

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

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

▶ **To cite this version:**

Eitan Altman, Izza Mounir, Fatim-Zahra Najid, Samir Perlaza. On the true number of COVID-19 infections: Effect of Sensitivity, Specificity and Number of Tests on Prevalence Ratio Estimation. [Research Report] RR-9344, INRIA Sophia Antipolis - Méditerranée. 2020. hal-02633844v3

HAL Id: hal-02633844
https://inria.hal.science/hal-02633844v3

Submitted on 4 Aug 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

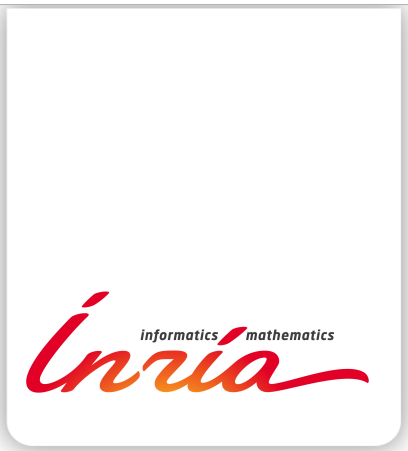


On the true number of COVID-19 infections: Effect of the Sensitivity, Specificity and Number of Tests on Prevalence Estimation

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 on Prevalence Estimation

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

Project-Team NEO

Research Report n° 9344 — version 2 — initial version May 2020 — revised version July 2020 — 28 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 Amiens Picardie - 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 prevalence ratio of a disease in a population that is tested with imperfect tests is given. The formula is in terms of the fraction of positive test results and test parameters, i.e., probability of true positives (sensitivity) and the probability of true negatives (specificity). The motivation of this work arises in the context of the COVID-19 pandemic in which estimating the number of infected individuals depends on the sensitivity and specificity of the tests. In this context, it is shown that approximating the prevalence ratio by the ratio between the number of positive tests and the total number of tested individuals leads to dramatically high estimation errors, and thus, unadapted public health policies. The relevance of estimating the prevalence ratio using the formula presented in this work is that precision increases with the number of tests. Two conclusions are drawn from this work. First, in order to ensure that a reliable estimation is achieved with a finite number of tests, 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 is different from one, the exact values of these parameters have very little impact on the estimation error.

Key-words: SARS-CoV-2; Covid-19; Cross-Sectional Studies; Prevalence Ratio; Sensitivity and Specificity; Molecular, Serological and Medical Imaging diagnostics; Number of Infections; False Positive and False Negative Probabilities; Policy-Making and 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 **5**

2 Case Study: SARS-CoV-2 **6**
2.1 Virological Tests 6
2.2 Serological Tests 7
2.3 Medical Imaging 7

3 Prevalence Ratio and Unreliable Tests **7**

4 Estimation of the Prevalence Ratio Using Unreliable Tests **9**
4.1 Main Result 9
4.2 Proof of Theorem 1 10
4.3 Connections to Maximum Likelihood Estimation 13

5 Final Remarks **13**
5.1 Relevance of the Sensitivity and Specificity 14
5.2 Tests whose Results are Useless 15
5.3 Impact of the Number of Tests. 24

6 Conclusions **24**

7 Further Research **24**
7.1 Beyond Binary Tests 25
7.2 Tests with Unknown Parameters 25
7.3 Non-Independent Tests 25
7.4 Finite Number of Tests and Budget Optimization 25

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, for example, encouraging social distancing, isolation of infected individuals, mobility restrictions, lockdowns, and contact tracing. The main objective is to guarantee that the number of infected individuals that develop critical forms of symptoms does not exceed the capacity of local health care systems. Nonetheless, most of the strategies to slow down the outbreak induce dramatic economical consequences, and thus, public health policies must be designed based on reliable predictions of the evolution of the pandemic to minimize undesired effects on the global economy. For doing so, estimating the values of variables such as the proportion of susceptible, infected and recovered individuals in the population, among other variables, is of paramount importance. This is due to the fact that such variables are the inputs of mathematical models that help to predict the evolution of the pandemic [1, 2], and thus, impact public health policy-making. Reliable estimations of these variables can be achieved in part by testing the population. Nonetheless, diagnosing SARS-CoV-2 is a challenging task given that designing highly reliable tests for massive testing is still an open research problem, c.f., [3, 4, 5].

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 a condition, for example, a SARS-CoV-19 infection. Alternatively, the latter is the probability with which a test is able to correctly identify the absence of such condition. Within this context, the main contribution of this work is a mathematical formula for estimating the fraction of individuals that exhibit the condition in a population in which every individual has been tested once with identical unreliable tests. In the following, this fraction is referred to as the prevalence ratio [2]. In these terms, the main result is Theorem 1 in Section 4, which presents an estimator of the prevalence ratio in terms of the sensitivity, specificity and the fraction of positive test results. More importantly, the estimation error induced by such estimator is proved to decrease with the number of tests.

The novelty of this work with respect to existing methods for estimating the prevalence ratio, for example, method of multipliers, capture and recapture methods, among others, c.f., [2, 6], is that it takes into account the effects of both false positive and false negative probabilities. More specifically, it takes into account the fact that some individuals that are infected could have been declared noninfected and vice versa. This consideration has already been discussed by several authors, c.f., [7, 8, 9, 10]. Nonetheless, a simple general formula for estimating prevalence ratios in terms of the sensitivity, specificity, and the fraction of positive test results is not available in current literature. This said, the prevalence ratio estimation presented in this work is based exclusively on the results of data obtained through testing campaigns with unreliable binary tests. No other assumption is taken into account. This breaks away from the studies based on mathematical regressions in which some assumptions on the random variables are often adopted and correctness is often the ground of vivid discussions, c.f., [11, 12, 13, 14]. In the particular case of SARS-CoV-19, very little is known about the underlying stochastic properties of the virus dissemination and thus, arbitrary assumptions might lead to estimation errors.

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 ratio between the number of positive tests and the total number of tested individuals is not a reliable estimation of the prevalence ratio;

(ii) 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 when a sufficiently large number of individuals is tested in the population (Lemma 1 in Section 4);

(iii) 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 prevalence ratio independently of the number of tested individuals (Lemma 7 in Section 4); and

(iv) When the objective is to estimate the prevalence ratio in a population, the key parameter for reducing the estimation error is the number of tests. That is, as long as the sum of the sensitivity and specificity is different than one, and a large number of test results is available, the exact values of both sensitivity and specificity have very little impact on the estimation error.

The remaining sections of this paper are organized as follows: Section 2 presents a brief overview of the tests for diagnosing SARS-CoV-2 and the reliability of the existing tests; Section 3 formulates the problem of estimating the prevalence ratio taking into account the sensitivity and specificity of the tests; Section 4 presents an estimator of the prevalence ratio using data obtained from unreliable tests, 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.

2 Case Study: 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 [15]. 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. [3, 5].

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., [16, 17]. 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 [16, 17], 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., [18, 19]; and (*b*) Reverse transcription loop-mediated isothermal amplification (RT-LAMP), c.f., [20, 21]; and (*c*) other techniques, c.f., [19, 22, 23].

