On the true number of COVID-19 infections: Effect of the Sensitivity, Specificity and Number of Tests

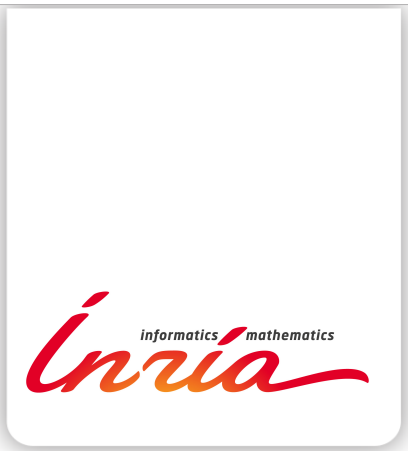
Eitan Altman, Izza Mounir, Fatim-Zahra Najid, Samir Perlaza

On the true number of COVID-19 infections: Effect of the Sensitivity, Specificity and Number of Tests

Eitan Altman, Izza Mounir, Fatim-Zahra Najid, and Samir M. Perlaza

RESEARCH REPORT
N° 9344
May 2020
Project-Team NEO

ISSN 0249-6399 ISRN INRIA/RR--9344--FR+ENG



On the true number of COVID-19 infections: Effect of the Sensitivity, Specificity and Number of Tests

Eitan Altman, Izza Mounir, Fatim-Zahra Najid, and Samir M. Perlaza

Project-Team NEO

Research Report n° 9344 — version 1 — initial version May 2020 — revised version June 2020 — 27 pages

---

Eitan Altman and Samir M. Perlaza are with INRIA, Centre de Recherche de Sophia Antipolis - Méditerranée, 2004 Route des Lucioles, 06902 Sophia Antipolis CEDEX, France. ({eitan.altman, samir.perlaza}@inria.fr)

Izza Mounir is with the Centre Hospitalier Universitaire de Nice - 30 Voie Romaine, 06000 Nice, France. (mounir.i@chu-nice.fr)

Fatim-Zahra Najid is with the Centre Hospitalier Universitaire d'Amiens - 1 Rue du Professeur Christian Cabrol, 80054 Amiens, France (najid.fatim-zahra@chu-amiens.fr)

Samir M. Perlaza is also with the Electrical Engineering Department, Princeton University, Princeton, NJ 08544, USA.

RESEARCH CENTRE
SOPHIA ANTIPOLIS – MÉDITERRANÉE

2004 route des Lucioles - BP 93
06902 Sophia Antipolis Cedex

Abstract: In this report, a formula for estimating the number of SARS-CoV-2 infections in a population is given. The formula is in terms of the test results and test parameters, i.e., probability of true positives (sensitivity) and the probability of true negatives (specificity). The motivation of this work lies on the fact that, depending on the sensitivity and specificity of the tests, the number of positive results might be drastically different to the number of infected individuals. From this perspective, counting the positive results is not a reliable information source for decision-making or policy-making. The relevance of the estimation presented in this work is that the precision increases with the number of tests. That is, large testing campaigns lead to a reliable estimation of the fraction of infected individuals in a population. Two conclusions are drawn from this work. First, in order to ensure that a reliable estimation is achieved, testing campaigns must be implemented with tests for which the sum of the sensitivity and the specificity is sufficiently different from one. Second, the key parameter for reducing the estimation error is the number of tests. For large number of tests, as long as the sum of the sensitivity and specificity do not add up to one, the exact values of these parameters have very little impact in the estimation error.

Key-words: Covid-19, SARS-CoV-2, Sensitivity, Specificity, PCR, Virological Test, Serological Test, Number of Infections, Estimation, False Negative, False Positive, Data Analysis, Policy-Making, Testing Campaigns.

Résumé : Ce rapport présente une formule mathématique pour estimer le nombre d'infections SARS-CoV-2 dans une population donnée. La formule utilise les résultats et les paramètres des tests, c'est-à-dire la probabilité de vrais positifs (sensibilité) et de vrais négatifs (spécificité). Selon la sensibilité et la spécificité des tests, le nombre de résultats positifs peut être radicalement différent du nombre d'individus infectés. Ainsi, le nombre final de résultats rendus positifs n'est pas une source d'information fiable pour la prise de décision ou l'élaboration des directives. Deux conclusions sont tirées de ce travail; afin de garantir l'obtention d'une estimation fiable, des campagnes de tests doivent être mises en oeuvre avec des tests pour lesquels la somme de la sensibilité et de la spécificité est significativement différente de un. De plus, il est prouvé qu'un nombre important de tests conduit à une estimation plus précise du nombre d'infectés. Pour un grand nombre de tests, tant que la somme de la sensibilité et de la spécificité n'est pas égale à un, les valeurs exactes de ces paramètres ont très peu d'impact sur l'erreur d'estimation.

Mots-clés : Covid-19, SARS-CoV-2, sensibilité, spécificité, PCR, test virologique, test sérologique, nombre d'infections, estimation, faux négatifs, faux positifs, analyse de données, élaboration de politiques, campagnes de tests.

Contents

1 Introduction

In the absence of a vaccination or effective medical treatment against the SARS-CoV-2, the global population must cohabitate with the virus. For succeeding in this task, different strategies to slow down the outbreak can be implemented, e.g., encouraging social distancing, isolation of infected individuals, mobility restrictions, lockdowns, contact tracing, *etc.* The main objective is to guarantee that the number of infected individuals that develop critical forms of symptoms does not exceed the capacity of the health care system. Nonetheless, strategies to slow down the outbreak induce dramatic economical consequences, and thus, policy-making requires precise knowledge of the variables describing the state of the pandemic. Among these variables are the number of infected individuals who are capable of transmitting the virus to others, and the number of individuals who have been infected and have developed an immunity agains the virus. Reliable estimations of these numbers can be achieved in part by continuously testing the population. Nonetheless, diagnosing SARS-CoV-2 is a challenging task given the nonnegligible probabilities of false positives and false negatives of the tests, c.f., [1] and [2].

In the general realm of epidemiology, the reliability of tests is measured in terms of two parameters: sensitivity and specificity. The former is the probability with which a test is able to correctly identify the presence of infection. The latter is the probability with which a test is able to correctly identify the absence of infection. Taking into account these two parameters the impact of testing can be analyzed at the individual and society levels.

At the individual level, testing with low quality tests does not bring enough information for decision-making. A test whose sensitivity is 75% would induce one out of four infected individuals to avoid self-quarenting and continue spreading the virus. A test whose specificity is 25% would induce three noninfected individuals out of four to apply a quarantine protocol without actual need. This is one of the reasons why countries have opted for testing only individuals exhibiting symptoms that are compatible with those produced by the SARS-CoV-2 [3]. At the society level, testing against SARS-CoV-2 brings valuable information. For instance, when a large number of tests is available and the parameters of the tests are known, the fraction of infected individuals among those who have been tested can be estimated with surprising precision, c.f., Theorem 1 in Section 4. This certainly eases policy-making at the regional or country level.

With the available knowledge about the SARS-CoV-2, tests are able to determine whether an individual is infected or has been infected by SARS-CoV-2. This essentially allows dividing the population into two groups: susceptible and infected individuals. Today, none of the tests allows determining whether an individual has developed immunity against the virus or whether it is capable of transmitting the virus, c.f., [3] and [4].

The main objective of this work is to present guidelines for calculating the number of infected individuals in a population based on testing results. More specifically, using the number of positive and negative tests, the true number of individuals infected by SARS-CoV-2 is estimated with an estimation error that decreases with the number of individuals that undergo the test. The motivation of this research is that the number of positive tests is not a reliable approximation to the number of infected individuals. This is essentially because infected individuals might obtain negative results (false negatives); and noninfected individuals might obtain positive results (false positivies). The underlying assumptions of this work are: (a) Tests exclusively distinguish between infected or susceptible individuals; (b) the result obtained by an individual is independent of the results obtained by others; and (c) a large number of individuals is tested using tests with identical parameters. Under these assumptions, the main conclusions of this work are:
(i) The number of positive tests might be drastically different to the number of infected individuals in a population depending on the sensitivity and specificity of the tests. Hence, the number of positive tests should not be used for decision-making or policy-making;

(*ii*) Theorem 1 in Section 4 presents an estimation of the number of infected individuals with an estimation error that decreases with the number of tests. That is, it is an asymptotically optimal estimation;
(*iii*) Testing campaigns using tests for which the sum of the sensitivity and specificity is different from one, always allow a reliable estimation of the number of infected individuals (Lemma 1 in Section 4);
(*iv*) Testing campaigns using a test for which the sum of the sensitivity and the specificity is equal to one, lead to data from which it is impossible to estimate the number of infected individuals (Lemma 2); and
(*v*) When the objective is to estimate the number of SARS-CoV-2 infections in a population, the key parameter for reducing the estimation error is the number of tests. As long as the sum of the sensitivity and specificity do not add up to one, the exact values of both sensitivity and specificity have very little impact in the estimation error.

The remaining sections of this report are organized as follows: Section 2 presents a brief overview of the tests for diagnosing SARS-CoV-2; Section 3 formulates the problem of estimating the number of infected individuals of a population; Section 4 presents the estimator of the fraction of infected individuals of a population and the proofs of the main results; Section 5 introduces some examples in which the impact of the sensitivity, specificity and number of tests on the estimation error is numerically analyzed; Section 6 concludes this work; and Section 7 establishes further research directions.

