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[1] | Daniel Andrés Díaz-Pachón*[1] | Chen Zhao[1] | J. Sunil Rao[1] | Ola Hössjer[2]

[1]Division of Biostatistics, University of Miami, Florida, United States of America
[2]Department of Mathematics, Stockholm University, Stockholm, Sweden

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 of prevalence (i.e., proportion of observed infected individuals in the sample) that can be very far from the true proportion of infected. In this work, we present a method of prevalence estimation that reduces both the effect of bias due to testing errors and oversampling of symptomatic individuals, eliminating it altogether in some scenarios. Moreover, this procedure considers stratified errors in which tests have different error rate profiles for symptomatic and asymptomatic individuals. This results in easily implementable algorithms, for which code is provided, that produce better prevalence estimates than other methods (in terms of reducing and/or removing bias), as demonstrated by formal results, simulations, 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, except for the hypothetical case of random errors, imperfect testing almost always distorts actual proportions. Second, it is 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.[2]

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.[3,4,5] Though this example corresponds to a non-probability COVID-19 sampling setting, the problem is of course more general. It applies not only to every form of prevalence estimation performed through testing —probabilistic or not— and even more general forms of selection bias.[6,7,8,9]

Recently, Díaz-Pachón and Rao introduced a correction for oversampling of the symptomatic group.[10] 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.

The article can be summarized as follows:

1. The researcher observes, from a sample of size N_T, the proportion of individuals whose test was positive. This constitutes the naïve estimator of prevalence $\tilde{\mathbf{p}}_T^{*,1}$.

2. The naïve estimator $\tilde{\mathbf{p}}_T^{*,1}$ is biased in two ways. First, it is subject to testing errors; and second, there is sample bias because symptomatic individuals are more likely to be tested than asymptomatic ones.

3. Under the assumption that, from a different and independent sample, testing error rates are estimated for group s as $\hat{\alpha}_s$ and $\hat{\beta}_s$ for false positives and false negatives, respectively, it is possible to obtain another estimator $\bar{p}_T^{*,1}$ that corrects the effect of errors.

4. From $\bar{p}_T^{*,1}$ it is possible to obtain another estimator $\hat{p}^{(1)}$ that reduces (and possibly removes) the sampling bias. This is achieved through applying the principle of maximum entropy to the fraction of symptomatic individuals with the disease, using the knowledge that symptomatic individuals are more likely to get tested.

With this summary and the information of Table 1, without having to go through the details that led to their derivation, the reader can obtain corrections following the steps of Algorithm 1 when no testing errors are present and all symptomatic individuals are sampled, Algorithm 2 when no testing errors are present and not all the symptomatic individuals are sampled, Algorithm 3 when testing errors are present and all symptomatic individuals are sampled (provided that the testing errors are estimated unbiasedly), and Algorithm 4 when there are testing errors and not all symptomatic individuals are sampled. Section 5 presents an example with real data. Proofs of all the results are consigned to the Appendix, as well as a set of simulations that assess the behavior of the four algorithms.

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 $i = 0, 1$. The proportion to be estimated is $p^{(1)} = p_0^{(1)} + p_1^{(1)}$, corresponding to the infected individuals, and the naïve estimator is biased because the proportion of symptomatic individuals p_1 is overestimated.

Quantity	Symptoms s and infection i $I_s^{(i)}$	Symptoms s I_s	Infection i $I^{(i)}$
population totals	$N_s^{(i)}$	N_s	$N^{(i)}$
population proportion	$p_s^{(i)}$	p_s	$p^{(i)}$
sampling probability	$q\left(I_s^{(i)}\right)$	$q\left(I_s\right)$	$q\left(I^{(i)}\right)$
sampling probability approximation	$q^*\left(I_s^{(i)}\right)$	$q^*\left(I_s\right)$	$q^*\left(I^{(i)}\right)$
sampling totals	$N_T^{s,i}$	$N_T^{s,*}$	$N_T^{*,i}$
naïve estimator (with errors and sampling bias)	$\tilde{\mathbf{p}}_T^{s,i}$	$\tilde{\mathbf{p}}_T^{s,*}$	$\tilde{\mathbf{p}}_T^{*,i}$
naïve estimator (only with sampling bias)	$\hat{\mathbf{p}}_T^{s,i}$	$\hat{\mathbf{p}}_T^{s,*}$	$\hat{\mathbf{p}}_T^{*,i}$
corrected estimator for errors	$\bar{p}_T^{s,i}$	$\bar{p}_T^{s,*}$	$\bar{p}_T^{*,i}$
corrected estimator for errors and sampling bias	$\hat{p}_s^{(i)}$	$\hat{p}_s$	$\hat{p}^{(i)}$

TABLE 1 A population of known size N is divided into symptomatic individuals ($s = 1$) and asymptomatic ones ($s = 0$), and non-infected individuals ($i = 0$) and infected ones ($i = 1$). The second column gives the notation for symptoms s and infection i, while the third column marginalizes symptoms, and the fourth one marginalizes infection. To facilitate reading, the notation is arranged as follows: **(a)** the letter q is only used for sampling probabilities; **(b)** all estimators are capped by tildes, bars, or hats; **(c)** bold caps refer to naïve estimators; **(d)** estimators not in bold are corrections of naïve estimators; **(e)** population proportions do not have any cap; **(f)** prevalence values are obtained replacing i in the last column by 1.

2.1 | Sampling probabilities

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

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

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

The sampling probability $q\left(I_s^{(i)}\right)$ of individuals with symptoms s and infection i is approximated by

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

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

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

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

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

where $N_T^{*,i} = N_T^{0,i} + N_T^{1,i}$. Notice that, except when all symptomatic individuals are sampled (in whose case N_1 is known), the approximations $q^*(\cdot)$ are not estimators of the sampling probabilities because the population values in their denominators are in general unknown. However, when $N \to \infty$,

$$q^*\left(I_s^{(i)}\right) \to q\left(I_s^{(i)}\right). \tag{6}$$

The total of sampled individuals $\sum_{j=1}^{N} T_j$ is defined as

$$N_T = \sum_{s,i} N_T^{s,i}. \tag{7}$$

Finally, we define the expected fraction of sampled individuals:

$$q = p_0 q(I_0) + p_1 q(I_1). \tag{8}$$

3 | NO TESTING ERRORS

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

$$\hat{\mathbf{p}}_T^{s,i} = f_{S^*|T}\left(I_s^{(i)} \mid T = 1\right), \tag{9}$$

the conditional probability of an individual belonging to the group $I_s^{(i)}$, given that they were sampled. The main goal of this paper is to provide a correction for $\hat{\mathbf{p}}_T^{s,i}$, under the assumption that symptomatic individuals are oversampled. A Bayesian approach, inspired from ideas in publication bias,[6] leads to

Proposition 1.

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

Then $N_s^{(i)}$, the population size of $I_s^{(i)}$, disappears from the sample estimator, and (A1) in the Appendix shows that all information in the sample about the group $I_s^{(i)}$ comes from the sampling mechanism $q\left(I_s^{(i)}\right)$. In fact, $p_s^{(i)}$ can be seen as the message sent, $\hat{\mathbf{p}}_T^{s,i}$ as the message received, and $q\left(I_s^{(i)}\right) = P(T = 1|I_s^{(i)})$ as the channel between them distorting the message.[11,12] This interpretation, taken from Shannon's information diagram, is particularly important to analyze bias as a modification of the information inherent to the prevalence parameter in Section D of the Appendix.[13]

Analogously to (9) with Proposition 1, the naive estimator of individuals with symptoms s, $\hat{\mathbf{p}}_T^{s,*}$, and the naive estimator of individuals with infection status i, $\hat{\mathbf{p}}_T^{*,i}$, are defined as

$$\hat{\mathbf{p}}_T^{s,*} = f_{S^*|T}\left(I_s \mid T = 1\right) = \frac{N_T^{s,*}}{N_T}, \tag{11}$$

$$\hat{\mathbf{p}}_T^{*,i} = f_{S^*|T}\left(I^{(i)} \mid T = 1\right) = \frac{N_T^{*,i}}{N_T}. \tag{12}$$

3.1 | Correction of sampling bias

According to (A1) in the Appendix, some information about the sampling mechanism is needed if any meaningful conclusion is going to be obtained. For the scenario considered in this article, this corresponds to symptomatic individuals being more prone to get tested than asymptomatic ones:

$$q\left(I_0\right) < q\left(I_1\right). \tag{13}$$

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

$$q\left(I_s^{(0)}\right) = q\left(I_s^{(1)}\right), \tag{14}$$

for $s = 0, 1$.

We consider two scenarios. First, when all symptomatic individuals are sampled. Second, building on the previous case, when not all symptomatic are tested, but they are overrepresented in the sample.

3.1.1 | All the symptomatic group is sampled

Díaz-Pachón and Rao studied the situation in which, for COVID-19, all symptomatic individuals are tested.[10] This scenario corresponds well to some subpopulations like those of universities or industries, in which all symptomatic individuals are required to get tested. In this case, (13) becomes

$$q\left(I_0\right) < q\left(I_1\right) = 1,$$

so the proportion p_1 of symptomatic in the population can be fully recovered from the sample as

$$p_1 = \hat{\mathbf{p}}_T^{1,*} \frac{N_T}{N}. \tag{15}$$

Since, by (14) the sample is assumed to be random among symptomatic individuals,

$$\hat{p}_1^{(1)} := p_1 \frac{N_T^{1,1}}{N_T^{1,*}} = p_1 \frac{\hat{\mathbf{p}}_T^{1,1}}{\hat{\mathbf{p}}_T^{1,*}}. \tag{16}$$

Using now (14) on the asymptomatic group, the prevalence among the asymptomatic is obtained as

$$\hat{p}_0^{(1)} := p_0 \frac{N_T^{0,1}}{N_T^{0,*}} = \left(1 - p_1\right) \frac{\hat{\mathbf{p}}_T^{0,1}}{\hat{\mathbf{p}}_T^{0,*}}. \tag{17}$$

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

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

and the random sampling inside each group gives that $E\hat{p}^{(1)} = p^{(1)}$, making the estimator unbiased.

Algorithm 1 summarizes the steps from observations to corrected estimate when there is no testing error and all symptomatic individuals are tested.

