Group testing as a strategy for COVID-19 epidemiological monitoring and community surveillance

Vincent Brault [1,4,☯], Bastien Mallein [2,4,☯], Jean-François Rupprecht [3,4,☯*]

1 Université Grenoble Alpes, CNRS, Grenoble INP, LJK, 38000 Grenoble, France,
2 Université Sorbonne Paris Nord, LAGA, UMR 7539, F-93430, Villetaneuse, France.
3 Aix Marseille Univ, CNRS, Centre de Physique Théorique, Turing Center for Living Systems, Marseille, France.
4 Members of the GROUPOOL & MODCOV19 initiatives.

☯These authors contributed equally to this work.
* rupprecht@cpt.univ-mrs.fr

Abstract

We propose an analysis and applications of sample pooling to the epidemiologic monitoring of COVID-19. We first introduce a model of the RT-qPCR process used to test for the presence of virus in a sample and construct a statistical model for the viral load in a typical infected individual inspired by large-scale clinical datasets. We present an application of group testing for the prevention of epidemic outbreak in closed connected communities. We then propose a method for the measure of the prevalence in a population taking into account the increased number of false negatives associated with the group testing method.

Author summary

Sample pooling consists in combining samples from multiple individuals into a single pool that is then tested using a unique test-kit. A positive test means that at least one individual within the pool is infected. Sample pooling could provide the means for rapid and massive testing for the presence of SARS-CoV2 among asymptomatic individuals. Here, we do not address any diagnostic problems - e.g. how to use a minimal number of tests to obtain an individual diagnostic - but rather focus on population-scale application of pooling. We first quantify the reduction of test sensitivity due to sample dilution and quantify the efficiency of large pools in (i) obtaining precise estimates of the proportion of infected individuals in the general population at reduced costs and (ii) implementing regular large-scale screenings beneficial in the early detection of epidemic outbreaks within communities (e.g. nursing homes or university campuses).

Introduction

Testing aims at revealing the presence of viral load of SARS-CoV-2 within infected individuals [1, 2]. At date, the most standard mean to reveal such viral load remains the *reverse transcription quantitative polymerase chain reaction* (RT-qPCR) tests [3]. Bottlenecks in the production of reactants used in RT-qPCR diagnostic testing [4, 5] contributed to the development of alternative techniques that provide a more rapid

diagnostic, e.g. lateral-flow antigen and RT-LAMP tests, yet at the expense of a reduced sensitivity compared to RT-qPCR tests.

In the wake of a COVID-19 second wave in Europe and at the current date , several countries have been implementing or are actively considering the implementation of *massive testing*, e.g. Slovakia, whereby repeated nation-wide screenings based on antigen tests occured on week-ends in November 2020 [6]; the Duchy of Luxembourg, whereby a nation-wide screening based on RT-qPCR tests is scheduled for Spring 2021 [7]; the city of Liverpool (United Kingdom) with operation Moonshot consisting in repeated city-wide screenings using a combination of testing strategies.

As COVID-19 infected individuals may be contagious without showing symptoms, tracing is particularly challenging; while individuals showing no symptoms throughout the infection appear to account for only 15% of infections [13–15]), pre-symptomatic infections appear to cause around 50% of infections, approximatively [8–10, 12].

Large-scale testing programs aim at addressing such challenge by allowing an earlier identification of asymptomatic and pre-symptomatic carriers [11]. In China, city-wide testing programs were reported in several cities including Wuhan (May 2020) [16] and Qingdao (October 2020) [17]. These cities relied on a technique called *sample pooling*, equivalently called *group testing*. The principle of group testing consists in combining samples from multiple individuals into a single pool that is then tested using a single test - which, in the COVID-19 context, amounts to using a single RT-PCR well and reactive kit. The pool sample is considered to be positive if and only if at least one individual in the group is infected.

Group testing has a long history that dates back to the seminal work by R. Dorfman in 1943 [18] in the context of syphilis detection, see [19] for a review.

Several teams across the world have developed group testing protocols for SARS-CoV-2 infected individuals using RT-qPCR tests. As early as February 2020, pools of 10 have been used over 2740 patients to detect 2 positive patients over the San Francisco Bay in California [20]. Late April, a report from Saarland University, Germany, indicated that positive sample with a relatively mild viral load from asymptomatic patients could still be detected within pools of 30 [21]. Further works suggest that RT-qPCR viral detection can been achieved in pools with a number of samples ranging from 5 to 64 [22–36].

In parallel, the theoretical literature on group testing for SARS-CoV-2 diagnostic is growing at a fast pace [4,37–42]. Most of the emphasis has been put on the binary (positive or negative) outcome of tests, with little or no regard on the viral load quantification [3]. Moreover, if the possibility of false negatives is sometimes considered, the increase in the rate of false negatives with dilution of samples due to group testing is not often taken into account [43].

In this article, we do not address any diagnostic problems, such as the question of determining optimal strategies to provide individual positive diagnostic using a minimal number of tests as solved by hypercube methods [23, 26, 43–45]) and the P-Best algorithm [24].

We rather propose to evaluate pooling strategies in the non-diagnostic contexts of *screening* and *surveillance*, as defined by the Centers of Disease Control (CDC, USA) terminology [46].

In Section II, we propose a group testing protocol that aims at the early detection of an epidemic outbreak in a closed community, such as nursing homes or universities. In the context surveillance, the size of pools is mainly determined by the maximal tolerated sensibility loss induced by the sample pooling process; the optimal pool size predicted according to a diagnostic criteria addressing the minimizing number of diagnostic is ill defined in such context, and does not provide an adapted answer to the question of determining whether a disease is present or not absent in a community.

In Section III, we provide a mathematical formula to estimate the viral prevalence (i.e.

the fraction of positive individuals among the tested population) based on pool results. Indeed, the laboratory in charge of the screening program may not have access to the subsequent diagnosis results from positive pools; within the American CDC surveillance protocol, non-clinical (e.g. veterinary) laboratories are allowed to perform pooling of samples for surveillance screenings but individual diagnosis is to be performed by a clinical laboratory abiding by the Clinical Laboratory Improvement Amendments (CLIA) rules. Some individuals may also refuse to comply to the diagnosis tests. In addition, diagnostic tests performed on positive pools may also turn negative [47]; we see at least 3 possible reasons for such discrepancies: (i) an inherent false-negative risk in diagnostic tests, (ii) a possible time delay between the screening and diagnosis tests (e.g. that could result in positive individuals in the pool turning negative in the diagnostic test) or (iii) the fact that the screening and diagnostic tests may not rely on the same sample collection - with screenings relying on self-collected nasal swabs or saliva collection, while diagnosis tests are most often performed on nasopharyngeal swabs samples with an arguably higher level of sensitivity.

We find that large pools are extremely efficient at estimating the prevalence. Such estimates could serve as a metric to scale prevention measures within a predefined graded response scheme - e.g. in a college university campus, the decision to switch to remote teaching could be triggered once a critical measured prevalence is reached. Surveillance protocols based on sample pooling have indeed been implemented in several universities across the United States, including Duke University [48] and the State University of New York [49]. Similar protocols have been defined for regular surveillance at Liège University (Belgium), as well as at Nottingham and Cambridge universities (United Kingdom) [50]; in the latter, samples are pooled by dormitories; if a pool turns positive, all individuals are requested to undergo isolation as potential case contact; a second diagnostic test is then performed to find the infected individuals [50].

Both Section II and Section III rely on a realistic models for the risk of false negatives induced by sample pooling. Estimating such risk is the objective of Section I, whereby we provide a short description of the RT-qPCR and a statistical model for its study. In Section I.2, we analyse the distribution of viral loads among a series of clinical datasets to estimate the averaged false-negative rates induced by the sample pooling process, assuming a linear dilution and a fixed positivity cycle threshold.

Results

I Models for sample pooling in RT-qPCR test

We present a mathematical model of the RT-qPCR test as well as a new censored-Gaussian method to fit distributions of viral load in the population. We apply our results to the estimation of the increased risk of false-negatives due to dilution.

I.1 Statistical model for the cycle threshold value (fixed a given viral concentration in the sample)

The RT-qPCR technique is a routine laboratory technique used to estimate the concentration of viral material in samples [51]. A qPCR machine typically returns a C_t value, which corresponds to $-\log_2$ of the initial number of DNA copies in the sample, up to an additive constant and measure error. It is measured as an estimated number of cycles needed for the intensity of the fluorescence of the sample to reach a target value (see Fig. 1). Combined measures of two viral RNA strands are also recommended [3]. Here we focus on a single RNA strand detection and we do not model here the possible errors at the reverse transcription stage, which could lead to some biased measure of the viral

Figure 1. (a) Sketch of an RT-qPCR fluorescence intensity signal for a positive patient without pooling (solid red line) a single positive patient in a pool of 64 patients (dashed red curve) and for a negative sample representing the response of an artefact (dotted magenta curve); as pooling dilutes the initial concentration, the pooled response (dashed red curve) is expected to be close to the translation $x \to x + 6$ from that of a single patient (solid red line). (b) Sketch of the distribution of threshold values for qPCR, either for individual testing (solid blue line) or in pools 64 (dashed red curve); part of the distribution crosses the limit of detection of the test (figured as the grey area) at the detection threshold d_{cens}.

load distribution. Depending on the RTqPCR device, the C_t value of the sample can shift by an additive constant; such constant can be estimated by measuring the C_t value of a standard solution of viral DNA to tare the measure. In any cases, some RTqPCR device might allow the detection of lower viral loads than others.

I.1.1 Model of the cycle threshold values for an individual sample

RT-qPCR tests are prone to amplify non-specific DNA sequences [51, 52] that can trigger an onset of fluorescence in a samples with no viral SARS-CoV-2 load. The fact that such spurious onset of fluorescence typically occurs beyond a relatively large critical number of cycles imposes the following condition on the diagnosis to minimize the risk of *false positives*: a reliable positive result can only be made if the C_t value is lower than a critical value, denoted d_{cens}. Here, the onset of fluorescence from virus-free samples will be modelled as if triggered by a vanishingly small artificial concentration, denoted ϵ_1.

We propose to model the number of cycles threshold value C_t as a random variable, denoted by Y, that depends on the viral load c in the measured sample as

$$Y = -\log_2(c + \epsilon_1) + \epsilon_2, \tag{1}$$

where we assume that (i) the risk of non-specific amplification (false-positive) ϵ_1 as log-normal distribution with parameters (ν, τ^2); and that (ii) the intrinsic variability in C_t measurement ϵ_2 is a centered Gaussian random variable with variance ρ^2.

As mentioned above, tests are considered to be reliably positive when $Y \leq d_{\text{cens}}$. To minimize the risk of false positives, the threshold d_{cens} (with cens for censoring) is chosen such that $\mathbb{P}(\epsilon_1 > 2^{-d_{\text{cens}}}) \ll 1$. Thus, using that as long as a and b are of different orders of magnitude, we have $\log(a + b) \approx \log(\max(a, b))$, we deduce that

$$Y \approx \min(-\log_2(c), d_{\text{cens}}) + \epsilon_2, \tag{2}$$

which obeys the law of a Gaussian random variable with variance ρ^2 and mean $-\log_2(c)$, censored at d_{cens}.

In the no false-positive risk limit ($\epsilon_1 \to 0$), the RT-qPCR threshold intensity of a negative patient ($c = 0$) would never be reached ($Y \to \infty$ as well as $d_{\text{cens}} = \infty$).

I.1.2 Model of the cycle threshold values for pooled samples

We now consider what happens when constructing a pooled sample of N samples. For each $i \leq N$, we write $Z_i = 1$ if the sample i contains a viral RNA load with concentration $C_i > 0$, and $Z_i = C_i = 0$ otherwise. In the rest of the paper, we assume that, in a combined sample created from a homogeneous mixing of the individual samples, the viral concentration reads:

$$C^{(N)} = \frac{1}{N} \sum_{j=1}^{N} Z_j C_j. \tag{3}$$

This assumption relies on the fact that infected individuals should have a sufficiently high number of viral copies per sample, so that taking a portion $1/N$ of a virus bearing sample brings a fraction $1/N$ of its viral charge. The result of the RT-qPCR measure of a grouped test with N individuals is then given by Eq. (1), with $c = C^{(N)}$, hence reads

$$Y^{(N)} = \min \left[\log_2 N - \log_2 \left(\sum_{j=1}^{N} Z_j C_j \right), d_{\text{cens}} \right] + \epsilon_2. \tag{4}$$

where $(Z_i, i \leq N)$ are i.i.d. Bernoulli random variables whose parameter is the prevalence of the disease in the population; $(C_i, i \leq N)$ are i.i.d. random variables corresponding to the law of the viral concentration within samples taken from a typical infected individual in the overall population.

Our model Eq. (4) is consistent with the experimental result of [22] as well as [33], whereby linear relations are found between the logarithm of the pool size and the measured C_t that are sufficiently distant from the identified detection threshold.

Remark I.1. If it were possible to combine samples without dilution (e.g. following the protocol of [34], whereby the exact same volume of each sample is added to the buffer solution as if the sample were being tested individually), Eq. (4) would then be replaced by

$$Y^{(N)} = \min \left[-\log_2 \left(\sum_{j=1}^{N} Z_j C_j \right), d_{\text{cens}} \right] + \epsilon_2, \tag{5}$$

in which case, theoretically, pool testing would never loose precision when the pool size increases. However, if the dilution effect occurs for pool sizes exceeding a threshold size K, Eq. (4) would be replaced by

$$Y^{(N)} = \min \left[\log_2 (\frac{N}{K})_+ - \log_2 \left(\sum_{j=1}^{N} Z_j C_j \right), d_{\text{cens}} \right] + \epsilon_2; \tag{6}$$

where $\log_2(N/K)_+ = 0$ if $N < K$ and $\log_2(N/K))$ otherwise; the analysis would then be similar to what is presented in the rest of the paper, yet with a lower false negative rate.

Remark I.2. We expect the RT-qPCR result to correspond to the sample with the highest viral load, up to a dilution-induced drift $\log_2(N)$, under the model hypothesis of Sec. I (cf. Fig. 1). Indeed, since the viral concentration in randomly selected infected individuals spans several order of magnitudes, we expect that

$$\log_2 \left(\frac{1}{N} \sum_{i=1,\dots,j} C_i \right) \approx \log_2 \left(\max_{i=1,\dots,j} C_i \right) - \log_2(N), \tag{7}$$

for j positive samples with concentration C_j diluted in a pool of N. In contrast with [53], we find, based Eq. (7), that the measured value of the pooled sample viral concentration cannot be used to estimate the number of infected individual within the pool. However, we point out that the RT-qPCR viral load measure could be used to improve efficiency and cross testing of smart pooling type diagnostic methods, which are beyond the scope of this paper. We plan to investigate this aspect in future work.

In order to determine the statistics of the measured cycle $Y^{(N)}$ in a group test of N individuals, we need a distribution for the value of C_j, the viral distribution of infected individuals in the population; this is the objective of the next section.

I.2 Statistical analysis of the population-level viral load

In this section, we model the C_t distributions extracted from a set of clinical datasets.

I.2.1 Clinical datasets

Here we considered four studies in our analysis:

1. The ImpactSaliva dataset [54] providing raw Ct measures from saliva samples (Yale University, USA) with the N1 gene as a target. A set of raw C_t data from $N = 180$ individuals is provided, including 45 C_ts values beyond the positivity threshold set at $C_t = 38$.