2 Diagnosing SARS-CoV-2

Tests for SARS-CoV-2 can be broadly divided into three groups: virological tests, serological tests, and tests based on medical imaging. Each of these groups provide information about different aspects of the infection and exhibit different reliability parameters.

2.1 Virological Tests

Virological tests inform about the presence of the SARS-CoV-2 virus genome in nasopharyngeal (nasal swab) or oropharyngeal swabs (oral swab), blood, anal swab, urine, stool, and sputum samples [5]. Individuals with positive virological tests are declared capable of contaminating others, and thus, virological tests are central in decision-making and policy-making, c.f. [1] and [2].

The reliability of virological tests in terms of sensitivity and specificity depends on a variety of parameters. These parameters include the type of clinical specimen, the materials and methods used for obtaining the specimens, specimen transportation, viral density of patients, and human errors in data processing in laboratories. In the case of respiratory specimens, viral density appears to play a central role in the sensitivity and specificity of virological tests, c.f., [6] and [7]. This stems from the fact that during the first week after infection, the virus can be detected by nasopharyngeal or oropharyngeal swabs. During the second week and later, the virus might disappear in the upper parts of the respiratory system and migrate to the bronchial tube and the lungs. From the studies in [6] and [7], it appears that specimens from the lower respiratory track increase the sensitivity and specificity of virological tests.

Virological tests are based on several techniques: (*a*) Reverse transcription polymerase chain reaction (RT-PCR), c.f., [8] and [9]; and (*b*) Reverse transcription loop-mediated isothermal amplification (RT-LAMP), c.f., [10] and [11]; and (*c*) other techniques, c.f., [9, 12] and [13].

2.2 Serological Tests

Serological tests determine whether an individual has developed anti-bodies or antigens against the SARS-CoV-2 virus. Nonetheless, an individual produces anti-bodies against SARS-CoV-2 only several days after contracting the infection. Typically, the time between infection and the production of anti-bodies ranges from seven to fourteen days, c.f., [14, 15] and [16]. Serological tests are based on the enzyme linked immunosorbent assay (ELISA) and exhibit high specificity and sensitivity, after fourteen days of infections [14]. This drastically limits the use of serological tests in the early detection of the infection and policy-making, c.f., [1, 3] and [4]. In a nutshell, on one hand, a serological test answers the question whether an individual is or has been infected. On the other hand, serological tests do not allow determining whether an individual has immunity to the SARS-CoV-2 virus or whether the individual is currently spreading the virus. Up to the day of publication of this report, serological tests are not considered for massive testing in France, c.f., [3] and [4].

2.3 Medical Imaging

Medical Imaging for detection of SARS-CoV-2 includes chest X-Ray and chest computed tomography (CT) scans, which reveal ground-glass opacities and consolidations in the periphery of the lungs of infected individuals [17]. Nonetheless, the sensitivity and specificity of CT depends on the experience of radiologists to distinguish SARS-CoV-2 pneumonia from non-SARS-CoV-2 pneumonia [18]. In [19], it is reported that the sensitivity of CT is better than the one achieved by RT-PCR tests.

3 Problem Formulation

Consider a population of individuals whose state is either susceptible (S) or infected (I) and assume that n individuals of this population are tested with the same type of test. Let the ground-truth state of such n individuals be represented by the vector $\boldsymbol{x} \triangleq \big(x_1,\, x_2,\, \ldots,\, x_n\big)$. That is, for all $t \in \{1, 2, \ldots, n\}$, it follows that $x_t \in \{I, S\}$ is the true state of the individual t. The result of testing individual t is denoted by $y_t \in \{I, S\}$. Hence, the outcome of a testing campaign over such population is a vector $\boldsymbol{y} \triangleq \big(y_1,\, y_2,\, \ldots,\, y_n\big) \in \{I, S\}^n$. Due to the fact that tests possess strictly positive probabilities of false negatives and false positives, the vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ might be different. That is, some individuals that are infected could have been declared susceptible and *vice versa*.

A central observation in this analysis is that a test for determining whether an individual is contaminated by SARS-CoV-2 can be modelled by a random transformation $P_{Y|X}$ for which the input and output sets are $\{I, S\}$. More specifically, if an individual whose state is $x \in \{I, S\}$ is tested, the result $y \in \mathcal{Y}$ is observed with probability $P_{Y|X}(y|x)$. Figure 1 shows the model of a test with binary inputs.

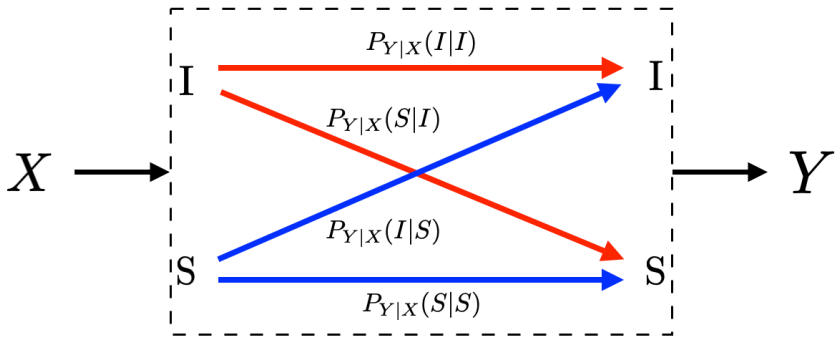


Figure 1: An SARS-CoV-2 test represented by a random transformation from $\{I, S\}$ into $\{I, S\}$ via the conditional probability distribution $P_{Y|X}$.

Using this notation, the sensitivity of the test is $P_{Y|X}(I|I)$; and the specificity of the test is $P_{Y|X}(S|S)$. The probability of false positive is $P_{Y|X}(I|S) = 1 - P_{Y|X}(S|S)$; and the probability of false negative is $P_{Y|X}(S|I) = 1 - P_{Y|X}(I|I)$. This said, a test is fully described by any of the following pairs of parameters:

- The sensitivity and the specificity;
- The sensitivity and the probability of false positive;
- The probability of false negative and the specificity; or
- The probability of false negative and the probability of false positive.

Note that in epidemiology, the parameters of a test are often expressed in percentages, whereas in mathematics, probability measures are expressed using positive reals in the interval [0,1]. In the following, both notations are indistinctly used.

Let X be random variable taking values in $\{I, S\}$ and denote by $P_X : \{I, S\} \to [0, 1]$ its probability distribution such that $P_X(I)$ is the ground-truth fraction of infected individuals in the population. For this reason, the probability distribution P_X is referred to as the *ground-truth input probability distribution*. Let Y be a second random variable taking values in $\{I, S\}$ such that its joint probability distribution with X is P_{XY} and for all $(x, y) \in \{I, S\}^2$,

$$P_{XY}(x, y) = P_X(x)P_{Y|X}(y|x), \tag{1}$$

where the conditional distribution $P_{Y|X}$ is the test. Often, the probability distribution P_Y is referred to as the *ground-truth output probability distribution* and it is obtained as the marginal of P_{XY}. That is, for all $y \in \{I, S\}$,

$$P_Y(y) = \sum_{x \in \{I,S\}} P_X(x)P_{Y|X}(y|x). \tag{2}$$

The problem consists in using the data $\boldsymbol{y}$ obtained through a testing campaign with tests whose parameters are modeled by $P_{Y|X}$ to determine the fraction $P_X(I)$ of infected individuals in the population. More formally, the problem can be stated as follows:

Consider two random variables X and Y with the joint probability distribution P_{XY} in (1). The problem consists in estimating the probability distribution P_X based only on n realizations y_1, y_2, ..., y_n of the random variable Y, with n a finite integer.

This problem is reminiscent to the problem of *population recovery* introduced in [20] and further studied in [21] and [22].

4 Main Results

Given the data $\boldsymbol{y} \in \{I, S\}^n$ collected during a test campaign, the fraction of the population reporting positive and negative tests form an empirical distribution denoted by $\bar{P}_Y^{(n)}$ on the set $\{I, S\}$ such that,

$$\bar{P}_Y^{(n)}(I) = \frac{1}{n}\sum_{t=1}^{n} \mathbb{1}_{\{I=y_t\}}, \text{ and} \tag{3a}$$

$$\bar{P}_Y^{(n)}(S) = \frac{1}{n}\sum_{t=1}^{n} \mathbb{1}_{\{S=y_t\}}, \tag{3b}$$

where $\mathbb{1}_{\{\cdot\}}$ is the indicator function. Essentially, $\bar{P}_Y^{(n)}$ is a counting probability measure. In the following, such probability measure is often referred to as the *output empirical distribution* obtained from the data $\boldsymbol{y}$.

Let $\hat{P}_X^{(n)}$ be a probability distribution on the set $\{I, S\}$ representing the estimation of P_X based on the data $\boldsymbol{y}$. The error induced by estimating P_X using $\hat{P}_X^{(n)}$ can be measured by the total variation between these two probability distributions, which is denoted by $\left\|P_X - \hat{P}_X^{(n)}\right\|_{\text{TV}}$ and satisfies,

$$\left\|P_X - \hat{P}_X^{(n)}\right\|_{\text{TV}} \triangleq \frac{1}{2}\left(\left|P_X(I) - \hat{P}_X^{(n)}(I)\right| + \left|P_X(S) - \hat{P}_X^{(n)}(S)\right|\right). \tag{4}$$

Using this notation, the following theorem presents the main result of this work.