Algorithm 1 Corrected estimator of prevalence without errors and all symptomatic individuals sampled

1. For $\hat{\mathbf{p}}_T^{1,*} = \hat{\mathbf{p}}_T^{1,0} + \hat{\mathbf{p}}_T^{1,1}$, take
$$p_1 = \hat{\mathbf{p}}_T^{1,*} \frac{N_T}{N}.$$
2. Make
$$\hat{p}_1^{(1)} = p_1 \frac{\hat{\mathbf{p}}_T^{1,1}}{\hat{\mathbf{p}}_T^{1,*}}.$$
3. Take $\hat{p}_0^{(1)} = \frac{\hat{\mathbf{p}}_T^{0,1}}{\hat{\mathbf{p}}_T^{0,*}} \left(1 - p_1\right)$, where $\hat{\mathbf{p}}_T^{0,*} = \hat{\mathbf{p}}_T^{0,0} + \hat{\mathbf{p}}_T^{0,1}$.
4. The estimated total prevalence is: $\hat{p}^{(1)} = \hat{p}_0^{(1)} + \hat{p}_1^{(1)}$.

Remark 1. Although the scenario of all sampled individuals is also considered by Díaz-Pachón and Rao in the context of COVID-19,[10] the correction obtained by Algorithm 1 is stronger than theirs in at least two aspects. First, Díaz-Pachón and Rao considered $m \geq 2$ categories of symptoms, so that, if m is large, an individual with all symptoms is highly likely to be infected; however, with two categories of symptoms, this would correspond to $p_1 = p_1^{(1)}$, which seems a very strong assumption. Second, Algorithm 1 takes into account the information from the sample to obtain $\hat{p}_0^{(1)}$ (see step 3 or (17)); instead, Díaz-Pachón and Rao proposed to take U uniformly distributed in the interval $[0, 1]$ and make $\hat{p}_0^{(1)} = U p_0$.

The following result shows that $\hat{\mathbf{p}}_T^{*,1}$ and $\hat{p}^{(1)}$ are asymptotically normal. More specifically, it is proven that $\hat{\mathbf{p}}_T^{*,1}$ is not a consistent estimator of the true prevalence $p^{(1)}$, but $\hat{p}^{(1)}$ is.

Theorem 1. Suppose $N \to \infty$ in such a way that $p_1 = N_1/N$ is kept fixed. Then

$$N^{1/2} \left(\hat{\mathbf{p}}_T^{*,1} - \frac{p_0^{(1)} q(I_0) + p_1^{(1)}}{p_0 q(I_0) + p_1} \right) \xrightarrow{\mathcal{L}} Z_0, \tag{19}$$

$$N^{1/2} \left(\hat{p}^{(1)} - p^{(1)} \right) \xrightarrow{\mathcal{L}} Z_1, \tag{20}$$

as $N \to \infty$, where $Z_0 \sim \mathcal{N}(0, V_{01} + V_{02})$, $Z_1 \sim \mathcal{N}(0, V_{03})$ are normally distributed, and V_{01}, V_{02}, and V_{03} are defined in the Appendix.

3.1.2 | Not all the symptomatic group is sampled

The main difference between this section and the previous one is that, since now the proportion of symptomatic individuals in the population is not known, it has to be estimated from the sample. Drawing inspiration from cosmological fine-tuning,[14,15] the approach will be to use the information in (13) to generate a maximum entropy distribution, which is "the least biased estimate possible on the given information."[16] Next we will use (14) to obtain estimators of prevalence inside each class of symptoms.

Theorem 2. For $\hat{\mathbf{p}}_T^{1,*}, p_1 \in (0,1)$, $q^*\left(I_0\right) < q^*\left(I_1\right)$ if and only if $\hat{\mathbf{p}}_T^{1,*} > p_1$.

Theorem 2 shows that, given the basic assumption (13), with high probability p_1 is bounded above by $\hat{\mathbf{p}}_T^{1,*}$. On the other hand, $\hat{\mathbf{p}}_T^{1,*} = N_T^{1,*}/N_T$ says that there are at least $N_T^{1,*}$ infected symptomatic individuals in the population. Therefore, it makes sense to bound p_1 as follows:

$$N_T^{1,*}/N \leq p_1 \leq \hat{\mathbf{p}}_T^{1,*}. \tag{21}$$

By the maximum entropy principle,[17] the corrected estimator of p_1 is taken to be the expectation of a uniform distribution over $\left(N_T^{1,*}/N, \hat{\mathbf{p}}_T^{1,*}\right)$. Formally, let U be a uniform distribution over the interval $\left(\hat{\mathbf{p}}_T^{1,*}\frac{N_T}{N}, \hat{\mathbf{p}}_T^{1,*}\right)$. The corrected estimator of p_1 is defined as

$$\hat{p}_1 := E(U) = \frac{\hat{\mathbf{p}}_T^{1,*}}{2}\left(\frac{N_T}{N}+1\right). \tag{22}$$

From this point on, we proceed analogously to Subsubsection 3.1.1, replacing p_1 with $\hat{p}_1$ in equations (16), (17), and (18), to obtain

$$\hat{p}_1^{(1)} := \hat{p}_1 \frac{N_T^{1,1}}{N_T^{1,*}} = \hat{p}_1 \frac{\hat{\mathbf{p}}_T^{1,1}}{\hat{\mathbf{p}}_T^{1,*}}, \tag{23}$$

$$\hat{p}_0^{(1)} := \hat{p}_0 \frac{N_T^{0,1}}{N_T^{0,*}} = \left(1-\hat{p}_1\right)\frac{N_T^{0,1}}{N_T^{0,*}} = \left(1-\hat{p}_1\right)\frac{\hat{\mathbf{p}}_T^{0,1}}{\hat{\mathbf{p}}_T^{0,*}}, \tag{24}$$

$$\hat{p}^{(1)} := \hat{p}_1^{(1)} + \hat{p}_0^{(1)}. \tag{25}$$

Algorithm 2 summarizes the procedure to produce estimators that correct the sampling bias when there is no testing error and not all symptomatic individuals are sampled.

Algorithm 2 Corrected estimator of prevalence without errors and not all symptomatic individuals sampled

1. For $\hat{\mathbf{p}}_T^{1,*} = \hat{\mathbf{p}}_T^{1,0} + \hat{\mathbf{p}}_T^{1,1}$, take
$$\hat{p}_1 = \frac{\hat{\mathbf{p}}_T^{1,*}}{2}\left(\frac{N_T}{N}+1\right).$$
2. Make
$$\hat{p}_1^{(1)} = \hat{p}_1 \frac{\hat{\mathbf{p}}_T^{1,1}}{\hat{\mathbf{p}}_T^{1,*}}.$$
3. Take $\hat{p}_0^{(1)} = \frac{\hat{\mathbf{p}}_T^{0,1}}{\hat{\mathbf{p}}_T^{0,*}}\left(1-\hat{p}_1\right)$, where $\hat{\mathbf{p}}_T^{0,*} = \hat{\mathbf{p}}_T^{0,0} + \hat{\mathbf{p}}_T^{0,1}$.
4. The estimated total prevalence is: $\hat{p}^{(1)} = \hat{p}_0^{(1)} + \hat{p}_1^{(1)}$.

Analogously to the previous asymptotic result, Theorem 3 shows that the naïve and corrected estimators are asymptotically normal. However, once again we find that the naïve estimator is not a consistent estimator of the true prevalence. Moreover, the corrected estimator will only be consistent if $E(\hat{p}_s) = p_s$, for $s = 0, 1$.

Theorem 3. Suppose $N \to \infty$ in such a way that $p_1 = N_1/N$ is kept fixed. Suppose additionally that, for $s = 0, 1$, there exists $\rho_s \in [0, 1]$, such that

$$\hat{p}_s \xrightarrow{p} \rho_s \tag{26}$$

as $N \to \infty$, where "$\xrightarrow{p}$" refers to convergence in probability. Then

$$N^{1/2}\left(\hat{\mathbf{p}}_T^{*,1} - \frac{p_0^{(1)}q(I_0) + p_1^{(1)}q(I_1)}{p_0 q(I_0) + p_1 q(I_1)}\right) \xrightarrow{\mathcal{L}} Z_2, \tag{27}$$

$$N^{1/2}\left(\hat{p}^{(1)} - \frac{\rho_0}{p_0}p_0^{(1)} - \frac{\rho_1}{p_1}p_1^{(1)}\right) \xrightarrow{\mathcal{L}} Z_3, \tag{28}$$

as $N \to \infty$, where $Z_2 \sim \mathcal{N}(0, V_{11} + V_{12})$, $Z_3 \sim \mathcal{N}(0, V_{13} + V_{14})$ are normally distributed random variables, and V_{11}, V_{12}, V_{13}, and V_{14} are given in the Appendix.

4 | WITH TESTING ERRORS

When testing errors are considered, the naïve estimators have an additional source of bias. Using (10), in this section, we present first the explicit form of the naïve estimators in the presence of sampling bias and testing errors stratified by symptoms. This is the most general form of naïve estimator considered in this paper. As a corollary, unstratified errors are also considered. With naïve estimators in this general form, we then present their respective corrections.

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:

$$\begin{aligned}
\tilde{\mathbf{p}}_T^{0,0} &= \left(1 - \breve{\alpha}_0\right)\hat{\mathbf{p}}_T^{0,0} + \breve{\beta}_0\hat{\mathbf{p}}_T^{0,1},\\
\tilde{\mathbf{p}}_T^{0,1} &= \breve{\alpha}_0\hat{\mathbf{p}}_T^{0,0} + \left(1 - \breve{\beta}_0\right)\hat{\mathbf{p}}_T^{0,1},\\
\tilde{\mathbf{p}}_T^{1,0} &= \left(1 - \breve{\alpha}_1\right)\hat{\mathbf{p}}_T^{1,0} + \breve{\beta}_1\hat{\mathbf{p}}_T^{1,1},\\
\tilde{\mathbf{p}}_T^{1,1} &= \breve{\alpha}_1\hat{\mathbf{p}}_T^{1,0} + \left(1 - \breve{\beta}_1\right)\hat{\mathbf{p}}_T^{1,1},
\end{aligned} \tag{29}$$

where $\breve{\alpha}_s$ and $\breve{\beta}_s$, the proportion of false positives and false negatives in the sample, for $s = 0, 1$, approximate α_s and β_s respectively.

