Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

ARTICLE TYPE

Correcting prevalence estimation for biased sampling with testing errors

Lili Zhou | Daniel Andrés Díaz-Pachón* | Chen Zhao | J. Sunil Rao

[1]Division of Biostatistics, University of Miami, Florida, United States of America

Correspondence
*Daniel Andrés Díaz-Pachón, 1120 NW 14th St, Room 1057, Miami, Florida 33136.
Email: Ddiaz3@miami.edu

Summary

Sampling for prevalence estimation of infection is subject to bias by both oversampling of symptomatic individuals and error-prone tests. This results in naïve estimators that can be very far from the truth. In this work, we present a method of prevalence estimation that removes the effect of testing errors and reduces the effect of oversampling symptomatic individuals. Moreover, this procedure considers stratified errors in which tests have different error rate profiles for symptomatic and asymptomatic individuals. The result is an easily implementable algorithm (for which code is provided) that produces better prevalence estimates than other methods, as demonstrated by simulation and on Covid-19 data from the Israeli Ministry of Health.

KEYWORDS:
Active information, bias correction, Covid-19, maximum entropy, prevalence, sampling, sampling bias, testing errors

1 | INTRODUCTION

Estimation of disease prevalence is challenging. First, imperfect testing always distorts actual proportions. Second, it's not uncommon to have to derive estimates from samples that under-represent or fail to capture subpopulations that are at greatest risk or of interest. An example is estimating the general population prevalence of chronic hepatitis C (HCV) because of the challenges of sampling from subpopulations of former and current injecting drug users, the homeless or incarcerated.[1] Other examples include the over-representation of symptomatic individuals in a sample since these individuals are more likely to get tested than asymptomatic ones, with which the final estimates of prevalence inflates, since symptomatic individuals are also more likely to be truly infected than asymptomatic ones.

This situation became clear during the recent Covid-19 pandemic: besides usual discussions of the error rates of PCR and rapid tests, surveillance mechanisms have usually relied on convenience sampling or contact tracing. Therefore sampling bias was also present. In the case of convenience sampling, because it passively waits for symptomatic individuals to get tested, whereas asymptomatic individuals have few reasons to do so. As for contact tracing, because it actively pursues infected individuals, ignoring the non-infected almost altogether. Besides this, contact tracing has also raised questions on privacy and individual liberties.[2,3,4] Though this example corresponds to a non-probability Covid-19 sampling setting, the problem is of course more general. It applies to every form of prevalence estimation performed through testing, either probabilistic or not.

Recently, Díaz-Pachón and Rao introduced a correction for oversampling of the symptomatic group.[5] It was a three-step procedure based on the assumption that all symptomatic individuals in the population were sampled and infected but it did not address the issue of imperfect testing (i.e. the presence of false positives and false negatives). This implies that the symptomatic and infected individuals in the sample corresponded to the total number of symptomatic individuals in the population. Thus the

asymptomatic group in the population was the complement of the total of symptomatic individuals in the sample. The prevalence among the asymptomatic group was then obtained as a uniform random variable among the asymptomatic individuals in the population, with no resource to the sample.

In this paper a method that is stronger in all aspects is presented. First, it does not assume that *all* symptomatic individuals are sampled, only that symptomatic individuals are overrepresented in the sample. Second, sample values among the asymptomatic are used to produce an estimator of prevalence that is informed by evidence. Third, testing errors are considered. And fourth, the proposed correction is extended to stratified errors by symptom status.

2 | SETTING

Consider a population $\mathcal{P}$ of size N that is divided into four categories: asymptomatic and non-infected individuals, $I_0^{(0)}$, with size $N_0^{(0)}$; asymptomatic and infected individuals, $I_0^{(1)}$, with size $N_0^{(1)}$; symptomatic and infected individuals, $I_1^{(1)}$, with size $N_1^{(1)}$; and symptomatic and non-infected individuals, $I_1^{(0)}$, with size $N_1^{(0)}$. The population total N is known, whereas $N_1^{(1)}$, $N_0^{(1)}$, $N_1^{(0)}$, and $N_0^{(0)}$ are unknown, though their sum is N.

The group of individuals with symptoms s in the population will be denoted by $I_s = I_s^{(0)} \cup I_s^{(1)}$, and its total by $N_s = N_s^{(0)} + N_s^{(1)}$, for $s = 0, 1$. Analogously, the group of individuals with infection status i in the population will be denoted by $I^{(i)} = I_0^{(i)} \cup I_1^{(i)}$, and its total by $N^{(i)} = N_0^{(i)} + N_1^{(i)}$, for $i = 0, 1$.

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

$$f_{S^*}\left(I_s^{(i)}\right) = p_s^{(i)}, \tag{1}$$

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

The proportion of individuals in the group I_s is then given by $p_s = p_s^{(0)} + p_s^{(1)}$, for $s = 0, 1$. And the proportion of individuals in the group $I^{(i)}$ is given by $p^{(i)} = p_0^{(i)} + p_1^{(i)}$, for $i = 0, 1$.

2.1 | Sampling probabilities

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

$$T_j \mid j \in I_s^{(i)} = \begin{cases} 1 & \text{with probability } p\left(I_s^{(i)}\right), \\ 0 & \text{with probability } 1 - p\left(I_s^{(i)}\right). \end{cases} \tag{2}$$

That is, an individual in the category $I_s^{(i)}$ will be tested with probability $p\left(I_s^{(i)}\right)$, for $s, i = 0, 1$.