Theorem 1. *Consider a population whose true ratio of infected (I) and susceptible (S) individuals is $P_X(I)$ and $P_X(S) = 1 - P_X(S)$, respectively, with $P_X(I) \in [0, 1]$. Assume that n individuals of such population are tested with a test $P_{Y|X}$ that satisfies*

$$P_{Y|X}(S|S) + P_{Y|X}(I|I) \neq 1, \tag{5}$$

which induces an output empirical probability distribution $\bar{P}_Y^{(n)}$ as in (3). Then, the probability distribution $\hat{P}_X^{(n)}$ on $\{I, S\}$, such that

$$\hat{P}_X^{(n)}(I) = \frac{1 - \bar{P}_Y^{(n)}(I) - P_{Y|X}(S|S)}{1 - P_{Y|X}(I|I) - P_{Y|X}(S|S)}, \text{ and} \tag{6}$$

$$\hat{P}_X^{(n)}(S) = \frac{\bar{P}_Y^{(n)}(I) - P_{Y|X}(I|I)}{1 - P_{Y|X}(I|I) - P_{Y|X}(S|S)}, \tag{7}$$

satisfies

$$\lim_{n\to\infty} \left\|P_X - \hat{P}_X^{(n)}\right\|_{\text{TV}} = 0 \text{ in probability.} \tag{8}$$

The proof of Theorem 1 uses simple arguments from probability theory. Nonetheless, such arguments have profound implications in the correct analysis of data obtained from testing campaigns against SARS-CoV-2. The proof as well as the implications of each step of the proof in the analysis of data obtained by testing campaigns are discussed in the following sections.

In a nutshell, Theorem 1 states that the value $\hat{P}_X^{(n)}(I)$ constitutes an estimation of the ground-truth fraction $P_X(I)$ of infected individuals of the population. Moreover, it shows that such estimation is asymptotically optimal. That is, $\hat{P}_X^{(n)}$ is a reliable estimation of P_X when the number of tests is sufficiently large.

4.1 The Relation between Infections and Positive Tests

The proof of Theorem 1 leverages the following intuition: Under the assumption that $\bar{P}_Y^{(n)}$, which is obtained from the data $\boldsymbol{y}$ as in (3), is a good estimation of the ground-truth output probability distribution P_Y, then, a distribution $\hat{P}_X^{(n)}$ that satisfy

$$\begin{pmatrix} \bar{P}_Y^{(n)}(I) \\ \bar{P}_Y^{(n)}(S) \end{pmatrix} = \begin{pmatrix} P_{Y|X}(I|I) & P_{Y|X}(I|S) \\ P_{Y|X}(S|I) & P_{Y|X}(S|S) \end{pmatrix} \begin{pmatrix} \hat{P}_X^{(n)}(I) \\ \hat{P}_X^{(n)}(S) \end{pmatrix}, \tag{9}$$

is a good estimation of the ground-truth input probability distribution P_X. This intuition builds upon the observation that the output distribution $\bar{P}_Y^{(n)}$ induced by the data, must be the marginal of a joint distribution consisting of the product of the conditional $P_{Y|X}$ and the input distribution. That is, for all $y \in \{I, S\}$,

$$\bar{P}_Y^{(n)}(y) = \sum_{x \in \{I,S\}} P_{Y|X}(y|x)\hat{P}_X^{(n)}(x),$$

which is equivalent to the system in (9). Note that the equality in (9) is a linear system of two equations with two variables, and thus, if it is consistent, it has either a unique solution or infinitely many solutions. The following lemmas summarize the conditions on the parameters of the test that lead to a unique solution to (9).

Lemma 1. *The following five statements are equivalent:*

- *The system of equations in* (9) *has a unique solution;*
- *The sensitivity* $P_{Y|X}(I|I)$ *and specificity* $P_{Y|X}(S|S)$ *satisfy*

$$P_{Y|X}(I|I) + P_{Y|X}(S|S) \neq 1; \tag{10a}$$

- *The sensitivity* $P_{Y|X}(I|I)$ *and the probability of false positive* $P_{Y|X}(I|S)$ *satisfy*

$$P_{Y|X}(I|I) \neq P_{Y|X}(I|S); \text{ and} \tag{10b}$$

- *The probability of false negative* $P_{Y|X}(S|I)$ *and the specificity* $P_{Y|X}(S|S)$ *satisfy*

$$P_{Y|X}(S|S) \neq P_{Y|X}(S|I). \tag{10c}$$

- *The probability of false positive* $P_{Y|X}(I|S)$ *and the probability of false negative* $P_{Y|X}(S|I)$ *satisfy*

$$P_{Y|X}(I|S) + P_{Y|X}(S|I) \neq 1. \tag{10d}$$

Proof. The proof of Lemma 1 follows from the fact that a unique solution to (9) is observed if and only if the determinant of the matrix

$$\begin{pmatrix} P_{Y|X}(I|I) & P_{Y|X}(I|S) \\ P_{Y|X}(S|I) & P_{Y|X}(S|S) \end{pmatrix}$$

is different from zero (Rouché - Fontené theorem [23]). That is,

$$P_{Y|X}(I|I)P_{Y|X}(S|S) - P_{Y|X}(S|I)P_{Y|X}(I|S) \neq 0. \tag{11}$$

The proof is complete by verifying that the expression in (11) is equivalent to those in (10). □

Note that all conditions in (10) are equivalent to each other, and thus, they are equivalent to the condition in (5).

Lemma 2. *The following five statements are equivalent:*

- *The system of equations in* (9) *has infinitely many solutions;*
- *The sensitivity* $P_{Y|X}(I|I)$ *and specificity* $P_{Y|X}(S|S)$ *satisfy*

$$P_{Y|X}(I|I) + P_{Y|X}(S|S) = 1; \tag{12a}$$

- *The sensitivity* $P_{Y|X}(I|I)$ *and the probability of false positive* $P_{Y|X}(I|S)$ *satisfy*

$$P_{Y|X}(I|I) = P_{Y|X}(I|S); \tag{12b}$$

- *The probability of false negative* $P_{Y|X}(S|I)$ *and the specificity* $P_{Y|X}(S|S)$ *satisfy*

$$P_{Y|X}(S|S) = P_{Y|X}(S|I); \text{ and} \tag{12c}$$

- *The probability of false positive* $P_{Y|X}(I|S)$ *and the probability of false negative* $P_{Y|X}(S|I)$ *satisfy*

$$P_{Y|X}(I|S) + P_{Y|X}(S|I) = 1. \tag{12d}$$

Proof. The proof of Lemma 2 follows from the theorem of Rouché and Fontené [23] that states that when the system in (9) is consistent, it has infinitely many solutions if the determinant of the matrix

$$\begin{pmatrix} P_{Y|X}(I|I) & P_{Y|X}(I|S) \\ P_{Y|X}(S|I) & P_{Y|X}(S|S) \end{pmatrix}$$

is not full rank. When such a matrix is not full rank, its determinant is zero. That is,

$$P_{Y|X}(I|I)P_{Y|X}(S|S) - P_{Y|X}(S|I)P_{Y|X}(I|S) = 0. \tag{13}$$

The proof is complete by verifying that the expression in (13) is equivalent to those in (12). □

When a unique solution to (9) exists, a unique estimation $\hat{P}_X^{(n)}$ of P_X that is consistent with the data $\boldsymbol{y}$ exists. Otherwise, nothing meaningful can be said about P_X based on the data. The following theorem summarizes this observation.

Lemma 3. *Consider a test* $P_{Y|X}$ *that satisfies at least one of the conditions in* (12)*. Hence, the following statements are equivalent:*

- *Given the output empirical distribution* $\bar{P}_Y^{(n)}$ *obtained from the data* $\boldsymbol{y}$ *as in* (3a)*, any probability distribution* $\hat{P}_X^{(n)}$ *on* $\{I, S\}$ *satisfies the equality in* (9)*;*
- *Two random variables* X *and* Y *whose joint probability distribution* P_{XY} *satisfies* (1) *have zero mutual information; and*
- *Two random variables* X *and* Y *whose joint probability distribution* P_{XY} *satisfies* (1) *are independent.*

Proof. The first statement is a consequence of Lemma 2; the second statement follows from the fact that under any of the assumptions in (12), the mutual information satisfies

$$\begin{aligned} \mathsf{I}(X;Y) &\triangleq \sum_{x\in\{I,S\}} \sum_{y\in\{I,S\}} P_X(x) P_{Y|X}(y|x) \log_2\left(\frac{P_{Y|X}(y|x)}{P_Y(y)}\right) && (14)\\ &= 0, && (15) \end{aligned}$$

where P_X, $P_{Y|X}$, and P_Y satisfy the equality in (1). The third statement follows from the fact that two random variables are independent if and only if their mutual information is zero. □

Note that Lemma 3 justifies the conclusion of Lemma 2. That is, when at least one of the conditions in (12) holds, the ground-truth output probability distribution P_Y does not contain any information about the ground-truth input probability distribution P_X. That is, any estimation of P_X based on the available data $\boldsymbol{y}$ satisfies (9), and thus, nothing can be said about P_X based on the data $\boldsymbol{y}$.

The following section describes the unique solution to (9), when it exists.

4.2 Estimation of the Number of Infections

The following lemma introduces explicit expressions for the estimation $\hat{P}_X^{(n)}$ of P_X based on the data $\boldsymbol{y}$ under the assumption that the test verifies at least one of the conditions in (10).