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

Corollary 1. If errors are not stratified, for $s = 0, 1$, the estimators of Proposition 2,

$$\begin{aligned}
\tilde{\mathbf{p}}_T^{s,0} &= \left(1 - \breve{\alpha}_s\right)\hat{\mathbf{p}}_T^{s,0} + \breve{\beta}_s\hat{\mathbf{p}}_T^{s,1},\\
\tilde{\mathbf{p}}_T^{s,1} &= \breve{\alpha}_s\hat{\mathbf{p}}_T^{s,0} + \left(1 - \breve{\beta}_s\right)\hat{\mathbf{p}}_T^{s,1},
\end{aligned} \tag{30}$$

are such that $\breve{\alpha}_s$ and $\breve{\beta}_s$, the proportion of false positives and negatives in the sample for symptom class s, approximate the probabilities α and β of false positives and false negatives, respectively, independently of s.

Remark 2. The right-hand side of (29) and (30) contains the contribution to the naïve estimator by each group in the sample weighted by the probability of their errors. However, in either case, the proportions observed by practitioner are the tilde terms $\tilde{\mathbf{p}}_T^{s,i}$ in the left-hand side. The hat terms in the right-hand side, corresponding to (10) are unknown to him.

4.1 | Correction of testing errors

According to Remark 2, when testing errors are considered, estimators that correct them are necessary before applying the correction to sampling bias. This section presents such estimators.

Proposition 3. For $s = 0, 1$, assume $\hat{\alpha}_s$ and $\hat{\beta}_s$ are estimators of α_s and β_s, respectively, obtained from different data, satisfying also that they are independent of $\check{\alpha}_s$ and $\check{\beta}_s$. Then the estimators with correction for testing errors

$$\begin{aligned} \bar{p}_T^{s,0} &= \hat{\mathbf{p}}_T^{s,*} \cdot \frac{\tilde{\mathbf{p}}_T^{s,0}/\hat{\mathbf{p}}_T^{s,*} - \hat{\beta}_s}{1 - \hat{\alpha}_s - \hat{\beta}_s} = \frac{\tilde{\mathbf{p}}_T^{s,0} - \hat{\beta}_s \tilde{\mathbf{p}}_T^{s,*}}{1 - \hat{\alpha}_s - \hat{\beta}_s}, \\ \bar{p}_T^{s,1} &= \hat{\mathbf{p}}_T^{s,*} \cdot \frac{\tilde{\mathbf{p}}_T^{s,1}/\hat{\mathbf{p}}_T^{s,*} - \hat{\alpha}_s}{1 - \hat{\alpha}_s - \hat{\beta}_s} = \frac{\tilde{\mathbf{p}}_T^{s,1} - \hat{\alpha}_s \tilde{\mathbf{p}}_T^{s,*}}{1 - \hat{\alpha}_s - \hat{\beta}_s} \end{aligned} \tag{31}$$

approximate $\hat{\mathbf{p}}_T^{s,0}$ and $\hat{\mathbf{p}}_T^{s,1}$ respectively. Moreover, $\bar{p}_T^{s,0}/\bar{p}_T^{s,*}$ and $\bar{p}_T^{s,1}/\bar{p}_T^{s,*}$ are consistent estimators of $p_s^{(0)}/p_s$ and $p_s^{(1)}/p_s$, if $\hat{\alpha}_s$ and $\hat{\beta}_s$ are in turn consistent estimators of α_s and β_s, respectively.

4.1.1 | All symptomatic group is sampled

When all the symptomatic group is sampled, if $\check{\alpha}_s$ and $\hat{\alpha}_s$ are unbiased estimators of α_s, and $\check{\beta}_s$ and $\hat{\beta}_s$ are unbiased estimators of β_s, for $s = 0, 1$, following Algorithm 1, we obtain Algorithm 3 substituting $\hat{\mathbf{p}}_T^{s,i}$ by $\bar{p}_T^{s,i}$.

4.1.2 | Not all the symptomatic group is sampled

Analogously to Subsubsection 3.1.2, replace $\hat{\mathbf{p}}_T^{s,i}$ with $\bar{p}_T^{s,1}$ in equations (22)-(25), to obtain

$$\hat{p}_1 := \frac{\bar{p}_T^{1,*}}{2}\left(\frac{N_T}{N} + 1\right), \tag{32}$$

$$\hat{p}_1^{(1)} := \hat{p}_1 \frac{\bar{p}_T^{1,1}}{\bar{p}_T^{1,*}}, \tag{33}$$

$$\hat{p}_0^{(1)} := \left(1 - \hat{p}_1\right) \frac{\bar{p}_T^{0,1}}{\bar{p}_T^{0,*}}, \tag{34}$$

$$\hat{p}^{(1)} := \hat{p}_1^{(1)} + \hat{p}_0^{(1)}. \tag{35}$$

This information can be used to generate Algorithm 4. Theorem 4 summarizes the asymptotic behavior of the estimators involved in this section.

Theorem 4. Suppose the conditions of Theorem 3 hold, and that the estimators prevalences $\bar{p}_T^{s,0}/\bar{p}_T^{s,*}$ and $\bar{p}_T^{s,1}/\bar{p}_T^{s,*}$ converge to

$$\bar{\rho}_s^{(0)} = \frac{\beta_s - \bar{\beta}_s}{1 - \bar{\alpha}_s - \bar{\beta}_s} + \frac{1 - \alpha_s - \beta_s}{(1 - \bar{\alpha}_s - \bar{\beta}_s)} \cdot \frac{p_s^{(0)}}{p_s}, \tag{36}$$

$$\bar{\rho}_s^{(1)} = \frac{\alpha_s - \bar{\alpha}_s}{1 - \bar{\alpha}_s - \bar{\beta}_s} + \frac{1 - \alpha_s - \beta_s}{(1 - \bar{\alpha}_s - \bar{\beta}_s)} \cdot \frac{p_s^{(1)}}{p_s} \tag{37}$$

as $N \to \infty$. Suppose further that the proportion of individuals wrongly classified in the sample $\check{\alpha}_s$, $\check{\beta}_s$, $s = 0, 1$ are independent. Then

$$N^{1/2}\left(\bar{p}_T^{*,1} - \frac{p_0 q(I_0)\bar{\rho}_0^{(1)} + p_1 q(I_1)\bar{\rho}_1^{(1)}}{q}\right) \xrightarrow{\mathcal{L}} Z_4, \tag{38}$$

$$N^{1/2}\left[\hat{p}^{(1)} - \left(\rho_0 \bar{\rho}_0^{(1)} + \rho_1 \bar{\rho}_1^{(1)}\right)\right] \xrightarrow{\mathcal{L}} Z_5, \tag{39}$$

as $N \to \infty$, where $\hat{p}^{(1)}$ is defined as in (35), $Z_4 \sim \mathcal{N}(0, V_5 + V_6)$, $Z_5 \sim \mathcal{N}(0, V_7 + V_8)$ are normally distributed random variables, and V_5, V_6, V_7, and V_8 are defined in the Appendix.

Remark 3. If stratification is ignored, throughout all this section just take $\alpha = \alpha_0 = \alpha_1$ and $\hat{\alpha} = \hat{\alpha}_0 = \hat{\alpha}_1$. On the other hand, $\check{\alpha}_0$ and $\check{\alpha}_1$ are still distinct quantities, since they correspond to the observed positive error rates of stratum $s = 0$ and $s = 1$ respectively. Then do analogously with the beta terms to obtain β, $\hat{\beta}$, $\check{\beta}_1$ and $\check{\beta}_2$.

Algorithm 3 Corrected estimator of prevalence with errors and all symptomatic individuals sampled

1. For $s = 0, 1$, make

$$\bar{p}_T^{s,0} = \frac{\tilde{\mathbf{p}}_T^{s,0} - \hat{\beta}_s \tilde{\mathbf{p}}_T^{s,*}}{1 - \hat{\alpha}_s - \hat{\beta}_s},$$
$$\bar{p}_T^{s,1} = \frac{\hat{\alpha}_s \tilde{\mathbf{p}}_T^{s,*} - \tilde{\mathbf{p}}_T^{s,1}}{\hat{\alpha}_s + \hat{\beta}_s - 1},$$

2. If $\bar{\alpha}_s = E\hat{\alpha}_s$ equals $E\check{\alpha}_s = \alpha_s$ and if $\bar{\beta}_s = E\hat{\beta}_s$ equals $E\check{\beta}_s = \beta_s$, for $\bar{p}_T^{1,*} = \bar{p}_T^{1,0} + \bar{p}_T^{1,1}$, take

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

3. Make

$$\hat{p}_1^{(1)} = \hat{p}_1 \frac{\bar{p}_T^{1,1}}{\bar{p}_T^{1,*}}.$$

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

Algorithm 4 Corrected estimator of prevalence with errors and not all symptomatic individuals sampled

1. For $s = 0, 1$, make

$$\bar{p}_T^{s,0} = \frac{\tilde{\mathbf{p}}_T^{s,0} - \hat{\beta}_s \tilde{\mathbf{p}}_T^{s,*}}{1 - \hat{\alpha}_s - \hat{\beta}_s},$$
$$\bar{p}_T^{s,1} = \frac{\hat{\alpha}_s \tilde{\mathbf{p}}_T^{s,*} - \tilde{\mathbf{p}}_T^{s,1}}{\hat{\alpha}_s + \hat{\beta}_s - 1},$$

2. If $\bar{\alpha}_s = E\hat{\alpha}_s$ equals $E\check{\alpha}_s = \alpha_s$ and if $\bar{\beta}_s = E\hat{\beta}_s$ equals $E\check{\beta}_s = \beta_s$, for $\bar{p}_T^{1,*} = \bar{p}_T^{1,0} + \bar{p}_T^{1,1}$, take

$$\hat{p}_1 = \frac{\bar{p}_T^{1,*}}{2} \left(\frac{N_T}{N} + 1 \right).$$

3. Make

$$\hat{p}_1^{(1)} = \hat{p}_1 \frac{\bar{p}_T^{1,1}}{\bar{p}_T^{1,*}}.$$

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

5 | DATA FROM THE ISRAELI MINISTRY OF HEALTH

In what follows, COVID-19 data from the Israeli Ministry of Health is considered.[18] 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 as the population, consisting of 99 232 tested individuals, among 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 2.

TABLE 2 Observed disease status by category of symptoms.