The sampling probability $p\left(I_s^{(i)}\right)$ of individuals with symptoms s and infection i is defined as

$$p\left(I_s^{(i)}\right) = \frac{N_T^{s,i}}{N_s^{(i)}}, \tag{3}$$

where $N_T^{s,i}$ is the number of tested individuals from group $I_s^{(i)}$. Analogously to (3), $p\left(I_s\right)$, the sampling probability among individuals with symptoms s, is defined as

$$p\left(I_s\right) = \frac{N_T^{s,*}}{N_s}, \tag{4}$$

where $N_T^{s,*} = N_T^{s,0} + N_T^{s,1}$. And $p\left(I^{(i)}\right)$, the sampling probability among individuals with infection status i, is defined as

$$p\left(I^{(i)}\right) = \frac{N_T^{*,i}}{N^{(i)}}, \tag{5}$$

where $N_T^{*,i} = N_T^{0,i} + N_T^{1,i}$. Finally, the total of sampled individuals $\sum_{j=1}^{N} T_j$ is defined as

$$N_T = \sum_{s,i} N_T^{s,i}. \tag{6}$$

3 | ESTIMATORS

In case there is no error in testing, the naïve estimator of $p_s^{(i)}$ can be naturally defined as

$$p_T^{s,i} = f_{S^*|T}\left(I_s^{(i)} \mid T = 1\right), \tag{7}$$

the conditional probability of an individual belonging to the group $I_s^{(i)}$, given that s/he was sampled. A Bayesian approach, inspired from ideas in publication bias,[6] leads to

Proposition 1.

$$p_T^{s,i} = \frac{N_T^{s,i}}{N_T}. \tag{8}$$

Then $N_s^{(i)}$, the population size of $I_s^{(i)}$, disappears from the sample estimator, and (1) in the Appendix shows that all information in the sample about the group $I_s^{(i)}$ comes from the sampling mechanism $p\left(I_s^{(i)}\right)$. In fact, $p_s^{(i)}$ can be seen as the message sent, $\tilde{p}_s^{(i)}$ as the message received, and $p\left(I_s^{(i)}\right)/P[T=1]$ as the channel between them distorting the message.[7,8]

Analogously to (7) with Proposition 1, the naive estimator of individuals with symptoms s, $p_T^{s,*}$, and the naive estimator of individuals with infection status i, $p_T^{*,i}$, are defined as

$$p_T^{s,*} = f_{S^*|T}\left(I_s \mid T = 1\right) = \frac{N_T^{s,*}}{N_T}, \tag{9}$$

$$p_T^{*,i} = f_{S^*|T}\left(I^{(i)} \mid T = 1\right) = \frac{N_T^{*,i}}{N_T}. \tag{10}$$

Equation (1) in the Appendix also says that some information about the sampling mechanism is needed if any meaningful observation is going to be obtained. For the scenario considered in this article, this corresponds to the intuition that, since symptomatic individuals are more prone to get tested than asymptomatic individuals, the probability of sampling from the symptomatic group is larger than the probability of sampling from the asymptomatic one:

$$p\left(I_0\right) < p\left(I_1\right). \tag{11}$$

Also corresponding to the intuition that infected and non-infected individuals inside each category are randomly sampled,

$$p\left(I_s^{(0)}\right) = p\left(I_s^{(1)}\right) = p\left(I_s\right), \tag{12}$$

for $s = 0, 1$.

3.1 | Naïve estimators with testing errors

Up to this point the analysis has not considered testing errors. The following result is obtained when the errors are introduced and stratified by symptoms:

Proposition 2. Let α_0 and β_0 be the false positive and false negative rate for asymptomatic individuals, respectively; and let α_1 and β_1 be the false positive and false negative rate for symptomatic individuals, respectively. The naïve estimators thus become:

$$\begin{aligned}
\tilde{p}_T^{0,0} &= \left(1-\alpha_0\right) p_T^{0,0} + \beta_0 p_T^{0,1}, \\
\tilde{p}_T^{0,1} &= \alpha_0 p_T^{0,0} + \left(1-\beta_0\right) p_T^{0,1}, \\
\tilde{p}_T^{1,0} &= \left(1-\alpha_1\right) p_T^{1,0} + \beta_1 p_T^{1,1}, \\
\tilde{p}_T^{1,1} &= \alpha_1 p_T^{1,0} + \left(1-\beta_1\right) p_T^{1,1}.
\end{aligned} \tag{13}$$

Analogously to previous definitions, let $\tilde{p}_T^{s,*} = \tilde{p}_T^{s,0} + \tilde{p}_T^{s,1}$ and $\tilde{p}_T^{*,i} = \tilde{p}_T^{0,i} + \tilde{p}_T^{1,i}$.

Remark 1. The right-hand side of (13) contains the contribution to the naïve estimator by each group in the sample weighted by the probability of their errors. The "tilde terms" $\tilde{p}_T^{s,i}$ are observed by the practitioner, but $p_T^{s,i}$ are unknown to him.

4 | CORRECTION

This section introduces an estimator that corrects bias induced by the testing errors and oversampling of symptomatic individuals. Section 4.1 uses the maximum entropy principle, together with (11) and (12) to correct the latter. Section 4.2 proposes an estimator that eliminates the bias induced by the former. Section 4.3 puts the two together in a simple algorithm that summarizes the findings.

