shingles).

Early in the pandemic, scientists were substantially interested in determining a correlate of protection (Krammer, 2021; Openshaw, 2022). An immunological marker, i.e., antibody level, can be measured in laboratory settings, and its determination leads to fewer trial participants and less cost than Phase 3 placebo-controlled efficacy trials. There is evidence from multiple Phase 3 trials that neutralizing antibody titer are correlates of protection for SARS-CoV-2, i.e., surrogate endpoints for vaccine efficacy against symptomatic COVID-19 illness. (Gilbert et al., 2022a; Cromer et al., 2022; Goldblatt et al., 2022). The FDA and European Medicines Agency recommended that the approval of new vaccine strains and booster doses for SARS-CoV-2 vaccines can be based in part on these correlates (mentioned in Gilbert et al., 2022a).

The analyses of Earle et al. (2021) and Khoury et al. (2021) are exemplars for the methodology of Chapter 3 and are examples of meta-analytic approaches to surrogate outcome evaluation (Joffe

and Greene, 2009). Prentice (1989) gave criteria for surrogacy when one trial is under consideration. In addition to the single-trial approach, surrogates can be validated with a meta-analysis of trials (discussed in Albert et al., 1998). Daniels and Hughes (1997), Buyse et al. (2000), and Gail et al. (2000) developed methods for the meta-analytic setting, where surrogates are assessed depending on the association between the treatment effect on the clinical endpoint and the treatment effect on the surrogate endpoint. Under appropriate conditions such as randomization of the trials, the association can be termed a "causal association" (Joffe and Greene, 2009). There are several frameworks for surrogate evaluation, including and not limited to principal surrogacy (Frangakis and Rubin, 2002; Gilbert et al., 2008). The meta-analytic model is a component of greater efforts to assess surrogates, detailed in the U.S. government COVID-19 response team Statistical Analysis Plan (USG COVID-19 Response Team / Coronavirus Prevention Network (CoVPN) Biostatistics Team et al., 2022).

### 1.5 Estimating equation estimators and data fusion

A data fusion study design "combines data from different sources to answer a question that could not be answered (as well) by data from subsets of these sources" (Cole et al., 2023). Fusion estimators estimate quantities of interest from fusion designs such as correcting for measurement error with auxiliary data (Rogan and Gladen, 1978), as in Chapters 2 and 4; bridged treatment effects that compare effects across different trials (Bareinboim and Pearl, 2016; Breskin et al., 2021; Shook-Sa et al., 2023), as in Chapter 3; and generalizability studies that use auxiliary covariate information from a target population (Cole and Stuart, 2010), as in all the papers of this dissertation.

Data fusion estimators are sometimes estimating equation estimators (EE estimators), also called M-estimators. M-estimation is typically attributed to Huber (1964) (see also Godambe, 1960; Huber, 1967). Huber (1964) considered independent and identically distributed (iid) data vectors $Y_1, \ldots, Y_n$ with common distribution function $F(y - \theta)$. Huber studied estimation of a parameter $\theta$ with a statistic $T = T_n(Y_1, \ldots, Y_n)$ that minimizes the quantity $\sum_i \rho(Y_i - T)$, where $\rho(Y_i, \theta)$ is a non-constant function. Huber called the maximum-likelihood-type estimator $T$ an M-estimator, as

it minimized an estimating equation analogously to how an MLE maximizes a likelihood. Huber (1967) proved the consistency and asymptotic normality of M-estimators with iid data. An EE estimator $\hat{\theta}$ is the solution to the vector equation $\sum_{i=1}^{n} \psi(Y_i, \theta) = 0$ (see primer by Stefanski and Boos, 2002). That is, the EE estimator $\hat{\theta}$ satisfies $\sum_{i=1}^{n} \psi(Y_i, \hat{\theta}) = 0$ where $Y_1, \ldots, Y_n$ are assumed to be independent but not necessarily identically distributed, and $\psi$ is a known function that does not depend on $i$ or $n$. The estimator $\hat{\theta}$ is usually called an M-estimator, as in Stefanski and Boos (2002), although Huber originally defined M-estimators as maximizing an EE (discussed in van der Vaart 1998, Chapter 5.1).

In the fusion setting, often data are independent across different sources but are not identically distributed, i.e., are not iid, because the distributions of important factors differ across data sources. Inagaki (1973) extended Huber's theorem of consistency and asymptotic normality of M-estimators for independent but not identically distributed data under regularity conditions for the estimating equations. For the non-identically distributed case, Yuan and Jennrich (1998) developed regularity conditions that may be more easily checked than those of Inagaki (1973). Under the conditions of Yuan and Jennrich (1998), EE estimators in the independent but not identically distributed setting are consistent and asymptotically normal, i.e., $\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} \text{Normal}(0, V(\theta_0))$ with $V(\theta_0) = A(\theta_0)^{-1} B(\theta_0) \left[A(\theta_0)^{-1}\right]^T$ as $n \to \infty$, where $\theta_0$ is the true value of $\theta$, $A(\theta) = E[\partial\psi/\partial\theta]$, and $B(\theta) = E[\psi\psi^T]$. The empirical sandwich variance estimator substitutes the 'bread' $A(\theta)$ and 'filling' $B(\theta)$ with their empirical counterparts.

The asymptotic sandwich variance $A(\theta)^{-1} B(\theta) \left[A(\theta)^{-1}\right]^T$ can sometimes be derived in closed form. However, the necessary derivations "may involve tedious and error-prone derivative and matrix calculations" (Saul and Hudgens, 2020). Software was created in R (Saul and Hudgens, 2020) and in Python (Zivich et al., 2022b) whereby analysts enter their vector of estimating equations $\psi$. The software then computes the point estimator $\hat{\theta}$ and the empirical sandwich variance estimator, i.e., computes the $A(\theta)$ and $B(\theta)$ matrices.

Though the independent and not identically distributed setting of Inagaki (1973), among others, could be considered a generalization of the iid setting of Huber (1967), the iid setting is assumed in

several standard references (e.g., Bang and Robins, 2005; Boos and Stefanski, 2013; Carroll et al., 2006; Lunceford and Davidian, 2004; Huber and Ronchetti, 2009). When references for iid setups such as these are cited for the asymptotic normality of fusion EE estimators in non-iid setups, the asymptotic distributions of the estimators could be called into question. In Chapter 4, clarity is provided regarding the consistency and asymptotic normality of data fusion estimators for non-iid data. In particular, the conditions of Yuan and Jennrich (1998) are described. Examples of fusion estimators claimed CAN in the literature are proven CAN in Chapter 4 via application of the Yuan and Jennrich (1998) regularity conditions.

# CHAPTER 2: ESTIMATING SARS-CoV-2 SEROPREVALENCE

## 2.1 Introduction

Estimating the proportion of people who have antibodies to severe acute respiratory syndrome coronavirus (SARS-CoV-2) is useful for tracking the pandemic's severity and informing public health decisions (Arora et al., 2021). Individuals may have detectable antibodies for different reasons, including prior infection or vaccination. Antibody levels within a person are dynamic, typically increasing after an infection or vaccination, and then eventually decreasing (waning) over time. Thus individuals may not have detectable antibodies if never (or very recently) infected or vaccinated, or if their antibody levels have waned below the limit of detection of the assay being employed. To the extent that antibody levels are associated with protection from infection with SARS-CoV-2 or COVID-19 disease (Earle et al., 2021; Khoury et al., 2021), seroprevalence estimates may be helpful in modeling the fraction of a population which may be immune or less susceptible to COVID-19. Likewise, cross-sectional seroprevalence estimates, combined with certain modeling assumptions and other data, may permit inference about other parameters such as the cumulative incidence of previous SARS-CoV-2 infection, infection fatality rate, or attack rate (Takahashi et al., 2023; Shioda et al., 2021; Buss et al., 2021; Perez-Saez et al., 2021; Brazeau et al., 2020).

Unfortunately, seroprevalence studies often suffer from at least two sources of bias: measurement error due to false positives and negatives, and selection bias due to non-probability sampling designs. Typically, blood tests for antibodies result in a continuous measure of a particular antibody response, such as that of immunoglobulin G, M, or A (IgG, IgM, or IgA). Dichotomizing antibody responses using a cut-off value almost always produces misclassification bias in the form of false positives and false negatives (Bouman et al., 2021). The following example from Sempos and Tian (2021) demonstrates how this measurement error can lead to biased seroprevalence estimates.

Suppose 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, with 99% specificity. Despite this assay's high sensitivity and specificity, it is straightforward to show that naively using the sample proportion of positive test results as a seroprevalence estimator would, in expectation, lead to a seroprevalence estimate of nearly 2% rather than 1%. To account for measurement error, sensitivity and specificity can be estimated and incorporated into the seroprevalence estimator, e.g., using the method popularized by Rogan and Gladen (1978) (see also Levy and Kass (1970); Marchevsky (1979)).

Many seroprevalence studies are conducted by non-probability sampling methods, which may lead to selection bias when characteristics that drive participation in the study are also risk factors for SARS-CoV-2 infection. Probability-based sampling studies are ideal because they are representative by design and often lead to less biased estimates than non-probability samples with post-hoc statistical adjustments (Shook-Sa et al., 2020; Accorsi et al., 2021). However, probability-based sampling may not always be feasible due to time and cost constraints. For this reason, seroprevalence studies often utilize convenience sampling by, for example, drawing blood samples from routine clinic visitors (e.g., Barzin et al., 2020; Stadlbauer et al., 2021) or using residual sera from blood donors (e.g., Uyoga et al., 2021) or commercial laboratories (e.g., Bajema et al., 2021). Convenience sample-based estimators often assume that each person in a covariate-defined stratum has an equal probability of being in the sample (Elliott and Valliant, 2017). Under this assumption, population seroprevalence of SARS-CoV-2 can be estimated with direct standardization (Havers et al., 2020; Barzin et al., 2020; Cai et al., 2022), though weighting methods such as calibration can be used (e.g., Bajema et al., 2021).

In this paper, methods are considered which combine standardization and the Rogan–Gladen adjustment to account for both measurement error and selection bias. The article is organized as follows. Section 2.2 reviews prevalence estimation under measurement error. Non-parametric and parametric standardized prevalence estimators and their large-sample properties are described in Section 2.3. Section 2.4 presents simulation studies to evaluate the empirical bias and 95% confi-

dence interval (CI) coverage of the standardized estimators across a range of assay characteristics and bias scenarios. The methods are then applied in Section 2.5 to three studies that estimate seroprevalence of SARS-CoV-2 in 2020 among all residents of New York City (NYC), all residents of Belgium, and asymptomatic residents of North Carolina. Section 2.6 concludes with a discussion. Proofs are in Appendix A.

## 2.2 Seroprevalence estimation under measurement error

### 2.2.1 Problem set-up

Let the true serology status for an individual in the target population be denoted by $Y$, with $Y = 1$ if the individual has antibodies against SARS-CoV-2 and $Y = 0$ otherwise. Our goal is to draw inference about the population seroprevalence $\pi = P(Y = 1)$. Because of error in the serology assay, $Y$ is not observed directly. Let the result of the serology assay be denoted by $X$, with $X = 1$ if the individual tests positive (according to the antibody assay used) and $X = 0$ otherwise. Three key quantities are sensitivity, the probability that a true positive tests positive, denoted by $\sigma_e = P(X = 1 \mid Y = 1)$; specificity, the probability that a true negative tests negative, denoted by $\sigma_p = P(X = 0 \mid Y = 0)$; and the population expectation of the serology assay outcome, denoted by $\rho = \mathbb{E}(X) = P(X = 1)$. Unless the assay has perfect sensitivity and specificity with $\sigma_e = \sigma_p = 1$, $\rho$ typically will not equal $\pi$ and $X$ will be a misclassified version of $Y$.

The sensitivity and specificity of a diagnostic test are commonly estimated by performing the assay on 'validation' samples of known true positives and true negatives, respectively. Specifically, measurements are taken on $n_1$ independent and identically distributed (iid) units from strata of the population where $Y = 1$ and on $n_2$ iid units from strata where $Y = 0$. Thus $n_1$ copies of $X$ are observed to estimate sensitivity and $n_2$ copies of $X$ are observed to estimate specificity. In the COVID-19 setting, samples from patients who had a case confirmed with reverse transcription polymerase chain reaction (PCR) testing are often assumed to be true positives. Remnant blood samples that were drawn in 2019 or earlier are often assumed to be true negatives. To estimate

seroprevalence in a target population, a 'main' study with $n_3$ iid copies of $X$ is then conducted, among which true infection status is unknown.

Assume, as is realistic in many SARS-CoV-2 studies, that there is no overlap between the units in each of the three studies. Let $\delta_i$ be an indicator of which study the $i$th individual's sample $X_i$ is from, with $\delta_i = 1$ for the sensitivity study, $\delta_i = 2$ for the specificity study, and $\delta_i = 3$ for the main study. Note that $\sum I(\delta_i = j) = n_j$ for $j = 1, 2, 3$, where $n = n_1 + n_2 + n_3$ and here and throughout summations are taken from $i = 1$ to $n$ unless otherwise specified. Assume $n_j/n \to c_j \in (0, 1)$ as $n \to \infty$.

### 2.2.2 Estimators and statistical properties

Let $\theta = (\sigma_e, \sigma_p, \rho, \pi)^T$. Consider the estimator $\hat{\theta} = (\hat{\sigma}_e, \hat{\sigma}_p, \hat{\rho}, \hat{\pi}_{RG})^T$, where $\hat{\sigma}_e = n_1^{-1} \sum I(\delta_i = 1)X_i$, $\hat{\sigma}_p = n_2^{-1} \sum I(\delta_i = 2)(1-X_i)$, $\hat{\rho} = n_3^{-1} \sum I(\delta_i = 3)X_i$, and $\hat{\pi}_{RG} = (\hat{\rho}+\hat{\sigma}_p-1)/(\hat{\sigma}_e+\hat{\sigma}_p-1)$. The prevalence estimator $\hat{\pi}_{RG}$ is motivated by rearranging the identity that $\rho = \pi\sigma_e + (1-\pi)(1-\sigma_p)$ and is sometimes referred to as the Rogan–Gladen (1978) estimator. Note the sample proportions $\hat{\sigma}_e$, $\hat{\sigma}_p$, and $\hat{\rho}$ are maximum likelihood estimators (MLEs) for $\sigma_e$, $\sigma_p$, and $\rho$, respectively, so $\hat{\pi}_{RG}$ is a function of the MLE of $(\sigma_e, \sigma_p, \rho)$ (see Appendix A.1 for details).

The estimator $\hat{\theta}$ can be expressed as the solution (for $\theta$) to the estimating equation vector

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

where here and below 0 denotes a column vector of zeros. Since the samples were selected from three different populations, the data $X_1, \ldots, X_n$ are not identically distributed and care must be taken to derive the large-sample properties of $\hat{\theta}$. In Appendix A.2, the estimator $\hat{\theta}$ is shown to be consistent and asymptotically normal. Specifically, as $n \to \infty$, $\sqrt{n}(\hat{\theta}-\theta) \to_d \mathcal{N}\left(0, \mathbb{A}(\theta)^{-1}\mathbb{B}(\theta)\mathbb{A}(\theta)^{-T}\right)$ and $\sqrt{n}(\hat{\pi}-\pi) \to_d \mathcal{N}\left(0, V_{\pi,RG}\right)$ assuming $\sigma_e > 1-\sigma_p$ (as discussed below), where $\mathbb{A}(\theta)^{-1}\mathbb{B}(\theta)\mathbb{A}(\theta)^{-T}$

is a covariance matrix with bottom right element

$$V_{\pi,RG} = \left\{ \frac{\pi^2 \sigma_e(1-\sigma_e)}{c_1} + \frac{(1-\pi)^2 \sigma_p(1-\sigma_p)}{c_2} + \frac{\rho(1-\rho)}{c_3} \right\} (\sigma_e + \sigma_p - 1)^{-2}. \quad (2.1)$$

The proof of consistency and asymptotic normality is similar to proofs from standard estimating equation theory (e.g., Boos and Stefanski, 2013, Equation 7.10), but because the data are not identically distributed the Lindeberg-Feller Central Limit Theorem (CLT) is used in place of the classical Lindeberg-Lévy CLT. Note that the asymptotic variance (2.1) consists of three components corresponding to the sensitivity, specificity, and main studies. In some circumstances, investigators may be able to decrease the variance of $\hat{\pi}_{RG}$ by increasing the sample sizes of the sensitivity or specificity studies compared to the main study (Larremore et al., 2020).

Let $\hat{V}_{\pi,RG}$ denote the plug-in estimator defined by replacing $\sigma_e, \sigma_p, \rho, \pi$, and $c_j$ in (2.1) with $\hat{\sigma}_e, \hat{\sigma}_p, \hat{\rho}, \hat{\pi}_{RG}$, and $n_j/n$ for $j = 1, 2, 3$, and note that $\hat{V}_{\pi,RG}/n$ is the variance estimator proposed by Rogan and Gladen (1978). By the continuous mapping theorem, $\hat{V}_{\pi,RG}$ is consistent for the asymptotic variance assuming $\sigma_e > 1 - \sigma_p$ and can be used to construct Wald-type CIs that asymptotically attain nominal coverage probabilities. In finite samples, Wald-type CIs can sometimes have erratic coverage properties when estimating a single binomial parameter (Brown et al., 2001; Dean and Pagano, 2015). In Section 2.4, simulations are conducted to assess the performance of the Wald-type CIs in seroprevalence estimation scenarios. Alternative approaches for constructing CIs are discussed in Section 2.6.

### 2.2.3 Truncation

The Rogan–Gladen estimator $\hat{\pi}_{RG}$ sometimes outside of $[0, 1]$ when (i) $\hat{\sigma}_e < 1 - \hat{\sigma}_p$, (ii) $\hat{\rho} < 1 - \hat{\sigma}_p$, or (iii) $\hat{\rho} > \hat{\sigma}_e$. Indeed, (ii) occurred in the ScreenNC study discussed in Section 2.5.3. Estimates are typically truncated to be inside $[0, 1]$ because the true population prevalence must exist in $[0, 1]$ (Hilden, 1979). In this article, all point estimates and bounds of interval estimates are so truncated. Note, though, that as the three sample sizes grow large the estimator $\hat{\pi}_{RG}$ yields estimates inside $[0, 1]$ almost surely unless $\sigma_e < 1 - \sigma_p$. In practice, settings where $\sigma_e < 1 - \sigma_p$ may be very

unlikely; in such scenarios, the probability of a positive test result is higher for seronegative persons than for seropositive persons, so such a measurement instrument performs worse in expectation than random guessing. Throughout this manuscript, it is assumed that $\sigma_e > 1 - \sigma_p$.

## 2.3 Standardized seroprevalence estimation

### 2.3.1 Problem setup

In some settings it may not be reasonable to assume the $n_3$ copies of $X$ from the main study constitute a random sample from the target population. Suppose instead that for each copy of $X$ a vector of discrete covariates $Z$ is observed, with $Z$ taking on $k$ possible values $z_1, \ldots, z_k$. The covariates $Z$ are of interest because seroprevalence may differ between the strata; for instance, $Z$ might include demographic variables such as age group, race, or gender. Denote the mean of $X$ in the $j$th stratum as $\rho_j = P(X = 1 \mid Z = z_j)$ and the sample size for the $j$th stratum as $n_{z_j} = \sum I(\delta_i = 3, Z_i = z_j)$, so $\sum_{j=1}^{k} n_{z_j} = n_3$.

The distribution of strata in the target population, if known, can be used to standardize estimates so they are reflective of the target population (for a review of direct standardization, see van Belle et al. (2004, Chapter 15)). Denote the proportion of the target population comprised by the $j$th stratum as $\gamma_j = P(Z = z_j)$ and suppose that these stratum proportions are known with each $\gamma_j > 0$ and $\sum_{j=1}^{k} \gamma_j = 1$. The stratum proportions are commonly treated as known based on census data or large probability-based surveys (Lohr, 2010, Ch. 4.4; Korn and Graubard, 1999, Ch. 2.6). Alternatively, $\gamma_1, \ldots, \gamma_k$ could be estimated, e.g., from a random sample of the target population, and the estimator of the seroprevalence estimator's variance could be appropriately adjusted to reflect the uncertainty in these estimated proportions.

Assume that all persons in a covariate stratum defined by $Z$ have the same probability of inclusion in the sample. Then the covariates $Z$ in the main study sample have a multinomial distribution with $k$ categories, sample size $n_3$, and an unknown sampling probability vector $(s_1, \ldots, s_k)^T$ where $\sum_{j=1}^{k} s_j = 1$. For $j = 1, \ldots, k$, the probability $s_j$ indicates the chance of a sampled individual being in stratum $j$. Note that if the main study were a simple random sample from the target

population, then the sampling probabilities would be equal to the stratum proportions (with $s_j = \gamma_j$ for $j = 1, \ldots, k$).

### 2.3.2 Non-parametric standardization

First, consider a seroprevalence estimator which combines non-parametric standardization and the Rogan–Gladen adjustment to account for both selection bias and measurement error. Note that $\rho$ is a weighted average of the stratum-conditional means $\rho_j$, where each weight is a known stratum proportion $\gamma_j$, i.e., $\rho = \sum_{j=1}^{k} \rho_j \gamma_j$. A non-parametric standardization estimator for $\rho$ using the sample stratum-conditional prevalences $\hat{\rho}_j = n_{z_j}^{-1} \sum I(Z_i = z_j, \delta_i = 3) X_i$ for $j = 1, \ldots, k$ is $\hat{\rho}_{SRG} = \sum_{j=1}^{k} \hat{\rho}_j \gamma_j$. A standardized prevalence estimator accounting for measurement error is $\hat{\pi}_{SRG} = (\hat{\rho}_{SRG} + \hat{\sigma}_p - 1)/(\hat{\sigma}_e + \hat{\sigma}_p - 1)$, which has been used in SARS-CoV-2 seroprevalence studies (Havers et al., 2020; Barzin et al., 2020; Cai et al., 2022).

Let $\theta_s = (\sigma_e, \sigma_p, \rho_1, \ldots, \rho_k, \rho, \pi)^T$. The estimator $\hat{\theta}_s = (\hat{\sigma}_e, \hat{\sigma}_p, \hat{\rho}_1, \ldots, \hat{\rho}_k, \hat{\rho}_{SRG}, \hat{\pi}_{SRG})^T$ solves the vector $\sum \psi(X_i, Z_i; \delta_i, \theta_s) = (\sum \psi_e, \sum \psi_p, \sum \boldsymbol{\psi_\rho}, \psi_\rho, \psi_\pi)^T = 0$ of estimating equations, where $\sum \psi_e, \sum \psi_p$, and $\psi_\pi$ are defined in Section 2.2; $\sum \boldsymbol{\psi_\rho}$ is a $k$-vector with $j$th element $\sum \psi_{\rho_j} = \sum I(Z_i = z_j, \delta_i = 3)(X_i - \rho_j)$; and $\psi_\rho = \sum_{j=1}^{k} \rho_j \gamma_j - \rho$. It follows that $\hat{\theta}_s$ is consistent and asymptotically normal and that $\sqrt{n}(\hat{\pi}_{SRG} - \pi) \to_d \mathcal{N}(0, V_{\pi,SRG})$ where

$$V_{\pi,SRG} = \left\{ \frac{\pi^2 \sigma_e (1 - \sigma_e)}{c_1} + \frac{(1 - \pi)^2 \sigma_p (1 - \sigma_p)}{c_2} + \sum_{j=1}^{k} \frac{\gamma_j^2 \rho_j (1 - \rho_j)}{c_3 s_j} \right\} (\sigma_e + \sigma_p - 1)^{-2}. \quad (2.2)$$

The asymptotic variance $V_{\pi,SRG}$ can be consistently estimated by the plug-in estimator $\hat{V}_{\pi,SRG}$ defined by replacing $\sigma_e, \sigma_p, \rho_j, s_j, \pi$, and $c_l$ in (2.2) with $\hat{\sigma}_e, \hat{\sigma}_p, \hat{\rho}_j, n_{z_j}/n_3, \hat{\pi}_{SRG}$, and $n_l/n$ for $j = 1, \ldots, k$ and $l = 1, 2, 3$. Consistency of $\hat{V}_{\pi,SRG}$ holds by continuous mapping, and a proof of asymptotic normality and justification of (2.2) are in Appendix A.3.

Standardization requires estimating the stratum-conditional mean of $X$, $\rho_j = P(X = 1 \mid Z = z_j)$. However, when $n_{z_j} = 0$ for some strata $j$, the corresponding estimator $\hat{\rho}_j$ is undefined, and $\hat{\rho}_{SRG}$ is then undefined as well. Values of $n_{z_j}$ may equal zero for two reasons. First, the study

design may exclude these strata ($s_j = 0$), a situation referred to as deterministic or structural nonpositivity (Westreich and Cole, 2010). Second, even if $s_j > 0$, random nonpositivity can occur if no individuals with $Z = z_j$ are sampled, which may occur if $s_j$ is small or if $n_3$ is relatively small. When nonpositivity arises, an analytical approach often employed entails "restriction" (Westreich and Cole, 2010), where the target population is redefined to consist only of strata $j$ for which $n_{z_j} > 0$. However, this redefined target population may be less relevant from a public health or policy perspective.

### 2.3.3 Parametric standardization

Rather than redefining the target population, an alternative strategy for combatting positivity violations is to fit a parametric model to estimate all stratum-conditional means $\rho_j$. Such parametric models allow inference to the original target population and, when they are correctly specified, typically outperform non-parametric approaches (Petersen et al., 2012; Rudolph et al., 2018; Zivich et al., 2022a). Assume the binary regression model $g(\rho_j) = \beta h(z_j)$ holds, where $g$ is an appropriate link function for a binary outcome like the logit or probit function; $\beta$ is a row vector of $p$ regression coefficients with intercept $\beta_1$; and $h(z_j)$ is a user-specified $p$-vector function of the $j$th stratum's covariate values that may include main effects and interaction terms, with $l$th element denoted $h_l(z_j)$ and $h_1(z_j)$ set equal to one to correspond to an intercept. Let $\text{supp}(z)$ be the covariate support in the sample, i.e., $\text{supp}(z) = \{z_j : n_{z_j} > 0\}$ with dimension $\dim\{\text{supp}(z)\} = \sum_{j=1}^{k} I(n_{z_j} > 0)$, and assume $p \leq \dim\{\text{supp}(z)\} \leq k$. (Note that $\dim\{\text{supp}(z)\} = k$ only when there is positivity, and in that case $\hat{\pi}_{SRG}$ can be used with no restriction needed.)

Under the assumed binary regression model, each $\rho_j$ is a function of the parameters $\beta$ and the covariates $z_j$ that define the $j$th stratum, denoted $\rho_j(\beta, z_j) = g^{-1}\{\beta h(z_j)\}$. A model-based standardized Rogan–Gladen estimator of $\pi$ is $\hat{\pi}_{SRGM} = (\hat{\rho}_{SRGM} + \hat{\sigma}_p - 1)/(\hat{\sigma}_e + \hat{\sigma}_p - 1)$, where $\hat{\rho}_{SRGM} = \sum_{j=1}^{k} \hat{\rho}_j(\hat{\beta}, z_j)\gamma_j$ and $\hat{\beta}$ is the MLE of $\beta$. Estimating equation theory can again be used to derive large-sample properties by replacing the $k$ equations for $\rho_1, \ldots, \rho_k$ from Section 2.3.2 with $p$ equations for $\beta_1, \ldots, \beta_p$ corresponding to the score equations from the binary regression.

Let $\theta_m = (\sigma_e, \sigma_p, \beta_1, \ldots, \beta_p, \rho, \pi)^T$ and $\hat{\theta}_m = (\hat{\sigma}_e, \hat{\sigma}_p, \hat{\beta}_1, \ldots, \hat{\beta}_p, \hat{\rho}_{SRGM}, \hat{\pi}_{SRGM})^T$. The estimator $\hat{\theta}_m$ solves the vector $\sum \psi(X_i, Z_i; \delta_i, \theta_m) = (\sum \psi_e, \sum \psi_p, \sum \psi_\beta, \psi_\rho, \psi_\pi)^T = 0$ of estimating equations, where $\sum \psi_e, \sum \psi_p$, and $\psi_\pi$ are as in Section 2.2; $\sum \psi_\beta$ is a $p$-vector with $j$th element $\sum \psi_{\beta_j} = \sum I(\delta_i = 3)\left[X_i - g^{-1}\left\{\beta h(Z_i)\right\}\right] h_j(Z_i)$; and $\psi_\rho = \sum_{j=1}^{k} g^{-1}\{\beta h(Z_j)\}\gamma_j - \rho$. It follows that $\hat{\theta}_m$ is consistent and asymptotically normal and $\sqrt{n}(\hat{\pi}_{SRGM} - \pi) \rightarrow_d \mathcal{N}(0, V_{\pi,SRGM})$. The asymptotic variance $V_{\pi,SRGM}$ can be consistently estimated by $\hat{V}_{\pi,SRGM}$, the lower right element of the empirical sandwich variance estimator of the asymptotic variance of $\hat{\theta}_m$. A proof of asymptotic normality and the empirical sandwich variance estimator are given in Appendix D. An R package for computing $\hat{\pi}_{SRG}$, $\hat{\pi}_{SRGM}$, and their corresponding variance estimators is available at `https://github.com/samrosin/rgStandardized`.

### 2.4 Simulation study

Simulation studies were conducted to compare $\hat{\pi}_{RG}$, $\hat{\pi}_{SRG}$, and $\hat{\pi}_{SRGM}$. Four data generating processes (DGPs) were considered, within which different scenarios were defined through full factorial designs that varied simulation parameters $\pi, \sigma_e, \sigma_p, n_1, n_2$, and $n_3$. These DGPs featured no selection bias (DGP 1), selection bias with two strata (DGP 2), and more realistic selection bias with 40 strata and 80 strata (DGPs 3 and 4).

For each DGP and set of simulation parameters, sensitivity and specificity validation samples of size $n_1$ and $n_2$ were generated with $X$ distributed Bernoulli with a mean of $\sigma_e$ or $1 - \sigma_p$, respectively. In DGPs 1 and 2, a main study of size $n_3$ was then generated where $Y$ was Bernoulli with mean $\pi$ and $X \mid Y$ was Bernoulli with mean $\sigma_e Y + (1 - \sigma_p)(1 - Y)$; in DGPs 3 and 4, $X$ was generated from the distribution of $X \mid Z$, as described below. Simulation parameter values were selected based on the seroprevalence studies described in Section 2.5. Sensitivity was varied in $\sigma_e \in \{.8, .99\}$, specificity in $\sigma_p \in \{.8, .95, .99\}$, and prevalence in $\pi \in \{.01, .02, \ldots, .20\}$. Sample sizes were $n_1 = 40$, $n_2 = 250$, and $n_3 = 2500$. The full factorial design led to 120 scenarios per DGP, and within each scenario 1,000 simulations were conducted unless otherwise specified. Performance was measured by: (a) mean bias, computed as the mean of $\hat{\pi} - \pi$ for

each estimator $\hat{\pi}$; (b) empirical coverage, i.e., whether the 95% Wald-type CIs based on each variance estimator $\hat{V}_\pi$ contained the true prevalence; (c) mean squared error (MSE), computed as the mean of $(\hat{\pi} - \pi)^2$ for each estimator $\hat{\pi}$. R code implementing the simulations is available at `https://github.com/samrosin/rgStandardized_ms`.

### 2.4.1 No selection bias

For DGP 1, 10,000 simulations were conducted to assess the performance of $\hat{\pi}_{RG}$ when no selection bias was present. The estimator $\hat{\pi}_{RG}$ was generally unbiased, as seen in Figure A.1. Performance improved as $\sigma_e$ and $\sigma_p$ tended toward 1, with $\sigma_p$ being a stronger determinant of bias. An exception to these results occurred when $\pi \leq 0.05$ and $\sigma_p \leq 0.95$, in which case $\hat{\pi}_{RG}$ overestimated the true prevalence. The Rogan–Gladen estimator without truncation was also evaluated in this DGP to determine if truncation caused the bias. While the non-truncated estimator was slightly biased, the magnitude of the bias was less than 0.002 in all scenarios, suggesting the bias of $\hat{\pi}_{RG}$ in low prevalence, low specificity settings is due largely to truncation.

Wald CIs based on $\hat{V}_{\pi,RG}$ attained nominal coverage in almost every scenario, as seen in Figure A.2. However, when some parameters were near their boundaries, coverage did not reach the nominal level. For instance, when $\pi$ was 0.01 and $\sigma_p$ was 0.99, 95% CIs covered in 90% and 91% of simulations for two values of $\sigma_e$. These variable CI coverage results concord with previous simulation studies evaluating $\hat{V}_{\pi,RG}$ (Lang and Reiczigel, 2014). The MSE of $\hat{\pi}_{RG}$, shown in Figure A.3, tended to increase with $\pi$ and decrease as $\sigma_e$ and $\sigma_p$ approached 1.

### 2.4.2 Low-dimensional selection bias

In DGP 2, the target population was comprised of two strata defined by a covariate $Z \in \{z_1, z_2\}$ with proportions $\gamma_1 = \gamma_2 = .5$. Within the main study, $Z$ was generated from a binomial distribution of sample size $n_3$ and sampling probabilities $(.2, .8)$. Individuals' serostatuses were generated from the conditional distribution $Y \mid Z$, which was such that $P(Y = 1 \mid Z = z_1) = 1.5\pi$ and

$P(Y = 1 \mid Z = z_2) = 0.5\pi$ for each value of $\pi$. In each simulation $\hat{\pi}_{RG}$ and $\hat{\pi}_{SRG}$ and their corresponding 95% CIs were computed.

The non-parametric standardized estimator $\hat{\pi}_{SRG}$ was empirically unbiased for true prevalences $\pi \geq 0.05$, as seen in Figure A.4, and 95% CIs based on $\hat{V}_{\pi,SRG}$ attained nominal coverage in almost every scenario, as seen in Figure A.5. As with $\hat{\pi}_{RG}$ in DGP 1, CI coverage for $\hat{\pi}_{SRG}$ was slightly less than the nominal level for very low $\pi$ and for $\sigma_p$ near the boundary, e.g., coverage was 91% for $\pi = .01$ and $\sigma_e = \sigma_p = .99$. MSE trends for $\hat{\pi}_{SRG}$ were similar to those of $\hat{\pi}_{RG}$ in DGP 1, as seen in Figure A.6. Figures A.4 through A.6 show that $\hat{\pi}_{RG}$ performed poorly under selection bias, with large negative bias, CI coverage far less than the nominal level in most cases, and much greater MSE than $\hat{\pi}_{SRG}$.