	Positive	Negative	Total
Symptomatic	1 754	108	1 862
Asymptomatic	6 639	90 731	97 370
Total	8 393	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. Although we do not have exact numbers of stratified false-positive and false-negative rates for PCR tests, a public report from UK Government Office for Science in 2020 indicated that the median false positive rate in the UK's COVID-19 RT-PCR testing program is 2.3% with IQR of 0.8% to 4.0%.[19] Moreover, Arévalo-Rodríguez et al stated that after they collected information among all patients from 34 studies, the summary estimate of the false-negative rate was 13% with range of 1.8% to 58%.[20] For the purpose of investigating how the estimates change with different assumed error rates, we will compare different combinations of error rates within the reasonable ranges according to the literature we found. The combinations of error rates in the population are assumed to have the values in (40). The actual number of individuals inside each group can be found in Table 3 after correcting Table 2 for these errors.

$$\begin{aligned}
&\textbf{Comb. 1} \quad \alpha_0 = 1\%, \quad \alpha_1 = 3\%, \quad \beta_0 = 20\%, \quad \beta_1 = 2\%;\\
&\textbf{Comb. 2} \quad \alpha_0 = 1\%, \quad \alpha_1 = 4\%, \quad \beta_0 = 10\%, \quad \beta_1 = 2\%;\\
&\textbf{Comb. 3} \quad \alpha_0 = 2\%, \quad \alpha_1 = 3\%, \quad \beta_0 = 15\%, \quad \beta_1 = 5\%.
\end{aligned} \tag{40}$$

TABLE 3 Real proportions under stratified errors with different error combinations

Error Comb.	Symptoms	Disease	Non-Disease	Total
	Sympt.	1 704	158	1 862
Comb. 1	**Asympt.**	24 719	72 651	97 370
	Total	26 423	72 809	99 232
	Sympt.	1 686	176	1 862
Comb. 2	**Asympt.**	15 646	81 724	97 370
	Total	17 332	81 900	99 232
	Asympt.	1 707	155	1 862
Comb. 3	**Asympt.**	20 116	77 254	97 370
	Total	21 823	77 409	99 232

The real prevalence with **Comb. 1** is then

$$p^{(1)} = (1\,704 + 24\,719)/99\,232 = 0.266, \tag{41}$$

and prevalence among the asymptomatic is $24\,719/97\,370 = 0.254$.

The real prevalence with **Comb. 2** is then

$$p^{(1)} = (1\,686 + 15\,646)/99\,232 = 0.175, \tag{42}$$

and prevalence among the asymptomatic is $15\,646/97\,370 = 0.161$.

The real prevalence with **Comb. 3** is then

$$p^{(1)} = (1\,707 + 20\,116)/99\,232 = 0.220, \tag{43}$$

and prevalence among the asymptomatic is $20\,116/97\,370 = 0.207$.

Additionally, in (44) we assume some sample error rates for the combinations (40). We emphasize that the actual values of (44) are "known unknowns" to the practitioner,[21] and it is precisely their effect what needs to be corrected.

$$\begin{array}{llllll} \textbf{Comb. 1} & \breve{\alpha}_0 = 0.7\%, & \breve{\alpha}_1 = 3.5\%, & \breve{\beta}_0 = 25\%, & \breve{\beta}_1 = 4\%; \\ \textbf{Comb. 2} & \breve{\alpha}_0 = 1.2\%, & \breve{\alpha}_1 = 3.6\%, & \breve{\beta}_0 = 12\%, & \breve{\beta}_1 = 3\%; \\ \textbf{Comb. 3} & \breve{\alpha}_0 = 2.5\%, & \breve{\alpha}_1 = 2.9\%, & \breve{\beta}_0 = 18\%, & \breve{\beta}_1 = 10\%. \end{array} \tag{44}$$

Finally, active information (defined in Appendix C) is used to compare how well Algorithms 3 and 4 and other proposed estimators in the literature are doing with respect to the real prevalence. The best estimator will be the one with active information $\hat{I}^+$ 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;[10] Diggle's Bayesian approach, which corrects for imperfect testing but ignores sampling bias;[22] and the Rogan-Gladen estimate, a frequentist method that only corrects for testing errors too.[23] Neither of the competitors corrects for sampling bias and testing errors at the same time. In fact, as much as we searched, we could not find a methodology that simultaneously corrects for imperfect testing and sampling bias; this will be reflected in the analysis. All of the results are presented in Table 8.

Each of the following protocols presents a table with the sample results. These correspond to the observations given by (29). These values, the population size, and the estimated error rates from a different study ($\hat{\alpha}_s$ and $\hat{\beta}_s$) will be the input of an R program, for which the code is available at https://github.com/kalilizhou/BiasCorrection.git, with the four algorithms proposed in this article. The program thus obtains the correction. As a simplifying assumption in the remaining of this section, we take $\hat{\alpha}_s = \alpha_s$ and $\hat{\beta}_s = \beta_s$, for $s = 0, 1$.

Sampling Protocol 1: In the first scenario, all symptomatic individuals are sampled, as considered by Díaz-Pachón and Rao.[10] The sample consists of 2483 individuals. Among these, 1862 (75%) are symptomatic and 621 (25%) are asymptomatic. The sample error rates are taken from (44). The observed sampling results, corresponding to (29), are given by Table 4.

TABLE 4 Stratified observed sample for 75% **symptomatic and** 25% **asymptomatic, with all symptomatic individuals sampled.**

Error Comb.	Symptoms	Positive	Negative	Totoal
Comb. 1	Symptomatics	1 641	221	1 862
	Asymptomatics	122	499	621
	Total	1 763	720	2 483
Comb. 2	Symptomatics	1 642	220	1 862
	Asymptomatics	94	527	621
	Total	1 736	747	2 483
Comb. 3	Symptomatics	1 541	321	1 862
	Asymptomatics	117	503	620
	Total	1 658	824	2 482

According to Table 4, for instance with **Comb. 1**, the naïve estimator is $\tilde{\mathbf{p}}_T^{1,*} = 1\,763/2\,483 \approx 0.71$. Since all the symptomatic group is sample, we use Algorithm 3, which produces the corrected estimator $\hat{p}^{(1)} = 0.248$. Table 8 presents these results as well as those of the other methods.

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 works very well, and RGE performs as poorly as the naïve estimators. Our Algorithm 3 is the best correction to the the naïve

estimate, producing the closest-to-zero actinfo. The corrected estimators of prevalence obtained from Algorithm 3 almost equal the real prevalence for all combinations of testing errors.

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 4 is followed.

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 3. Table 5 shows the observed sample. The summary of results under different methods is shown in Table 8.

TABLE 5 Stratified observed sample for 75% **symptomatic and** 25% **asymptomatic (not all symptomatic individuals sampled).**

Error Comb.	Symptoms	Positive	Negative	Total
	Sympt.	132	18	150
Comb. 1	**Asympt.**	10	40	50
	Total	142	58	200
	Sympt.	132	18	150
Comb. 2	**Asympt.**	8	42	50
	Total	140	60	200
	Sympt.	125	25	150
Comb. 3	**Asympt.**	9	41	50
	Total	134	66	200

Table 8 shows that, in this sampling scenario, our Algorithm 4 still has the best performance. In fact, Diggle's and Rogan-Gladen's estimates do as poorly as the naïve estimate. Algorithm 4 beats its competitors because it is the only one that corrects for sampling 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 the correction, as reflected by the active information.

Sampling Protocol 3: In this scenario there are 100 symptomatic and 100 asymptomatic individuals. Again, the proportions inside each group were taken from Table 3. The observed sample is given in Table 6. After correcting the estimates, the summary of results under different methods for this sampling protocol is also presented in Table 8.

TABLE 6 Stratified observed sample for 50% **symptomatic and** 50% **asymptomatic.**

Error Comb.	Symptoms	Positive	Negative	Total
	Sympt.	89	11	100
Comb. 1	**Asympt.**	19	81	100
	Total	108	92	200
	Sympt.	89	11	100
Comb. 2	**Asympt.**	15	85	100
	Total	104	96	200
	Sympt.	83	17	100
Comb. 3	**Asympt.**	19	81	100
	Total	102	98	200

Compared to Protocol 2, Protocol 3 has less sampling bias. Therefore, all the methods perform better than in the previous scenario. But Algorithm 4 still works better than competitors.

Sampling Protocol 4: This sample is truly random, with $N_T = 200$, and it is obtained from Table 3. The observed sample is presented in Table 7. The results of the different methods for this scenario are presented in Table 8.

TABLE 7 Stratified observed random sample of size 200.

Error Comb.	Symptoms	Positive	Negative	Total
Comb. 1	**Sympt.**	3	1	4
	Asympt.	39	157	196
	Total	42	158	200
Comb. 2	**Sympt.**	3	1	4
	Asympt.	30	166	196
	Total	33	167	200
Comb. 3	**Sympt.**	3	1	4
	Asympt.	37	159	196
	Total	40	160	200

TABLE 8 Comparison among methods of all sampling protocols with different combinations of stratified error rates.

			Corrected Estimates $\hat{p}^{(1)}$ (Active Information $\hat{I}^{+}$)				
Error Comb.	**Real Prevalence**	**Sampling Protocol**	**Naïve**	**Díaz-Rao**	**Diggle**	**Rogan-Gladen**	**Algorithms 3 and 4**
Comb. 1	**p = 0.266**	1	0.710 (0.981)	0.212 (-0.229)	-	0.729 (1.009)	**0.248 (-0.069)**
		2	0.710 (0.981)	-	0.735 (1.016)	0.729 (1.009)	**0.486 (0.602)**
		3	0.540 (0.707)	-	0.535 (0.699)	0.564 (0.751)	**0.398 (0.401)**
		4	0.210 (-0.237)	-	0.003 (-4.667)	**0.249 (-0.066)**	0.244 (-0.086)
Comb. 2	**p = 0.175**	1	0.699 (1.387)	0.167 (-0.045)	-	0.711 (1.402)	**0.173 (-0.011)**
		2	0.700 (1.388)	-	0.695 (1.379)	0.712 (1.403)	**0.441 (0.926)**
		3	0.520 (1.091)	-	0.493 (1.035)	0.530 (1.109)	**0.344 (0.679)**
		4	0.165 (-0.057)	-	0.003 (-4.248)	0.173 (-0.014)	0.168 (-0.047)
Comb. 3	**p = 0.220**	1	0.668 (1.111)	0.204 (-0.076)	-	0.701 (1.159)	**0.215 (-0.019)**
		2	0.670 (1.114)	-	0.690 (1.143)	0.703 (1.161)	**0.448 (0.713)**
		3	0.510 (0.841)	-	0.503 (0.826)	0.537 (0.892)	**0.371 (0.524)**
		4	0.200 (-0.095)	-	0.003 (-4.477)	**0.215 (-0.024)**	0.209 (-0.050)

In this scenario, without sampling bias, all estimates perform extremely well except Diggle's estimate. Diggle's Bayesian estimates grossly overcorrect to the point of removing more than 4 nats of information with respect to the real prevalences. On the other hand, Rogan-Gladen's frequentist estimates are optimal, and Algorithm 4 work quite well. The results of these two methods are very close to the real prevalence.