4.1 | Correction of sampling bias

This section ignores the presence of testing errors. The approach will be to use the maximum entropy principle, which is "the least biased estimate possible on the given information."[9] Such given information corresponds to the overrepresentation of symptomatic individuals (11) and the random sampling of infected and non-infected individuals inside each category of symptoms (12). The next theorem shows that (11) provides an upper bound to p_1 and $p_1^{(1)}$.

Theorem 1. For $p_T^{1,*}, p_1 \in (0,1)$, $p\left(I_0\right) < p\left(I_1\right)$ if and only if $p_T^{1,*} > p_1$.

Theorem 1 shows that, given the basic assumption (11), p_1 is bounded above by $p_T^{1,*}$. On the other hand, $p_T^{1,*} = N_T^{1,*}/N_T$ says that there are at least $N_T^{1,*}$ infected symptomatic individuals in the population. Therefore,

$$N_T^{1,*}/N \leq p_1 \leq p_T^{1,*}. \tag{14}$$

By the maximum entropy principle,[10] the corrected estimator of p_1 is taken to be the expectation of a uniform distribution over $\left(N_T^{1,*}/N, p_T^{1,*}\right)$. Formally, let U be a uniform distribution over the interval $\left(p_T^{1,*}\frac{N_T}{N}, p_T^{1,*}\right)$. The corrected estimator of p_1 is defined as

$$\hat{p}_1 := E(U) = \frac{p_T^{1,*}}{2}\left(\frac{N_T}{N} + 1\right). \tag{15}$$

Since, by (12), the sample is assumed to be random among symptomatic individuals,

$$\hat{p}_1^{(1)} := \hat{p}_1 \frac{N_T^{1,1}}{N_T^{1,*}} = \hat{p}_1 \frac{p_T^{1,1}}{p_T^{1,*}}. \tag{16}$$

Again, using (12), but now on the asymptomatic group, the prevalence for this group is obtained as

$$\hat{p}_0^{(1)} := \hat{p}_0 \frac{N_T^{0,1}}{N_T^{0,*}} = \left(1 - \hat{p}_1\right) \frac{N_T^{0,1}}{N_T^{0,*}}. \tag{17}$$

Taking (16) and (17), the final sampling-bias corrected prevalence is then taken to be

$$\hat{p} := \hat{p}^{(1)} = \hat{p}_1^{(1)} + \hat{p}_0^{(1)}. \tag{18}$$

Hössjer et al proved that $\hat{p}^{(1)}$ converges asymptotically to $E\left(\hat{p} \mid \hat{p}_0, \hat{p}_1\right)$:[11]

Theorem 2 (Hössjer et al, Theorem 1)**.** Suppose $N \to \infty$ in such a way that, for $s = 0, 1$, $p_s = N_s/N$ is fixed and the sampling probabilities satisfy that $p\left(I_s^{(i)}\right) = p\left(I_s\right)$. Assume also that there exists p_s' such that $\hat{p}_s$ converges in probability to p_s' for all s as $N \to \infty$. Then

$$\sqrt{N}\left[\hat{p} - E\left(\hat{p} \mid \hat{p}_0, \hat{p}_1\right)\right] \to_{\mathcal{L}} N\left(0, V\right), \tag{19}$$

where "$\to_{\mathcal{L}}$" implies convergence in distribution, and

$$V = \sum_{s=0}^{1} \frac{\left(p_s'\right)^2}{p_s} \frac{1 - p\left(I_s\right)}{p\left(I_s\right)} \frac{p_s^{(1)}}{p_s}\left(1 - \frac{p_s^{(1)}}{p_s}\right).$$

Remark 2. Theorem 2 does not say that $\hat{p} \to p^{(1)}$. It says that the corrected estimator behaves as well as $\hat{p}_1 = E(U)$ estimates N_1/N. This corresponds well to the maximum entropy assumption: the more useful information is at hand, the highest the reduction of entropy, and therefore the better the correction.

4.2 | Error-free estimator

According to Remark 1, when testing errors are considered, estimators that correct them are necessary before applying the correction to sampling bias. This section presents such estimators.

Theorem 3. For $s = 0, 1$, assume α_s and β_s are known, and let $\beta_s \leq 1 - \alpha_s$. The estimators

$$\bar{p}_T^{s,0} = \frac{\tilde{p}_s^{(0)} - \beta_s \tilde{p}_s}{1 - \alpha_s - \beta_s}, \tag{20}$$

$$\bar{p}_T^{s,1} = \frac{\alpha_s \tilde{p}_s - \tilde{p}_s^{(1)}}{\alpha_s + \beta_s - 1}, \tag{21}$$

of $p_s^{(0)}$ and $p_s^{(1)}$, respectively, are unbiased for errors, where $\tilde{p}_0 = \tilde{p}_0^{(0)} + \tilde{p}_0^{(1)}$ and $\tilde{p}_1 = \tilde{p}_1^{(0)} + \tilde{p}_1^{(1)}$. Thus $\bar{p}_T^{s,i}$ is an estimator of $p_T^{s,i}$.

