medRxiv preprint doi: https://doi.org/10.1101/2021.11.12.21266254; this version posted November 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

Revisiting the estimation of Covid-19 prevalence: Implications for rapid testing

Lili Zhou [*1], Daniel Andrés Díaz–Pachón [†1], and J. Sunil Rao [‡1]

[1]Division of Biostatistics - University of Miami, Don Soffer Clinical Research Center, 1120 NW 14th St, Miami FL, 33136

November 12, 2021

Abstract

Surveillance studies for Covid-19 prevalence estimation are subject to sampling bias due to oversampling of symptomatic individuals and error-prone tests, particularly rapid antigen tests which are known to have high false negative rates for asymptomatic individuals. This results in naïve estimators which can be very far from the truth.In this work, we present a method that removes these two sources of error directly. Moreover, our procedure can be easily extended to the stratified error situation in which a test has very different error rate profiles for symptomatic and asymptomatic individuals as is the case for rapid antigen testing. The result is an easily understandable four-step algorithm that produces much more reliable prevalence estimates as demonstrated on data from the Israeli Ministry of Health. Thus it may re-open the debate about whether we are under-valuing rapid testing as a surveillance tool and may have policy implications in Third-World countries or disadvantaged communities where access to PCR testing may be less accessible.

1 Introduction

Surveillance testing for COVID-19 remains an effective strategy for understanding viral spread in a population even as the vaccines have changed the focus of most media coverage. Since the virus will likely eventually enter a state of endemicity, the need for testing will not vanish. In fact, it will be a necessary tool in order to understand when spikes arise and more vigilance to contain spread is needed.

[*]lxz516@miami.edu
[†]Ddiaz3@miami.edu
[‡]JRao@miami.edu

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2021.11.12.21266254; this version posted November 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

COVID-19 testing usually comes in one of three possible ways: serological, rapid antigen, or PCR (so-called molecular) testing. Serological tests, by their counting of antibodies are more apt to detect past COVID-19 infection, and rapid antigen tests and PCR tests can detect current viral infection. However, since rapid antigen tests are known to have very high false negative rates amongst asymptotic individuals,[1] they have fallen out of favor as as surveillance tool and are preferentially used now to test symptomatic individuals and to determine whether that individual is infectious or not. This is a shame because rapid tests can be much more easily deployed, can be administered at home, do not require lab-based assays, and are much more cost effective. This could have made them ideal tools to use particularly in Third-World countries, which remain vulnerable to the virus due to persistently low vaccination rates.

As for sampling strategies, contact-tracing ensures that individuals who were in close contact to other individuals who tested positive are also tested. But contact tracing has important drawbacks, either if implemented manually or digitally,[2, 3] and it has raised questions on privacy and individuals liberties.[4] The world has thus relied mostly upon convenience or surveillance sampling (i.e. not random sampling) to estimate the prevalence of the disease. Convenience here implies that typically there is over-sampling of symptomatic individuals, and since the probability of testing positive for these individuals is higher than for asymptomatic individuals, this results in an over-estimation of the population prevalence using the biased sample naïve prevalence estimate of the proportion in the sample who tested positive.

In our previous work we derived a correction to remove the bias described above.[5] However, we did not address the issue of imperfect testing (i.e. false positives and negatives). In this work, we provide a solution for both issues and importantly discover a correction framework when we examine the special case of stratified errors (by symptom status) which results in dramatic corrections to the naïve sample prevalence estimates — even in situations where the error rates are as high as for rapid testing. This then naturally begs the question if we have been under-valuing the role of rapid testing for surveillance of COVID-19 and re-opens the debate on whether such a tool could be deployed effectively to track the spread of the virus, particularly in Third-World countries.

Our methodology can then be summarized as follows: First there is a population prevalence. Second, a sample is taken from the population. Third, individuals in the sample are tested. With these testing totals prevalence is estimated. Now, the population prevalence is in general unknown, that is why we do sampling. However, convenience sampling is biased towards symptomatic individuals. Moreover, tests are imperfect so they have false positive and false negatives rates. Then the naïve estimator taken after testing has incorporated the bias from the sampling strategy and the errors from testing. The goal of this article is to correct these two sources of error.