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., [24, 25, 26]. Serological tests are based on the enzyme linked immunosorbent assay (ELISA) and exhibit high specificity and sensitivity, after fourteen days of infections [24]. This drastically limits the use of serological tests in the early detection of the infection and policy-making, c.f., [3, 4]. In a nutshell, on the 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 paper, serological tests are not considered for massive testing in France, c.f., [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 [27]. 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 [28]. In [29], it is reported that the sensitivity of CT is better than the one achieved by RT-PCR tests.

3 Prevalence Ratio and Unreliable Tests

Consider a population subset of n individuals whose state is either susceptible (S) or infected (I) and assume that all individuals of this population subset are tested with the same type of test. Let the actual 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 modeled 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 this binary-input binary-output model.

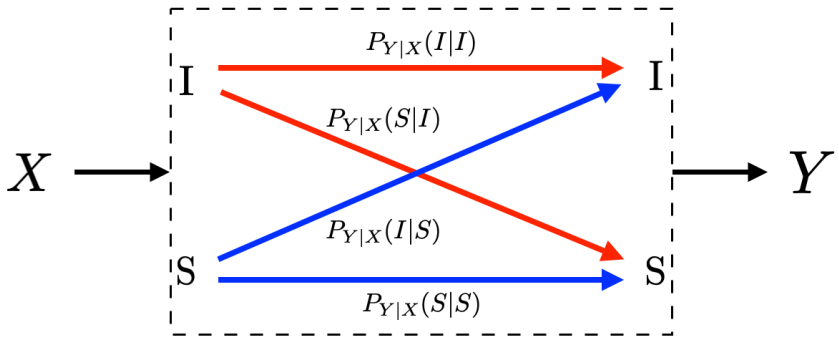


Figure 1: A SARS-CoV-2 test represented by a random transformation from $\{I, S\}$ into $\{I, S\}$ via the conditional probability distribution $P_{Y|X}$.

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 a false positive is $P_{Y|X}(I|S) = 1-P_{Y|X}(S|S)$; and the probability of a 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 a false positive;
- The probability of a false negative and the specificity; or
- The probability of a false negative and the probability of a false positive.

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 actual fraction of infected individuals among the n individuals. That is, $P_X(I)$ is the *prevalence ratio* of SARS-CoV-19 in this population subset. 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. See, for instance, Figure 1. 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 in which parameters are modeled by $P_{Y|X}$ to determine the fraction $P_X(I)$ of infected individuals in the population, i.e., the prevalence ratio. 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 [30] and further studied in [31, 32].

4 Estimation of the Prevalence Ratio Using Unreliable Tests

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,

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

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

where $\mathbb{1}_{\{\cdot\}}$ is the indicator function. Essentially, $\bar{P}_Y^{(n)}$ is a counting probability measure for which the values $\bar{P}_Y^{(n)}(I)$ and $\bar{P}_Y^{(n)}(S)$ represent the fraction of positive and negative test results. Hence, $\bar{P}_Y^{(n)}(I) + \bar{P}_Y^{(n)}(S) = 1$. In the following, such probability measure is often referred to as the *output empirical distribution* obtained from the data $\boldsymbol{y}$.

Let $\hat{P}_X^{(n)} : \{I, S\} \to \mathbb{R}$ be a function 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, which is denoted by $\left\|P_X - \hat{P}_X^{(n)}\right\|_{\text{TV}}$ and satisfies,

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

$$= \left|P_X(I) - \hat{P}_X^{(n)}(I)\right|. \tag{5}$$

Note that in the case of binary tests, the total variation is simply the absolute difference between the actual prevalence ratio $P_X(I)$ and the estimate $\hat{P}_X^{(n)}(I)$.

4.1 Main Result

The following theorem presents the main result of this work.

Theorem 1. *Consider a population of n individuals 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 all individuals of such population are tested with a test $P_{Y|X}$ that satisfies*

$$P_{Y|X}(S|S) + P_{Y|X}(I|I) \neq 1. \tag{6}$$

Let $\bar{P}_Y^{(n)}$ be the resulting output empirical probability distribution in (3) *and assume that $\bar{P}_Y^{(n)}(I)$ satisfies the following condition,*

$$\min\{P_{Y|X}(I|I), P_{Y|X}(I|S)\} \leqslant \bar{P}_Y^{(n)}(I) \leqslant \max\{P_{Y|X}(I|I), P_{Y|X}(I|S)\}. \tag{7}$$

Then, the estimator $\hat{P}_X^{(n)} : \{I, S\} \to \mathbb{R}$ of P_X, such that

$$\hat{P}_X^{(n)}(I) = \frac{1 - \bar{P}_Y^{(n)}(I) - P_{Y|X}(S|S)}{1 - P_{Y|X}(I|I) - P_{Y|X}(S|S)}, \text{ and} \tag{8a}$$

$$\hat{P}_X^{(n)}(S) = \frac{\bar{P}_Y^{(n)}(I) - P_{Y|X}(I|I)}{1 - P_{Y|X}(I|I) - P_{Y|X}(S|S)}, \tag{8b}$$

forms a probability measure that satisfies

$$\lim_{n\to\infty} \left\| P_X - \hat{P}_X^{(n)} \right\|_{\mathrm{TV}} = 0 \text{ in probability.} \tag{9}$$

In a nutshell, Theorem 1 states that the value $\hat{P}_X^{(n)}(I)$ constitutes an estimation of the prevalence ratio $P_X(I)$. Moreover, it shows that such estimation is asymptotically optimal. That is, approximating P_X by $\hat{P}_X^{(n)}$ induces an error that vanishes when the number of tests n increases. Nonetheless, despite the fact that $\hat{P}_X^{(n)}(I) + \hat{P}_X^{(n)}(S) = 1$, it holds that for a small number of tests n, $\hat{P}_X^{(n)}(I)$ and $\hat{P}_X^{(n)}(S)$ do not necessarily form a probability measure. That is, it might be observed that either $\hat{P}_X^{(n)}(I) < 0$ and $\hat{P}_X^{(n)}(S) > 1$; or $\hat{P}_X^{(n)}(I) > 1$ and $\hat{P}_X^{(n)}(S) < 0$. Later, in Lemma 4, it is shown that with a large number of test results, the fraction of positive results $\bar{P}_Y^{(n)}(I)$ satisfies the inequalities in (7). Note also that the condition in (7) is necessary and sufficient to observe that $0 \leqslant \hat{P}_X^{(n)}(I) \leqslant 1$ in Theorem 1. This highlights the need for a sufficiently large number of tests in order to obtain a valid estimation of $P_X(I)$ using Theorem 1.

Finally, note that the formulas in (8) 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 a false positive and a false negative, or any combination of the parameters describing the test. The following corollary shows the formulas in (8) in terms of the probabilities of a false positive $P_{Y|X}(I|S)$ and a false negative $P_{Y|X}(S|I)$.