### 2.4.3 More realistic selection bias

#### 2.4.3.1 DGP 3

DGPs 3 and 4 compared $\hat{\pi}_{SRG}$ and $\hat{\pi}_{SRGM}$ in scenarios with larger numbers of strata. In DGP 3, three covariates were defined as $Z_1 \in \{z_{10}, z_{11}\}$, $Z_2 \in \{z_{20}, z_{21}, z_{22}, z_{23}\}$, and $Z_3 \in \{z_{30}, z_{31}, z_{32}, z_{33}, z_{34}\}$, leading to $k = 40$ strata with proportions $(\gamma_1, \ldots, \gamma_{40})$. Within the main study, $Z$ was generated as multinomial with size $n_3$ and known sampling probabilities. Figure 2.1(a) shows the structure of selection bias in DGP 3 by comparing the stratum proportions and sampling probabilities. Some low-prevalence strata that frequently occur in the population were oversampled, while most remaining strata were undersampled. Individuals' test results were generated from the conditional distribution $X \mid Z$, where

$$\begin{aligned}\text{logit}\{P(X = 1 \mid Z)\} = \beta_0 + \beta_1 I(Z_1 = z_{11}) + \beta_2 I(Z_2 = z_{20}) + \beta_3 I(Z_2 = z_{21}) \\ + \beta_4 I(Z_3 = z_{30}) + \beta_5 I(Z_3 = z_{31}).\end{aligned}$$

The parameters $\beta_1 = -1,\ \beta_2 = -.6,\ \beta_3 = .8,\ \beta_4 = .6$, and $\beta_5 = .4$ were set to reflect differential prevalences by stratum, while a "balancing intercept" $\beta_0$ (Rudolph et al., 2021) was set to different values so that $\pi$ equalled (approximately) $\{.01, .02, \ldots, .20\}$. The non-parametric estimator $\hat{\pi}_{SRG}$

and corresponding CI were computed using a restricted target population when random nonpositivity arose; the values of $\pi$ used to compute bias and coverage were based on the total (unrestricted) population, which is the parameter of interest. The parametric estimator $\hat{\pi}_{SRGM}$ was computed with a correctly-specified logistic regression model, with parameters estimated using maximum likelihood.

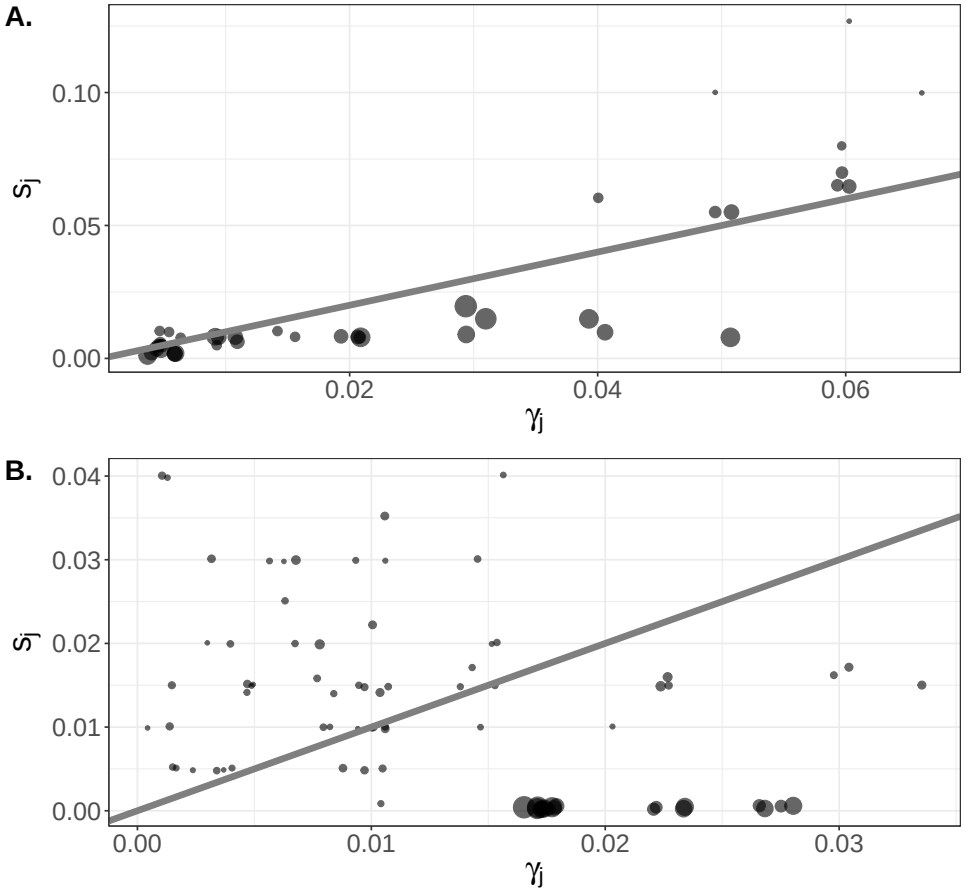


Figure 2.1: Panels A and B represent selection bias in the simulation studies of DGPs 3 and 4, described in Sections 2.4.3.1 and 2.4.3.2, respectively. Circle size is proportional to prevalence. Points are jittered slightly for legibility, and the diagonal lines denote equality between $\gamma_j$ (stratum proportion) and $s_j$ (sampling probability).

Both $\hat{\pi}_{SRG}$ and $\hat{\pi}_{SRGM}$ performed well in this scenario. Figure 2.2 shows that the estimators were generally empirically unbiased, though modest bias occurred when $\sigma_p = 0.8$ and $\pi$ was low. As in DGP 2, $\hat{\pi}_{RG}$ exhibited substantial bias and the CIs based on $\hat{\pi}_{RG}$ did not attain nominal

coverage. Figure A.7 shows 95% CIs based on either $\hat{V}_{\pi,SRG}$ or $\hat{V}_{\pi,SRGM}$ attained nominal coverage, with slight under-coverage for $\pi < 0.05$, similar to the results from DGPs 1 and 2. For $\pi = .01$ and $\sigma_p = .99$, coverage was 92% and 90% based on $\hat{V}_{\pi,SRG}$ and 91% and 90% based on $\hat{V}_{\pi,SRGM}$ for $\sigma_e \in \{.8, .99\}$, respectively. The two standardized estimators had roughly equivalent MSE (Figure A.8). On average across all 120 scenarios, positivity was present in 89% (range of 86%-92%) of simulated datasets, i.e., these datasets included all strata in the target population.

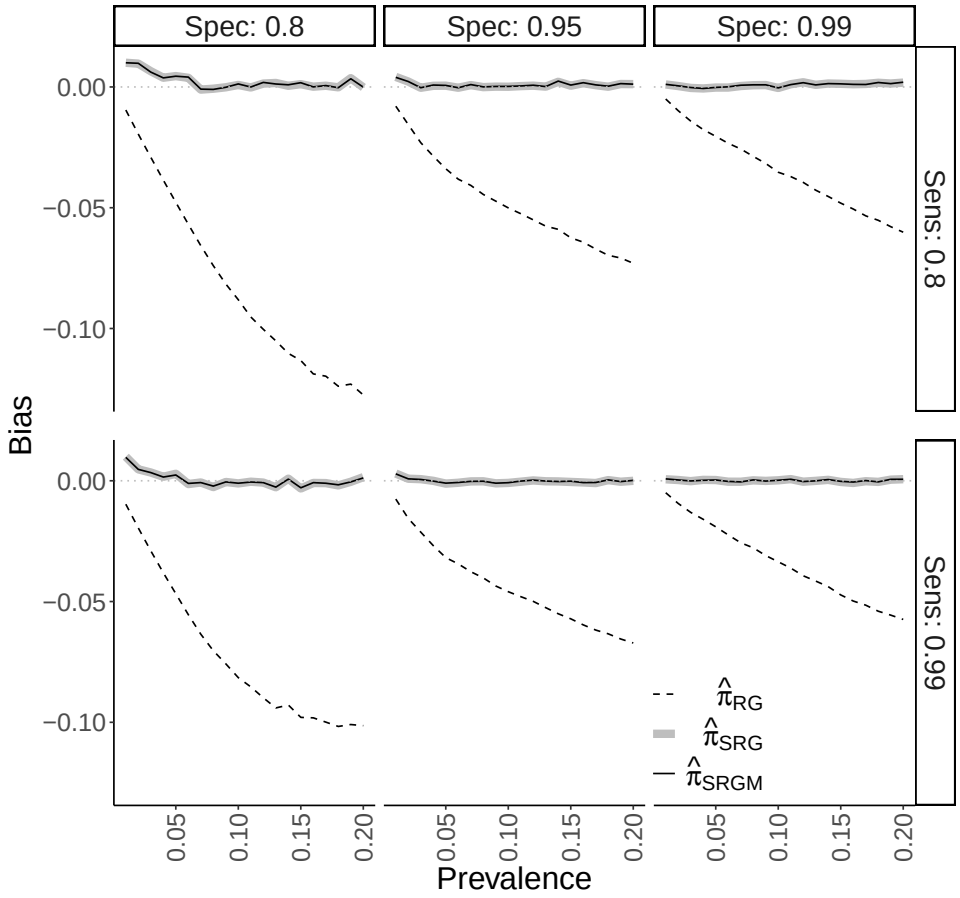


Figure 2.2: Empirical bias of the Rogan–Gladen ($\hat{\pi}_{RG}$), non-parametric standardized ($\hat{\pi}_{SRG}$), and logistic regression standardized ($\hat{\pi}_{SRGM}$) estimators from simulation study for DGP 3, described in Section 2.4.3.1. The six facets correspond to a given combination of sensitivity ('Sens') and specificity ('Spec').

#### 2.4.3.2 DGP 4

Data were generated as in DGP 3, but the inclusion of a fourth covariate $Z_4 \in \{z_{40}, z_{41}\}$ led to 80 strata. The conditional distribution $X \mid Z$ was such that $\text{logit}\{P(X = 1 \mid Z)\} = \nu h(Z)$, where $h(Z)$ here contains the same terms as in DGP 3 plus a main effect for $I(Z_4 = z_{41})$ with corresponding coefficient $\nu_6$. Regression parameters were a balancing intercept $\nu_0$, $\nu_1 = -1$, $\nu_2 = 3.25$, $\nu_3 = .8$, $\nu_4 = .6$, $\nu_5 = .4$, and $\nu_6 = .1$. The larger value for $\nu_2$, as compared to $\beta_2$, led to a stronger relationship between $X$ and $Z$ than was present in DGP 3. Figure 2.1(b) displays selection bias in DGP 4. Some of the highest-prevalence and most commonly-occurring strata were undersampled to a greater degree than occurred in DGP 3, so in this sense there was more selection bias in DGP 4. The parametric estimator $\hat{\pi}_{SRGM}$ was again computed with a correctly-specified logistic regression model using maximum likelihood for parameter estimation.

Results for the DGP 4 simulations are shown in Figure 2.3 and Figures A.9 through A.11. Figure 2.3 shows that only $\hat{\pi}_{SRGM}$ was generally unbiased under DGP 4, although there was positive bias when $\sigma_p = .8$ and $\pi < .1$. The non-parametric $\hat{\pi}_{SRG}$ typically had a moderately negative bias. Nonpositivity almost always occurred (in either all or all but one of the simulations, for each of the 120 scenarios). The worse bias of $\hat{\pi}_{SRG}$ may be explained by restriction leading to bias under nonpositivity. CIs based on $\hat{V}_{\pi,SRGM}$ typically attained nominal or close-to-nominal coverage, unlike those based on $\hat{V}_{\pi,SRG}$ or $\hat{V}_{\pi,RG}$, as seen in Figure A.9. For instance, when $\sigma_p = 0.8$, the lowest coverage for CIs based on $\hat{V}_{\pi,SRGM}$ was 92% across all 40 combinations of $\sigma_e$ and $\pi$. However, the $\hat{V}_{\pi,SRGM}$ based CIs exhibited undercoverage when $\sigma_p = .99$ and prevalence $\pi$ was low. For example, coverage of these CIs was only 59% when $\pi = 0.01$ and $\sigma_e = \sigma_p = .99$; Figure A.10 shows this undercoverage is due to 39% of the $\hat{\pi}_{SRGM}$ estimates being truncated to 0 with corresponding CIs which were overly narrow. Note that $\hat{V}_{\pi,SRGM}$ was negative for two simulations in a single scenario, and these "Heywood cases" (Kolenikov and Bollen, 2012) were ignored in the coverage calculation for that scenario. The MSE of $\hat{\pi}_{SRGM}$ tended to be less than that of $\hat{\pi}_{SRG}$ (Figure A.11).

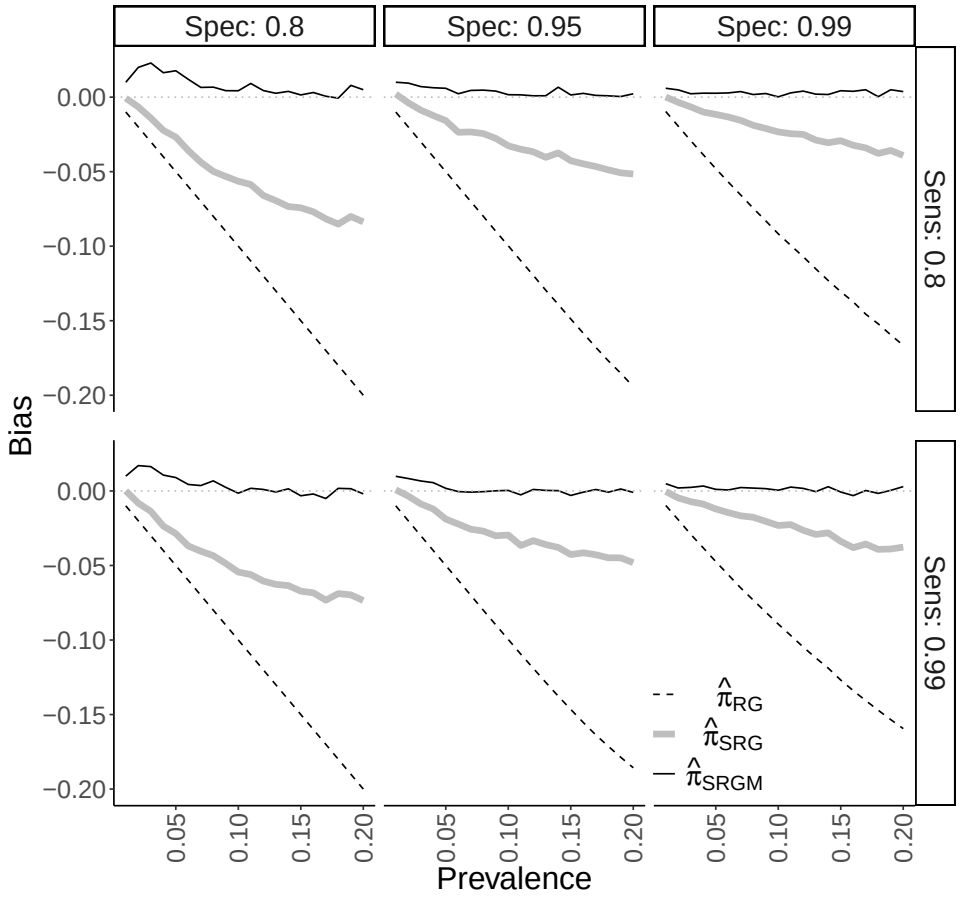


Figure 2.3: Bias results from simulation study on DGP 4, described in Section 2.4.3.2. Figure layout is as in Figure 2.2.

In summary, both the non-parametric and parametric standardized estimators $\hat{\pi}_{SRG}$ and $\hat{\pi}_{SRGM}$ had low empirical bias and close to nominal 95% CI coverage when there was positivity or near positivity. As the number of covariates, amount of selection bias, and potential for nonpositivity increased, the (correctly-specified) parametric $\hat{\pi}_{SRGM}$ generally maintained its performance while $\hat{\pi}_{SRG}$ had greater empirical bias and the intervals based on $\hat{V}_{\pi,SRG}$ did not attain nominal coverage levels.

### 2.4.4 Model misspecification

The performance of $\hat{\pi}_{SRGM}$ was assessed in scenarios similar to DGPs 3 and 4, but under model misspecification. Here, the true conditional distributions of $Y \mid Z$ were $\text{logit}\{P(Y = 1 \mid$

$Z)\} = \beta h(Z)$ and $\text{logit}\{P(Y = 1 \mid Z)\} = \nu h(Z)$, where $\beta h(Z)$ and $\nu h(Z)$ are the specifications used in the models for $\text{logit}\{P(X = 1 \mid Z)\}$ in DGPs 3 and 4, respectively. The test results $X$ were generated from $X \mid Y$ as in DGPs 1 and 2. The results shown in Figures A.12 through A.17 demonstrate that, in terms of bias, 95% CI coverage, and MSE, inferences based on $\hat{\pi}_{SRGM}$ were generally robust to this misspecification. For DGP 3 the true $X \mid Z$ distribution was $\Pr(X = 1 \mid Z) = [\text{logit}^{-1}\{\beta h(Z)\} + \sigma_p - 1]/[\sigma_e + \sigma_p - 1]$, while the model-based estimator incorrectly assumed $\Pr(X = 1 \mid Z) = \text{logit}^{-1}\{\beta h(Z)\}$, and likewise for DGP 4 with $\nu h(Z)$ replacing $\beta h(Z)$. Thus the degree of misspecification was determined by $\sigma_e$ and $\sigma_p$, with values farther from 1 leading to greater misspecification. For all simulation scenarios $\sigma_e \geq 0.8$ and $\sigma_p \geq 0.8$ such that the overall degree of misspecification was generally mild, potentially explaining the robustness of $\hat{\pi}_{SRGM}$ to model misspecification in these simulations.

Robustness of the model-based estimator $\hat{\pi}_{SRGM}$ was also assessed when the model was misspecified by omitting a variable. Under DGP 3, $\hat{\pi}_{SRGM}$ was estimated based on three misspecified logistic regression models, each omitting one of the three variables $Z_1$, $Z_2$, or $Z_3$ (i.e., omitting all indicator variables that included the variable). Results displayed in Figure A.18 show the empirical bias of $\hat{\pi}_{SRGM}$ depended on which variable was omitted, with substantial bias when $Z_2$ was not included in the model. These results demonstrate that $\hat{\pi}_{SRGM}$ may not be robust to model misspecification due to variable omission.

## 2.5 Applications

### 2.5.1 NYC seroprevalence study

The methods were applied to a seroprevalence study in New York City (NYC) that sampled patients at Mount Sinai Hospital from February 9 to July 5, 2020 (Stadlbauer et al., 2021). Patients were sampled from two groups: (1) a 'routine care' group visiting the hospital for reasons unrelated to COVID-19, including obstetric, gynaecologic, oncologic, surgical, outpatient, cardiologic, and other regular visits; (2) an 'urgent care' group of patients seen in the emergency department or admitted to the hospital for urgent care. Analyses were stratified by these two care groups. The

urgent care group may have included individuals seeking care for moderate-to-severe COVID-19 (Stadlbauer et al., 2021); this would potentially violate the assumption of equal sampling probabilities within strata, but standardized analysis of the urgent care group is included here to demonstrate the methods. The routine care group was thought to be more similar to the general population (Stadlbauer et al., 2021). Serostatus was assessed using a two-step enzyme-linked immunosorbent assay (ELISA) with estimated sensitivity of $\hat{\sigma}_e = 0.95$ from $n_1 = 40$ PCR-confirmed positive samples and estimated specificity of $\hat{\sigma}_p = 1$ from $n_2 = 74$ negative controls, 56 of which were pre-pandemic and 18 of which did not have confirmed SARS-CoV-2 infection.

In this analysis, the samples were grouped into five collection rounds of approximately equal length of time. The demographics considered were sex, age group, and race. Sex was categorized as male/female, with one individual of indeterminate sex excluded. Five age groups were $[0, 20), [20, 40), [40, 60), [60, 80)$, and $[80, 103]$. Race was coded as Asian, Black or African American, Other, and White, with 446 individuals of unknown race excluded. After exclusions, the sample size ranged from $n_3 = 937$ to 1576 in the routine care group and $n_3 = 622$ to 955 in the urgent care group across the collection rounds. The target population for standardization was NYC (8,336,044 persons), with stratum proportions and population size obtained from the 2019 American Community Survey (US Census Bureau, 2019). Table 2.1 compares the distributions of sex, age group, and race in the routine and urgent care groups to the NYC population. Women were overrepresented in the routine care group relative to the general population of NYC. Persons aged 0-19 were underrepresented in both groups, and persons aged 60 and older were overrepresented. Persons with race classified as Other were overrepresented in both groups relative to the NYC population. There was slight nonpositivity in four of the five collection rounds for the routine group and five of the five rounds for the urgent care group, and $\hat{\pi}_{SRG}$ made inference to restricted target populations. The model-based estimators $\hat{\pi}_{SRGM}$ included main effects for sex, age group, and race and an interaction term between sex and age group.

Seroprevalence estimates are presented in Figure 2.4. Adjusting for assay sensitivity and specificity resulted in slightly higher estimates and slightly wider CIs than the naive estimator

Table 2.1: Demographic comparisons of the New York City (NYC) seroprevalence study (Feb 9 - July 5, 2020) routine and urgent care group samples with the NYC population. Data on the NYC population are from the 2019 American Community Survey(US Census Bureau, 2019). Sample size is denoted by $n$.

| | | Routine care | | Urgent care | | NYC | |
|---|---|---|---|---|---|---|---|
| | | $n$ | % | $n$ | % | $n$ | % |
| | | 6,348 | 100 | 3,898 | 100 | 8,336,044 | 100 |
| Sex | Female | 4,274 | 67 | 1,789 | 46 | 4,349,715 | 52 |
| | Male | 2,074 | 33 | 2,109 | 54 | 3,986,329 | 48 |
| Age | 0-19 | 238 | 4 | 93 | 2 | 1,887,268 | 23 |
| | 20-39 | 2,624 | 41 | 551 | 14 | 2,608,394 | 31 |
| | 40-59 | 1,396 | 22 | 1,065 | 27 | 2,080,599 | 25 |
| | 60-79 | 1,780 | 28 | 1,633 | 42 | 1,426,301 | 17 |
| | 80+ | 310 | 5 | 556 | 14 | 333,482 | 4 |
| Race | Asian | 562 | 9 | 217 | 6 | 1,202,530 | 14 |
| | Black or African-American | 1,287 | 20 | 1,051 | 27 | 2,057,795 | 25 |
| | Other | 1,774 | 28 | 1,514 | 39 | 1,528,503 | 18 |
| | White | 2,725 | 43 | 1,116 | 29 | 3,547,216 | 43 |

$\hat{\rho}$. Standardization had the largest impact on the estimates in the third round, when $\hat{\pi}_{RG}$ and the standardized estimators differed by as much as 9 percentage points. The standardized estimates were accompanied by wider CIs relative to $\hat{\rho}$ and $\hat{\pi}_{RG}$, reflecting greater uncertainty associated with estimating seroprevalence when not assuming the main study data constitute a random sample from the target population. The CIs were wider as time went on and the point estimates approached 0.5, with the narrower intervals in collection periods 1 and 2.

### 2.5.2 Belgium seroprevalence study

The standardized Rogan–Gladen methods were applied to a nationwide SARS-CoV-2 seroprevalence study in Belgium conducted across seven week-long collection rounds between March and October 2020 (Herzog et al., 2022). The final collection round took place before the first vaccine authorization in the European Union in December 2020. Residual sera were collected in a stratified random sample from private laboratories encompassing a wide geographical network, with stratification by age group (10 year groups from 0-9, 10-19, . . . , 90-plus), sex (male or female), and region (Wallonia, Flanders, or Brussels). The presence of SARS-CoV-2 IgG antibodies was

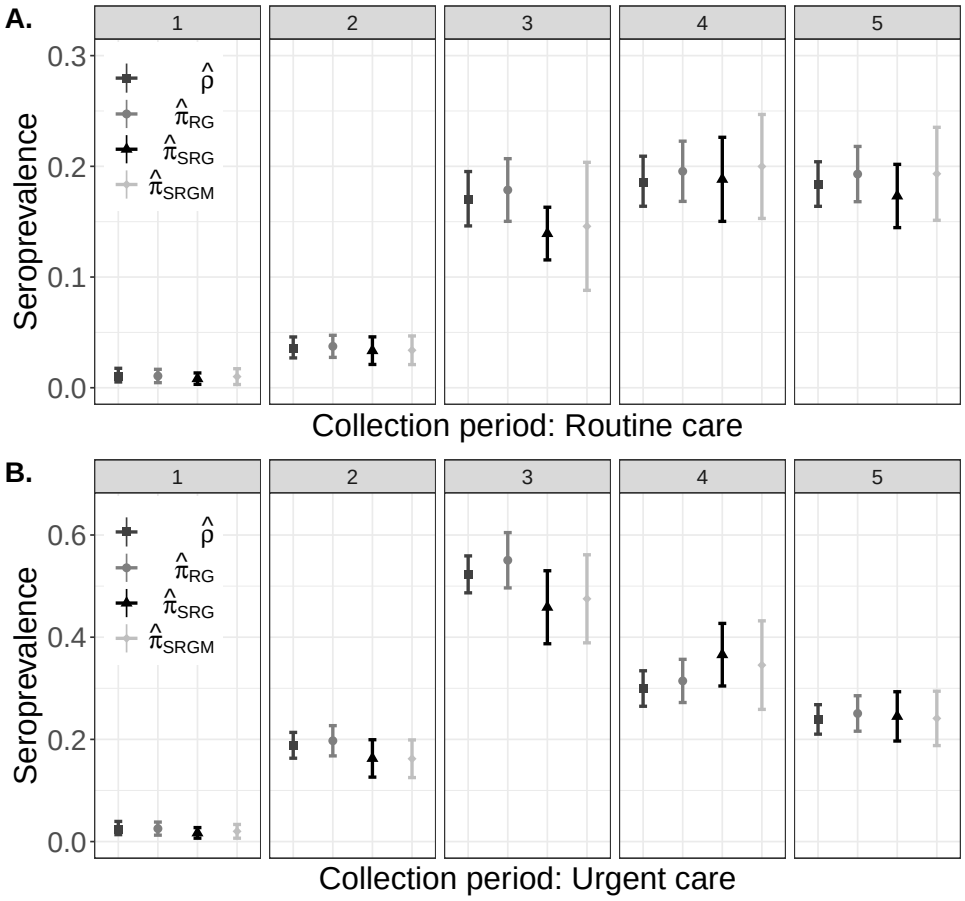


Figure 2.4: Estimates and corresponding 95% confidence intervals for each of five collection rounds for the NYC seroprevalence study (Stadlbauer et al., 2021), stratified by routine and urgent care groups, described in Section 2.5.1.

determined using a semi-quantitative EuroImmun ELISA. Based on validation studies of $n_1 = 181$ reverse transcription PCR-confirmed COVID-19 cases and $n_2 = 326$ pre-pandemic negative controls, sensitivity and specificity were estimated to be $\hat{\sigma}_e = .851$ and $\hat{\sigma}_p = .988$ (Herzog et al., 2022, Table S1.1). The number of samples for assessing seroprevalence varied between $n_3 =$ 2,960 and $n_3 =$ 3,910 across the seven collection rounds.

In this analysis, nationwide seroprevalence in Belgium was estimated during each collection round standardized by age group, sex, and province (11 total), using 2020 stratum proportion data from the Belgian Federal Planning Bureau (2021). Province was used rather than region to match the covariates selected for weighting in Herzog et al. (2022). Table 2.2 compares the sex,

Table 2.2: Demographic comparisons of the 2020 Belgium seroprevalence study sample participants in collection rounds 1 (30 March - 5 Apr) and 7 (9 - 12 Sept) with the Belgium population. Data on the Belgium population are from the Federal Planning Bureau (2021). Sample size is denoted by $n$.

| | | Round 1 | | Round 7 | | Belgium | |
|---|---|---|---|---|---|---|---|
| | | $n$ | % | $n$ | % | $n$ | % |
| | | 3,910 | 100 | 2,966 | 100 | 11,492,641 | 100 |
| Sex | Female | 2,111 | 54 | 1,589 | 54 | 5,832,577 | 51 |
| | Male | 1,799 | 46 | 1,377 | 46 | 5,660,064 | 49 |
| Age | 0-9 | 36 | 1 | 68 | 2 | 1,269,068 | 11 |
| | 10-19 | 294 | 8 | 405 | 14 | 1,300,254 | 11 |
| | 20-29 | 436 | 11 | 402 | 14 | 1,407,645 | 12 |
| | 30-39 | 461 | 12 | 397 | 13 | 1,492,290 | 13 |
| | 40-49 | 468 | 12 | 397 | 13 | 1,504,539 | 13 |
| | 50-59 | 498 | 13 | 400 | 13 | 1,590,628 | 14 |
| | 60-69 | 507 | 13 | 406 | 14 | 1,347,139 | 12 |
| | 70-79 | 506 | 13 | 204 | 7 | 924,291 | 8 |
| | 80-89 | 493 | 13 | 160 | 5 | 539,390 | 5 |
| | 90+ | 211 | 5 | 127 | 4 | 117,397 | 1 |
| Province | Antwerp | 819 | 21 | 473 | 16 | 1,869,730 | 16 |
| | Brussels | 204 | 5 | 288 | 10 | 1,218,255 | 11 |
| | East Flanders | 388 | 10 | 392 | 13 | 1,525,255 | 13 |
| | Flemish Brabant | 261 | 7 | 317 | 11 | 1,155,843 | 10 |
| | Hainaut | 245 | 6 | 271 | 9 | 1,346,840 | 12 |
| | Liege | 515 | 13 | 425 | 14 | 1,109,800 | 10 |
| | Limburg | 318 | 8 | 280 | 9 | 877,370 | 8 |
| | Luxembourg | 254 | 7 | 177 | 6 | 286,752 | 3 |
| | Namur | 352 | 9 | 170 | 6 | 495,832 | 4 |
| | Walloon Brabant | 145 | 4 | 101 | 3 | 406,019 | 4 |
| | West Flanders | 409 | 10 | 72 | 2 | 1,200,945 | 10 |

age group, and province distributions in collection rounds 1 and 7 to the Belgian population as a whole. The seroprevalence study samples were similar to the population, although the study overrepresented older persons and underrepresented younger persons relative to the population. In six of the seven collection rounds nonpositivity arose, with between 2 to 15 of the 220 strata not sampled, so restricted target populations were used for computation of $\hat{\pi}_{SRG}$. For $\hat{\pi}_{SRGM}$, each logistic regression model had main effects for age group, sex, and province, as well as an interaction term between age group and sex.

Figure 2.5 displays point estimates and CIs for $\hat{\pi}_{RG}$, $\hat{\pi}_{SRG}$, and $\hat{\pi}_{SRGM}$ alongside those for the unadjusted, or naive, sample prevalence $\hat{\rho}$ for each collection round (with exact Clopper-Pearson

95% CIs). The naive estimates $\hat{\rho}$ were typically greater than the other three estimates and had narrower CIs. The greatest differences were between $\hat{\rho}$ and $\hat{\pi}_{RG}$, which can be attributed to (estimated) measurement error in the assay. Both standardized estimates $\hat{\pi}_{SRG}$ and $\hat{\pi}_{SRGM}$ were similar in value to $\hat{\pi}_{RG}$ in most collection periods. These estimates, in combination with the stratified random sampling design, suggest that the magnitude of measurement error in this study may have been larger than that of selection bias.

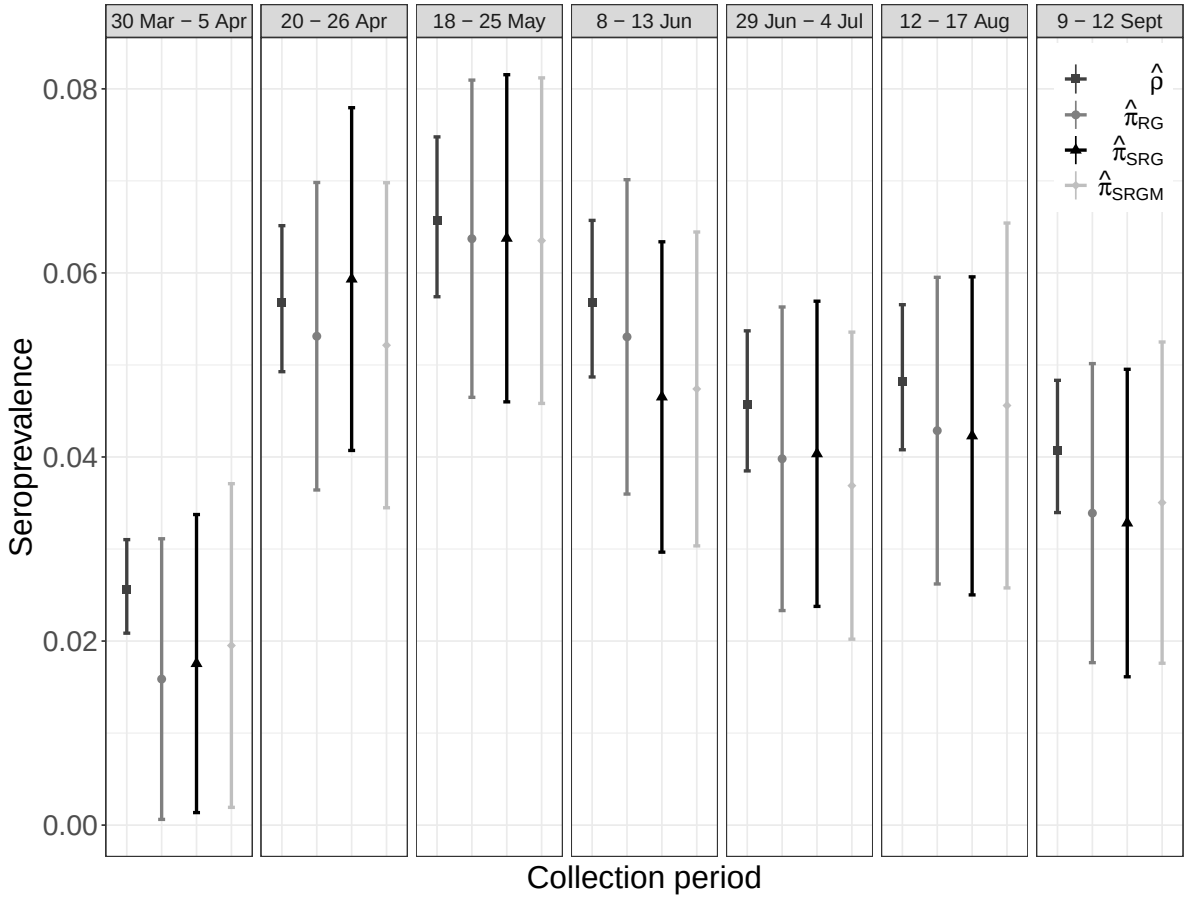


Figure 2.5: Estimates and corresponding 95% confidence intervals for each of seven collection rounds for the 2020 Belgian seroprevalence study (Herzog et al., 2022), described in Section 2.5.2.

### 2.5.3 North Carolina seroprevalence study

The standardization methods of Section 2.3 were also applied to ScreenNC, which tested a convenience sample of $n_3 = 2{,}973$ asymptomatic patients age 20 and older in North Carolina

(NC) for antibodies to SARS-CoV-2 between April to June 2020 (Barzin et al., 2020), before the authorization of vaccines in the United States. These patients were seeking unrelated medical care at eleven sites in NC associated with the University of North Carolina (UNC) Health Network. The presence of antibodies was determined with the Abbott Architect SARS-CoV-2 IgG assay. Based on validation studies of $n_1 = 40$ reverse transcription PCR confirmed positive patients and $n_2 = 277$ pre-pandemic serum samples assumed to be negative, sensitivity was estimated as $\hat{\sigma}_e = 1$ and specificity as $\hat{\sigma}_p = 0.989$.