2. A dataset constructed on a histogram from Lennon et al. [55] based on 2179 nasopharyngeal samples from residents and staff members of nursing homes during the April to May 2020 period; the N1 gene was used as a target (Massachusetts,USA). Up to date and to the best of our knowledge, Lennon et al. [55] is the largest to study date to present separated C_t histograms between symptomatic ($N = 739$) and asymptomatic ($N = 1440$) individuals at the time of testing.

3. A dataset constructed on a histogram from Jones et al. [56] based on $N = 3598$ nasopharyngeal swabs samples from individuals with various age at La Charité hospital (Berlin, Germany) during the March to early April 2020 period; two target genes (E gene and RdRp) are mentionned in [1]).

4. A dataset constructed on a histogram from Cabrera et al. [28,57] based on $N = 852$ infected nasopharyngeal swabs samples from residents and staff members of nursing homes in Galicia (Spain) during the March to May 2020 period ; the Open Reading Frame 1b (ORF1b) was used as a target gene.

As the precise distribution of data points within each bars of the histogram are unknown in the datasets 2,3 and 4, we assume that points were distributed uniformly in their histogram bar class. We have verified the robustness of our fit estimator for several distribution of points which lead to consistent values for our model parameters (see SI Sec. 2. B).

I.2.2 Censored Gaussian model fits

As we expect the measure error ϵ_2 of the qPCR to be small with respect to the width of the histogram classes, we set $\rho = 0$ in the rest of the section.

Mixture model The shape of the histograms in Fig. 2 suggest that the law of the viral load should be distributed according to a mixture of three or more Gaussian distributions. We performed fitting using standard Gaussian distributions models which we refer to as the naive model.

However, as the dataset histograms usually exhibit a sudden drop in the number of detected cases around a C_t value denoted d_{att} (e.g. $d_{\text{att}} = 35.6$ for the Jones et al. dataset), that we refer to as the attenuation threshold. We explain these drops by a loss of sensibility of the measure for samples with C_t value between d_{att} and d_{cens} (the limit of detection). We model this loss of sensibility by a fixed probability q of detection above level d_{att}.

Censored models To model a partial lack of detection of low viral load (C_t higher than a threshold d_{att}), we introduce the partially censored Gaussian variable as a building block for the representation of the density of the viral load in infected patients.

We assume that if the sampled C_t value is lower than the attenuation threshold d_{att}; if the value is higher than d_{att}, the sample will be detected with probability q, and its measure will be registered. Otherwise, it will be discarded as a (false) negative, with

Figure 2. (a) Representation of the density for the classical mixing Gaussian model (dashed lines) and the partially censored model (solid lines) each composed as a sum of 3 components for the Gaussian model (orange/green/red dashed lines) and the partially censored model (orange/green/red solid lines); (purple vertical line) location of the threshold $d_{\text{att}} \approx 35.6$. Data based on the histogram presented in [56]. (b) Focus on the false negative region, with the estimated false negative probability in the partially censored model (solid line) due to the defect of detection above the threshold d_{att} (red color filled area). (c) Mixing Gaussian model on the ImpactSaliva dataset presented in [54]. (d-e) Mixing Gaussian model on the ImpactSaliva dataset presented in [55] for (d) asymptomatic and (e) symptomatic individuals at the moment of the test.

probability $1-q$. The parameter q represents the probability of detection of a viral load that falls below the detection threshold of some RT-qPCR measures.

The assumption that the probability of detection only depends on whether the C_t value is higher than a fixed threshold is of course an important simplification, as one would expect lower viral loads to be more difficult to detect than higher ones. However, the simplicity of this model allows us to study it as a three parameters statistical model, and to construct simple estimators for these parameters. Additionally, it fits rather well the available data, and fitting a more complicate censorship model would require a lot of measures of C_t values close to the detection threshold d_{att}. The quantity d_{att} is fixed based on the observed distribution of C_t values in datasets.

To avoid the problem of modelling of the partial censorship, a solution that we implement here as a comparison tool, is *to forget* the values after the threshold and we perform the fit on the *completely censored* model (i.e. with $q = 0$) to the remaining data. See the Method section for further discussion.

Application to the Jones et al. dataset [56] We apply here the statistical analysis described in the previous section to simulated data based on the values for the viral load distribution found in [56] with a mixture model and a censoring threshold $d_{\text{att}} \approx 35.6$ (so the two rightmost bars in the histogram of Fig. 2, that appear much smaller than the nearby values, are supposed to be censored). It is reasonable to assume that the censoring threshold has the same value for each sub-population, as it depends on the test methodology rather than on the tested individuals. In Fig. 2, we represent the histogram with the density for the mixture.

We observe that the separation in sub-populations and the resulting densities are very close to the ones obtained in the naive classical Gaussian mixture model, constructed without taking into account the detection threshold. The principal difference between the naive and censored models consists, for the later, in a larger variance that extends above the threshold. To a lesser extent, the sub-population with a median concentration can also exceed the threshold. It is worth mentioning that as expected, the probability of detection below the threshold value is sensibly the same for all three clusters (around 20%).

As a result, using the computed estimates (see SI. Tab. S3) and the model, we can calculate a theoretical false negative rate, see SI. Eq. [S3]: in this case, the value is approximately 3.8% (represented by the red area on the Fig. 2b); it mostly belongs to the third cluster. Such false-negative estimate remains to be treated with caution.

To validate the censored model, we can verify that if one (i) erases the data to the right of a certain value and (ii) uses the totally censored model on the remaining data, a similar estimate should be obtained for the parameters. We refer to SI. Fig. S7 for the density obtained using the censored mixture estimation with $d_{\text{att}} \approx 35.6,\ 34.4$ and 33.2 (removing the first two, the third, then the fourth rightmost bars in the histogram). We observe that the first and second components are globally unchanged. The mean

and standard deviation of the last component are almost the same for $d_{\text{att}} \approx 34.4$ and $d_{\text{att}} \approx 35.6$ (see SI. Tab. S4); only the proportions naturally decrease with the threshold. On the other hand, the mean moves slightly to the left for $d_{\text{att}} \approx 33.2$; this is due to the fact that we loose the information of the largest bars of this component. It might also be caused by our ignorance of the exact distribution of C_t values within classes of the histogram (we recall that we assume that it is a uniform distribution).

Note that if we were to set the threshold at $d_{\text{att}} \approx 34.4$ as threshold for the partially censored model without erasing data, the optimization procedure `nlm` would not converge. This is further indication that a detection drop happens in the neighbourhood of 35.6.

Application to the other datasets We applied a similar statistical analysis to the other datasets listed in Sec.I.2.1 and consistently found either two to three sub-populations using our algorithm. The estimation obtained for the Gaussian fit of the C_t distribution they obtained is given in SI Tab. S3.

In datasets of smaller size [22, 58], the statistical resolution does not allow us to distinguish between several sub-populations; we rather found that the distribution of C_t corresponds to a single Gaussian with standard deviation σ in the 5 to 6 range.

I.2.3 Interpretation of the Gaussian mixture model

We propose an interpretation of the observed Gaussian decomposition of the log-viral load of individuals based on the viral load temporal evolution within individuals.

A first model for the individual viral load evolution Following [59], we consider a piece-wise linear model for the temporal evolution of the mean detected C_t as a function of time after infection $t > 0$,

$$E[C_t(t)] = \begin{cases} \infty & t \leq t_o \quad \text{(incubation)}, \\ d_{\text{cens}} - \frac{\Delta C_{\max}}{t_p - t_o}\,(t - t_o) & t_o < t \leq t_p, \quad \text{(growth)}, \\ d_{\text{cens}} - \Delta C_{\max} + \frac{\Delta C_{\max}}{t_r - t_p}\,(t - t_p) & t_p < t \leq t_r \quad \text{(decay)}, \\ \infty & t > t_f, \quad \text{(recovered)} \end{cases} \tag{8}$$

In our first model, we only consider asymptomatic individuals, for which we set $\Delta t_{\text{decay}}^{(\text{asymp})} = t_r^{(\text{asymp})} - t_p \approx 7$ days [59]. All other parameters are indicated in Table 1.

Distribution of testing times For individuals that remain asymptomatic throughout the infection, we expect the testing time distribution to be random, which means the viral load distribution should depend strongly on the rate of new infections.

We denote by $G(t)$ the distribution of testing times t after infection; for asymptomatic individuals, we consider an exponential distribution $G(t) \propto \exp(-t/\tau)$; one may assume that $\tau > 0$ can be approximated by the observed decrease rate of the incidence.

We simulate the viral load measured in a population of $N = 4,000$ infected individuals samples at random times t_i, $1 \leq i \leq N$, distributed according to a model distribution of testing times. We further assumed that the measured viral load law follows a Gaussian distribution $C_i = \mathcal{N}_{d_{\text{cens}}}(E[C_t(t_i)], \sigma)$, $1 \leq i \leq N$, where $E[C_t(t_i)]$ is given by Eq. (8), and $\mathcal{N}_{d_{\text{cens}}}$ is a Gaussian variable conditioned to values inferior to a model limit of detection threshold (set to $d_{\text{cens}} = 35$); σ is a constant noise amplitude that models an intrinsic dispersion of the viral load among infected individuals.

Considering the model viral load evolution of Eq. (8) (referred to as Model 1 in Fig. 3), exponential $G(t)$ distribution will fail to account the two Gaussian peaks distribution observed in the asymptomatic dataset reported in [55], see Fig. 3b.

Figure 3. (a) Model 1 for the evolution of the viral load post-infection. (b) Modelled distribution in the infection age at the moment of the test for (red) symptomatic individuals as a Gamma function; (blue-cyan) fully asymptomatic (i.e. throughout the infection) individuals either as (blue) a constant if the new infections rate is a constant with time or (cyan) as an exponential if the rate of new infections decays exponentially with time (characteristic decay time $\tau = 10$ days). (c) In the model 1 context, the distribution of the viral load in asymptomatic individuals is relatively uniform. (d) Model 2 for the evolution of the viral load post-infection distinguishing between symptomatic and asymptomatic (combining [59] and [60]). Parameters estimate are provided in SI. (e) In the model 2 context, the distribution of the viral load in asymptomatic individuals shows 2 peaks at high and low viral loads. (f) The distribution of the viral load in symptomatic individuals is less bimodal than the observed asymptomatic distribution.

Second model for the individual viral load evolution To account for the observed peaks in the viral load distribution, we propose the existence of two flat C_t phase that would correspond to the behaviour of the viral (i) near the peak of viral excretion (ii) during a relatively long late infectious phase. Our piece-wise model then reads:

$$E[C_t(t)] = \begin{cases} \infty & t \leq t_o & \text{(incubation)}, \\ d_{\text{cens}} - \frac{\Delta C_{\max}}{t_{p_1} - t_o}\,(t - t_o) & t_o < t \leq t_{p_1}, & \text{(growth)}, \\ d_{\text{cens}} - \Delta C_{\max} & t_{p_1} < t \leq t_{p_2}, & \text{(peak)}, \\ d_{\text{cens}} - \Delta C_{\max} + \frac{\Delta C_{\max}}{t_r - t_{p_2}}\,(t - t_{p_2}) & t_{p_2} < t \leq t_r & \text{(decay)}, \\ d_{\text{cens}} - \Delta C_{\min} & t_r < t \leq t_f, & \text{(late)}, \\ \infty & t > t_f, & \text{(recovered)} \end{cases} \tag{9}$$

Following [59], we consider different estimates for the decay durations between symptomatic ($\Delta t_{\text{decay}}^{(\text{symp})} = t_r^{(\text{symp}} - t_p \approx 10$ days) and asymptomatic ($\Delta t_{\text{decay}}^{(\text{asymp})} = t_r^{(\text{asymp})} - t_p \approx 7$ days) individuals. We consider the same scaling for the duration of the late infectious phase as the one for the decay time (see parameters in Table 1). The piece-wise model considered in [60] is also nearly flat during the late infectious phase at large time; yet in contrast to Eq. (9), [60] considers a sharp evolution at peak phase.

Results Based on Eq. (9), we find that the viral load distribution for asymptomatic individuals displays two peaks at high and low viral loads, see Fig. 3e - in agreement with our analysis of the dataset Lennon et al. [55], see Fig. 2d. An exponential decrease in the number of new cases favors the proportion of individual at high C_ts, see Fig. 3e. The distribution of the viral load in symptomatic individuals is less bimodal than the observed asymptomatic distribution, in agreement with our analysis of the symptomatic dataset from Lennon et al. [55], see Fig.2e.

In [60], a similar results was obtained; a decrease in the incidence rate is shown to be associated to an increase in the proportion of individuals with high C_t value.

Regarding symptomatic individuals, we assume the distribution of testing time to be modelled as a Gamma distribution $G(t) = \Gamma_{\alpha,\beta}(t)$ with parameters $\alpha = 2$ and $\beta = 3$ day^{-1}, see Fig 3; for simplicity, we consider such distribution to be independent of the epidemic status, although a realistic model could include an additional time delay in getting a test during high incidence phases [61]. Based on both models Eq. (9) and (8), we observe the relatively equally distributed viral load, see Fig. 3f; such distribution is in qualitative agreement with the behaviour observed in Fig. 2e.

Multi-Gaussian expression Here we interpret the observed multi-Gaussian distribution of the viral load in terms of a simple analytical model. In the absence of noise

Table 1. Table of values used in Fig. 3 for our viral load evolution models.

Symbol	Meaning	Date/Value
$\Delta C_{\min}$	C_t difference with threshold of the long time plateau	2
$\Delta C_{\max}$	C_t difference with threshold of the peak plateau	13
t_0	Incubation time	Day 2
$t_p, t_{p_1}, t_{p,2}$	Peak time	Day $5, 5, 7$
t_r	Model 1 - Decay time	Day 11
$t_r^{(\text{symp})}$	Model 2 - Decay time (symptomatic)	Day 11
$t_r^{(\text{asymp})}$	Model 2 - Decay time (asymptomatic)	Day 14
$t_f^{(\text{symp})}$	Model 2 - End of infection time (symptomatic)	Day 16
$t_f^{(\text{asymp})}$	Model 2 - End of infection time (asymptomatic)	Day 20
σ	Noise on the measured C_t	2
$G(t)$	Distribution of testing times after infection	N.A.
τ	Decay time in the rate of new infections	0 or 10 days

$\sigma = 0$, the distribution of viral loads corresponding to Eq. (9) reads, for any $x \leq d_{\text{cens}}$

$$f_{nn}(x) = \left(\int_{t_{p_2}}^{t_{p_1}} G(t)dt\right)\delta(x - \Delta C_{\max}) + \left(\int_{t_r}^{t_f} G(t)dt\right)\delta(x - \Delta C_{\min}) \tag{10}$$

$$+ \left[\frac{G(t_1(x))}{t_{p_1} - t_0} + \frac{G(t_2(x))}{t_p - t_{p_2}}\right] \mathbb{1}(x > C_{\min})\mathbb{1}(x < C_{\max}), \tag{11}$$

where δ is a Dirac delta-function; $t_1(x)$ and $t_2(x)$ are the two dates such that $C_t(t) = x$, when applicable. In the presence of a noise of amplitude $\sigma(x)$, the measured viral load density reads:

$$f(z) = \mathcal{N} \int_{C_{\min}}^{C_{\max}} dx \int_{-\infty}^{\infty} dy \exp\left(-\frac{y^2}{2\sigma(x)^2}\right) f_{nn}(x - y)\delta(z - x - y), \tag{12}$$

with $\mathcal{N}$ a normalization constant. We therefore expect to obtain a multi-Gaussian distribution for the distribution of viral loads, with two weights being proportional to the time spent at the peak viral phase and at minimal elimination phase for the smallest and largest C_t means, respectively .