Corollary 1. *Consider a population of n individuals 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 all individuals of such population are tested with a test $P_{Y|X}$ that satisfies* (6)*. Let $\bar{P}_Y^{(n)}$ be the resulting output empirical probability distribution in* (3) *and assume that $\bar{P}_Y^{(n)}(I)$ satisfies condition* (7)*. Then, the estimator $\hat{P}_X^{(n)} : \{I, S\} \to \mathbb{R}$ of P_X, such that*

$$\hat{P}_X^{(n)}(I) = \frac{\bar{P}_Y^{(n)}(I) - P_{Y|X}(I|S)}{1 - P_{Y|X}(S|I) - P_{Y|X}(I|S)}, \text{ and} \tag{10a}$$

$$\hat{P}_X^{(n)}(I) = \frac{1 - P_{Y|X}(S|I) - \bar{P}_Y^{(n)}(I)}{1 - P_{Y|X}(S|I) - P_{Y|X}(I|S)}, \tag{10b}$$

forms a probability measure that satisfies (9)*.*

4.2 Proof of Theorem 1

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 valid estimation of the ground-truth output probability distribution P_Y, i.e., it satisfies (7), then a distribution $\hat{P}_X^{(n)}$ that satisfies

$$\begin{pmatrix} \bar{P}_Y^{(n)}(I) \\ \bar{P}_Y^{(n)}(S) \end{pmatrix} = \begin{pmatrix} P_{Y|X}(I|I) & P_{Y|X}(I|S) \\ P_{Y|X}(S|I) & P_{Y|X}(S|S) \end{pmatrix} \begin{pmatrix} \hat{P}_X^{(n)}(I) \\ \hat{P}_X^{(n)}(S) \end{pmatrix}, \tag{11}$$

is a good estimation of the 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 (11).

With this intuition in mind, the proof proceeds as follows. First, it is shown that under the condition in (6), there exists a unique pair $(\hat{P}_X^{(n)}(I), \hat{P}_X^{(n)}(S))$ that satisfies the equality in (11). This is essentially due to the fact that the equality in (11) forms 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.

Lemma 1. *Consider the empirical output distribution $\bar{P}_Y^{(n)}$ in (3) obtained by a test described by the conditional probability distribution $P_{Y|X}$. Then, the following five statements are equivalent:*

- *The system of equations in (11) has a unique solution;*
- *The sensitivity $P_{Y|X}(I|I)$ and specificity $P_{Y|X}(S|S)$ satisfy*
$$P_{Y|X}(I|I) + P_{Y|X}(S|S) \neq 1; \tag{12a}$$
- *The sensitivity $P_{Y|X}(I|I)$ and the probability of a false positive $P_{Y|X}(I|S)$ satisfy*
$$P_{Y|X}(I|I) \neq P_{Y|X}(I|S); \text{ and} \tag{12b}$$
- *The probability of a false negative $P_{Y|X}(S|I)$ and the specificity $P_{Y|X}(S|S)$ satisfy*
$$P_{Y|X}(S|S) \neq P_{Y|X}(S|I). \tag{12c}$$
- *The probability of a false positive $P_{Y|X}(I|S)$ and the probability of a false negative $P_{Y|X}(S|I)$ satisfy*
$$P_{Y|X}(I|S) + P_{Y|X}(S|I) \neq 1. \tag{12d}$$

Proof. The proof of Lemma 1 follows from the fact that a unique solution to (11) 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 [33]). That is,

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

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

Note that all conditions in (12) are equivalent to each other, and thus, they are equivalent to the condition in (6).

The proof of Theorem 1 continues by showing that when such a unique solution exists, it is identical to the one shown in (8).

Lemma 2. *Consider a test $P_{Y|X}$ that satisfies at least one of the conditions in (12). Then, under the assumption that the empirical output distribution $\bar{P}_Y^{(n)}$ in (3) satisfies (7), the unique probability distribution $\hat{P}_X^{(n)}$ that satisfies (11) is:*

$$\hat{P}_X^{(n)}(I) = \frac{1 - \bar{P}_Y^{(n)}(I) - P_{Y|X}(S|S)}{1 - P_{Y|X}(I|I) - P_{Y|X}(S|S)}, \text{ and} \tag{14a}$$

$$\hat{P}_X^{(n)}(S) = \frac{\bar{P}_Y^{(n)}(I) - P_{Y|X}(I|I)}{1 - P_{Y|X}(I|I) - P_{Y|X}(S|S)}. \tag{14b}$$

Proof. The proof of Lemma 2 follows from solving the system of equations in (11) and observing that $\hat{P}_X^{(n)}$ is a probability measure if and only if condition (7) holds. □

The rest of the proof of Theorem 1 consists of showing that the error vanishes with the number of test results. This is shown in three steps. The first step consists of showing 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 scaling factor.

Lemma 3. *Consider a test $P_{Y|X}$ that satisfies at least one of the conditions in* (12)*. Then, under the assumption that the empirical output distribution $\bar{P}_Y^{(n)}$ in* (3) *satisfies* (7)*, the estimation $\hat{P}_X^{(n)}$ in* (8) *of P_X satisfies*

$$\left\|P_X - \hat{P}_X^{(n)}\right\|_{\text{TV}} = \frac{1}{\left|1 - P_{Y|X}(I|I) - P_{Y|X}(S|S)\right|} \left\|P_Y - \bar{P}_Y^{(n)}\right\|_{\text{TV}}, \tag{15}$$

where P_X and P_Y are the input and output probability distributions in (1) *and* (2)*, respectively.*

Proof. The proof of Lemma 3 follows from the definition of total variation in (4) and from equalities in (14). □

Note that Lemma 3 proves the intuition over which the proof of Theorem 1 is based on. That is, if $\bar{P}_Y^{(n)}$ is sufficiently close to P_Y, then $\hat{P}_X^{(n)}$ must be sufficiently close to P_X. The following lemma shows that the more test results are available, the closer $\bar{P}_Y^{(n)}$ and P_Y are in total variation.

Lemma 4. *Consider a test $P_{Y|X}$ that satisfies at least one of the conditions in* (12)*. Then, the empirical output distribution $\bar{P}_Y^{(n)}$ in* (3) *satisfies*

$$\lim_{n\to\infty} \left\|P_Y - \bar{P}_Y^{(n)}\right\|_{\text{TV}} = 0 \text{ in probability,} \tag{16}$$

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

Proof. The proof of Lemma 4 is a consequence of the Theorem of Glivenko and Cantelli [34]. □

Finally, from Lemma 3 and Lemma 4, it holds that by increasing the number of tests, the error of approximating P_X by $\hat{P}_X^{(n)}$ in (14) can be made arbitrarily small. The following lemma leverages this observation.

Lemma 5. *Consider a test $P_{Y|X}$ that satisfies at least one of the conditions in* (12)*. Then, under the assumption that the empirical output distribution $\bar{P}_Y^{(n)}$ in* (3) *satisfies* (7)*, the input distribution P_X and the estimation $\hat{P}_X^{(n)}$ in* (8) *satisfy*

$$\lim_{n\to\infty} \left\|P_X - \hat{P}_X^{(n)}\right\|_{\text{TV}} = 0 \text{ in probability.} \tag{17}$$

Proof. The proof of Lemma 5 is an immediate consequence of both Lemma 3 and Lemma 4. □

This completes the proof of Theorem 1.