In our analysis, seroprevalence was estimated in two relevant target populations. First, standardization was made to the population patients accessing the UNC Health Network during a similar timeframe (21,901 patients from February to June of 2020). The main study sample differed from this UNC target population in terms of age group, race, and sex characteristics, as seen in Table 2.3, and meta-analyses suggested that prevalence of COVID-19 infections differed between levels of these covariates in some populations (Pijls et al., 2021; Mackey et al., 2021), supporting the covariates' use in standardization. Note that several racial classifications, including patient refused and unknown, were reclassified as 'Other'. Second, standardization was made to the 2019 NC population over the age of 20 (7,873,971 persons) using covariate data from the American Community Survey (US Census Bureau, 2019). The assumption of equal sampling probabilities may be less reasonable for this target population because not all NC residents are in the UNC Health Network and because there were some geographic areas where no patients in the study sample were from. There was no sample data in the main study for two covariate strata that existed in the UNC Health Network, so restriction was used for $\hat{\pi}_{SRG}$. Logistic regression models with main effects for sex, race, and age group were used to compute $\hat{\pi}_{SRGM}$; interaction effects were not included as the small number of positive test results could have led to model overfit.

The sample proportion of positive tests was $\hat{\rho} = 24/2973 = 0.81\%$. The sample false positive rate was $1 - \hat{\sigma}_p = 1.08\%$, so the data are, at first appearance, consistent with a population prevalence of 0%. Indeed, the Rogan–Gladen seroprevalence estimate was $\hat{\pi}_{RG} = 0\%$ (95% CI 0%, 1.00%). Likewise, the UNC target population had non-parametric and parametric standardized estimates

Table 2.3: Demographic comparisons of the ScreenNC study sample, UNC Hospitals patient population, and North Carolina population aged 20+. Data on the NC population are from the 2019 American Community Survey (US Census Bureau, 2019). Several racial classifications including Patient Refused and Unknown were reclassified as Other. Sample size is denoted by $n$.

| | | ScreenNC | | UNC Hospitals | | NC | |
|---|---|---|---|---|---|---|---|
| | | $n$ | % | $n$ | % | $n$ | % |
| | | 2,973 | 100 | 21,901 | 100 | 7,873,971 | 100 |
| Sex | Female | 1,955 | 66 | 13,926 | 64 | 4,108,603 | 52 |
| | Male | 1,018 | 34 | 7,975 | 36 | 3,765,368 | 48 |
| Race | Asian | 67 | 2 | 460 | 2 | 230,759 | 3 |
| | Black or Af.-Am. | 395 | 13 | 5,109 | 23 | 1,640,311 | 21 |
| | Other | 311 | 10 | 1,799 | 8 | 455,600 | 6 |
| | White or Cauc. | 2,200 | 74 | 14,533 | 66 | 5,547,301 | 70 |
| Age | 20-29 | 342 | 12 | 2060 | 9 | 1,400,918 | 18 |
| | 30-39 | 599 | 20 | 2,763 | 13 | 1,344,647 | 17 |
| | 40-49 | 518 | 18 | 3,382 | 15 | 1,351,156 | 17 |
| | 50-59 | 602 | 20 | 4,200 | 19 | 1,360,357 | 17 |
| | 60-69 | 489 | 17 | 4,548 | 21 | 1,228,123 | 16 |
| | 70-79 | 310 | 11 | 3,325 | 15 | 806,002 | 10 |
| | 80+ | 77 | 3 | 1,623 | 7 | 382,768 | 5 |

of $\hat{\pi}_{SRG} = 0\%$ (0%, 1.11%) and $\hat{\pi}_{SRGM} = 0\%$ (0%, 1.13%), and the NC target population had corresponding estimates of 0% (0%, 1.10%) and 0% (0%, 1.11%). All estimates were truncated into $[0, 1]$. The closeness of the standardized and unstandardized results may be due to the small number of positive test results and similarities between the sample and the target populations. Note that the limited violations of positivity and modest demographic differences (Table 2.3) make this application more similar to DGP 3 than DGP 4. Simulation results suggest the standardized estimators and corresponding CIs may perform well in settings similar to DGP 3, even if the true prevalence is low; e.g., see the lower right facets of Figure 2 and Figure A.7, where $\sigma_e = \sigma_p = .99$.

## 2.6 Discussion

Non-parametric and model-based standardized Rogan–Gladen estimators were examined, and their large-sample properties and consistent variance estimators were derived. While motivated by SARS-CoV-2 seroprevalence studies, the methods considered are also applicable to estimation of proportions for settings where validation data can be used to estimate the measurement instrument's

sensitivity and specificity and when standardization is helpful. Simulation studies demonstrated that both standardized Rogan–Gladen methods had low empirical bias and nominal CI coverage in the majority of practical settings. The empirical results in Section 2.4 highlight the tradeoffs inherent in choosing which method to use for a seroprevalence study. The parametric standardized estimator $\hat{\pi}_{SRGM}$ was empirically unbiased even when the number of strata and covariates, and with them the potential for random nonpositivity, increased. A drawback to $\hat{\pi}_{SRGM}$ is the need to correctly specify the form of a regression model. On the other hand, the non-parametric standardized estimator $\hat{\pi}_{SRG}$ does not require model specification and performed well in scenarios with lower amounts of selection bias and nonpositivity. As the number of strata and covariates grew, however, $\hat{\pi}_{SRG}$ was empirically biased and its corresponding 95% CIs did not attain nominal coverage.

In this paper, the validation samples were not standardized, so the methodology implicitly assumed that $X \not\perp\!\!\!\perp Z$ but $X \perp\!\!\!\perp Z \mid Y$, i.e., sensitivity and specificity are constant across covariates $Z$. Sensitivity and specificity could be corrected for selection bias. In particular, the proposed nonparametric standardization estimator corrects for selection bias and then for measurement error. An alternative estimator that changes the order of corrections is $\tilde{\pi} = (\tilde{\rho} + \tilde{\sigma}_p - 1)/(\tilde{\sigma}_e + \tilde{\sigma}_p - 1)$ with tildes denoting standardized estimators, (i.e., $\tilde{\sigma}_e$ could be $\sum_{j=1}^{k} \widehat{\Pr}(Y = 1 \mid X = 1, Z = z_j) \Pr(Z = z_j)$). Test performance sometimes varies across subgroups in the spectrum effect phenomenon (Mulherin and Miller, 2002). For instance, the $n_1$ individuals in the validation sample of true positives might be based on hospitalized patients who on average have more severe disease than the general population of seropositive individuals that are sampled in the main study, and if so sensitivity may be overestimated and seroprevalence may be underestimated. Takahashi et al. (2023) used an estimator with $\tilde{\rho}$ and $\tilde{\sigma}_e$. However, stratum-specific data from the sensitivity and specificity datasets are necessary for this alternative estimator, and the validation datasets are typically much smaller than the main study dataset, as in all three examples in Section 2.5. Moreover, it is unclear if covariate data is commonly collected on sensitivity and specificity samples.

The limitations of Wald-type confidence intervals as they relate to parameters near their boundary values are mentioned in Sections 2.2.2 and 2.4. Alternative confidence intervals could

be considered based on the bootstrap (Cai et al., 2022), Bayesian posterior intervals (Gelman and Carpenter, 2020), test inversion (DiCiccio et al., 2022), or fiducial confidence distributions (Bayer et al., 2023). In particular, extensions of CIs designed to guarantee coverage at or above the nominal levels (Bayer et al., 2023; Lang and Reiczigel, 2014) could be developed to accommodate potential selection bias due to unknown sampling probabilities.

Extensions of the estimators in this paper could be considered which make additional assumptions (e.g., smoothness, monotonicity) about the longitudinal nature of seroprevalence. Another possible extension could consider variations in assay sensitivity, which may depend on a variety of factors such as the type of assay used; the recency of infection or vaccination of an individual; disease severity in infected individuals; the type and dose of vaccine for vaccinated individuals; and so forth. Additionally, inverse probability of sampling weights (Lesko et al., 2017) or inverse odds of sampling weights (Westreich et al., 2017) could be considered rather than standardization. Standardization and weighting methods could possibly be combined for the development of a doubly robust Rogan–Gladen estimator.

# CHAPTER 3: DOUBLY STANDARDIZED SURROGATE ENDPOINTS FOR SARS-CoV-2 VACCINES

## 3.1 Introduction

In the COVID-19 pandemic, vaccine efficacy estimates from Phase 3 trials supported authorization and approval of vaccines by the U.S. Food and Drug Administration and regulatory bodies worldwide. The trials took several months and were conducted under equipoise that enabled random assignment to treatment or placebo for tens of thousands of participants; similar placebo-controlled randomization for evaluation of vaccines and boosters for COVID-19 variants may be unethical because of proven vaccine efficacy. There was thus substantial interest in determining a correlate of protection, i.e., a surrogate immunological vaccine outcome that reliably correlates with vaccine efficacy (Krammer, 2021; Openshaw, 2022). Such a correlate can expedite the approval of vaccines and can be measured in the laboratory in fewer trial participants. There is strong evidence for both serum anti-spike IgG concentration and neutralizing antibody titer as correlates of protection, i.e., possible surrogate endpoints for vaccine efficacy against symptomatic COVID-19 illness. (Gilbert et al., 2022a; Cromer et al., 2022; Goldblatt et al., 2022). The FDA and European Medicines Agency have recommended approval of new vaccine strains and booster doses for SARS-CoV-2 vaccines be based on these surrogate endpoints (Gilbert et al., 2022a).

Beyond SARS-CoV-2, the use of surrogate endpoints for drug and vaccine development has received great attention over the past 40 years. Between 2010 and 2012, 45% of FDA approvals were based on a surrogate endpoint (U.S. Food and Drug Administration, 2017). Examples include hemoglobin A1c for diabetes, hemagglutination inhibition antibody titer for influenza vaccines, and progression-free or disease-free survival for several kinds of cancer (U.S. Food and Drug Administration, 2022).

Statistical methods for the evaluation and validation of surrogate endpoints can be divided into methods for a single trial and meta-analytic methods for a collection of trials. The earliest developments, including Prentice's (1989) landmark paper that defined and gave criteria for surrogacy, considered the single-trial setting. Thereafter a National Institute of Allergy and Infectious Disease workshop on HIV surrogate endpoints recognized limitations of the single-trial approach and the importance of replicating surrogate validation in a meta-analysis of trials (Albert et al., 1998). Some of the first methodological developments for the meta-analytic setting were by Daniels and Hughes (1997), Buyse et al. (2000), and Gail et al. (2000). These methods fit in the "causal association" paradigm, where "good" surrogates are ones with strong associations between the treatment effect on the surrogate endpoint and the treatment effect on the clinical endpoint (Joffe and Greene, 2009). Such meta-analytically validated surrogates have been named "general surrogates" and deemed the highest level of a hierarchy of surrogates (Qin et al., 2007; Gilbert et al., 2008). Conceptual frameworks and methods for the single-trial setting and connections between the single-trial and meta-analytic paradigms were reviewed by Joffe and Greene (2009) and Alonso et al. (2015). As an example of the meta-analytic approach, Figure 3.6 displays a meta-analysis of seven COVID-19 vaccine trials that found the surrogate endpoint of neutralizing antibody titers to be strongly associated with vaccine efficacy against symptomatic COVID-19 (Earle et al., 2021). Khoury et al. (2021) found evidence for neutralizing antibodies as a correlate of protection using a meta-analytic approach as well.

Drawing a simple comparison of COVID-19 vaccine trial efficacy estimates is problematic without considering the factors affecting the trial context and design, including characteristics of a study's population (Rapaka et al., 2022). The SARS-CoV-2 vaccine trials underrepresented older adults and the inclusion of racial/ethnic groups differed across trials (Rapaka et al., 2022; Michiels et al., 2022). It is important to measure factors such as age, sex, and sociodemographic group as possible effect modifiers for COVID-19 vaccine efficacy (World Health Organization, 2021).

In this paper, methods are presented to standardize trial-based estimates of surrogate and clinical effects to a common target population that may differ from the target populations of individual

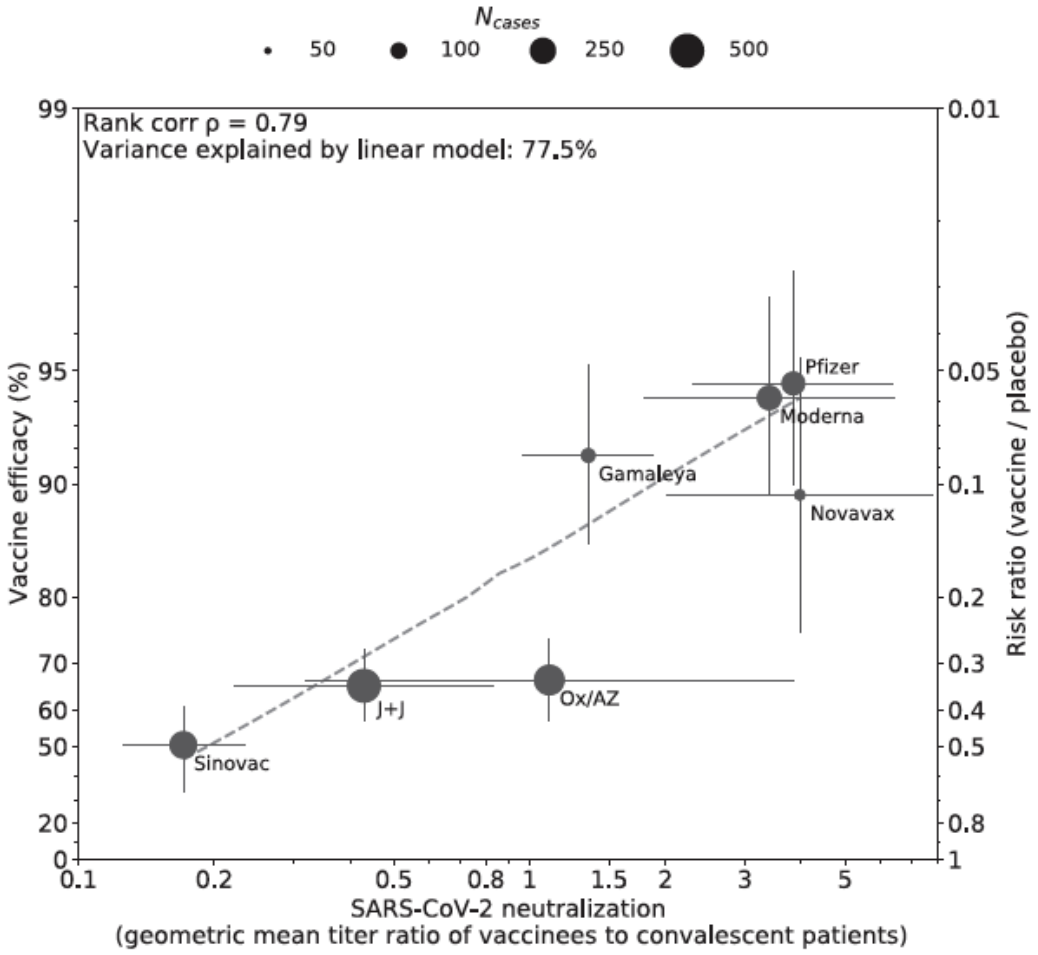


Figure 3.6: Figure 1(c) from Earle et al. (2021). Titers of SARS-CoV-2 neutralizing antibodies, calibrated to human convalescent sera measured with the same assay, are highly correlated with vaccine efficacy against symptomatic COVID-19 in seven trials. Error bars indicate 95% confidence intervals and the dashed line is a nonparametric LOESS regression fit. Further details can be found in Earle et al. (2021). The figure is reproduced under a Creative Commons CC-BY license.

trials in terms of demographic, geographic, and clinical factors. A direct standardization (i.e., parametric g-formula) estimator of the causal association between the clinical treatment effect and the surrogate treatment effect in a target population is proposed in Section 3.2. The causal association parameter is shown to be identifiable from the observed data under a set of identification assumptions, and three choices for the parameter are considered. An estimator is proposed for each example parameter, and with the asymptotic scheme where the trial sample sizes grow large but the number of trials is constant, two of the estimators are consistent and asymptotically normal. A hypothesis test for detection of violations of a key exchangeability assumption is proposed. The estimators are evaluated empirically in simulation studies guided by data from five Phase 3 vaccine trials (USG COVID-19 Response Team / Coronavirus Prevention Network (CoVPN) Biostatistics Team et al., 2022) in Section 3.3. Section 3.4 concludes with a discussion. Technical details and supplemental information are in Appendix B.

## 3.2 Methodology

### 3.2.1 Preliminaries

Consider randomized placebo-controlled trials of a treatment class, e.g., SARS-CoV-2 vaccines, with each trial measuring a clinical (primary) endpoint and a possible surrogate endpoint. Denote the clinical endpoint for individual $i$ as $Y_i$ and the surrogate endpoint as $S_i$. Let treatment be $A_i \in \mathscr{A} = \{0, 1, \ldots, m\}$, with $A_i = 0$ denoting receipt of placebo and $A_i = a$ denoting receipt of treatment in trial $a$ for $a \in \mathscr{A}_{-0} = \{1, \ldots, m\}$. An individual's potential outcomes if they were to receive treatment $A = a$ are $Y_i^a$ and $S_i^a$ for the clinical and surrogate outcomes, respectively. Suppose a covariate vector $X_i$ is observed for participants in all trials. Denote the trial by $T_i \in \mathscr{T} = \{1, \ldots, m\}$. Assume without loss of generality that trial $T = 1$ is a random sample from the target population, denoting $\mathscr{A}_{-1} = \{0, 2, \ldots, m\}$. Let $R_i = I(T_i = 1)$ be an indicator that a participant is randomly sampled from the target population. Subscripts for individuals $i$ are mostly omitted until Section 3.2.3 for notational ease. In trial $t$, the observed data are $n_t$ independent and identically

distributed (iid) observations of $(T = t, X, Y, S, A, R)$ for all $t \in \mathscr{T}$. The data combined across trials are assumed to be independent but not necessarily identically distributed. Let $n = \sum_{t \in \mathscr{T}} n_t$.

The treatment effects on the clinical and surrogate outcomes (clinical, and surrogate, effects) in the target population can be defined as $\tau^a = g\left(E[Y^a \mid R = 1], E[Y^0 \mid R = 1]\right)$ and $\delta^a = h\left(E[S^a \mid R = 1], E[S^0 \mid R = 1]\right)$ respectively for functions $g$ and $h$ of the mean potential outcomes such that $g\left(E[Y^a \mid R = 1], E[Y^a \mid R = 1]\right)$ and $h\left(E[S^a \mid R = 1], E[S^a \mid R = 1]\right)$ are equal to the null values of the effect measures $\tau^a$ and $\delta^a$. For example, if the functions are defined to be $g(x, y) = 1 - x/y$ and $h(x, y) = x - y$ and if $Y$ is binary, the effects would be vaccine efficacy $\tau^a = 1 - E[Y^a \mid R = 1]/E[Y^0 \mid R = 1]$ (one minus a causal risk ratio) and an average treatment effect $\delta^a = E[S^a \mid R = 1] - E[S^0 \mid R = 1]$. Defining $\boldsymbol{\tau} = (\tau^1, \ldots, \tau^m)$ and $\boldsymbol{\delta} = (\delta^1, \ldots, \delta^m)$, the causal association parameter $\rho = f(\boldsymbol{\tau}, \boldsymbol{\delta})$ quantifies the association between the treatment-level clinical and surrogate effects with a function $f$. In general, the analyst chooses the function $f$ such that $\rho$ is suitable for the scientific application. In the context of estimating correlates of protection for SARS-CoV-2 vaccine trials, three choices of $f$, and thus three choices of the target parameter $\rho$, are considered in Section 3.2.4.

### 3.2.2 Causal association parameter identifiability

The causal association parameter $\rho$ is identifiable under the following assumptions, wherein $a \in \mathscr{A}$ and $t \in \mathscr{T}$ except when specified otherwise.

**Assumption 3.1** (No measurement error)**.**

**Assumption 3.2** (Causal consistency)**.** $Y = Y^a I(A = a)$ *and* $S = S^a I(A = a)$.

For notational parsimony, let $O$ be a generic outcome variable such that, within each assumption or equation, $O$ can be substituted by the clinical outcome $Y$ or by the surrogate outcome $S$. Thus assumption 3.2 can be written as $O = O^a I(A = a)$, for instance.

**Assumption 3.3** (No interference)**.** *For individuals* $i$ *and* $i'$ *where* $i \neq i'$, $O_i^a$ *are unaffected by whether individual* $i'$ *receives treatment* $a$ *or* $a'$ *for all* $a, a' \in \mathscr{A}$.

Let $F_X(x \mid T)$ be the distribution function of $X \mid T$.

**Assumption 3.4** (Conditional exchangeability of treatment in each trial)**.** $O^a \perp\!\!\!\perp A \mid T, X$.

**Assumption 3.5** (Positivity of treatment assignment in each trial)**.** $\Pr(A = a \mid T, X) > 0$ *for* $X = x$ *with* $dF_X(x \mid T = t) > 0$.

**Assumption 3.6** (Conditional exchangeability of trial)**.** $O^a \perp\!\!\!\perp T \mid X$ *for* $a \in \mathscr{A}_{-0}$.

**Assumption 3.7** (Positivity of trial participation)**.** $\Pr(T = t \mid X = x) > 0$ *for* $x$ *with* $dF_X(x) > 0$.

**Assumption 3.8** (Conditional exchangeability of trial for placebo)**.** $O^0 \perp\!\!\!\perp T \mid (A = 0, X)$.

**Assumption 3.9** (Positivity of trial participation within covariate strata for placebo)**.** $\Pr(T = t \mid A = 0, X) > 0$ *for* $x$ *with* $dF_X(x) > 0$.

**Assumption 3.10** (Asymptotic non-negligibility)**.** $n_t/n \to \pi_t \in (0, 1)$ *as* $n \to \infty$.

Assumption 3.2 usually holds in settings where there is no treatment variation (Cole and Frangakis, 2009). Multi-site randomized trials such as the SARS-CoV-2 vaccine trials that motivated this work may feature limited interference between trial participants, supporting assumption 3.3. Assumption 3.4 is guaranteed by marginal or stratified randomization in each trial, and assumption 3.5 is guaranteed by marginal randomization or plausible with stratified randomization in each trial. Assumptions 3.6 and 3.8 are the critical links from the trials to the target population. Assumption 3.8 is not necessary, but may be realistic in SARS-CoV-2 vaccine trials that used identical saline placebos (USG COVID-19 Response Team / Coronavirus Prevention Network (CoVPN) Biostatistics Team et al., 2022) and enables pooling of the placebo data across trials. A diagnostic test for plausibility of assumption 3.8 is proposed in Section 3.2.5. Assumptions 3.7 and 3.9 state that, in each trial, there is a positive probability of inclusion for all covariate combinations in the target population. These two assumptions are likely reasonable for the large-scale SARS-CoV-2 vaccine trials, and if they are violated, several approaches for changing the target parameter or using parametric models are possible (Petersen et al., 2012; Westreich and Cole, 2010). Assumption 3.10 is commonly made in the biased sampling literature (e.g., Bickel and Kwon, 2001).

The identification of the potential outcome means of the clinical and surrogate endpoints in the target population ('target potential outcome means') is shown in Appendix B.1. That is, $E[O^a \mid R = 1]$ is equal to a function of the observable random variables $(A, O, T, X)$ for $a \in \mathscr{A}$. The causal association parameter $\rho = f(\boldsymbol{\tau}, \boldsymbol{\delta})$ is identifiable because it is a function of the treatment-level effects $\tau^a$ and $\delta^a$, which are in turn identifiable because they are functions of the target potential outcome means.

### 3.2.3 Estimation and inference

Let $\hat{\rho}$ be the plug-in causal association estimator $\hat{\rho} = f(\hat{\boldsymbol{\tau}}, \hat{\boldsymbol{\delta}})$, where $\hat{\boldsymbol{\tau}} = (\hat{\tau}^1, \ldots, \hat{\tau}^m)$, $\hat{\boldsymbol{\delta}} = (\hat{\delta}^1, \ldots, \hat{\delta}^m)$, $\hat{\tau}^a = g(\hat{E}[Y^a \mid R = 1], \hat{E}[Y^0 \mid R = 1])$, and $\hat{\delta}^a = h(\hat{E}[S^a \mid R = 1], \hat{E}[S^0 \mid R = 1])$. Direct standardization, i.e., Robins' (1986) parametric g-formula (see also Hernán and Robins, 2020, Chapter 13) can be used to estimate each target potential outcome mean $E[O^a \mid R = 1]$ for $a \in \mathscr{A}_{-1}$, i.e., $\hat{E}[O^a \mid R = 1] = n_1^{-1} \sum_{i=1}^{n} \{I(T_i = 1)\hat{E}[O \mid X = x, A = a]\}$. For treatment $a \in \mathscr{A}_{-1}$, constructing the estimator entails first fitting conditional mean models within subsets of participants with treatment $a$ and then standardizing the model estimates to the covariate distribution of $X$ in the target population. The covariate distribution of $X$ in the target population is estimated from the empirical distribution function of $X$ in trial 1. The data used for estimation of $\hat{E}[O^a \mid R = 1]$ combine data from the trial(s) where $A = a$ with covariate data from trial 1. With the exception of $\hat{E}[O^1 \mid R = 1]$, estimation of $\hat{E}[O^a \mid R = 1]$ requires pooling data from multiple trials. Observations of participants in these trials are independent, but not necessarily identically distributed. In this paper, parametric outcome models $\hat{E}[Y \mid X = x, A = a; \hat{\theta}_a]$ and $\hat{E}[S \mid X = x, A = a; \hat{\eta}_a]$ indexed by finite-dimensional parameters $\theta_a$ and $\eta_a$ for $a \in \mathscr{A}_{-1}$ are fit, assumed correctly specified, and estimated by maximum likelihood. For treatment $A = 1$, which is given in trial 1 which is a random sample of the target population, the sample mean from trial 1, $\hat{E}[O^1 \mid R = 1] = \{\sum_{i=1}^{n} I(A_i = 1)O_i\} / \sum_{i=1}^{n} I(A_i = 1)$, is a natural estimator of $E(O^1 | R = 1)$.

For a finite set of $m$ trials, as $n \to \infty$ and $m$ remains fixed, $\widehat{E}\,[O^a \mid R = 1]$ is, under regularity conditions, strongly consistent, i.e., converges almost surely to $E[O^a \mid R = 1]$, and asymptotically

normal (Dahabreh et al., 2022). The causal association estimator $\hat{\rho}$ can be expressed as a function $j$ of the $2m + 2$ treatment-specific estimators. If $j$ is continuous, then the causal association estimator $\hat{\rho}$ is strongly consistent by the continuous mapping theorem. If the function $j$ has $2m+2$ continuous first partial derivatives at the true value of the $2m + 2$ parameters, with at least one derivative nonzero, then $\hat{\rho}$ is asymptotically normal by the delta method (Rao, 1973, Chapter 6). For many choices of the causal association parameter $\rho$, an estimating equations approach can be used to consistently estimate the asymptotic variance of $\hat{\rho}$ with an empirical sandwich variance estimator (Stefanski and Boos, 2002), from which Wald-type confidence intervals can be computed.

### 3.2.4 Causal association parameter choice

In the context of correlates of protection for SARS-CoV-2 vaccines, consider three causal association parameters: the slope of a linear regression of $\boldsymbol{\tau}$ on $\boldsymbol{\delta}$, a Pearson correlation coefficient, and a Spearman rank correlation coefficient.

Consider the simple linear regression of $\boldsymbol{\tau}$ on $\boldsymbol{\delta}$. Let $(\beta_0, \beta) = \arg\min_{(b_0,b)} \sum_{a=1}^{m} \{\tau^a - (b_0 + b\delta^a)\}^2$ for scalars $(b_0, b)$ and $(\beta_0, \beta)$. The slope parameter $\beta$ summarizes a dose-response relationship between the surrogate and treatment effects, such that larger values of $\beta$ are indicative of stronger surrogates. The ordinary least squares estimator $\hat{\beta}$ satisfies $(\hat{\beta}_0, \hat{\beta}) = \arg\min_{(b_0,b)} \sum_{a=1}^{m} \{\hat{\tau}^a - (b_0 + b\hat{\delta}^a)\}^2$. The estimator $\hat{\beta}$ is continuous and has the continuously differentiable property defined in Section 3.2.3, so $\hat{\beta}$ is a strongly consistent and asymptotically normal estimator of $\beta$. Estimating equations are defined in Appendix B.2.1 from which a sandwich variance estimator of $\mathrm{Var}(\hat{\beta})$ and Wald-type confidence intervals for $\beta$ can be obtained. The causal association parameter $\beta$ is analogous to parameters considered by Daniels and Hughes (1997) and Dai and Hughes (2012) in meta-analytic settings. In the SARS-CoV-2 vaccine setting, regressions of clinical effects on surrogate effects were fit by Earle et al. (2021, Figure 1c) and Khoury et al. (2021, Figure 1a).

The Pearson correlation coefficient $\rho_p$ measures the linear correlation between $\boldsymbol{\tau}$ and $\boldsymbol{\delta}$. Stronger surrogates have values of $\rho_p$ closer to 1. The plug-in correlation coefficient estimator is $\hat{\rho}_p = \mathrm{Cov}(\hat{\boldsymbol{\tau}}, \hat{\boldsymbol{\delta}})/\mathrm{SD}(\hat{\boldsymbol{\tau}})\mathrm{SD}(\hat{\boldsymbol{\delta}})$, where $\mathrm{Cov}$ is defined to be the sample covariance and $\mathrm{SD}$ the sample

standard deviation. The estimator $\hat{\rho}_p$ is continuous and has the continuously differentiable property defined in Section 3.2.3, and so is strongly consistent for $\rho_p$ and asymptotically normal. The variance of $\hat{\rho}_p$ can be estimated using the stacked estimating equations given in Appendix B.2.1. The squared coefficient $\rho_p^2$ is analogous to the squared correlation considered by Dai and Hughes (2012) and the $R^2_{trials(r)}$ parameter of Buyse et al. (2000) when the latter parameter is computed from a linear model.

Spearman correlation has been used in the evaluation of surrogate endpoints, e.g., for SARS-CoV-2 vaccine correlates of protection (Earle et al., 2021), colon cancer (Sargent et al., 2007), and colorectal cancer (Tang et al., 2007). The Spearman rank correlation $\rho_s$ is a nonparametric measure of association between $\boldsymbol{\tau}$ and $\boldsymbol{\delta}$ defined as Pearson's correlation coefficient between the rank statistics $R(\boldsymbol{\tau})$ and $R(\boldsymbol{\delta})$, where $R(\boldsymbol{\tau})$ is defined so that the values of $\tau$ are ranked from highest to lowest and $R(\tau^a)$ is the rank of $\tau^a$. Each $(\hat{\tau}^a, \hat{\delta}^a)$ pair is asymptotically distributed bivariate normal as $n$ grows large, as shown in Appendix B.2.2. However, under the present asymptotic regime with a finite and fixed number of trials $m$, the plug-in estimator $\hat{\rho}_s = \mathrm{Cov}(R(\hat{\boldsymbol{\tau}}), R(\hat{\boldsymbol{\delta}}))/\mathrm{SD}(R(\hat{\boldsymbol{\tau}}))\mathrm{SD}(R(\hat{\boldsymbol{\delta}}))$ has a discrete distribution and a finite support as $n \to \infty$. In this case, interval estimation approaches based on asymptotic normal approximations are inappropriate. A bootstrap approach can be used for interval estimation of $\rho_s$ can be used, with an algorithm of resampling the data from each trial $t \in \mathscr{T}$, estimating $\hat{\rho}_s^{(b)}$ for each of $b = 1, \ldots, B$ samples, and constructing percentile bootstrap intervals using $\hat{\rho}_s^{(1)}, \ldots, \hat{\rho}_s^{(B)}$.

### 3.2.5 Diagnostic test of assumption of conditional exchangeability of trial for placebo

In this section, a diagnostic test is proposed to assess assumption 3.8 of conditional exchangeability of trial for placebo, i.e., $O^0 \perp\!\!\!\perp T \mid (A = 0, X)$. The assumption implies $E[O^0 \mid A = 0, T = 1, X] = \cdots = E[O^0 \mid A = 0, T = m, X]$, in turn implying a null hypothesis of

$$H_0 : E[O \mid A = 0, T = 1, X] = \cdots = E[O \mid A = 0, T = m, X] = E[O \mid A = 0, X].$$

by assumption 3.2 of causal consistency. An alternative hypothesis is

$$H_1 : E[O \mid A = 0, T = t, X] \neq E[O \mid A = 0, T = t', X] \text{ for some } t, t' \in \mathscr{T} \text{ with } t \neq t'.$$

These statistical hypothesis can be tested from the data.

When each endpoint $Y$ and $S$ has a distribution in an exponential family, a simple method for testing the null hypothesis uses nested generalized linear models. Denote a vector of covariates containing $X$, $T$, and possibly interactions between $X$ and $T$ by $X_T$. For each outcome $O$, the nested regression models are $E[O \mid X, A; \xi_0] = k_O^{-1}(\xi_0 X)$ and $E[O \mid X, T, A; \xi_T] = k_O^{-1}(\xi_T X_T)$ with an appropriate link function $k_O$ and row vectors of regression parameters $\xi_0$ and $\xi_T$. When $O = Y$, $\xi_0 = \theta_0$ defined in Section 3.2.3, and when $O = S$, $\xi_0 = \eta_0$ from Section 3.2.3. The null hypothesis $H_0$ corresponds to the regression coefficient on $T$ equaling 0. In this paper, $H_0$ is tested with a likelihood ratio test (LRT), although the Wald or score tests can be used. Denoting the true parameter by $\xi$, the LRT hypothesis test is $H_0 : \xi = \xi_0$ vs. $H_1 : \xi = \xi_T$. Letting the log-likelihoods of the models be $\ell(\xi_0)$ and $\ell(\xi_T)$, respectively, the LRT statistic is $LRT_O = -2[\ell(\hat{\xi}) - \ell(\hat{\xi}_T)]$. Under the asymptotic regime with fixed $m$, as $n \to \infty$ the test statistic has an asymptotic distribution of $LRT_O \xrightarrow{d} \chi^2_{df}$, where $df = \dim(\xi_T) - \dim(\xi_0)$.

The $p$-value corresponding to the likelihood ratio test statistic can be obtained with a permutation-type approach, and likely have the advantage of superior type I error control (Hayes, 1996). An algorithm for the permutation approach is as follows: permute the trial variable $T$ and compute the test statistic to generate a permutation distribution of size $n_{perm}$ of the test statistic, and then take $p = \#(\widetilde{LRT}_O > LRT_O)/n_{perm}$ as the $p$-value, where $\widetilde{LRT}_O$ are the $n_{perm}$ permuted values and $LRT_O$ is the observed test statistic.