4.3 | Algorithm

The correction procedure can be motivated as follows. The researcher observes the total of positive and negative individuals among the symptomatic and asymptomatic groups after testing, and s/he does not know neither the real prevalence nor the sampling scheme, except for the fact that the symptomatic group is oversampled (11). Thus the process to correct the estimator runs backwards: starting with the naïve estimator (that includes bias from errors and sampling), the sampling proportions are recovered (getting rid of the errors), to finally produce a corrected estimator (reducing the sampling bias). Algorithm 1, for which code is available at https://github.com/kalilizhou/BiasCorrection.git, summarizes the procedure to obtain a corrected estimator of prevalence.

Algorithm 1 Corrected estimator of prevalence

1. For $s = 0, 1$, make
$$\bar{p}_T^{s,0} = \frac{\tilde{p}_s^{(0)} - \beta_s \tilde{p}_s}{1 - \alpha_s - \beta_s},$$
$$\bar{p}_T^{s,1} = \frac{\alpha_s \tilde{p}_s - \tilde{p}_s^{(1)}}{\alpha_s + \beta_s - 1}.$$
2. For $\bar{p}_T^{1,*} = \bar{p}_T^{1,0} + \bar{p}_T^{1,1}$, take
$$\hat{p}_1 = \frac{\bar{p}_T^{1,*}}{2}\left(\frac{N_T}{N} + 1\right). \tag{22}$$
3. Make
$$\hat{p}_1^{(1)} = \hat{p}_1 \frac{N_T^{1,1}}{N_T^{1,*}}.$$
4. Take $\hat{p}_0^{(1)} = \frac{\bar{p}_T^{0,1}}{\bar{p}_T^{0,*}}\left(1 - \hat{p}_1\right)$, where $\bar{p}_T^{0,*} = \bar{p}_T^{0,0} + \bar{p}_T^{0,1}$.
5. The estimated total prevalence is: $p = \hat{p}_0^{(1)} + \hat{p}_1^{(1)}$.

Remark 3. If stratification is ignored, just take $\alpha = \alpha_0 = \alpha_1$ and $\beta = \beta_0 = \beta_1$ in Step 1 of Algorithm 1.

Remark 4. If error in testing is not of interest, and only sampling bias is being considered, Algorithm 1 can still be used, starting from step 2.

Remark 5. If only testing errors are under consideration, and sampling bias is ignored, Step 1 of Algorithm 1 provides a correction.

Remark 6. As specified in Remark 2, under the assumption of maximum entropy a natural way to increase the precision of the estimator is to add knowledge whenever it is available. The scenario considered by Díaz-Pachón and Rao in which all the symptomatic individuals are tested (as it is required in most universities and companies in the U.S.) is one example.[5] In this case, the only modification of Algorithm 1 is that $\hat{p}_1$ in (22) becomes

$$\hat{p}_1 = \tilde{p}_T^{1,*} \frac{N_T}{N}. \tag{23}$$

No other changes are required. Notice however that, under this modification, Algorithm 1 does a better job than the procedure considered by Díaz-Pachón and Rao,[5] since Algorithm 1 neither assumes the absence of symptomatic individuals without the disease nor ignores the evidence from the sample to estimate prevalence between the classes of symptoms, as becomes clear from steps 3 and 4. This superiority, as well as comparisons with other mechanisms, will be analyzed in Section 6.

5 | SIMULATION

This section uses simulation to analyze the asymptotic behavior of the corrected estimator. The population has the following features:

- The proportion of positive cases with symptoms $p_1^{(1)}$ is 0.15,
- the proportion of negative cases with symptoms $p_1^{(0)}$ is 0.05,
- the proportion of positive cases without symptoms $p_0^{(1)}$ is 0.05,
- and proportion of negative cases without symptoms $p_0^{(0)}$ is 0.75.

Thus, the prevalence is $p_0^{(1)} + p_1^{(1)} = 0.2$, the proportion of symptomatic individuals in the population is $p_1^{(1)} + p_1^{(0)} = 0.2$, so the proportion of asymptomatic in the population is $p_0^{(1)} + p_0^{(0)} = 0.8$. The asymptomatic false positive rate is taken to be $\alpha_0 = 0.01$, the symptomatic false positive rate is $\alpha_1 = 0.05$, the asymptomatic false negative rate is $\beta_0 = 0.1$, and the symptomatic false negative rate is $\beta_1 = 0.05$.

The proportion of the asymptomatic patients being tested is $p(I_0)$ is 0.1, while the proportion of the symptomatic patients being tested is $p(I_1)$ is 0.9. (Notice that, in spite of the selection of the sampling probabilities for this simulation, there is no requirement that $p(I_0) + p(I_1) = 1$.)

The naïve estimator from the sample, $\tilde{p}_T^{*,1}$, and the corrected estimator, $\hat{p}$, are listed in Table 1 for increasing population values, which suggests that the estimated values are gradually converging; i.e., $\tilde{p}_T^{*,1} \to 0.512$ and $\hat{p} \to 0.362$.