4.3 Connections to Maximum Likelihood Estimation

In this section, it is shown that the estimator presented in Theorem 1 is also the *maximum likelihood estimator*. For doing so, note that under the assumption that the prevalence ratio is $\hat{P}_X^{(n)}(I) \in [0,1]$, the probability of observing $y \in \{I,S\}$, as the result of testing any of the individuals of the population with a test described by the conditional probability distribution $P_{Y|X}$ is:

$$\hat{P}_Y^{(n)}(y) \triangleq \sum_{x\in\{I,S\}} P_{Y|X}(y|x)\hat{P}_X^{(n)}(x). \tag{18}$$

From this perspective, the probability of observing the vector $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)$, as the result of a testing campaign over a population of n individuals is

$$\left(\hat{P}_Y^{(n)}(I)\right)^{n\bar{P}_Y^{(n)}(I)} \left(1-\hat{P}_Y^{(n)}(I)\right)^{n\left(1-\bar{P}_Y^{(n)}(I)\right)}, \tag{19}$$

where $\bar{P}_Y^{(n)}$ and $\hat{P}_Y^{(n)}$ are defined in (3) and (18), respectively. Hence, the log-likelihood function $L : \{I,S\}^n \times [0,1] \to [0,1]$ is for all $\boldsymbol{y} \in \{I,S\}^n$ and $\hat{P}_X^{(n)}(I) \in [0,1]$,

$$\begin{aligned}
L(\boldsymbol{y}, \hat{P}_X^{(n)}(I)) &= n\Big(\bar{P}_Y^{(n)}(I)\ln\left(\hat{P}_Y^{(n)}(I)\right) + \left(1-\bar{P}_Y^{(n)}(I)\right)\ln\left(1-\hat{P}_Y^{(n)}(I)\right)\Big) && (20)\\
&= n\Bigg(\bar{P}_Y^{(n)}(I)\ln\left(\hat{P}_Y^{(n)}(I)\right) - \bar{P}_Y^{(n)}(I)\ln\left(\bar{P}_Y^{(n)}(I)\right) && (21)\\
&\quad + \left(1-\bar{P}_Y^{(n)}(I)\right)\ln\left(1-\hat{P}_Y^{(n)}(I)\right) - \left(1-\bar{P}_Y^{(n)}(I)\right)\ln\left(1-\bar{P}_Y^{(n)}(I)\right) && (22)\\
&\quad + \bar{P}_Y^{(n)}(I)\ln\left(\bar{P}_Y^{(n)}(I)\right) + \left(1-\bar{P}_Y^{(n)}(I)\right)\ln\left(1-\bar{P}_Y^{(n)}(I)\right)\Bigg) && (23)\\
&= -n\Big(H\left(\bar{P}_Y^{(n)}\right) + D\left(\bar{P}_Y^{(n)}||\hat{P}_Y^{(n)}\right)\Big), && (24)
\end{aligned}$$

where $H\left(\bar{P}_Y^{(n)}\right)$ denotes the entropy of the probability distribution $\bar{P}_Y^{(n)}$; and $D\left(\bar{P}_Y^{(n)}||\hat{P}_Y^{(n)}\right)$ denotes the Kullback–Leibler divergence between the distributions $\bar{P}_Y^{(n)}$ and $\hat{P}_Y^{(n)}$. Given that $D\left(\bar{P}_Y^{(n)}||\hat{P}_Y^{(n)}\right) > 0$, it follows that

$$L(\boldsymbol{y}, \hat{P}_X^{(n)}(I)) \leqslant -nH\left(\bar{P}_Y^{(n)}\right), \tag{25}$$

where the equality holds if and only if $D\left(\bar{P}_Y^{(n)}||\hat{P}_Y^{(n)}\right) = 0$. That is, when both $\bar{P}_Y^{(n)}$ and $\hat{P}_Y^{(n)}$ are identical. This observation leads to the conclusion that the log-likelihood function is maximized when the assumed prevalence ratio $\hat{P}_X^{(n)}(I)$ is such that $\bar{P}_Y^{(n)}$ in (3) and $\hat{P}_Y^{(n)}$ in (18) are identical, which is induces the system of equations in (11) and in which the unique solution is formed by the equalities in (8). This proves that the estimator in Theorem (1) is the unique maximum likelihood estimator.

5 Final Remarks

This section highlights some of the conclusions drawn from Lemma 1–5 using a numerical analysis in particular examples. In the following examples, the data is artificially generated. That is,

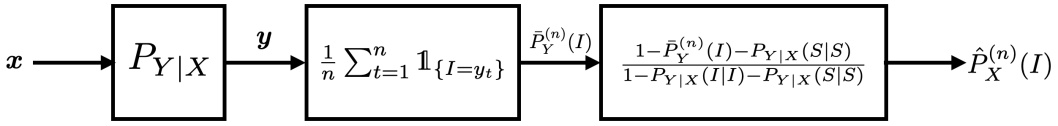


Figure 2: Relation between the input vector $\boldsymbol{x} = (x_1, x_2, \ldots, x_n) \in \{I, S\}^n$ (state of the individuals); the output vector $\boldsymbol{y} = (y_1, y_2, \ldots, y_n) \in \{I, S\}^n$ (result of the tests); the calculation of the fraction of positive tests $\bar{P}_Y^{(n)}(I)$ in (3); and estimation of the prevalence ratio $\hat{P}_X^{(n)}(I)$ in (8a).

for a given prevalence ratio $P_X(I)$, 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 $Y_t \sim P_{Y|X=x_t}$ and represents the result of the test of individual t. Using the vector $\boldsymbol{y}$, the fraction of positive tests $\bar{P}_Y^{(n)}(I)$ is calculated using (3); and the estimation $\hat{P}_X^{(n)}(I)$ of the prevalence ratio $P_X(I)$ is calculated using (8a). Figure 2 shows this procedure. From this perspective, the analysis is based on simulated testing campaigns. Note that the use of simulated data allows knowing the actual prevalence ratio, which enables analyzing the estimation error. This is rarely possible with data from actual testing campaigns.

Example 1. *Consider a population of* $n = 10,000$ *individuals with prevalence* $P_X(I) = 0.4$. *Assume that all individuals are tested with identical tests* $P_{Y|X}$.

Example 2. *Consider a population of* $n = 100,000$ *individuals with prevalence* $P_X(I) = 0.4$. *Assume that all individuals are tested with identical tests* $P_{Y|X}$.

Example 3. *Consider a population of* $n = 100,000,000$ *individuals with prevalence* $P_X(I) = 0.4$. *Assume that all individuals are tested with identical tests* $P_{Y|X}$.

In Figure 3–8, the actual prevalence ratio $P_X(I)$ 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 (2) is plotted with a dashed red line. In Figures 3, 5 and 7, these values are plotted as a function of the specificity $P_{Y|X}(S|S)$ for a fixed sensitivity. Alternatively, in Figures 4, 6 and 8, these values are plotted as a function of the sensitivity $P_{Y|X}(I|I)$ for a fixed specificity. For each of the examples, one vector $\boldsymbol{x} \in \{I, S\}^n$ is generated. In all figures, Figure 3–8, each plotted point of $\bar{P}_Y^{(n)}(I)$ (blue diamonds) and $\hat{P}_X^{(n)}(I)$ (red circles) is calculated using a single vector $\boldsymbol{y}$ generated by the same vector $\boldsymbol{x}$, according to the corresponding values of sensitivity $P_{Y|X}(I|I)$ and specificity $P_{Y|X}(S|S)$, as described above. In the following sections, some remarks based on these examples are presented.