Generality Our results are robust to reasonable variations of parameters. We expect our interpretation to be robust to a large class of viral load models that exhibit a sharp viral load rise, plateau at a high level and long decay time.

I.3 Estimation of the population-averaged false negative rate induced by pooling

One positive individual within the pool The distribution of the viral load of a single positive sample within a pool of several negative samples appears as shifted towards higher C_t-values, see Fig. 1. A pooled sample returns positive if the average concentration is smaller than d_{att} with probability 1, or if the average concentration is between d_{att} and d_{cens} with probability q; thus using the observation of Sec. I.2.2, infection will be detected in a group of N individuals typically if at least one individual in the group has a viral load larger than $N2^{-d_{\text{cens}}}$. Therefore, there is a risk that low viral load samples (that would have been tested positive using individual tests) would no longer be positive in pool tests. Similarly to Eq. (2), we express the increased rate of false negative

Figure 4. (a-c) Relative increase in the false negative risk $(1-\Phi(d^{(N)}_{\text{cens}}))/(1-\Phi(d^{(1)}_{\text{cens}}))$ in pools of size N including a single infected individual whose viral distribution is estimated using the naive (solid line), partially censored (circle) or fully censored (crossed line) fitting method of the following datasets from (a) Watkins et al. [54], (b) Jones et al. [56] and (b) Lennon et al. [55]. In (a), we superpose the clinical estimation of the risk of false negative provided in [54] (red crosses). Here, in contrast to [26], we do not change the threshold level of positivity compared to the individual test.

due to pooling as $\mathbb{P}(-\log_2(C)+\epsilon_2 \geq d_{\text{cens}}-\log_2(N))+(1-q)\mathbb{P}(d_{\text{att}}-\log_2(N)\leq -\log_2(C)+\epsilon_2 \leq d_{\text{cens}}-\log_2(N))$, where $\log_2(C)$ is the viral concentration of the positive individual. For simplicity we neglect the measurement error of the qPCR, i.e. considering that $\rho=0$, thus an expression for the increased rate of false negatives reads $(1-\Phi(d^{(N)}_{\text{cens}}))+(1-q)(\Phi(d^{(N)}_{\text{cens}})-\Phi(d^{(N)}_{\text{att}})$, where

$$\Phi(z)=\mathbb{P}\left[-\log_2(C)\leq z\right], \tag{13}$$

and $d^{(N)}_{\text{cens}}=d_{\text{cens}}-\log_2(N)$, $d^{(N)}_{\text{att}}=d_{\text{att}}-\log_2(N)$. It is worth noting that a simple upper bound is obtained by setting $q=0$, i.e. considering that the test is systematically negative when the C_t value is larger than d_{att} (or in other words, by setting $d_{\text{cens}}=d_{\text{att}}$). This is the choice made when using the completely censored model, and the formula we will use in the rest of the article for the false negative probability is $1-\Phi(d^{(N)}_{\text{cens}})$.

Discussion The naive and censored fitting models predict lead to two different cumulative distribution expressions Φ. This impact the estimation of the relative false-negative risks defined as $1-\Phi(d^{(N)}_{\text{cens}})/(1-\Phi(d^{(N)}_{\text{cens}}))$

- For the Watkins et al. [54] dataset, Fig. 4a, our estimate of the relative increase in the false negative is consistent with the quantification performed on pools of a single positive samples in pools of 5, 10 and 20.
- For the Lennon et al. [55] dataset we find that false-negative rate is higher for asymptomatic individuals than for symptomatic ones, see Fig. 4, in agreement with a higher proportion of individuals at a low viral load in the former category. In the uncensored model, we make the assumption that the histogram obtained was not subject to any attenuation, while in the censored mode, we consider the sharp drop around $C_t=35.6$ are being caused by false negative results.
- For the Jones et al. dataset [56], we find that, when estimated by the censored model, the false negative risk function $\Phi(d^{(N)}_{\text{cens}})$ grows quicker as the pool size increases than in the uncensored model, see Fig. 4. This is mainly partly due to the fact that the censored model makes the assumption that $d_{\text{cens}}\approx 35.6$, whereas in the uncensored model, the assumption made is that $d_{\text{cens}}\approx 37.3$.

Multiple positive individuals within the pool The case of a pool of N samples that contains $k>1$ positive individual is particularly relevant as pooling may be achieved on individuals living in the same household, as in [47], or students sharing the same residence hall, as mentionned in [62]; in these cases, the fact that one individual is infected increases the probability that more individuals in the pool are infected as well. Such correlation has been clinically found to lower the risk of false negative risk, an effect coined *hitchhiking* in [47]. In SI Sec. II, we provide estimation for the false-negative rate in pools, we expect to depend on the prevalence among the tested individuals.

Choosing a correct statistical model for the distribution of C_t values has a critical impact on the estimation of the false negative risk, but a less critical one on the estimation of the efficiency of screening strategies, as discussed in the next Result section.

II Group testing and epidemic outbreak surveillance

We now consider some applications of group testing to the early detection of an epidemic outbreak within a community (with a total number of individuals denoted A) that is interconnected and reasonably closed to the outside (e.g. schools, nursing homes, detention centres).

Here, we focus on the false negative risk. False positives are also a concern in low prevalence settings whereby positive predictive value might be low. However, positives appear very rare in RT-qPCR tests - with an estimated higher bound at 0.01% [63].

II.1 Risk mitigation from a single pre-symptomatic individual

We first consider the impact of group testing strategy, consisting in k group test with pools of N individuals, on the early time of the outbreak $t \ll \lambda^{-1}$. With a unique infected individual in the population, the detection probability reads

$$\mathbb{P}\left[+|k \text{ tests}\right] = kN\Phi_0(d_{\text{cens}}^{(N)})/A, \quad \text{with } kN \leq A, \tag{14}$$

where $\Phi_0(d_{\text{cens}}^{(N)})$ is defined according to Eq. (13), with the difference that the assumed viral load of the patient 0, corresponding to that measured at early times, may need not be equal to the distribution estimated in Eq. (21) based on clinical data. For simplicity, we will assume in the following that Φ_0 is the cumulative distribution of a log-normal viral load distribution $\log\mathrm{N}(\mu_0, \sigma_0)$ of mean μ_0 and variance σ_0.

We first consider of a patient 0 with a weak viral load ($\mu_0 \approx 30$), see Fig. 5a. Such low viral load can model the case of a presymptomatic individual, e.g. with a testing time distribution $G(t)$ distributed in the $t = t_0$ to $t = t_p - 2$ days. In Fig. 5, we represent the evolution of the probability to detect the patient 0 as a function of the total number of sampled individuals in a population of size $A = 120$. We observe that if μ_0 is close enough to d_{cens}, i.e. if the viral load of the patient 0 is close to being undetectable, then there will exist an optimal size for the pools. When N becomes too large the risk of false negative overcomes the potential benefits of testing larger portions of the community (see Fig. 5a). In contrast, if the viral load of patient 0 is slightly higher, the detection probability becomes a monotonic function of the pool size N, indicating that larger pools are always beneficial. Additionally, if using multiple tests increases the detection probability when the viral load is close to the detection threshold, using multiple tests has a smaller impact when the viral load gets easier to detect.

Here we first considered the case of a patient 0 with a weak viral load; however, [64] indicates that a large fraction of presymptomatic individuals detected in a nursing homes had relatively high viral loads (wuth a mean C_t in the $\mu_0 \approx 20 - 25$ range), which tends to indicate that screening methods based on pooling would be even more efficient than suggested in Fig. 5a. We then considered the case of single individual with a viral load distributed according to the fits of the datasets Lennon et al. [55] and Jones et al. [56]; in these instances too, we find that no optimum exists for these two viral load distributions and that large pool sizes are always optimal, see Fig. 5b.

II.2 Risk mitigation from a cluster of infected individuals

We now consider an epidemic outbreak involving a number Q of infected individuals within a community campus of size $A = 4,000$ at the day T of a screening program.

Figure 5. Detection probability within a community of 120 as a function of the total number of sampled individuals $M = k \times N$, where k is the total number of tests used and N the number of samples pooled together in a test (a) Case of a single patient 0 with low viral load; $k = 5$ (red dotted line); $k = 4$ (orange line with arrow), $k = 3$ (purple line with circles); $k = 2$ (dashed green line); $k = 1$ (solid blue line) for several values in the parameters describing the viral load of the patient 0 at the onset of contagiosity, expressed in terms of a normal distribution in C_t (the number of RT-qPCR amplification cycles) with a standard deviation σ_0 and a mean μ_0 and a threshold at a value denoted d_{cens} satisfying: $\mu_0 = d_{\text{cens}} - 1$ (top row), modelling a patient 0 with a very low viral concentration, $\mu_0 = d_{\text{cens}} - 3$ (middle row), $\mu_0 = d_{\text{cens}} - 5$ (bottom row); $\sigma_0 = 2$ (left column); $\sigma_0 = 6$ (right column). (b) Case of a single patient 0 with a viral load distributed datasets(left) for the three fitting methods used to describe the asymptomatic dataset corresponding to Lennon et al. [55], for $k = 1$ (blue) and $k = 5$ (red)and (right) comparing the datasets of Lennon et al. [55] and Jones et al. [56] for the naive fitting method (upper curve $k = 5$, lower curve $k = 1$).

Table 2. Table with standard parameter values (with std. the abbreviation of standard deviation).

Symbol	Meaning	Value
d_{cens}	Maximal cycle number	SI Table 4-5,7-9
μ_i, σ_i, p_i	Viral load (in C_t) distribution fits	SI Table 4-5,7-9
ρ	RT-qPCR measurement error (std.)	0
A	Total number in the community	120 or 4000
N	Pool size	1–128
Q	Threshold number of infected individuals	20
μ_0 ; σ_0	Viral load (in C_t), patient 0 (mean, std.)	$30 - 35$
k	Number of tests used per day	$1 - 5$

Our objective is to find an estimate of the pool size and testing cost that will ensure detection of at least one individual within the cluster within a maximal tolerated number of days denoted D. The probability to detect the outbreak using k pooled tests of size N simply reads:

$$\mathbb{P}(\text{1-day detection}) \approx \left[1 - \binom{A-Q}{Nk} / \binom{A}{Nk}\right] \Phi_0(d_{\text{cens}}^{(N)}). \tag{15}$$

In turn, the probability of detection between Day T and Day $T + D$ then reads:

$$\mathbb{P}(\text{D-day detection}) \approx 1 - (1 - \mathbb{P}(\text{1-day detection}))^D . \tag{16}$$

For simplicity, here we considered that Q remains a constant; we do not model the spread of the infection between the day T and the day $T + D$; such spread would only increase the probability of detection, making Eq. (16) a lower bound estimate. A more elaborated model exploring the question of the optimal testing frequency the presence of an epidemic spread is discussed in SI Sec. IV.

For the surveillance program to be efficient, detection should be highly probable within a time window should be smaller than the typical time scale of apparition of first symptoms within the forming cluster.

We now consider a reasonable order of magnitude estimation of detection probability Eq. (16) with $D = 3$ days and $\Phi_0(d_{\text{cens}}^{(N)})$ estimated using the Jones et al. dataset. With $A = 4,000$ and setting a threshold of $Q = 20$ infected individuals, we find that with

$k = 16$ pools of $N = 16$ individuals - i.e. a total number of $N \times k = 256$ sampled individuals per day - the one-day detection probability Eq. (15) reaches 72 %. The 3-day success probability, as defined through Eq. (16) then reaches 99%. With $k = 4$ pools of $N = 16$ individuals, corresponding to 64 sampled individuals per day, the one-day detection probability of Eq. (15) is only at 27%; the 3-day success probability, as defined through Eq. (16), reaches 62%; yet the 3-day detection probability reaches 85% if the threshold is raised to $Q = 40$ infected individuals.

In the next section, we intend to build an estimator for the prevalence based on the currently available results of pools.

III Measuring the prevalence using group testing

We investigate in this section the measure of the prevalence of the disease in a population using a group testing strategy. We first consider the assumption of *perfect* tests, i.e. with no risks of false negative nor false positive.

III.1 Measuring the prevalence in the absence of false-negatives

We assume that we have n pool tests of size N which allow us to sample, at random, nN individuals within a population. Each of these pools is then tested using the perfect tests. For all $i \leq n$, we write $X_i^{(N)} = 1$ if the ith test is positive (i.e. if and only if at least one of the N individuals in the ith pool is infected), and $X_i^{(N)} = 0$ otherwise. We denote by p the (unknown) proportion of infected individuals in the population. then $(X_i^{(N)}, i \leq n)$ forms an independent and identically distributed (i.i.d.) sequence of Bernoulli random variables with parameter $1 - (1-p)^N$.

Writing $\overline{X}_n^{(N)} = \frac{1}{n}\sum_{j=1}^{N} X_j^{(N)}$, the quantity $1-(1-\overline{X}_n^{(N)})^{1/N}$ is a strongly consistent and asymptotically normal estimator of the p. Following the seminal derivation proposed in [65] (reproduced in SI), one finds that the confidence interval of asymptotic level $1-\alpha$ reads

$$\mathrm{CI}_{1-\alpha}(p) = \left[1 - (1-\overline{X}_n^{(N)})^{1/N} \pm \frac{q_\alpha (1-\overline{X}_n^{(N)})^{1/N-1}\sqrt{\overline{X}_n^{(N)}(1-\overline{X}_n^{(N)})}}{\sqrt{n}N}\right], \qquad (17)$$

where q_α is the quantile of order $1-\alpha/2$ of the standard Gaussian random variable.

The precision of the measure of prevalence decays as $n^{-1/2}$, with a prefactor that depend on the prevalence p and the number N of individual per pool. There exists an optimal choice of N that minimizes the value of this prefactor, largely improving the precision of the measure. Again following [65], one shows that the prefactor in Eq. (S12) is minimal when the number of mixed samples per pool is equal to:

$$N_{\mathrm{opt}}^{(\mathrm{perf})} = -\frac{c_\star}{\log(1-p)} \iff (1-p)^{N_{\mathrm{opt}}^{(\mathrm{perf})}} \approx 0.20, \qquad (18)$$

where $c_\star = 2 + W(-2e^{-2}) \approx 1.59$ and W is the Lambert W function. Specifically, the size of the pools is optimal when approximately 80% of the tests made on the groups turn positive, in sharp contrast with the diagnostic Dorfman criterion [18].

In a recent guideline [66], the European Center of Disease Control presents a seemingly different expression for the prevalence confidence intervals; however, we point out that these estimators for the confidence intervals width become asymptotically equivalent in the limit of a large number of individuals N.