To add some formality, consider a population $\mathcal{P}$ of size N that is divided into three categories: asymptomatic and non-infected individuals, $I_0^{(0)}$, with size $N_0^{(0)}$; asymptomatic and infected individuals, $I_0^{(1)}$, with size $N_0^{(1)}$; and

medRxiv preprint doi: https://doi.org/10.1101/2021.11.12.21266254; this version posted November 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

symptomatic and infected individuals, $I_1^{(1)}$, with size $N_1^{(1)}$. It is possible to have a fourth group of symptomatic and non-infected individuals, $I_1^{(0)}$, but we reasonably set it to have zero size ($N_1^{(0)} = 0$), since developing a constellation of COVID-19 specific symptoms and not having the disease is not common. Notice that the sum of all the individuals in these groups is N.

As for the sample, if an individual belongs to the category $I_s^{(i)}$, she will be tested with probability $p\left(I_s^{(i)}\right)$, for $s, i = 0, 1$. We also set $p\left(I_1^{(0)}\right)$ to be 0, since it was assumed that $N_1^{(0)} = 0$. Therefore, for the three non-empty categories, calling $N_T^{s,i}$ the number of individuals tested from the group $I_s^{(i)}$, we obtain

$$p\left(I_s^{(i)}\right) = \frac{N_T^{s,i}}{N_s^{(i)}}, \tag{1}$$

where $N_T^{1,0} = 0$.

With this setting, inspired by previous work on detection and correction of publication bias,[6] it is possible to obtain that the naïve estimators $\tilde{N}_s^{(i)}$ of totals for each group are (see the Methods):

$$\begin{aligned}
\tilde{N}_0^{(0)} &= (1-\alpha)N_T^{0,0} + \beta N_T^{0,1}, \\
\tilde{N}_0^{(1)} &= \alpha N_T^{0,0} + (1-\beta)N_T^{0,1}, \\
\tilde{N}_1^{(0)} &= \beta N_T^{1,1}, \\
\tilde{N}_1^{(1)} &= (1-\beta)N_T^{1,1},
\end{aligned} \tag{2}$$

with α being the probability of a false positive, and β, the probability of a false negative. This set of equations is very intuitive. For instance, the naïve group of asymptomatic and non-infected is formed by the real group of asymptomatic and non-infected who were not false positives in the test and the group of asymptomatic and infected individuals in the population who were false negatives. Interestingly, the third group, corresponding to the naïve estimator of symptomatic and non-infected individuals is nonzero, since false negatives will make the naïve estimate positive.

On the other hand, notice that the naïve estimators in the left-hand side are determined by the testing errors and the sampled individuals of each group. Therefore, if the sample is not random, there will be bias. This is easy to see when we assume that there is no error in testing, in whose case the previous naïve estimators are the total sampled from each group. In fact, to reflect our initial hypothesis of overrepresentation in the sample of the symptomatic group, we assume that $p\left(I_1^{(1)}\right) \geq q \geq 1/2$. Thus, q is a parametric value saying that at least half the symptomatic individuals were sampled.

Notice that the assumption $q \geq 1/2$ is not unreal in convenience testing, particularly in developed societies. In fact, most universities and large companies in the US require now that all symptomatic individuals get tested, with which $p\left(I_1^{(1)}\right) = 1$. However, underdeveloped populations (countries mainly, but also

medRxiv preprint doi: https://doi.org/10.1101/2021.11.12.21266254; this version posted November 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

possibly subpopulations in developed countries, like illegal immigrants) might not achieve this goal. The fact that, as we will see below, the correction, even in the presence of large false-negatives rates, is so effective, should motivate sampling at least half of the symptomatic individuals to allow the estimation of population prevalence.

Now, provided $\beta \leq 1-\alpha$, we propose the corrected estimators of prevalence $\hat{N}_s^{(1)}$ in Algorithm 1. Some comments are in order:

Algorithm 1 Estimation of total number of infected

1. Estimate $\hat{N}_T^{1,1}$ as $\tilde{N}_1^{(0)} + \tilde{N}_1^{(1)}$.
2. For $1/2 \leq q \leq 1$, take $\hat{N}_1^{(1)}$ as a uniform random variable in the set
$$\mathbf{U} = \{\hat{N}_T^{1,1}, \hat{N}_T^{1,1}+1, \ldots, \lfloor q^{-1}\hat{N}_T^{1,1} \rfloor\}, \tag{3}$$
where $\lfloor \cdot \rfloor$ is the integer part.
3. Make
$$\hat{N}_T^{0,0} = \frac{\tilde{N}_0^{(0)} - \beta\tilde{N}_0}{1-\alpha-\beta},$$
$$\hat{N}_T^{0,1} = \frac{\alpha\tilde{N}_0 - \tilde{N}_0^{(1)}}{\alpha+\beta-1},$$
where $\tilde{N}_0$ is the number of asymptomatic individuals in the sample.
4. Take $\hat{N}_0^{(1)} = \frac{\hat{N}_T^{0,1}}{N_T - \hat{N}_T^{1,1}} N$, with N_T being the sample size.
5. Take $\hat{N}^{(1)}$, the estimate number of infected, as $\hat{N}_1^{(1)} + \hat{N}_0^{(1)}$.