5.1 Relevance of the Sensitivity and Specificity

One of the main observations to be highlighted from this numerical analysis is that there exists an important difference between the fraction of positive tests $\bar{P}_Y^{(n)}(I)$ and the actual prevalence ratio $P_X(I)$ due to the sensitivity and specificity of the tests. This difference is clearly depicted in Figure 3–8, which together with the mathematical analysis presented before, highlights the conclusion that the fraction of positive tests should not be used as an estimation of the prevalence ratio in public health policy-making.

The following lemma determines the influence of the sensitivity and specificity on $\bar{P}_Y^{(n)}(I)$. For doing so, note that from Lemma (2), it holds that the fraction of individuals reporting positive tests $\bar{P}_Y^{(n)}(I)$ satisfies:

$$\bar{P}_Y^{(n)}(I) = 1 - P_{Y|X}(S|S)\left(1 - \hat{P}_X^{(n)}(I)\right) - \left(1 - \hat{P}_{Y|X}(I|I)\right)\hat{P}_X^{(n)}(I). \tag{26}$$

Lemma 6. *Consider a test $P_{Y|X}$ that satisfies at least one of the conditions in* (12)*. Then, given the empirical output distribution $\bar{P}_Y^{(n)}$ in* (3) *and assuming that it satisfies* (7)*, the following statements hold:*

- *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 a 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 a false negative of the test $P_{Y|X}(S|I)$.*

Proof. The proof of Lemma 6 consists in verifying that the derivative of $\bar{P}_Y^{(n)}$ in (26) 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 statements in Lemma 6 become evident in Figures 3, 5 and 7. In these figures, it is shown that the fraction of positive tests increases with the sensitivity; where as in Figure 4, 6 and 8, it is shown that the fraction of positive tests decreases with the specificity, c.f., Lemma 6. From this perspective, tests might lead to countings in which the fraction of individuals reporting positive testing results $\bar{P}_Y^{(n)}(I)$ is bigger than the actual prevalence ratio $P_X(I)$, i.e., $\bar{P}_Y^{(n)}(I) > P_X(I)$. Alternatively, tests might also lead to estimations in which the fraction of individuals reporting positive testing results $\bar{P}_Y^{(n)}(I)$ is smaller than the actual prevalence ratio $P_X(I)$, i.e., $\bar{P}_Y^{(n)}(I) < P_X(I)$. These observations highlight the relevance of using the estimation $\hat{P}_X^{(n)}$ of P_X for decision and policy making instead of $\bar{P}_Y^{(n)}$, which includes false positives and false negatives.

5.2 Tests whose Results are Useless

In Figure 3–8, the value of the sensitivity $P_{Y|X}(I|I)$ and specificity $P_{Y|X}(S|S)$ that satisfy $P_{Y|X}(I|I) + P_{Y|X}(S|S) = 1$ are plotted with a blue dash-dot vertical 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. The following lemmas shed some light into this singularity.

Lemma 7. *Consider the empirical output distribution $\bar{P}_Y^{(n)}$ in* (3) *obtained by a test described by the conditional probability distribution $P_{Y|X}$. Then, the following five statements are equivalent:*

- *The system of equations in* (11) *has infinitely many solutions;*

- *The sensitivity* $P_{Y|X}(I|I)$ *and specificity* $P_{Y|X}(S|S)$ *satisfy*

$$P_{Y|X}(I|I) + P_{Y|X}(S|S) = 1; \tag{27a}$$

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

$$P_{Y|X}(I|I) = P_{Y|X}(I|S); \tag{27b}$$

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

$$P_{Y|X}(S|S) = P_{Y|X}(S|I); \text{ and} \tag{27c}$$

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

$$P_{Y|X}(I|S) + P_{Y|X}(S|I) = 1. \tag{27d}$$

Proof. The proof of Lemma 7 follows from the theorem of Rouché and Fontené [33] that states that when the system in (11) 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,

$$P_{Y|X}(I|I)P_{Y|X}(S|S) - P_{Y|X}(S|I)P_{Y|X}(I|S) = 0. \tag{28}$$

The proof is completed by verifying that the expression in (28) is equivalent to those in (27). □

When at least one of the equalities in (27) is satisfied, nothing meaningful can be said about P_X based on the data. This is essentially because any probability distribution $\hat{P}_X^{(n)}$ satisfies the equality in (11). The following lemma reinforces this statement in terms of information measures.

Lemma 8. *Consider a test* $P_{Y|X}$ *that satisfies at least one of the conditions in* (27)*. Hence, the following statements are equivalent:*

- *Given the output empirical distribution* $\bar{P}_Y^{(n)}$ *obtained from the data* $\boldsymbol{y}$ *as in* (3)*, any probability distribution* $\hat{P}_X^{(n)}$ *on* $\{I, S\}$ *satisfies the equality in* (11)*;*
- *Two random variables* X *and* Y*, in which the joint probability distribution* P_{XY} *satisfies* (1)*, have zero mutual information; and*
- *Two random variables* X *and* Y*, in which the joint probability distribution* P_{XY} *satisfies* (1)*, are independent.*

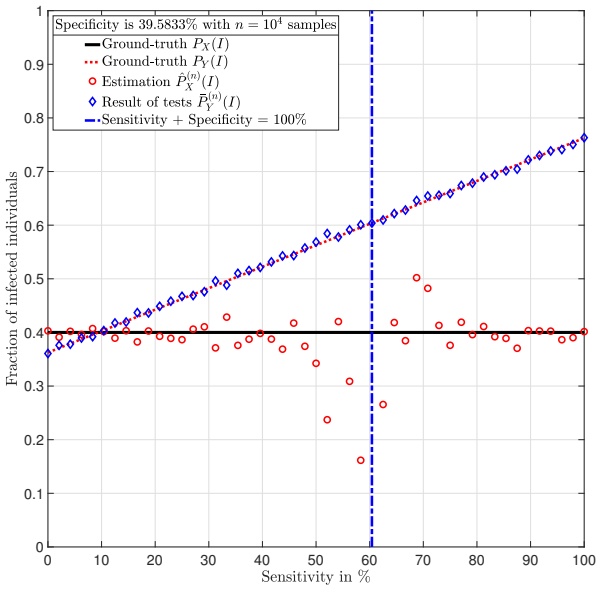
Proof. The first statement is a consequence of Lemma 7; the second statement follows from the fact that under any of the assumptions in (27), the mutual information satisfies

$$\begin{aligned} \mathsf{I}(X;Y) &\triangleq \sum_{x\in\{I,S\}} \sum_{y\in\{I,S\}} P_X(x)P_{Y|X}(y|x) \log_2\left(\frac{P_{Y|X}(y|x)}{P_Y(y)}\right) && (29) \\ &= 0, && (30) \end{aligned}$$

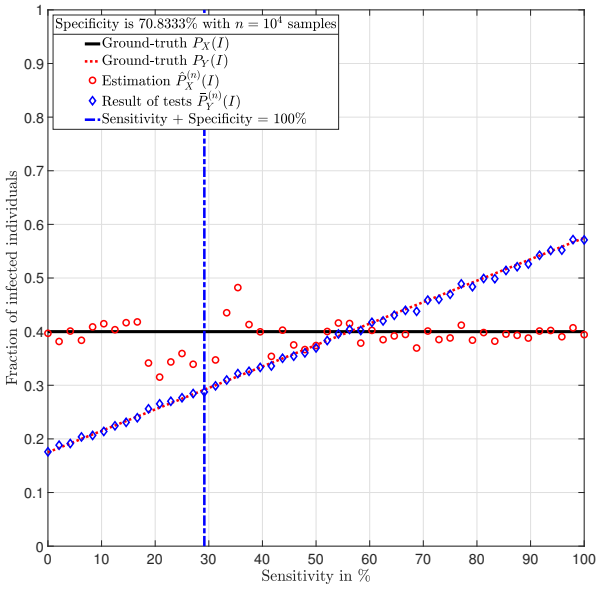
where P_X, $P_{Y|X}$, and P_Y satisfy the equality in (1). The third statement follows from the fact that two random variables are independent if and only if their mutual information is zero. □

Lemma 8 shows that when at least one of the conditions in (27) holds, the output probability distribution P_Y does not provide any information about the input probability distribution P_X. That is, nothing can be said about P_X based on the data $\boldsymbol{y}$.

