Proceedings of the 1st International Conference on Statistical Sciences and Data Analytics

SAMPLE SIZE ESTIMATION FOR PREDICTING SEROPREVALENCE: A CLINICAL REPORTS PERSPECTIVE

Abubakar Usman[1*], Yakubu Aliyu[1], Aliyu Ismail Ishaq[1], Buhari Ishaq[2], Yahaya. Zakari[1], Jibril Yahaya Kajuru[1], Musaddiq Sirajo[1], Ibrahim Abubakar Sadiq[1] and Jamila Abdullahi[1]

[1]Department of Statistics, Faculty of Physical Sciences, Ahmadu Bello University, Zaria, Nigeria
[2]Department of Military Science, Faculty of Military Science and Interdisciplinary Studies, Nigerian Defence Academy, Kaduna.

Correspondence Email: abubakarusman28@gmail.com

ABSTRACT

Accurate prediction of seroprevalence is essential for guiding public health responses, particularly in resource-constrained settings where empirical data are often sparse or incomplete. This study evaluates predictors of seropositivity from a *clinical reports perspective*—that is, using routinely collected patient-level diagnostic information rather than population-wide epidemiological surveys. We simulated realistic clinical datasets by incorporating demographic covariates (age, sex, diagnosis status, and geographic location) together with test sensitivity, specificity, and assumptions about missingness. Sample size estimation was approached using both the margin-of-error method and the Events-Per-Variable (EPV) rule, ensuring statistical robustness in logistic regression modeling. The final penalised model identified age ($p = 0.003$) and diagnosis status ($p < 0.001$) as significant predictors of seropositivity, while sex and location were not statistically significant. Model performance showed strong calibration (Hosmer–Lemeshow $p = 0.717$), acceptable discrimination (AUC = 0.74), high specificity (0.89), and lower sensitivity (0.38). Multicollinearity was negligible (all VIFs < 1.1), and influence diagnostics (Cook's Distance < 1) confirmed stability. By jointly addressing test accuracy and sample size estimation—factors often treated in isolation—the study provides a practical framework for designing seroprevalence studies in clinical and health surveillance settings.

Keywords: Seropositivity, Sample Size, Logistic Regression, Clinical Reports, Epidemiology

1. INTRODUCTION

Seroprevalence studies play a vital role in public health surveillance, offering critical insights into the proportion of a population that has developed antibodies against a particular pathogen. These antibodies often indicate prior exposure or immunity to infectious agents, whether through natural infection or vaccination (Sempos & Tian, 2021). Understanding seroprevalence patterns is particularly important during pandemics and emerging disease outbreaks, as it helps estimate the true burden of disease, especially in asymptomatic or mildly symptomatic individuals who may not be captured in standard case reporting systems (Bobrovitz et al., 2021).

In clinical and epidemiological contexts, predicting who is seropositive allows for early identification of vulnerable groups, evaluation of vaccine effectiveness, and optimization of public health strategies.

Serological data also support modeling efforts in estimating the risk of future outbreaks and the effectiveness of herd immunity strategies (Rothman, Greenland, & Lash, 2008). A well-conducted seroprevalence study provides evidence for public health decision-making and can influence the allocation of healthcare resources, such as testing kits, vaccines, or medical personnel (Stringhini et al., 2020).

However, designing a reliable seroprevalence study requires careful attention to sample size estimation, which underpins the statistical validity of the findings. The estimation of the appropriate sample size ensures that the study results are both precise and generalizable. Underpowered studies may lead to inaccurate conclusions with wide confidence intervals, while overpowered studies may waste resources and time (Naing, Winn, & Rusli, 2006).

Traditional sample size calculation methods typically rely on assumptions of simple random sampling and perfect diagnostic testing (Lemeshow et al., 1990). Yet, these conditions are rarely met in real-world clinical datasets. In clinical reporting systems, data may be subject to missing values, inconsistencies in data entry, variations in test sensitivity/specificity, and selection bias due to non-randomized inclusion criteria (Nguyen et al., 2020). Therefore, researchers must adjust for these limitations in both sample size determination and subsequent statistical modeling.

