Potential for bias in (sero)prevalence estimates when not accounting for test sensitivity and specificity

Sarah R Haile
Epidemiology, Biostatistics and Prevention Institute (EBPI),
University of Zurich, Zurich, Switzerland

June 8, 2023

Abstract

Objectives: The COVID-19 has led to many studies of seroprevalence. A number of methods exist in the statistical literature to correctly estimate disease prevalence in the presence of diagnostic test misclassification, but these methods seem to be less known and not routinely used in the public health literature. We aimed to show how widespread the problem is in recent publications, and to quantify the magnitude of bias introduced when correct methods are not used.

Methods: We examined a sample of recent literature to determine how often public health researcher did not account for test performance in estimates of seroprevalence. Using straightforward calculations, we estimated the amount of bias introduced when reporting the proportion of positive test results instead of using sensitivity and specificity to estimate disease prevalence.

Results: Of the seroprevalence studies sampled, 80% failed to account for sensitivity and specificity. Expected bias is often more than is desired in practice, ranging from 1% to 10%.

Conclusions: Researchers conducting studies of prevalence should correctly account for test sensitivity and specificity in their statistical analysis.

Keywords:

Bias, Prevalence, Seroprevalence, Diagnostic Tests, Statistical Methods

1 Introduction

Since the beginning of the SARS-CoV-2 pandemic, thousands of papers have been published detailing seroprevalence estimates in various populations [1]. As a reader and sometimes reviewer of such publications, I noticed that while some researchers used simple approaches such as proportions or logistic regression, others used complicated methods like Bayesian hierarchical models. This made me wonder how often these methods are used in epidemiological studies and what, if any, degree of bias was introduced by using one method or the other.

As diagnostic tests are not 100% accurate, it is expected that some small number of test results will be either false positives or false negatives. Using a simple proportion of the number of positive diagnostic tests over the total number of tests ignores any misclassification inherent to the test. In the case where there are similar numbers of true positives and true negatives in

the population, the bias introduced by using the proportion of positive tests to estimate the proportion of subjects with the disease may not be very high. However, if the rate of false positives differs greatly from that of false negatives, the bias may be quite large.

For example, in Table 1, 35.5% (355/1000) of subjects had a positive test result, but the true disease prevalence in this population is 30.0% (196/1000). So there is a bias of 5.5% because there are many more false positive test results (n = 70) than false negatives (n = 15). Other examples of this phenomenon are found in the literature [2, 3]. Statisticians often talk about sensitivity, 95% in this example, and specificity, 90%, of the diagnostic in relation to these quantities (described in more detail below), but it is accepted that without a "gold standard" diagnostic tool, it is difficult to accurately assess disease prevalence.

Accounting for such misclassification in the interpretation of diagnostic tests is certainly not new in the literature. A straightforward method of adjusting observed prevalence is available [4], which gives a maximum likelihood estimate of true prevalence assuming predefined test sensitivity and specificity. The Rogan-Gladen correction has been extended to compute confidence intervals [5, 6]. Recently, an adaptation of the Rogan-Gladen correction that accounts for sampling bias, for example if only hospitalized subjects as opposed to the general population have been tested, has been proposed [7, 8, 9]. Bayesian approaches have also been developed [**?**, e.g.,]]Joseph, Berkvens, Gelman2020. A comparison of Bayesian and frequentist methods [10] showed that Bayesian methods are to be preferred, or the method of [4] with confidence intervals of [6].

Despite this extensive treatment of the misclassification problem in the statistical literature, many public health researchers appear to not realize they may be publishing biased results or know what to do about it. In what follows, a brief review of some recent publications of COVID-19 seroprevalence will be described to explore the extent of incorrect methodology used in epidemiological studies. I quantify the bias introduced in using the proportion of positive tests to estimate the proportion of subjects with disease in order to emphasize the need for methods other than simply using the proportion of positive tests (a naive approach, assuming that all subjects with a positive test are also disease positive, and that there are no false negatives). I will describe key concepts, and derive an estimate of the bias, as well as a range of prevalences where such naive estimates show low bias. Bias estimates will be described according to test sensitivity and specificity, and I will apply these results to a real example of SARS-CoV-2 seroprevalence in children. A brief discussion will follow.