Table 8 summarizes the results of all protocols and all combinations of testing errors. Our proposed Algorithms 3 (first row in each combination) and 4 (rows 2-4 in each combination) are always the best in Sampling Protocols 1-3 and perform as well as the naïve estimator under random sampling (Protocol 4), according to the actinfo assessment. The result shows the promising ability of our proposed algorithms to correct for both sampling bias and testing errors in prevalence estimation.

6 | 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. Such biased samples

are here shown to be inconsistent in terms of not converging to the true proportion of infected individuals. In addition, Algorithms 1, 2, 3, and 4 go further and present corrections both for sampling bias and testing errors. Such corrections either eliminate bias completely (Algorithms 1 and 3), or reduce it substantially (Algorithms 2 and 4) when testing error rates are known or can be estimated consistently. However, the methodology generalizes easily regardless of how the biased samples resulted.

A limitation of our study is that we do not estimate error rates directly from our sample, but take the estimator from a previous independent sample. 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.[22] This naturally results in increased variability of the prevalence estimate and relies on a reasonable prior distribution being elicited for the prevalence.

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.[24] 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.[25,26] 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.[27]

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,[28] 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.

In First-World countries, particularly in urban areas, testing practices seem to be well-described by oversampling of symptomatic individuals, sometimes even testing the whole group in a subpopulation, as it is the case with COVID-19 testing in universities and industries. A possible extension, however, is to consider the opposite situation in which the symptomatic group is under-sampled, producing an estimator that is biased because it underestimates prevalence. Such scenario is certainly relevant for COVID-19 too in several Third World countries, and even in difficult-to-reach subpopulations of First World Countries.

Author contributions

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

Financial disclosure

D.A.D.P., C.Z., and J.S.R. acknowledge the support of the Copeland Foundation Award 2022 from the Department of Public Health Sciences at the University of Miami.

Conflict of interest

The authors declare no potential conflict of interests.

Data availability statement

Code to implement Algorithms 1-4 is available at https://github.com/kalilizhou/BiasCorrection.git. The data used in Section 5 is publicly available at https://github.com/nshomron/COVIDpred.

Acknowledgements

The authors thank the comments from the two anonymous reviewers that greatly improved the quality and readability of this manuscript.

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

A GENERAL PROOFS

Proof of Proposition 1.

$$\begin{aligned}
\hat{\mathbf{p}}_T^{s,i} &= f_{S^*|T}\left(I_s^{(i)} \mid T=1\right) \\
&= \frac{P\left[T=1 \mid S^* = I_s^{(i)}\right]}{P[T=1]} f_{S^*}\left(I_s^{(i)}\right) \\
&= \frac{q\left(I_s^{(i)}\right)}{P[T=1]} f_{S^*}\left(I_s^{(i)}\right) \\
&= \frac{N_T^{s,i}/N_s^{(i)}}{N_T/N} \frac{N_s^{(i)}}{N} \\
&= \frac{N_T^{s,i}}{N_T}.
\end{aligned} \tag{A1}$$

Remember that terms without subindex T are here population values, whereas terms with the subindex T are sample values. Notice in the fourth and fifth steps that $N_s^{(i)}$ cancels. Therefore, all the remaining information about $I_s^{(i)}$ comes from $N_T^{s,i}$, the number of tested (sampled) individuals with symptoms s and infectious status i. Now, from the third equality, this value is seen to come from $q\left(I_s^{(i)}\right)$, the sampling probability of the group $I_s^{(i)}$. Therefore, all knowledge of $I_s^{(i)}$ comes from whatever knowledge we have about the sample mechanism $q\left(I_s^{(i)}\right)$. □

Proof of Theorem 2.

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

where the fourth step used that $p_0 + p_1 = 1 = \hat{\mathbf{p}}_T^{0,*} + \hat{\mathbf{p}}_T^{1,*}$. □

Proof of Proposition 2. This proposition follows directly from the definition of testing errors. Consider for instance the first equation of (29). It stipulates that whereas the proportion of sampled individuals with $s = i = 0$ is $\hat{\mathbf{p}}_T^{0,0}$, the reported fraction $\tilde{\mathbf{p}}_T^{0,0}$ of sampled individuals with $s = i = 0$ differs from $\hat{\mathbf{p}}_T^{0,0}$ by an amount $\breve{\beta}_0 \hat{\mathbf{p}}_T^{0,1} - \breve{\alpha}_0 \hat{\mathbf{p}}_T^{0,0}$, where the first term $\breve{\beta}_0 \hat{\mathbf{p}}_T^{0,1}$ is the fraction of $(0,1)$-individuals wrongly classified as $(0,0)$, whereas the second term $\breve{\alpha}_0 \hat{\mathbf{p}}_T^{0,0}$ is the fraction of $(0,0)$-individuals wrongly classified as $(0,1)$. The other three equations of (29) are motivated analogously. □

Proof of Proposition 3. It follows from equations (29) and (31) that $\bar{p}_T^{s,*} = \tilde{\mathbf{p}}_T^{s,*} = \hat{\mathbf{p}}_T^{s,*}$ for $s = 0, 1$. This, and another application of (29) and (31) gives

$$\frac{\bar{p}_T^{s,0}}{\bar{p}_T^{s,*}} = \frac{\tilde{\mathbf{p}}_T^{s,0}/\hat{\mathbf{p}}_T^{s,*} - \hat{\beta}_s}{1 - \hat{\alpha}_s - \hat{\beta}_s} = \frac{\hat{\mathbf{p}}_T^{s,0}}{\hat{\mathbf{p}}_T^{s,*}} \cdot \frac{1 - \breve{\alpha}_s - \breve{\beta}_s}{1 - \hat{\alpha}_s - \hat{\beta}_s} + \frac{\breve{\beta}_s - \hat{\beta}_s}{1 - \hat{\alpha}_s - \hat{\beta}_s}. \tag{A2}$$

Since $q(I_s^{(0)}) = q(I_s^{(1)})$ by (14), it follows that $\hat{\mathbf{p}}_T^{s,0}/\hat{\mathbf{p}}_T^{s,*}$ is a consistent estimator of $p_s^{(0)}/p_s$, and by assumption, $\hat{\alpha}_s$ and $\hat{\beta}_s$ are consistent estimators of α_s and β_s respectively. Moreover, Lemma 2 below implies that $\breve{\alpha}_s$ and $\breve{\beta}_s$ are consistent estimators of α_s and β_s as well. From this and (A2) it follows that $\bar{p}_T^{s,0}/\bar{p}_T^{s,*}$ is a consistent estimator of $p_s^{(0)}/p_s$. The fact that $\bar{p}_T^{s,1}/\bar{p}_T^{s,*}$ is a consistent estimator of $p_s^{(1)}/p_s$ is proved in the same way. □

B ASYMPTOTICS

As a preparation, we prove the following lemma that will be used as assumption in the main result of this section:

Lemma 1. $q\left(I_s^{(0)}\right) = q\left(I_s^{(1)}\right) = q\left(I_s\right)$.

Proof. The first equality is obtained by assumption (14). In order to prove the second equality, we use that a randomly chosen individual from I_s belongs to $I_s^{(i)}$ with probability $p_s^{(i)}/p_s$ for $i = 0, 1$. Conditioning on which subcohort of I_s the individual belongs to, it follows that

$$q(I_s) = \frac{p_s^{(0)}}{p_s} q(I_s^{(0)}) + \frac{p_s^{(1)}}{p_s} q(I_s^{(1)}).$$

Therefore, since $q(I_s)$ is a weighed average of $q(I_s^{(0)})$ and $q(I_s^{(1)})$, the second equality of Lemma 1 follows from the first one. □

Hössjer et al. (2023) proved a couple of theorems that we will use to prove the asymptotic results for the estimators discussed in this paper.[27] We present them here for completeness, fitting their notation to ours.

Theorem 5 (Theorem 1 of Hössjer et al. (2023))**.** Let $N \to \infty$ in such a way that N_1/N is always fixed, that Lemma 1 holds for fixed $q(I_0)$ and $q(I_1)$, and that there exists $\rho_s \in [0, 1]$ such that $\hat{p}_s \xrightarrow{p} \rho_s$, as $N \to \infty$, for $s = 0, 1$. Then

$$\begin{aligned} N^{1/2}\left(\hat{\mathbf{p}}_T^{*,1} - \frac{p_0^{(1)} q(I_0) + p_1^{(1)} q(I_1)}{q}\right) &\xrightarrow{\mathcal{L}} \mathcal{N}(0, V_1 + V_2), \\ N^{1/2}\left(\hat{p}^{(1)} - \left(\frac{\rho_0 p_0^{(1)}}{p_0} + \frac{\rho_1 p_1^{(1)}}{p_1}\right)\right) &\xrightarrow{\mathcal{L}} \mathcal{N}(0, V_3 + V_4), \\ N^{1/2}(\hat{p}_s - \rho_s) &\xrightarrow{\mathcal{L}} \mathcal{N}(0, C_{00}) \end{aligned} \tag{B3}$$

as $N \to \infty$, where q is defined in (8),

$$\begin{aligned} V_1 &= \sum_{s=0}^{1} \frac{p_s q(I_s)\left[1 - q(I_s)\right] \frac{p_s^{(1)}}{p_s}\left(1 - \frac{p_s^{(1)}}{p_s}\right)}{\left(p_0 q(I_0) + p_1 q(I_1)\right)^2}, \\ V_2 &= \sum_{s=0}^{1} \frac{p_s q(I_s)\left[1 - q(I_s)\right]\left(\frac{p_s^{(1)}}{p_s} - \frac{p_0^{(1)} q(I_0) + p_1^{(1)} q(I_1)}{p_0 q(I_0) + p_1 q(I_1)}\right)^2}{\left(p_0 q(I_0) + p_1 q(I_1)\right)^2} \\ V_3 &= \sum_{s=0}^{1} \rho_s^2 p_s^{-1} \frac{1 - q(I_s)}{q(I_s)} \frac{p_s^{(1)}}{p_s}\left(1 - \frac{p_s^{(1)}}{p_s}\right), \\ V_4 &= C_{00}\left(\frac{p_0^{(1)}}{p_0} - \frac{p_1^{(1)}}{p_1}\right)^2, \end{aligned}$$

and

$$C_{00} = \frac{(1+q)^2 B_{11} + \left(\frac{p_1 q(I_1)}{q}\right)^2 \left[p_0 q(I_0)(1-q(I_0)) + p_1 q(I_1)(1-q(I_1))\right] + 2(1+q)\frac{p_1 q(I_1)}{q}\Sigma_{\pi p1}}{4},$$

$$B_{11} = \frac{p_1 q(I_1)\left\{q^2\left[1-q(I_1)\right] - 2qp_1 q(I_1)\left[1-q(I_1)\right] + p_1 q(I_1) p_0 q(I_0)\left[1-q(I_0)\right] + p_1^2 q^2(I_1)\left[1-q(I_1)\right]\right\}}{q^4},$$

$$\Sigma_{\pi p1} = \frac{p_1 q(I_1)\left[1-q(I_1)\right]}{q} - \frac{p_1 q(I_1)\left\{p_0 q(I_0)\left[1-q(I_0)\right] + p_1 q(I_1)\left[1-q(I_1)\right]\right\}}{q^2}.$$