With respect to the first step, its estimator is very easy to see from equations (2).

With respect to the second step, had we known that all the symptomatic group was sampled, as Diaz and Rao did, then $\hat{N}_1^{(1)} = N_T^{1,1}$.[5] The methodology here is more realistic. Given our imperfect knowledge, the only unbiased assumption we can make is that $\hat{N}_1^{(1)}$ is uniformly distributed in the set $\mathbf{U}$.[7, 8, 9] Any other distribution, unless additional knowledge is at hand, will introduce bias.

With respect to the third step, it is important to notice that $N_T^{0,0}$ and $N_T^{0,1}$ do *n*ot depend on $\hat{N}_1^{(1)}$. However, they do depend on the restriction on β and α which is in general satisfied, even with rapid antigen testing.

As for the fourth step, since we do not have a reason to think otherwise, we assume that the sample is random for the asymptomatic individuals. $\hat{N}_0^{(1)}$ is thus estimated accordingly.

medRxiv preprint doi: https://doi.org/10.1101/2021.11.12.21266254; this version posted November 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

What the researcher observes are the total of positive and negative individuals among symptomatic and asymptomatic ones after testing. The real prevalence is unknown to her, as is the sampling scheme. Thus the process to correct the estimator runs backwards: starting with the naïve estimator (that includes errors and bias), we then recover the sampling proportions (getting rid of the errors), to finally produce the correct estimator (getting rid of the bias).

1.1 Stratified errors

Notice we can make a stratification of errors by group of symptoms. In this case the naïve estimators become:

$$\begin{aligned}
\tilde{N}_0^{(0)} &= (1-\alpha_0)N_T^{0,0} + \beta_0 N_T^{0,1}, \\
\tilde{N}_0^{(1)} &= \alpha_0 N_T^{0,0} + (1-\beta_0)N_T^{0,1}, \\
\tilde{N}_1^{(0)} &= \beta_1 N_T^{1,1}, \\
\tilde{N}_1^{(1)} &= (1-\beta_1)N_T^{1,1},
\end{aligned} \tag{4}$$

In this scenario, the correction is analogous to Algorithm 1, with the sole difference that α and β now become α_0 and β_0 for the observed values in the asymptomatic group, as made explicit in Algorithm 2. Notice how this framework allows for the unique error profiles of rapid testing where accuracies vary greatly between asymptomatic and symptomatic individuals. For rapid testing, β_0 can be as high as 50% and β_1 on the order of 10%.[1] Typically, α_1 and α_0 remain small.

2 Data from the Israeli Ministry of Health

In what follows, we consider data from the Israeli Ministry of Health.[10] Accordingly, we start with a sample of 99,232 tested individuals of whom 1,862 were symptomatic (have shortness of breath or have at least three of four symptoms: cough, fever, sore throat, and headache) and 97,370 were asymptomatic. Among the total tested individuals, it was possible to identify 8,393 infections through PCR testing. Among the individuals who tested positive, 1,754 were symptomatic. The characteristics of the data set are presented in Table 1.

	Positive	Negative	Total
Symptomatic	$\tilde{N}_1^{(1)} = 1754$	$\tilde{N}_1^{(0)} = 108$	$\tilde{N}_1 = 1862$
Asymptomatic	$\tilde{N}_0^{(1)} = 6639$	$\tilde{N}_0^{(0)} = 90731$	$\tilde{N}_0 = 97370$
Total	$\tilde{N}^{(1)} = 8393$	$\tilde{N}^{(0)} = 90839$	$\tilde{N} = 99232$

Table 1: Observed disease status by category of symptoms

medRxiv preprint doi: https://doi.org/10.1101/2021.11.12.21266254; this version posted November 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