2 Methods

To start, I introduce some notation. Disease status, D, is denoted 1 if a subject has the disease in question (or for the case of seroprevalence, has antibodies for it), and 0 otherwise. Similarly, the result of the diagnostic test, Y, is given as 1 if the subject tests positive for the disease, and 0 otherwise. FP is often used to refer to false positive test results, and similarly FN for false negatives, TN for true negatives and TP for true positives.

Prevalence is the probability of having the disease of interest, $P = Pr(D = 1)$. Often in prevelance studies, this probability is studied at a specific point in time, giving so-called point prevalence[11]. Seroprevalence, a related concept, looks at the proportion of individuals in the population have antibodies for a specific disease, for example, SARS-CoV-2 [12]. Sensitivity, denoted Se, sometimes also called the true positive fraction (TPF), is the probability of having a positive test result, given that the subject has the disease, $Pr(Y = 1|D = 1)$ [2]. On the other hand, specificity, Sp is the probability of having a negative test result when a subject does not have the disease, $Pr(Y = 0|D = 0)$ (sometimes 1 - specificity is discussed, which is often

referred to as false positive fraction, or FPF [13]). In real settings where true disease status is known via another method, sometimes referred to as the "gold standard", Se can be computed as $TP/(TP+FN)$, where TP is the number of true positives and FN is the number of true negatives. Similarly, Sp can be computed as $1-FP/(FP+TN)$.

The proportion of positive tests can be expressed as

$$Pr(Y=1) = (FP+TP)/(FP+TP+TN+FN),$$

while the disease prevalence in the sample can be expressed as

$$Pr(D=1) = (FN+TP)/(FP+TP+TN+FN).$$

The difference between these two quantities is simply $(FP-FN)/(FP+TP+TN+FN)$, that is, the proportion of false positives minus the proportion of false negatives.

According to the definition of joint probability $Pr(A,B) = Pr(A|B)Pr(B)$, the proportion of false positives can be written as

$$Pr(Y=1, D=0) = Pr(Y=1|D=0)Pr(D=0),$$

which simplifies to $(1-P)(1-Sp)$. In a similar fashion, the proportion of false negatives can be written as

$$Pr(Y=0, D=1) = Pr(Y=0|D=1)Pr(D=1),$$

which simplifies to $P(1-Se)$. The bias when using the proportion of positive tests, $Pr(Y=1)$, to estimate the proportion with disease, $Pr(D=1)$, is therefore $(1-P)(1-Sp)-P(1-Se)$ or equivalently $1-Sp+P(Sp+Se-2)$.

Suppose we want to guarantee that the bias is no larger than, say, $\delta = 0.02$, that is +/- 2% in either direction. We can solve

$$-\delta \leq 1-Sp+P(Sp+Se-2) \leq \delta \tag{1}$$

for P, to get:

$$\max\left(\frac{\delta+Sp-1}{Sp+Se-2}, 0\right) \leq P \leq \min\left(\frac{-\delta+Sp-1}{Sp+Se-2}, 1\right). \tag{2}$$

The lower bound will be 0 if $\delta \geq 1-Sp$, while the upper bound will be 1 if $\delta \geq 1-Se$. Therefore, if both Se and Sp are very high, say 99% or higher, then the proportion of positive tests is a good estimate of the true prevalence. If only Se is that high, this is will be true only when the true prevalence is quite high, and conversely if only Sp is very high, this will be true only when true prevalence is quite low. When neither Se nor Sp is high, the proportion of positive tests may or may not be a good estimate of the true prevalence.