Despite the singularity, the values of specificity and sensitivity in which the sum is close to one, i.e., around the singularity, are also worthy of discussion. Note that for some $1 > \epsilon > 0$, the absolute 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 observations are justified by the fact that the total variation $\left\|P_X - \hat{P}_X^{(n)}\right\|_{\text{TV}}$ is equal to $\left\|P_Y - \bar{P}_Y^{(n)}\right\|_{\text{TV}}$ up to a constant factor, as shown in Lemma 3. Such a factor is indeed $\frac{1}{\left|1-P_{Y|X}(I|I)-P_{Y|X}(S|S)\right|}$, and thus, larger errors are expected around the singularity for the same finite numbers of tests n. This is evident in the numerical analysis. In Example 1, i.e., Figures 3 and 4, around the singularity, the estimations $\hat{P}_X^{(n)}$ of P_X appear more disperse than the estimations in Example 3, i.e., Figures 7 and 8.

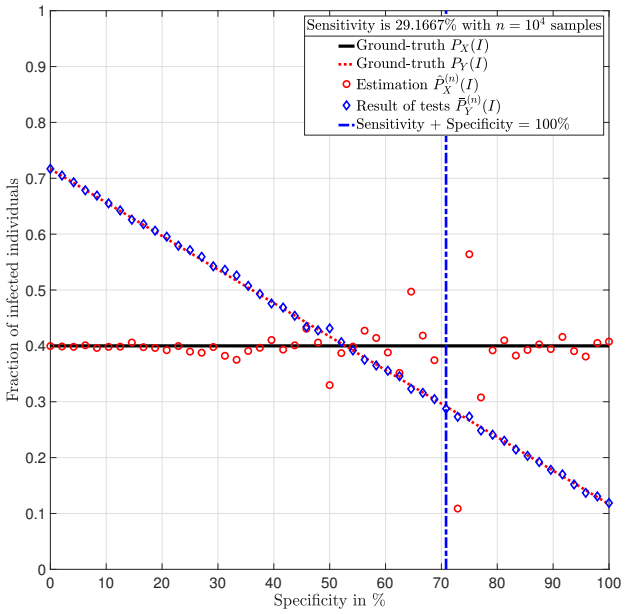


(a) Specificity is 39.6 %

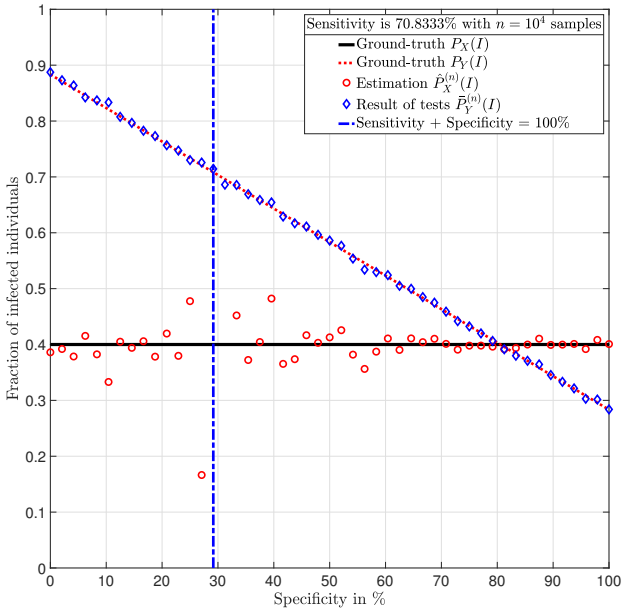


(b) Specificity is 70.8 %

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).

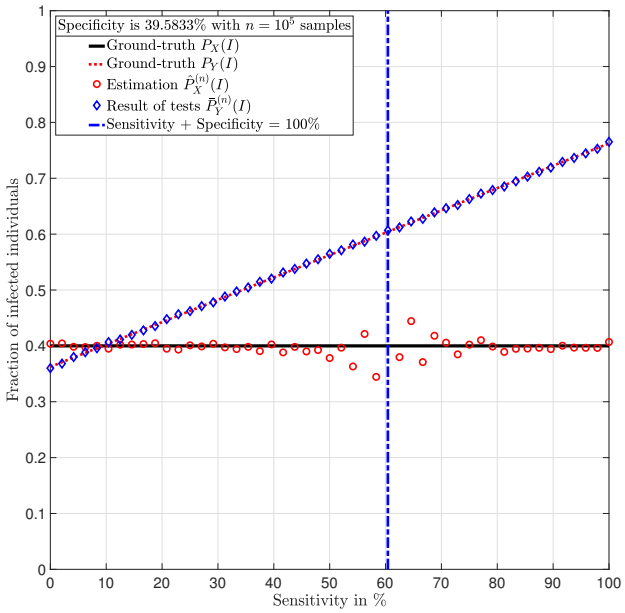


(a) Sensitivity is 29.2 %

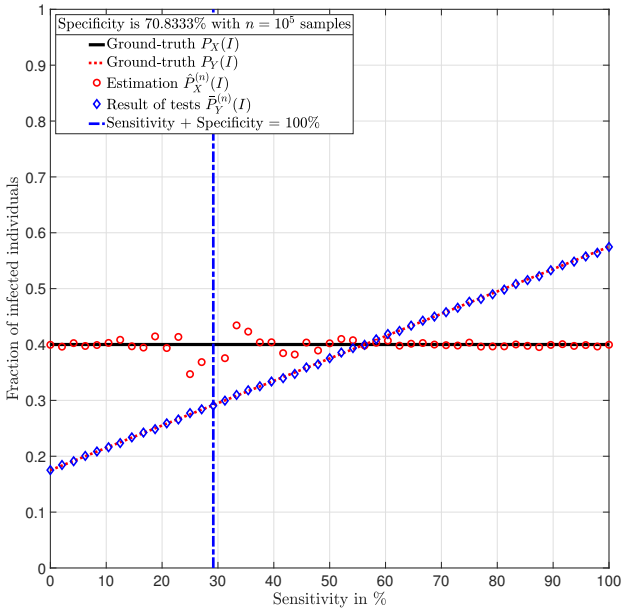


(b) Sensitivity is 70.8 %

Figure 4: 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).