If we measure the prevalence of the population using group testing, choosing $N = N_{\mathrm{opt}}^{(\mathrm{perf})}$ for the size of the groups, then measuring with a given precision the prevalence

Figure 6. (a,b) Total number of tests (red) and total number of sampled individuals (blue) in order to estimate a prevalence of $p = 1\%$ with a $\pm 0.2\%$ precision with 95% confidence interval as a function of the pool size N for the perfect case (dashed lines) with no false negative versus the case with false negatives (solid lines) estimated according to the Lennon et al. asymptomatic dataset [55]. In (a) N ranges from 0 to 25; in (b) N ranges from 0 to 128. The optimal pool size $N_{\rm opt}^{\rm (perf)}$ is beyond the N-axis limit.

will require significantly less tests than if we were to use one test per sampled individual (i.e. if $N = 1$). On the other hand, using this group testing method increases the total number of individuals needed to be sampled, which also has a cost to be considered. However, one can observe that the bottom of the valley of the (red) functions plotted in Fig. 6, that represent the number of tests needed as a function of the size of the pool, is rather wide and flat. There is therefore a large variety of quasi-optimal pool sizes that can be chosen with minimal diminution of the precision in the measure of the prevalence.

In Fig. 6, we consider the case of a prevalence at $p = 1\%$, in which case the the optimal pool size $N = N_{\rm opt}^{\rm (perf)}$ is larger than 255. Choosing a pool size of $N = 20$ requires almost a 100% increase in the total number Nn of sampled individuals but more than a 10 fold decrease in the total number of required tests, see Fig. 6.

In Table 3, for illustrative purposes, we consider another example of a high prevalence setting ($p = 3\%$) in which expect we expect the optimal pool size (minimizing the number of tests needed) to be in the clinically achievable range $N_{\rm opt}^{\rm (perf)} \approx 50$ [22]. At optimality, the number of tests needed is divided by 20 as compared to individual testing. In this case, the total number of individuals that need to be sampled is more than doubled compared to individual testing ($N = 1$), see Fig. 6. Choosing instead a pool size of $N = 20$ requires almost the same number of tests, yet at a cost of almost a 30% increase in the total number Nn of sampled individuals. The same observation holds for different values of the prevalence, see SI. Fig. S13.

Table 3. Table of the pool size as a function of the number of tests for a prevalence of 3% measured with a precision of 0.2% at a 95% confidence interval, for both perfect tests (with no false negatives, see Sec. III.1) and imperfect tests (with false negatives estimated using the Jones et al. dataset; model parameters defined in Table 2); computed using Eqs. [17] and [19].

Pool size N	Perfect tests		Imperfect tests	
	Number of tests n	Sample size nN	Number of tests n	Sample size nN
1	29100	29100	29464	29464
2	14775	29550	15069	30138
3	10003	30009	10261	30783
5	6191	30955	6411	32055
10	3350	33500	3530	35300
20	1973	39460	2130	42600
30	1561	46830	1716	51480
50	1349	67450	1525	76250
100	1884	188400	2235	223500
200	10378	2075600	13105	2621000

III.2 Measuring the prevalence including false negatives

As discussed in Eq. (4), we model the concentration of the pooled sample as the average of the individual sample loads; and we assume that viral concentration becomes undetectable below a given threshold. Therefore, creating groups has the effect of increasing the false negative rate, which has to be quantified. We then use this estimation to un-bias the estimator of the prevalence in the overall population based on group testing, and study its impact on the optimal choice of group sizes.

Assuming a false negative rate of $1-\Phi(d_{\text{cens}}^{(N)})$ in pool testing with groups of size N, we observe that $1-(1-\overline{X}_n^{(N)})^{1/N}$ (as defined using the notation of Section III) is a consistent estimator of $p\Phi(d_{\text{cens}}^{(N)})$. As a result, the confidence interval constructed for the prevalence p now reads

$$\text{CI}_{1-\alpha}(p) = \left[\frac{1-(1-\overline{X}_n^{(N)})^{1/N}}{\Phi(d_{\text{cens}}^{(N)})} \pm \frac{q_\alpha}{\sqrt{n}} \frac{(1-\overline{X}_n^{(N)})^{1/N-1}\sqrt{\overline{X}_n^{(N)}\left(1-\overline{X}_n^{(N)}\right)}}{N\Phi(d_{\text{cens}}^{(N)})} \right]. \quad (19)$$

For the numerical applications presented in Fig. 6 and Table 3, we consider a viral load C that is distributed according to Eq. (21) using the Jones et al. parameter fits. As expected, due to false negatives, we find that the number of tests needed to reach a given precision on the prevalence is increased; however we this increase is moderate.

In particular, the optimal pool size value, $N_{\text{opt}}^{(\text{imper})}$, that minimizes the number of tests needed to reach a given precision level, is close to the value $N_{\text{opt}}^{(\text{perf})}$, defined in Eq. (18).

Similarly, one can observe that using a different distribution with similar mean and variance for $-\log_2 C$ as Eq. (21) would lead to moderate changes of the values estimated in Table 3. While modelling of the viral load of an infected individual is crucial to un-bias the estimator of the prevalence via group testing, the practical implementation of such group testing strategy, i.e. the choice of the group size N and the number n of tests to use, is relatively independent of the precise statistical properties of the viral load distribution. We therefore obtained similar results as for the optimal pool size for the prevalence measurement using the viral distribution extracted from the Lennon et al. and ImpactSaliva datasets.

Based on Eq. (19), in Box 2. we propose an iterative method to estimate p, which, during a survey, allows for on-the-fly adaptations of the pool size.

III.3 Group testing and Bayesian inference of the prevalence in sub-categories of the population

The viral prevalence may vary significantly among specific categories within the overall population. In particular, a prevalence reaching 5% was measured among the health care workers population in a hospital [67], which we expect to be significantly higher than the estimate prevalence within the general population.

Here we show that we do not specifically need pool samples from individuals from homogeneous categories in order to recover the distribution of prevalence within these categories.

The protocol described in Box 2 can be adapted to study different prevalences in specific sub-populations, provided that the number of individuals of each subpopulation is known for every grouped sample. In SI. Fig. S15 , we evaluate, as function of the number of tests, the credibility intervals on the prevalence within two categories of the population: one at $p_1 = 5\%$ representing 20% of the total population (a value inspired

Box 2: A protocol of prevalence determination

We propose the following procedure for the measure of prevalence via group testing:

1. Start from an a priori estimate for the prevalence ($\hat{p}_0$).
2. Based on the value of $\hat{p}_0$, estimate the number N of individuals in the pool that minimizes the total number of tests needed to achieve the estimation of the prevalence p at the targeted precision and confidence interval.
3. Construct a number of n pools containing each N individuals selected at random in the general population, with n the number of tests available for the measure.
4. Count the number of positive tests and compute the average $\overline{X}_n^{(N)}$.
5. An improved estimate of the prevalence then reads: $\hat{p}_1 = 1 - (1 - \overline{X}_n^{(N)})^{1/N}$ (cf Sec. III.1).

Note that this method can easily be adapted into a Bayesian algorithm, with the number N of individuals tested modified at each iteration of the procedure.

by [67]), the other being at $p_2 = 0.5\%$ (a value inspired by [68]). More information on this adaptative protocol can be found in the SI.

Remark III.1. Note that once a difference in prevalence is noted from the epidemiological study of the general population, testing can be adapted to construct groups containing only members of one subpopulation to attain similar levels of precision for the prevalence of the sub-populations. The prevalence in the general population can then be recovered by averaging the estimators of the sub-populations. The advantage of these adaptative settings is that the existence of a difference of prevalence in populations can be tested before deployment of resources needed to measure them specifically.

Discussion

We consider the effect of sample dilution in RT-qPCR grouped tests and we propose a model to describe the risk of false negatives as a function of the pool size. We present a procedure to analyse experimental datasets for the viral load of patients. Inspired by the clinical study [56], we expect the statistics of the number of amplification cycles to be well described as a mixture of 2 to 3 Gaussian variables censored at the RT-qPCR sensibility limit. We interpret this decomposition in terms of a simple model for the evolution in the viral load from samples of infected individuals.

We then considered the interest of group testing methods for large-scale screenings in communities. We have used a minimal set of parameters in order for analytical calculations to be tractable. Including more parameters (e.g. considering a time-dependent infection rate or viral load for patients after their infection, graph of relationship within the community) would be needed in order to obtain conclusive results to be used as healthcare guidelines. In this direction, based on stochastic simulations encompassing a large set of parameters, [69] also concludes on the efficiency of group testing in preventing epidemic outbreaks in health care structures.

Several recent papers indicate RT-qPCR tests based on saliva samples are highly-sensitive [70–75]. Saliva collection appears well accepted by the population techniques [76] while decreasing the cost and risks of sample collection. In this context, saliva sample

pooling, which demonstrates reduced loss of sensitivity even in large pool sizes [54] and has been massively used in the State University of New York, appears as a promising solution for regular large-scale surveillance programs.

Group testing could provide the means for regular and massive screenings allowing the early detection of asymptomatic and pre-symptomatic individuals – a particularly crucial task to succeed in the containment of the epidemic [12, 59, 77]. We expect group testing for SARS-CoV2 to remain relevant throughout the upcoming vaccination era, in particular as a tool to track the evolution of viral variants.

Method

Here we clarify the method used to fit the viral load distribution datasets. We define the *partially censored Gaussian model*, denoted by $\mathcal{CN}_{d_{\text{att}}}(\mu, \sigma, q)$, with μ and σ the mean and standard deviation of the Gaussian variable before censorship and q the detection probability above the threshold. If we denote by X the random variable, $f_{\mu,\sigma}$ (resp. $F_{\mu,\sigma}$) the density (resp. the cumulative distribution function) of a Gaussian law $\mathcal{N}(\mu, \sigma)$ then the density of X is defined for every $x \in \mathbb{R}$ by:

$$f_X(x) = \frac{f_{\mu,\sigma}(x)}{q + (1-q)F_{\mu,\sigma}(d_{\text{att}})} \times \begin{cases} 1 \text{ if } x \leq d_{\text{att}}, \\ q \text{ otherwise.} \end{cases} \tag{20}$$

We also define the *fully censored Gaussian model*, denoted by $\mathcal{CN}_{d_{\text{att}}}(\mu, \sigma) = \mathcal{CN}_{d_{\text{att}}}(\mu, \sigma, 0)$, such that the fitting density is defined for every $x \in \mathbb{R}$ by

$$f_X(x) = \frac{f_{\mu,\sigma}(x)}{F_{\mu,\sigma}(d_{\text{att}})} \mathbb{1}_{\{x \leq d_{\text{att}}\}}, \tag{21}$$

where $\mathbb{1}_{\{x \leq d_{\text{att}}\}}$ is the indicator function equal to 1 if $x \leq d_{\text{att}}$, and 0 otherwise. This analysis allows to test several values of d_{att}, the fact that estimates of μ and σ remain stable for different values of d_{att} justifies the validity of the censored Gaussian mixture model.

Remark III.2. In the absence of censorship (i.e. in the limits $q \to 1$ or $d_{\text{att}} \to +\infty$), we check that Eq. (20) converges to a Gaussian density distribution.

Due to the presence of the cumulative distribution function of a Gaussian law in the denominator in the normalization constant, it is not possible to obtain analytical forms of the parameter estimators. Nevertheless, we can estimate the parameters using an optimization algorithm like the R function nlm (available in [78]) which implements a Newton-type algorithm. In SI, we provide the proof of a theorem that guarantees the quality of our maximal likelihood estimators.

Supporting information

S1 Appendix Supporting analysis and proofs document.

S2 Code Codes used in the paper.

S3 Appendix Excel sheet with numerical values that were used to generate viral load histograms.

Acknowledgements

We wish to thank members of the MODCOV initiative and in particular Françoise Praz and Florence Debarre who gave us numerous helpful comments. We thank Philippe Hupé for his critical reading; Marie-Claude Potier, Marc Sanson, Agnès Delaunay-Moisan and Jean-Yves Thuret for insightful discussions on RT-qPCR tests as well as on the interest of saliva samples; Catherine Hill for valuable discussions regarding epidemiology.

References

1. Wölfel R, Corman VM, Guggemos W, Seilmaier M, Zange S, Müller MA, et al. Virological assessment of hospitalized patients with COVID-2019. Nature. 2020;doi:10.1038/s41586-020-2196-x.

2. World Health Organization. Laboratory testing for 2019 novel coronavirus (2019-nCoV) in suspected human cases. WHO - Interim guidance. 2020;2019(January):1–7.

3. Corman VM, Landt O, Kaiser M, Molenkamp R, Meijer A, Chu DKW, et al. Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR. Eurosurveillance. 2020;25(3):1–8. doi:10.2807/1560-7917.ES.2020.25.3.2000045.

4. Gollier C, Gossner O. Group Testing against Covid-19 *. Covid Economics. 2020;(2):32–42.

5. Pouwels KB, Roope LSJ, Barnett A, Hunter DJ, Nolan TM, Clarke PM. Group Testing for SARS-CoV-2: Forward to the Past? PharmacoEconomics - Open. 2020;(0123456789):2–5. doi:10.1007/s41669-020-00217-8.

6. Holt E. Slovakia to test all adults for SARS-CoV-2. The Lancet. 2020;396(10260):1386–1387. doi:10.1016/S0140-6736(20)32261-3.

7. Health Ministry of the Grand Duchy of Luxembourg. Large Scale Testing; 2020. Available from: `https://covid19.public.lu/dam-assets/covid-19/documents/strat{é}gie-lst/strategie-LST-Fr.pdf`.

8. Mizumoto K, Kagaya K, Zarebski A, Chowell G. Estimating the asymptomatic proportion of coronavirus disease 2019 (COVID-19) cases on board the Diamond Princess cruise ship, Yokohama, Japan, 2020. Eurosurveillance. 2020;25(10):1–5. doi:10.2807/1560-7917.ES.2020.25.10.2000180.

9. Bi Q, Wu Y, Mei S, Ye C, Zou X, Zhang Z, et al. Epidemiology and transmission of COVID-19 in 391 cases and 1286 of their close contacts in Shenzhen, China: a retrospective cohort study. The Lancet Infectious Diseases. 2020;3099(20):1–9. doi:10.1016/S1473-3099(20)30287-5.

10. Bai Y, Yao L, Wei T, Tian F, Jin DY, Chen L, et al. Presumed Asymptomatic Carrier Transmission of COVID-19. JAMA. 2020;323(14):1406.

11. Lavezzo E, Franchin E, Ciavarella C, Cuomo-dannenburg G, Barzon L, Sciro M, et al. Suppression of COVID-19 outbreak in the municipality of Vo ', Italy. medRxiv. 2020; p. 1–23. doi:10.1101/2020.04.17.20053157.

12. Ferretti L, Wymant C, Kendall M, Zhao L, Nurtay A, Abeler-Dörner L, et al. Quantifying SARS-CoV-2 transmission suggests epidemic control with digital contact tracing. Science. 2020;368(6491):6936. doi:10.1126/science.abb6936.

13. Byambasuren O, Cardona M, Bell K, Clark J, McLaws ML, Glasziou P. Estimating the extent of asymptomatic COVID-19 and its potential for community transmission: Systematic review and meta-analysis. Official Journal of the Association of Medical Microbiology and Infectious Disease Canada. 2020;COVID-19:Accepted versio. doi:10.3138/jammi-2020-0030.

14. Johansson MA, Quandelacy TM, Kada S, Prasad PV, Steele M, Brooks JT. SARS-CoV-2 Transmission From People Without COVID-19 Symptoms. 2021; p. 1–8. doi:10.1001/jamanetworkopen.2020.35057.