Population	1000	10000	100000	1000000
$\tilde{p}_T^{*,1}$	0.513 (0.023)	0.512 (0.007)	0.512 (0.002)	0.512 (0.0001)
$\hat{p}$	0.362 (0.018)	0.363 (0.006)	0.362 (0.002)	0.362 (0.001)

TABLE 1 Estimated sample prevalence and population prevalence

5.1 | Active information: the index

Active information (actinfo) was introduced in search problems to quantify the amount of Shannon information introduced by the programmer in a search problem.[12,13,14] In machine learning, it has been used to show that no algorithm performs well for a large class of problems, in agreement with the so-called No Free Lunch Theorems.[15,16,17] It has also been used for mode hunting,[18,19] and to compare neutral to non-neutral models in population genetics.[20] Following the recommendation of Hössjer et al, here active information is used as a measure of bias.[11] The idea is as follows: active information is defined as

$$I^+ = \log\left(\hat{p}/p^{(1)}\right), \tag{24}$$

where the logarithm is taken to be in base e, so that information is measured in nats. Thus defined, active information measures the amount of Shannon information of the estimator $\hat{p}$ to the true proportion $p^{(1)}$, and it is the quantity that is averaged in the Kullback-Leibler divergence[21]. That is, if the true proportion is overestimated, the active information will be positive and large; if the true proportion is underestimated, the active information will be negative; and if the true proportion is accurately estimated, the active information will be around zero.[22,23]

Moeover, active information can be decomposed into two parts, $I^+ = I_T^+ + I_C^+$, where $I_T^+ = \log\left(\tilde{p}_T^{*,1}/p^{(1)}\right)$ measures the difference in information from the biased estimate to the real prevalence, and $I_C^+ = \log\left(\hat{p}/p_T^{*,1}\right)$ measures the difference in information from the correction to the naïve estimator.[11] Their empirical versions are listed in Table 2.

The active information for the biased estimator is $I_T^+ \approx 0.94$. The active information for the correction $I_C^+ \approx -0.34$ reduces the bias, producing $I^+ \approx 0.59$.[22]

Population	1000	10000	100000	1000000
$\hat{I}_T^+$	0.942	0.940	0.939	0.939
$\hat{I}_C^+$	-0.347	-0.345	-0.345	-0.345
$\hat{I}^+$	0.595	0.595	0.594	0.594

TABLE 2 Average active information of 1000 simulations

6 | DATA FROM THE ISRAELI MINISTRY OF HEALTH

In what follows, Covid-19 data from the Israeli Ministry of Health is considered.[24] The Ministry of Health publicly released data for individuals tested for Covid-19 via a PCR assay from a nasal swab sample collected between March 22, 2020 and April 7, 2020. The dataset contains information on the test date, test result, clinical symptoms, gender of the individual, known contact with an infected individual and a binary indicator of whether the individual was 60 years of age or older. Symptoms include cough, fever, sore throat, shortness of breath and headache. For the purposes of illustrating the methodology, we will consider this the population consisting of 99 232 tested individuals of whom 1862 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 8393 infections through PCR testing. Among the individuals who tested positive, 1754 were symptomatic. The characteristics of the data set are presented in Table 3.

TABLE 3 Observed disease status by category of symptoms.

	Positive	Negative	Total
Symptomatic	1754	108	1862
Asymptomatic	6639	90 731	97 370
Total	8393	90 839	99 232

Error rates will be stratified by symptoms. Thus, let α_0 and α_1 be the false positive rate for asymptomatic and symptomatic individuals, respectively, and β_0 and β_1, the false negative rate for asymptomatic and symptomatic individuals, respectively. For the purpose of this example, $\alpha_0 = 0.1\%, \alpha_1 = 0.5\%, \beta_0 = 10\%$, and $\beta_1 = 5\%$. The actual number of individuals inside each group can be found in Table 4, after correcting for these errors.

The real prevalence is then

$$p = (1\,751 + 15\,705)/99\,232 = 0.176, \tag{25}$$

and prevalence among the asymptomatic is $15\,705/97\,370 = 0.161$.

TABLE 4 Real proportions under stratified errors with $\alpha_0 = 0.1\%, \alpha_1 = 0.5\%, \beta_0 = 10\%$, and $\beta_1 = 5\%$.

	Disease	Non-Disease	Total
Symptomatic	1751	111	1862
Asymptomatic	15 705	81 665	97 370
Total	53 754	81 776	99 232

Finally, active information (24) is used to compare how well Algorithm 1 and other estimators proposed in the literature are doing with respect to the real prevalence. The best estimator will be the one with active information I^+ closer to 0. The competitors will be the method proposed by Díaz-Pachón and Rao, which assumes all symptomatic individuals are sampled, correcting only for sample bias and ignoring testing errors;[5] Diggle's Bayesian approach, which corrects for imperfect testing but ignores sampling bias;[25] and the Rogan-Gladen estimate, a frequentist method that only corrects for testing errors too.[26] Neither of the competitors corrects for sampling bias and testing errors at the same time. As much as we search, we could not find a methodology that simultaneously corrects for imperfect testing and sampling bias; this will be reflected in the analysis.

Sampling Protocol 1: In the first scenario, all symptomatic individuals are sampled, as considered by Díaz-Pachón and Rao.[5] 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 3. As for the asymptomatic group, the sampling proportions are taken according to Table 4. The observations of this setting are summarized in Table 5.

TABLE 5 Stratified sample with 75% symptomatic and 25% asymptomatic (sample all symptomatic individuals).

	Positive	Negative	Total
Symptomatic	1754	108	1862
Asymptomatic	100	521	621
Total	1854	629	2483

According to Table 5, the naïve estimator is $\tilde{p}^{(1)} = 1854/2483 \approx 0.75$. Using Algorithm 1 with the modification (23), the corrected estimator is $\hat{p} = 0.185$. Table 6 presents these results as well as those of the other methods.

TABLE 6 Results of Sampling Protocol 1.