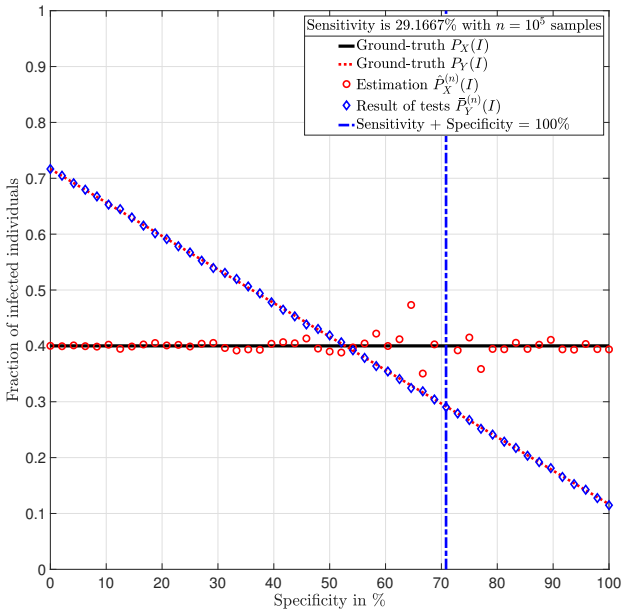


(a) Specificity is 39.6 %

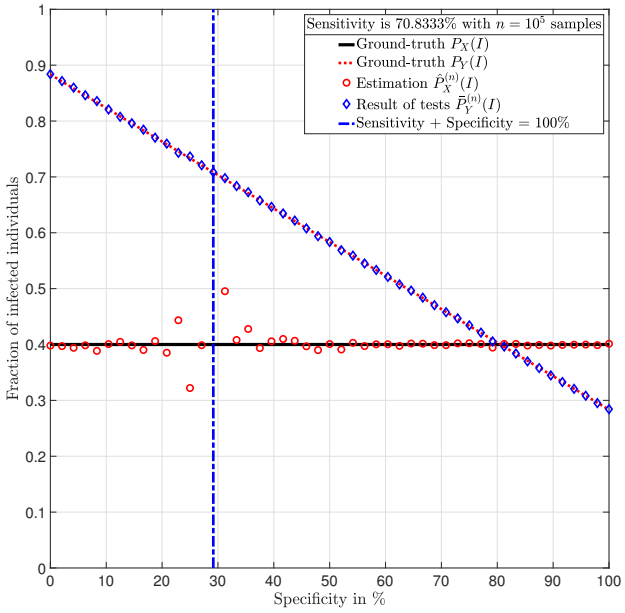


(b) Specificity is 70.8 %

Figure 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).

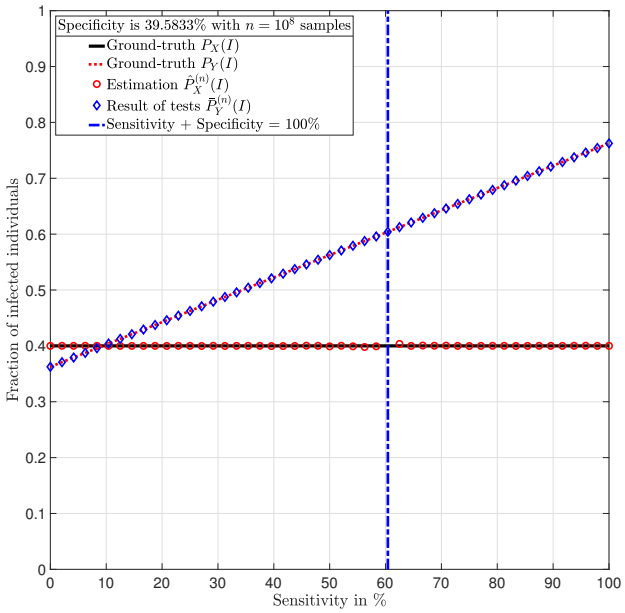


(a) Sensitivity is 29.2 %

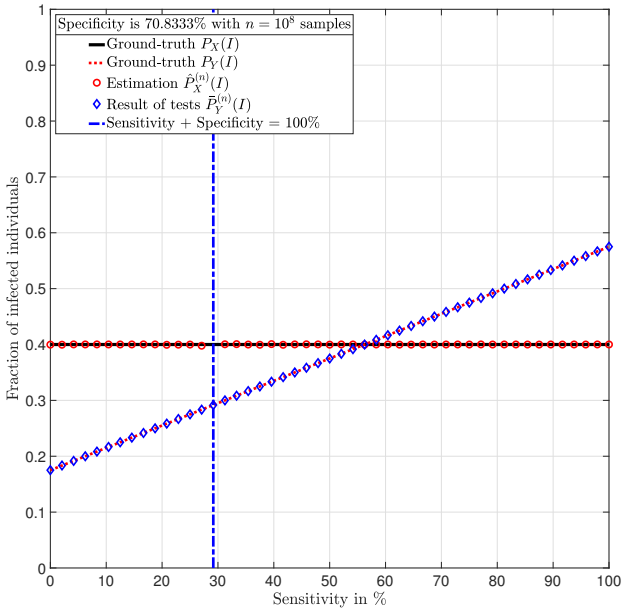


(b) Sensitivity is 70.8 %

Figure 6: Population in which the fraction of individuals infected with SARS-CoV-2 is $P_X = 0.4$ and $n = 100,000$ individuals are tested (Example 2).

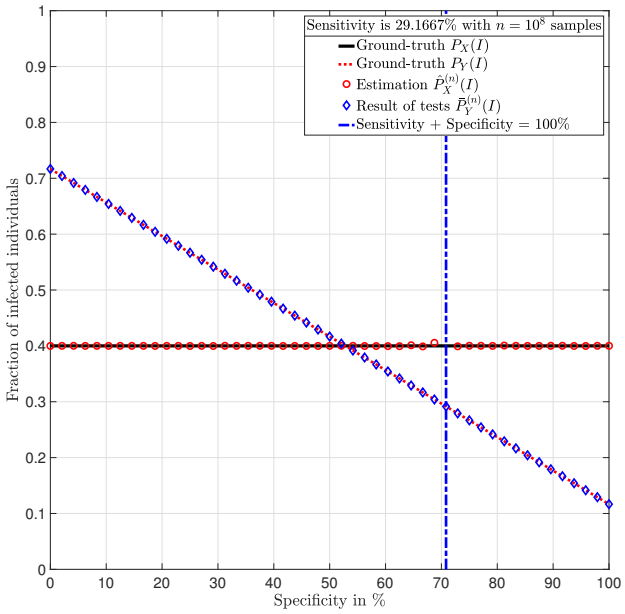


(a) Specificity is 39.6 %

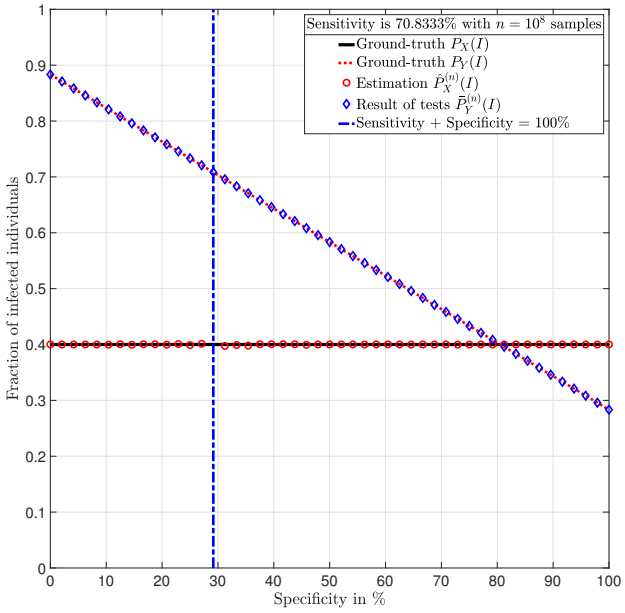


(b) Specificity is 70.8 %

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).