Proof of Theorem 1. Since $\hat{p}_1 = p_1$ and $\hat{p}_0 = p_0 = 1 - p_1$ are known, convergence in probability $\hat{p}_s \xrightarrow{p} \rho_s$ trivially holds with $\rho_s = p_s$, whereas the asymptotic weak limit of $N^{1/2}(\hat{p}_s - p_s)$ in (B3) degenerates to a one point distribution at 0 ($C_{00} = 0$). Since $q(I_1) = 1$, V_1, V_2, and V_3, in Theorem 5 are readily simplified to

$$V_{01} = \frac{p_0 q(I_0)\left[1-q(I_0)\right]\frac{p_0^{(1)}}{p_0}\left(1-\frac{p_0^{(1)}}{p_0}\right)}{\left(p_0 q(I_0) + p_1\right)^2}, \tag{B4}$$

$$V_{02} = \frac{p_0 q(I_0)\left[1-q(I_0)\right]\left(\frac{p_0^{(1)}}{p_0} - \frac{p_0^{(1)} q(I_0) + p_1^{(1)}}{p_0 q(I_0) + p_1}\right)^2}{\left(p_0 q(I_0) + p_1\right)^2}, \tag{B5}$$

$$V_{03} = p_0 \frac{1-q(I_0)}{q(I_0)} \frac{p_0^{(1)}}{p_0}\left(1-\frac{p_0^{(1)}}{p_0}\right), \tag{B6}$$

whereas $V_4 = 0$ since $C_{00} = 0$. □

Proof of Theorem 3. It is obtained directly from Theorem 5 by substituting V_1, V_2, V_3, and V_4 for V_{11}, V_{12}, V_{13}, and V_{14}, respectively, where

$$V_{11} = \sum_{s=0}^{1} \frac{p_s q(I_s)\left[1-q(I_s)\right]\frac{p_s^{(1)}}{p_s}\left(1-\frac{p_s^{(1)}}{p_s}\right)}{\left(p_0 q(I_0) + p_1 q(I_1)\right)^2}, \tag{B7}$$

$$V_{12} = \sum_{s=0}^{1} \frac{p_s q(I_s)\left[1-q(I_s)\right]\left(\frac{p_s^{(1)}}{p_s} - \frac{p_0^{(1)} q(I_0) + p_1^{(1)} q(I_1)}{p_0 q(I_0) + p_1 q(I_1)}\right)^2}{\left(p_0 q(I_0) + p_1 q(I_1)\right)^2}, \tag{B8}$$

$$V_{13} = \sum_{s=0}^{1} \rho_s^2 p_s^{-1} \frac{1-q(I_s)}{q(I_s)} \frac{p_s^{(1)}}{p_s}\left(1-\frac{p_s^{(1)}}{p_s}\right), \tag{B9}$$

$$V_{14} = C_{00}\left(\frac{p_0^{(1)}}{p_0} - \frac{p_1^{(1)}}{p_1}\right)^2. \tag{B10}$$

□

B.1 Asymptotics with testing errors

Before proving the asymptotic results with testing errors, some previous assumptions and results are used. First, we assume that the existing estimators of error rates are asymptotically normal. That is, there exists $\bar{\boldsymbol{\alpha}} = (\bar{\alpha}_0, \bar{\alpha}_1)$ and $\bar{\boldsymbol{\beta}} = (\bar{\beta}_0, \bar{\beta}_1)$ such that, for $\hat{\boldsymbol{\alpha}} = (\hat{\alpha}_0, \hat{\alpha}_1)$ and $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1)$,

$$N^{1/2}\left(\hat{\boldsymbol{\alpha}} - \bar{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}}\right) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \begin{pmatrix} \Omega_{\alpha\alpha} & \Omega_{\alpha\beta} \\ \Omega_{\alpha\beta}^T & \Omega_{\beta\beta} \end{pmatrix}\right) \tag{B11}$$

as $N \to \infty$, where each of the terms in the variance-covariance matrix is a 2×2 matrix, and

$$\begin{aligned} \Omega_{\alpha\alpha} &= \left(\Omega_{\alpha\alpha rs}\right)_{r,s=0,1} \\ \Omega_{\beta\beta} &= \left(\Omega_{\beta\beta rs}\right)_{r,s=0,1} \\ \Omega_{\alpha\beta} &= \left(\Omega_{\alpha\beta rs}\right)_{r,s=0,1}. \end{aligned} \tag{B12}$$

We also assume that $\check{\alpha}_0$, $\check{\alpha}_1$, $\check{\beta}_0$, and $\check{\beta}_1$ are all independent. Moreover, the following Lemma will be used:

Lemma 2 (Lemma 2 from Hössjer et al. (2023))**.** The proportions of false positive and negatives in the sample satisfy

$$N^{1/2}(\check{\alpha}_s - \alpha_s) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma_{\alpha\alpha s}) \tag{B13}$$

and

$$N^{1/2}(\check{\beta}_s - \beta_s) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma_{\beta\beta s}) \tag{B14}$$

respectively as $N \to \infty$, with

$$\begin{aligned} \Sigma_{\alpha\alpha s} &= \alpha_s(1-\alpha_s)q / \left[p_s q(I_s)\left(1 - p_s^{(1)}/p_s\right)\right], \\ \Sigma_{\beta\beta s} &= \beta_s(1-\beta_s)qp_s / \left[p_s q(I_s) p_s^{(1)}\right]. \end{aligned}$$

After transforming the notation of Hössjer et al. (2023) to our notation, the asymptotic results of Theorem 4 are a direct consequence of the following result from Hössjer et al. (2023):

Theorem 6 (Theorem 2 of Hössjer et al. (2023))**.** Suppose the conditions of Theorem 3 hold, and additionally that the estimators $\bar{p}_T^{s,0}/\bar{p}_T^{s,*}$ and $\bar{p}_T^{s,1}/\bar{p}_T^{s,*}$ of prevalences of unaffected and affected in symptom group s converge to

$$\bar{\rho}_s^{(0)} = \frac{\beta_s - \bar{\beta}_s}{1 - \bar{\alpha}_s - \bar{\beta}_s} + \frac{1 - \alpha_s - \beta_s}{(1 - \bar{\alpha}_s - \bar{\beta}_s)} \cdot \frac{p_s^{(0)}}{p_s}, \tag{B15}$$

$$\bar{\rho}_s^{(1)} = \frac{\alpha_s - \bar{\alpha}_s}{1 - \bar{\alpha}_s - \bar{\beta}_s} + \frac{1 - \alpha_s - \beta_s}{(1 - \bar{\alpha}_s - \bar{\beta}_s)} \cdot \frac{p_s^{(1)}}{p_s} \tag{B16}$$

respectively as $N \to \infty$. Suppose further that the proportion of individuals wrongly classified in the sample $\check{\alpha}_s$, $\check{\beta}_s$, $s = 0, 1$ are independent. Then

$$N^{1/2}\left(\bar{p}_T^{*,1} - \frac{p_0 q(I_0)\bar{\rho}_0^{(1)} + p_1 q(I_1)\bar{\rho}_1^{(1)}}{q}\right) \xrightarrow{\mathcal{L}} \mathcal{N}(0, V_5 + V_6), \tag{B17}$$

$$N^{1/2}\left[\hat{p}^{(1)} - \left(\rho_0\bar{\rho}_0^{(1)} + \rho_1\bar{\rho}_1^{(1)}\right)\right] \xrightarrow{\mathcal{L}} \mathcal{N}(0, V_7 + V_8) \tag{B18}$$

as $N \to \infty$, where $\hat{p}^{(1)}$ is defined as in (35),

$$\begin{aligned} V_5 &= \sum_{r,s} \frac{p_r q(I_r) p_s q(I_s)}{q^2} \bar{A}_{rs}, \\ V_6 &= \sum_{r,s} \bar{\rho}_r^{(1)} \bar{\rho}_s^{(1)} B_{rs}, \\ V_7 &= \sum_{r,s} \rho_r \rho_s \bar{A}_{rs}, \\ V_8 &= \sum_{r,s} \bar{\rho}_r^{(1)} \bar{\rho}_s^{(1)} C_{rs}, \end{aligned}$$

whereas

$$\begin{aligned} B_{11} = B_{00} &= -B_{01} = -B_{10}, \\ C_{00} = C_{11} &- C_{01} = -C_{10}, \end{aligned}$$

are defined in Theorem 5. Moreover,

$$\begin{aligned} \bar{A}_{ss} &= K_{s1}^2 A_{ss} + K_{s2}^2 \Sigma_{\alpha\alpha s} + K_{s3}^2 \Sigma_{\beta\beta s} \\ &+ K_{s4}^2 \Omega_{\alpha\alpha ss} + K_{s5}^2 \Omega_{\beta\beta ss} + 2K_{s4}K_{s5}\Omega_{\alpha\beta ss}. \end{aligned} \tag{B19}$$

and

$$\begin{aligned} \bar{A}_{rs} &= K_{r4}K_{s4}\Omega_{\alpha\alpha rs} + K_{r5}K_{s5}\Omega_{\beta\beta rs} \\ &+ K_{r4}K_{s5}\Omega_{\alpha\beta rs} + K_{r5}K_{s4}\Omega_{\alpha\beta sr} \end{aligned} \tag{B20}$$

when $r \neq s$, with

$$\begin{aligned} K_{s1} &= \left(1-\alpha_s-\beta_s\right)/K_s, \\ K_{s2} &= \left(1-\frac{p_s^{(1)}}{p_s}\right)/K_s, \\ K_{s3} &= -p_s^{(1)}/\left(p_s K_s\right), \\ K_{s4} &= \left[\alpha_s+\bar{\beta}_s-1+\frac{p_s^{(1)}\left(1-\alpha_s-\beta_s\right)}{p_s}\right]/K_s^2, \\ K_{s5} &= \left[\alpha_s-\bar{\alpha}_s+\frac{p_s^{(1)}\left(1-\alpha_s-\beta_s\right)}{p_s}\right]/K_s^2, \\ K_s &= 1-\bar{\alpha}_s-\bar{\beta}_s, \end{aligned} \tag{B21}$$

and finally

$$A_{ss} = \frac{\left(1-q(I_s)\right)\frac{p_s^{(1)}}{p_s}\left(1-\frac{p_s^{(1)}}{p_s}\right)}{p_s q(I_s)}. \tag{B22}$$