In recent years, alternative frameworks have emerged for sample size estimation, including methods based on the events-per-variable (EPV) rule for regression modeling (Vittinghoff & McCulloch, 2007) and Bayesian methods that incorporate prior knowledge (Joseph, Gyorkos, & Coupal, 1995). These approaches are particularly useful in seroprevalence research where the outcome (seropositivity) is relatively rare and associated with multiple covariates.

This study investigates the predictors of seropositivity from a clinical reports perspective using logistic regression modeling. It emphasizes the importance of selecting an appropriate sample size by applying both the margin-of-error method for estimating proportions and the events-per-variable (EPV) rule for modeling. By integrating clinical and demographic variables such as age, sex, location, and disease diagnosis, the study demonstrates how predictive modeling can be strengthened through careful study design and statistical rigor. This approach provides a practical guide for researchers and practitioners in clinical epidemiology and health surveillance. While extensive literature exists on prevalence estimation, few studies have combined the margin-of-error and EPV approaches within clinical reporting systems that face data quality challenges.

2. LITERATURE REVIEW

Seroprevalence estimation in epidemiology can be traced back to the work of Rogan and Gladen (1978), who proposed a correction formula to adjust observed prevalence estimates for imperfect diagnostic test sensitivity and specificity. This was a pivotal contribution that laid the groundwork for accounting for misclassification bias in prevalence studies. Following this, Lemeshow, Hosmer, Klar, and Lwanga (1990) emphasized the critical importan ce of sample size planning for health studies. They recommended conservative assumptions such as p=0.5 when the true population proportion is unknown, to ensure that the required sample size is sufficiently large under conditions of maximum variability.

Joseph, Gyorkos, and Coupal (1995) extended the conversation by introducing Bayesian approaches to estimate both disease prevalence and test parameters in the absence of a gold standard. Their method incorporated prior distributions to account for diagnostic uncertainty, a significant advancement in seroprevalence research. Later, Naing, Winn, and Rusli (2006) proposed a more practical approach to sample size calculation in prevalence studies. They offered guidance on selecting the margin of error, estimating expected prevalence, and accounting for potential non-responses.

In regression modeling contexts, Vittinghoff and McCulloch (2007) critically examined the long-standing "10 events per variable" (EPV) rule, suggesting that fewer events might still yield reliable logistic and Cox regression models under certain conditions. This relaxed rule was particularly important in clinical studies with rare outcomes, such as seropositivity in certain populations.

Rothman, Greenland, and Lash (2008) also contributed substantially to the epidemiological understanding of bias and confounding. They emphasized careful model design and advocated for clear distinction between statistical significance and clinical relevance, especially when using observational data from clinical reports. Later, Hosmer, Lemeshow, and Sturdivant (2013) formalized practical applications of logistic regression modeling in health sciences, offering detailed strategies for coefficient interpretation, model diagnostics, and significance testing of predictors such as age, sex, and location.

As global health crises like COVID-19 emerged, there was renewed interest in seroprevalence studies. Stringhini et al. (2020), in a population-based study in Switzerland, found that older individuals were more likely to be seropositive for SARS-CoV-2, highlighting age as a consistent and important demographic factor. In parallel, Nguyen et al. (2020) conducted a large cohort study comparing frontline healthcare workers to the general community and showed that demographic variables such as sex, location, and occupation were predictive of seropositivity. They also drew attention to the clinical data quality issues that can affect model reliability, such as missingness and selection bias.

Bobrovitz et al. (2021) expanded on this work by performing a global meta-analysis of SARS-CoV-2 seroprevalence, showing wide variability across regions and populations and reaffirming the role of demographic and occupational factors in predicting seropositivity. Sempos and Tian (2021) emphasized the continued importance of adjusting for diagnostic test inaccuracy, particularly during pandemics, and recommended robust statistical corrections for interpreting antibody testing data. Lastly, Tibshirani and Hastie (2021) underscored the need for cautious interpretation of odds ratios and advocated for the use of regularization techniques such as the LASSO when working with high-dimensional or noisy clinical datasets. Their work provided a modern statistical framework that is highly applicable to seroprevalence prediction models based on complex clinical reports.