Algorithm 2 Estimation of total number of infected with errors by stratum

1. Estimate $\hat{N}_T^{1,1}$ as $\tilde{N}_1^{(0)} + \tilde{N}_1^{(1)}$.
2. For $1/2 \leq q \leq 1$, take $\hat{N}_1^{(1)}$ as a uniform random variable in the set
$$\mathbf{U} = \{\hat{N}_T^{1,1}, \hat{N}_T^{1,1} + 1, \ldots, \lfloor q^{-1}\hat{N}_T^{1,1} \rfloor\}, \tag{5}$$
where $\lfloor \cdot \rfloor$ is the integer part.
3. Make
$$\hat{N}_T^{0,0} = \frac{\tilde{N}_0^{(0)} - \beta_0 \tilde{N}_0}{1 - \alpha_0 - \beta_0},$$
$$\hat{N}_T^{0,1} = \frac{\alpha_0 \tilde{N}_0 - \tilde{N}_0^{(1)}}{\alpha_0 + \beta_0 - 1},$$
where $\tilde{N}_0$ is the number of asymptomatic individuals in the sample.
4. Take $\hat{N}_0^{(1)} = \frac{\hat{N}_T^{0,1}}{N_T - \hat{N}_T^{1,1}} N$, with N_T being the sample size.
5. Take $\hat{N}^{(1)}$, the estimate number of infected, as $\hat{N}_1^{(1)} + \hat{N}_0^{(1)}$.

We also know that PCR testing has false negatives rates $\beta = 0.26$ and false positives rate $\alpha = 0.003$.[11] Therefore we can obtain the real number of individuals for each group, as in Table 2.

	Disease	Non-Disease	Total
Symptomatic	1777	85	1862
Asymptomatic	30209	67161	97370
Total	31986	67246	99232

Table 2: Real total of individuals per group

If a whole population is tested with PCR (as in this data set) and the error rates for PCR are known (and we know them), just by correcting for the errors we can obtain the real values in the population. We then will assume that Table 2 represents a population total and we will take samples from it. Accordingly, the real prevalence is

$$p = 31986/99232 = 0.32, \tag{6}$$

and prevalence among the asymptomatic is $30209/97370 = 0.31$. Finally, calling $\tilde{p}$ the naïve estimator, and $\hat{p}$ the corrected estimator, we define the ratio of absolute errors,

$$\text{RAE} = \frac{|\tilde{p} - p|}{|\hat{p} - p|},$$

medRxiv preprint doi: https://doi.org/10.1101/2021.11.12.21266254; this version posted November 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

for $\hat{p} \neq p$. This ratio will be larger than 1 when the correction works better than the naïve estimator, will be less than 1 when the naïve estimator does a better job than the correction, and will be 1 when the two estimators behave similarly.

In what follows we will assume $q = 1/2$. The uniform random variable in step 2 of Protocols 1 and 2 will be replaced by its expected value: $1.5 \times \hat{N}_T^{1,1}$.

2.1 Without stratified errors

Sampling Protocol 1: Here the sample is made 100% of symptomatic individuals. We also assume that all the symptomatic group was sampled. Thus, the sample consists of 1862 individuals. Table 3 summarizes this information.

	Positive	Negative	Total
Symptomatic	1754	108	1862
Asymptomatic	0	0	0
Total	1754	108	1862

Table 3: 100% symptomatic individuals tested.

In this extreme case, the naïve estimate of prevalence is $\tilde{p} = 0.94$. The total corrected prevalence is $\hat{p} = (1862/99232)1.5 \approx 0.028$. Then the ratio of absolute errors is

$$\text{RAE} = \frac{|0.94 - 0.32|}{|0.32 - 0.028|} = 2.12,$$

which shows that even for so bad a sample the correction behaves better than the naïve estimator.

Sampling Protocol 2: This sample is made of 75% symptomatic individuals, and 25% asymptomatic ones. We also assume that all the 1862 symptomatic individuals were sampled. Then the asymptomatic sampled are 621. Since we do not have more information about the asymptomatic group, we assume that they are sampled at random from the population (Table 2); then 199 are asymptomatic disease positive and 422 are asymptomatic disease negative. However, the *observed values* (based on testing, not on true disease status) are given in Table 4.

	Positive	Negative	Total
Symptomatic	1754	108	1862
Asymptomatic	149	472	621
Total	1903	580	2483

Table 4: Observed values when 75% symptomatic and 25% asymptomatic individuals are tested.