15. Buitrago-Garcia D, Egli-Gany D, Counotte MJ, Hossmann S, Imeri H, Ipekci AM, et al. Occurrence and transmission potential of asymptomatic and presymptomatic SARS-CoV-2 infections: A living systematic review and meta-analysis. PLoS medicine. 2020;17(9):e1003346. doi:10.1371/journal.pmed.1003346.

16. Post-lockdown SARS-CoV-2 nucleic acid screening in nearly ten million residents of Wuhan, China. Nature Communications. 2020;11(1):1–7. doi:10.1038/s41467-020-19802-w.

17. Xing Y, Wong GWK, Ni W, Hu X, Xing Q. Rapid Response to an Outbreak in Qingdao, China. New England Journal of Medicine. 2020;February(Coorespondance):e129. doi:10.1056/NEJMc2032361.

18. Dorfman R. The Detection of Defective Members of Large Populations. The Annals of Mathematical Statistics. 1943;.

19. Aldridge M, Johnson O, Scarlett J. Group Testing: An Information Theory Perspective. Foundations and Trends in Communications and Information Theory. 2019;15(3-4):196–392. doi:10.1561/0100000099.

20. Hogan CA, Sahoo MK, Pinsky BA. Sample Pooling as a Strategy to Detect Community Transmission of SARS-CoV-2. JAMA. 2020;323(19):1967. doi:10.1001/jama.2020.5445.

21. Lohse S, Pfuhl T, Berkó-Göttel B, Rissland J, Geißler T, Gärtner B, et al. Pooling of samples for testing for SARS-CoV-2 in asymptomatic people. The Lancet Infectious Diseases. 2020;3099(20).

22. Yelin I, Aharony N, Shaer Tamar E, Argoetti A, Messer E, Berenbaum D, et al. Evaluation of COVID-19 RT-qPCR test in multi-sample pools. Clinical Infectious Diseases. 2020;doi:10.1093/cid/ciaa531.

23. Ben-Ami R, Klochendler A, Seidel M, Sido T, Gurel-Gurevich O, Yassour M, et al. Pooled RNA extraction and PCR assay for efficient SARS-CoV-2 detection. medRxiv. 2020; p. 2020.04.17.20069062. doi:10.1101/2020.04.17.20069062.

24. Shental N, Levy S, Wuvshet V, Skorniakov S, Shalem B, Ottolenghi A, et al. Efficient high-throughput SARS-CoV-2 testing to detect asymptomatic carriers. Science Advances. 2020;6(37):eabc5961. doi:10.1126/sciadv.abc5961.

25. Schmidt M, Hoehl S, Berger A, Zeichhardt H, Hourfar K, Ciesek S, et al. FACT- Frankfurt adjusted COVID-19 testing- a novel method enables high-throughput SARS-CoV-2 screening without loss of sensitivity. medRxiv. 2020; p. 2020.04.28.20074187. doi:10.1101/2020.04.28.20074187.

26. Mutesa L, Ndishimye P, Butera Y, Souopgui J, Uwineza A, Rutayisire R, et al. A pooled testing strategy for identifying SARS-CoV-2 at low prevalence. Nature. 2020;doi:10.1038/s41586-020-2885-5.

27. Torres I, Albert E, Navarro D. Pooling of Nasopharyngeal Swab Specimens for SARS-CoV-2 detection by RT-PCR. Journal of Medical Virology. 2020; p. 25971. doi:10.1002/jmv.25971.

28. Cabrera JJ, Rey S, Perez S, Martinez-Lamas L, Cores-Calvo O, Torres J, et al. Pooling for SARS-CoV-2 control in care institutions. medRxiv. 2020;doi:10.1101/2020.05.30.20108597.

29. Wacharapluesadee S, Kaewpom T, Ampoot W, Ghai S, Khamhang W, Worachotsueptrakun K, et al. Evaluating the efficiency of specimen pooling for PCR-based detection of COVID-19. Journal of Medical Virology. 2020; p. 0–1. doi:10.1002/jmv.26005.

30. Martin A, Storto A, Andre B, Mallory A, Dangla R, Visseaux B, et al. High-sensitivity COVID-19 group testing by digital PCR. arXiv. 2020;.

31. Khodare A, Padhi A, Gupta E, Agarwal R, Dubey S, Sarin SK. Optimal size of sample pooling for RNA pool testing: an avant-garde for scaling up SARS CoV 2 testing. medRxiv. 2020;doi:10.1101/2020.06.11.20128793.

32. Wernike K, Keller M, Conraths FJ, Mettenleiter TC, Groschup MH, Beer M. Pitfalls in SARS-CoV-2 PCR diagnostics. bioRxiv. 2020; p. 2020.06.03.132357. doi:10.1101/2020.06.03.132357.

33. Gan Y, Du L, Faleti OD, Huang J, Xiao G, Lyu X, et al. Sample Pooling as a Strategy of SARS-COV-2 Nucleic Acid Screening Increases the False-negative Rate. medRxiv. 2020;.

34. Griesemer SB, Van Slyke G, St George K. Assessment of sample pooling for clinical SARS-CoV-2 testing. bioRxiv. 2020;.

35. Chong BSW, Tran T, Druce J, Ballard SA, Simpson JA, Catton M. Sample pooling is a viable strategy for SARS-CoV-2 detection in low-prevalence settings. Pathology. 2020;doi:10.1016/j.pathol.2020.09.005.

36. Christoff AP, Cruz GNF, Sereia AFR, Boberg DR, de Bastiani DC, Yamanaka LE, et al. Swab pooling for large-scale RT-qPCR screening of SARS-CoV-2. medRxiv. 2020; p. 2020.09.03.20187732. doi:10.1101/2020.09.03.20187732.

37. Hanel R, Thurner S. Boosting test-efficiency by pooled testing strategies for SARS-CoV-2. arXiv. 2020;(1):3–5.

38. Deckert A, Bärnighausen T, Kyei N. Pooled-sample analysis strategies for COVID-19 mass testing: a simulation study. Bull World Health Organ. 2020;(April). doi:10.2471/BLT.20.257188.

39. Sinnott-Armstrong N, Klein D, Hickey B. Evaluation of Group Testing for SARS-CoV-2 RNA. medRxiv. 2020;doi:10.1101/2020.03.27.20043968.

40. Verdun CM, Fuchs T, Harar P, Elbrächter D, Fischer DS, Berner J, et al. Group testing for SARS-CoV-2 allows for up to 10-fold efficiency increase across realistic scenarios and testing strategies. medRxiv. 2020; p. 2020.04.30.20085290. doi:10.1101/2020.04.30.20085290.

41. Narayanan K, Frost I, Heidarzadeh A, Tseng KK, Banerjee S, John J, et al. Pooling RT-PCR or NGS samples has the potential to cost-effectively generate estimates of COVID-19 prevalence in resource limited environments. medRxiv. 2020; p. 2020.04.03.20051995. doi:10.1101/2020.04.03.20051995.

42. Bilder CR, Iwen PC, Abdalhamid B, Tebbs JM, McMahan CS. Tests in short supply? Try group testing. Significance. 2020;17(3):15–16. doi:10.1111/1740-9713.01399.

43. Furon T. The illusion of group testing. [Research Report] RR-9164, Inria Rennes Bretagne Atlantique. 2018; p. 1–19.

44. Barillot E, Lacroix B, Cohen D. Theoretical analysis of library screening using a N-dimensional pooling strategy. Nucleic Acids Research. 1991;19(22):6241–6247. doi:10.1093/nar/19.22.6241.

45. Thierry-Mieg N. Pooling in systems biology becomes smart. Nature Methods. 2006;3(3):161–162. doi:10.1038/nmeth0306-161.

46. Centers for Disease Control and Prevention. Interim Guidance for Use of Pooling Procedures in SARS-CoV-2 Diagnostic, Screening, and Surveillance Testing; 2020. Available from: `https://www.cdc.gov/coronavirus/2019-ncov/lab/pooling-procedures.html`.

47. Barak N, Ben-Ami R, Sido T, Perri A, Shtoyer A, Rivkin M, et al. Lessons from applied large-scale pooling of 133,816 SARS-CoV-2 RT-PCR tests. medRxiv. 2020;.

48. Denny TN, Andrews L, Bonsignori M, Cavanaugh K, Datto MB, Deckard A, et al. Implementation of a Pooled Surveillance Testing Program for Asymptomatic SARS-CoV-2 Infections on a College Campus — Duke University, Durham, North Carolina, August 2–October 11, 2020. MMWR Morbidity and Mortality Weekly Report. 2020;69(46):1743–1747.

49. Trimble NS. How Upstate Medical University used spit and grit to make 'game-changer' coronavirus test; 2021.

50. Fogarty A, Joseph A, Shaw D. Pooled saliva samples for COVID-19 surveillance programme. The Lancet Respiratory Medicine. 2020;3099(20):2599–2600. doi:10.1016/s2213-2600(20)30444-6.

51. Forootan A, Sjöback R, Björkman J, Sjögreen B, Linz L, Kubista M. Methods to determine limit of detection and limit of quantification in quantitative real-time PCR (qPCR). Biomolecular Detection and Quantification. 2017;12(April):1–6. doi:10.1016/j.bdq.2017.04.001.

52. Ruiz-Villalba A, van Pelt-Verkuil E, Gunst QD, Ruijter JM, van den Hoff MJ. Amplification of nonspecific products in quantitative polymerase chain reactions (qPCR). Biomolecular Detection and Quantification. 2017;14(July):7–18. doi:10.1016/j.bdq.2017.10.001.

53. Cleary B, Hay JA, Blumenstiel B, Harden M, Cipicchio M, Bezney J, et al. Using viral load and epidemic dynamics to optimize pooled testing in resource constrained settings. medRxiv. 2020; p. 2020.05.01.20086801. doi:10.1101/2020.05.01.20086801.

54. Watkins AE, Fenichel EP, Weinberger DM, Vogels CBF, Brackney DE, Casanovas-Massana A, et al. Pooling saliva to increase SARS-CoV-2 testing capacity. medRxiv. 2020;doi:10.1101/2020.09.02.20183830.

55. Lennon NJ, Bhattacharyya RP, Mina MJ, Rehm HL, Hung DT, Smole S, et al. Comparison of viral levels in individuals with or without symptoms at time of COVID-19 testing among 32,480 residents and staff of nursing homes and assisted living facilities in Massachusetts. medRxiv. 2020; p. 2020.07.20.20157792.

56. Jones TC, Mühlemann B, Talitha V, Marta Z, Hofmann J, Stein A, et al. An analysis of SARS-CoV-2 viral load by patient age. Preprint Charité Hospital. 2020;.

57. Cabrera Alvargonzalez JJ, Rey Cao S, Pérez Castro S, Martinez Lamas L, Cores Calvo O, Torres Pinõn J, et al. Pooling for SARS-CoV-2 control in care institutions. BMC Infectious Diseases. 2020;20(1):1–6. doi:10.1186/s12879-020-05446-0.

58. Liu Y, Yan LM, Wan L, Xiang TX, Le A, Liu JM, et al. Viral dynamics in mild and severe cases of COVID-19. The Lancet Infectious diseases. 2020;(20).

59. Kissler SM, Tedijanto C, Goldstein E, Grad YH, Lipsitch M. Projecting the transmission dynamics of SARS-CoV-2 through the postpandemic period. Science. 2020;5793(April):eabb5793. doi:10.1126/science.abb5793.

60. Hay JA, Kennedy-Shaffer L, Kanjilal S, Lipsitch M, Mina MJ. Estimating epidemiologic dynamics from single cross-sectional viral load distributions. medRxiv. 2020;001121:2020.10.08.20204222.

61. Pullano G, Domenico LD, Sabbatini CE, Valdano E, Turbelin C, Debin M, et al. Underdetection of COVID-19 cases in France in the exit phase following lockdown. Medrxiv. 2020; p. 1–13.

62. Mahase E. Covid-19: Universities roll out pooled testing of students in bid to keep campuses open. BMJ (Clinical research ed). 2020;370:m3789. doi:10.1136/bmj.m3789.

63. Hogan CA, Gombar S, Wang H, Röltgen K, Shi RZ, Holubar M, et al. Large-Scale Testing of Asymptomatic Healthcare Personnel for Severe Acute Respiratory Syndrome Coronavirus 2. Emerging Infectious Diseases. 2021;27(1):1–2. doi:10.3201/eid2701.203892.

64. Arons MM, Hatfield KM, Reddy SC, Kimball A, James A, Jacobs JR, et al. Presymptomatic SARS-CoV-2 Infections and Transmission in a Skilled Nursing Facility. New England Journal of Medicine. 2020;382(22):2081–2090. doi:10.1056/NEJMoa2008457.

65. Thompson KH. Estimation of the Proportion of Vectors in a Natural Population of Insects. Biometrics. 1962;18(4):568. doi:10.2307/2527902.

66. Prevention ECfD. Methodology for estimating point prevalence of SARS-CoV-2 infection by pooled RT-PCR testing. 2020;(May).

67. Barrett ES, Horton DB, Roy J, Gennaro ML, Brooks A, Tischfield J, et al. Prevalence of SARS-CoV-2 infection in previously undiagnosed health care workers at the onset of the U.S. COVID-19 epidemic. medRxiv. 2020; p. 2020.04.20.20072470. doi:10.1101/2020.04.20.20072470.

68. Gudbjartsson DF, Helgason A, Jonsson H, Magnusson OT, Melsted P, Norddahl GL, et al. Spread of SARS-CoV-2 in the Icelandic Population. New England Journal of Medicine. 2020; p. NEJMoa2006100. doi:10.1056/NEJMoa2006100.

69. Smith DR, Duval A, Pouwels KB, Guillemot D, Fernandes J, Huynh BT, et al. How best to use limited tests? Improving COVID-19 surveillance in long-term care. medRxiv. 2020;.

70. Wyllie AL, Fournier J, Casanovas-Massana A, Campbell M, Tokuyama M, Vijayakumar P, et al. Saliva is more sensitive for SARS-CoV-2 detection in COVID-19 patients than nasopharyngeal swabs. medRxiv. 2020;(2):2020.04.16.20067835. doi:10.1101/2020.04.16.20067835.

71. Azzi L, Carcano G, Gianfagna F, Grossi P, Gasperina DD, Genoni A, et al. Saliva is a reliable tool to detect SARS-CoV-2. Journal of Infection. 2020;(xxxx):1–6. doi:10.1016/j.jinf.2020.04.005.

72. To KKW, Tsang OTY, Leung WS, Tam AR, Wu TC, Lung DC, et al. Temporal profiles of viral load in posterior oropharyngeal saliva samples and serum antibody responses during infection by SARS-CoV-2: an observational cohort study. The Lancet Infectious Diseases. 2020;20(5):565–574. doi:10.1016/S1473-3099(20)30196-1.

73. Williams E, Bond K, Zhang B, Putland M, Williamson DA. Saliva as a non-invasive specimen for detection of SARS-CoV-2. Journal of Clinical Microbiology. 2020;50(April). doi:10.1128/JCM.00776-20.