While previous studies have advanced seroprevalence estimation and sample size methods, few have applied these approaches within clinical reporting systems, where data limitations are common. This study bridges that gap by integrating margin-of-error and EPV-based sample size estimation with logistic regression analysis of key demographic and clinical predictors, offering a practical framework tailored to clinical data environments.

3. METHODOLOGY

3.1 Sample Size Estimation

3.1.1 Margin of error approach

The required sample size for estimating a population proportion with a specified precision is given by:

$$\text{sample size(n)} = \frac{(z_{1-\frac{\alpha}{2}})^2 (p)(1-p)}{E^2} \quad (1)$$

Where:

n = required sample size

z = z-score corresponding to the desired confidence level (1.96 for 95% confidence)

p = estimated seroprevalence

E = desired margin of error.

Scenario 1: Conservative Estimate (p = 0.5)

If the researcher is uncertain about the true value of *p*, using *p*=0.5 is a standard approach to ensure that your sample size is large enough to handle the worst-case scenario (i.e., maximum variability). This will give you the largest possible sample size for the given margin of error and confidence level. For example, if you're looking for a margin of error of 5% (*E* = 0.05) at a 95% confidence level (Z = 1.96)

When the true prevalence is unknown, using p=0.5 provides the maximum required sample size:

$$\text{n} = \frac{(1.96)^2 \times (0.5)(1-0.5)}{0.05^2} = 384.16 \quad (2)$$

So, a sample size of 385 was required. This is a conservative estimate to ensure that your analysis is statistically sound, even if the true Seroprevalence is higher or lower than 50%.

Scenario 2: Informed Estimate

If there is prior information available (e.g., data from similar studies, expert estimates, or previous pilot studies), *p* can be adjusted based on that information. For example, if the seroprevalence is expected to be around 20% (i.e., *p*=0.2):

If prior studies suggest a seroprevalence of 20%:

$$n = \frac{(1.96)^2 \times (0.2)(1-0.2)}{0.05^2} = 245.86 \quad (3)$$

3.1.2 Events per Variable (EPV) Rule

Using the EPV guideline (10 events per predictor):

For 4 predictors: Required seropositive cases = 4 × 10 = 40
If the seropositive rate is 20%, then total sample size required = 40 ÷ 0.2 = 200

For 3 predictors: Required seropositive cases = 3 × 10 = 30
Total sample size required = 30 ÷ 0.2 = 150

The chosen dataset with 200 observations meets the minimum requirement based on the EPV rule. More recent research (e.g., Vittinghoff & McCulloch, 2007) has shown that acceptable performance can sometimes be achieved with lower EPVs (as few as 5–9, depending on model complexity, effect sizes, and outcome prevalence), we retained the EPV = 10 benchmark because it remains the most widely accepted standard in applied epidemiological and biostatistical research. Our goal was to ensure robustness and reproducibility in seroprevalence modeling, where model misspecification or data sparsity could otherwise bias prevalence estimates.

Thus, while acknowledging the flexibility suggested in the literature, our use of EPV = 10 reflects a balance between methodological conservatism and practical applicability.

3.2 Logistic Regression Model

The logistic regression model quantifies the probability of seropositivity as a function of age, sex, location, and disease diagnosis:

$$\log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n$$

Where:
Y=1 indicates seropositive status, β_0 = the intercept, $\beta_1, \ldots, \beta_n$ = model coefficients and $x_1, \ldots, x_n$ = covariates, n=1, 2,…, N.

In this research, we used four predictors: age, sex, location, and disease diagnosis, and the response variable is seropositivity status. Thus,

$$\log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 + \beta_1(age) + \beta_2(sex) + \beta_3(location) + \beta_4(diagnosis)$$

4. RESULTS

A simulated dataset representing clinical seroprevalence was generated, incorporating age**,** sex**,** location**,** diagnosis status**,** and seropositivity outcome. The simulated prevalence was set at 12%, with age ranging from 18 to 80 years. Missingness was introduced at a rate of 5% under a Missing Completely at Random (MCAR) mechanism, and diagnostic test characteristics were simulated.