## 3.3 Simulations

### 3.3.1 Simulation settings

A simulation study was conducted to evaluate the performance of the three causal association estimators. Motivated by the USG Phase 3 COVID-19 vaccine trial data, five trials were considered and trial 1 was considered to be a random sample from the target population. Each trial had $n_t$ participants for $t \in \{1, \ldots, 5\}$. Two baseline covariates $X = (X_1, X_2)$ were generated with $X_1$ an indicator for being at high risk of severe COVID-19 illness based on baseline trial factors (USG COVID-19 Response Team / Coronavirus Prevention Network (CoVPN) Biostatistics Team et al., 2022) and $X_2$ age in years. The at-risk for COVID-19 indicator was generated from a Bernoulli distribution with probabilities of 0.3, 0.4, and 0.2 for trials $T = 1$, $T \in \{2, 3\}$, and $T \in \{4, 5\}$, respectively. Age in years was generated from a normal distribution with means $55$, $65$, and $45$ and variances $22.3^2$, $24^2$, and $18^2$ for trials $T = 1$, $T \in \{2, 4\}$, and $T \in \{3, 5\}$ respectively. Age was then truncated into $[18, 85]$. For trial $T = 1$, the above parameters were based on simulation guidelines from the CoVPN Statistical Analysis Plan (USG COVID-19 Response Team / Coronavirus Prevention Network (CoVPN) Biostatistics Team et al., 2022).

Potential outcomes were generated as $Y^a \sim \text{Bernoulli}\left(\text{expit}(X'\theta_a)\right)$ and $S^a \sim \text{Normal}\left(X'\gamma_a, \sigma_a^2\right)$ for values of 3-vectors $\theta_a$ and $\gamma_a$ and scalar $\sigma_a^2$ for $a \in \mathscr{A}$. Treatment assignment was random, i.e., within each trial, $\Pr(A = 0) = \Pr(A = a) = 0.5$ where $a = t \in \{1, 2, 3, 4, 5\}$. The observed endpoints were $Y = Y^a I(A = a)$ and $S = S^a I(A = a)$ by assumption 3.2. The true value of each causal association parameter was determined empirically based on the average potential outcomes from 20,000,000 participants in trial $T = 1$, the random sample from the target population.

Three simulation scenarios were considered, each with different parameter values of $\theta_a$, $\gamma_a$, and $\sigma_a^2$. Scenarios 1 and 2 had sample sizes of $n_t = 3000$ for all trials while scenario 3 had $n_t = 8000$ for all trials $t \in \{1, \ldots, 5\}$. In the first two scenarios, the potential outcome mean for the true outcome in the target population $E[Y^a \mid R = 1]$ ('target potential outcome mean'), was

$E[Y^0 \mid R = 1] = 0.49$ for placebo. The target potential outcome means $E[Y^a \mid R = 1]$ for $a = 1$ through $a = 5$ in these scenarios varied from 0.02 to 0.10 in scenario 1 and 0.03 to 0.40 in scenario 2. The vaccine efficacy values $\tau^1, \ldots, \tau^5$ ranged from 0.81 to 0.96 in scenario 1 and from 0.20 to 0.93 in scenario 2. In scenario 3, the target potential outcome means were closer to the sample event rates from the USG COVID-19 vaccine trials, ranging from .002 for .034 among treatments and equaling 0.049 for placebo. In scenarios 1 and 3, the true causal association parameters were all approximately equal to 0, while in scenario 2, the true causal association parameters were approximately $\beta = .18$, $\rho_p = .90$, $\rho_s = .90$. The parameter values used to generate the simulations and the true values of each $\tau^a$ and $\delta^a$ are listed in Table B.1.

For evaluation of the diagnostic test of Section 3.2.5, a fourth scenario was simulated where assumption 3.8 of conditional exchangeability of trial for placebo did not hold and thus the null hypothesis of Section 3.2.5 was false. Let the potential outcomes be indexed by both treatment and trial. Potential outcomes were generated as $Y^{a,t} \sim \text{Bernoulli}\left(\text{expit}(X'\theta_{at})\right)$ and $S^{a,t} \sim \text{Normal}\left((X'\gamma_{at}), \sigma^2\right)$ for fixed values of 3-vectors $\theta_{at}$ and $\gamma_{at}$ and scalar $\sigma^2_{at}$ for all combinations of $a \in \mathscr{A}$ and $t \in \mathscr{T}$ under consideration, i.e., for each trial $t \in \{1, 2, 3, 4, 5\}$, $a \in \{0, t\}$. The parameter values for this scenario are listed in Table B.2.

For each of the first three scenarios, the proposed estimators $\hat{\beta}$, $\hat{\rho}_p$, and $\hat{\rho}_s$ were computed over 1000 iterations, using correctly specified logistic regression models $E[Y \mid X, A]$ and linear regression models $E[S \mid X, A]$ for each estimator. For $\hat{\beta}$ and $\hat{\rho}_p$, the empirical sandwich variance estimator was used to estimate standard errors from which 95% Wald-type confidence intervals were constructed. Sandwich variance estimates were computed with the `geex` package (Saul and Hudgens, 2020) in R (R Core Team, 2022) using the 'simple' numerical differentiation method from the `numDeriv` package . For $\hat{\rho}_s$, the percentile bootstrap interval with 10,000 bootstrap samples was used to compute 95% confidence intervals, and the standard error of $\hat{\rho}_s$ was estimated from the bootstrap distribution.

For comparison, naive point estimators of $\beta$, $\rho_p$, and $\rho_s$ were computed with each potential outcome mean estimated without standardization, i.e., the sample proportion $\hat{E}[O^a \mid R = 1] =$

$\sum_{i=1}^{n} I(A_i = a)O_i / \sum_{i=1}^{n} I(A_i = a)$ for $a \in \mathscr{A}$. For all four scenarios, the diagnostic test of Section 3.2.5 was performed, with both permutation $p$-values based on 10,000 permutations and asymptotic $p$-values computed for the likelihood ratio test statistics $LRT_Y$ and $LRT_S$.

### 3.3.2 Simulation results

The proposed estimators $\hat{\beta}$ and $\hat{\rho}_p$ generally had little to no bias and the 95% confidence intervals attained nominal coverage, while the naive estimators were substantially biased, as displayed in Table 3.4. The average estimated standard errors of $\hat{\beta}$ and $\hat{\rho}_p$, were generally similar to the empirical standard errors. An exception was $\rho_p$ in scenario 3, where bias of $\hat{\rho}_p$ was .006 and empirical confidence interval coverage was 91%.

In the first scenario, 6 of the 1000 simulations had complete or quasi-complete separation of the true endpoint $Y$ by the at-risk indicator $X_1$ within at least one treatment group $A = a$. In these cases, the maximum likelihood estimate for the regression coefficient corresponding to $X_1$ did not exist or its estimated variance was extremely large, leading to computational challenges for variance estimation of $\hat{\beta}$ and $\hat{\rho}_p$. For this scenario, 6 additional simulations were run so there were 1000 simulations.

| | | | Naive estimator | Proposed estimator | | | |
|---|---|---|---|---|---|---|---|
| Scenario | $n_t$ | Parameter | Bias | Bias | ESE | ASE | CI |
| 1 | 3000 | $\beta = 0.00$ | -0.060 | 0.000 | 0.01 | 0.01 | 0.95 |
| | | $\rho_p = 0.00$ | -0.760 | 0.003 | 0.10 | 0.10 | 0.95 |
| 2 | 3000 | $\beta = 0.18$ | -0.122 | 0.000 | 0.01 | 0.01 | 0.96 |
| | | $\rho_p = 0.90$ | -0.446 | -0.001 | 0.02 | 0.02 | 0.95 |
| 3 | 8000 | $\beta = 0.00$ | 0.004 | 0.000 | 0.03 | 0.03 | 0.95 |
| | | $\rho_p = 0.00$ | 0.042 | 0.006 | 0.17 | 0.17 | 0.91 |

Table 3.4: Average bias of the naive and proposed estimators, and empirical standard error (ESE), average estimated standard error (ASE), and 95% confidence interval (CI) coverage (%) of the estimators of the causal association parameters $\beta$ and $\rho_p$ from 1000 simulations of three scenarios described in Section 3.3.1. True values of the causal association parameters were determined empirically based on the average potential outcomes from 20,000,000 participants in trial $T = 1$, the random sample from the target population.

Performance of the proposed Spearman correlation estimator $\hat{\rho}_s$ was mixed, as displayed in Table 3.5. The point estimator was, on average, biased in two of the three scenarios, albeit less

biased than the naive estimator. The estimator was median-unbiased in all three scenarios, though, i.e., the empirical median of $\hat{\rho}_s$ was approximately equal to $\rho_s$. Interval estimation was challenging. The percentile bootstrap had 100% or near-100% coverage in all three scenarios. In scenario 2, all 1000 point estimates were $\hat{\rho}_s = 0.9$, perhaps due to the wide spread in the true values of $\tau^a$ and relatively small variability induced by the data generating mechanism (see Table B.1), and so the empirical SE was 0. Most intervals were simply the point set $\{0.9\}$, and the average interval width was less than 0.01. However, in scenarios 1 and 3, the intervals were wide, with average widths of 0.56 and 0.38, respectively. Interval estimation for the Spearman correlation is challenging for low sample sizes. For instance, considering estimation of the Spearman correlation for data pairs $(X_i, Y_i)$, $i = 1, \ldots, n$ (and no nuisance parameters), Borkowf (2000) evaluated several interval estimation methods in a simulation study and found that none worked well in general for $n = 5$ or $n = 8$.

| | | Naive est. bias | | Proposed est. bias | | Proposed estimator | | |
|---|---|---|---|---|---|---|---|---|
| Scen | $\rho_s$ | Mean | Med | Mean | Med | ESE | ASE | Coverage |
| 1 | 0.00 | -0.78 | -0.80 | -0.05 | 0.00 | 0.164 | 0.163 | 0.999 |
| 2 | 0.90 | -0.28 | -0.30 | 0.00 | 0.00 | 0.000 | 0.007 | 1.000 |
| 3 | 0.00 | -0.07 | -0.20 | 0.06 | 0.00 | 0.144 | 0.136 | 0.999 |

Table 3.5: Mean and median (Med) bias of the naive and proposed estimators (est.), and empirical standard error (ESE), average bootstrap standard error (ASE), and 95% confidence interval coverage of the proposed estimator of the causal association parameter $\rho_s$ from 1000 simulations of three scenarios (Scen) described in Section 3.3.1. True values of the causal association parameters were determined empirically based on the average potential outcomes from 20,000,000 participants in trial $T = 1$, the random sample from the target population.

The diagnostic test for violations of exchangeability assumptions proposed in Section 3.2.5 generally performed as expected. The null hypothesis of homogeneity was true in scenarios 1-3, and in these scenarios the Type I error of the test was approximately equal to the nominal significance level of .05 for both the tests based on $LRT_Y$ and $LRT_S$. In scenario 4, the null hypothesis of homogeneity was false, and the diagnostic test had high power, correctly rejecting the null in more than 99.5% of simulations for both $LRT_Y$ and $LRT_S$. The permutation and asymptotic tests gave similar results.

| | Perm $p$-value | | Asy $p$-value | |
|---|---|---|---|---|
| Scenario | $LRT_Y$ | $LRT_S$ | $LRT_Y$ | $LRT_S$ |
| 1 | .061 | .058 | .061 | .058 |
| 2 | .051 | .043 | .051 | .044 |
| 3 | .054 | .054 | .054 | .056 |
| 4 | 99.6 | 99.9 | 99.7 | 99.9 |

Table 3.6: Type I error (scenarios 1-3) and power (scenario 4) of the diagnostic hypothesis test described in Section 3.2.5 for 1,000 simulations of four scenarios described in Section 3.3.1.

## 3.4 Discussion

In this chapter, three causal association parameters were considered for the evaluation of surrogate endpoints while standardizing both true and surrogate effects to a target population. These meta-analytic parameters were shown to be identifiable from the observed data and direct standardization (parametric g-formula) point and variance estimators for each parameter were proposed. The regression slope and Pearson correlation estimators $\hat{\beta}$ and $\hat{\rho}_p$ were shown to be strongly consistent and asymptotically normal for the true parameter. A diagnostic test procedure was proposed to measure the conditional exchangeability of trial assumption, using likelihood ratio tests based on nested generalized linear models. Simulation studies were conducted motivated by U.S. government Phase 3 COVID-19 trial data (USG COVID-19 Response Team / Coronavirus Prevention Network (CoVPN) Biostatistics Team et al., 2022). The estimators $\hat{\beta}$ and $\hat{\rho}_p$ were empirically unbiased and their interval estimators attained nominal coverage in most simulation scenarios.

The Spearman correlation estimator $\hat{\rho}_s$ had empirical bias in some scenarios, although it was empirically median-unbiased. The percentile bootstrap confidence intervals for $\rho_s$ over-covered substantially, with nearly 100% coverage in all scenarios, and may have been wider than necessary. These limitations of the Spearman estimators encourage the complementary use of all three causal association parameters and estimators for surrogate endpoint evaluation. Future work could explore alternative interval estimation methods for $\rho_s$ such as the bias corrected and accelerated bootstrap (reviewed by Diciccio and Romano, 1988), transformation approaches (Bishara and Hittner, 2012),

or quantile-based methods (Borkowf, 2000). Improved interval estimation of $\rho_p$ for the low target potential outcome mean setting may be possible as well.

This paper considers a binary clinical endpoint, but trials usually have right-censoring from loss to follow up and administrative censoring. Thus evaluation of time-to-event clinical endpoints using, e.g., Cox proportional hazards outcome models, is useful. That said, for situations with rare endpoints and low rates of loss to follow up, such as the SARS-CoV-2 vaccine trials, treating the endpoint as binary may be satisfactory (USG COVID-19 Response Team / Coronavirus Prevention Network (CoVPN) Biostatistics Team et al., 2022).

In the COVID-19 setting, the circulating SARS-CoV-2 variants and strains and the force of infection affect vaccine efficacy, and these factors may differ across trials (Rapaka et al., 2022). Phase 3 correlates analyses have been limited to COVID-19 caused by original or variant lineages and ongoing studies are measuring antibody titers CoPs for omicron and future lineages (Gilbert et al., 2022a). To the extent that calendar times of trials overlap, calendar time could be included as a covariate. The possible lack of robustness when variables such as calendar time are not included is important. Sensitivity analyses could be performed such as computation of nonparametric bounds for the potential outcome means (Cole et al., 2019) and/or the causal association parameter.

Future research will apply the method to Phase 3 trial data from 5 or 6 USG / COVID-19 Response Team trials. The antibody markers in the trials were measured with a case-cohort design. The immunogenicity subcohorts consist of breakthrough cases at Day 59 and baseline positives and a random sample of trial participants stratified by key baseline covariates. Future analysis will adjust for the case-cohort design using inverse probability of sampling weights (USG COVID-19 Response Team / Coronavirus Prevention Network (CoVPN) Biostatistics Team et al., 2022).

# CHAPTER 4: ON APPLICATION OF ESTIMATING EQUATIONS TO DATA FUSION

## 4.1 Introduction

Combining data from multiple sources to improve estimation is of interest in many fields of research including biostatistics, epidemiology, statistics, and computer science. For instance, data from a randomized trial and an observational study might be fused in order to generalize or transport causal effects estimated in the randomized trial to a target population represented by an observational cohort. In these cases, the data are independent across the multiple sources, but often are not identically distributed (Pearl and Bareinboim, 2014; Bareinboim and Pearl, 2016). Several systematic reviews and tutorials of data fusion–also called data integration, data combination, and combining information–are available ( Cole et al., 2023; Colnet et al., 2023; Dahabreh, 2023; Degtiar and Rose, 2023; Ridder and Moffitt, 2007, Section 6; Shi et al., 2023. In this chapter, the implications of the non-identically distributed data for statistical inference of estimating equation estimators is considered.

Both generally and in data fusion, an estimator $\hat{\theta}$ of a parameter $\theta \in \Theta \subset \mathbb{R}^p$ can frequently be expressed as the solution to a vector of equations $\sum_{i=1}^{n} \psi(O_i; \theta) = 0$ with $\psi$ a known vector function of the observed data $O_i$ for unit $i$, with $i = 1, \ldots, n$ and $n$ total units summing over all samples. The equations are known as the estimating equations (EEs) or estimating functions (Godambe, 1960). Maximum likelihood estimators, the generalized method of moments, least squares, and generalized estimating equations can all be expressed as solutions to $\sum \psi(O_i; \theta) = 0$ for suitable choices of the $\psi$ function. The estimator $\hat{\theta}$ is termed an EE estimator in this chapter, although the term M-estimator is widely used; note that M-estimators were originally defined as maximizing an EE (Huber 1964; discussed in van der Vaart 1998, Chapter 5.1). Stefanski and Boos (2002) provide a tutorial on estimation and inference with EE estimators/M-estimators.

In this chapter, the consistency and asymptotic normality (CAN) of EE estimators is considered, i.e., proving $\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, V(\theta_0))$ with $V(\theta_0) = A(\theta_0)^{-1}B(\theta_0)\left[A(\theta_0)^{-1}\right]^T$ as $n \to \infty$, where $\theta_0$ is the true value of $\theta$ and the components $A(\theta_0)$ and $B(\theta_0)$ of the 'sandwich' variance $V(\theta_0)$ are defined in Section 4.2. The asymptotic normality of $\hat{\theta}$ is the basis for an empirical sandwich variance estimator $V(\hat{\theta})$ that replaces unknown quantities in $V(\theta_0)$ with their estimates (Stefanski and Boos, 2002). When data are independent and identically distributed (iid) from a probability model and regularity conditions for the EEs hold, then $\hat{\theta}$ is consistent and asymptotically normality (Huber, 1967), well-known results (Bang and Robins, 2005; Boos and Stefanski, 2013; Carroll et al., 2006; Lunceford and Davidian, 2004; Huber and Ronchetti, 2009). When the data $O_i$ are iid, regularity conditions are assumed for an arbitrary summand from $\sum \psi(O_i; \theta)$. Inagaki (1973) extended Huber's results to the case of independent but not necessarily identically distributed (inid) data under conditions similar to those in Huber (1967), except on each summand $\psi(O_i; \theta)$. For the non-identically distributed case, Yuan and Jennrich (1998) (hereafter YJ) developed regularity conditions YJ described as "natural yet general" and "more direct and easier to verify than those in Inagaki (1973)". Note that semiparametric and empirical process theory (van der Vaart, 1998; Kosorok, 2008) can be used to demonstrate consistency and asymptotic normality.

In data fusion, data are often inid across sources and a nuance of the standard EE theory is that it does not imply consistency and asymptotic normality for iid data. The inid EE theory of (Inagaki, 1973) or YJ does imply CAN. In this chapter the YJ conditions for consistency and asymptotic normality of EE estimators using non-identically distributed data are described in Section 4.2. Here it is shown the YJ conditions hold for two common fusion designs in Section 4.3 and thus the estimators are indeed CAN. A brief discussion concludes the chapter.

## 4.2 Estimating equation asymptotic theory

In this section the EE framework of YJ is described. Consider the EEs $\sum_{i=1}^{n} \psi(O_i; \theta) = 0$ with the data $O_i$ assumed to be independent but not necessarily identically distributed. The solution to

the EEs is the EE estimator $\hat{\theta}$. Let $G_n(\theta) = n^{-1} \sum \psi(O_i; \theta)$ with summations taken from $i = 1$ to $n$ and $\dot{G}_n(\theta) = \partial G_n(\theta)/\partial\theta$.

**Condition 4.1.** *$G_n(\theta_0) \xrightarrow{a.s.} 0$ (converges almost surely) as $n \to \infty$.*

**Condition 4.2.** *There is a neighborhood of $\theta_0$ on which with probability one $-G_n(\theta)$ is continuously differentiable; and $-\dot{G}_n(\theta)$ converges uniformly to a non-stochastic limit which is non-singular at $\theta_0$.*

**Condition 4.3.** *$\sqrt{n}G_n(\theta_0) \xrightarrow{d} \mathcal{N}(0, B(\theta_0))$ (converges in distribution).*

Under conditions 4.1 and 4.2, $\hat{\theta} \xrightarrow{a.s.} \theta_0$, i.e., $\hat{\theta}$ is a strongly consistent estimator of $\theta$ (YJ Theorem 2). Under conditions 4.1, 4.2, and 4.3, $\hat{\theta}$ is asymptotically normal, i.e., $\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, V(\theta_0))$ (YJ Theorem 4).

The alternative regularity conditions 4.4 and 4.5 from Yuan and Jennrich (1998) were

**Condition 4.4.** *For all $i$, $E[\psi(O_i; \theta_0)] = 0$ and $\mathrm{Cov}\{\psi(O_i; \theta_0)\} = B_i(\theta_0)$ with $\bar{B}(\theta_0) = n^{-1} \sum B_i(\theta_0) \to B(\theta_0)$ for positive definite $B(\theta_0)$.*

**Condition 4.5.** *For all $B_i(\theta_0)$, there exist positive numbers $\delta$ and $N$ such that for all $i$, $E\left|\psi(O_i; \theta_0)^T \{I + B_i(\theta_0)\}^{-1} \psi(O_i; \theta_0)\right|^{1+\delta} \leq N$*

Condition 4.4 implies condition 4.1 (YJ Theorem 5) and conditions 4.4 and 4.5 together imply condition 4.3 (YJ Theorem 6). Importantly, the conditions must hold for all individuals $i$, whereas in the iid setting conditions similar conditions can omit the subscript on $i$ (as detailed in Boos and Stefanski, 2013, Chapter 7.5). Under condition 4.4, an unbiased set of EEs are used and the variance of the EE vector converges to a positive definite matrix.

Let $\Theta$ denote a compact neighborhood of $\theta_0$.

**Condition 4.6.** *With probability one, the EEs $\psi(O_i; \theta)$ are twice continuously differentiable on $\Theta$.*

**Condition 4.7.** *For each $\theta \in \Theta$, $E[-\dot{G}_n(\theta)] \xrightarrow{p} -\dot{G}(\theta) = A(\theta)$ with $A(\theta_0)$ non-singular, and $-\dot{G}_n(\theta) \xrightarrow{a.s.} A(\theta)$.*

**Condition 4.8.** *For each $i$, there exists a function $q(O_i;\theta)$ such that $\|-\ddot{\psi}(O_i;\theta)\| \leq q(O_i;\theta)$ and* $\Pr(\sup_{i\geq 1} q(O_i;\theta) < \infty) = 1$.

Under conditions 4.6 through 4.8, $-\dot{G}_n(\theta) \xrightarrow{a.s.} -\dot{G}(\theta)$ uniformly and can be used for checking condition 4.2 (Yuan, 1997; Yuan and Jennrich, 1998, Lemma 6).

## 4.3 Examples

### 4.3.1 Prevalence estimation under misclassification bias with external validation data

Consider prevalence estimation of a binomial parameter $\pi = \Pr(Y = 1)$ where the variable of interest $Y$ is unobserved. The observed data $X$ are measured by an instrument subject to misclassification bias, i.e., has possibly imperfect sensitivity $Se = \Pr(X = 1 \mid Y = 1)$ and specificity $Sp = \Pr(X = 0 \mid Y = 0)$. For instance, $Y$ may be disease status and $X$ the result of a laboratory assay. The population proportion of positive results is $p = \Pr(X = 1)$. The parameter of interest $\pi$ can be expressed as a function of the nuisance parameters $\theta = (Se, Sp, p)$, as $\pi = f(\theta) = (p + Sp - 1)/(Se + Sp - 1)$.

Suppose a main study is drawn as a random sample from the population of interest. Also suppose validation studies are conducted on samples from strata of the population where it is known that $Y = 1$ and that $Y = 0$ (true positives and true negatives), in order to estimate the sensitivity and specificity of the measurement instrument. There are three independent samples with $X_1, \ldots, X_{n_1} \overset{iid}{\sim} \text{Bernoulli}(Se)$ for the sensitivity validation study, $X_{n_1+1}, \ldots, X_{n_1+n_2} \overset{iid}{\sim} \text{Bernoulli}(1 - Sp)$ for the specificity validation study, and $X_{n_1+n_2+1}, \ldots, X_{n_3} \overset{iid}{\sim} \text{Bernoulli}(p)$ for the main study with $n = n_1 + n_2 + n_3$. Let $S$ denote the sample each unit is in with $S_i = 1$ for the sensitivity study, $S_i = 2$ for the specificity study, and $S_i = 3$ for the main study, with $S$ treated as a fixed constant. Assume $(n_1/n, n_2/n, n_3/n) \to (c_1, c_2, c_3) \in (0, 1)^3$ as $n \to \infty$. The observed data are $O_i = (X_i, S_i)$ for $i = 1, \ldots, n$. The sample proportions $\widehat{Se} = n_1^{-1} \sum I(S_i = 1)X_i$, $\widehat{Sp} = n_2^{-1} \sum I(S_i = 2)(1 - X_i)$, and $\hat{p} = n_3^{-1} \sum I(S_i = 3)X_i$ are consistent estimators of the nuisance parameters. The plug-in estimator of $\pi$ is $\hat{\pi} = (\hat{p} + \widehat{Sp} - 1)/(\widehat{Se} + \widehat{Sp} - 1)$ (Rogan and Gladen, 1978; Marchevsky, 1979).

Consider the asymptotic distribution of $\hat{\pi}$. Let $\hat{\theta} = (\widehat{Se}, \widehat{Sp}, \hat{p})$. The estimator $\hat{\theta}$ solves the vector of estimating equations

$$\psi(O_i;\theta) = \begin{pmatrix} \psi_{Se} \\ \psi_{Sp} \\ \psi_p \end{pmatrix} = \begin{pmatrix} I(S_i = 1)\{X_i - Se\} \\ I(S_i = 2)\{(1 - X_i) - Sp\} \\ I(S_i = 3)\{X_i - p\} \end{pmatrix} = 0$$

(Cole et al., 2023; Rosin et al., 2023). Note $E[\psi_{Se}] = E[X_i - Se \mid S_i = 1]\Pr(S_i = 1) = 0$ by the law of total expectation and likewise $E[\psi_{Sp}] = E[\psi_p] = 0$. Let $B_i(\theta_0) = \mathrm{Cov}\{\psi(O_i;\theta_0)\}$ with

$$B_i(\theta_0) = \mathrm{diag}\left(I(S_i = 1)Se(1 - Se), I(S_i = 2)Sp(1 - Sp), I(S_i = 3)p(1 - p)\right),$$

$$\bar{B}(\theta_0) = n^{-1}\sum B_i(\theta_0) = n^{-1}\,\mathrm{diag}\left(n_1 Se(1 - Se), n_2 Sp(1 - Sp), n_3 p(1 - p)\right),$$

and

$$\bar{B}(\theta_0) \to \mathrm{diag}\left(c_1 Se(1 - Se), c_2 Sp(1 - Sp), c_3 p(1 - p)\right).$$

When the population parameters $Se$, $Sp$, and $p$ are not on the boundaries of their parameter spaces, i.e., assuming $(Se, Sp, p) \in (0, 1)^3$, the limiting matrix is positive definite because $c_1, c_2, c_3 > 0$ and Condition 4.4 holds.

For all $i$,

$$
\begin{aligned}
&E\left|\psi(O_i;\theta_0)^T\{I+B_i(\theta_0)\}^{-1}\psi(O_i;\theta_0)\right|^2 \\
&= E\left|\frac{I(S_i=1)(X_i-Se)^2}{1+I(S_i=1)Se(1-Se)}+\frac{I(S_i=2)((1-X_i)-Sp)^2}{1+I(S_i=2)Sp(1-Sp)}+\frac{I(S_i=3)(X_i-p)^2}{1+I(S_i=3)p(1-p)}\right|^2 \\
&\leq E\left|\frac{I(S_i=1)(X_i-Se)^2}{1+Se(1-Se)}+\frac{I(S_i=2)((1-X_i)-Sp)^2}{1+Sp(1-Sp)}+\frac{I(S_i=3)(X_i-p)^2}{1+p(1-p)}\right|^2 \\
&= E\left[\frac{I(S_i=1)(X_i-Se)^4}{(1+Se(1-Se))^2}\right]+E\left[\frac{I(S_i=2)((1-X_i)-Sp)^4}{(1+Sp(1-Sp))^2}\right]+E\left[\frac{I(S_i=3)(X_i-p)^4}{(1+p(1-p))^2}\right] \\
&+2E\left[\frac{I(S_i=1)I(S_i=2)(X_i-Se)^2((1-X_i)-Sp)^2}{(1+Se(1-Se))}(1+Sp(1-Sp))\right] \\
&+2E\left[\frac{I(S_i=1)I(S_i=3)(X_i-Se)^2(X_i-p)^2}{(1+Se(1-Se))}(1+p(1-p))\right] \\
&+2E\left[\frac{I(S_i=2)I(S_i=3)((1-X_i)-Sp)^2(X_i-p)^2}{(1+Sp(1-Sp))(1+p(1-p))}\right] \\
&= E\left[\frac{(X_i-Se)^4}{(1+Se(1-Se))^2}\mid S_i=1\right]\Pr(S_i=1) \\
&+E\left[\frac{((1-X_i)-Sp)^4}{(1+Sp(1-Sp))^2}\mid S_i=2\right]\Pr(S_i=2) \\
&+E\left[\frac{(X_i-p)^4}{(1+p(1-p))^2}\mid S_i=3\right]\Pr(S_i=3) \qquad (4.3) \\
&\leq E[(X_i-Se)^4\mid S_i=1]\Pr(S_i=1)+E[((1-X_i)-Sp)^4\mid S_i=2]\Pr(S_i=2) \\
&+E[(X_i-p)^4\mid S_i=3]\Pr(S_i=3) \\
&\leq \Pr(S_i=1)+\Pr(S_i=2)+\Pr(S_i=3) \\
&\leq 3
\end{aligned}
$$

with (4.3) holding by the law of total probability so Condition 4.5 holds with $\delta=1$ and $=3$. The matrix $-\dot{\psi}_i(O_i;\theta)=\text{diag}(I(S=1),I(S=2),I(S=3))$ and $-\ddot{\psi}(O_i;\theta)=0$, so the EEs are infinitely continuously differentiable and conditions 4.6 and 4.8 hold. The matrix $E[-\dot{G}_n(\theta)]$ equals

$$
n^{-1}\sum \text{diag}\left(I(S_i=1),I(S_i=2),I(S_i=3)\right)=n^{-1}\,\text{diag}(n_1,n_2,n_3)\to\text{diag}(c_1,c_2,c_3)=A(\theta)
$$

since $S$ is treated as a fixed constant and $A(\theta_0)$ is non-singular because $c_1, c_2, c_3 \in \{0, 1\}$ (with the convergence non-stochastic). The matrix $-\dot{G}_n(\theta) \to A(\theta)$ by the same logic, so condition 4.7 holds.

Thus the nuisance estimator $\hat{\theta}$ is CAN with $\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, V(\theta_0))$. The prevalence estimator $\hat{\pi}$ is CAN with asymptotic variance $\dot{f}(\theta_0)V_{\theta_0}\dot{f}(\theta_0)^T$ by the delta method.

### 4.3.2 Generalizing randomized trials using inverse probability of sampling weights

#### 4.3.2.1 Problem setup and notation

Suppose a randomized trial is conducted with $n$ participants and the goal is generalization of the trial results to a target population of size $N$, from which $m < N$ individuals are randomly sampled. Let $A$ denote a treatment indicator with $A = 1$ for treatment and $A = 0$ otherwise. Define $Y_i^a$ as the potential outcome if individual $i$ were to receive treatment $A = a$, $a \in \{0, 1\}$. Covariates $X_i$ are observed on all $n + m$ individuals combining the trial and the sample from the cohort. For parsimony, presume the trial is marginally randomized, although extensions to stratified trials by $X$ are straightforward. Denote the randomization probabilities by $\pi_a = \Pr(A = a \mid S = 1)$ for $a \in \{0, 1\}$. Let $S_i$ indicate trial participation with $S_i = 1$ for the trial and $S_i = 0$ otherwise. The main parameter of interest is the population average treatment effect $\tau = E[Y^1 - Y^0] = \mu_1 - \mu_0$ with $\mu_1$ and $\mu_0$ denoting the mean potential outcomes in the target population (assuming the potential outcomes are integrable).

Following Buchanan et al. (2018), the observed data are $n+m$ independent (inid) copies of $O_i = (S_i, X_i, S_i(A_i, Y_i))$ for individuals $i = 1, \ldots, n+m$, and the data are stacked such that the data are $n$ iid copies of $(Y_i, X_i, A_i \mid S_i = 1)$ from the trial and $m$ iid copies of $(X_i \mid S_i = 0)$ from the random sample from the target population. Several assumptions enabling causal identification. The stable unit treatment value assumption (Rubin, 1980) assumes $Y_i = Y_i^1 A + Y_i^0(1 - A)$ (causal consistency) and assumes no interference between individuals, i.e., one individual's potential outcomes are not affected by another individual's treatment assignment. Assume treatment exchangeability $Y_i^a \perp\!\!\!\perp A_i \mid S_i = 1$ and treatment positivity $\Pr(A_i = a \mid S_i = 1, X_i) > 0$ which are guaranteed

by marginal randomization. Assume sampling exchangeability $Y_i^a \perp\!\!\!\perp S_i \mid X_i$ and trial positivity $\Pr(S_i = 1 \mid X_i = x) \geq b > 0$ for some constant $b$ and all $x$ such that $dF_X(x) > 0$, where $F_X(x)$ is the distribution function of $X$. For elucidation of the asymptotic theory, assume the sampling weights (scores) $w_i(X_i) = \Pr(S_i = 1 \mid X_i)$ are known functions of the covariates, although generally when combining data the weights must be estimated, e.g., from a parametric model. The sampling weights are positive by trial positivity.

Inverse probability of sampling weighting (IPSW) is an approach for estimation of the sampling (selection) scores (Cole and Stuart, 2010; Stuart et al., 2010; Lesko et al., 2017; Buchanan et al., 2018; Egami and Hartman, 2021; Colnet et al., 2022). An IPSW estimator of $\tau$ with Hájek-type stabilized weights is

$$\hat{\tau} = \hat{\mu}_1 - \hat{\mu}_0 = \frac{\sum_i S_i Y_i A_i / w_i}{\sum_i S_i A_i / w_i} - \frac{\sum_i S_i Y_i (1 - A_i) / w_i}{\sum_i S_i (1 - A_i) / w_i}$$

with summations taken from $i = 1$ to $n + m$ (Buchanan et al., 2018; Colnet et al., 2023, Appendix E.1.1) (Buchanan et al., 2018). Let $\hat{\theta} = (\hat{\mu}_1, \hat{\mu}_0)$ denote the estimator of $\theta = (\mu_1, \mu_0)$. Buchanan et al. (2018) introduced unbiased estimating equations for $\theta$:

$$\psi(O_i; \theta) = \begin{pmatrix} S_i A_i (Y_i - \mu_1) / w_i(X_i) \\ S_i (1 - A_i)(Y_i - \mu_0) / w_i(X_i) \end{pmatrix} = 0.$$

Put $\Pr(S_i = 1) = n/N$ as the marginal probability of an individual in the target population participating in the trial and, following Buchanan et al. (2018), assume $\lim_{n,N \to \infty} n/N = \Pr(S = 1)$. Similarly put the limiting ratio of the trial size to the combined size of the trial and the sample from the target as $\lim_{n,m \to \infty} n/(n+m) = c \in (0, 1)$. Assume $E_{Y,X|S=1}[(Y - \mu_a)^m / w(X)^m \mid S = 1] \in (0, \infty)$ for $m \in \{2, 3\}$ and $a \in \{0, 1\}$. Lastly, assume the empirical distribution function of $X$ converges almost surely to a distribution function.