	Naïve	Díaz-Rao	Diggle	Rogan-Gladen	Algorithm 1
$\hat{p}$	0.750	0.177	–	0.750	0.185
I^+	1.45	0.005	–	1.45	0.05

In this case, Diggle's correction was not implemented because it involves combinations in its logarithm that are difficult to approximate when the sample is moderately large. Under the assumption of sampling all symptomatic individuals, Díaz-Rao algorithm works better than all others, and RGE performs as poorly as the naïve estimator. However, Algorithm 1 also corrects very well the naïve estimate, producing small active information. Both Díaz-Rao and Algorithm 1 are close to the real prevalence, and there is not statistical difference between them for this scenario.[22]

For the next protocols, the assumption that all symptomatic individuals were sampled is removed, which implies that the Diaz-Rao correction cannot be assessed and Algorithm 1 is followed without modifications.

Sampling Protocol 2: The sample consists of 200 individuals. Among these, 150 (75%) are symptomatic and 50 (25%) are asymptomatic. For both the symptomatic and asymptomatic groups, the sampling proportions are taken according to Table 4. With this information, we can observe Table 7. The summary of results under different methods is shown in Table 8.

Table 8 shows that Algorithm 1 has the best performance, without being optimal. In fact, Diggle's and Rogan-Gladen's estimates do as poorly as the naïve estimate. Algorithm 1 beats its competitors because it is the only one that corrects for sampling

TABLE 7 Stratified sample with 75% symptomatic and 25% asymptomatic (not all symptomatic individuals sampled).

	Positive	Negative	Total
Symptomatic	141	9	150
Asymptomatic	8	42	50
Total	149	51	200

TABLE 8 Results of Sampling Protocol 2.

	Algorithm 1	Naïve	Diggle	Rogan-Gladen
$\hat{p}$	0.464	0.745	0.805	0.750
I^+	0.97	1.44	1.52	1.45

bias, whereas the other two only correct for testing errors. Notice that the additional information of Protocol 1 (knowing that all symptomatic individuals were sampled), in comparison to Protocol 2, greatly improves the performance of Algorithm 1, as reflected by the active information.

Sampling Protocol 3: In this scenario there are 100 symptomatic and 100 asymptomatic individuals. Again Table 4 reflects the proportions inside each group for this protocol. Table 9 is obtained. With Table 9 as base, the summary of results under different methods for this sampling protocol is presented in Table 10.

TABLE 9 Stratified observed totals with 50% symptomatic and 50% asymptomatic.

	Positive	Negative	Total
Symptomatic	94	6	100
Asymptomatic	16	84	100
Total	110	90	200

TABLE 10 Results of Sampling Protocol 3.

	Algorithm 1	Naïve	Diggle	Rogan-Gladen
$\hat{p}$	0.37	0.55	0.59	0.50
I^+	0.74	1.14	1.21	1.04

Thus, compared to the Sampling Protocol 2, with less sampling bias, all the methods perform better. Rogan-Glade's estimates performs better than Diggle's, reducing the naïve bias. Algorithm 1 still works better than competitors, correcting a half of the naïve overestimation.

Sampling Protocol 4: This sample is truly random, with $N_T = 200$. Table 4 is used to determine the proportions of all groups. With these values Table 11 is obtained. The results of the different methods for this scenario are presented in Table 12.

TABLE 11 Stratified observed totals from a random sample of size 200.

	Positive	Negative	Total
Symptomatic	4	1	5
Asymptomatic	32	163	195
Total	36	164	200

TABLE 12 Results of Sampling Protocol 4.

	Algorithm 1	Naïve	Diggle	Rogan-Gladen
$\hat{p}$	0.19	0.18	0.193	0.025
I^+	0.07	0.02	0.09	-1.952

Of course, in this scenario the naïve estimate is optimal. Rogan-Gladen's frequentist estimate grossly overcorrects to the point of removing almost 2 nats of information with respect to the real prevalence. On the other hand, Diggle's Bayesian approach and Algorithm 1 work pretty well and no statistical difference is observed between them and the naïve estimate.

7 | DISCUSSION

Timely and accurate prevalence estimation of a disease is one of the most fundamental concepts in epidemiology and its importance is because it provides a measure of disease burden in a population at a particular point in time. It can also be part of a compendium of measures used to inform public health prevention policies to help slow the spread of disease through the population. To provide prevalence estimates that are reliable and generalizable, the sample must be comprehensive enough to capture all relevant subpopulations in the general population and as mentioned, for a number of diseases this can be challenging because many of these sub-populations can be hard-to-reach. Thus, sampling bias corrections are needed. Interestingly, this paper has presented new methodology where biased samples result due to over-sampling of symptomatic individuals. In addition, Algorithm 1 goes further and presents a correction both for sampling bias and testing errors. However, the methodology generalizes easily regardless of how the biased samples resulted.

A limitation of our study is that error rates for tests are assumed to be 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.[25] 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 the setting in this paper, in which sampling bias is also an issue.

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.[27] This is sometimes called group testing. However, this implicitly assumes random sampling of pools which is clearly not the case considered here.

Another approach is to use population seroprevalence complex surveys.[28,29] 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 has been studied.[11]

For some diseases it is becoming more common to use administrative data to estimate disease prevalence since for many countries these data cover large proportions of the population. Examples include Canada, Denmark and Italy among others. This requires some effort to properly assemble these data sources,[30] but they have to date not proven as useful for emerging diseases like Covid-19 where surveillance studies dominated the earlier days of the pandemic.