Lemma 4. *Consider a test* $P_{Y|X}$ *that satisfies at least one of the conditions in* (10)*. Then, given the empirical output distribution* $\bar{P}_Y^{(n)}$ *obtained from the data* $\boldsymbol{y} \triangleq \big(y_1, y_2, \ldots, y_n\big)$ *as in* (3)*, the unique input distribution* $\hat{P}_X^{(n)}$ *that satisfies* (9) *is:*

$$\hat{P}_X^{(n)}(I) = \frac{1 - \bar{P}_Y^{(n)}(I) - P_{Y|X}(S|S)}{1 - P_{Y|X}(I|I) - P_{Y|X}(S|S)}, \text{ and} \quad (16a)$$

$$\hat{P}_X^{(n)}(S) = \frac{\bar{P}_Y^{(n)}(I) - P_{Y|X}(I|I)}{1 - P_{Y|X}(I|I) - P_{Y|X}(S|S)}. \quad (16b)$$

Proof. The proof of Lemma 4 follows from solving the system of equations in (9). □

The formulas in (17) are given in terms of the sensitivity $P_{Y|X}(I|I)$ and specificity $P_{Y|X}(S|S)$ of the test. Nonetheless, it can be expressed in terms of the probabilities of false positive and false negative, or any combination of the parameters describing the test. The following corollary shows the formulas in terms of the probabilities of false positive and false negative.

Corollary 1. *Consider a test $P_{Y|X}$ that satisfies at least one of the conditions in* (10)*. Then, given the empirical output distribution $\bar{P}_Y^{(n)}$ obtained from the data $\boldsymbol{y} \triangleq \left(y_1,\, y_2,\, \ldots,\, y_n\right)$ as in* (3)*, the unique input distribution $\hat{P}_X^{(n)}$ that satisfies* (9) *is:*

$$\hat{P}_X^{(n)}(I) = \frac{\bar{P}_Y^{(n)}(I) - P_{Y|X}(I|S)}{1 - P_{Y|X}(S|I) - P_{Y|X}(I|S)},\ \textit{and} \tag{17a}$$

$$\hat{P}_X^{(n)}(I) = \frac{1 - P_{Y|X}(S|I) - \bar{P}_Y^{(n)}(I)}{1 - P_{Y|X}(S|I) - P_{Y|X}(I|S)}. \tag{17b}$$

The formulas obtained in Lemma 4 allow studying the impact of each of the test parameters on the number of positive and negative results.

4.3 Impact of the Parameters of the Test

From Lemma (4), it holds that the fraction of individuals reporting positive tests $\bar{P}_Y^{(n)}(I)$ satisfies:

$$\bar{P}_Y^{(n)}(I) = 1 - P_{Y|X}(S|S)\left(1 - \hat{P}_X^{(n)}(I)\right) - \left(1 - \hat{P}_{Y|X}(I|I)\right)\hat{P}_X^{(n)}(I). \tag{18}$$

The following lemma determines the influence of the parameters of the test on $\bar{P}_Y^{(n)}(I)$.

Lemma 5. *Consider a test $P_{Y|X}$ that satisfies at least one of the conditions in* (10)*. Then, for a population whose fraction of infected individuals P_X is fixed, the following statements are true when n individuals of the population are tested using the test $P_{Y|X}$:*

- *The fraction $\bar{P}_Y^{(n)}(I)$ of positive tests linearly decreases with the specificity of the test $P_{Y|X}(S|S)$;*
- *The fraction $\bar{P}_Y^{(n)}(I)$ of positive tests linearly increases with the probability of false positive of the test $P_{Y|X}(I|S)$;*
- *The fraction $\bar{P}_Y^{(n)}(I)$ of positive tests linearly increases with the sensitivity of the test $P_{Y|X}(I|I)$; and*
- *The fraction $\bar{P}_Y^{(n)}(I)$ of positive tests linearly decreases with the probability of false negative of the test $P_{Y|X}(S|I)$.*

Proof. The proof of Lemma 5 consists in verifying that the derivative of $\bar{P}_Y^{(n)}$ in (18) with respect to $P_{Y|X}(S|S)$ is negative; with respect to $P_{Y|X}(I|S)$ is positive; with respect to $P_{Y|X}(I|I)$ is positive; and with respect to $P_{Y|X}(S|I)$ is negative. □

The importance of Lemma 5 is that it shows that tests might lead to optimistic or pessimistic estimations of the ground-truth fraction of infected individuals $P_X(I)$. Consider the case of low specificity and high sensitivity tests. In such a case, most of infected individuals observe positive results (true positives) but also many individual who are not infected observe positive testing results (false positives). Alternatively, in the case of high specificity and low sensitivity, most of susceptible individuals observe negative testing results (true negatives) but also many infected individuals observe negative testing results (false negatives).

4.4 Estimation Error and Large Number of Tests

The error induced by estimating P_X by using $\hat{P}_X^{(n)}$, which is based on the data $\boldsymbol{y}$, can be quantified by any distance, e.g., total variation, or pseudo-distance, e.g., Kullback-Divergence, between the probability distributions P_X and $\hat{P}_X^{(n)}$. The following lemma shows that the total variation between P_X and $\hat{P}_X^{(n)}$, denoted by $\left\|P_X - \hat{P}_X^{(n)}\right\|_{\text{TV}}$ is equivalent to the total variation between P_Y and $\bar{P}_Y^{(n)}$, denoted by $\left\|P_Y - \bar{P}_Y^{(n)}\right\|_{\text{TV}}$, up to a constant factor.

Lemma 6. *Consider a test $P_{Y|X}$ that satisfies at least one of the conditions in* (10)*. Then, for a population whose fraction of infected individuals P_X is fixed, the estimation $\hat{P}_X^{(n)}$ in* (17) *satisfies*

$$\left\|P_X - \hat{P}_X^{(n)}\right\|_{\text{TV}} = \frac{1}{\left|1 - P_{Y|X}(I|I) - P_{Y|X}(S|S)\right|} \left\|P_Y - \bar{P}_Y^{(n)}\right\|_{\text{TV}}, \tag{19}$$

where P_Y is the output probability distribution in (2)*.*

Proof. The proof of Lemma 6 follows from the definition of total variation in (4) and from equalities in (17). □

Lemma 6 reinforces some of the conclusions drawn form previous lemmas. Note for instance that:
(a) When the sum of the sensitivity $P_{Y|X}(I|I)$ and the specificity $P_{Y|X}(S|S)$ is equal to one, the total variation between P_X and $\hat{P}_X^{(n)}$ exhibits a singularity;
(b) When the sensitivity and the specificity are both equal to one, i.e., $P_{Y|X}(I|I) = P_{Y|X}(S|S) = 1$, the total variation between P_X and $\hat{P}_X^{(n)}$ is identical to the total variation between P_Y and $\bar{P}_Y^{(n)}$;
(c) The smaller the total variation $\left\|P_Y - \bar{P}_Y^{(n)}\right\|_{\text{TV}}$, the smaller the total variation $\left\|P_X - \hat{P}_X^{(n)}\right\|_{\text{TV}}$.

The observation (a) reinforces the main conclusion of Lemma 2. The observations (b) and (c) are reminiscent to the intuition discussed in Section 4.1 to establish the equality in (9). Therein, it was argued that if $\bar{P}_Y^{(n)}$ is sufficiently close to P_Y, then $\hat{P}_X^{(n)}$ must be sufficiently close to P_X, which is now formally proved. The following lemma leverages these observations to allow concluding on the ground-truth fraction of infected individuals of the whole population.

Lemma 7. *Consider a test $P_{Y|X}$ that satisfies at least one of the conditions in* (10)*. Then the empirical output distribution satisfies*

$$\lim_{n\to\infty} \left\|P_Y - \bar{P}_Y^{(n)}\right\|_{\text{TV}} = 0 \text{ in probability,} \tag{20}$$

where P_Y is the ground-truth output probability distribution in (2)*.*

Proof. The proof of Lemma 7 is a consequence of the Theorem of Glivenko and Cantelli [24] □

Finally, from Lemma 6 and Lemma 7, it holds that by increasing the number of tests, the error of approximating P_X by $\hat{P}_X^{(n)}$ in (17) can be made arbitrarily small. The following lemma leverages this observation.

Lemma 8. *Consider a test $P_{Y|X}$ that satisfies at least one of the conditions in* (10)*. Then, the ground-truth input distribution P_X and its estimation $\hat{P}_X^{(n)}$ in* (17) *satisfy*

$$\lim_{n\to\infty} \left\|P_X - \hat{P}_X^{(n)}\right\|_{\text{TV}} = 0 \text{ in probability.} \tag{21}$$

Proof. The proof of Lemma 8 is an immediate consequence of both Lemma 6 and Lemma 7. □

In the following section, some of the conclusions drawn from the lemmas above are used to provide some guidelines for conveniently choosing the tests in terms of their parameters.