#### 4.3.2.2 Consistency and asymptotic normality

The unbiasedness of the estimating equations is directly implied from a proof of consistency of IPSW estimators (Colnet et al., 2022, Theorem A.2) as well as more research on asymptotics of IPSW estimators (Buchanan et al., 2018; Egami and Hartman, 2021). A proof is straightforward and is omitted. The variance of the EEs for individual $i$ is

$$\begin{aligned}
B_i(\theta) &= \mathrm{Var}(\psi(O_i;\theta)) \\
&= E\left[S_i \,\mathrm{diag}(A_i(Y_i-\mu_1)^2, (1-A_i)(Y_i-\mu_0)^2)/w_i(X_i)^2\right] \\
&= E\left[\mathrm{diag}(A_i(Y_i-\mu_1)^2, (1-A_i)(Y_i-\mu_0)^2)/w_i(X_i)^2 \mid S=1\right]\Pr(S_i=1) \\
&= \frac{n}{N} E_{Y,X|S=1}\left[\mathrm{diag}(\pi_1(Y-\mu_1)^2, \pi_0(Y-\mu_0)^2)/w(X)^2 \mid S=1\right]
\end{aligned}$$

with the last equality holding because the data are iid conditional on $S$, so

$$\begin{aligned}
\bar{B}(\theta) &= (n+m)^{-1}\sum_{i=1}^{n+m} B_i(\theta) = (n+m)^{-1}\sum_{i=1}^{n} B_i(\theta) \\
&= \frac{n}{n+m}\frac{n}{N} E_{Y,X,A|S=1}\left[\mathrm{diag}(A(Y-\mu_1)^2, (1-A)(Y-\mu_0)^2)/w(X)^2 \mid S=1\right] \\
&\xrightarrow{a.s.} c\Pr(S=1) E_{Y,X,A|S=1}\left[\mathrm{diag}(A(Y-\mu_1)^2, (1-A)(Y-\mu_0)^2)/w(X)^2 \mid S=1\right] \\
&= B(\theta)
\end{aligned}$$

with the second equality holding because the data are iid conditional on $S$. The asymptotic variance $B(\theta)$ is positive definite by the assumption that $c \in (0,1)$ and finitude of $E[(Y-\mu_a)^2/w(X)^2 \mid S=1]$. The $\Pr(S=1)$ term in the variance is only identifiable from the data if the target population size $N$ is defined (see Colnet et al., 2023, Section 2.1 and Appendix C for a discussion of identifiability). Weighted logistic regression (Scott and Wild, 1986; Buchanan et al., 2018; Egami and Hartman, 2021) incorporates the target population size in estimation.

A special case of assumption 4.5 is the boundness of $E\|\psi(O_i;\theta_0)\|^3$ (YJ). Letting $\delta = 1/2$,

$$\begin{aligned}
E\|\psi(O_i;\theta_0)\|^3 = &= E\left[(S_iA_i(Y_i-\mu_1)^2/w_i(X_i)^2 + S_i(1-A_i)(Y_i-\mu_0)^2/w_i(X_i)^2)^{3/2}\right] \\
&= E[S_iA_i(Y_i-\mu_1)^3/w_i(X_i)^3 + S_i(1-A_i)(Y_i-\mu_0)^3/w_i(X_i)^3] \\
&= E[\pi_1(Y_i-\mu_1)^3/w_i(X_i)^3 + \pi_0(Y_i-\mu_0)^3/w_i(X_i)^3 \mid S_i=1]\Pr(S_i=1) \\
&= \frac{n}{N}E_{Y,X|S=1}[\pi_1(Y-\mu_1)^3/w(X)^3 + \pi_0(Y-\mu_0)^3/w(X)^3 \mid S=1]
\end{aligned}$$

is bounded by finitude of $E[(Y-\mu_a)^3/w(X)^3 \mid S=1]$ so condition 4.5 holds.

All derivatives of $-\dot{\psi}(O_i;\theta) = \mathrm{diag}(S_iA_i/w_i, S_i(1-A_i)/w_i)$ equal 0 so $\psi(O_i;\theta)$ is infinitely continuously differentiable and condition 4.6 holds. The matrix

$$\begin{aligned}
E[-\dot{G}_n(\theta)] &= (n+m)^{-1}\sum_{i=1}^{n+m} E\left[\mathrm{diag}(A_i, 1-A_i)S_i/w_i(X_i))\right] \\
&= (n+m)^{-1}\sum_{i=1}^{n} E[\mathrm{diag}(A_i, 1-A_i)/w_i(X_i) \mid S_i=1]\Pr(S_i=1) \\
&= \frac{n}{n+m}\frac{n}{N}\,\mathrm{diag}(\pi_1,\pi_0)E_{X|S=1}\left[1/w_i(X_i) \mid S=1\right] \\
&\xrightarrow{a.s.} c\Pr(S=1)\,\mathrm{diag}(\pi_1,\pi_0)E_{X|S=1}[1/w(X) \mid S=1] \\
&= A(\theta)
\end{aligned}$$

with the third equality from iid data conditional on $S$. The limit $A(\theta_0)$ is non-singular by the assumption $c \in (0,1)$ and by trial positivity. From the above, it is straightforward that $-\dot{G}_n(\theta) \xrightarrow{a.s.} A(\theta)$ by the strong law of large numbers and thus condition 4.7 holds. Since $-\ddot{\psi}(O_i;\theta) = 0$ condition 4.8 holds. Thus $\hat{\theta}$ is CAN and $\hat{\tau}$ is CAN by the delta method.

### 4.4 Discussion

In the prevalence estimation example, an assumption was made that the true parameter $\theta_0 = (Se_0, Sp_0, p_0)$ was on the interior of the parameter space $[0,1]^3$ (not on the boundary). The parameter boundary is important in prevalence estimation because point estimates of $Se$ and $Sp$ are sometimes

1 and point estimates of the probability of testing positive are near 0 at the beginning of pandemics. The parameter interior assumption is a limitation of the approach used in YJ, as it was necessary to ensure that the asymptotic variance of the nuisance estimator was positive definite. Andrews (1999) introduced an approach to extremum estimation (of which M-estimation is a subclass) when the parameter is on the boundary that could be used in fusion problems. However, for the particular case of prevalence estimation with misclassification bias, new developments motivated by the COVID-19 pandemic suggest interval estimation approaches not based on asymptotic normal approximations have advantages (Gelman and Carpenter, 2020; DiCiccio et al., 2022; Cai et al., 2022; Bayer et al., 2023, etc.).

It could be possible to prove that the EE estimator is a uniformly consistent estimator, a stronger convergence than the almost sure convergence used in this chapter. A proof of uniform consistency might require stronger conditions. For instance, a condition such as the uniform convergence of $G_n(\theta_0)$ might be necessary, rather than the almost sure convergence in condition 4.1.While these examples and the example in Yuan and Jennrich (1998) used the alternative conditions, it could be interesting to use the those of Inagaki (1973) or conditions 4.1 through 4.3. Lastly, future directions for this research could investigate the asymptotic distribution of the inverse probability of sampling weights estimator when the regression coefficients are estimated.

# APPENDIX A: TECHNICAL DETAILS FOR CHAPTER 2

This appendix contains proofs and figures supplemental to the main text of Chapter 2.

## A.1 Maximum likelihood estimation of $(\sigma_e, \sigma_p, \rho)$

Consider estimation of the parameter vector $(\sigma_e, \sigma_p, \rho)$. The likelihood is a product of three binomial distributions corresponding to the sensitivity, specificity, and main study datasets. Letting $T_1 = \sum_{i=1}^{n_1} X_i$, $T_2 = \sum_{i=n_1+1}^{n_1+n_2} X_i$, and $T_3 = \sum_{i=n_1+n_2+1}^{n_1+n_2+n_3} X_i$, the log-likelihood is proportional to

$$T_1 \log \sigma_e + (n_1 - T_1)\log(1-\sigma_e) + (n_2 - T_2)\log \sigma_p + T_2 \log(1-\sigma_p) + T_3 \log \rho + (n_3 - T_3)\log(1-\rho).$$

Therefore the maximum likelihood estimator (MLE) of $(\sigma_e, \sigma_p, \rho)$ is $(\hat{\sigma}_e, \hat{\sigma}_p, \hat{\rho})$ where $\hat{\sigma}_e = T_1/n_1$, $\hat{\sigma}_p = (n_2 - T_2)/n_2$, and $\hat{\rho} = T_3/n_3$.

## A.2 Proofs for Section 2.2

### A.2.1 Proof of asymptotic normality

Taylor expansion of $n^{-1}\sum \psi(X_i; \delta_i, \hat{\theta})$ around the true parameter $\theta$ yields

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

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

$$\sqrt{n}(\hat{\theta} - \theta) = \left\{ n^{-1}\sum -\dot{\psi}(X_i; \delta_i, \theta) \right\}^{-1} \sqrt{n} \left\{ n^{-1}\sum \psi(X_i; \delta_i, \theta) \right\} + \sqrt{n} R^* \tag{A.1}$$

where $R^*$ is a new remainder defined below. It is shown below that $n^{-1}\sum -\dot{\psi}(X_i; \delta_i, \theta) \to_p \mathbb{A}(\theta)$, $\sqrt{n}\{n^{-1}\sum \psi(X_i; \delta_i, \theta)\} \to_d \mathcal{N}\{0, \mathbb{B}(\theta)\}$, and $\sqrt{n} R^* \to_p 0$, where $\mathbb{A}(\theta)$ and $\mathbb{B}(\theta)$ are defined below. Therefore, by Slutsky's theorem, $\sqrt{n}(\hat{\theta} - \theta) \to_d \mathcal{N}\left\{0, \mathbb{A}(\theta)^{-1}\mathbb{B}(\theta)\mathbb{A}(\theta)^{-T}\right\}$.

First, define

$$\begin{aligned}\mathbb{A}_n(X,\delta,\theta) &= n^{-1}\sum -\dot{\psi}(X_i;\delta_i,\theta)\\ &= n^{-1}\sum \begin{bmatrix} I(\delta_i=1) & 0 & 0 & 0\\ 0 & I(\delta_i=2) & 0 & 0\\ 0 & 0 & I(\delta_i=3) & 0\\ \pi & -1+\pi & -1 & \sigma_e+\sigma_p-1 \end{bmatrix}\end{aligned}$$

and let

$$\mathbb{A}(\theta) = \begin{bmatrix} c_1 & 0 & 0 & 0\\ 0 & c_2 & 0 & 0\\ 0 & 0 & c_3 & 0\\ \pi & -1+\pi & -1 & \sigma_e+\sigma_p-1 \end{bmatrix}.$$

As $n\to\infty$, $\mathbb{A}_n(X,\delta,\theta)\to\mathbb{A}(\theta)$ by the assumption that $n^{-1}\sum I(\delta_i=j)=n_j/n\to c_j\in(0,1)$ for $j\in\{1,2,3\}$.

Second, for brevity, let $\psi_e$ denote $\psi_e(X_i;\delta_i,\theta_i)$, and similarly for $\psi_p,\psi_\rho$, and $\psi_\pi$. Define

$$\begin{aligned}\mathbb{B}_n(X,\delta,\theta) &= n^{-1}\sum \mathbb{E}\left\{\psi(X_i;\delta_i,\theta)\psi(X_i;\delta_i,\theta)^T\right\}\\ &= n^{-1}\sum \mathbb{E}\begin{bmatrix} \psi_e^2 & 0 & 0 & \psi_e\psi_\pi\\ 0 & \psi_p^2 & 0 & \psi_p\psi_\pi\\ 0 & 0 & \psi_\rho^2 & \psi_\rho\psi_\pi\\ \psi_\pi\psi_e & \psi_\pi\psi_p & \psi_\pi\psi_\rho & \psi_\pi^2 \end{bmatrix}\end{aligned}$$

and let

$$\mathbb{B}(\theta) = \begin{bmatrix} c_1\sigma_e(1-\sigma_e) & 0 & 0 & 0 \\ 0 & c_2\sigma_p(1-\sigma_p) & 0 & 0 \\ 0 & 0 & c_3\rho(1-\rho) & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

Note that $\mathbb{B}(\theta) = \lim_{n\to\infty} \mathbb{B}_n(X, \delta, \theta)$ as $\mathbb{E}(\psi_e^2) = \mathbb{E}\{I(\delta_i = 1)(X_i - \sigma_e)^2\} = I(\delta_i = 1)\sigma_e(1-\sigma_e)$ and likewise for $\mathbb{E}(\psi_p^2)$ and $\mathbb{E}(\psi_\rho^2)$. It follows that $\sqrt{n}\left\{n^{-1}\sum \psi(X_i; \delta_i, \theta)\right\} \to_d \mathcal{N}\{0, \mathbb{B}(\theta)\}$ by the Lindeberg-Feller CLT. In particular, let $s_n^2 = \sum\{\mathrm{Var}(\psi_e) + \mathrm{Var}(\psi_p) + \mathrm{Var}(\psi_\rho) + \mathrm{Var}(\psi_\pi)\} = n_1\sigma_e(1-\sigma_e) + n_2\sigma_p(1-\sigma_p) + n_3\rho(1-\rho)$ and let $\|\cdot\|$ denote the Euclidean norm. Note that $\max_i \|\psi(X_i; \delta_i, \theta)\| \leq \sqrt{2}$ as the maximum magnitude of each element of $\psi(X_i; \delta_i, \theta)$ is 1, and for a given observation $X_i$ it is always true that two of the indicators $\delta_1, \delta_2, \delta_3$ equal zero and the third indicator equals one. Therefore for all $\epsilon > 0$,

$$\lim_{n\to\infty} s_n^{-2} \sum \mathbb{E}\left\{\|\psi(X_i; \delta_i, \theta)\|^2 I(\|\psi(X_i; \delta_i, \theta)\| \geq \epsilon s_n)\right\} \leq \lim_{n\to\infty} 2s_n^{-2} \sum P(\sqrt{2} \geq \epsilon s_n) = 0,$$

implying the Lindeberg condition holds.

Third, it remains to prove $\sqrt{n}R^* \to_p 0$. The outline of the proof of Boos and Stefanski (2013) Theorem 7.2 can be followed, but their assumption of identically distributed data must be removed. Consider the second-order Taylor series expansion of the $j$th element of the vector $n^{-1}\sum \psi(X_i; \delta_i, \hat{\theta})$, denoted $n^{-1}\sum \psi_j(X_i; \delta_i, \hat{\theta})$, around the true value $\theta$:

$$\begin{aligned} 0 = n^{-1}\sum \psi_j(X_i; \delta_i, \hat{\theta}) = n^{-1}\sum \psi_j(X_i; \delta_i, \theta) + n^{-1}\sum \dot{\psi}_j(X_i; \delta_i, \theta)(\hat{\theta} - \theta) \\ + \frac{1}{2}(\hat{\theta} - \theta)^T n^{-1}\sum \ddot{\psi}_j(X_i; \delta_i, \widetilde{\theta}_j)(\hat{\theta} - \theta) \end{aligned}$$

where $\widetilde{\theta}_1, \ldots, \widetilde{\theta}_4$ are on the line segment joining $\hat{\theta}$ and $\theta$ and $\ddot{\psi}_j(X_i; \delta_i, \theta_j)$ is a $4 \times 4$ matrix with entry $(j, k)$ equal to $\partial^2\psi_j(X_i; \delta_i, \theta_j)/\partial\theta_j\partial\theta_k$ for $j, k \in \{1, 2, 3, 4\}$. Writing these 4 equations in matrix notation yields

$$0 = n^{-1}\sum \psi(X_i; \delta_i, \theta) + \widetilde{R}(\hat{\theta} - \theta) \tag{A.2}$$

where $\widetilde{R} = \left\{ n^{-1} \sum \dot{\psi}(X_i; \delta_i, \theta) + (1/2)\widetilde{Q} \right\}$ and

$$\widetilde{Q} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ \sigma_e - \hat{\sigma}_e & \sigma_p - \hat{\sigma}_p & 0 & 0 \end{bmatrix}$$

is the $4 \times 4$ matrix with $j$th row given by $(\hat{\theta} - \theta)^T n^{-1} \sum \ddot{\psi}_j(X_i; \delta_i, \widetilde{\theta}_j)$. Note that $\widetilde{Q} \to_p 0_{4\times 4}$ by the Weak Law of Large Numbers, where in general $0_{r\times c}$ denotes an $r \times c$ matrix of zeros. Also note that $n^{-1} \sum \dot{\psi}(X_i; \delta_i, \theta) \to_p -\mathbb{A}(\theta)$, where $-\mathbb{A}(\theta)$ is nonsingular under the assumption that $\sigma_e > 1 - \sigma_p$. It follows that as $n \to \infty$, $\widetilde{R}$ is invertible with probability one. On the set $S_n$ where $\widetilde{R}$ is invertible, (A.2) can be rearranged to yield

$$\hat{\theta} - \theta = \left( -\widetilde{R} \right)^{-1} \left\{ n^{-1} \sum \psi(X_i; \delta_i, \theta) \right\}. \tag{A.3}$$

Define

$$\widetilde{R}^* = \frac{1}{1+g} \left\{ n^{-1} \sum -\dot{\psi}(X_i; \delta_i, \theta) \right\}^{-1} (1/2)\widetilde{Q} \left\{ n^{-1} \sum -\dot{\psi}(X_i; \delta_i, \theta) \right\}^{-1}$$

where $g = \mathrm{Tr}\left[ -(1/2)\widetilde{Q} \left\{ n^{-1} \sum -\dot{\psi}(X_i; \delta_i, \theta) \right\}^{-1} \right]$, and note that $\widetilde{R}^* \to_p 0_{4\times 4}$ by Slutsky's theorem because $\widetilde{Q} \to_p 0_{4\times 4}$. Since $(1/2)\widetilde{Q}$ has rank one, an application of the Sherman-Morrison-Woodbury formula (Miller, 1981) can be used to show $\left( -\widetilde{R} \right)^{-1} = \left\{ n^{-1} \sum -\dot{\psi}(X_i; \delta_i, \theta) \right\}^{-1} + \widetilde{R}^*$. Thus, substituting this expression for $(-\widetilde{R})^{-1}$ into (A.3) and multiplying by $\sqrt{n}$ yields (A.1), where $R^* = \widetilde{R}^* \left\{ n^{-1} \sum \psi(X_i; \delta_i, \theta) \right\}$. Therefore, because $\lim_{n\to\infty} \Pr(S_n) = 1$, $\widetilde{R}^* \to_p 0_{4\times 4}$, and $\sqrt{n} \left\{ n^{-1} \sum \psi(X_i; \delta_i, \theta) \right\} \to_d \mathcal{N}(0, \mathbb{B}(\theta))$, by Slutsky's theorem $\sqrt{n} R^* \to_p 0$.

### A.2.2 Computation of asymptotic variance

Since $\mathbb{A}(\theta)$ is lower triangular, it follows that

$$\mathbb{A}(\theta)^{-1} = \begin{bmatrix} 1/c_1 & 0 & 0 & 0 \\ 0 & 1/c_2 & 0 & 0 \\ 0 & 0 & 1/c_3 & 0 \\ -\dfrac{\pi}{c_1(\sigma_e+\sigma_p-1)} & \dfrac{1-\pi}{c_2(\sigma_e+\sigma_p-1)} & \dfrac{1}{c_3(\sigma_e+\sigma_p-1)} & (\sigma_e+\sigma_p-1)^{-1} \end{bmatrix},$$

and therefore $\mathbb{A}(\theta)^{-1}\mathbb{B}(\theta)\mathbb{A}(\theta)^{-T}$ equals

$$\begin{bmatrix} \sigma_e(1-\sigma_e)/c_1 & 0 & 0 & * \\ 0 & \sigma_p(1-\sigma_p)/c_2 & 0 & * \\ 0 & 0 & \rho(1-\rho)/c_3 & * \\ * & * & * & V_{\pi,RG} \end{bmatrix},$$

where * denotes quantities not expressed explicitly and

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

## A.3 Proofs for Section 2.3.2

### A.3.1 Proof of asymptotic normality

The Taylor expansion of $n^{-1}\sum \psi(X_i, Z_i; \delta_i, \hat{\theta}_s)$ around the true parameter $\theta_s$ yields

$$\sqrt{n}(\hat{\theta}_s - \theta_s) = \left\{ n^{-1}\sum -\dot{\psi}(X_i, Z_i; \delta_i, \theta_s) \right\}^{-1} \sqrt{n} \left\{ n^{-1}\sum \psi(X_i, Z_i; \delta_i, \theta_s) \right\} + \sqrt{n}R^*,$$

which is similar in form to (A.1), except here the estimating equations are dependent on covariates $Z$. The remainder $R^*$ here is distinct from that in Appendix A.2.1, and in general symbols may be reused and notation may not hold the same meaning across appendices. Below it is established,

using an analogous approach to that of Appendix A.2.1, that $n^{-1}\sum -\dot{\psi}(X_i, Z_i; \delta_i, \theta_s) \to_p \mathbb{A}(\theta_s)$, $\sqrt{n}\{n^{-1}\sum \psi(X_i, Z_i; \delta_i, \theta_s)\} \to_d \mathcal{N}\{0, \mathbb{B}(\theta_s)\}$, and $\sqrt{n}R^* \to_p 0$. Therefore, by Slutsky's theorem, $\sqrt{n}(\hat{\theta}_s - \theta_s) \to_d \mathcal{N}\{0, \mathbb{A}(\theta_s)^{-1}\mathbb{B}(\theta_s)\mathbb{A}(\theta_s)^{-T}\}$.

First, define

$$\mathbb{A}_n(X, Z, \delta, \theta_s) = n^{-1}\sum \left\{-\dot{\psi}(X_i, Z_i; \delta_i, \theta_s)\right\} = \begin{bmatrix} A' & 0_{(k+2)\times 2} \\ C & D \end{bmatrix}$$

as a block matrix where $A' = \text{diag}\,(n_1/n, n_2/n, n_{z_1}/n, \ldots, n_{z_k}/n)$ is $(k+2) \times (k+2)$,

$$C = \begin{bmatrix} 0 & 0 & -\gamma_1 & \ldots & -\gamma_k \\ \pi & -1+\pi & 0 & \ldots & 0 \end{bmatrix}$$

is $2 \times (k+2)$, and

$$D = \begin{bmatrix} 1 & 0 \\ -1 & \sigma_e + \sigma_p - 1 \end{bmatrix}.$$

Let

$$\mathbb{A}(\theta_s) = \begin{bmatrix} A & 0_{(k+2)\times 2} \\ C & D \end{bmatrix}$$

where $A = \text{diag}\{c_1, c_2, c_3 s_1, \ldots, c_3 s_k\}$, and note that $\mathbb{A}(\theta_s) = \lim_{n\to\infty} \mathbb{A}_n(X, Z, \delta, \theta_s)$ since $n_j/n \to c_j$ for $j \in \{1, 2, 3\}$. Note that $s_1, \ldots, s_k$ are all nonzero due to positivity, without which $\hat{\pi}_{SRG}$ is undefined.

Second, define

$$\mathbb{B}_n(X, Z, \delta, \theta_s) = n^{-1}\sum \mathbb{E}\{\psi(X_i, Z_i; \delta_i, \theta_s)\psi(X_i, Z_i; \delta_i, \theta_s)^T\} = \begin{bmatrix} E' & 0_{(k+2)\times 2} \\ 0_{2\times(k+2)} & 0_{2\times 2} \end{bmatrix},$$

where

$$E' = \text{diag}\{n_1\sigma_e(1-\sigma_e)/n, n_2\sigma_p(1-\sigma_p)/n, n_{z_1}\rho_1(1-\rho_1)/n, \ldots, n_{z_k}\rho_k(1-\rho_k)/n\},$$

since $\mathbb{E}(\psi^2_{\rho_j}) = \mathbb{E}\{I(Z_i = z_j, \delta_i = 3)(X_i - \rho_j)^2\} = I(Z_i = z_j, \delta_i = 3)\rho_j(1-\rho_j)$. Define $\mathbb{B}(\theta_s)$ to have the same form as $\mathbb{B}_n(X, Z, \delta, \theta_s)$ except $E'$ is replaced by

$$E = \text{diag}\{c_1\sigma_e(1-\sigma_e), c_2\sigma_p(1-\sigma_p), c_3 s_1\rho_1(1-\rho_1), \ldots, c_3 s_k\rho_k(1-\rho_k)\}.$$

Since $\mathbb{B}(\theta_s) = \lim_{n\to\infty}\mathbb{B}_n(X, Z, \delta, \theta_s)$, $\sqrt{n}\{n^{-1}\sum\psi(X_i, Z_i; \delta_i, \theta_s)\} \to_d \mathcal{N}\{0, \mathbb{B}(\theta_s)\}$ by the Lindeberg-Feller CLT. In particular, let

$$\begin{aligned} v_n^2 &= \sum\{\text{Var}(\psi_e) + \text{Var}(\psi_p) + \text{Var}(\psi_{\rho_1}) + \cdots + \text{Var}(\psi_{\rho_k}) + \text{Var}(\psi_\rho) + \text{Var}(\psi_\pi)\} \\ &= n_1\sigma_e(1-\sigma_e) + n_2\sigma_p(1-\sigma_p) + \sum_{j=1}^{k} n_{z_j}\rho_j(1-\rho_j), \end{aligned}$$

and note that $\max_i\|\psi(X_i, Z_i; \delta_i, \theta_s)\| \leq \sqrt{3}$ because at most one of $\psi_e, \psi_p, \psi_{\rho_1}, \ldots, \psi_{\rho_k}$ is nonzero and each element of $\psi(X_i, Z_i; \delta_i, \theta_s)$ has a maximum value of one. Therefore for all $\epsilon > 0$,

$$\lim_{n\to\infty} v_n^{-2}\sum\mathbb{E}\left\{\|\psi(X_i; \delta_i, \theta)\|^2 I(\|\psi(X_i; \delta_i, \theta)\| \geq \epsilon v_n)\right\} \leq \lim_{n\to\infty} 3v_n^{-2}\sum P(\sqrt{3} \geq \epsilon v_n) = 0,$$

implying the Lindeberg condition holds.

Third, let $\widetilde{Q}_s$ be analogous to $\widetilde{Q}$ from Appendix A.2.1. Denote the $j$th entry of $\theta_s$ as $\theta_{s_j}$ and note that $|\partial^2\psi_j(X_i, Z_i; \delta_i, \theta_s)/\partial\theta_{s_j}\theta_{s_l}| \leq 1$ for all $j, l \in \{1, 2, \ldots, k+4\}$. Thus there exists a function $g(X_i, Z_i, \delta_i)$ such that, for each $\theta_s^*$ in a neighborhood of $\theta_s$, $\left|\partial^2\psi_j(X_i, Z_i; \delta_i, \theta_s)/\partial\theta^*_{s_j}\partial\theta^*_{s_l}\right| \leq g(X_i, Z_i, \delta_i)$ holds for all $X_i, Z_i, \delta_i$ where $\int g(X_i, Z_i, \delta_i)dF(X_i, Z_i, \delta_i) < \infty$; e.g., let $g(X_i, Z_i, \delta_i) = 2$. Therefore, each entry in $\widetilde{Q}_s$ is bounded by $\|\hat{\theta}_s - \theta_s\|n^{-1}\sum g(X_i, \delta_i) = o_p(1)$, so $\widetilde{Q}_s \to_p 0_{(k+4)\times(k+4)}$. This fact can be used to prove $\sqrt{n}R^* \to_p 0$ with the same technique as in Appendix A.2.1, so we omit the rest of the proof.

### A.3.2 Computation of asymptotic variance

Note that $\mathbb{A}(\theta_s)$ is block lower triangular, and thus

$$\mathbb{A}(\theta_s)^{-1} = \begin{bmatrix} A^{-1} & 0_{(k+2)\times 2} \\ -D^{-1}CA^{-1} & D^{-1} \end{bmatrix}$$

and therefore

$$\mathbb{A}(\theta_s)^{-1}\mathbb{B}(\theta_s)\mathbb{A}(\theta_s)^{-T} = \begin{bmatrix} A^{-1}EA^{-1} & -A^{-1}EA^{-1}C^TD^{-T} \\ -D^{-1}CA^{-1}EA^{-1} & D^{-1}CA^{-1}EA^{-1}C^TD^{-T} \end{bmatrix}.$$

The bottom right submatrix of $\mathbb{A}(\theta_s)^{-1}\mathbb{B}(\theta_s)\mathbb{A}(\theta_s)^{-T}$, and specifically the bottom right element of that submatrix, is of primary interest. Since $A$ and $E$ are both diagonal,

$$A^{-1}EA^{-1} = \text{diag}\{c_1^{-1}\sigma_e(1-\sigma_e), c_2^{-1}\sigma_p(1-\sigma_p), c_3^{-1}s_1^{-1}\rho_1(1-\rho_1), \ldots, c_3^{-1}s_k^{-1}\rho_k(1-\rho_k)\}.$$

Next, note

$$D^{-1}C = \begin{bmatrix} 0 & 0 & -\gamma_1 & \cdots & -\gamma_k \\ \dfrac{\pi}{\sigma_e+\sigma_p-1} & \dfrac{-1+\pi}{\sigma_e+\sigma_p-1} & \dfrac{-\gamma_1}{\sigma_e+\sigma_p-1} & \cdots & \dfrac{-\gamma_k}{\sigma_e+\sigma_p-1} \end{bmatrix}.$$

Therefore, $D^{-1}CA^{-1}EA^{-1}$ equals

$$\begin{bmatrix} 0 & 0 & -\gamma_1 c_3^{-1}s_1^{-1}\rho_1(1-\rho_1) & \cdots & -\gamma_k c_3^{-1}s_k^{-1}\rho_1(1-\rho_1)) \\ \dfrac{\pi\sigma_e(1-\sigma_e)}{c_1(\sigma_e+\sigma_p-1)} & \dfrac{(-1+\pi)\sigma_p(1-\sigma_p)}{c_2(\sigma_e+\sigma_p-1)} & -\dfrac{\gamma_1\rho_1(1-\rho_1)}{c_3 s_1(\sigma_e+\sigma_p-1)} & \cdots & -\dfrac{\gamma_k\rho_k(1-\rho_k)}{c_3 s_k(\sigma_e+\sigma_p-1)} \end{bmatrix}$$

and thus

$$D^{-1}CA^{-1}EA^{-1}C^TD^{-T} = \begin{bmatrix} * & * \\ * & V_{\pi,SRG} \end{bmatrix}$$

where * denotes quantities not expressed explicitly and

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

### A.4 Proof for Section 3.3

The proof that $\theta_m$ is asymptotically normal is given for the case of logistic regression, but it extends to any link function $g(\cdot)$ appropriate for binary regression. Recall that the $(p+4)$-vector of estimating equations is

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

In the above vector $\sum \psi_e, \sum \psi_p$, and $\psi_\pi$ are identical to the equations used in Appendices A.2 and A.3; $\psi_\rho = \sum_{j=1}^{k} \text{logit}^{-1}\{\beta h(Z_j)\}\gamma_j - \rho$; and $\sum \psi_\beta$ is a $p$-vector with $j$th element $\sum \psi_{\beta_j} = \sum I(\delta_i = 3) \left[ X_i - \text{logit}^{-1}\{\beta h(Z_i)\} \right] h_j(Z_i)$. The Taylor expansion of $n^{-1} \sum \psi(X_i, Z_i; \delta_i, \hat{\theta}_m)$ around the true parameter $\theta_m$ yields

$$\sqrt{n}(\hat{\theta}_m - \theta_m) = \left\{ n^{-1} \sum -\dot{\psi}(X_i, Z_i; \delta_i, \theta_m) \right\}^{-1} \sqrt{n} \left\{ n^{-1} \sum \psi(X_i, Z_i; \delta_i, \theta_m) \right\} + \sqrt{n} R^*.$$

The rest of the proof is similar to Appendices A.2.1 and A.3.1. Namely, first define

$$\mathbb{A}_n(X, Z, \delta, \theta_m) = n^{-1} \sum \left\{ -\dot{\psi}(X_i, Z_i; \delta_i, \theta_m) \right\} = \begin{bmatrix} A' & 0_{2\times p} & 0_{2\times 2} \\ 0_{p\times 2} & B' & 0_{p\times 2} \\ C & D & E \end{bmatrix}$$

as a block matrix where $A' = \text{diag}(n_1/n, n_2/n)$, $B'$ is $p \times p$ with entry $(j,k)$ equal to $n^{-1} \sum I(\delta_i = 3) h_j(Z_i) h_k(Z_i) \exp\{\beta h(Z_i)\} / \left[1 + \exp\{\beta h(Z_i)\}\right]^2$,

$$C = \begin{bmatrix} 0 & 0 \\ \pi & -1 + \pi \end{bmatrix},$$

$$D = \begin{bmatrix} -\sum_{j=1}^{k} h_1(Z_j) \dfrac{\exp\{\beta h(Z_i)\}}{\left[1 + \exp\{\beta h(Z_i)\}\right]^2} \gamma_j & \cdots & -\sum_{j=1}^{k} h_p(Z_j) \dfrac{\exp\{\beta h(Z_i)\}}{\left[1 + \exp\{\beta h(Z_i)\}\right]^2} \gamma_j \\ 0 & \cdots & 0 \end{bmatrix}$$

is $2 \times p$, and

$$E = \begin{bmatrix} 1 & 0 \\ -1 & \sigma_e + \sigma_p - 1 \end{bmatrix}.$$

Let $\mathbb{A}(\theta_m)$ have the same form as $\mathbb{A}_n(X, Z, \delta, \theta_m)$, replacing $A'$ with $A = \text{diag}(c_1, c_2)$ and $B'$ with $B$, with entry $(j,k)$ of $B$ equal to $c_3 \mathbb{E}\left(h_j(Z) h_k(Z) \exp\{\beta h(Z_i)\} / \left[1 + \exp\{\beta h(Z_i)\}\right]^2\right)$. Note that $\mathbb{A}_n(X, Z, \delta, \theta_m) \to_p \mathbb{A}(\theta_m)$ by the Weak Law of Large Numbers.

Second, define

$$\mathbb{B}_n(X, Z, \delta, \theta_m) = n^{-1} \sum \mathbb{E}(\psi(X_i, Z_i; \delta_i, \theta_m) \psi(X_i, Z_i; \delta_i, \theta_m)^T) = \begin{bmatrix} F' & 0_{2\times p} & 0_{2\times 2} \\ 0_{p\times 2} & G' & 0_{p\times 2} \\ 0_{2\times 2} & 0_{2\times p} & 0_{2\times 2} \end{bmatrix}$$

where $F' = \text{diag}\left\{n_1 \sigma_e (1 - \sigma_e)/n, n_2 \sigma_p (1 - \sigma_p)/n\right\}$ and $G'$ is a $p \times p$ matrix with entry $(j,k)$ equal to $(n_3/n)\mathbb{E}\left(h_j(Z) h_k(Z) \left[X - \text{logit}^{-1}\{\beta h(Z)\}\right]^2\right)$. Let $\mathbb{B}(\theta_m)$ be of the same form as $\mathbb{B}_n(X, Z, \delta, \theta_m)$ except with $F$ and $G$ replacing $F'$ and $G'$, where $F = \text{diag}\{c_1 \sigma_e (1 - \sigma_e), c_2 \sigma_p (1 - \sigma_p)\}$ and $G$ is identical to $G'$ except with $c_3$ replacing $n_3/n$ in each element. Noting that $\mathbb{B}(\theta_m) = \lim_{n\to\infty} \mathbb{B}_n(X, Z, \delta, \theta_m)$, it follows that $\sqrt{n}\left\{n^{-1} \sum \psi(X_i, Z_i; \delta_i, \theta_m)\right\} \to_d \mathcal{N}\left\{0, \mathbb{B}(\theta_m)\right\}$ by the Lindeberg-Feller CLT. Specifically, note that the range of the inverse logit function is $[0, 1]$ and that all elements of $h(Z)$, the user-specified function of the covariates, must be finite. Then the Lindeberg condition holds by the same logic as in Appendices A.2.1 and A.3.1.