One simple way to reduce this bias, if no dependence on covariates is assumed, is to use the Rogan-Gladen correction [4]. Assuming an observed fraction P_{obs} of positive test results, the corrected prevalence is

$$P_{RG} = \frac{P_{obs}+Sp-1}{Se+Sp-1}. \tag{3}$$

In a small number of cases, primarily when the sample size and the prevalence are both small[14, 15], the Rogan-Gladen correction will yield values less than 0 or greater then 1. However, even if this "clipped" version has some bias, the variance will be smaller.

The brief review of recent studies of seroprevalence in the literature started with a pubmed

(https://pubmed.ncbi.nlm.nih.gov/) search for "covid-19 seroprevalence", which yielded 1704 publications, 589 of which were published in 2022. Language or journal of publication were not exclusion criteria. A random sample of 200 of those 589 publications were examined in further detail. The following information was extracted: 1) whether the aim of the study was to assess COVID-19 seroprevalence in humans, 2) the sensitivity and 3) specificity of the diagnostic test, 4) the reported seroprevalence estimate, and 5) which statistical methods were used to calculate seroprevalence. Assuming approximately half of the publications assess COVID-10 seroprevalence in humans, 95% confidence intervals will then be no wider than $\pm$7.5% when the observed rate of non-adjustment for test sensitivity and specificity is around 80%.

3 Results

To examine the methods actually used in seroprevalence studies in the literature, I selected a random sample of 200 of the 589 publications published in 2022 which may have estimated COVID-19 seroprevalence (see flowchart in supplementary material). 92 papers were excluded because they did not directly estimate COVID-19 seroprevalence in humans. Of the remaining 108 publications (Supplementary Material), 87 (80%, 95% CI 72% to 87%) did not adjust for test performance in any way, while the remaining 21 (20%, 13% to 28%) adjusted for test performance (13 using the Rogen-Gladen correction, 6 using a Bayesian approach, 1 using a bootstrap approach, and 1 using some form of adjustment which was not clearly specified). Of the 87 that did not adjust for test performance, 45 (52%) reported sensitivity and specificity, and 42 did not report any test characteristics (expect perhaps the name and manufacturer). Based on the reported sensitivity and specificity, expected bias ranged from -14.8% to 9.6%. 19 of the publications reporting seroprevalence to within +/-1% of the true value despite not using any adjustment, while the remaining 26 (58%) needed adjustment for test performance (8 of those were not even within +/-5%). It could be inferred therefore that approximately 24 of the 42 publications not reporting test performance are also in need of adjusted seroprevalence estimates to account for test performance, even though all of those publications reported naive estimates. So, while the need to adjust seroprevalance estimates for test performance is well known the the statistical literature, the vast majority of published analyses on this topic fail to account for it when they should have. This problem is also not restricted to “low quality” journals, as such analyses can be found also in many prominent journals.

Next, using the result bias $= 1 - Sp + P(Sp + Se - 2)$ described above, I calculated the expected bias for a range of reasonable combinations of sensitivity, specificity and disease prevalence (Table 2, Figure 1). When sensitivity and specificity were both 90%, bias was as high as 10%, especially near prevalences of 0% or 100% (bottom row of Table 2, red line in leftmost panel of Figure 1). When specificity was 90%, a bias of 10% could be expected with small prevalences near 0% even if sensitivity was 99% (e.g. 3rd line of Table 2). The least bias, 1%, could be expected where sensitivity and specificity were both 99% (1st line of Table 2).

Next, using the bounds of prevalence as derived above, I explored where the maximum tolerated bias is limited to 1%, 2.5% and 5% (Figure 2). When Se and Sp are each 90%, bias is within a tolerance of 1% only very close to 50% disease prevalence, within 2.5% tolerance in the range of 38% - 62% disease prevalence and to within 5% tolerance as long as disease prevalence is between 25% and 75%. When the desired tolerance is 1%, the range of disease prevalence where a naive approach will yield unbiased results is fairly narrow in all cases, unless Se and Sp are each at least 99%. Outside of these ranges, using the proportion of positive test results to estimate seroprevalence will be too biased, and more sophisticated

analysis methods should be used.