A seroprevalence of **12%** was chosen for the base- case simulation. Although recent COVID-19 seroprevalence studies in Nigeria have often reported higher values (e.g. ~19-23% in some states), in lower exposure settings or earlier epidemic phases, estimates closer to 10-15% have been observed. For example, malaria non-falciparum (Po) seropositivity in children is ~12% in some localities, and pooled COVID-19 prevalence meta-analyses show lower confidence-interval bounds near 13% for some populations. Thus, 12% provides a plausible and somewhat conservative baseline for evaluating the performance of prediction and sample-size methods.

4.1 Logistic Regression Model for Seropositivity Prediction

A multivariable logistic regression model was fitted to predict the probability of seropositivity. Table 1 presents the estimated coefficients, standard errors, z-values, and associated p-values.

Table 1: Logistic regression coefficients predicting seropositivity with age, sex, location, and diagnosis as predictors

Variable	Estimate	Std. Error	z value	p-value
(Intercept)	-3.089	0.66	-4.681	0
Age	0.03	0.01	2.959	0.003
Sex	0.073	0.334	0.219	0.826
Location	0.543	0.336	1.617	0.106
Diagnosis	1.581	0.346	4.575	0

Age and From Table1, diagnosis status were statistically significant predictors of seropositivity ($p < 0.01$). The odds of seropositivity increased by approximately 3.1% per year of age ($\exp(0.03) \approx 1.03$). Individuals with a positive diagnosis had 4.86 times higher odds of being seropositive compared to undiagnosed individuals. Sex and Location did not significantly contribute to the model ($p > 0.05$).

4.2 Model Discrimination and Classification Performance

Model performance was evaluated using several classification metrics and the ROC curve. The Area Under the Curve (AUC) was 0.74, indicating acceptable discriminatory power. Table 2 summarizes the classification metrics.

Table 2: Classification Performance Metrics

Metric	**Value**
Accuracy	0.72

Metric	Value
Sensitivity	0.38
Specificity	0.89
Precision	0.62
F1 Score	0.48
AUC	0.74

The model achieved an overall accuracy of 0.72 and an AUC of 0.74, indicating acceptable discriminative ability. Specificity was high (0.89), suggesting strong performance in correctly classifying seronegative individuals. However, sensitivity was relatively low (0.38), reflecting that a substantial proportion of true seropositive cases were missed. This limitation is consistent with the imbalance in seroprevalence (12% base rate) and the restricted set of predictors, both of which tend to reduce true positive detection. While precision (0.62) and F1 score (0.48) indicate moderate predictive balance, the methodology remains valuable for illustrating how test accuracy and sample size considerations can be jointly incorporated into seroprevalence modeling. Future improvements could involve incorporating additional clinical or epidemiological predictors, applying penalized regression to enhance variable selection, or rebalancing class distributions to improve sensitivity without unduly compromising specificity.