The values of Table 4 were obtained using Equation (2). Notice however for the symptomatic group that, since all its individuals were tested, the first row is identical to that of Table 1.

medRxiv preprint doi: https://doi.org/10.1101/2021.11.12.21266254; this version posted November 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

In this case, the naïve estimate of prevalence, obtained from Table 4, is $\tilde{p} = 0.766$. As for the correction, following Algorithm 1, we obtain that $\hat{p} = 0.028 + 0.32 = 0.348$, where the first term is the estimated prevalence among the symptomatic, and the second, prevalence among the asymptomatic.

The ratio of absolute errors now becomes:

$$\text{RAE} = \frac{|0.766 - 0.32|}{|0.348 - 0.32|} = 15.93,$$

with which the superiority of the correction is clearly seen.

Sampling Protocol 3: This sample contains 50% symptomatic and 50% asymptomatic individuals. Since we again assume that all the symptomatic were sampled, we have a sample of size 3724. Under a similar analysis to that of the previous scenario, the observed values are presented in Table 5.

	Positive	Negative	Total
Symptomatic	1754	108	1862
Asymptomatic	431	1431	1862
Total	2185	1539	3724

Table 5: 50% symptomatic and 50% asymptomatic in the sample

The naïve estimate is thus $\tilde{p} = 2185/3724 = 0.59$. And using Algorithm 1, we obtain that the corrected estimate is again $\hat{p} = 0.34$. In this case, the ratio of absolute errors becomes

$$\text{RAE} = \frac{|0.59 - 0.32|}{|0.34 - 0.32|} = 13.5.$$

Again, a huge improvement when using the correction.

Sampling Protocol 4: This sample is truly random, with $N_T = 50,000$. Using the proportions from Table 2 to obtain $N_T^{1,*}$, $N_T^{0,0}$ and $N_T^{0,1}$, and then Equations 2, after testing we obtain the observations in Table 6.

	Positive	Negative	Total
Symptomatic	830	108	938
Asymptomatic	11,365	37,697	49,062
Total	12,195	37,941	50,000

Table 6: Random sample of size 50,000.

Equations (2) actually produce that the number of symptomatic observed positive are 244, however, we restrict them to 108 because that is the number observed in Table 1. The naïve estimate out of Table 6 is $\tilde{p} = 12,195/50,000 = 0.2439$.

In this case, $\hat{p}_1 = (938/99,232)1.5 = 0.014$, and $\hat{p}_0 = 0.31$. Therefore, the total corrected estimator is $\hat{p} = 0.324$. The ratio of absolute errors will be:

$$\text{RAE} = \frac{|0.32 - 0.2439|}{|0.324 - 0.32|} = 19.025.$$

medRxiv preprint doi: https://doi.org/10.1101/2021.11.12.21266254; this version posted November 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

2.2 With stratified errors

In some more realistic scenarios like doing rapid antigen testing, considering stratifying errors, we give the notations that α_0 and α_1 are the false positive rate for asymptomatic and symptomatic individuals, respectively, and β_0 and β_1 are the false negative rate for asymptomatic and symptomatic individuals, respectively. In our sampling protocols, the values are given by $\alpha_0 = 0.1\%, \alpha_1 = 0.5\%, \beta_0 = 50\%$, and $\beta_1 = 10\%$. Therefore, analogous to what we did with Table 2 previously, we can obtain a new table showing as the totals for this new set of errors. Table 7 summarizes this information.

	Disease	Non-Disease	Total
Symptomatic	1756	106	1862
Asymptomatic	51,998	45,372	97,370
Total	53,754	45,478	99,232

Table 7: Real proportions under stratified errors with $\alpha_0 = 0.1\%, \alpha_1 = 0.5\%, \beta_0 = 50\%$, and $\beta_1 = 10\%$.

According to Table 7, the true prevalence with stratifying errors is $p = 0.54$. 94% individuals are infected in the symptomatic group, and 53% are infected in the asymptomatic group.

Sampling Protocol 1: The sample consists of the 1862 symptomatic individuals. In this extreme case, the naïve estimate of prevalence is $\tilde{p} = 1$. Thus, the observed totals are presented in Table 8.

	Positive	Negative	Total
Symptomatic	1754	108	1862
Asymptomatic	0	0	0
Total	1754	108	1862

Table 8: Stratified sample with only symptomatic individuals.