5 Examples and Remarks

This section highlights some of the conclusions drawn from Lemma 1 - Lemma 8 using a numerical analysis in particular examples. In each example, the data is artificially generated. That is, for a given fixed ground-truth input probability distribution P_X on $\{I, X\}$, an n-dimensional vector $\boldsymbol{x} = \big(x_1, x_2, \ldots, x_n\big) \in \{I, S\}^n$ is generated such that for all $t \in \{1, 2, \ldots, n\}$, x_t is a realization of a random variable $X \sim P_X$ and represents the state of individual t. Given a test $P_{Y|X}$ an n-dimensional vector $\boldsymbol{y} = \big(y_1, y_2, \ldots, y_n\big) \in \{I, S\}^n$ is generated such that for all $t \in \{1, 2, \ldots, n\}$, y_t is the realization of a random variable $Y \sim P_{Y|X=x_t}$ and represents the result of the test of individual t. The analysis is based on the data collected by the testing campaigns, that is, the vector $\boldsymbol{y}$.

Example 1. *Consider a population for which the fraction of infected individuals is* $P_X(I) = 0,4$. *Assume that* $n = 10\,000$ *individuals of such population are diagnosed using a test* $P_{Y|X}$.

Example 2. *Consider a population for which the fraction of infected individuals is* $P_X(I) = 0,4$. *Assume that* $n = 100\,000$ *individuals of such population are diagnosed using a test* $P_{Y|X}$.

Example 3. *Consider a population for which the fraction of infected individuals is* $P_X(I) = 0,4$. *Assume that* $n = 100\,000\,000$ *individuals of such population are diagnosed using a test* $P_{Y|X}$.

In Figure 2, Figure 4, and Figure 6, the ground-truth fraction $P_X(I)$ of infected individuals in the population is plotted with a straight black line; the estimation $\hat{P}_X^{(n)}$ of P_X is plotted with red circles; the fraction of positive tests $\bar{P}_Y^{(n)}(I)$ is plotted with blue diamonds; and the value of $P_Y(I)$ in (18) is plotted with a dashed red line as a function of the specificity $P_{Y|X}(S|S)$ for a fixed sensitivity.

In Figure 3, Figure 5, and Figure 7, the ground-truth fraction $P_X(I)$ of infected individuals in the population is plotted with a straight black line; the estimation $\hat{P}_X^{(n)}$ of P_X is plotted with red circles; the fraction of positive tests $\bar{P}_Y^{(n)}(I)$ is plotted with blue diamonds; and the value of $P_Y(I)$ in (18) is plotted with a dashed red line as a function of the sensitivity $P_{Y|X}(I|I)$ for a fixed specificity.

5.1 The Correct Use of Data for Policy-Making Against SARS-CoV-2

One of the main observations to be highlighted from this numerical analysis is that there exists an important difference between the number of positive tests in a population and the number of infected individuals. Figure 2 - Figure 7 show the ratio between the number of positive tests and the total number of tests (fraction of positive tests), i.e., $\bar{P}_Y^{(n)}$; and the fraction of infected individuals in the population, i.e., P_X. Therein, the gaps between both $\bar{P}_Y^{(n)}$ and P_X in Example 1 - Example 3 are evident. This reinforces the conclusion that number of positive tests should not be used for policy-making and decision-making.

5.2 Estimation of the Ground-Truth

Figure 2 - Figure 7 show that when the parameters of the test satisfy at least one of the conditions in (10), it is always possible to obtain an estimation $\hat{P}_X^{(n)}(I)$ of the ground-truth fraction $P_X(I)$ of infected individuals. This is independently of the exact values of the parameters. More importantly, the reliability of such estimation increases with the number of test results, c.f., Example 2 and Example 3.

The implications of this observation are very important in practical terms. This shows that if the objective of a testing campaign against SARS-CoV-2 is to determine the fraction of infected individuals, the key parameter is the number of tests. This is essentially because testing with low quality tests (low sensitivity and low specificity) or high quality tests (high sensitivity and high specificity) leads to identical results in terms of the estimation, when a large number of tests is performed. The condition that is essential for the estimation is that the sum of the sensitivity and specificity of the test is different from one, c.f., Lemma 1. Note for instance the singularities shown in Figure 2 - Figure 7 in blue dot-dash lines.

5.3 Relevance of the Sensitivity and Specificity

In Figure 2, Figure 4, and Figure 6, it is shown that the number of positive tests increases with the sensitivity, c.f., Lemma 5. This is a consequence of the fact that higher sensitivity leads to more infected individuals observing positive testing results. Alternatively, in Figure 3, Figure 5, and Figure 7, it is shown that the number of positive tests decreases with the specificity, c.f., Lemma 5. This is a consequence of the fact that higher specificity reduces the number of noninfected individuals observing positive testing results.

From this perspective, tests might lead to estimations in which the fraction of individuals reporting positive testing results $\bar{P}_Y^{(n)}(I)$ is bigger than the ground-truth fraction of infected individuals $P_X(I)$, i.e., $\bar{P}_Y^{(n)}(I) > P_X(I)$. In this case, it is said that the test is pessimistic. This is the case of low specificity and high sensitivity tests. Alternatively, tests might lead to estimations in which the fraction of individuals reporting positive testing results $\bar{P}_Y^{(n)}(I)$ is smaller than the ground-truth fraction of infected individuals $P_X(I)$, i.e., $\bar{P}_Y^{(n)}(I) < P_X(I)$. In this case, it is said that the test is optimistic. This is the case of high specificity and low sensitivity.

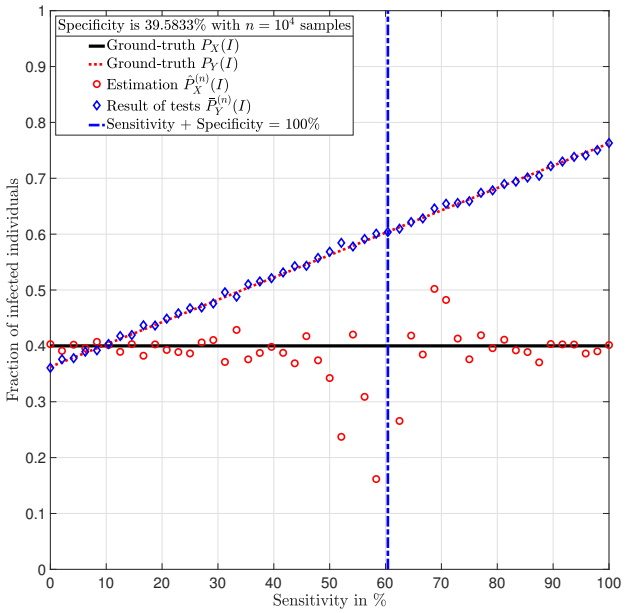
This analysis highlights the relevance of using the estimation $\hat{P}_X^{(n)}$ of P_X for decision and policy making rather than $\bar{P}_Y^{(n)}$, which includes false positives and false negatives.

5.4 Tests Whose Results are Useless

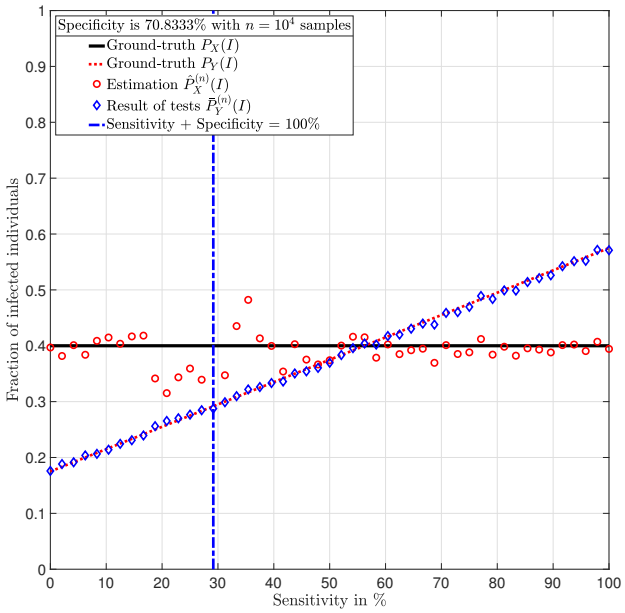
In Figure 2, Figure 4, and Figure 6, the value of the sensitivity $P_{Y|X}(I|I)$ that satisfies $P_{Y|X}(I|I)+P_{Y|X}(S|S)=1$ is plotted with a blue dash-dot line. Alternatively, in Figure 3, Figure 5, and Figure 7, the value of the specificity $P_{Y|X}(S|S)$ that satisfies $P_{Y|X}(I|I)+P_{Y|X}(S|S)=1$ is plotted with a blue dash-dot line. Note that for these specific values of sensitivity and specificity, the estimation $\hat{P}_X^{(n)}(I)$ of $P_X(I)$ is not plotted, as infinitely many estimations exist, c.f., Lemma 2 and Lemma 3. More importantly, note that for some $\epsilon > 0$, the difference $\left|P_X(I) - \hat{P}_X^{(n)}(I)\right|$ is bigger when the sensibility and specificity satisfy $\left|1 - P_{Y|X}(S|S) - P_{Y|X}(I|I)\right| < \epsilon$ than when these parameters satisfy $\left|1 - P_{Y|X}(S|S) - P_{Y|X}(I|I)\right| > \epsilon$. These observation is justified by the fact that the total variation $\left\|P_X - \hat{P}_X^{(n)}\right\|_{\text{TV}}$ in (19) exhibits a singularity when $P_{Y|X}(S|S)+P_{Y|X}(I|I)=1$, c.f., Lemma 6, and thus, larger errors are expected around the singularity for finite numbers of tests n.