Finally, $\sqrt{n}R^* \to_p 0$ can be proved as in Appendix A.3.1. Letting $\theta_{m_j}$ denote the $j$th entry of $\theta_m$, the proof follows from the fact that $|\partial^2\psi_j(X_i, Z_i; \delta_i, \theta_m)/\partial\theta_{m_j}\partial\theta_{m_k}| < \infty$ for all $j, k \in \{1, 2, \ldots, p+4\}$. Therefore, by Slutsky's Theorem

$$\sqrt{n}(\hat{\theta}_m - \theta_m) \to_d \mathcal{N}\left\{0, \mathbb{A}(\theta_m)^{-1}\mathbb{B}(\theta_m)\mathbb{A}(\theta_m)^{-T}\right\} \text{ and } \sqrt{n}(\hat{\pi} - \pi) \to_d \mathcal{N}(0, V_{\pi,SRGM})$$

where the bottom right element of $\mathbb{A}(\theta_m)^{-1}\mathbb{B}(\theta_m)\mathbb{A}(\theta_m)^{-T}$ is denoted as $V_{\pi,SRGM}$. The asymptotic variance of $\hat{\theta}_m$ can be consistently estimated by the empirical sandwich variance estimator $\mathbb{A}(\hat{\theta}_m)^{-1}\mathbb{B}(\hat{\theta}_m)\mathbb{A}(\hat{\theta}_m)^{-T}$; the bottom right element of the sandwich estimator is $\hat{V}_{\pi,SRGM}$.

## A.5 Supplementary figures

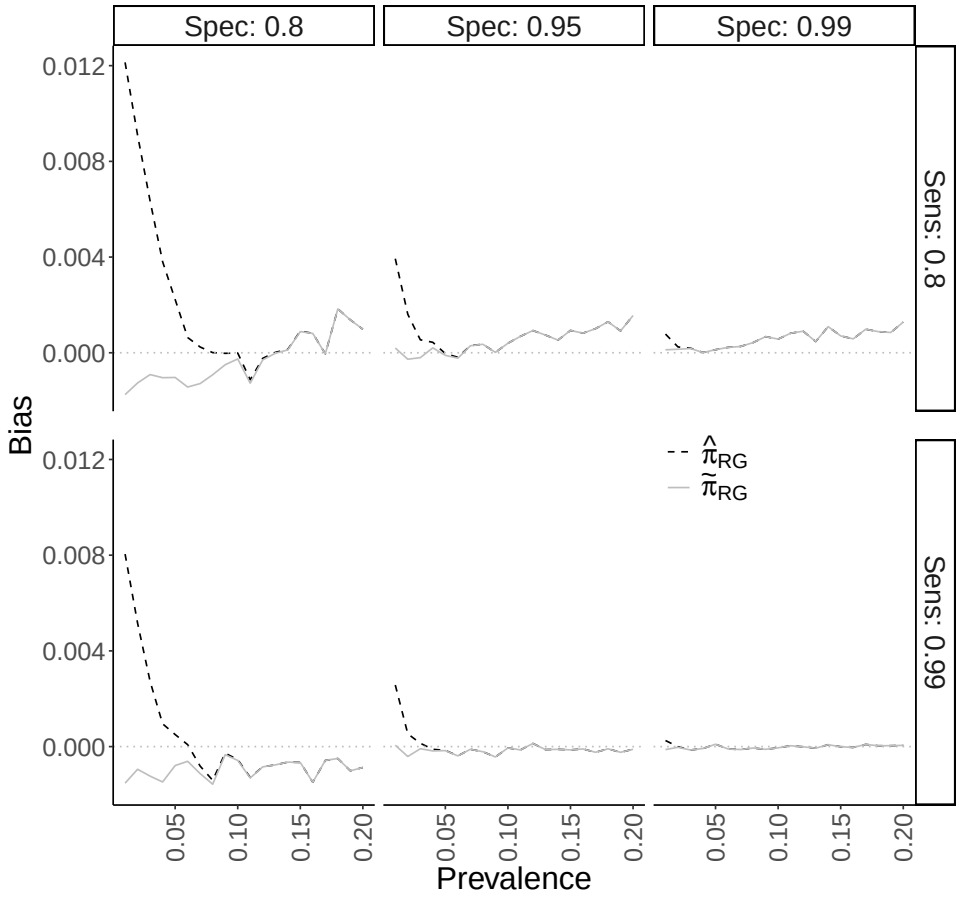


Figure A.1: Empirical bias of the Rogan-Gladen ($\hat{\pi}_{RG}$) estimator and the non-truncated Rogan-Gladen estimator ($\tilde{\pi}_{RG}$) from simulation study for DGP 1, described in Section 2.4. The six facets correspond to a given combination of sensitivity $\sigma_e$ ('Sens') and specificity $\sigma_p$ ('Spec'). 10,000 simulations were conducted for this scenario.

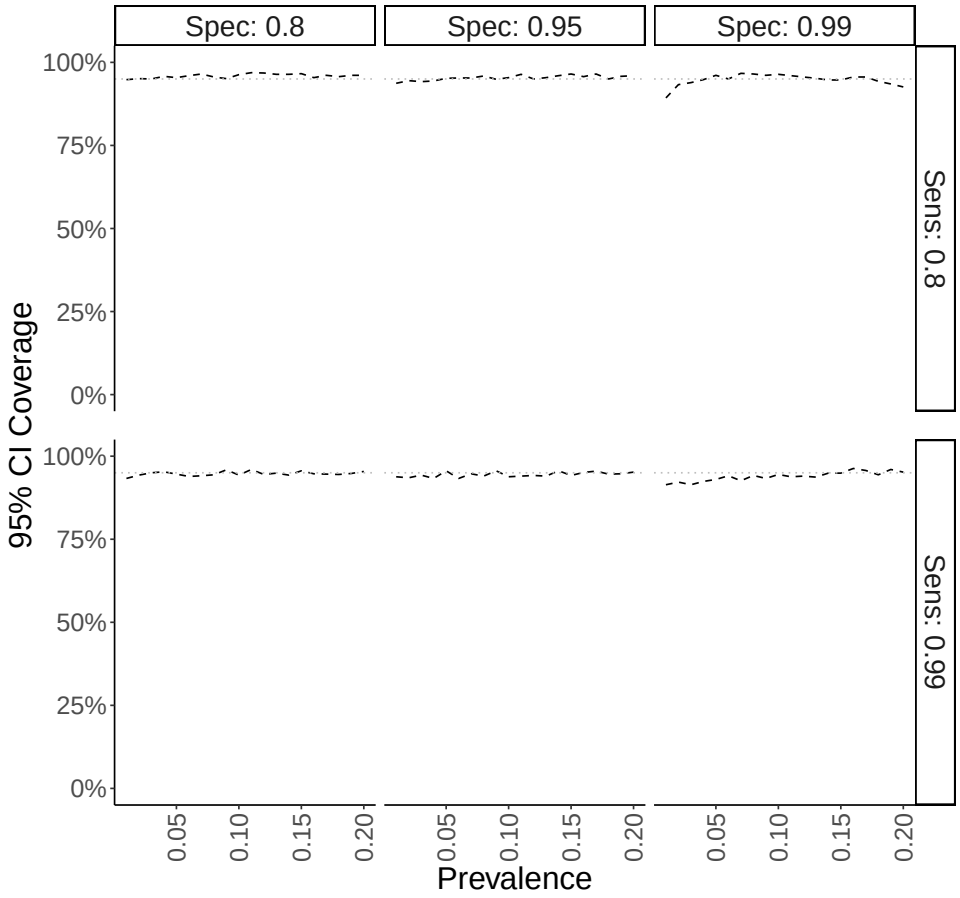


Figure A.2: Confidence interval coverage of the Rogan-Gladen ($\hat{\pi}_{RG}$) estimator from simulation study for DGP 1, described in Section 2.4.1. 10,000 simulations were conducted for this scenario.

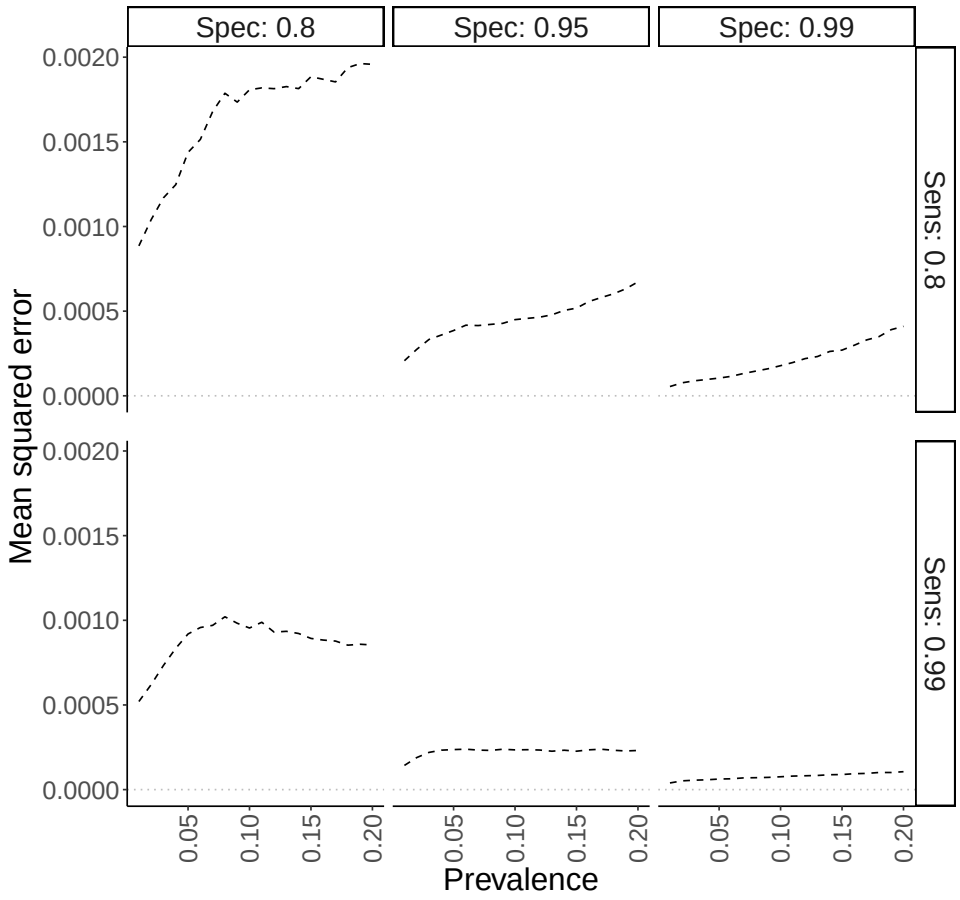


Figure A.3: Mean squared error of the Rogan-Gladen ($\hat{\pi}_{RG}$) estimator from simulation study for DGP 1, described in Section 2.4.1. 10,000 simulations were conducted for this scenario.

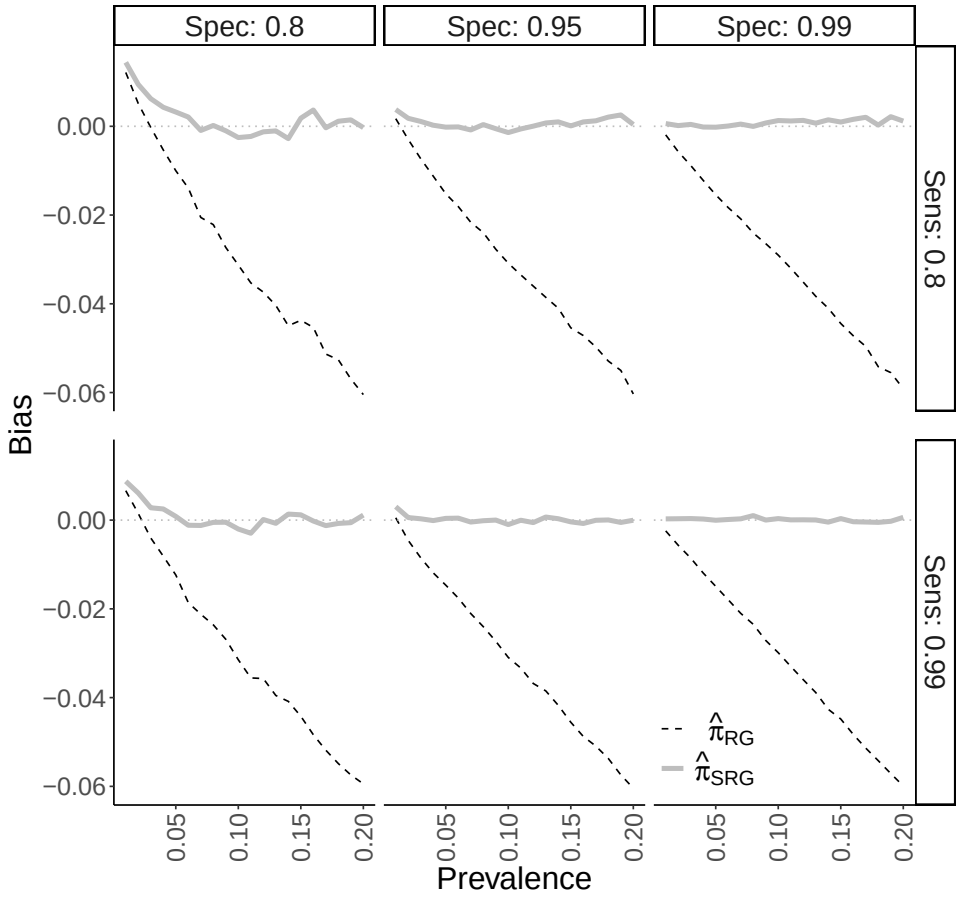


Figure A.4: Empirical bias of the Rogan-Gladen ($\hat{\pi}_{RG}$) and nonparametric standardised ($\hat{\pi}_{SRG}$) estimators from simulation study for DGP 2, described in Section 2.4.2.

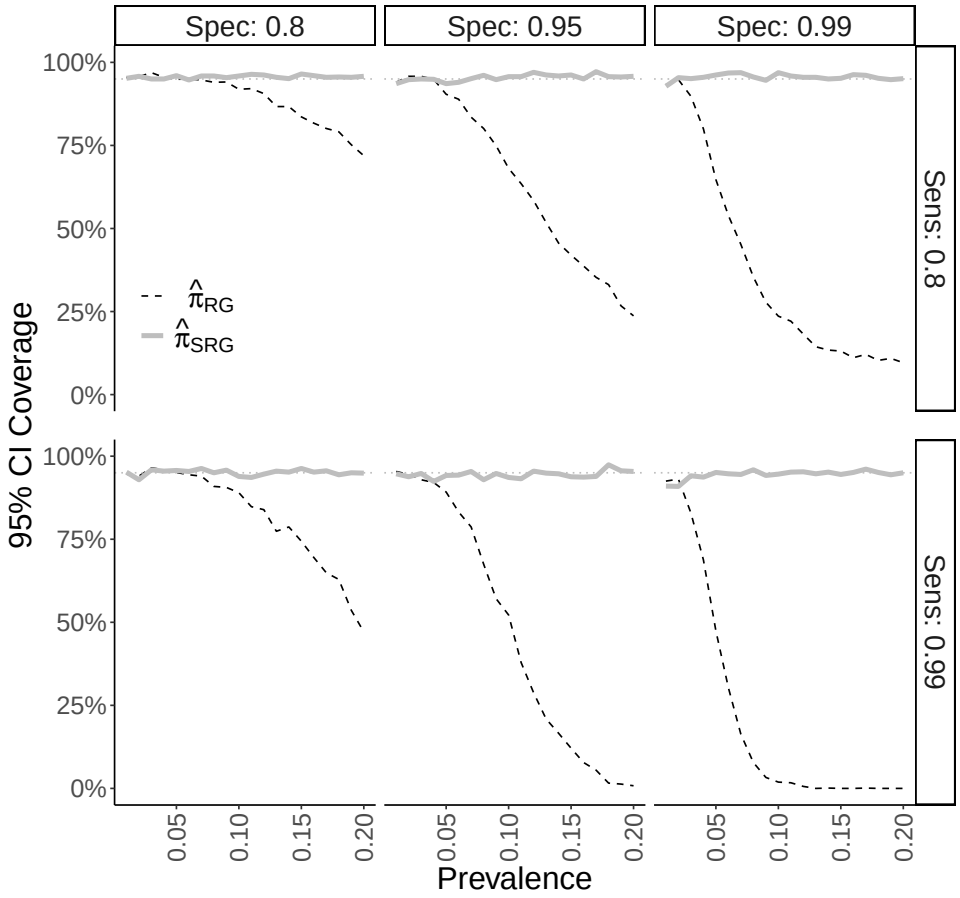


Figure A.5: Confidence interval coverage of the Rogan-Gladen ($\hat{\pi}_{RG}$) and nonparametric standardised ($\hat{\pi}_{SRG}$) estimators from simulation study for DGP 2, described in Section 2.4.2.

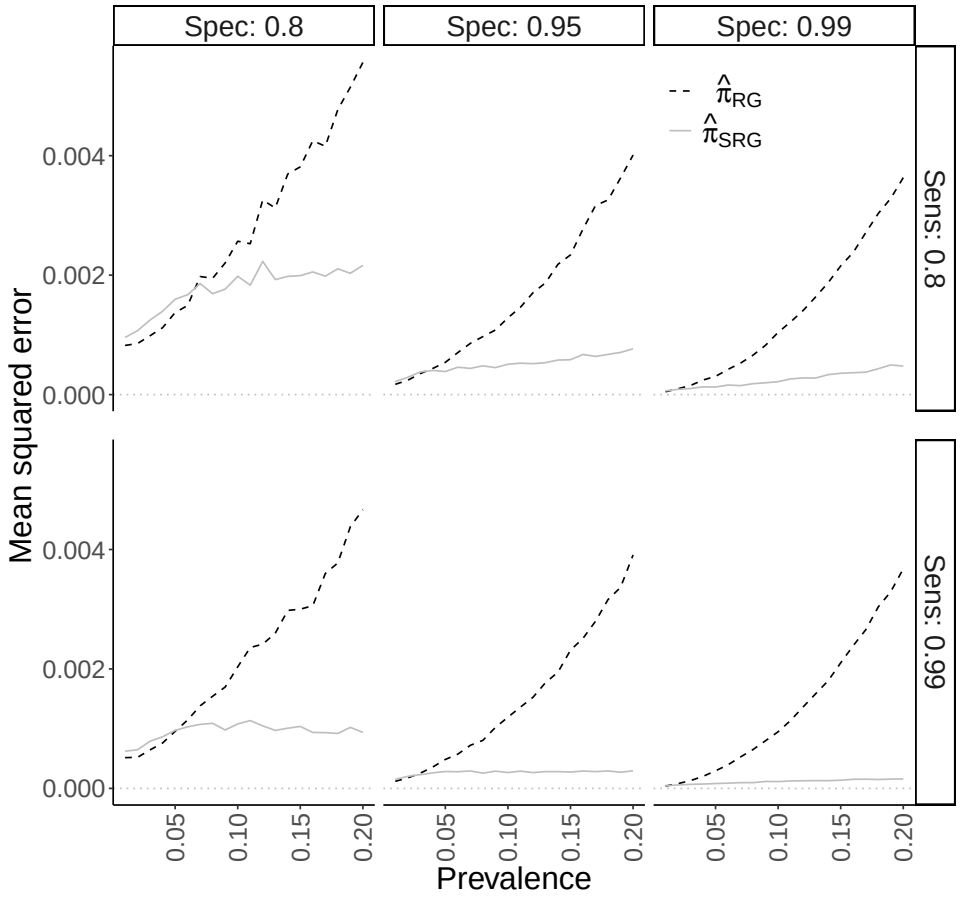


Figure A.6: Mean squared error of the Rogan-Gladen ($\hat{\pi}_{RG}$) and nonparametric standardised ($\hat{\pi}_{SRG}$) estimators from simulation study for DGP 2, described in Section 2.4.2.

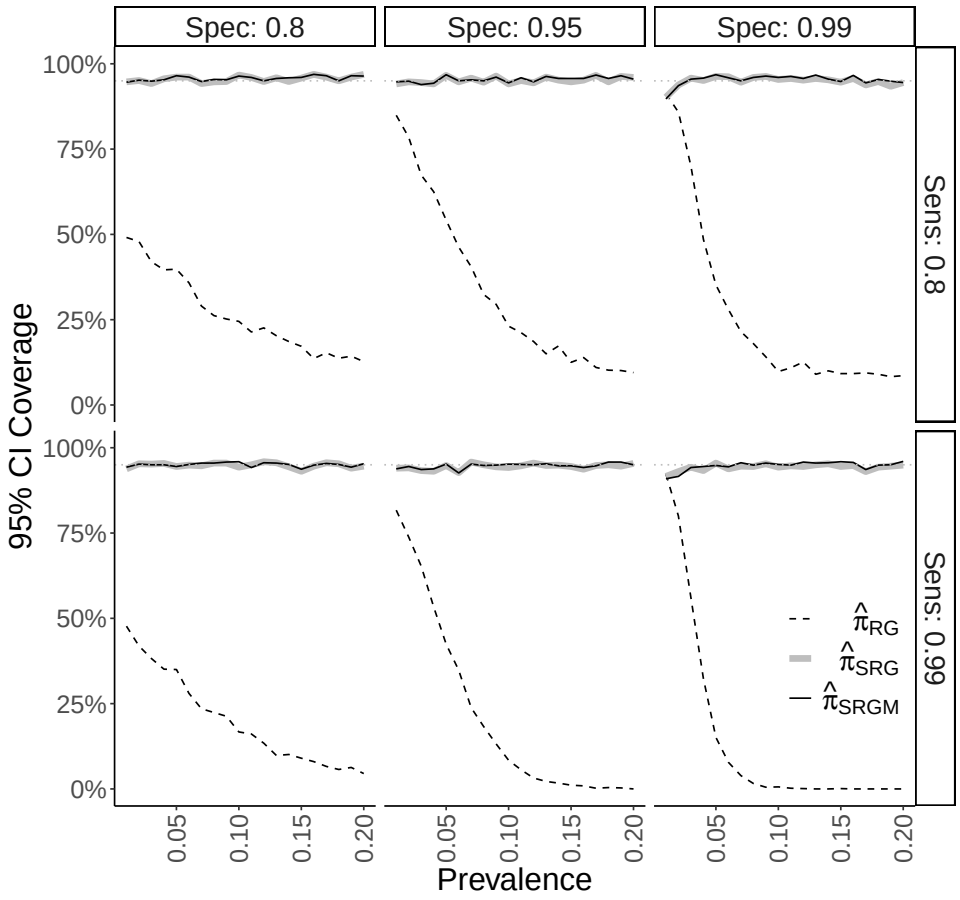


Figure A.7: Confidence interval coverage of the Rogan-Gladen ($\hat{\pi}_{RG}$), nonparametric standardised ($\hat{\pi}_{SRG}$), and parametric standardised ($\hat{\pi}_{SRGM}$) estimators from simulation study for DGP 3, described in Section 2.4.3.1.

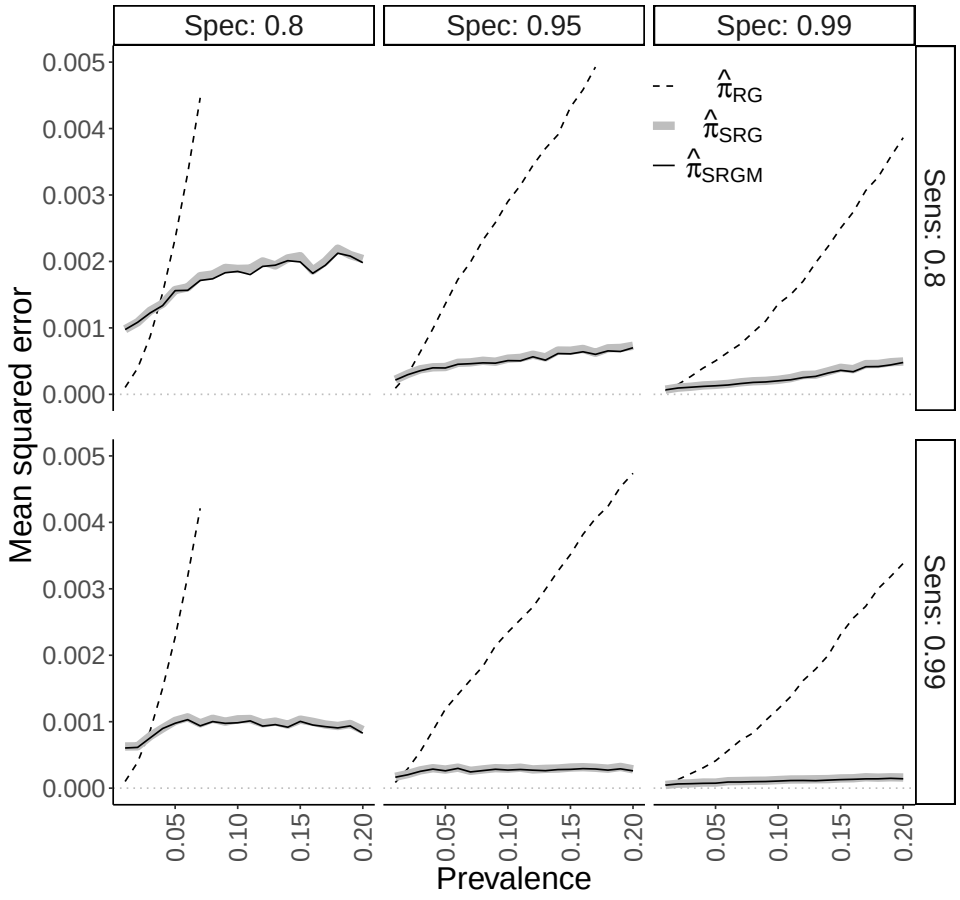


Figure A.8: Mean squared error of the Rogan-Gladen ($\hat{\pi}_{RG}$), nonparametric standardised ($\hat{\pi}_{SRG}$), and parametric standardised ($\hat{\pi}_{SRGM}$) estimators from simulation study for DGP 3, described in Section 2.4.3.1. The $y$-axis is truncated at 0.005 for ease of distinguishing $\hat{\pi}_{SRG}$ and $\hat{\pi}_{SRGM}$.

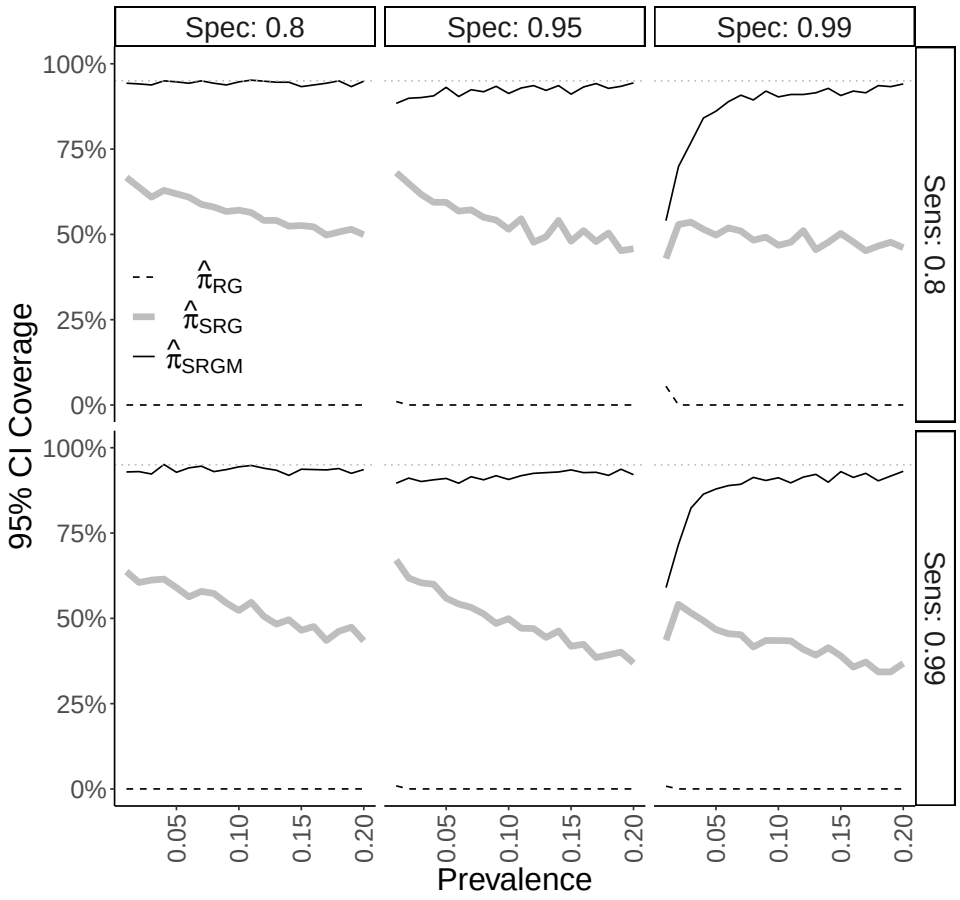


Figure A.9: Confidence interval coverage of the Rogan-Gladen ($\hat{\pi}_{RG}$), nonparametric standardised ($\hat{\pi}_{SRG}$), and parametric standardised ($\hat{\pi}_{SRGM}$) estimators from simulation study for DGP 4, described in Section 2.4.3.2.

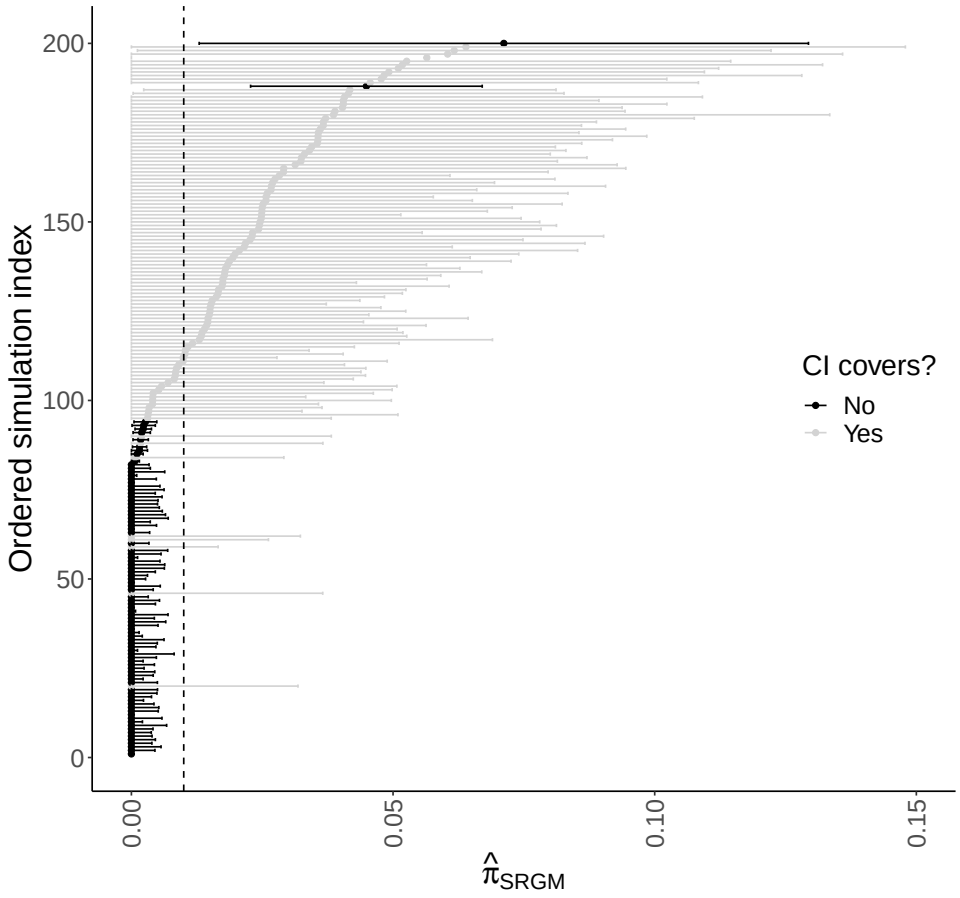


Figure A.10: A random sample of 200 point estimates $\hat{\pi}_{SRGM}$ and 95% confidence interval estimates based on $\hat{V}_{\pi,SRGM}$ from DGP 4, where the data were generated with $\sigma_e = \sigma_p = .99$ and $\pi = .01$. 59% of the intervals covered the true value of $\pi = .01$. The $x$-axis is truncated at 0.015 for visibility of the estimates.

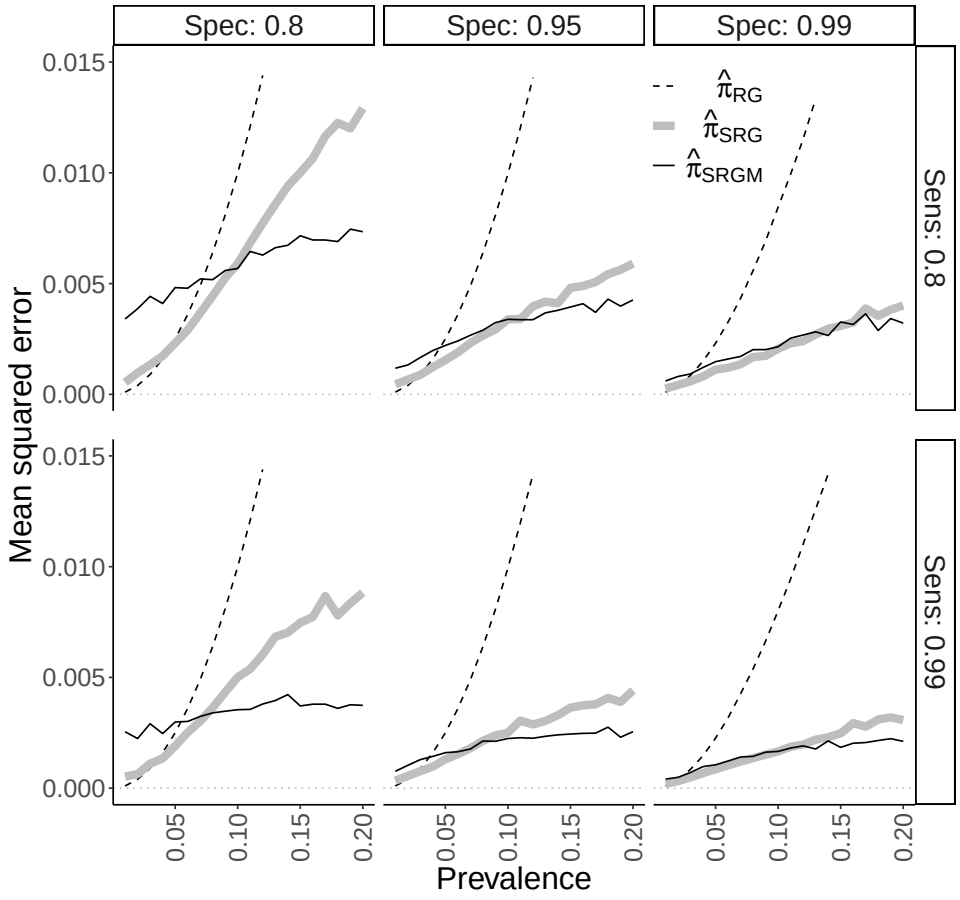


Figure A.11: Mean squared error of the Rogan-Gladen ($\hat{\pi}_{RG}$), nonparametric standardised ($\hat{\pi}_{SRG}$), and parametric standardised ($\hat{\pi}_{SRGM}$) estimators from simulation study for DGP 4, described in Section 2.4.3.2. The $y$-axis is truncated at 0.015 for ease of distinguishing $\hat{\pi}_{SRG}$ and $\hat{\pi}_{SRGM}$.