(a) Sensitivity is 29.2 %



(b) Sensitivity is 70.8 %

Figure 8: 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).

5.3 Impact of the Number of Tests.

Figure 3–8 show that when the parameters of the test satisfy at least one of the conditions in (12) and there exist a sufficiently large number of test results, it is always possible to obtain an estimation $\hat{P}_X^{(n)}(I)$ of the prevalence ratio $P_X(I)$. This is independent of the exact values of the specificity and sensitivity as long as (12) holds. More importantly, the reliability of such estimation increases with the number of test results. For instance, compare the estimations in Examples 1 and 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 prevalence ratio, the quality of the tests is not important. 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 error, when a large number of tests is performed. Nonetheless, when a low number of tests is available, it is worth noting that when the sensitivity $P_{Y|X}(I|I)$ and specificity $P_{Y|X}(S|S)$ satisfy $\left|1 - P_{Y|X}(S|S) - P_{Y|X}(I|I)\right| > 1 - \epsilon$, for some $0 < \epsilon < 1$, the smaller ϵ, the smaller the estimation error of the prevalence ratio, c.f., Lemma 3. This observation is of paramount importance as it implies that smaller estimation errors are observed when the sum of the sensitivity and specificity is bounded away from one. This said, the key parameter for reducing the estimation error is the number of tests.

6 Conclusions

In this work, it has been shown that estimating the prevalence ratio of a condition, for example, a SARS-CoV-19 infection, by the ratio between the number of positive test results and the total number of tests leads to excessive estimation errors when tests are unreliable. This is simply due to the fact that unreliable tests, i.e., tests in which probabilities of false positives and false negatives are nonzero, lead to some individuals exhibiting the condition to observe negative test results (false negatives), and some individuals who do not exhibit the condition to observe positive results (false positives). From this perspective, an estimation of the prevalence ratio using data obtained from tests must take into account both the sensitivity and the specificity of the tests. Theorem 1 provides an estimation of the prevalence ratio with an estimation error that decreases with 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, always allow a reliable estimation of the prevalence ratio (Lemma 1 in Section 4) subject to a sufficiently large number of individuals being tested. Alternatively, testing campaigns using tests 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 prevalence ratio even with infinitely many tests (Lemma 7 in Section 4).

A final conclusion is that for estimating the prevalence ratio of a given condition, i.e., a SARS-CoV-2 infection, 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 is different than one, 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 (11) 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* [35].

7.2 Tests with Unknown Parameters

One of the assumptions adopted for developing the results of this report 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* [36] might play an essential role.

References

[1] F. Brauer, C. Castillo-Chávez, and Z. Feng, *Mathematical Models in Epidemiology*, 1st ed. New York, NY, USA: Springer, 2019.

[2] K. J. Rothman and S. Greenland, *Modern epidemiology*, 3rd ed. Philadelphia, PA, USA: Lippincott Williams & Wilkins, 2008.

[3] 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*, pp. 1–5, Apr. 2020.

[4] 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.

[5] 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*, p. 100130, Apr. 2020.

[6] M. Hickman and C. Taylor, "Indirect methods to estimate prevalence," in *Epidemiology of Drug Abuse*, Z. Sloboda, Ed. Boston, MA, USA: Springer, 2005, ch. 8, pp. 113–131.

[7] M. Staquet, M. Rozencweig, Y. J. Lee, and F. M. Muggia, "Methodology for the assessment of new dichotomous diagnostic tests," *Journal of Chronic Diseases*, vol. 34, no. 12, pp. 599 – 610, Dec. 1981.

[8] P. Diggle, "Estimating prevalence using an imperfect test," *Epidemiology Research International*, vol. 2011, pp. 1–6, Oct. 2011.

[9] S. Lachish, A. M. Gopalaswamy, S. C. L. Knowles, and B. C. Sheldon, "Site-occupancy modelling as a novel framework for assessing test sensitivity and estimating wildlife disease prevalence from imperfect diagnostic tests," *Methods in Ecology and Evolution*, vol. 3, no. 2, pp. 339–348, Apr. 2012.

[10] R. M. Cannon, "Sense and sensitivity–designing surveys based on an imperfect test," *Preventive veterinary medicine*, vol. 49, no. 3–4, pp. 141–163, May 2001.

[11] T. Skov, J. Deddens, M. Petersen, and L. Endahl, "Prevalence proportion ratios: Estimation and hypothesis testing," *International Journal of Epidemiology*, vol. 27, no. 1, pp. 91– 95, Mar. 1998.

[12] A. D. Penman and W. D. Johnson, "Complementary log-log regression for the estimation of covariate-adjusted prevalence ratios in the analysis of data from cross-sectional studies," *Biometrical journal*, vol. 51, no. 3, pp. 433– 442, Jun. 2009.

[13] W. Chen, L. Qian, J. Shi, and M. Franklin, "Comparing performance between log-binomial and robust Poisson regression models for estimating risk ratios under model misspecification," *BMC Medical Research Methodology*, vol. 18, no. 1, pp. 1–12, 2018.

[14] M. R. Petersen and J. A. Deddens, "A comparison of two methods for estimating prevalence ratios," *BMC Medical Research Methodology*, vol. 8, no. 9, pp. 1– 9, Feb. 2008.

[15] 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.

[16] 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*, pp. 1–17, Feb. 2020.

[17] 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.

[18] 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*, pp. 1–14, Apr. 2020.

[19] 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, pp. 139–1142, Aug. 2020.

[20] 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, pp. 1–31, Apr. 2020.

[21] 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*, pp. 1–7, Apr. 2020.

[22] 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, p. 104423, Jul. 2020.

[23] 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*, pp. 1–20, May 2020.

[24] 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*, pp. 1–23, Apr. 2020.

[25] 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.

[26] 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.

[27] 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.

[28] 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*, p. 200823, Mar. 2020.

[29] 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*, p. 200642, Feb. 2020.

[30] 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.

[31] 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.

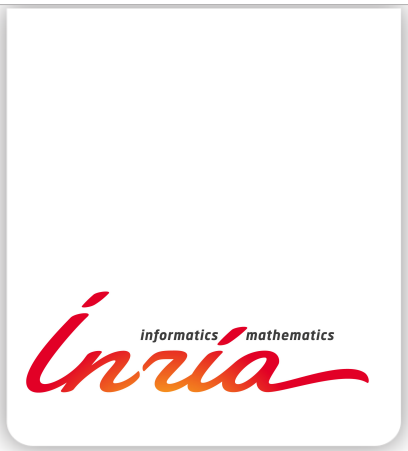
[32] A. De, M. Saks, and S. Tang, "Noisy population recovery in polynomial time," in *Proc. of the 57th Annual Symposium on Foundations of Computer Science (FOCS)*, New Brunswick, NJ, USA, Oct. 2016, pp. 675–684.

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

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

[35] 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.

[36] 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