5.5 Impact of the Number of Tests.

In Example 1, i.e., Figure 2, and Figure 3, the estimations $\hat{P}_X^{(n)}$ of P_X appear more disperse than the estimations in Example 3, i.e., Figure 6 and Figure 7. This is even more evident for the estimations around the singularity plotted with a blue dot-dash line. This observation is in line with the conclusion of Lemma 8. That is, the total variation $\left\|P_X - \hat{P}_X^{(n)}\right\|_{\mathrm{TV}}$, which quantifies the error on the estimation of P_X by $\hat{P}_X^{(n)}$, reduces with the number of tests. This observation is of paramount importance as it shows that as long as the sum of the sensitivity and specificity do not add up to one, the exact values of both sensitivity and specificity have very little impact in the estimation when the number of tests is sufficiently large. The the key parameter for reducing the estimation error is the number of tests.

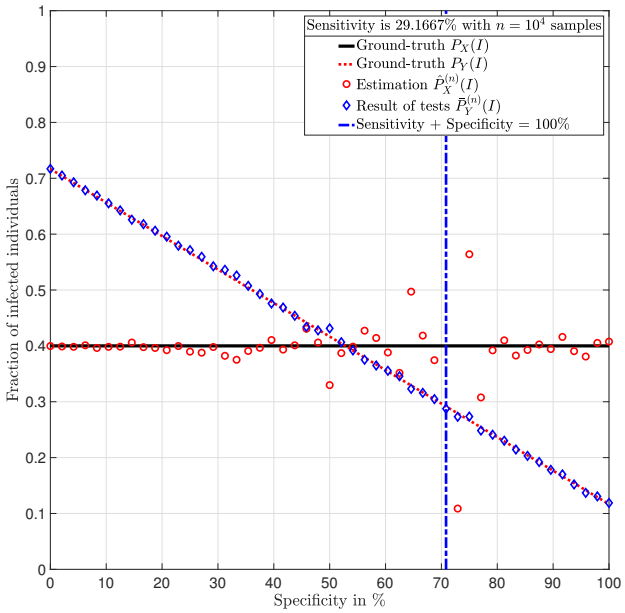


(a) Specificity is 39.6 %

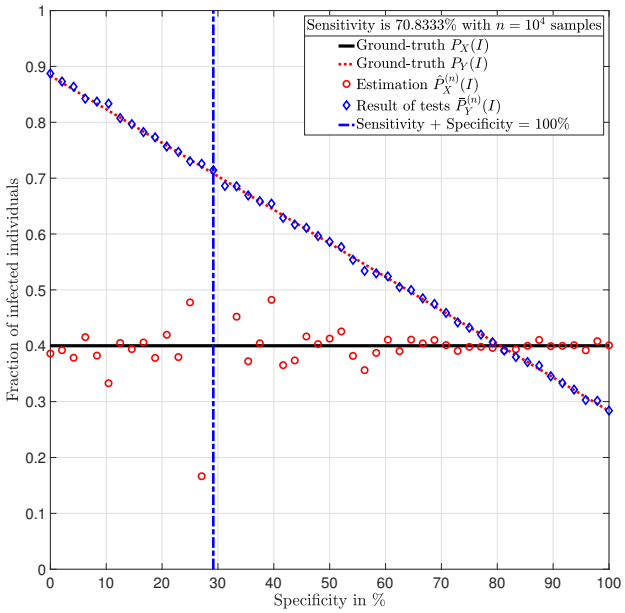


(b) Specificity is 70.8 %

Figure 2: Population in which the fraction of individuals infected with SARS-CoV-2 is $P_X = 0,4$ and $n = 10\,000$ individuals are tested (Example 1).

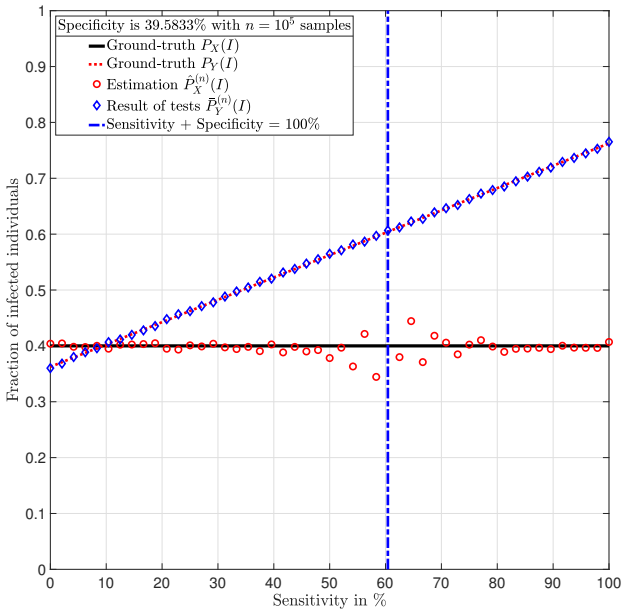


(a) Sensitivity is 29.2 %

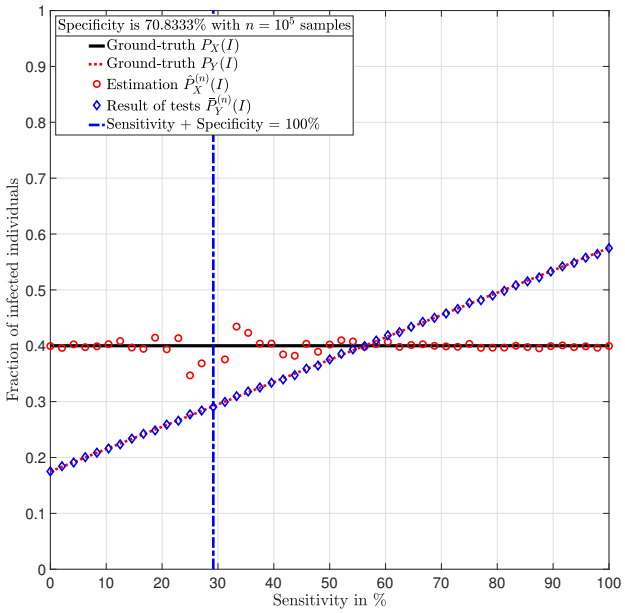


(b) Sensitivity is 70.8 %

Figure 3: Population in which the fraction of individuals infected with SARS-CoV-2 is $P_X = 0,4$ and $n = 10\,000$ individuals are tested (Example 1).

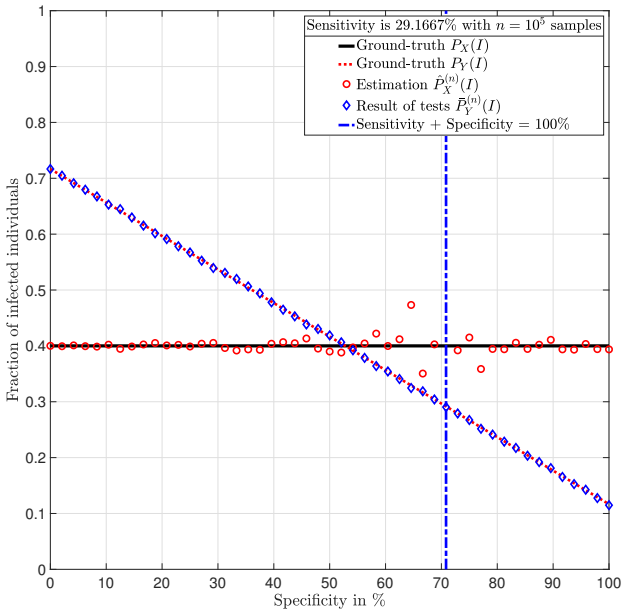


(a) Specificity is 39.6 %

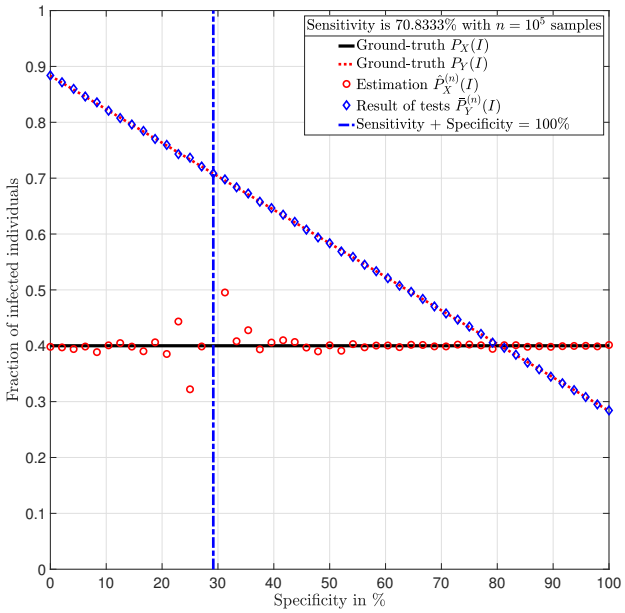


(b) Specificity is 70.8 %

Figure 4: Population in which the fraction of individuals infected with SARS-CoV-2 is $P_X = 0,4$ and $n = 100\,000$ individuals are tested (Example 2).