As an example of this, take the Ciao Corona study [16], a school-based longitudinal study of seroprevalence in Swiss school children with 5 rounds of SARS-CoV-2 antibody testing between June 2020 and June 2022, covering a range of seroprevalences in the population (Trial Registration: ClinicalTrials.gov NCT04448717). The study was approved by the Ethics Committee of the Canton of Zurich, Switzerland (2020-01336). All participants provided written informed consent before being enrolled in the study. The antibody test used has a sensitivity of 94% in children, and a specificity of 99.2%. In June 2020, 98 / 2473 (4.0%) of subjects showed as seropositive, compared to 154 / 2500 (6.2%) in October 2021, 17.3% (426 / 2453) in March 2021, 48.5% (910 / 1876) in November 2021, and 94.5% (2008 / 2125) in June 2022. Given the diagnostic test characteristics, absolute bias can be expected to be less than 1% in the range of 0% - 26.5% disease prevalence, and less than 2% for disease prevalence of up to 41.2%. These results imply that reported seroprevalence estimates based on a naive logistic approach are likely relatively unbiased for the first 3 rounds of Ciao Corona antibody testing (0.5%, 0.4% and -0.4% respectively), but that after that any seroprevalence estimates that do not adjust for test characteristics are likely quite biased (-2.4% and -5.6%). In order to adjust for covariates and survey sampling weights, we corrected the seroprevalence estimates using a Bayesian hierarchical model approach in all rounds of testing.

4 Discussion

I have demonstrated that average bias in prevalence estimates can be higher than desired, as high as 10%, when using a naive approach of calculation based on the proportion of positive test results, even if sensitivity and specificity are 90% or higher. Further, I have derived a range of disease prevalence values for which the naive approach gives reasonably unbiased prevalence estimates. A brief look into the literature indicates that many public health researchers are not aware of methods for reducing this potential bias, and do not correct for this in their own studies of prevalence. Nor do peer reviewers and editors seem to notice this widespread problem.

Taken together, the results emphasize the necessity in public health research to not simply report raw proportions of positive tests, even if those are adjusted for demographic characteristics using logistic regression. Since disease prevalence is of course not known precisely prior to study conduct, the most straightforward approach is then to plan statistical methods so that sensitivity and specificity are accounted for. Even if other sources of bias (e.g. sampling bias, or sampling variation) are accounted for, the results of seroprevalence studies will continue to be biased if analyses do not also account for test sensitivity and specificity. Care should also be taken in reading publications reporting (sero)prevalence estimates to ensure that suitable statistical methods have been used.

These results are based on the definitions of sensitivity and specificity only and require no complicated derivations. I have not adjusted for demographic characteristics, such as age and gender, or used weighting to approximate the target population, as is typical in surveys of disease prevalence. However, such adjustment cannot alleviate any general concerns of bias as presented here. The bias demonstrated here is also an average bias, and observed bias may vary more or less depending on the size of the sample. The results do not account for other possible issues with a diagnostic test [17, 18, 19], that can often not be corrected with statistical methods (e.g. when the validation sample, on which the sensitivity and specificity estimates are based, is not similar to the population of interest). Average bias is given, which does not account for sampling bias or variation, as has been described elsewhere [8]. Additionally, the

examination of publications reporting on seroprevalence studies in the literature was a sample of relevant studies, not a systematic review, and therefore the actual proportion of studies reporting only naive seroprevalence estimates may be somewhat different than what has been described here.