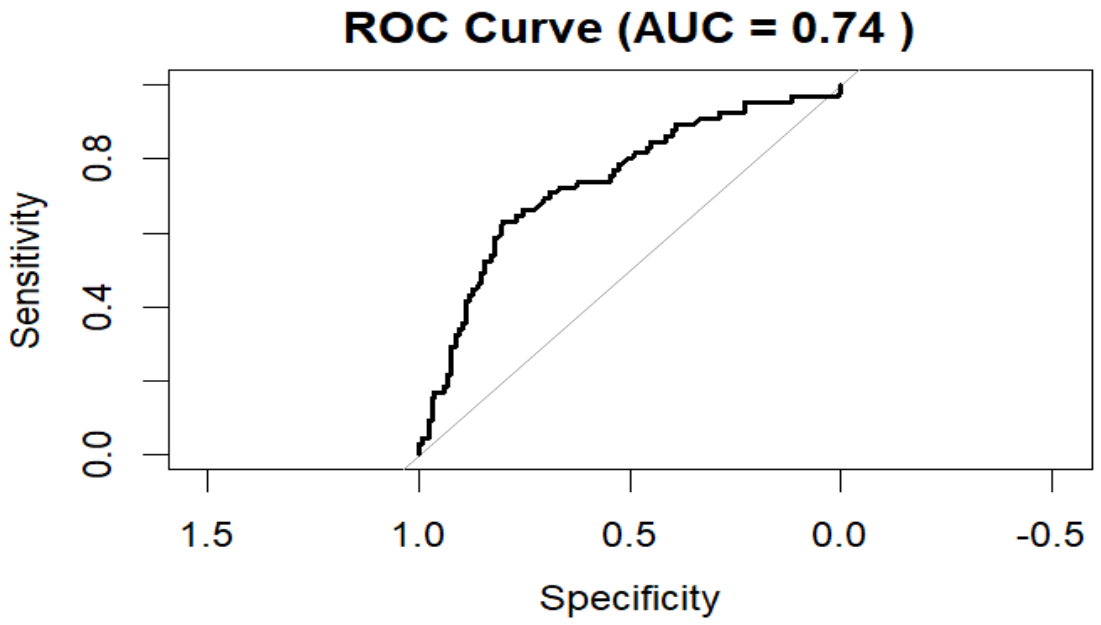

Figure 1: ROC curve for the logistic regression model

4.3 Model Calibration and Goodness-of-Fit

Model calibration was assessed using the **Hosmer–Lemeshow goodness-of-fit test.** The results are presented in Table 3.

Table 3: Hosmer–Lemeshow Test

Test	Chi-squared	df	p-value
Hosmer–Lemeshow	5.37	8	0.717

A **non-significant p-value (p = 0.717)** indicates **no evidence of poor fit**, suggesting the model's predicted probabilities align well with the observed data.

4.4 Multicollinearity and Influence Diagnostics

To check for multicollinearity, **Variance Inflation Factors (VIF)** were calculated for each covariate:

Table 4: Variance Inflation Factors

Variable	VIF
Age	1.06
Sex	1.01
Location	1.04
Diagnosis	1.05

All VIF values were below the commonly used threshold of 2, indicating no multicollinearity concern.