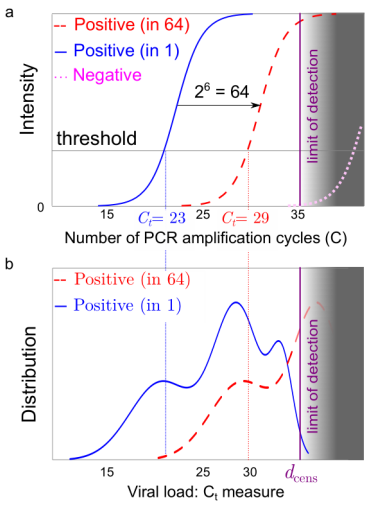
74. Khurshid Z, Zohaib S, Joshi C, Moin SF, Sohail M, Speicher DJ, et al. Saliva as a non-invasive sample for the detection of SARS-CoV-2 : a systematic review. medRxiv. 2020;.

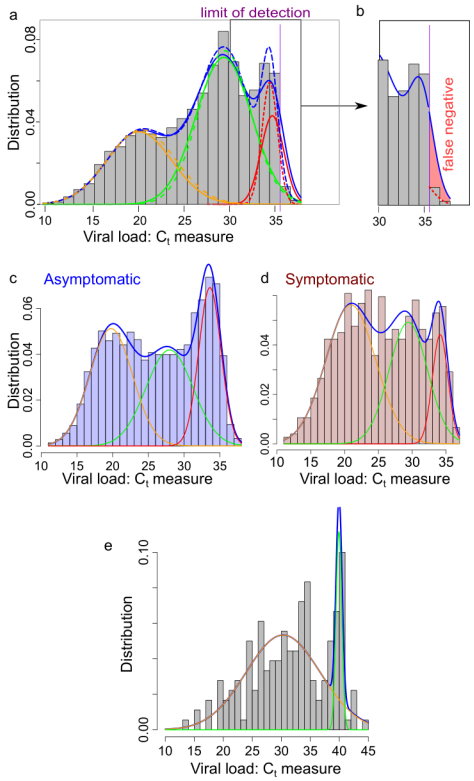
75. Hanson KE, Barker AP, Hillyard DR, Gilmore N, Barrett JW, Orlandi RR, et al. Self-Collected Anterior Nasal and Saliva Specimens versus Healthcare Worker-Collected Nasopharyngeal Swabs for the Molecular Detection of SARS-CoV-2. Journal of Clinical Microbiology. 2020;(October):1–5. doi:10.1128/jcm.01824-20.

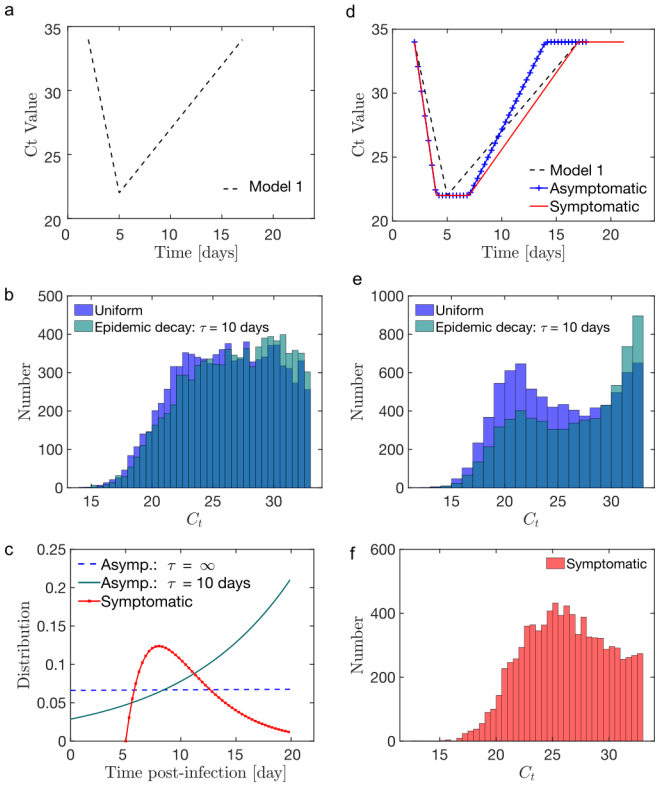
76. Siegler AJ, Hall E, Luisi N, Zlotorzynska M, Wilde G, Sanchez T, et al. Willingness to Seek Diagnostic Testing for SARS-CoV-2 With Home , Drive-through , and Clinic-Based Specimen Collection Locations. 2020;doi:10.1093/ofid/ofaa269.

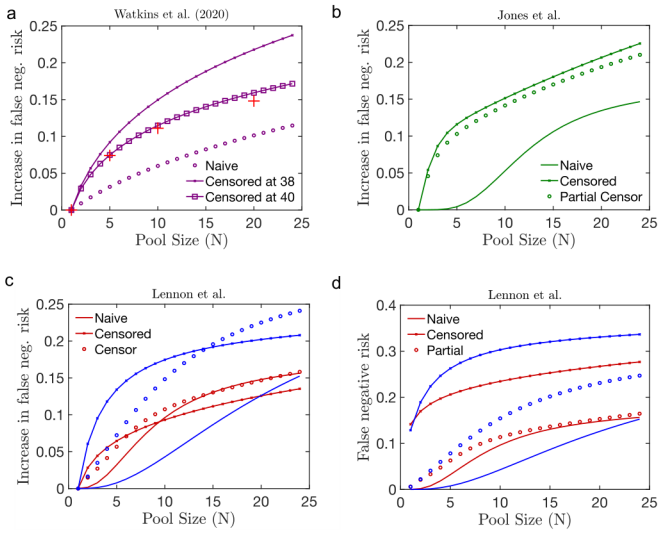
77. Treibel TA, Manisty C, Burton M, McKnight Á, Lambourne J, Augusto JB, et al. COVID-19: PCR screening of asymptomatic health-care workers at London hospital. The Lancet. 2020;6736(20):19–20. doi:10.1016/S0140-6736(20)31100-4.

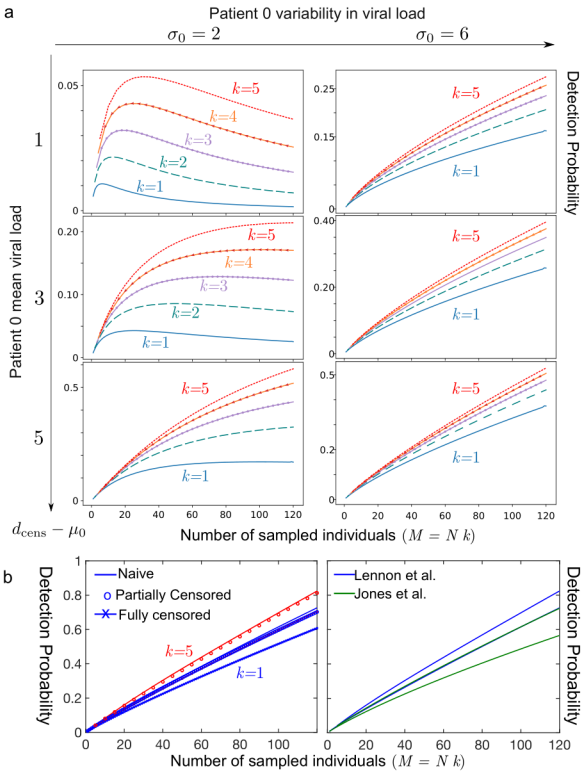
78. R Core Team. R: A Language and Environment for Statistical Computing; 2020. Available from: `https://www.R-project.org/`.

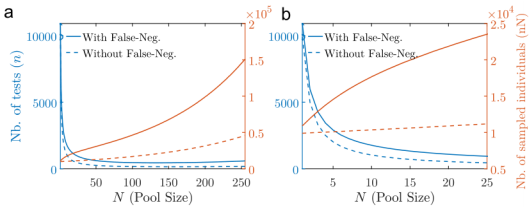

a
Positive (in 64)
Positive (in 1)
Negative
Intensity
2^6 = 64
threshold
limit of detection
0
15
C_t= 23
25
C_t= 29
35
Number of PCR amplification cycles (C)
b
Positive (in 64)
Positive (in 1)
Distribution
limit of detection
15
25
30
d_{cens}
Viral load: C_t measure


a
limit of detection
b
false negative
Distribution
0.00
0.04
0.08
10
15
20
25
30
35
Viral load: C_t measure
c
Asymptomatic
d
Symptomatic
0.00
0.02
0.04
0.06
e
0.10
40
45


a
Ct Value
Time [days]
Model 1
d
Ct Value
Time [days]
Model 1
Asymptomatic
Symptomatic
b
Number
C_t
Uniform
Epidemic decay: $\tau = 10$ days
e
Number
C_t
Uniform
Epidemic decay: $\tau = 10$ days
c
Distribution
Time post-infection [day]
Asymp.: $\tau = \infty$
Asymp.: $\tau = 10$ days
Symptomatic
f
Number
C_t
Symptomatic


a
Watkins et al. (2020)
Increase in false neg. risk
Pool Size (N)
Naive
Censored at 38
Censored at 40
b
Jones et al.
Increase in false neg. risk
Pool Size (N)
Naive
Censored
Partial Censor
c
Lennon et al.
Increase in false neg. risk
Pool Size (N)
Naive
Censored
Censor
d
Lennon et al.
False negative risk
Pool Size (N)
Naive
Censored
Partial


a
Patient 0 variability in viral load
$\sigma_0 = 2$
$\sigma_0 = 6$
Patient 0 mean viral load
1
3
5
$d_{\text{cens}} - \mu_0$
Detection Probability
k=5
k=4
k=3
k=2
k=1
Number of sampled individuals ($M = N\,k$)
b
Naive
Partially Censored
Fully censored
Lennon et al.
Jones et al.
Detection Probability
Number of sampled individuals ($M = N\,k$)


a
b
With False-Neg.
Without False-Neg.
Nb. of tests (n)
Nb. of sampled individuals (nN)
N (Pool Size)
N (Pool Size)
$\times 10^5$
$\times 10^4$
10000
5000
0
50
100
150
200
250
2
1.5
1
0.5
5
10
15
20
25
2.5

Supplementary Information

Vincent Brault, Bastien Mallein, Jean-Francois Rupprecht

I Fit of viral load distributions

I.1 Censored Gaussians

In this section, we present the simple mixing models of Sec. I.2, as well as some complementary graphs and the estimations obtained for the parameters of this models.

I.1.1 Theorem I.1

Theorem I.1. *The estimators $(\widehat{\mu}, \widehat{\sigma}, \widehat{q})$ of (μ, σ, q) obtained by maximisation of the likelihood ratio are strongly consistent and asymptotically normal.*

The properties of the maximum likelihood estimators is a consequence of the fact that the (partially) censored Gaussian model belongs to the family of exponential laws (c.f. [1, Chapter 9] and SI I.1.1). To check the quality of the approximation of the estimators by `nlm`, we simulate variable sizes of samples distributed according to the censored Gaussian model. The values of these estimations are plotted in SI B.3.

Proof. To prove the lemma I.1, we observe that for every $x \in \mathbb{R}$ we have the following decomposition of the density f_X if $q > 0$:

$$f_X(x) = b(\eta) \exp\left[\langle \eta, T(x) \rangle\right], \tag{S1}$$

with $< \cdot, \cdot >$ is the scalar product, $\eta = \left(\frac{\mu}{\sigma^2}, \quad -\frac{1}{2\sigma^2}, \quad \ln q\right)$ the natural parameters,

$$T(x) = \left(x, \quad x^2, \quad \mathbb{1}_{\{x > d_{\text{cens}}\}}\right),$$

the sufficient statistics and

$$b(\eta) = \frac{1}{q + (1-q)F_{\mu,\sigma}(s)} \times \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\mu^2}{2\sigma^2}\right).$$

□

For the totally censured model, we have the same decomposition with the third parameters and taking $q = 0$. Thanks the decomposition in Eq. (S1), the (partially) censured model belongs to the family of exponential laws and the maximum likelihood estimators are strongly consistent and asymptotically normal.

I.1.2 Simulations

To study the quality of the estimators defined in Sec. I.2.2, we simulated 10^4 samples of size $n \in \{10^2, 10^3, 10^4, 10^5\}$ of variables following the model $\mathcal{CN}_{d_{\text{cens}}}(0, 1, q)$ with $d_{\text{cens}} \in \{-2, -1, 0, 1, 2, 3\}$ and $q \in \{0, 0.1, 0.5, 0.9\}$. We provide boxplots estimations of the parameters in Fig. S1 and a zoom on significant part in Fig. S2. Note that these parameters ($\mu = 0, |d_{\text{cens}}| \leq 3$) are very different from the ones expected for C_t values, but the model can be straightforwardly adapted by an affine transformation to measured parameters of interest.

Observe from Fig. S1 that the estimations are generally close to the parameters but we can sometimes have very large deviations. We find that the more n increases, the better the estimator. The threshold seems to have a weak influence on the estimation of the partially censored model but, for the fully censored model, we see that the more d_{cens} increases and the more the quality of the estimators increases; especially when $d_{\text{cens}} = -2$ which represents approximately the 2.3% quantile. Note that we observe large deviations in the partially censored model when d_{cens} is equal to 2; this may seem counter-intuitive since we have access to around 97.7% of uncensored Gaussian information. However, this leaves few observations for the estimation of p (which we observe on the graphs of the last line) and this weakens in this case the model because censorship no longer really has any reason to be. We therefore recommend using the model only when the number of observations after censorship is sufficient to estimate the parameter p.