The question remains as to how best to account for diagnostic test sensitivity and specificity when estimating disease prevalence. A nice outline of some appropriate methods along with implementation in R [20] code is given by [21, 22, 10]. To calculate corrected confidence intervals for prevalence in studies where covariates do not need to be adjusted for, and no survey weights are needed, the R package `bootComb` [23] and website "epitools" (`https://epitools.ausvet.com.au/trueprevalence`) are available, while Bayesian methods are available in `prevalence` [24]. Using the Rogan-Gladen correction with bootstrap confidence intervals, or the Bayesian correction in the `prevalence` package are appropriate when there is no need to adjust for any other factors. Adjusting for covariates, adjusting for sampling bias or variation, or application of post-stratification weights (among other issues) may unfortunately need to be done without the use of such prepackaged code. Collaboration with experienced statisticians is invaluable in ensuring that correct analysis techniques are used so that unbiased prevalence estimates can be reported.

Conflict of Interest Statement

No conflicts of interest have been declared.

Funding statement

The Ciao Corona study, used in our example, is part of Corona Immunitas research network, coordinated by the Swiss School of Public Health (SSPH+), and funded by fundraising of SSPH+ that includes funds of the Swiss Federal Office of Public Health and private funders (ethical guidelines for funding stated by SSPH+ will be respected), by funds of the Cantons of Switzerland (Vaud, Zurich, and Basel) and by institutional funds of the Universities. Additional funding, specific to this study is available from the University of Zurich Foundation. No additional funding was acquired for the analysis presented here.

Contributions

SRH initiated the analysis, developed the methodology, performed the statistical analysis, and wrote the manuscript. No others meeting criteria for authorship have been omitted.

Acknowledgements

Thank you to Thomas Radtke, Julia Braun and Milo Puhan for their critical comments, and to Agnė Ulytė for suggesting the literature review.

References

[1] I Bergeri, MG Whelan, H Ware, L Subissi, A Nardone, HC Lewis, Z Li, X Ma, M Valenciano, B Cheng, L Al Ariqi, A Rashidian, J Okeibunor, T Azim, P Wijesinghe, L-V Le, A Vaughan, R Pebody, A Vicari, T Yan, M Yanes-Lane, C Cao, DA Clifton, MP Cheng, J Papenburg,

D Buckeridge, N Bobrovitz, RK Arora, MD Van Kerkhove, and Unity Studies Collaborator Group. Global SARS-CoV-2 seroprevalence from January 2020 to April 2022: A systematic review and meta-analysis of standardized population-based studies. *PLOS Medicine*, 19(11):1–24, November 2022.

[2] DG Altman and JM Bland. Statistics notes: Diagnostic tests 1: sensitivity and specificity. *BMJ*, 308(1552), 1994.

[3] L Joseph, TW Gyorkos, and L Coupal. Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard. *American Journal of Epidemiology*, 141(3):263–72, 1995.

[4] WJ Rogan and B Gladen. Estimating prevalence from the results of a screening test. *American Journal of Epidemiology*, 107(41), 1978.

[5] RA Lew and PS Levy. Estimation of prevalence on the basis of screening tests. *Statistics in Medicine*, 8:1225–1230, 1989.

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

[7] L Böttcher, M R D'Orsogna, and T Chou. Using excess deaths and testing statistics to determine COVID-19 mortalities. *European Journal of Epidemiology*, 36(5):545–558, 2021.

[8] L Böttcher, MR D'Orsogna, and T Chou. A statistical model of covid-19 testing in populations: effects of sampling bias and testing errors. *Philosophical Transactions of the Royal Society A*, 380(2214):20210121, 2022.

[9] P N Patrone and A J Kearsley. Classification under uncertainty: data analysis for diagnostic antibody testing. *Mathematical Medicine and Biology: A Journal of the IMA*, 38(3):396–416, 2021.

[10] M Flor, M Weiss, T Selhorst, C Müller-Graf, and M Greiner. Comparison of Bayesian and frequentist methods for prevalence estimation under misclassification. *BMC Public Health*, 2020.

[11] National Institute of Mental Health. What is prevalence? https://www.nimh.nih.gov/health/statistics/what-is-prevalence, May 2023. accessed 31.05.2023.