Notice that, if $\bar{\alpha}_s = \alpha_s$ and $\bar{\beta}_s = \beta_s$ (that is, if the error rates are estimated consistently), then $\bar{p}_T^{s,i}/p_T^{s,*}$ is a consistent estimator of $\bar{\rho}_s^{(i)} = p_s^{(i)}/p_s$ for $i = 0, 1$. In particular, $\bar{p}_T^{s,1}/p_T^{s,*}$ is a consistent estimator of the prevalence $\bar{\rho}_s^{(1)} = p_s^{(i)}/p_s$ among the individuals of symptom group s.

C 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.[29,30,31] 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.[32,33,34] It has also been used for mode hunting,[35,36], and to compare neutral to non-neutral models in population genetics.[37]

We now use active information to analyze the bias. Through the eyes of actinfo, the bias is formally seen as the addition (if the parameter is overestimated) or subtraction (if the parameter is underestimated) of relevant information in the estimation of the parameter. Formally, active information is defined as

$$\hat{I}^+ = \log\left(\hat{p}^{(1)}/p^{(1)}\right), \tag{C23}$$

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 $\hat{p}^{(1)}$ to the true proportion $p^{(1)}$, and it is the quantity that is averaged in the Kullback-Leibler divergence.[38] 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.[39,40] Because of Theorem 4, we interpret (C23) as an approximation of $I^+ = \log[(\rho_0\bar{\rho}_0^{(1)} + \rho_1\bar{\rho}_1^{(1)})/p^{(1)}]$.

D 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 15%,
- the proportion of negative cases with symptoms $p_1^{(0)}$ is 5%,
- the proportion of positive cases without symptoms $p_0^{(1)}$ is 5%,
- and proportion of negative cases without symptoms $p_0^{(0)}$ is 75%.

Thus, the prevalence is $p_0^{(1)} + p_1^{(1)} = 20\%$, the proportion of symptomatic individuals in the population is $p_1^{(1)} + p_1^{(0)} = 20\%$, so the proportion of asymptomatic in the population is $p_0^{(1)} + p_0^{(0)} = 80\%$.

D.1 Correction without testing errors

We will run the simulation for multiple proportions of asymptomatic individuals getting sampled and will compare the results using a boxplot. The true prevalence will be known in the simulation, allowing us to evaluate the accuracy of our estimators.

D.1.1 All the symptomatic group is sampled

Assuming a population size of 10 000, we initially sampled all symptomatic individuals. However, we increased the sample size by including more asymptomatic individuals as we changed the proportion of those getting tested. This resulted in both the corrected and naive estimators approaching the true value, and the variance of the estimators decreasing, as shown in Figure D1.

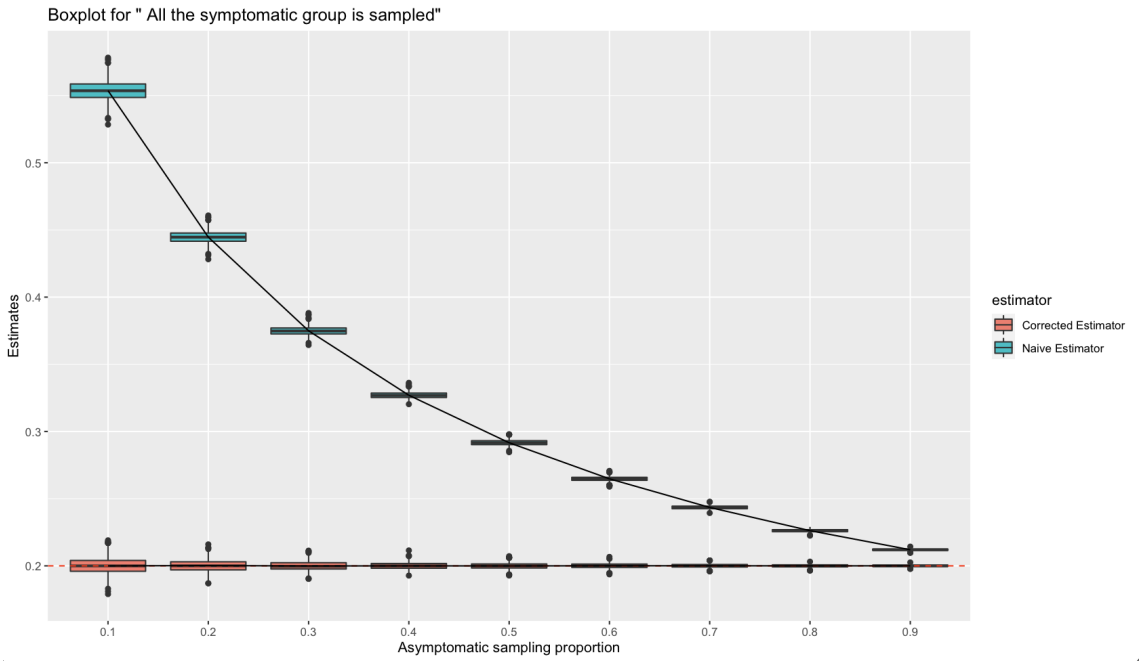


FIGURE D1 All the symptomatic group is sampled

D.2 Not all the symptomatic group is sampled

In this scenario, we assume that only 70% of symptomatic individuals underwent testing. From Figure D2, as the number of asymptomatic individuals in the total testing sample size increases, we observe that the corrected estimator converges to the true value faster than the naïve estimator. Thus we see once again that the testing error correction improves the accuracy of prevalence estimation. By Table D1, it can be seen that our estimated symptomatic rate in population $\hat{p}_1$ decreases as the number of asymptomatic individuals in the sample increases.

TABLE D1 Estimated proportion of symptomatic in the population $\hat{p}_1$ (E(U))

$\hat{\mathbf{p}}_T^{0,*}$	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
E(U)	0.3881	0.3031	0.2538	0.2221	0.1995	0.1829	0.1701	0.1596	0.1513

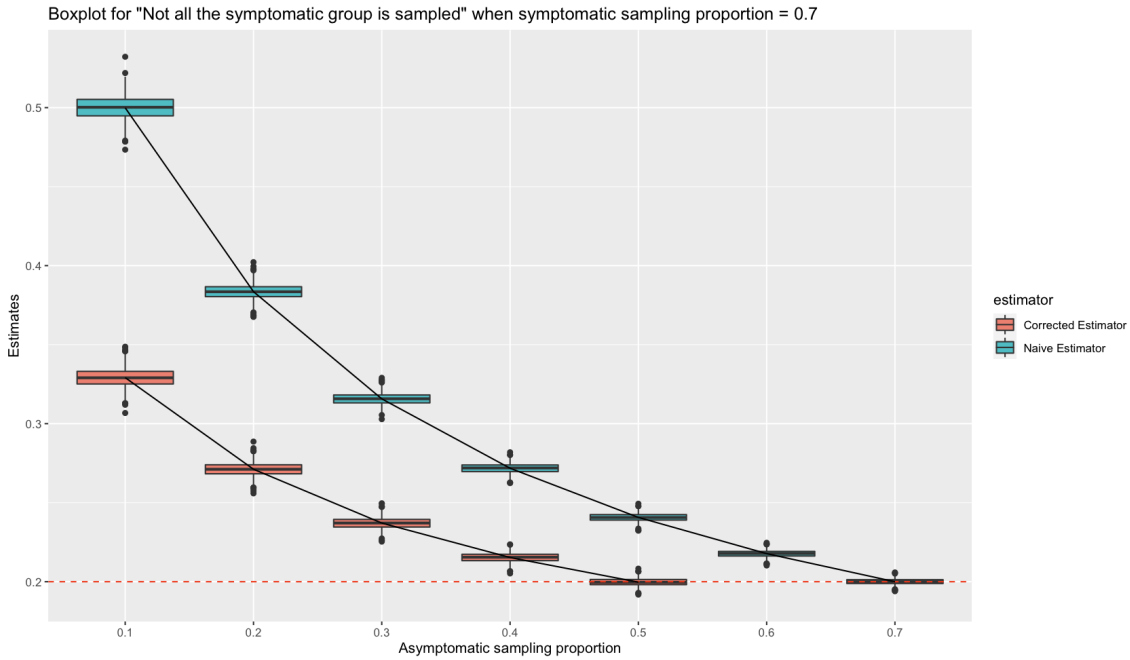


FIGURE D2 Not all the symptomatic group is sampled

E CORRECTION OF TESTING ERRORS

In the previous section, we explored the simulation of the correction of sampling error for a population without considering testing errors. In this section, we extend our analysis including testing error for asymptomatic and symptomatic individuals separately. Specifically, we will model the false positive and false negative rates for both groups in testing using normal distributions. The false positive rate for asymptomatic individuals $\check{\alpha}_0$ is assumed to follow a normal distribution with mean 0.01 and variance 0.0001, while the false negative rate for asymptomatic individuals $\check{\beta}_0$ follows a normal distribution with mean 0.2 and variance 0.0001. Similarly, the false positive rate for symptomatic individuals $\check{\alpha}_1$ is assumed to follow a normal distribution with mean 0.03 and variance 0.0001, while the false negative rate for symptomatic individuals $\check{\beta}_1$ follows a normal distribution with mean 0.02 and variance 0.0001. The real value of the parameters is assumed to be the mean of these distributions. We will consider two scenarios: one in which all symptomatic individuals are sampled, and another in which not all symptomatic individuals are sampled.