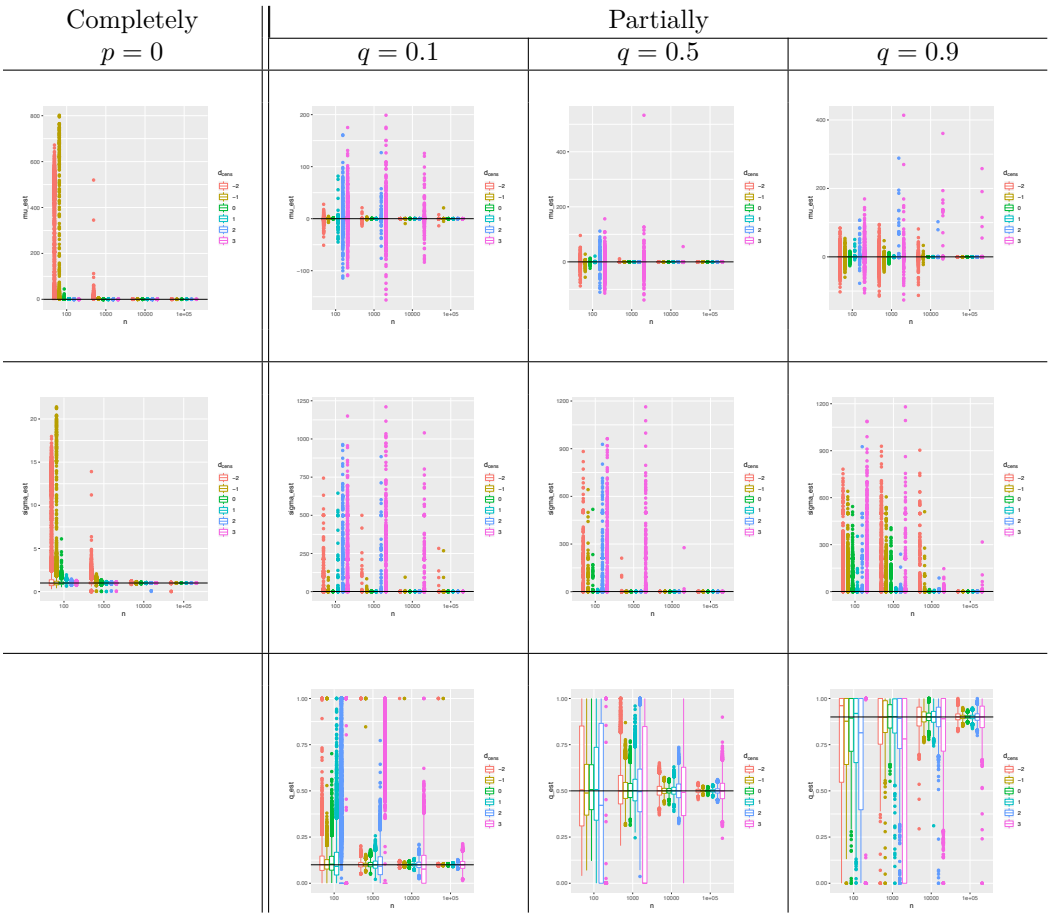


Figure S1. Boxplots of the estimations of μ (first row), σ (second row) and p (last row ; only for partially censored model) in function of model (columns), the size n of sample (x-axis) and the value of the threshold s (color). The true value is symbolised by the horizontal black line. Analysis is performed on a controlled dataset, as explained in B.1.3

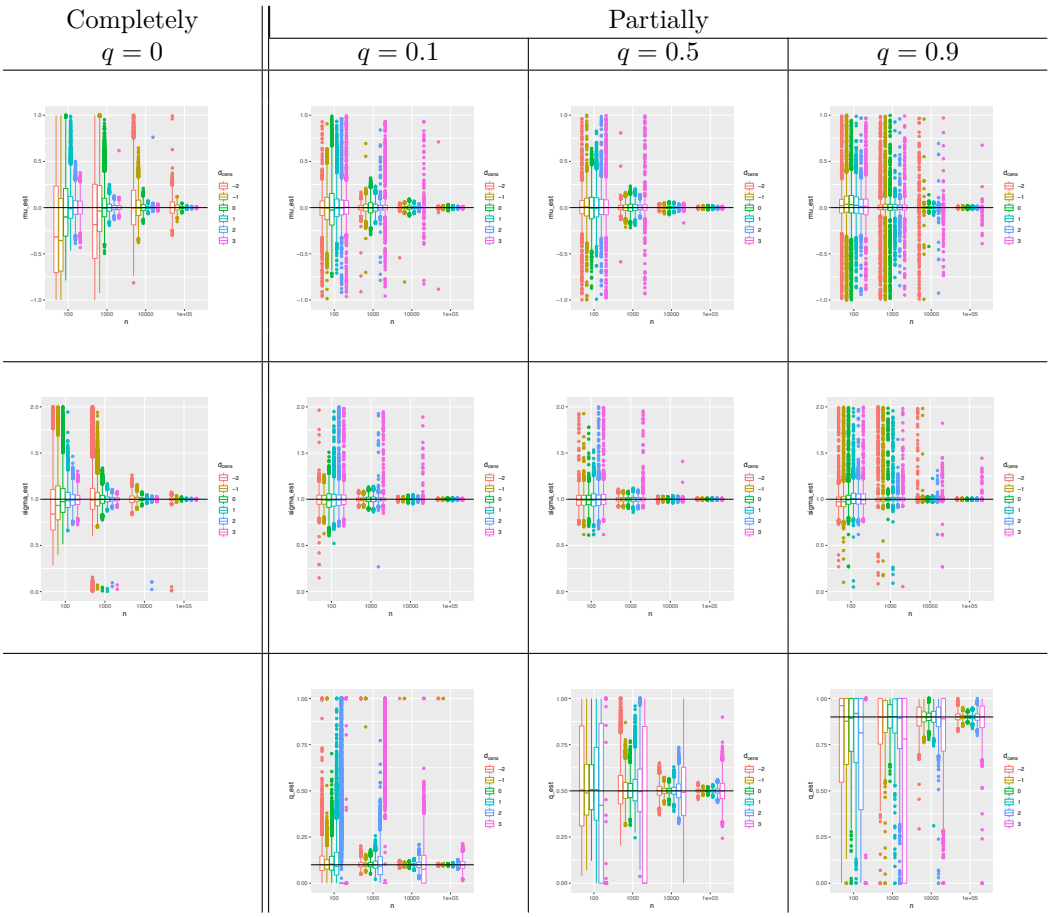


Figure S2. Zoom on boxplots of the estimations of μ (first row), σ (second row) and q (last row ; only for partially censored model) in function of model (columns), the size n of sample (x-axis) and the value of the threshold s (color). The true value is symbolised by the horizontal black line. Analysis is performed on a controlled dataset, as explained in B.1.3

I.2 Analysis of the Jones et al. dataset [2]

I.2.1 Naive method

In this section, we trace the density estimated by a simple mixture of Gaussian variable presented in the main text, Sec. I.2.2. An estimation of the parameters of this mixture are given in Table S3.

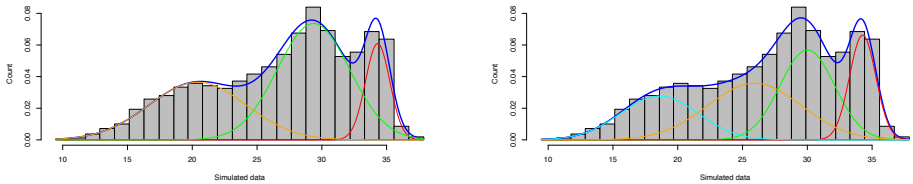


Figure S3. Representation of the histogram from Jones et al. [2] with the densities estimated with 3 classes (on the left) and 4 classes (on the right): the color lines (other than blue) represent the density of each component and the blue line the density of the mixture.

As we do not have access to raw data, we performed simulations to generate a reconstructed datasets with consistent histograms to Fig. 1 from [2], with randomized position of the points within each class. We applied the above procedure to 100 independently reconstructed data, in order to limit the influence of the random part. Among these 100 simulations, we obtain 95 times 3 clusters and 5 times 4 clusters. When there are 3 clusters, the estimation of the parameters is very stable (standard deviation less than 0.03 for each) but there is a little more variability in the case of 4 clusters in particular for the two classes with the largest averages (but the standard deviation does not exceed 0.25).

I.2.2 Censored mixture model

In this section, we present the complementary graphs of Sec. I.2. The statistical model presented here has the following density defined for all $x \in \mathbb{R}$ by:

$$f(x) = \sum_{k=1}^{3} \pi_k \frac{f_{\mu_k,\sigma_k}(x)}{q_k + (1-q_k)F_{\mu_k,\sigma_k}(d_{\text{cens}})} \left[1 + (q_k - 1)\mathbb{1}_{\{x > d_{\text{cens}}\}}\right]. \tag{S2}$$

where $f_{\mu_k,\sigma_k}(x)$ is the Gaussian density of mean μ_k and variable σ_k and $F_{\mu_k,\sigma_k}(d_{\text{cens}})$ the corresponding cumulative distribution at the limit of detection. With the model in Eq. (S2), we can estimate the theoretical false negative rate by the following formula:

$$\mathbb{P}\left(\text{false negative}\right) = \sum_{k=1}^{3} \pi_k \left[1 - F_{\mu_k,\sigma_k}(d_{\text{cens}})\right](1 - q_k). \tag{S3}$$

We point out that the completely censured mixture model has the same density than the Eq. (S2) in the limit $q_k = 0$.

I.3 Analysis of the Lennon et al. dataset [4]

In these datasets, there are two populations : symptomatic and asymptomatic. In the Table S4, we represent the number of clusters selected by the procedure on 100 resampling of the histogram.

Table S3. Estimated parameters for the naive Gaussian mixture fit and the censored Gaussian mixture fits defined in Eq. (S2), for the datasets available in [2] and [3]. Note the consistency of the estimations, in particular in the partially and completely censored models.

[2]

Model	$q_{i\,i=1..3}$	μ_1	σ_1	π_1	μ_2	σ_2	π_2	μ_3	σ_3	π_3
Naive		20.41	3.74	0.34	29.43	2.81	0.52	34.32	0.89	0.14
Partially	0.2	20.14	3.60	0.32	29.35	2.96	0.53	34.78	1.32	0.14
Completely		20.13	3.60	0.33	29.41	3.02	0.54	34.81	1.31	0.13

[3]

Model	$q_{i\,i=1..3}$	μ_1	σ_1	π_1	μ_2	σ_2	π_2	μ_3	σ_3	π_3
Naive		19.75	2.05	0.20	25.61	2.99	0.39	34.28	2.36	0.40
Partially	0.4	20.16	2.19	0.26	26.03	2.58	0.43	34.54	2.66	0.41
Completely		20.55	3.45	0.31	26.33	2.11	0.24	34.41	2.98	0.43

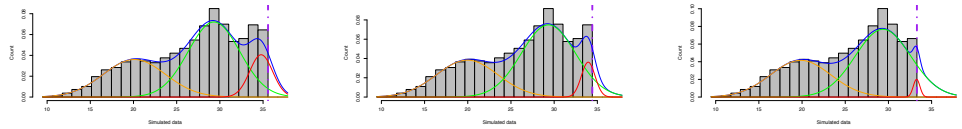


Figure S4. Density of the fits of the censored model with three components (obtained when erasing data to the right of the threshold) with a threshold at 35.6 (left), 34.4 (middle) and 33.2 (right): the orange, green and red lines represent the density of each component and the blue line the density of the mixture. The histogram correspond to the one presented in [2].

Table S4. Estimated parameters for the censored Gaussian mixture fit define in Eq. (S2) for different values of the threshold d_{cens}, applied to reconstructed data data with same distribution as in [2] erased above d_{cens}.

d_{cens}	μ_1	σ_1	π_1	μ_2	σ_2	π_2	μ_3	σ_3	π_3
35.6	20.13	3.60	0.33	29.41	3.02	0.54	34.81	1.31	0.13
34.4	20.13	3.61	0.35	29.35	2.99	0.57	34.21	1.03	0.08
33.2	19.97	3.56	0.03	29.40	3.14	0.59	33.21	1.16	0.48

Table S4. Repartition of the number of clusters according to the considered symptomatic or asymptomatic dataset in [4].

	Clusters	
	2	3
Symptomatic	33%	67%
Asymptomatic	0%	100%

For the symptomatic population, the 3 cluster decomposition is selected twice more often than 2 cluster one. For the asymptomatic population, 3 clusters were selected for every resampling. In the main text, we consider a 3 cluster decomposition for both datasets.

In Figure S5, we represents the estimations the mixture densities for each distribution.

For the censored mixture, we obtain the following estimations (see table S5 and Figure S6).

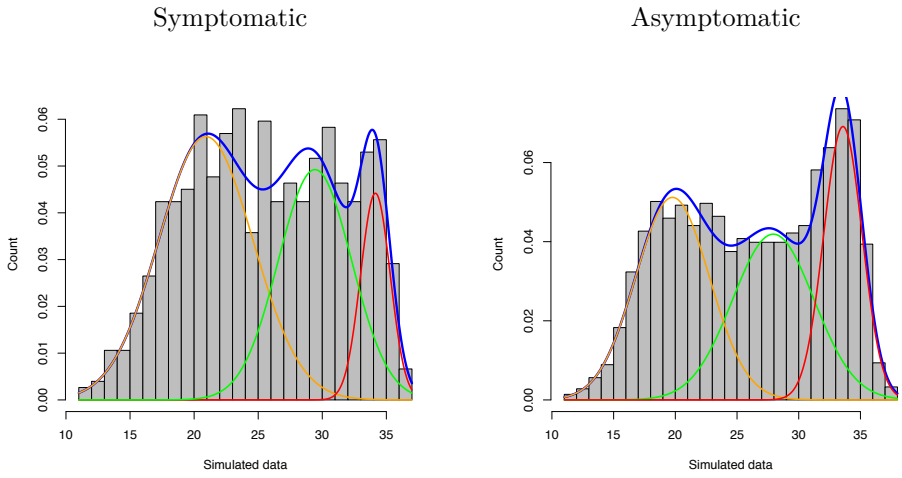


Figure S5. Representation of the histogram for each distribution (symptomatic on left and asymptomatic on right) with the estimation of the mixture densities.

Table S5. Estimated parameters for the censored Gaussian mixture fit define in Eq. (S2) for different values of the threshold d_{cens}, applied to reconstructed data data with same distribution as in Lennon at al. [4] erased above d_{cens}.

	d_{cens}	μ_1	σ_1	π_1	μ_2	σ_2	π_2	μ_3	σ_3	π_3
sympto	36	19.98	3.41	0.36	26.68	4.48	0.38	33.5	3.3	0.24
	35	20.41	3.47	0.46	29.59	4.15	0.45	92.01	10.09	0.07
	34	19.7	3.28	0.36	26.16	4.39	0.41	34.96	4.58	0.22
	33	19.03	3.08	0.32	24.9	3.71	0.45	31.53	2.66	0.22
asympto	37	19.24	2.81	0.31	27.41	4.23	0.43	33.56	1.59	0.24
	36	19.19	2.79	0.3	27.45	4.66	0.46	33.68	1.78	0.23
	35	18.47	2.49	0.23	24.3	3.77	0.38	34.83	3.74	0.37
	34	18.75	2.6	0.28	25.35	4.03	0.44	34.76	3.15	0.26

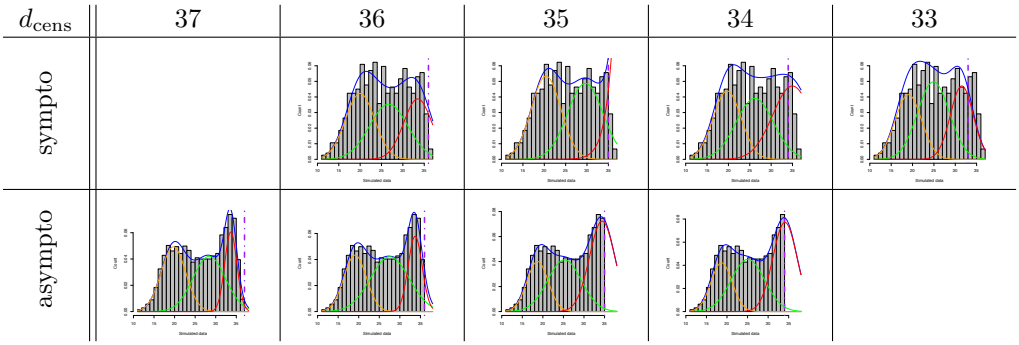


Figure S6. Density of the fits of the censored model with three components (obtained when erasing data to the right of the threshold) with a threshold (columns) for the two datasets (rows): the orange, green and red lines represent the density of each component and the blue line the density of the mixture. The histogram correspond to the one presented in Lennon et al. [4].

For the censored mixture, we obtain the following estimations (see table S6 and Figure S7).