[12] D McConnell, C Hickey, N Bargary, L Trela-Larsen, C Walsh, M Barry, and R Adams. Understanding the challenges and uncertainties of seroprevalence studies for sars-cov-2. *Int. J. Environ. Res. Public Health*, 18(9):4640, 2021.

[13] MS Pepe. *The Statistical Evaluation of Medical Tests for Classification and Prediction*. OUP, 2003.

[14] J Hilden. A further comment on "estimating prevalence from theresults of a screening test'. *American Journal of Epidemiology*, 109(6):721–3, 1979.

[15] EI Kritsotakis. On the importance of population-based serological surveys of SARS-CoV-2 without overlooking their inherent uncertainties. *Public Health in Practice*, 1:10013, 2020.

[16] SR Haile, A Raineri, S Rueegg, T Radtke, A Ulytė, MA Puhan, and S Kriemler. Heterogeneous evolution of SARS-CoV-2 seroprevalence in school-age children: Results from the school-based cohort study Ciao Corona in November-December 2021 in the canton of Zurich. *Swiss Medical Weekly*, 153(1):40035, 2023.

[17] S Takahashi, B Greenhouse, and I Rodriguez-Barraquer. Are seroprevalence estimates for severe acute respiratory syndrome coronavirus 2 biased? *Journal of Infectious Diseases*, 222:1772–5, 2020.

[18] S Burgess, MJ Ponsford, and D Gill. Are we underestimating seroprevalence of SARS-CoV-2? *BMJ*, 370(m3364), 2020.

[19] EK Accorsi, X Qiu, E Rumpler, L Kennedy-Shaffer, R Kahn, K Joshi, E Goldstein, MJ Stensrud, R Niehus, M Cevik, and M Lipsitch. 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, 2021.

[20] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2022.

[21] FI Lewis and PR Torgerson. A tutorial in estimating the prevalence of disease in humans and animals in the absence of a gold standard diagnostic. *Emerging Themes in Epidemiology*, 9(9), 2012.

[22] PJ Diggle. Estimating prevalence using an imperfect test. *Epidemiology Research International*, 608719, 2011.

[23] MYR Henrion. bootComb—an R package to derive confidence intervals for combinations of independent parameter estimates. *International Journal of Epidemiology*, 50(4):1071–76, 2021.

[24] Brecht Devleesschauwer, Paul Torgerson, Johannes Charlier, Bruno Levecke, Nicolas Praet, Sophie Roelandt, Suzanne Smit, Pierre Dorny, Dirk Berkvens, and Niko Speybroeck. *prevalence: Tools for prevalence assessment studies.*, 2022. R package version 0.4.1.

Tables and Figures

Table 1: Typical example of 2x2 table comparing diagnostic test results and disease status.

	Test neg	Test pos	total
Disease neg	630	70	700
Disease pos	15	285	300
total	645	355	1000

Table 2: Estimated bias (in percentage points) for selected combinations of sensitivity, specificity and disease prevalence.

Se	Sp	P = 2%	P = 10%	P = 30%	P = 50%	P = 90%	P = 98%
99%	99%	1.0%	0.8%	0.4%	0.0%	-0.8%	-1.0%
99%	95%	4.9%	4.4%	3.2%	2.0%	-0.4%	-0.9%
99%	90%	9.8%	8.9%	6.7%	4.5%	0.1%	-0.8%
95%	99%	0.9%	0.4%	-0.8%	-2.0%	-4.4%	-4.9%
95%	95%	4.8%	4.0%	2.0%	0.0%	-4.0%	-4.8%
95%	90%	9.7%	8.5%	5.5%	2.5%	-3.5%	-4.7%
90%	99%	0.8%	-0.1%	-2.3%	-4.5%	-8.9%	-9.8%
90%	95%	4.7%	3.5%	0.5%	-2.5%	-8.5%	-9.7%
90%	90%	9.6%	8.0%	4.0%	0.0%	-8.0%	-9.6%