According to Table 8, the biased estimator is $\tilde{p} = 0.94$. The correction is $\hat{p}(1754/99,232)1.5 = 0.0265$. With this information, the ratio of absolute errors is

$$\text{RAE} = \frac{|0.94 - 0.54|}{|0.54 - 0.00265|} = 0.74.$$

We see that, due to the fact that the heavy false negative rate corrects a little the very bad features of the sample, the naïve estimator does slightly better than the corrected estimate.

Sampling Protocol 2: The sample consists of 2483 individuals. Among these, 1862 (75%) are symptomatic and 621 (25%) are asymptomatic. Since all the symptomatic group was sampled and tested, the observations for this group coincide with those of Table 1. As for the asymptomatic group, the sampling proportions are taken according to Table 7, with which, according to Equations (4), we observe the naïve totals in Table 9.

medRxiv preprint doi: https://doi.org/10.1101/2021.11.12.21266254; this version posted November 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

	Positive	Negative	Total
Symptomatic	1754	108	1862
Asymptomatic	194	427	621
Total	1948	535	2483

Table 9: Stratified sample with 75% symptomatic and 25% asymptomatic.

Then the naïve estimate of prevalence is $\tilde{p} = 1948/2483 = 0.78$.

Using $\hat{p}_1$ is 0.0265 and $\hat{p}_0$ is 0.53. Therefore, $\hat{p} = 0.5565$. The ratio of absolute errors is thus

$$\text{RAE} = \frac{|0.78 - 0.54|}{|0.54 - 0.5565|} = 14.54,$$

which again shows the huge improvements of the correction with respect to the biased estimator.

Sampling Protocol 3: In this scenario we have 1862 symptomatic and 1862 asymptomatic individuals. Again Table 1 tells the behavior of the symptomatic when tested. As for the asymptomatic, opposite to the proportions in Table 7, and the previous example, we now make a twist and assume that 53% asymptomatic non-infected and 47% asymptomatic infected were tested. However, remember, this is unknown to the observer, since all she can see is Table 10), obtained from Equations (4).

	Positive	Negative	Total
Symptomatic	1754	108	1862
Asymptomatic	536	1326	1862
Total	2290	1434	3724

Table 10: Stratified observed totals with 50% symptomatic and 50% asymptomatic.

The naïve estimate is $\tilde{p} = 2290/3724 \approx 0.614$. The corrected estimate, using Algorithm 2, is $\hat{p} = 0.0265 + 0.44 = 0.4665$. In this case, the ratio of absolute errors becomes

$$\text{RAE} = \frac{|0.614 - 0.54|}{|0.54 - 0.4665|} \approx 1.$$

Therefore, in this scenario both estimators behave almost identically.

Sampling Protocol 4: This sample is also truly random. Say $N_T = 30,000$. Among these, $30,000(1862/99,232) \approx 563$ are symptomatic. Then, 29,437 are asymptomatic. We use Table 7 to determine the proportions sampled per group for asymptomatic, obtaining $N_T^{0,1} = 29,437(0.53) = 15,602$ infected and $N_T^{0,0} = 13,835$. With these values we map back to the observations in Table 11.

medRxiv preprint doi: https://doi.org/10.1101/2021.11.12.21266254; this version posted November 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

	Positive	Negative	Total
Symptomatic	507	56	563
Asymptomatic	9134	20,303	29,437
Total	9641	20,359	30,000

Table 11: Stratified observed totals from a random sample of size 30,000.

The naïve estimate is $\tilde{p} = 0.32$.

As for the correction, $\hat{p}_1 = (563/99,232)1.5 = 0.008$ and $\hat{p}_0 = 15,602/(30,000 - 563) = 0.53$. Therefore, $\hat{p} = 0.538$. With this, the ratio of absolute errors is

$$\text{RAE} = \frac{|0.32 - 0.54|}{|0.54 - 0.538|} = 110.$$

3 Methods

For $s, i = 0, 1$, define $p_s^{(i)}$ as $N_s^{(i)}/N$, that is, the proportion of individuals with symptoms s and infection status i. More formally, we can define a random element S^* taking values in the set $\mathbf{I} = \left\{I_0^{(0)}, I_0^{(1)}, I_1^{(1)}\right\}$, with density given by

$$f_{S^*}\left(I_s^{(i)}\right) = p_s^{(i)},$$

and $p_0^{(0)} + p_0^{(1)} + p_1^{(1)} = 1$.