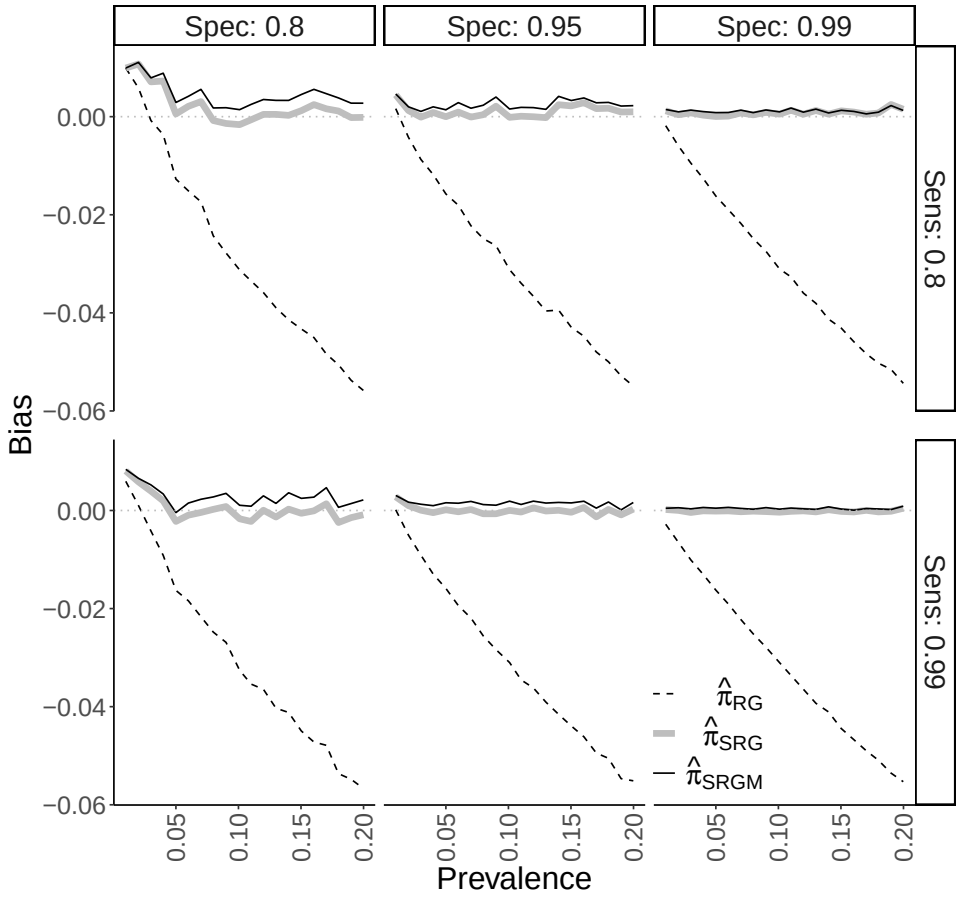


Figure A.12: Empirical bias of the estimators from simulation study for DGP 3 under model misspecification, described in Section 2.4.4.

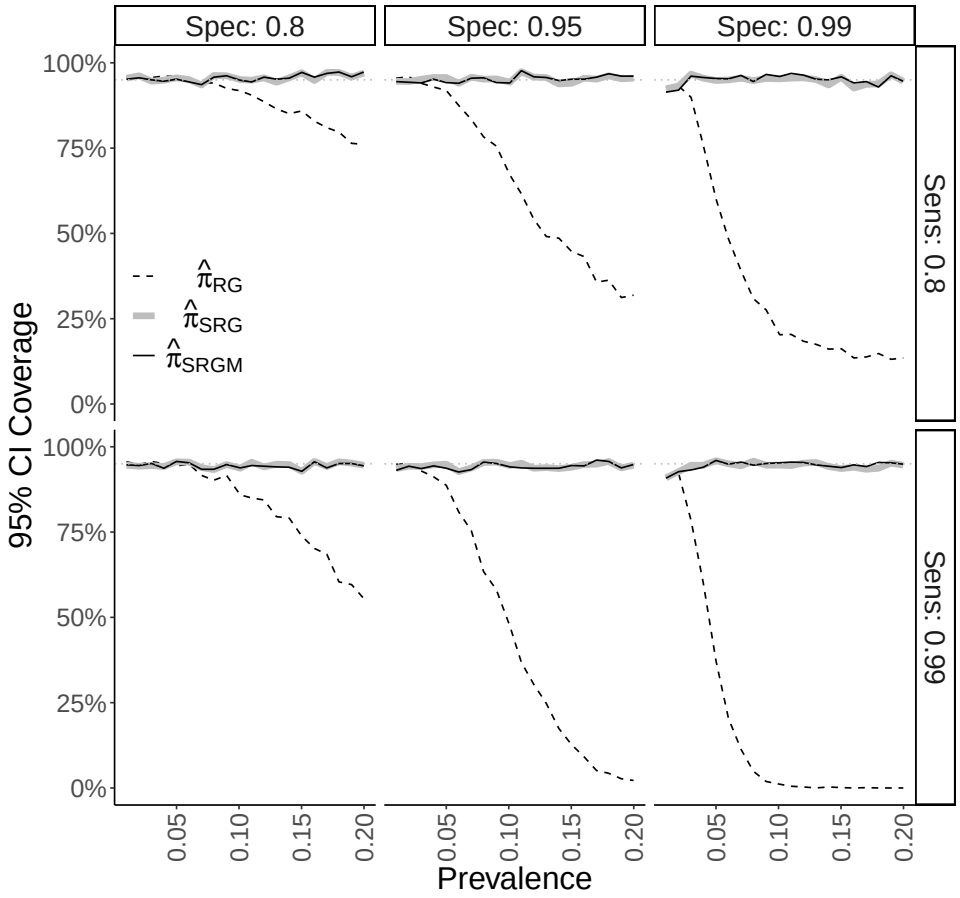


Figure A.13: Confidence interval coverage of the estimators from simulation study for DGP 3 under model misspecification, described in Section 2.4.4.

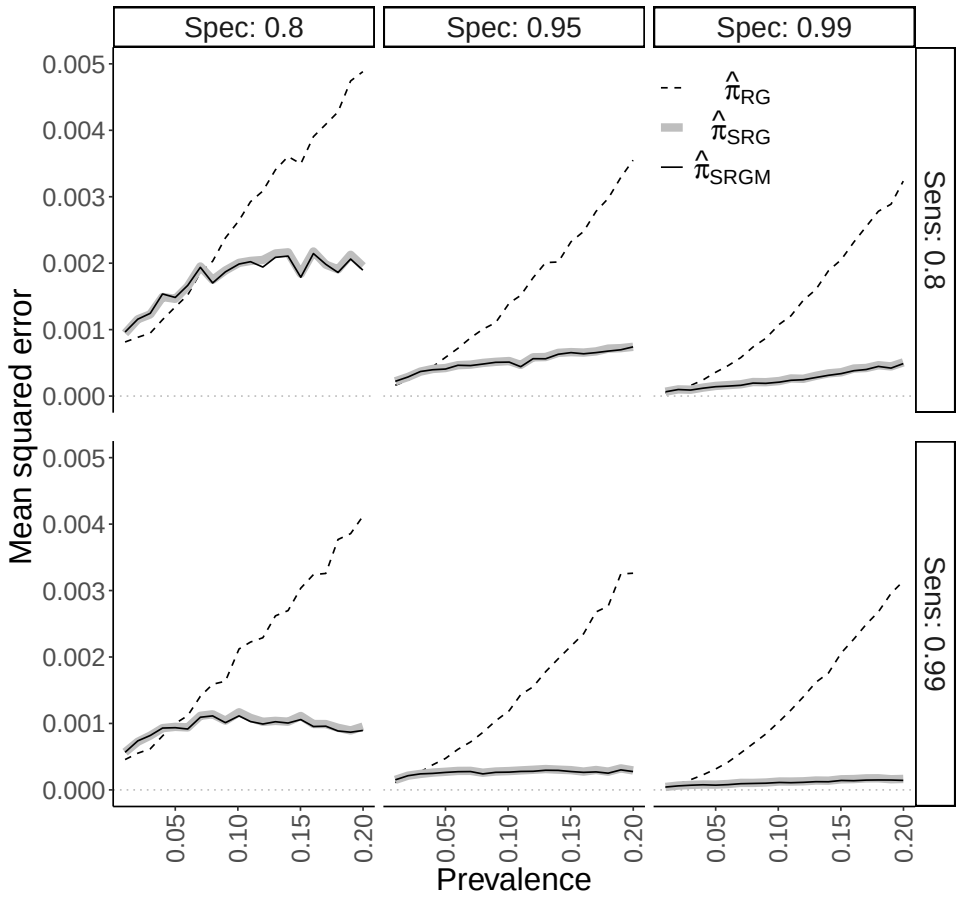


Figure A.14: Mean squared error of the Rogan-Gladen ($\hat{\pi}_{RG}$), nonparametric standardised ($\hat{\pi}_{SRG}$), and parametric standardised ($\hat{\pi}_{SRGM}$) estimators from simulation study for DGP 3 under model misspecification, described in Section 2.4.4. The $y$-axis is truncated at 0.005 for ease of distinguishing $\hat{\pi}_{SRG}$ and $\hat{\pi}_{SRGM}$.

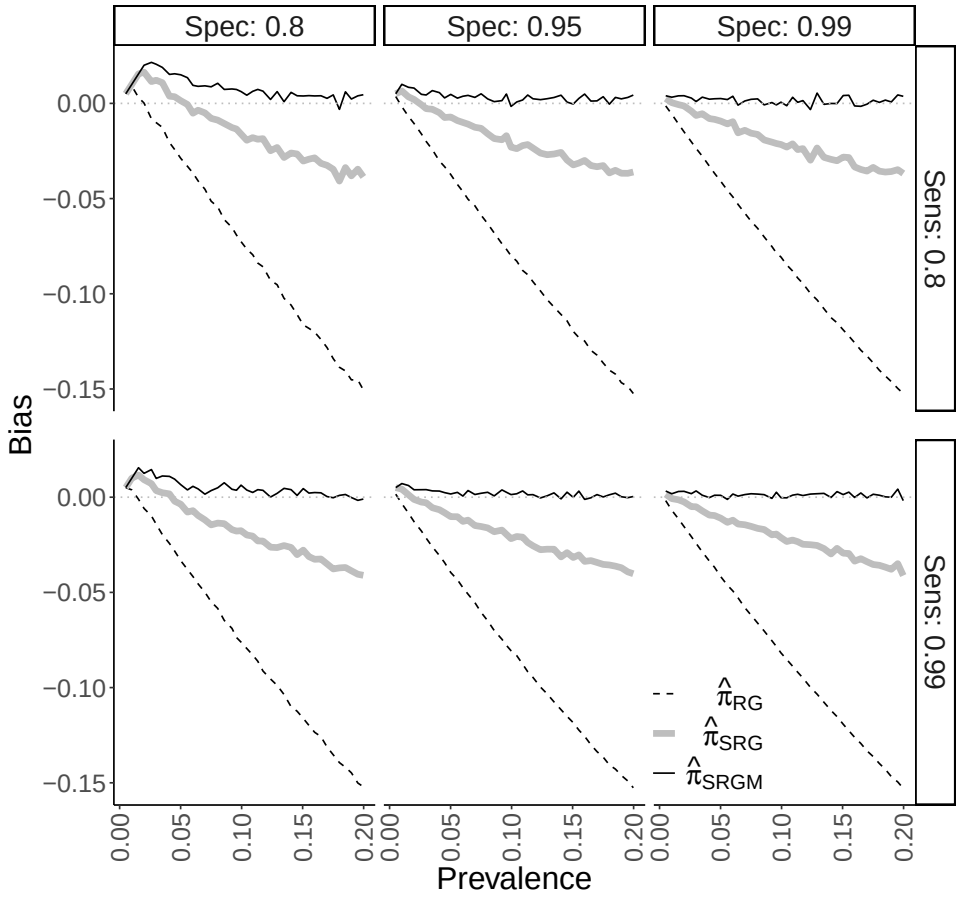


Figure A.15: Empirical bias of the estimators from simulation study for DGP 4 under model misspecification, described in Section 2.4.4.

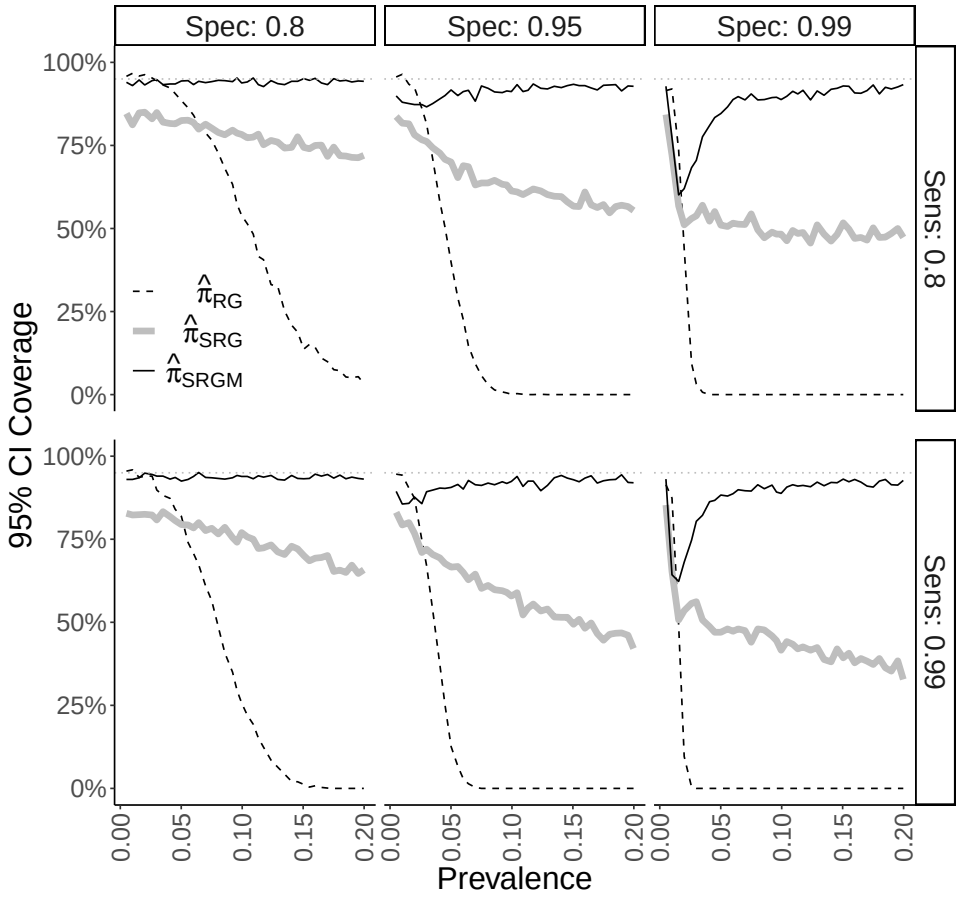


Figure A.16: Confidence interval coverage of the estimators from simulation study for DGP 4 under model misspecification, described in Section 2.4.4.

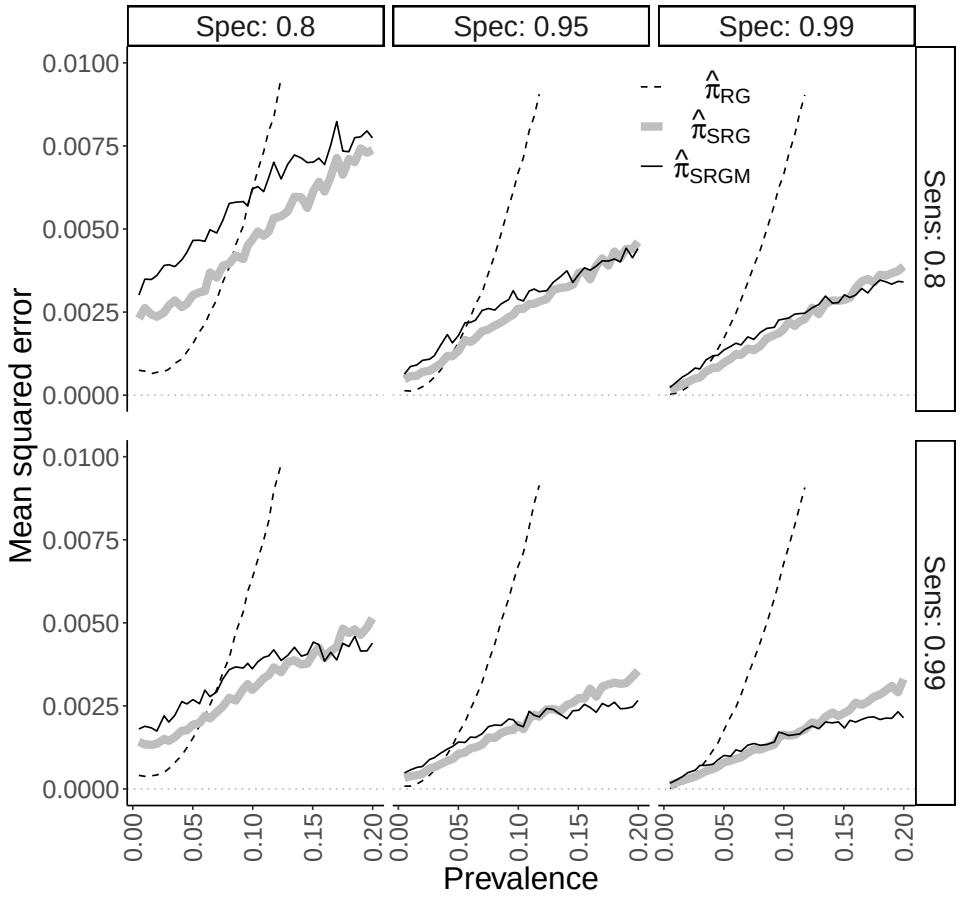


Figure A.17: Mean squared error of the Rogan-Gladen ($\hat{\pi}_{RG}$), nonparametric standardised ($\hat{\pi}_{SRG}$), and parametric standardised ($\hat{\pi}_{SRGM}$) estimators from simulation study for DGP 4 under model misspecification, described in Section 2.4.4. The $y$-axis is truncated at 0.01 for ease of distinguishing $\hat{\pi}_{SRG}$ and $\hat{\pi}_{SRGM}$.

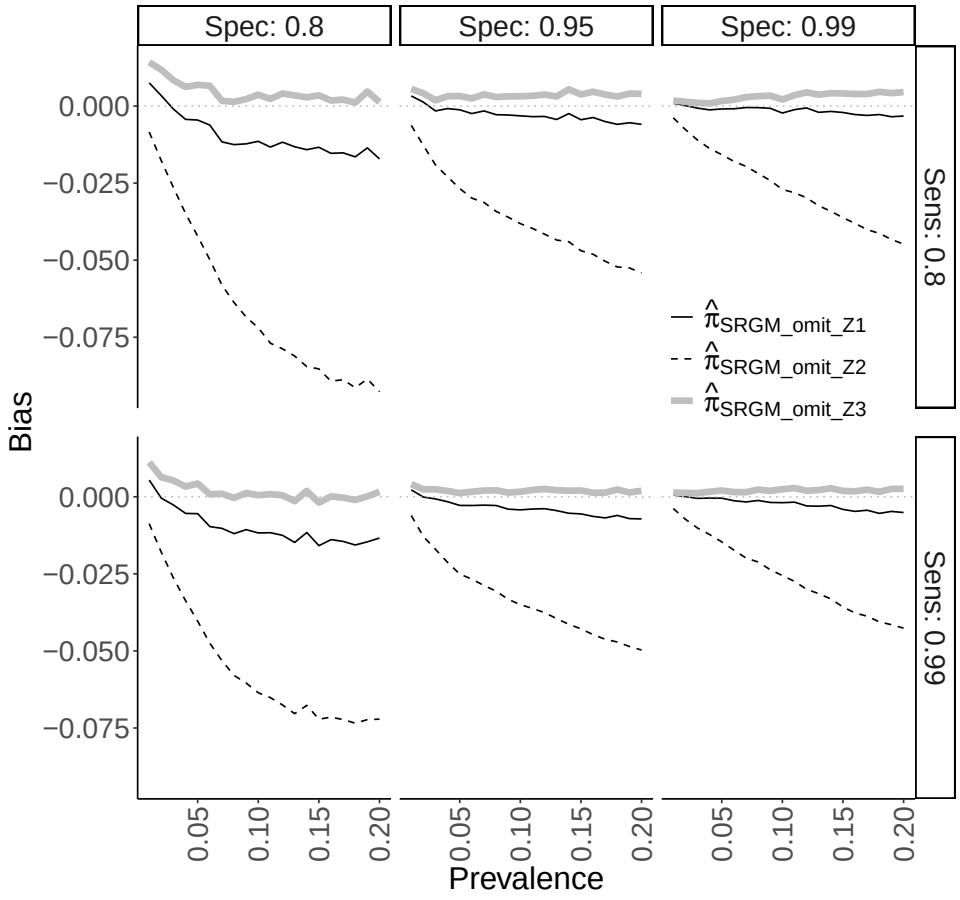


Figure A.18: Empirical bias in simulation study for DGP 3 with the model-based estimator $\hat{\pi}_{SRGM}$ misspecified by omitting $Z_1$, $Z_2$, and $Z_3$, respectively, described in Section 2.4.4.

## APPENDIX B: TECHNICAL DETAILS FOR CHAPTER 3

This appendix contains material supplemental to the main text of Chapter 3.

### B.1 Identification of target potential outcome means

For all treatments $a \in \mathscr{A}$, $E[O^a \mid R = 1] = E[O^a \mid T = 1]$ by the assumption of random sampling from the target population.

For treatments $a \in \mathscr{A}_{-0}$, $T$ is a function of $A$, i.e., $A = a$ implies $T = a$, implying that $Z \perp\!\!\!\perp T \mid A$ for any variable $Z$. The target potential outcome means for these treatments equals

$$E[O^a \mid T = 1] = E[E[O^a \mid T = 1, X] \mid T = 1] \tag{B.1}$$

$$= E[E[O^a \mid T = a, X] \mid T = 1] \tag{B.2}$$

$$= E[E[O^a \mid A = a, T = a, X] \mid T = 1] \tag{B.3}$$

$$= E[E[O^a \mid A = a, X] \mid T = 1] \tag{B.4}$$

$$= E[E[O \mid A = a, X] \mid T = 1]. \tag{B.5}$$

All equalities require assumption 3.1 of no measurement error. The equality (B.1) holds by iterated expectation, (B.2) by assumption 3.6 of conditional exchangeability of trial, (B.3) by assumption 3.4 of conditional exchangeability of treatment within trial, (B.4) because $T$ is a function of $A$ for $a \in \mathscr{A}_{-0}$, and (B.5) by assumption 3.2 of causal consistency. Treatment $A = 1$ is only measured in trial $T = 1$, enabling simplification of (B.5) to $E[O \mid A = 1, X]$.

For placebo $A = 0$,

$$E[O^0 \mid T = 1] = E[E[O^0 \mid T = 1, X] \mid T = 1] \tag{B.6}$$

$$= E[E[O^0 \mid A = 0, T = 1, X] \mid T = 1] \tag{B.7}$$

$$= E[E[O^0 \mid A = 0, X] \mid T = 1] \tag{B.8}$$

$$= E[E[O \mid A = 0, X] \mid T = 1] \tag{B.9}$$

with (B.6) holding by iterated expectation, (B.7) by assumption 3.4 of conditional exchangeability of treatment within trial, (B.8) by assumption 3.8 of conditional exchangeability of trial for placebo, and (B.9) by assumption 3.2 of causal consistency.

## B.2 Asymptotic distributions of causal association estimators

### B.2.1 Asymptotic distribution of $\hat{\rho}$

Let $\mu_a = E[Y^a \mid R = 1]$ and $\lambda_a = E[S^a \mid R = 1]$ for $a \in \mathscr{A}$, and denote $\boldsymbol{\mu} = (\mu_0, \mu_1, \ldots, \mu_m)$ and $\boldsymbol{\lambda} = (\lambda_0, \lambda_1, \ldots, \lambda_m)$. Denote $\boldsymbol{\theta} = (\theta_0, \theta_2, \ldots, \theta_m)$ and $\boldsymbol{\eta} = (\eta_0, \eta_2, \ldots, \eta_m)$ where each $\theta_a$ and $\eta_a$ are regression parameters for outcome models for $Y$ and for $S$ respectively, as in Section 3.2.3. Let $\boldsymbol{\zeta} = (\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\mu}, \boldsymbol{\lambda}, \rho)$ for a general causal association parameter $\rho = f(\boldsymbol{\tau}, \boldsymbol{\delta})$. The estimator $\hat{\boldsymbol{\zeta}}$ uses the estimators for component parameter $\theta_a$, $\eta_a$, $\mu_a$, $\lambda_a$, and $\rho$ defined in the main text. The estimator $\hat{\boldsymbol{\zeta}}$ is the solution to the vector of estimating equations $\sum \psi(Y_i, S_i, X_i, A_i, R_i; \boldsymbol{\zeta}) = 0$ where

$$\psi(Y, S, X, A, R; \boldsymbol{\zeta}) = \begin{pmatrix} \psi_\theta(Y, X, A, R = 0; \boldsymbol{\theta}) \\ \psi_\eta(Y, X, A, R = 0; \boldsymbol{\eta}) \\ \psi_\mu(Y, X, A, R = 1; \boldsymbol{\theta}, \boldsymbol{\mu}) \\ I(A = 1)(Y - \mu_1) \\ \psi_\lambda(S, X, A, R = 1; \boldsymbol{\eta}, \boldsymbol{\lambda}) \\ I(A = 1)(S - \lambda_1) \\ \psi_\rho(\boldsymbol{\tau}, \boldsymbol{\delta}; \boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\mu}, \boldsymbol{\lambda}) \end{pmatrix} = 0$$

with $\psi_\theta = (\psi_{\theta_0}, \psi_{\theta_2}, \ldots, \psi_{\theta_m})$ the vector of score equations for the $m$ regressions of $Y$ on $X$, $\psi_\eta = (\psi_{\eta_0}, \psi_{\eta_2}, \ldots, \psi_{\eta_m})$ the vector of score equations for the $m$ regressions of $S$ on $X$, $\psi_\mu = (\psi_{\mu_0}, \psi_{\mu_2}, \ldots, \psi_{\mu_m})$ with $\psi_{\mu_a} = I(R = 1)\left(\hat{E}[Y \mid X = x, A = a; \theta_a] - \mu_a\right)$ and $\psi_\lambda = (\psi_{\lambda_0}, \psi_{\lambda_2}, \ldots, \psi_{\lambda_m})$ with $\psi_{\lambda_a} = I(R = 1)\left(\hat{E}[S \mid X = x, A = a; \eta_a] - \lambda_a\right)$ using the predicted values from the regressions. The final estimating equation $\psi_\rho(\boldsymbol{\tau}, \boldsymbol{\delta})$ depends on the choice of causal association parameter $\rho$. For the linear regression slope $\beta$ considered in Section 3.2.4, $\psi_\rho = \psi_\beta(\boldsymbol{\tau}, \boldsymbol{\delta}) = \mathrm{Cov}(\hat{\boldsymbol{\tau}}, \hat{\boldsymbol{\delta}}) / \mathrm{Var}(\hat{\boldsymbol{\delta}}) - \beta$ where Cov is defined to be the sample covariance between

the estimates $\hat{\boldsymbol{\tau}}$ and $\hat{\boldsymbol{\delta}}$ and Var the sample variance. For the Pearson correlation $\rho_p$ considered in Section 3.2.4, $\psi_\rho = \psi_{\rho_p}(\boldsymbol{\tau}, \boldsymbol{\delta}) = \mathrm{Cov}(\hat{\boldsymbol{\tau}}, \hat{\boldsymbol{\delta}})/\{\mathrm{SD}(\hat{\boldsymbol{\tau}})\,\mathrm{SD}(\hat{\boldsymbol{\delta}})\} - \rho_p$.

For specific choices of the causal association parameter $\rho$ and the final estimating function $\psi_\rho$, such as the two choices $\beta$ and $\rho_p$ described above, the estimator $\hat{\boldsymbol{\zeta}}$ is strongly consistent and asymptotically normal, as shown in Sections 3.2.3 and 3.2.4. Specifically, for either of these choices of $\rho$,

$$\sqrt{n}(\hat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}) \xrightarrow{d} \mathcal{N}(0, \Sigma) \text{ and } \sqrt{n}(\hat{\rho} - \rho) \xrightarrow{d} \mathcal{N}(0, V_\rho),$$

where the asymptotic variance of $\rho$ is $V_\rho$, the lower right element of $\Sigma$, with $\Sigma = A(\boldsymbol{\zeta})^{-1}B(\boldsymbol{\zeta})\left[A(\boldsymbol{\zeta})^{-1}\right]^T$, $A(\boldsymbol{\zeta}) = E[-\partial\psi(\boldsymbol{\zeta})/\partial\boldsymbol{\zeta}^T]$, and $B(\boldsymbol{\zeta}) = E[\psi(\boldsymbol{\zeta})\psi(\boldsymbol{\zeta})^T]$.

### B.2.2 Asymptotic distribution of $(\hat{\tau}^a, \hat{\delta}^a)$

An estimating equations approach can be used to show the asymptotic normality of $\hat{\boldsymbol{\tau}}$ and of $\hat{\boldsymbol{\delta}}$ and the asymptotic bivariate normality of each $(\hat{\tau}^a, \hat{\delta}^a)$ pair. Denote $\boldsymbol{\zeta}_{\tau,\delta} = (\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\mu}, \boldsymbol{\lambda}, \boldsymbol{\tau}, \boldsymbol{\delta})$ where $\boldsymbol{\theta}$, $\boldsymbol{\eta}$, $\boldsymbol{\mu}$, $\boldsymbol{\lambda}$ are defined in Appendix B.2.1. The estimator $\hat{\boldsymbol{\zeta}}_{\tau,\delta}$ is the solution to the vector of estimating equations $\sum \psi(Y_i, S_i, X_i, A_i, R_i; \boldsymbol{\zeta}_{\tau,\delta}) = 0$ which replaces the last element $\psi_\rho$ from the vector of estimating equations $\sum \psi(Y_i, S_i, X_i, A_i, R_i; \boldsymbol{\zeta})$ from Appendix B.2.1 with the elements $\psi_\tau(Y, X, A, R; \boldsymbol{\theta}, \boldsymbol{\mu}, \boldsymbol{\tau})$ and $\psi_\delta(S, X, A, R; \boldsymbol{\eta}, \boldsymbol{\lambda}, \boldsymbol{\delta})$. The vector $\psi_\tau = (\psi_{\tau_1}, \ldots, \psi_{\tau_m})$ with $\psi_{\tau_a} = (1 - \mu_a/\mu_0) - \tau_a$ and the vector $\psi_\delta = (\psi_{\delta_1}, \ldots, \psi_{\delta_m})$ with $\psi_{\delta_a} = (\lambda_a - \lambda_0) - \delta_a$. By the continuity of the $g$ and $h$ functions that define $\tau^a$ and $\delta^a$ (as defined in Section 3.2.1 of the main text), each $\tau^a$ and $\delta^a$ is strongly consistent and asymptotically normal. Thus $\hat{\boldsymbol{\zeta}}_{\tau,\delta}$ is also strongly consistent and asymptotically normal. As a consequence, each $(\hat{\tau}^a, \hat{\delta}^a)$ pair is asymptotically bivariate normal.

## B.3 Simulation settings

| Scenario | $n_t$ | $a$ | $\tau_a$ | $\delta_a$ | $\theta_{a0}$ | $\theta_{a1}$ | $\theta_{a2}$ | $\mu_{a0}$ | $\mu_{a1}$ | $\mu_{a2}$ | $\sigma_a^2$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 3000 | 0 | – | – | -6 | 1.5 | .1 | -5.5 | 1.5 | .1 | $0.64^2$ |
| | | 1 | 0.81 | 3.13 | -8.18 | 1.5 | .082 | -1.135 | 1 | .08 | $0.28^2$ |
| | | 2 | 0.96 | 3.56 | -10 | 1.5 | .028 | -.908 | 1.5 | .081 | $0.64^2$ |
| | | 3 | 0.92 | 3.48 | -9.5 | 1 | .0808 | -1.33 | 1 | .09 | $0.64^2$ |
| | | 4 | 0.94 | 3.04 | -10.08 | 1.15 | .09 | -2.356 | 1.15 | .1 | $0.64^2$ |
| | | 5 | 0.88 | 4.30 | -9.36 | 1.5 | .09 | -1.197 | 1.5 | .1 | $0.64^2$ |
| 2 | 3000 | 0 | – | – | -6 | 1.5 | .1 | -5.5 | 1.5 | .1 | $0.64^2$ |
| | | 1 | 0.20 | 1.29 | -6.69 | 1.5 | .1 | -2.971 | 1 | .08 | $0.28^2$ |
| | | 2 | 0.45 | 0.69 | -6.42 | 1.5 | .081 | -3.775 | 1.5 | .081 | $0.64^2$ |
| | | 3 | 0.93 | 4.09 | -9.7 | 1 | .088 | -0.718 | 1 | .09 | $0.64^2$ |
| | | 4 | 0.59 | 2.58 | -8.1 | 1.15 | .1 | -2.818 | 1.15 | .1 | $0.64^2$ |
| | | 5 | 0.86 | 3.83 | -9.52 | 1.5 | .095 | -1.668 | 1.5 | .1 | $0.64^2$ |
| 3 | 8000 | 0 | – | – | 10.38 | 1.5 | 0.1 | -5.849 | 1.5 | 0.1 | $0.64^2$ |
| | | 1 | 0.96 | 4.00 | -12.5 | 1.5 | 0.81 | -0.61 | 1 | 0.08 | $0.28^2$ |
| | | 2 | 0.82 | 3.66 | -10.72 | 1.3 | 0.081 | -1.152 | 1.5 | 0.081 | $0.64^2$ |
| | | 3 | 0.73 | 2.66 | -10.82 | 1 | 0.09 | -2.494 | 1 | 0.09 | $0.64^2$ |
| | | 4 | 0.45 | 1.34 | -10.9 | 1.15 | 0.1 | -4.404 | 1.15 | 0.1 | $0.64^2$ |
| | | 5 | 0.31 | 5.13 | -10.54 | 1.1 | 0.099 | -0.715 | 1.5 | 0.1 | $0.64^2$ |

Table B.1: Parameter values used in the simulation study described in Section 3.3. The true values of $\tau_a$ and $\delta_a$ for $a \in \{1, 2, 3, 4, 5\}$ were determined empirically based on the average potential outcomes from 20,000,000 participants in trial $T = 1$, the random sample from the target population, and are presented rounded to two decimal places. The parameter values $\theta_{aj}$ and $\mu_{aj}$ for $a \in \mathscr{A}$ and $j \in \{0, 1, 2\}$ were simulation inputs. Dashes indicate entries that are not applicable.

| $a$ | $t$ | $\theta_{0t0}$ | $\theta_{0t1}$ | $\theta_{0t2}$ | $\mu_{0t0}$ | $\mu_{0t1}$ | $\mu_{0t2}$ | $\sigma_{0t}^2$ |
|---|---|---|---|---|---|---|---|---|
| 0 | 1 | -6 | 1.5 | .1 | -5.5 | 1.5 | .1 | $0.64^2$ |
| 0 | 2 | -5.6 | 1.5 | .1 | -5.4 | 1.5 | .1 | $0.64^2$ |
| 0 | 3 | -5.8 | 1.5 | .1 | -5.45 | 1.5 | .1 | $0.64^2$ |
| 0 | 4 | -6.2 | 1.5 | .1 | -5.55 | 1.5 | .1 | $0.64^2$ |
| 0 | 5 | -6.4 | 1.5 | .1 | -5.6 | 1.5 | .1 | $0.64^2$ |

Table B.2: Parameter values used in the simulation studies described in Section 3.3, when assumption 3.8 of conditional exchangeability of trial for placebo did not hold.