E.1 All symptomatic group is sampled

In this simulation, we assume that all symptomatic individuals are sampled for the testing group. We will adjust the proportion of asymptomatic individuals getting sampled from 0.1 to 0.9 to observe the effect of testing error correction on prevalence estimation. Based on the description in the previous section, we use the $\hat{\alpha}_0$, $\hat{\alpha}_1$, $\hat{\beta}_0$ and $\hat{\beta}_1$ obtained from other study as parameters for testing error correction. Here, we assume that $\hat{\alpha}_0$, $\hat{\alpha}_1$, $\hat{\beta}_0$, $\hat{\beta}_1$ follow a uniform distribution with a mean of the true values $\alpha_0 = 0.01$, $\alpha_1 = 0.03$, $\beta_0 = 0.2$ and $\beta_1 = 0.02$ in the simulation study.

From Figure E3, we expect to see that our corrected estimators are very close to the true value, while the naive estimator is approaching the true value as the proportion of asymptomatic individuals increases. Due to the additional variability introduced by testing error, we observe a larger variance of the corrected estimators compared to Algorithm 1.

E.2 Not all the symptomatic group is sampled

We also simulated an scenario where not all symptomatic individuals in the population were sampled for testing, accounting for testing error. Specifically, we assumed that 70% of symptomatic individuals in the population would go for a test.

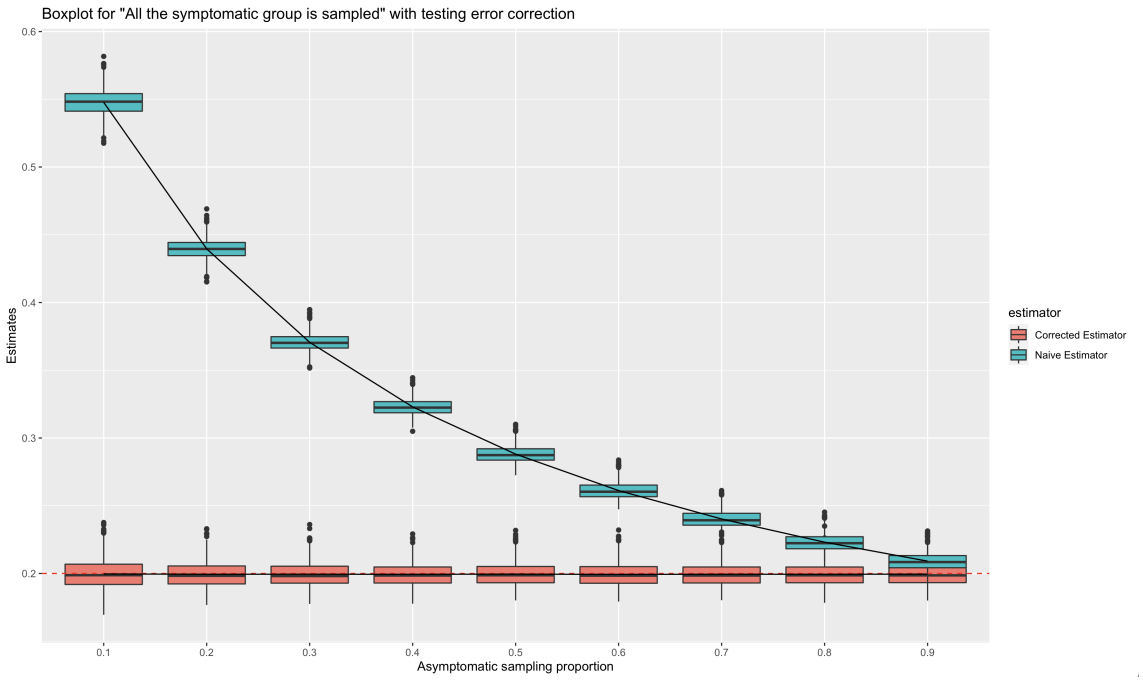
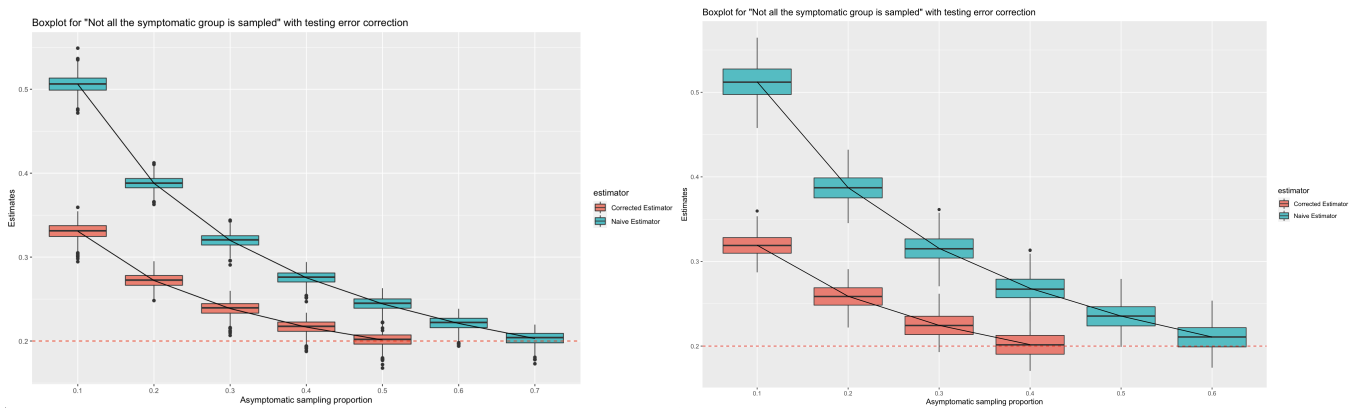


FIGURE E3 All the symptomatic group is sampled with testing error



(a) Unbiased correction parameters

(b) Biased correction parameters

FIGURE E4 Not all the symptomatic group is sampled with testing error

Here we consider two parameter settings for the correction parameters obtained from other studies: the first one is unbiased, where $\hat{\alpha}_0$, $\hat{\alpha}_1$, $\hat{\beta}_0$ and $\hat{\beta}_1$ follow a uniform distribution with mean 0.01,0.03,0.2 and 0.02 (the true values), respectively; the second one is biased, where $\hat{\alpha}_0$, $\hat{\alpha}_1$, $\hat{\beta}_0$ and $\hat{\beta}_1$ follow a uniform distribution with mean 0.05, 0.1, 0.3, and 0.1, respectively. From the results in Figures E4a and E4b, it can be observed that the estimates from both settings converge to the true values as the proportion of asymptomatic increases in the sample, but the corrected estimate from the second setting has a larger variance.

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. Alleva G, Arbia G, Falorsi PD, Zuliani A. A sample approach to the estimation of the critical parameters of the SARS-CoV-2 epidemics: an operational design with a focus on the Italian health system.. tech. rep., University of Sapienza; Rome: 2020.

3. 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

4. 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

5. 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

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. Kundu R, Shi X, Morrison J, Mukherjee B. A Framework for Understanding Selection Bias in Real-World Healthcare Data. *arXiv* 2023. doi: 10.48550/arXiv.2304.04652

8. Cortes C, Mohri M, Riley M, Rostamizadeh A. Sample Selection Bias Correction Theory. In: Heidelberg: Springer. 2008 (pp. 38-53).

9. Liu AC, Scholtus S, Waal TD. Correcting Selection Bias in Big Data by Pseudo-Weighting. *Journal of Survey Statistics and Methodology* 2022; smac029. doi: 10.1093/jssam/smac029

10. 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

11. Barbier J. High-dimensional inference: a statistical mechanics perspective. *Ithaca: Viaggio nella Scienza* 2020; XVI(99-137).

12. Hössjer O, Díaz-Pachón DA, Rao JS. A Formal Framework for Knowledge Acquisition: Going beyond Machine Learning. *Entropy* 2022; 24(10): 1469. doi: 10.3390/e24101469

13. Shannon CE. A Mathematical Theory of Communication. *Bell System Technical Journal* 1948; 27(3): 379-423.

14. Díaz-Pachón DA, Hössjer O, Marks II RJ. Is Cosmological Tuning Fine or Coarse?. *Journal of Cosmology and Astroparticle Physics* 2021; 2021(07): 020. doi: 10.1088/1475-7516/2021/07/020

15. Díaz-Pachón DA, Hössjer O, Marks II RJ. Sometimes size does not matter. *Foundations of Physics* 2023; 53(1). doi: 10.1007/s10701-022-00650-1

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

17. 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

18. 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

19. Emergencies fSAG. GOS: Impact of false positives and negatives, 3 June 2020. tech. rep., Government Office for Science; London: 2020.

20. Arevalo-Rodríguez I, Buitrago-García D, Simancas-Racines D, et al. False-negative results of initial RT-PCR assays for COVID-19: A systematic review. *PLoS ONE* 2020; 15(12): e0242958. doi: 10.1371/journal.pone.0242958

21. Haug S, Marks II RJ, Dembski WA. Exponential Contingency Explosion: Implications for Artificial General Intelligence. *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 2022; 52(5): 2800–2808. doi: 10.1109/TSMC.2021.3056669

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

23. 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

24. 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

25. 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.

26. 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

27. Hössjer O, Díaz-Pachón DA, Chen Z, Rao JS. An Information Theoretic Approach to Prevalence Estimation and Missing Data. *arXiv* 2022. doi: 10.48550/arXiv.2206.05120

28. 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

29. 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

30. 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

31. 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

32. 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

33. Montañez GD. A Unified Model of Complex Specified Information. *BIO-Complexity* 2018; 2018(4): 1-26. doi: 10.5048/BIO-C.2018.4

34. 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

35. 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

36. 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* 2023; 45(4): 4637-4649. doi: 10.1109/TPAMI.2022.3195462

37. 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

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

39. 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

40. Díaz-Pachón DA, Hössjer O. Assessing, testing and estimating the amount of fine-tuning by means of active information. *Entropy* 2022; 24(10): 1323. doi: 10.3390/e24101323