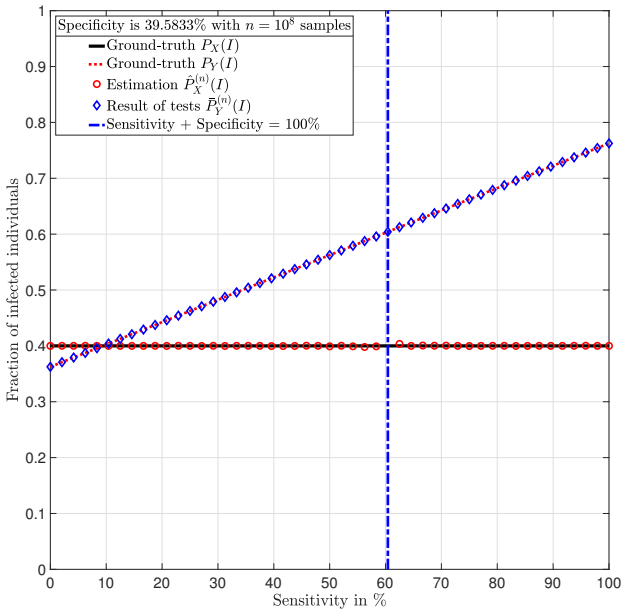


(a) Sensitivity is 29.2 %

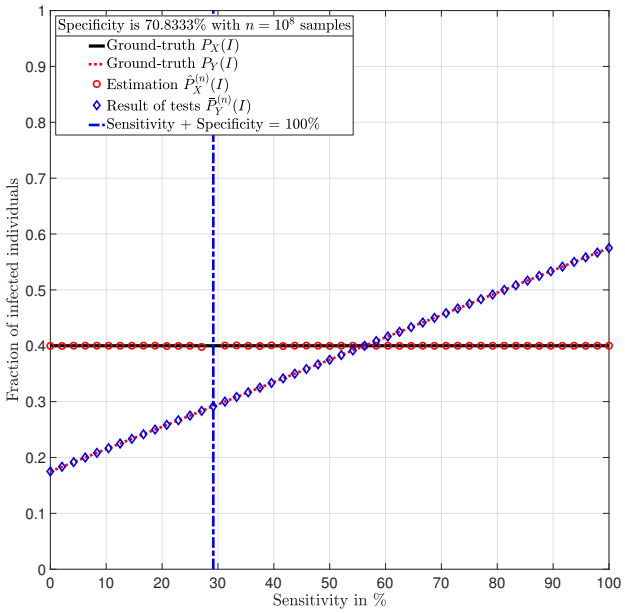


(b) Sensitivity is 70.8 %

Figure 5: Population in which the fraction of individuals infected with SARS-CoV-2 is $P_X = 0,4$ and $n = 100\,000$ individuals are tested (Example 2).

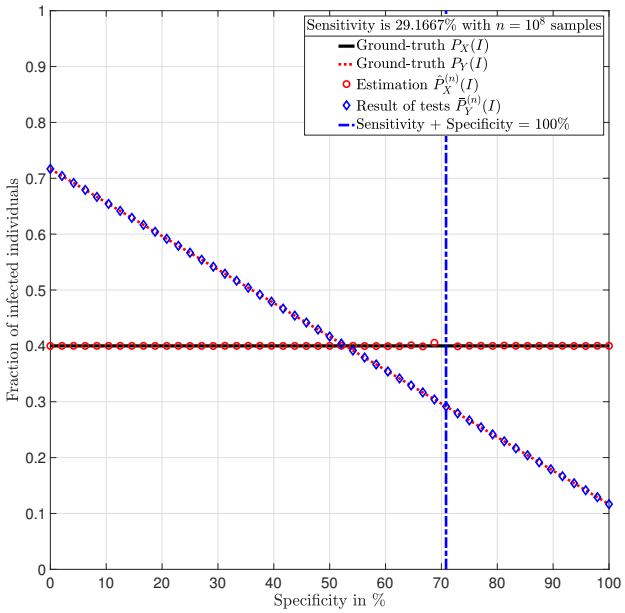


(a) Specificity is 39.6 %

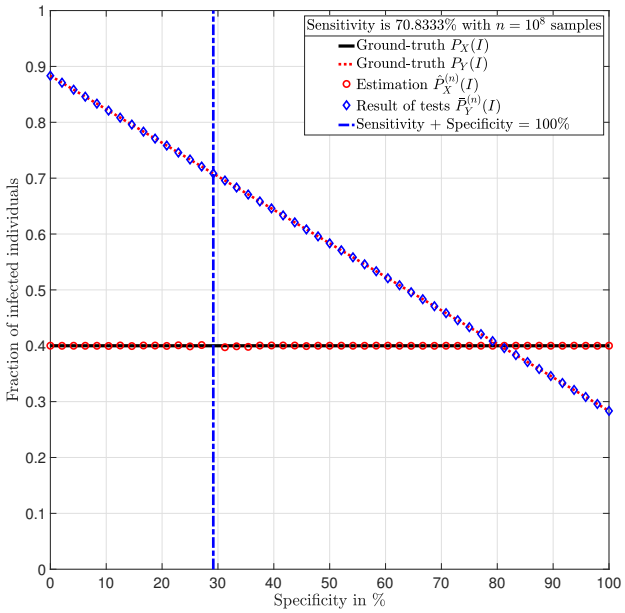


(b) Specificity is 70.8 %

Figure 6: Population in which the fraction of individuals infected with SARS-CoV-2 is $P_X = 0,4$ and $n = 100\,000\,000$ individuals are tested (Example 3).



(a) Sensitivity is 29.2 %



(b) Sensitivity is 70.8 %

Figure 7: Population in which the fraction of individuals infected with SARS-CoV-2 is $P_X = 0,4$ and $n = 100\,000\,000$ individuals are tested (Example 3).

6 Conclusions

In this work, it has been shown that the use of the number of positive tests obtained from testing campaigns against SARS-CoV-2 might lead to erroneous conclusions on the number of infected individuals in a population. This is due to the fact that the number of positive tests might be drastically different to the number of infected individuals depending on the sensitivity and specificity of the tests. From this perspective, an estimation of the number of infected individuals using the data is essential. Theorem 1 provides an estimation of the fraction of infected individuals with an estimation error that decreases with the number of tests. That is, the estimation error can be made arbitrarily small by increasing the number of tests.

Another important conclusion of this work is that testing campaigns using tests for which the sum of the sensitivity and specificity is different from one (100%), always allow a reliable estimation of the number of infected individuals (Lemma 1 in Section 4). Alternatively, testing campaigns using tests for which the sum of the sensitivity and the specificity is equal to one (100%), lead to data from which it is impossible to estimate the number of infected individuals (Lemma 2 in Section 4).

A final conclusion is that for estimating the number of SARS-CoV-2 infections in a population, the key parameter for reducing the estimation error is the number of tests. Surprisingly, as long as the sum of the sensitivity and specificity of the tests do not add up to one (100%), the exact values of both sensitivity and specificity have very little impact in the estimation when the number of tests is sufficiently large.

7 Further Research

The results presented in this work exhibit many limitations and thus, further research is needed to relax certain assumptions that might not be necessarily realistic. The following sections describe several research paths in this direction.

7.1 Beyond Binary Tests

This work has been developed considering that tests can exclusively distinguish between infected and susceptible individuals. That is, the input and the output of the random transformation $P_{Y|X}$ in (1) are binary. Nonetheless, with the advancement on the knowledge about the SARS-CoV-2, in the near future, it would be possible to distinguish more states, e.g., immune; infected and contagious; infected and noncontagious; among others. This extension is trivial as long as the matrix it induces in the system in (9) is invertible. The matrix is not invertible when an additional state is considered, e.g., undetermined. The state undetermined can be a state in which the test is not capable of classifying the individual among the input states, and thus, a new state is considered at the output. The mathematical model would be far from trivial and reminiscent to that of *population recovery with lossy observations* [25].

7.2 Tests with Unknown Parameters

One of the assumptions adopted for developing the results of this paper is that the sensitivity and the specificity of the tests are assumed to be known. Nonetheless, despite the data obtained from the provider of the tests, the method and preparation of the staff responsible of taking the samples, as well as, the manipulation and transportation of samples, play a central role in the final sensitivity and specificity of the tests. From this perspective, formulating the problem in

which the random transformation $P_{Y|X}$ in (1) is not completely known is an important research direction.

7.3 Non-Independent Tests

The results obtained in this work rely on the assumption that the testing results obtained by each individual are independent of all other individuals. This assumption neglects obvious interactions between individual, e.g., members of the same family. An important research direction is the case in which such interactions are taken into account and thus, correlations between the input random variables are considered. This would allow to consider the existence of clusters among the population and thus, refine the estimation of the number of infected individuals in the population.

7.4 Finite Number of Tests and Budget Optimization

An essential constraint that has been neglected in this study is the cost of performing a test. Hence, a relevant questions is: given a number a number tests n and the knowledge of the existing interactions among the individuals, what are the n individuals that must be tested to reduce the estimation error of the number of infected individuals. In this direction, the consideration of the correlation between the state of each individual is essential, which leads to a nontrivial mathematical problem in which elements of *group testing* [26] might play an essential role.

References

[1] D. B. Vinh, X. Zhao, K. L. Kiong, T. Guo, Y. Jozaghi, C. Yao, J. M. Kelley, and E. Hanna, "Overview of COVID-19 testing and implications for otolaryngologists," *Head & Neck*, vol. 0, no. 0, pp. 1–5, Apr. 2020.

[2] J. C. Kelly, M. Dombrowksi, M. O'neil-Callahan, A. S. Kernberg, A. I. Frolova, and M. J. Stout, "False-negative COVID-19 testing: Considerations in obstetrical care," *American Journal of Obstetrics and Gynecology MFM*, vol. 0, no. 0, p. 100130, Apr. 2020.