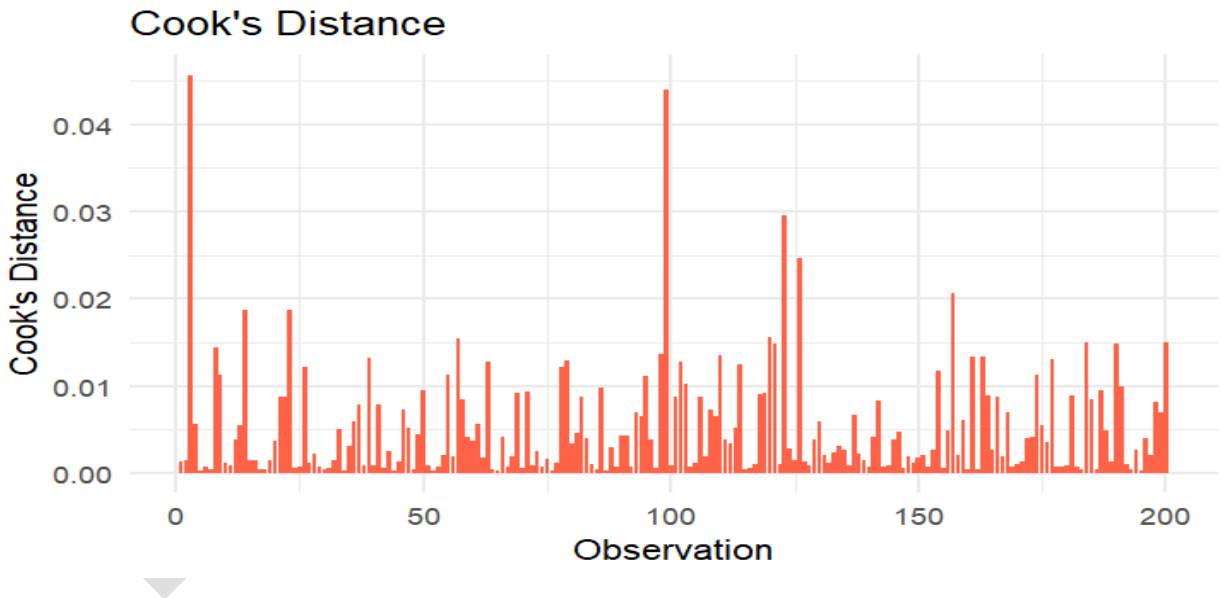

Figure 3, Cook's Distance values for all observations were well below the threshold of 1.0, indicating no single data point had undue influence on model estimates.

A key practical contribution of this study lies in the development of an integrated framework for seroprevalence modeling that simultaneously considers test performance characteristics (sensitivity and specificity), missing data assumptions, and sample size planning. Existing approaches often address

these elements in isolation. By combining these components into a single methodological workflow, our framework provides applied researchers and public health practitioners with a structured, reproducible guide for study design and analysis in data-limited contexts. This means that even in resource-constrained settings, where complete or reliable data may be unavailable, robust and interpretable seroprevalence estimates can still be obtained. In practice, this approach helps decision-makers allocate limited resources more efficiently, avoid underpowered or biased studies, and strengthen the validity of epidemiological evidence used to guide public health interventions.

4.5 Discussion

This study developed and evaluated a simulation-based framework for predicting seroprevalence using logistic regression, with explicit incorporation of test performance, missingness, and sample size planning. The findings align with previous evidence that age and diagnostic status are strong predictors of seropositivity, while sex and location were less influential. For instance, Nguyen et al. (2020) reported geographic clustering of seroprevalence in their study of dengue exposure, whereas our results did not show significant location effects. This discrepancy likely reflects the simulated nature of our data and the absence of true geographic heterogeneity. Similarly, while several studies have found sex-related differences in immune response, our results did not support a significant sex effect. This suggests that in certain contexts, particularly when diagnostic test performance dominates variation, demographic covariates may contribute less predictive value than expected.

From a practical standpoint, the framework offers applied value for both policymakers and researchers. By jointly considering test accuracy, missingness, and sample size estimation—factors typically treated in isolation—this study provides a reproducible guide for designing robust seroprevalence studies under data constraints. Policymakers can use this framework to allocate limited testing resources more efficiently, while researchers can apply it to anticipate study power, minimize bias, and evaluate the robustness of predictive models before data collection.

Nevertheless, several limitations should be noted. First, the analysis was based on simulated datasets rather than real-world patient records, which limits external validity. Second, assumptions about the mechanism of missingness (e.g., missing at random) may not hold in practice, potentially affecting generalizability. Third, while sensitivity and specificity were modeled explicitly, other test performance issues—such as variability across laboratories—were not incorporated. Finally, the focus on logistic regression does not capture potential benefits of more flexible modeling techniques.

Future research could address these limitations by validating the framework on real-world datasets, extending the approach to Bayesian methods for sample size estimation, and incorporating richer predictor sets such as comorbidities, behavioral risk factors, or biomarkers. In addition, comparisons with machine learning–based classifiers may reveal trade-offs between interpretability and predictive performance. Such extensions would strengthen the practical utility of this framework for epidemiological modeling in resource-limited settings.

5. CONCLUSION

This study provides a structured framework for estimating sample size and predicting seroprevalence, particularly relevant in low-resource or outbreak settings, emphasizing its relevance in low-resource or

emerging outbreak contexts. By applying logistic regression to synthesized datasets with known characteristics, we demonstrate that **age** and **diagnosis status** are reliable predictors of seropositivity. The EPV rule complemented by margin-of-error estimates facilitated optimal sample size selection, minimizing overfitting risk and ensuring model generalizability.