For the j-th individual in the population ($0 < j \leq N$), we assume a Bernoulli random variable:

$$T_j \mid j \in I_s^{(i)} = \begin{cases} 1 & \text{with probability } p\left(I_s^{(i)}\right), \\ 0 & \text{with probability } 1 - p\left(I_s^{(i)}\right). \end{cases} \tag{7}$$

Calling N_T the sample size, it is easily seen that $N_T = \sum_{j=1}^{N} T_j$, and we can define an uncondtional binary random variable

$$T = \begin{cases} 1 & \text{with probability } N_T/N, \\ 0 & \text{with probability } 1 - N_T/N. \end{cases} \tag{8}$$

Now, if there is no error in sampling,

$$\tilde{N}_s^{(i)} = N_T^{s,i}, \tag{9}$$

so that the bias is induced by the number of individuals from $I_s^{(i)}$ in the sample. Seen from another perspective, we have the following proposition:

Proposition 1.

$$\tilde{p}_s^{(i)} = \frac{p\left(I_s^{(i)}\right)}{P[T = 1]} p_s^{(i)}$$

medRxiv preprint doi: https://doi.org/10.1101/2021.11.12.21266254; this version posted November 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

Proof. Notice that, by definition,

$$\tilde{p}_s^{(i)} = f_{S^*|T}\left(I_s^{(i)} \mid T = 1\right).$$

Therefore, after applying applying Bayes rule at the RHS, we obtain

$$\begin{aligned}\tilde{p}_s^{(i)} &= \frac{P\left[T = 1 \mid S^* = I_s^{(i)}\right]}{P[T=1]} f_{S^*}\left(I_s^{(i)}\right) \\ &= \frac{p\left(I_s^{(i)}\right)}{P[T=1]} p_s^{(i)} \\ &= \frac{N_T^{s,i}}{N_T}. \end{aligned} \tag{10}$$

□

From (10), it is clear that $N_s^{(i)}$, the population size of $I_s^{(i)}$, disappeared from the sample. Therefore, the importance of Proposition 1 is to show that all information we have in the sample about $N_s^{(i)}$ comes from the sampling mechanism $p\left(I_s^{(i)}\right)$. In fact, we can think of $p_s^{(i)}$ as the message sent, $\tilde{p}_s^{(i)}$ as the received message, and $p\left(I_s^{(i)}\right)/P[T=1]$ as the channel between them distorting the message.

Up to this point the analysis has been done for testing without errors. However, notice that (4) is obtained directly from (9), once we insert testing errors.

3.1 Correction

Since the group of symptomatic and non-infected individuals is empty, it is clear that $\tilde{N}_1^{(0)}$ is exclusively made of false negatives coming from the symptomatic and infected individuals. Therefore, the real number of symptomatic and infected individuals is $\tilde{N}_1^{(0)} + \tilde{N}_1^{(0)} = N_T^{1,1}$, the total of symptomatic individuals sampled. Notice that here β disappears from the analysis, so we can safely ignore it in our estimation of the number of individuals in $I_1^{(1)}$.

From Proposition 1, it is clear that if we do not know anything about $p\left(I_s^{(i)}\right)$, it will be impossible to correct the bias. Thus we need a reasonable assumption, like $p\left(I_1^{1)}\right) \geq q \geq 1/2$. With this assumption we use the principle of maximum entropy to obtain a correct estimator $\hat{N}_1^{(1)}$.

$\hat{N}_T^{0,1}$ and $\hat{N}_T^{0,0}$ in the third step of Algorithm 1 can be obtained from the first two equations in (4), since $\tilde{N}_0^{(1)}$ and $\tilde{N}_0^{(0)}$ are known. Therefore, we have two equations with two unknowns. The requirement of $\alpha \leq 1 - \beta$ ensures that the values of $\hat{N}_T^{0,1}$ and $\hat{N}_T^{0,0}$ are non-negative.

The estimator of the number of infected individuals among the asymptomatic is thus obtained after assuming that they were randomly sampled in what remains of the sample once the symptomatic individuals have been removed.

medRxiv preprint doi: https://doi.org/10.1101/2021.11.12.21266254; this version posted November 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

4 Discussion

There are a couple of limitations to our study. First, we have assumed throughout that error rates for tests are known a priori. If this is not the case, then at least under the random sampling situation, prevalence can still be estimated using a Bayesian approach described by Diggle.[12] This naturally results in increased variability of the prevalence estimate and relies on a reasonable prior distribution being elicited for the prevalence. This approach has not been extended to our situation here where sampling bias is also an issue. Second, we assumed that the symptomatic without the disease group is negligible and size of sampled symptomatic individuals is at least half the population value. As for the former, we expect this to not be violated for COVID-19.