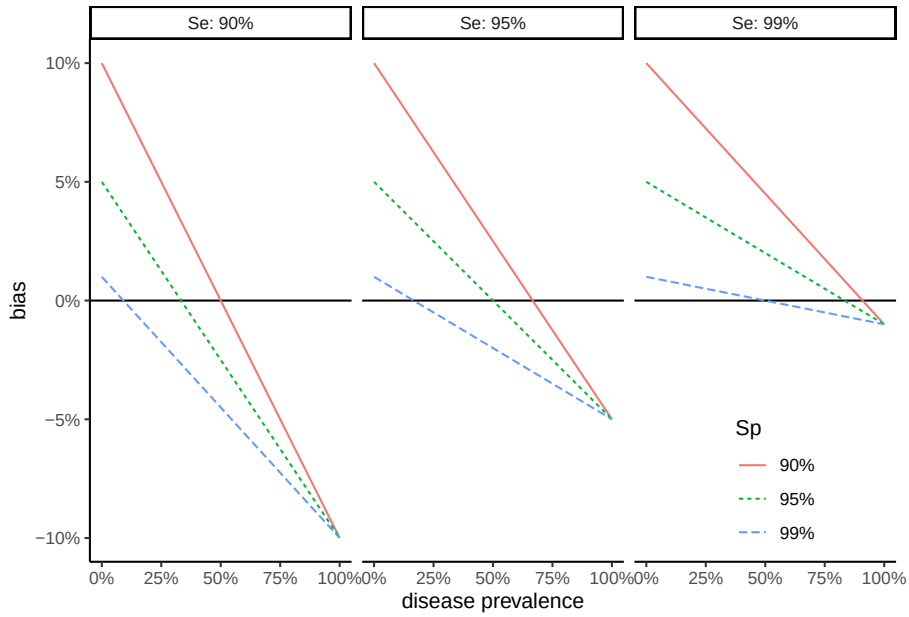

Figure 1: Estimated bias in prevalence estimate for selected combinations of sensitivity, specificity and true disease prevalence

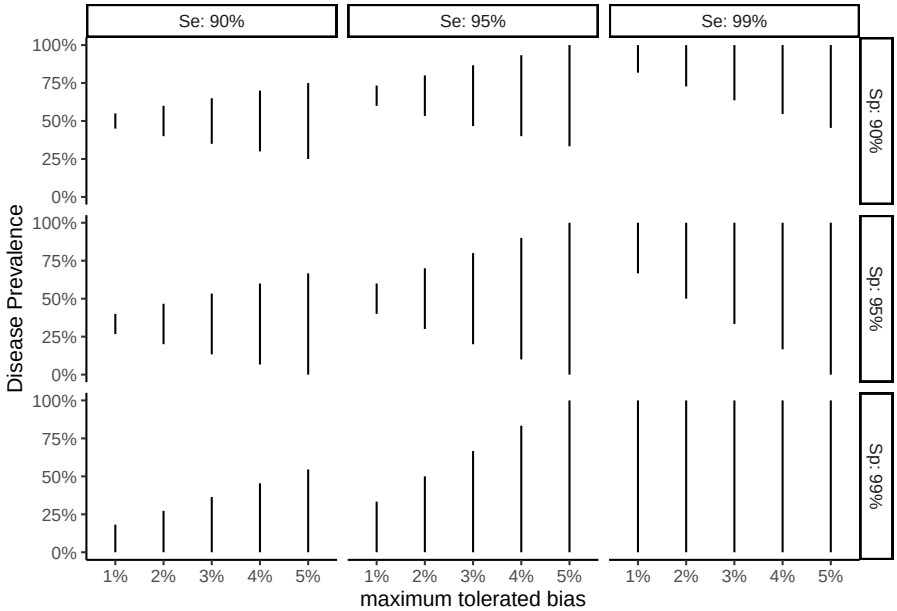

Figure 2: Range of true disease prevalence where the rate of positive tests is a close approximation of disease prevalence, to within maximum absolute tolerated bias