Author contributions

D. A. D. P. and J. S. R. conceptualized the methodology framework and the paper. L. Z. and D. A. D. P. developed the methodology details, L. Z. ran the examples and produced the R code, and C. Z. ran the simulations.

Financial disclosure

None reported.

Conflict of interest

The authors declare no potential conflict of interests.

Data availability statement

Code to implement Algorithm 1 is available at https://github.com/kalilizhou/BiasCorrection.git. The data used in Section 6 is publicly available at https://github.com/nshomron/covidpred.

How to cite this article: Zhou L., Díaz-Pachón D. A., Zhao C., and Rao J. S. (2022), Correcting prevalence estimation for biased sampling with testing errors, , *2022;00:1–13*.

APPENDIX

Proof of Proposition 1.

$$\begin{aligned}
\mathbf{p}_T^{s,i} &= f_{S^*|T}\left(I_s^{(i)} \mid T=1\right) \\
&= \frac{P\left[T=1 \mid S^*=I_s^{(i)}\right]}{P[T=1]} f_{S^*}\left(I_s^{(i)}\right) \\
&= \frac{\mathbf{p}\left(I_s^{(i)}\right)}{P[T=1]} f_{S^*}\left(I_s^{(i)}\right) \\
&= \frac{N_T^{s,i}/N_s^{(i)}}{N_T/N} \frac{N_s^{(i)}}{N} \\
&= \frac{N_T^{s,i}}{N_T} \\
&= p_T^{s,i},
\end{aligned} \tag{1}$$

where the approximation step uses (1), (3), and (6). □

Proof of Proposition 2. The result follows from Proposition 1 once testing errors are taken into account. □

Proof of Theorem 1.

$$\begin{aligned}
p(I_0) < p(I_1) &\Leftrightarrow \frac{N_T^{0,*}}{N_0} < \frac{N_T^{1,*}}{N_1} \\
&\Leftrightarrow \frac{N_1/N}{N_0/N} < \frac{N_T^{1,*}/N_T}{N_T^{0,*}/N_T} \\
&\Leftrightarrow \frac{p_1}{p_0} < \frac{p_T^{1,*}}{p_T^{0,*}} \\
&\Leftrightarrow \frac{1-p_1}{p_1} > \frac{1-p_T^{1,*}}{p_T^{1,*}} \\
&\Leftrightarrow \frac{1}{p_1} > \frac{1}{p_T^{1,*}} \\
&\Leftrightarrow p_T^{1,*} > p_1,
\end{aligned}$$

where the fourth step used that $p_0 + p_1 = 1 = p_T^{0,*} + p_T^{1,*}$. □

Proof of Theorem 2. The result follows from Theorem 1 in Hössjer et al.[11]. □

Proof of Theorem 3. Since the left-hand side of (13) is obtained from the sample, and the errors are known, the first two equations of (13) have two unknowns: $p_T^{0,0}$ and $p_T^{0,1}$. Analogously, the last two equations of (13) have two unknowns: $p_T^{1,0}$ and

$p_T^{1,1}$. Now, for $s = 0, 1$,

$$\begin{aligned} E\left(\tilde{p}_T^{s,0}\right) &= \frac{E\left(\tilde{p}_s^{(0)}\right) - \beta_s E\left(\tilde{p}_s\right)}{1-\alpha_s-\beta_s} \\ &= \frac{(1-\alpha_s)E\left(p_T^{s,0}\right) + \beta_s E\left(p_T^{s,1}\right) - \beta_s E\left(p_T^{s,0}+p_T^{s,1}\right)}{1-\alpha_s-\beta_s} \\ &= E\left(p_T^{s,0}\right), \end{aligned}$$

and

$$\begin{aligned} E\left(\tilde{p}_T^{s,1}\right) &= \frac{\alpha_s E\left(\tilde{p}_s\right) - E\left(\tilde{p}_s^{(1)}\right)}{\alpha_s+\beta_s-1} \\ &= \frac{\alpha_s E\left(p_T^{s,0}+p_T^{s,1}\right) - \alpha_s E\left(p_T^{s,0}\right) - (1-\beta_s)E\left(p_T^{s,1}\right)}{\alpha_s+\beta_s-1} \\ &= E\left(p_T^{s,1}\right). \end{aligned}$$

□

References

1. Tan S, Makela S, Heller D, et al. A Bayesian evidence synthesis approach to estimate disease prevalence in hard-to-reach populations: hepatitis C in New York City.. *Epidemics* 2018; Jun(23): 96-109. doi: 10.1016/j.epidem.2018.01.002

2. Hellewell J, al e. Feasibility of controlling COVID-19 outbreaks by isolation of cases and contacts. *Lancet Global Health* 2020; 8(4): e488–e496. doi: 10.1016/S2214-109X(20)30074-7

3. Mancastroppa M, Castellano C, Vezzani A, Burioni R. Stochastic sampling effects favor manual over digital contact tracing. *Nature Communications* 2021; 12(1919). doi: 10.1038/s41467-021-22082-7

4. Bengio Y, Janda R, Yu YW, et al. The need for privacy with public digital contact tracing during the COVID-19 pandemic. *The Lancet Digital Health* 2020; 2(7): e342-e344. doi: 10.1016/S2589-7500(20)30133-3

5. Díaz-Pachón DA, Rao JS. A simple correction for COVID-19 sampling bias. *Journal of Theoretical Biology* 2021; 512: 110556. doi: 10.1016/j.jtbi.2020.110556

6. Andrews I, Kasy M. Identification of and Correction for Publication Bias. *American Economic Review* 2019; 109(8): 2766-2794. doi: 10.1257/aer.20180310

7. Barbier J. Inferenza ad alta dimensionalità: una prospettiva di meccanica statistica. *Ithaca: Viaggio nella Scienza* 2020; XVI(99-137).