# BIBLIOGRAPHY


Accorsi, E. K., Qiu, X., Rumpler, E., Kennedy-Shaffer, L., Kahn, R., Joshi, K., Goldstein, E., Stensrud, M. J., Niehus, R., Cevik, M., and Lipsitch, M. (2021). How to detect and reduce potential sources of biases in studies of SARS-CoV-2 and COVID-19. *European Journal of Epidemiology*, 36:179–196.

Addetia, A., Crawford, K. H. D., Dingens, A., Zhu, H., Roychoudhury, P., Huang, M.-L., Jerome, K. R., Bloom, J. D., and Greninger, A. L. (2020). Neutralizing antibodies correlate with protection from SARS-CoV-2 in humans during a fishery vessel outbreak with a high attack rate. *Journal of Clinical Microbiology*, 58(11):e02107–20.

Agresti, A. and Coull, B. A. (1998). Approximate is better than "exact" for interval estimation of binomial proportions. *The American Statistician*, 52(2):119–126.

Albert, J. M., Ioannidis, J. P., Reichelderfer, P., Conway, B., Coombs, R. W., Crane, L., Demasi, R., Dixon, D. O., Flandre, P., Hughes, M. D., et al. (1998). Statistical issues for HIV surrogate endpoints: Point/counterpoint. *Statistics in Medicine*, 17(21):2435–2462.

Alonso, A., Van der Elst, W., Molenberghs, G., Buyse, M., and Burzykowski, T. (2015). On the relationship between the causal-inference and meta-analytic paradigms for the validation of surrogate endpoints. *Biometrics*, 71(1):15–24.

Andrews, D. W. (1999). Estimation when a parameter is on a boundary. *Econometrica*, 67(6):1341–1383.

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

Bajema, K. L., Wiegand, R. E., Cuffe, K., Patel, S. V., Iachan, R., Lim, T., Lee, A., Moyse, D., Havers, F. P., Harding, L., Fry, A. M., Hall, A. J., Martin, K., Biel, M., Deng, Y., Meyer III, W. A., Mathur, M., Kyle, T., Gundlapalli, A. V., Thornburg, N. J., Petersen, L. R., and Edens, C. (2021). Estimated SARS-CoV-2 seroprevalence in the US as of September 2020. *JAMA Internal Medicine*, 181(4):450–460.

Bang, H. and Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973.

Bareinboim, E. and Pearl, J. (2016). Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113(27):7345–7352.

Barzin, A., Schmitz, J. L., Rosin, S., Sirpal, R., Almond, M., Robinette, C., Wells, S., Hudgens, M., Olshan, A., Deen, S., Krejci, P., Quackenbush, E., Chronowski, K., Cornaby, C., Goins, J., Butler, L., Aucoin, J., Boyer, K., Faulk, J., Alston-Johnson, D., Page, C., Zhou, Y., Fiscus, L., Damania, B., Dittmer, D. P., and Peden, D. B. (2020). SARS-CoV-2 seroprevalences among a southern U.S. population indicates limited asymptomatic spread under physical distancing measures. *mBio*, 11(5):e02426–20.

Bayer, D., Fay, M., and Graubard, B. (2023). Confidence intervals for prevalence estimates from complex surveys with imperfect assays. *Statistics in Medicine*, 42(11):1822–1867. In press.

Bickel, P. J. and Kwon, J. (2001). Inference for semiparametric models: Some questions and an answer. *Statistica Sinica*, 11:863–886.

Bishara, A. J. and Hittner, J. B. (2012). Testing the significance of a correlation with nonnormal data: Comparison of Pearson, Spearman, transformation, and resampling approaches. *Psychological Methods*, 17(3):399.

Boos, D. D. and Stefanski, L. A. (2013). *Essential Statistical Inference: Theory and Methods*. Springer, New York, NY.

Borkowf, C. B. (2000). A new nonparametric method for variance estimation and confidence interval construction for Spearman's rank correlation. *Computational Statistics & Data Analysis*, 34(2):219–241.

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

Brazeau, N., Verity, R., Jenks, S., Fu, H., Whittaker, C., Winskill, P., Dorigatti, I., Walker, P., Riley, S., Schnekenberg, R., Heltgebaum, H., Mellan, T., Mishra, S., Unwin, H., Watson, O., Cucunuba, P. Z., Baguelin, M., Whittles, L., Bhatt, S., Ghani, A., Ferguson, N., and Okell, L. (2020). Report 34: COVID-19 infection fatality ratio: Estimates from seroprevalence. Technical report, Imperial College London.

Breskin, A., Cole, S. R., Edwards, J. K., Brookmeyer, R., Eron, J. J., and Adimora, A. A. (2021). Fusion designs and estimators for treatment effects. *Statistics in Medicine*, 40(13):3124–3137.

Brown, L. D., Cai, T. T., and DasGupta, A. (2001). Interval estimation for a binomial proportion. *Statistical Science*, 16(2):101–133.

Bryson, M. C. (1976). The Literary Digest poll: Making of a statistical myth. *The American Statistician*, 30(4):184–185.

Buchanan, A. L., Hudgens, M. G., Cole, S. R., Mollan, K. R., Sax, P. E., Daar, E. S., Adimora, A. A., Eron, J. J., and 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.

Buss, L. F., Prete Jr., C. A., Abrahim, C. M. M., Mendrone Jr., A., Salomon, T., Almeida-Neto, C. d., França, R. F. O., Belotti, M. C., Carvalho, M. P. S. S., Costa, A. G., Crispim, M. A. E., Ferreira, S. C., Fraiji, N. A., Gurzenda, S., Whittaker, C., Kamaura, L. T., Takecian, P. L., Peixoto, P. d. S., Oikawa, M. K., Nishiya, A. S., Rocha, V., Salles, N. A., Santos, A. A. d. S., da Silva, M. A., Custer, B., Parag, K. V., Barral-Netto, M., Kraemer, M. U. G., Pereira, R. H. M., Pybus, O. G., Busch, M. P., Castro, M. C., Dye, C., Nascimento, V. H., Faria, N. R., and Sabino, E. C. (2021). Three-quarters attack rate of SARS-CoV-2 in the Brazilian Amazon during a largely unmitigated epidemic. *Science*, 371(6526):288–292.

Buyse, M., Molenberghs, G., Burzykowski, T., Renard, D., and Geys, H. (2000). The validation of surrogate endpoints in meta-analyses of randomized experiments. *Biostatistics*, 1(1):49–67.

Cai, B., Ioannidis, J. P. A., Bendavid, E., and Tian, L. (2022). Exact inference for disease prevalence based on a test with unknown specificity and sensitivity. *Journal of Applied Statistics*. In press.

Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective*. Chapman and Hall/CRC.

Cevik, M., Bamford, C., and Ho, A. (2020). COVID-19 pandemic—a focused review for clinicians. *Clinical Microbiology and Infection*, 26(7):842–847.

Cole, S. R., Edwards, J. K., Breskin, A., Rosin, S., Zivich, P. N., Shook-Sa, B. E., and Hudgens, M. G. (2023). Illustration of 2 fusion designs and estimators. *American Journal of Epidemiology*, 192(3):467–474.

Cole, S. R. and Frangakis, C. E. (2009). The consistency statement in causal inference: A definition or an assumption? *Epidemiology*, 20(1):3–5.

Cole, S. R., Hudgens, M. G., Edwards, J. K., Brookhart, M. A., Richardson, D. B., Westreich, D., and Adimora, A. A. (2019). Nonparametric bounds for the risk function. *American Journal of Epidemiology*, 188(4):632–636.

Cole, S. R. and Stuart, E. A. (2010). Generalizing evidence from randomized clinical trials to target populations: The ACTG 320 trial. *American Journal of Epidemiology*, 172(1):107–115.

Colnet, B., Josse, J., Varoquaux, G., and Scornet, E. (2022). Causal effect on a target population: A sensitivity analysis to handle missing covariates. *Journal of Causal Inference*, 10(1):372–414.

Colnet, B., Mayer, I., Chen, G., Dieng, A., Li, R., Varoquaux, G., Vert, J.-P., Josse, J., and Yang, S. (2023). Causal inference methods for combining randomized trials and observational studies: a review. Accepted manuscript, *Statistical science*. arXiv:2011.08047 [stat].

Cox, R. (2013). Correlates of protection to influenza virus, where do we go from here? *Human Vaccines & Immunotherapeutics*, 9(2):405–408.

Cromer, D., Steain, M., Reynaldi, A., Schlub, T. E., Wheatley, A. K., Juno, J. A., Kent, S. J., Triccas, J. A., Khoury, D. S., and Davenport, M. P. (2022). Neutralising antibody titres as predictors of protection against SARS-CoV-2 variants and the impact of boosting: A meta-analysis. *The Lancet Microbe*, 3(1):e52–e61.

Dahabreh, I. J. (2023). Combining information to answer epidemiological questions about a target population. *American Journal of Epidemiology*. In press.

Dahabreh, I. J., Robertson, S. E., Petito, L. C., Hernán, M. A., and Steingrimsson, J. A. (2022). Efficient and robust methods for causally interpretable meta-analysis: Transporting inferences from multiple randomized trials to a target population. *Biometrics*. In press.

Dai, J. Y. and Hughes, J. P. (2012). A unified procedure for meta-analytic evaluation of surrogate end points in randomized clinical trials. *Biostatistics*, 13(4):609–624.

Daniel, R. M. (2018). Double Robustness. In *Wiley StatsRef: Statistics Reference Online*, pages 1–14. John Wiley & Sons.

Daniels, M. J. and Hughes, M. D. (1997). Meta-analysis for the evaluation of potential surrogate markers. *Statistics in Medicine*, 16(17):1965–1982.

Dean, N. and Pagano, M. (2015). Evaluating confidence interval methods for binomial proportions in clustered surveys. *Journal of Survey Statistics and Methodology*, 3(4):484–503.

Degtiar, I. and Rose, S. (2023). A review of generalizability and transportability. *Annual Review of Statistics and Its Application*, 10(1):501–524.

DiCiccio, T. J., Ritzwoller, D. M., Romano, J. P., and Shaikh, A. M. (2022). Confidence intervals for seroprevalence. *Statistical Science*, 37(3):306–321.

Diciccio, T. J. and Romano, J. P. (1988). A review of bootstrap confidence intervals. *Journal of the Royal Statistical Society: Series B (Methodological)*, 50(3):338–354.

Earle, K. A., Ambrosino, D. M., Fiore-Gartland, A., Goldblatt, D., Gilbert, P. B., Siber, G. R., Dull, P., and Plotkin, S. A. (2021). Evidence for antibody as a protective correlate for COVID-19 vaccines. *Vaccine*, 39:4423–4428.

Egami, N. and Hartman, E. (2021). Covariate selection for generalizing experimental results: Application to a large-scale development program in Uganda. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 184(4):1524–1548.

Elliott, M. R. and Valliant, R. (2017). Inference for nonprobability samples. *Statistical Science*, 32(2):249–264.

Federal Planning Bureau (2021). Population projections 2020-2070. Dataset.

Frangakis, C. E. and Rubin, D. B. (2002). Principal stratification in causal inference. *Biometrics*, 58(1):21–29.

Gail, M. H., Pfeiffer, R., Van Houwelingen, H. C., and Carroll, R. J. (2000). On meta-analytic assessment of surrogate outcomes. *Biostatistics*, 1(3):231–246.

Gart, J. J. and Buck, A. A. (1966). Comparison of a screening test and a reference test in epidemiologic studies: II. A probabilistic model for the comparison of diagnostic tests. *American Journal of Epidemiology*, 83(3):593–602.

Gastwirth, J. L. (1987). The statistical precision of medical screening procedures: Application to polygraph and AIDS antibodies test data. *Statistical Science*, 2(3):213–222.

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

Gilbert, P. B., Donis, R. O., Koup, R. A., Fong, Y., Plotkin, S. A., and Follmann, D. (2022a). A Covid-19 milestone attained—a correlate of protection for vaccines. *New England Journal of Medicine*, 387(24):2203–2206.

Gilbert, P. B., Montefiori, D. C., McDermott, A. B., Fong, Y., Benkeser, D., Deng, W., Zhou, H., Houchens, C. R., Martins, K., Jayashankar, L., Castellino, F., Flach, B., Lin, B. C., O'Connell, S., McDanal, C., Eaton, A., Sarzotti-Kelsoe, M., Lu, Y., Yu, C., Borate, B., van der Laan, L. W. P., Hejazi, N. S., Huynh, C., Miller, J., El Sahly, H. M., Baden, L. R., Baron, M., De La Cruz, L., Gay, C., Kalams, S., Kelley, C. F., Andrasik, M. P., Kublin, J. G., Corey, L., Neuzil, K. M., Carpp, L. N., Pajon, R., Follmann, D., Donis, R. O., Koup, R. A., Immune Assays Team, Moderna, Inc. Team, Coronavirus Vaccine Prevention Network (CoVPN)/Coronavirus Efficacy (COVE) Team, and United States Government (USG)/CoVPN Biostatistics Team (2022b). Immune correlates analysis of the mRNA-1273 COVID-19 vaccine efficacy clinical trial. *Science*, 375(6576):43–50.

Gilbert, P. B., Qin, L., and Self, S. G. (2008). Evaluating a surrogate endpoint at three levels, with application to vaccine development. *Statistics in Medicine*, 27(23):4758–4778.

Godambe, V. P. (1960). An optimum property of regular maximum likelihood estimation. *The Annals of Mathematical Statistics*, 31(4):1208–1211.

Goldblatt, D., Alter, G., Crotty, S., and Plotkin, S. A. (2022). Correlates of protection against SARS-CoV-2 infection and COVID-19 disease. *Immunological Reviews*, 310(1):6–26.

Greiner, M. and Gardner, I. A. (2000). Application of diagnostic tests in veterinary epidemiologic studies. *Preventive Veterinary Medicine*, 45(1):43–59.

Groves, R. M. (2006). Nonresponse rates and nonresponse bias in household surveys. *Public Opinion Quarterly*, 70(5):646–675.

Havers, F. P., Reed, C., Lim, T., Montgomery, J. M., Klena, J. D., Hall, A. J., Fry, A. M., Cannon, D. L., Chiang, C.-F., Gibbons, A., Krapiunaya, I., Morales-Betoulle, M., Roguski, K., Ur Rasheed, M. A., Freeman, B., Lester, S., Mills, L., Carroll, D. S., Owen, S. M., Johnson, J. A., Semenova, V., Blackmore, C., Blog, D., Chai, S. J., Dunn, A., Hand, J., Jain, S., Lindquist, S., Lynfield, R., Pritchard, S., Sokol, T., Sosa, L., Turabelidze, G., Watkins, S. M., Wiesman, J., Williams, R. W., Yendell, S., Schiffer, J., and Thornburg, N. J. (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):1576–1586.

Hayes, A. F. (1996). Permutation test is not distribution-free: Testing $H_0 : \rho = 0$. *Psychological Methods*, 1(2):184.

Hemenway, D. (1997). Survey research and self-defense gun use: An explanation of extreme overestimates. *The Journal of Criminal Law and Criminology*, 87(4):1430–1445.

Hens, N., Shkedy, Z., Aerts, M., Faes, C., Damme, P. V., and Beutels, P. (2012). *Modeling Infectious Disease Parameters Based on Serological and Social Contact Data: A Modern Statistical Perspective*. Springer Science & Business Media.

Hernán, M. A., Hernández-Díaz, S., and Robins, J. M. (2004). A structural approach to selection bias. *Epidemiology*, 15(5):615–625.

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

Herzog, S. A., Bie, J. D., Abrams, S., Wouters, I., Ekinci, E., Patteet, L., Coppens, A., Spiegeleer, S. D., Beutels, P., Damme, P. V., Hens, N., and Theeten, H. (2022). Seroprevalence of IgG antibodies against SARS-CoV-2 – a serial prospective cross-sectional nationwide study of residual samples, Belgium, March to October 2020. *Eurosurveillance*, 27(9):1–9.

Hilden, J. (1979). A further comment on "Estimating prevalence from the results of a screening test". *American Journal of Epidemiology*, 109(6):721–722.

Huber, P. J. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 221–234. University of California Press.

Huber, P. J. and Ronchetti, E. M. (2009). *Robust Statistics*. John Wiley & Sons, Inc., Hoboken, NJ, 2nd edition.

Hudgens, M. G. and Halloran, M. E. (2008). Toward causal inference with interference. *Journal of the American Statistical Association*, 103(482):832–842.

Inagaki, N. (1973). Asymptotic relations between the likelihood estimating function and the maximum likelihood estimator. *Annals of the Institute of Statistical Mathematics*, 25(1):1–26.

Joffe, M. M. and Greene, T. (2009). Related causal frameworks for surrogate outcomes. *Biometrics*, 65(2):530–538.

Johnson, N. P. A. S. and Mueller, J. (2002). Updating the Accounts: Global Mortality of the 1918-1920 "Spanish" Influenza Pandemic. *Bulletin of the History of Medicine*, 76(1):105–115. Publisher: Johns Hopkins University Press.

Khoury, D. S., Cromer, D., Reynaldi, A., Schlub, T. E., Wheatley, A. K., Juno, J. A., Subbarao, K., Kent, S. J., Triccas, J. A., and Davenport, M. P. (2021). Neutralizing antibody levels are highly predictive of immune protection from symptomatic SARS-CoV-2 infection. *Nature Medicine*, 27:1205–1211.

Kolenikov, S. and Bollen, K. A. (2012). Testing negative error variances: Is a Heywood case a symptom of misspecification? *Sociological Methods & Research*, 41(1):124–167.

Korn, E. L. and Graubard, B. I. (1999). *Analysis of Health Surveys*. John Wiley & Sons, 1st edition.

Kosorok, M. R. (2008). *Introduction to Empirical Processes and Semiparametric Inference.* Springer, New York, NY.

Krammer, F. (2020). SARS-CoV-2 vaccines in development. *Nature*, 586(7830):516–527.

Krammer, F. (2021). A correlate of protection for SARS-CoV-2 vaccines is urgently needed. *Nature Medicine*, 27(7):1147–1148.

Lang, Z. and Reiczigel, J. (2014). Confidence limits for prevalence of disease adjusted for estimated sensitivity and specificity. *Preventive Veterinary Medicine*, 113(1):13–22.

Larremore, D. B., Fosdick, B. K., Bubar, K. M., Zhang, S., Kissler, S. M., Metcalf, C. J. E., Buckee, C. O., and Grad, Y. H. (2021). Estimating SARS-CoV-2 seroprevalence and epidemiological parameters with uncertainty from serological surveys. *eLife*, 10:e64206.

Larremore, D. B., Fosdick, B. K., Zhang, S., and Grad, Y. H. (2020). Jointly modeling prevalence, sensitivity and specificity for optimal sample allocation. bioRxiv. `http://biorxiv.org/lookup/doi/10.1101/2020.05.23.112649`.

Lesko, C. R., Buchanan, A. L., Westreich, D., Edwards, J. K., Hudgens, M. G., and Cole, S. R. (2017). Generalizing study results: a potential outcomes perspective. *Epidemiology*, 28(4):553–561.

Levy, P. S. and Kass, E. H. (1970). A three-population model for sequential screening for bacteriuria. *American Journal of Epidemiology*, 91(2):148–154.

Lipsitch, M. and Kahn, R. (2021). Interpreting vaccine efficacy trial results for infection and transmission. *Vaccine*, 39(30):4082–4088.

Lohr, S. L. (2010). *Sampling: Design and Analysis*. Chapman & Hall/CRC, 2nd edition.

Lunceford, J. K. and Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. *Statistics in Medicine*, 23(19):2937–2960.

Mackey, K., Ayers, C. K., Kondo, K. K., Saha, S., Advani, S. M., Young, S., Spencer, H., Rusek, M., Anderson, J., Veazie, S., Smith, M., and Kansagara, D. (2021). Racial and ethnic disparities in COVID-19-Related infections, hospitalizations, and deaths. *Annals of Internal Medicine*, 174(3):362–373.

Manuel, D. G., Rosella, L. C., and Stukel, T. A. (2010). Importance of accurately identifying disease in studies using electronic health records. *BMJ*, 341:c4226.

Marchevsky, N. (1979). Re: "Estimating prevalence from the results of a screening test". *American Journal of Epidemiology*, 109(6):720–721.

Maxim, L. D., Niebo, R., and Utell, M. J. (2014). Screening tests: A review with examples. *Inhalation Toxicology*, 26(13):811–828.

Messam, L. L. M., Branscum, A. J., Collins, M. T., and 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.

Michiels, H., Vandebosch, A., and Vansteelandt, S. (2022). Estimation and interpretation of vaccine efficacy in COVID-19 randomized clinical trials. *Statistical Communications in Infectious Diseases*, 14(1).

Miller, K. S. (1981). On the inverse of the sum of matrices. *Mathematics Magazine*, 54(2):67–72.

Mulherin, S. A. and Miller, W. C. (2002). Spectrum bias or spectrum effect? Subgroup variation in diagnostic test evaluation. *Annals of Internal Medicine*, 137(7):598–602.

Neison, F. G. P. (1844). On a method recently proposed for conducting inquiries into the comparative sanatory condition of various districts, with illustrations, derived from numerous places in Great Britain at the period of the last census. *Journal of the Statistical Society of London*, 7(1):40–68.

Office of the Commissioner (2020a). FDA takes additional action in fight against COVID-19 by issuing Emergency Use Authorization for second COVID-19 vaccine. Press release.

Office of the Commissioner (2020b). FDA takes key action in fight against COVID-19 by issuing Emergency Use Authorization for first COVID-19 vaccine. Press release.

Openshaw, P. J. M. (2022). Using correlates to accelerate vaccinology. *Science*, 375(6576):22–23.

Pearl, J. and Bareinboim, E. (2014). External validity: From do-calculus to transportability across populations. *Statistical Science*, 29(4):579–595.

Perez-Saez, J., Zaballa, M.-E., Yerly, S., Andrey, D. O., Meyer, B., Eckerle, I., Balavoine, J.-F., Chappuis, F., Pittet, D., Trono, D., Kherad, O., Vuilleumier, N., Kaiser, L., Guessous, I., Stringhini, S., and Azman, A. S. (2021). Persistence of anti-SARS-CoV-2 antibodies: immunoassay heterogeneity and implications for serosurveillance. *Clinical Microbiology and Infection*, 27(11):1695.e7–1695.e12.

Petersen, M. L., Porter, K. E., Gruber, S., Wang, Y., and van der Laan, M. J. (2012). Diagnosing and responding to violations in the positivity assumption. *Statistical Methods in Medical Research*, 21(1):31–54.

Pijls, B. G., Jolani, S., Atherley, A., Derckx, R. T., Dijkstra, J. I. R., Franssen, G. H. L., Hendriks, S., Richters, A., Venemans-Jellema, A., Zalpuri, S., and Zeegers, M. P. (2021). Demographic risk factors for COVID-19 infection, severity, ICU admission and death: a meta-analysis of 59 studies. *BMJ Open*, 11(1):e044640.

Plotkin, S. A. (2010). Correlates of Protection Induced by Vaccination. *Clinical and Vaccine Immunology : CVI*, 17(7):1055–1065.

Plotkin, S. A. (2020). Updates on immunologic correlates of vaccine-induced protection. *Vaccine*, 38(9):2250–2257.

Prentice, R. L. (1989). Surrogate endpoints in clinical trials: Definition and operational criteria. *Statistics in Medicine*, 8(4):431–440.

Qin, L., Gilbert, P. B., Corey, L., McElrath, M. J., and Self, S. G. (2007). A framework for assessing immunological correlates of protection in vaccine trials. *The Journal of Infectious Diseases*, 196(9):1304–1312.

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rao, C. R. (1973). *Linear Statistical Inference and its Applications*, volume 2. Wiley, New York.

Rapaka, R. R., Hammershaimb, E. A., and Neuzil, K. M. (2022). Are some COVID-19 vaccines better than others? Interpreting and comparing estimates of efficacy in vaccine trials. *Clinical Infectious Diseases*, 74(2):352–358.

Ridder, G. and Moffitt, R. (2007). The econometrics of data combination. *Handbook of Econometrics*, 6:5469–5547.

Robins, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7(9):1393–1512.

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

Rosin, S., Shook-Sa, B. E., Cole, S. R., and Hudgens, M. G. (2023). Estimating SARS-CoV-2 seroprevalence. *Journal of the Royal Statistical Society, Series A (Statistics in Society)*. In press.

Rubin, D. B. (1980). Comment on "Randomization analysis of experimental data: The Fisher randomization test". *Journal of the American Statistical Association*, 75(371):591–593.

Rudolph, J. E., Cole, S. R., and Edwards, J. K. (2018). Parametric assumptions equate to hidden observations: comparing the efficiency of nonparametric and parametric models for estimating time to AIDS or death in a cohort of HIV-positive women. *BMC Medical Research Methodology*, 18(1):142.

Rudolph, J. E., Edwards, J. K., Naimi, A. I., and Westreich, D. J. (2021). Simulation in practice: The balancing intercept. *American Journal of Epidemiology*, 190(8):1696–1698.

Sargent, D. J., Patiyil, S., Yothers, G., Haller, D. G., Gray, R., Benedetti, J., Buyse, M., Labianca, R., Seitz, J. F., O'Callaghan, C. J., et al. (2007). End points for colon cancer adjuvant trials: Observations and recommendations based on individual patient data from 20,898 patients enrolled onto 18 randomized trials from the ACCENT Group. *Journal of Clinical Oncology*, 25(29):4569–4574.

Saul, B. C. and Hudgens, M. G. (2020). The calculus of M-estimation in R with geex. *Journal of Statistical Software*, 92(2):10.18637/jss.v092.i02.

Scott, A. J. and Wild, C. (1986). Fitting logistic models under case-control or choice based sampling. *Journal of the Royal Statistical Society: Series B (Methodological)*, 48(2):170–182.

Sempos, C. T. and Tian, L. (2021). Adjusting coronavirus prevalence estimates for laboratory test kit error. *American Journal of Epidemiology*, 190(1):109–115.

Shi, X., Pan, Z., and Miao, W. (2023). Data integration in causal inference. *Wiley Interdisciplinary Reviews: Computational Statistics*, 15(1):e1581.

Shioda, K., Lau, M. S., Kraay, A. N., Nelson, K. N., Siegler, A. J., Sullivan, P. S., Collins, M. H., Weitz, J. S., and Lopman, B. A. (2021). Estimating the cumulative incidence of SARS-CoV-2 infection and the infection fatality ratio in light of waning antibodies. *Epidemiology*, 32(4):518–524.

Shook-Sa, B. E., Boyce, R. M., and 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.

Shook-Sa, B. E., Zivich, P. N., Rosin, S. P., Edwards, J. K., Adimora, A. A., Hudgens, M. G., and Cole, S. R. (2023). Fusing trial data for treatment comparisons: Single versus multi-span bridging. arXiv. `https://doi.org/10.48550/arXiv.2305.00845`.

Shoukri, M. M. (2003). *Measures of Interobserver Agreement and Reliability*. CRC Press, New York.

Stadlbauer, D., Tan, J., Jiang, K., Hernandez, M. M., Fabre, S., Amanat, F., Teo, C., Arunkumar, G. A., McMahon, M., Capuano, C., Twyman, K., Jhang, J., Nowak, M. D., Simon, V., Sordillo, E. M., van Bakel, H., and Krammer, F. (2021). Repeated cross-sectional sero-monitoring of SARS-CoV-2 in New York City. *Nature*, 590(7844):146–150.

Stefanski, L. A. and Boos, D. D. (2002). The calculus of M-estimation. *The American Statistician*, 56(1):29–38.

Stuart, E. A., Cole, S. R., Bradshaw, C. P., and Leaf, P. J. (2010). The Use of Propensity Scores to Assess the Generalizability of Results from Randomized Trials. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 174(2):369–386.

Takahashi, S., Peluso, M. J., Hakim, J., Turcios, K., Janson, O., Routledge, I., Busch, M. P., Hoh, R., Tai, V., Kelly, J. D., et al. (2023). Sars-cov-2 serology across scales: A framework for unbiased estimation of cumulative incidence incorporating antibody kinetics and epidemic recency. *American Journal of Epidemiology*, page kwad106.

Tang, P. A., Bentzen, S. M., Chen, E. X., and Siu, L. L. (2007). Surrogate end points for median overall survival in metastatic colorectal cancer: literature-based analysis from 39 randomized controlled trials of first-line chemotherapy. *Journal of Clinical Oncology*, 25(29):4562–4568.

US Census Bureau (2019). American Community Survey 1-year estimates, Public Use Microdata Sample. Dataset.

U.S. Food and Drug Administration (2017). FDA facts: Biomarkers and surrogate endpoints. `https://www.fda.gov/about-fda/innovation-fda/fda-facts-biomarkers-and-surrogate-endpoints`. Accessed: 2023-03-25.

U.S. Food and Drug Administration (2022). Table of surrogate endpoints that were the basis of drug approval or licensure. https://www.fda.gov/drugs/development-resources/table-surrogate-endpoints-were-basis-drug-approval-or-licensure. Accessed: 2023-03-25.

USG COVID-19 Response Team / Coronavirus Prevention Network (CoVPN) Biostatistics Team, Gilbert, P. B., Fong, Y., Benkeser, D., Andriesen, J., Borate, B., Carone, M., Carpp, L. N., Diaz, I., Fay, M. P., Fiore-Gartland, A., Hejazi, N. S., Huang, Y., Huang, Y., Hyrien, O., Janes, H. E., Juraska, M., Li, K., Luedtke, A., Nason, M., Randhawa, A. K., van der Laan, L., Williamson, B., Zhang, W., and Follmann, D. (2022). USG COVID-19 Response Team / CoVPN vaccine efficacy trial immune correlates Statistical Analysis Plan. https://figshare.com/articles/online_resource/CoVPN_OWS_COVID-19_Vaccine_Efficacy_Trial_Immune_Correlates_SAP/13198595. Version 0.4.

Uyoga, S., Adetifa, I. M. O., Karanja, H. K., Nyagwange, J., Tuju, J., Wanjiku, P., Aman, R., Mwangangi, M., Amoth, P., Kasera, K., Ng'ang'a, W., Rombo, C., Yegon, C., Kithi, K., Odhiambo, E., Rotich, T., Orgut, I., Kihara, S., Otiende, M., Bottomley, C., Mupe, Z. N., Kagucia, E. W., Gallagher, K. E., Etyang, A., Voller, S., Gitonga, J. N., Mugo, D., Agoti, C. N., Otieno, E., Ndwiga, L., Lambe, T., Wright, D., Barasa, E., Tsofa, B., Bejon, P., Ochola-Oyier, L. I., Agweyu, A., Scott, J. A. G., and Warimwe, G. M. (2021). Seroprevalence of anti–SARS-CoV-2 IgG antibodies in Kenyan blood donors. *Science*, 371(6524):79–82.

van Belle, G., Fisher, L. D., Heagerty, P. J., and Lumley, T. (2004). *Biostatistics: A Methodology for the Health Sciences*. John Wiley & Sons, 2nd edition.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press, Cambridge, UK.

Weber, D. J., Rutala, W. A., Fischer, W. A., Kanamori, H., and Sickbert-Bennett, E. E. (2016). Emerging infectious diseases: Focus on infection control issues for novel coronaviruses (Severe Acute Respiratory Syndrome-CoV and Middle East Respiratory Syndrome-CoV), hemorrhagic fever viruses (Lassa and Ebola), and highly pathogenic avian influenza viruses, A(H5N1) and A(H7N9). *American Journal of Infection Control*, 44(5):e91–e100.

Westreich, D. and Cole, S. R. (2010). Invited Commentary: Positivity in practice. *American Journal of Epidemiology*, 171(6):674–677.

Westreich, D., Edwards, J. K., Lesko, C. R., Stuart, E., and Cole, S. R. (2017). Transportability of trial results using inverse odds of sampling weights. *American Journal of Epidemiology*, 186(8):1010–1014.

Wilson, J. M. G. and Jungner, G. (1968). Principles and practice of screening for disease. Technical report, World Health Organization.

Wilson, M. E. (1995). Travel and the emergence of infectious diseases. *Emerging Infectious Diseases*, 1(2):39–46.

Winship, C. and Mare, R. D. (1992). Models for sample selection bias. *Annual Review of Sociology*, 18:327–350.

World Health Organization (2021). Evaluation of COVID-19 vaccine effectiveness. https://www.who.int/publications/i/item/WHO-2019-nCoV-vaccine_effectiveness-measurement-2021.1. Interim guidance.

World Health Organization (2022). Coronavirus (Covid-19) data. Website. https://www.who.int/data.

Yuan, K.-H. (1997). A theorem on uniform convergence of stochastic functions with applications. *Journal of Multivariate Analysis*, 62(1):100–109.

Yuan, K.-H. and Jennrich, R. I. (1998). Asymptotics of estimating equations under natural conditions. *Journal of Multivariate Analysis*, 65(2):245–260.

Zivich, P. N., Cole, S. R., and Westreich, D. (2022a). Positivity: Identifiability and estimability. arXiv. https://doi.org/10.48550/arXiv.2207.05010.

Zivich, P. N., Klose, M., Cole, S. R., Edwards, J. K., and Shook-Sa, B. E. (2022b). Delicatessen: M-estimation in Python. arXiv. https://doi.org/10.48550/arXiv.2203.11300.