Our correction proves to be very effective in many situations that would be encountered in practice. As we argued in the Introduction, this re-opens the debate about the utility of widespread rapid testing as a surveillance tool particularly in third world countries where PCR testing may be too expensive to implement widely. The scenario(s) where the correction does not improve upon the naïve estimate are those where the error rates are so large relative to the information in the sample that the correction is blurred and appears negligible.

Sample pooling has also been proposed as an efficient way to estimate population prevalence because if the disease prevalence is low, then little information is accrued from individual tests.[13] This is sometimes called group testing. However, this implicitly assumes random sampling of pools which is clearly not the case in what we are considering here.

Another approach that has been taken is to use population seroprevalence complex surveys. [14, 15] While inherently much more difficult to conduct and analyze, these can also suffer from non-ignorable non-response which can lead to biased estimates of prevalence. Indeed, biased sampling can be more generally cast within a missing data framework and the impact of different missing data mechanisms studied.

References

[1] Ian W. Pray and Laura Ford et al. Performance of an Antigen-Based Test for Asymptomatic and Symptomatic SARS-CoV-2 Testing at Two University Campuses – Wisconsin, September- October 2020. *MMWR Morbidity and Mortality Weekly Report*, 69(5152):1642–1647, 2021.

[2] Joel Hellewell and et al. Feasibility of controlling COVID-19 outbreaks by isolation of cases and contacts. *Lancet Global Health*, 8(4):e488–e496, 2020.

[3] Marco Mancastroppa, Claudio Castellano, Alessandro Vezzani, and Raffaella Burioni. Stochastic sampling effects favor manual over digital contact tracing. *Nature Communications*, 12(1919), 2021.

[4] Yoshua Bengio, Richard Janda, Yun William Yu, Daphne Ippolito, Max Jarvie, Dan Pilat, Brooke Struck, Sekoul Krastev, and Abhinav Sharma.

medRxiv preprint doi: https://doi.org/10.1101/2021.11.12.21266254; this version posted November 14, 2021. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

The need for privacy with public digital contact tracing during the COVID-19 pandemic. *The Lancet Digital Health*, 2(7):e342–e344, 2020.

[5] Daniel Andrés Díaz-Pachón and J. Sunil Rao. A simple correction for COVID-19 sampling bias. *Journal of Theoretical Biology*, 512:110556, March 2021.

[6] Isaiah Andrews and Maximilian Kasy. Identification of and Correction for Publication Bias. *American Economic Review*, 109(8):2766–2794, 2019.

[7] E. T. Jaynes. Information Theory and Statistical Mechanics. *Physical Review*, 106(4):620–630, May 1957.

[8] E. T. Jaynes. Information Theory and Statistical Mechanics II. *Physical Review*, 108(2):171–190, 1957.

[9] Daniel Andrés Díaz-Pachón and Robert J. Marks II. Generalized active information: Extensions to unbounded domains. *BIO-Complexity*, 2020(3):1–6, 2020.

[10] Yazeed Zoabi, Shira Deri-Rozov, and Noam Shomron. Machine learning-based prediction of COVID-19 diagnosis based on symptoms. *npj Digital Medicine*, 4(3), 2021.

[11] Andrew N. Cohen, Bruce Kessel, and Michael G. Milgroom. Diagnosing SARS-CoV-2 infection: the danger of over-reliance on positive test results. *Preprint*, 2020.

[12] P. J. Diggle. Estimating prevalence using an imperfect test. *Epidemiology Research International*, page 608719, 2011.

[13] Ola Brynildsrud. Covid-19 prevalence estimation by random sampling in population - optimal sample pooling under varying assumptions about true prevalence. *BMC Medical Research Methodology*, 20:196, 2020.

[14] Julio Morales Carabaña. Datos de encuesta para estimar la prevalencia de COVID-19. Un estudio piloto en Madrid capital. *Revista española de salud pública*, 94(17 de noviembre):e202011159, 2020.

[15] Vinícius Bonetti Franceschi, Andressa Schneiders Santos, and Andressa Barreto Glaeser et al. Population-based prevalence surveys during the Covid-19 pandemic: A systematic review. *Reviews in Medical Virology*, 31(4):e2200, 2021.