8. Hössjer O, Díaz-Pachón DA, Rao JS. Active Information, Learning, and Knowledge Acquisition. *PsyArXiv* 2022. doi: 10.31234/osf.io/qt5kw

9. Jaynes ET. Information Theory and Statistical Mechanics. *Physical Review* 1957; 106(4): 620-630. doi: 10.1103/PhysRev.106.620

10. Díaz-Pachón DA, Marks II RJ. Generalized active information: Extensions to unbounded domains. *BIO-Complexity* 2020; 2020(3): 1-6. doi: 10.5048/BIO-C.2020.3

11. Hössjer O, Díaz-Pachón DA, Chen Z, Rao JS. Active information, missing data, and prevalence estimation. *arXiv* 2022. doi: 10.48550/arXiv.2206.05120

12. Dembski WA, Marks II RJ. Bernoulli's Principle of Insufficient Reason and Conservation of Information in Computer Search. *Proceedings of the 2009 IEEE International Conference on Systems, Man, and Cybernetics. San Antonio, TX* 2009: 2647-2652. doi: 10.1109/ICSMC.2009.5346119

13. Dembski WA, Marks II RJ. Conservation of Information in Search: Measuring the Cost of Success. *IEEE Transactions Systems, Man, and Cybernetics - Part A: Systems and Humans* 2009; 5(5): 1051-1061. doi: 10.1109/TSMCA.2009.2025027

14. Dembski WA, Marks II RJ. The Search for a Search: Measuring the Information Cost of Higher Level Search. *Journal of Advanced Computational Intelligence and Intelligent Informatics* 2010; 14(5): 475-486. doi: 10.20965/jaciii.2010.p0475

15. Montañez GD. The famine of forte: Few search problems greatly favor your algorithm. *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* 2017: 477-482. doi: 10.1109/SMC.2017.8122651

16. Montañez GD. A Unified Model of Complex Specified Information. *BIO-Complexity* 2018; 2018(4): 1-26.

17. Wolpert DH, MacReady WG. No Free Lunch Theorems for Optimization. *IEEE Transactions on Evolutionary Computation* 1997; 1(1): 67-82. doi: 10.1109/4235.585893

18. Díaz-Pachón DA, Sáenz JP, Rao JS, Dazard JE. Mode hunting through active information. *Applied Stochastic Models in Business and Industry* 2019; 35(2): 376-393. doi: 10.1002/asmb.2430

19. Liu T, Díaz-Pachón DA, Rao JS, Dazard JE. High Dimensional Mode Hunting Using Pettiest Component Analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 2022; Accepted. doi: 10.1109/TPAMI.2022.3195462

20. Díaz-Pachón DA, Marks II RJ. Active Information Requirements for Fixation on the Wright-Fisher Model of Population Genetics. *BIO-Complexity* 2020; 2020(4): 1-6. doi: 10.5048/BIO-C.2020.4

21. Cover TM, Thomas JA. *Elements of Information Theory*. Wiley. second ed. 2006.

22. Díaz-Pachón DA, Sáenz JP, Rao JS. Hypothesis testing with active information. *Statistics & Probability Letters* 2020; 161: 108742. doi: 10.1016/j.spl.2020.108742

23. Díaz-Pachón DA, Hössjer O. Assessing and Testing Fine-Tuning by Means of Active Information. *Submitted* 2022.

24. Zoabi Y, Deri-Rozov S, Shomron N. Machine learning-based prediction of COVID-19 diagnosis based on symptoms. *npj Digital Medicine* 2021; 4(3). doi: 10.1038/s41746-020-00372-6

25. Diggle PJ. Estimating prevalence using an imperfect test. *Epidemiology Research International* 2011: 608719. doi: 10.1155/2011/608719

26. Rogan WJ, Gladen B. Estimating prevalence from the results of a screening test. *American Journal of Epidemiology* 1978; 107(1): 71-76. doi: 10.1093/oxfordjournals.aje.a112510

27. Brynildsrud O. COVID-19 prevalence estimation by random sampling in population - optimal sample pooling under varying assumptions about true prevalence. *BMC Medical Research Methodology* 2020; 20: 196. doi: 10.1186/s12874-020-01081-0

28. Carabaña JM. Datos de encuesta para estimar la prevalencia de COVID-19. Un estudio piloto en Madrid capital. *Revista española de salud pública* 2020; 94(17 de noviembre): e202011159.

29. Franceschi VB, Santos AS, al. eABG. Population-based prevalence surveys during the Covid-19 pandemic: A systematic review. *Reviews in Medical Virology* 2021; 31(4): e2200. doi: rmv.2200

30. Ward MM. Estimating Disease Prevalence and Incidence Using Administrative Data: Some Assembly Required. *Journal of Rheumatology* 2013; 40(8): 1241-1243. doi: 10.3899/jrheum.130675