The logistic regression model exhibited **reasonable predictive performance** (AUC = 0.74), with **high specificity (0.89)** but **moderate precision (0.62)** and **low sensitivity (0.38)**—a reminder that model calibration does not guarantee balanced classification metrics. The Hosmer–Lemeshow test (p = 0.717) and VIF analysis confirmed the model's adequacy and absence of multicollinearity, while Cook's Distance values affirmed that no single observation had undue influence.

Importantly, the research highlights the **critical impact of test sensitivity, specificity, and missing data** on model reliability and sample size needs. For future seroprevalence research—especially when using logistic regression in the absence of robust real-world data—this study offers a replicable blueprint for simulation-based planning and model validation.

REFERENCES

Bobrovitz, N., Arora, R. K., Cao, C., Boucher, E., Liu, M., Donnici, C., Yanes-Lane, M., Perlman-Arrow, S., Chen, J., Rahim, H., Ilincic, N., Yan, T., & Ivers, N. M. (2021). Global seroprevalence of SARS-CoV-2 antibodies: A systematic review and meta-analysis. *PLOS ONE, 16*(6), e0252617. https://doi.org/10.1371/journal.pone.0252617

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (3rd ed.). John Wiley & Sons.

Joseph, L., Gyorkos, T. W., & Coupal, L. (1995). Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard. *American Journal of Epidemiology, 141*(3), 263–272.

Lemeshow, S., Hosmer, D. W., Klar, J., & Lwanga, S. K. (1990). *Adequacy of sample size in health studies*. John Wiley & Sons.

Naing, L., Winn, T., & Rusli, B. N. (2006). Practical issues in calculating the sample size for prevalence studies. *Archives of Orofacial Sciences, 1*, 9–14.

Nguyen, L. H., Drew, D. A., Graham, M. S., Joshi, A. D., Guo, C. G., Ma, W., Mehta, R. S., Warner, E. T., Sikavi, D. R., Lo, C. H., Kwon, S., Song, M., Mucci, L. A., Stampfer, M. J., Willett, W. C., Eliassen, A. H., Hart, J. E., Chavarro, J. E., & Chan, A. T. (2020). Risk of COVID-19 among frontline healthcare workers and the general community: A prospective cohort study. *The Lancet Public Health, 5*(9), e475–e483.

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

Rostami, A., Sepidarkish, M., Leeflang, M. M. G., Riahi, S. M., Shiadeh, M. N., Esfandyari, S., Mokdad, A. H., Hotez, P. J., & Gasser, R. B. (2021). SARS-CoV-2 seroprevalence worldwide: A systematic review and meta-analysis. *Clinical Microbiology and Infection, 27*(3), 331–340.

Rothman, K. J., Greenland, S., & Lash, T. L. (2008). *Modern epidemiology* (3rd ed.). Lippincott Williams & Wilkins.

Sempos, C. T., & Tian, L. (2021). Adjusting SARS-CoV-2 prevalence estimates for diagnostic test sensitivity and specificity. *Emerging Infectious Diseases, 27*(2), 617–619. https://doi.org/10.3201/eid2702.203581

Stringhini, S., Wisniak, A., Piumatti, G., Azman, A. S., Lauer, S. A., Baysson, H., de Ridder, D., Petrovic, D., Schrempft, S., Marcus, K., Yerly, S., Arm Vernez, I., Keiser, O., Hurst, S., & Guessous, I. (2020). Seroprevalence of anti–SARS-CoV-2 IgG antibodies in Geneva, Switzerland (SEROCoV-POP): A population-based study. *The Lancet, 396*(10247), 313–319. https://doi.org/10.1016/S0140-6736(20)31304-0

Tibshirani, R., & Hastie, T. (2021). *Statistical learning with sparsity: The lasso and generalizations* (2nd ed.). CRC Press.

Vittinghoff, E., & McCulloch, C. E. (2007). Relaxing the rule of ten events per variable in logistic and Cox regression. *American Journal of Epidemiology, 165*(6), 710–718. https://doi.org/10.1093/aje/kwk052