[3] HAS, "Place des tests sérologiques rapides (TDR, TROD, autotests) dans la stratégie de prise en charge de la maladie COVID- 19," Haute Autorité de Santé (HAS), Tech. Rep., Apr. 2020.

[4] ——, "Place des tests sérologiques dans la stratégie de prise en charge de la maladie COVID-19," Haute Autorité de Santé (HAS), Tech. Rep., Apr. 2020.

[5] W. Wang, Y. Xu, R. Gao, R. Lu, K. Han, G. Wu, and W. Tan, "Detection of SARS-CoV-2 in different types of clinical specimens," *JAMA*, vol. 323, no. 18, pp. 1843–1844, May 2020.

[6] Y. Yang, M. Yang, C. Shen, F. Wang, J. Yuan, J. Li, M. Zhang, Z. Wang, L. Xing, J. Wei, L. Peng, G. Wong, H. Zheng, M. Liao, K. Feng, J. Li, Q. Yang, J. Zhao, Z. Zhang, L. Liu, and Y. Liu, "Evaluating the accuracy of different respiratory specimens in the laboratory diagnosis and monitoring the viral shedding of 2019-nCoV infections," *medRxiv*, vol. 0, no. 0, pp. 1–17, Feb. 2020.

[7] L. Zou, F. Ruan, M. Huang, L. Liang, H. Huang, Z. Hong, J. Yu, M. Kang, Y. Song, J. Xia, Q. Guo, T. Song, J. He, H.-L. Yen, M. Peiris, and J. Wu, "SARS-CoV-2 viral load in upper respiratory specimens of infected patients," *New England Journal of Medicine*, vol. 382, no. 12, pp. 1177–1179, Mar. 2020.

[8] P. B. van Kasteren, B. van der Veer, S. van den Brink, L. Wijsman, J. de Jonge, A. van den Brandt, R. Molenkamp, C. B. Reusken, and A. Meijer, "Comparison of commercial RT-PCR diagnostic kits for COVID-19," *bioRxiv*, vol. 0, no. 0, pp. 1–14, Apr. 2020.

[9] T. Ishige, S. Murata, T. Taniguchi, A. Miyabe, K. Kitamura, K. Kawasaki, M. Nishimura, H. Igari, and K. Matsushita, "Highly sensitive detection of SARS-CoV-2 RNA by multiplex rRT-PCR for molecular diagnosis of COVID-19 by clinical laboratories," *Clinica Chimica Acta*, vol. 507, no. 0, pp. 139–1142, Aug. 2020.

[10] Y. Baek, J. Um, K. J. Antigua, J.-H. Park, Y. Kim, S. Oh, Y.-i. Kim, W.-S. Choi, S. Kim, J. Jeong, B. S. Chin, H. Nicolas, J.-Y. Ahn, K. Shin, Y. K. Choi, J.-S. Park, and M.-S. Song, "Development of a reverse transcription-loop-mediated isothermal amplification as a rapid early-detection method for novel SARS-CoV-2," *Emerging Microbes and Infections*, vol. 9, no. 0, pp. 1–31, Apr. 2020.

[11] C. Yan, J. Cui, L. Huang, B. Du, L. Chen, G. Xue, S. Li, W. Zhang, L. Zhao, Y. Sun, H. Yao, N. Li, H. Zhao, Y. Feng, S. Liu, Q. Zhang, D. Liu, and J. Yuan, "Rapid and visual detection of 2019 novel coronavirus (SARS-CoV-2) by a reverse transcription loop-mediated isothermal amplification assay," *Clinical Microbiology and Infection*, vol. 0, no. 0, pp. 1–7, Apr. 2020.

[12] N. Merindol, G. Pépin, C. Marchand, M. Rheault, C. Peterson, A. Poirier, C. Houle, H. Germain, and A. Danylo, "SARS-CoV-2 detection by direct rRT-PCR without RNA extraction," *Journal of Clinical Virology*, vol. 128, no. 0, p. 104423, Jul. 2020.

[13] M. N. Esbin, O. N. Whitney, C. S., A. Maurer, X. Darzacq, and R. Tjian, "Overcoming the bottleneck to widespread testing: A rapid review of nucleic acid testing approaches for COVID-19 detection," *RNA Journal*, vol. 0, no. 0, pp. 1–20, May 2020.

[14] F. Xiang, X. Wang, X. He, Z. Peng, B. Yang, J. Zhang, Q. Zhou, H. Ye, Y. Ma, H. Li, X. Wei, P. Cai, and W.-L. Ma, "Antibody Detection and Dynamic Characteristics in Patients with COVID-19," *Clinical Infectious Diseases*, vol. 0, no. 0, pp. 1–23, Apr. 2020.

[15] T. Hoffman, K. Nissen, J. Krambrich, B. Rönnberg, D. Akaberi, M. Esmaeilzadeh, E. Salaneck, J. Lindahl, and A. Lundkvist, "Evaluation of a COVID-19 IgM and IgG rapid test: An efficient tool for assessment of past exposure to SARS-CoV-2," *Infection Ecology and Epidemiology*, vol. 10, no. 1, p. 1754538, Apr. 2020.

[16] Z. Zainol Rashid, S. N. Othman, M. N. Abdul Samat, U. K. Ali, and K. K. Wong, "Diagnostic performance of COVID-19 serology assays," *Malays J Pathol.*, vol. 42, no. 1, pp. 13–21, Apr. 2020.

[17] M. Chung, A. Bernheim, X. Mei, N. Zhang, M. Huang, X. Zeng, J. Cui, W. Xu, Y. Yang, Z. A. Fayad, A. Jacobi, K. Li, S. Li, and H. Shan, "CT imaging features of 2019 novel coronavirus (2019-nCoV)," *Radiology*, vol. 295, no. 1, pp. 202–207, Apr. 2020.

[18] H. X. Bai, B. Hsieh, Z. Xiong, K. Halsey, J. W. Choi, T. M. L. Tran, I. Pan, L.-B. Shi, D.-C. Wang, J. Mei, X.-L. Jiang, Q.-H. Zeng, T. K. Egglin, P.-F. Hu, S. Agarwal, F. Xie, S. Li, T. Healey, M. K. Atalay, and W.-H. Liao, "Performance of radiologists in differentiating COVID-19 from viral pneumonia on chest CT," *Radiology*, vol. 0, no. 0, p. 200823, Mar. 2020.

[19] T. Ai, Z. Yang, H. Hou, C. Zhan, C. Chen, W. Lv, Q. Tao, Z. Sun, and L. Xia, "Correlation of chest CT and RT-PCR testing in coronavirus disease 2019 (COVID-19) in china: A report of 1014 cases," *Radiology*, vol. 0, no. 0, p. 200642, Feb. 2020.

[20] Z. Dvir, A. Rao, A. Wigderson, and A. Yehudayoff, "Restriction access," in *Proc. of the 3rd Innovations in Theoretical Computer Science Conference*, Jan. 2012, pp. 19–33.

[21] S. Lovett and J. Zhang, "Improved noisy population recovery, and reverse Bonami-Beckner inequality for sparse functions," in *Proc. of the Forty-Seventh Annual ACM Symposium on Theory of Computing*, Portland, Oregon, USA, Jun. 2015, pp. 137–142.

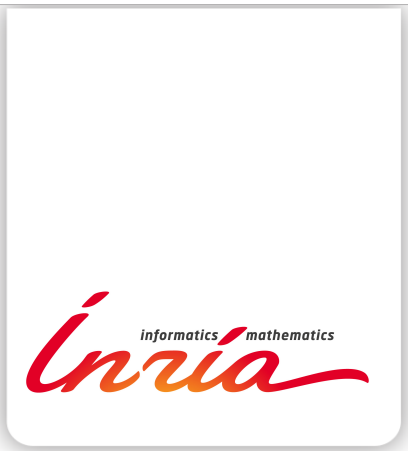
[22] A. De, M. Saks, and S. Tang, "Noisy population recovery in polynomial time," in *Proc. of the 57th Annual Symposium on Foundations of Computer Science (FOCS)*, New Brunswick, NJ, USA, Oct. 2016, pp. 675–684.

[23] I. R. Shafarevich and A. O. Remizov, *Linear Algebra and Geometry*, 1st ed. Berlin, Germany: Springer, 2012.

[24] R. B. Ash and C. A. Doléans-Dade, *Probability and Measure Theory*, 2nd ed. Burlington, MA, USA: Harcourt/Academic Press, 1999.

[25] A. Wigderson and A. Yehudayoff, "Population recovery and partial identification," in *Proc. of the 53rd Annual Symposium on Foundations of Computer Science*, New Brunswick, NJ, USA, Oct. 2012, pp. 390–399.

[26] M. Aldridge, O. Johnson, and J. Scarlett, "Group testing: An information theory perspective," *Foundations and Trends in Communications and Information Theory*, vol. 15, no. 3-4, pp. 196–392, Dec. 2019.



RESEARCH CENTRE
SOPHIA ANTIPOLIS – MÉDITERRANÉE

2004 route des Lucioles - BP 93
06902 Sophia Antipolis Cedex

Publisher
Inria
Domaine de Voluceau - Rocquencourt
BP 105 - 78153 Le Chesnay Cedex
inria.fr

ISSN 0249-6399