Table S6. Estimated parameters for the partially censored Gaussian mixture fit define in Eq. (S2) for different values of the threshold d_{cens}, applied to reconstructed data data with same distribution as in Lennon et al. [4] erased above d_{cens}.

	d_{cens}	q_i	μ_1	σ_1	π_1	μ_2	σ_2	π_2	μ_3	σ_3	π_3
symp	36	0.39	20.6	3.52	0.47	29.59	3.61	0.43	34.26	1.18	0.09
	35	0.76	21.28	3.8	0.55	30.26	2.82	0.34	34.54	0.96	0.09
asymp	37	0.51	19.55	2.93	0.35	27.8	3.75	0.38	33.56	1.6	0.25
	36	0.41	19.53	2.91	0.35	27.79	3.82	0.38	33.67	1.74	0.25
	35	0.82	19.64	2.96	0.36	28.01	3.62	0.37	33.67	1.6	0.25

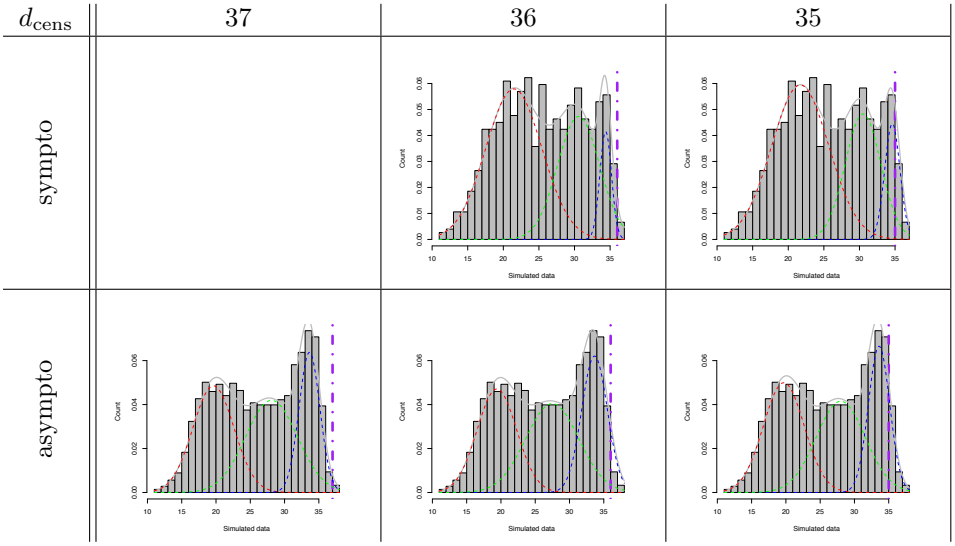


Figure S7. Density of the fits of the partially censored model with three components (obtained when erasing data to the right of the threshold) with a threshold (columns) for the two datasets (rows): the orange, green and red lines represent the density of each component and the blue line the density of the mixture. The histogram correspond to the one presented in Lennon et al. [4]

I.4 Analysis of the ImpactSaliva dataset (Watkins et al. [5])

For the censure data, we refer to Table S7 and Fig. S8. For the partially censored, we refer to Table S8 and the Fig. S9.

Table S7. Estimated parameters for the censored Gaussian mixture fit define in Eq. (S2) for different values of the threshold d_{cens}, applied to reconstructed data data with same distribution as in the ImpactSaliva (Watkins et al. [5]) dataset, erased above d_{cens}.

d_{cens}	μ_1	σ_1	π_1	μ_2	σ_2	π_2
44	30.8	6.82	0.86	40	0.47	0.13
43	31.19	7.1	0.87	40.02	0.45	0.12
42	30.79	6.83	0.86	40	0.47	0.13
41	31.19	7.09	0.86	40.05	0.52	0.13
40	31.67	7.37	0.96	39.49	0.02	0.03

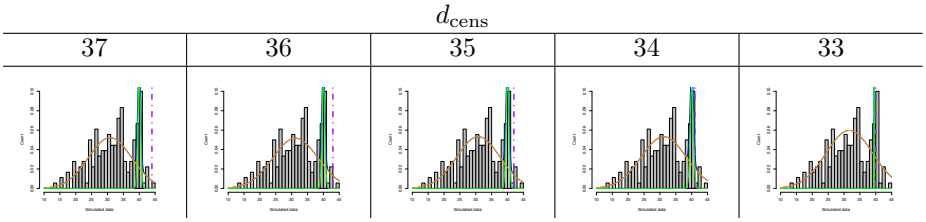


Figure S8. Density of the fits of the censored model with two components (obtained when erasing data to the right of the threshold) with a threshold (columns) for the two datasets (rows): the orange and green lines represent the density of each component and the blue line the density of the mixture. The histogram correspond to the one presented in ImpactSaliva dataset [5].

Table S8. Estimated parameters for the partially censored Gaussian mixture fit define in Eq. (S2) for different values of the threshold d_{cens}, applied to reconstructed data data with same distribution as in the ImpactSaliva dataset [5] erased above d_{cens}.

d_{cens}	q_i	μ_1	σ_1	π_1	μ_2	σ_2	π_2
44	0.95	30.74	6.75	0.86	40	0.47	0.13
43	0.95	31.13	7.04	0.87	40.01	0.46	0.12
42	0.76	30.46	6.5	0.86	39.99	0.48	0.13
41	0.08	30.69	6.68	0.86	40.05	0.6	0.13

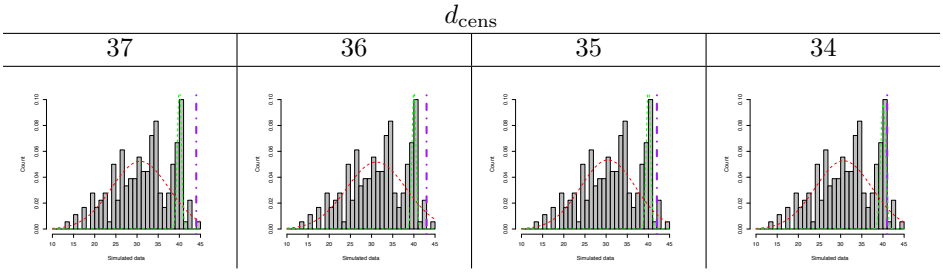


Figure S9. Density of the fits of the partially censored model with two components (obtained when erasing data to the right of the threshold) with a threshold (columns) for the two datasets (rows): the orange and green lines represent the density of each component and the blue line the density of the mixture. The histogram corresponds to the ImpactSaliva dataset.

II Estimation of the false-negative risk in the presence of multiple positive individuals in the pool

We treat here the case of a pool of N samples that contains $k > 1$ positive individuals. We also consider the risk of defective sampling (e.g. that the swabs fails to collect viral load in an infected individual), which we denote ζ. The probability of having a negative pool result given that there is k positive samples within the pool reads, according to the model presented in Eq. (4):

$$\mathbb{P}\left[-|k+\right] = \sum_{j=1}^{k} \binom{k}{j} \zeta^{k-j}(1-\zeta)^{j} \mathbb{P}\left[\log_2\left(\sum_{i=1,\dots,j} C_i/N\right) > d_{\text{cens}}\right]. \tag{S4}$$

Under the two assumptions that:

1. the viral load distribution spans several order of magnitudes (e.g. log-normal distributed), so that, following Eq. (7):

$$\mathbb{P}\left[\log_2\left(\sum_{i=1,\dots,j} C_i/N\right) > d_{\max}\right] = \mathbb{P}\left[\min_{i=1,\dots,j}(\log_2(C_i)) > d_{\max}^{(N)}\right], \tag{S5}$$

   with $d_{\max}^{(N)} = d_{\text{cens}} - \log_2(N)$.

2. the viral loads between the k infected individuals are independent, in which case:

$$\mathbb{P}\left[\min_{i=1,\dots,j}(\log_2(C_i)) > d_{\max}^{(N)}\right] = \mathbb{P}\left[\log_2(C_1) > d_{\max}^{(N)}\right]^{j}, \tag{S6}$$

we find that Eq. (S4) takes the simple expression:

$$\mathbb{P}\left[-|k+\right] = \left(\zeta + (1-\zeta)(1 - \mathbb{P}\left[\log_2(C_1) < d_{\max}^{(N)}\right]\right)^{k}. \tag{S7}$$

In Fig. S10, in the case of correlated samples, we find that the false negative risk in pooling is greatly reduced if there is more than one positive sample in the pool. The origin of such false-negative reduction is the large variability in viral load and the fact that the amplification technique is particularly sensitive to the highest viral load in the sample. Such false-negative reduction is robust to the presence of a finite risk of defective sampling $\zeta = 5\%$.

In addition, one may expect the number of positive k to be distributed according to a binomial distribution with a parameter p corresponding to the prevalence of the disease. Conditioned on the probability that there is at least one individual that is infected within the pool, the conditional probability that $k \geq 1$ is the number of infected individuals then reads

$$\mathbb{P}\left[k+|+\right] = \frac{1}{1-(1-p)^N}\binom{N}{k} p^k (1-p)^{N-k}, \tag{S8}$$

which leads to the following expression for the averaged probability that the pool test turns negative although there is at least one positive individuals in the community

$$\mathbb{P}\left[-|+\right] = \frac{1}{1-(1-p)^N} \sum_{k=1}^{N} \binom{N}{k} p^k (1-p)^{N-k} \left(\zeta + (1-\zeta)(1 - \mathbb{P}\left[\log_2(C_1) < d_{\max}^{(N)}\right]\right)^{k}, \tag{S9}$$

$$= \frac{1}{1-(1-p)^N} \left\{ \left(p\left(\zeta + (1-\zeta)(1 - \mathbb{P}\left[\log_2(C_1) < d_{\max}^{(N)}\right]\right) + 1 - p\right)^{N} - (1-p)^N \right\}. \tag{S10}$$

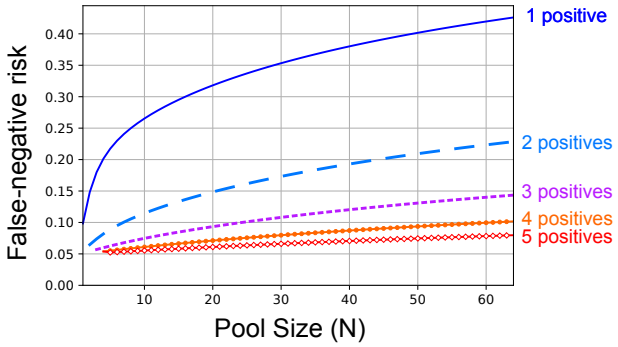


Figure S10. Evaluation of the total risk of false negatives estimated according to Eq. (S4) as a function of the pool size N for several values of the number of positive samples in the pool $k = 1$ (blue solid line); $k = 2$ (cyan dashed line); $k = 3$ (magenta line); $k = 4$ (diamond orange line); $k = 5$ (circle red line). We consider a risk that the sample is defective $\zeta = 0.05$.

As shown in Fig. S11, the averaged false-negative probability risk is not necessarily a monotonous function.

A similar non-linear relation between the false-negative rate and the underlying population prevalence is also reported in [6].

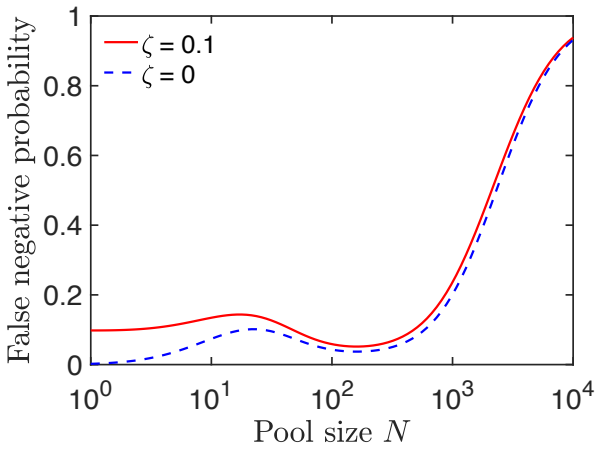


Figure S11. Example of a counter-intuitive evolution in averaged community false negative risk of false negatives, as defined through Eq. (S10), as a function of the pool size N (solid red curve) with a defective sampling probability $\zeta = 0.10$. (dashed blue curve) without defective sampling., considering a single Gaussian distribution of viral loads with $\mu = 27$ and $\sigma = 2$.

III Measuring the prevalence with ideal tests

We present here some of results obtained from the computations made in Sec. III, where we assumed perfect group testing and used it to measure prevalence in the population. Note that with a perfect test, the question of early detection of an outbreak in a community becomes much simpler : one just need to test everyone at regular time intervals with a single test.

III.1 Proof of the confidence intervals for the prevalence measurement

We recapitulate a derivation that closely follows the one of a seminal paper, [7]. We assume that we have n tests at our disposal. Given $N \in \mathbb{N}$, we sample nN individuals at random in the general population, and organize n pools of N individuals. Each of these pools is then tested using the perfect tests. For all $i \leq n$, we write $X_i^{(N)} = 1$ if the ith test is positive (i.e. if and only if at least one of the N individuals in the ith pool is infected), and $X_i^{(N)} = 0$ otherwise. We denote by p the (unknown) proportion of infected individuals in the population, then $(X_i^{(N)}, i \leq n)$ forms an independent and identically distributed (i.i.d.) sequence of Bernoulli random variables with parameter $1-(1-p)^N$.

Lemma III.1. *Writing* $\overline{X}_n^{(N)} = \frac{1}{n}\sum_{j=1}^N X_j^{(N)}$, *the quantity* $1-(1-\overline{X}_n^{(N)})^{1/N}$ *is a strongly consistent and asymptotically normal estimator of* p. *A confidence interval of asymptotic level* $1-\alpha$ *is*

$$\mathrm{CI}_{1-\alpha}(p) = \left[1-(1-\overline{X}_n^{(N)})^{1/N} \pm \frac{q_\alpha (1-\overline{X}_n^{(N)})^{1/N-1}\sqrt{\overline{X}_n^{(N)}(1-\overline{X}_n^{(N)})}}{\sqrt{n}N}\right], \quad \text{(S11)}$$

where q_α *is the quantile of order* $1-\alpha/2$ *of the standard Gaussian random variable.*

Proof. Note that $(X_j^{(N)}, j \leq n)$ is a standard Bernoulli model, hence $\overline{X}_n^{(N)}$ is a consistent and asymptotically normal estimator of $f(p) = 1-(1-p)^N$. Hence, using that f^{-1} is $\mathcal{C}^1$ and Slutsky's lemma, we deduce all the above properties of the estimator $f^{-1}(\overline{X}_n^{(N)})$ of p. □

Remark III.2. As $\lim_{n\to\infty} 1-(1-\overline{X}_n^{(N)})^{1/N} = p$ almost surely, for any $N \in \mathbb{N}$ the width of the confidence interval defined in Lemma III.1 satisfies

$$\frac{2q_\alpha (1-\overline{X}_n^{(N)})^{1/N-1}\sqrt{\overline{X}_n^{(N)}(1-\overline{X}_n^{(N)})}}{\sqrt{n}N} \underset{n\to\infty}{\sim} \frac{2q_\alpha}{\sqrt{n}}\frac{(1-p)}{N}\sqrt{\frac{1-(1-p)^N}{(1-p)^N}} \quad \text{a.s.} \quad \text{(S12)}$$

III.2 Proof of the optimal size for the prevalence measurement

The width of the confidence interval defined in Eq. (S11) behaves asymptotically, by law of large numbers, as $2(1-p)q_\alpha f(p,N)/\sqrt{n}$ where

$$f(p,n,N) := \frac{1}{N}\cdot\sqrt{\frac{1-(1-p)^N}{(1-p)^N}}.$$

An optimal choice for the value of N given p can thus be chosen as the value of N minimizing $f(p,\cdot)$. Indeed, this choice minimizes the width of the confidence interval for

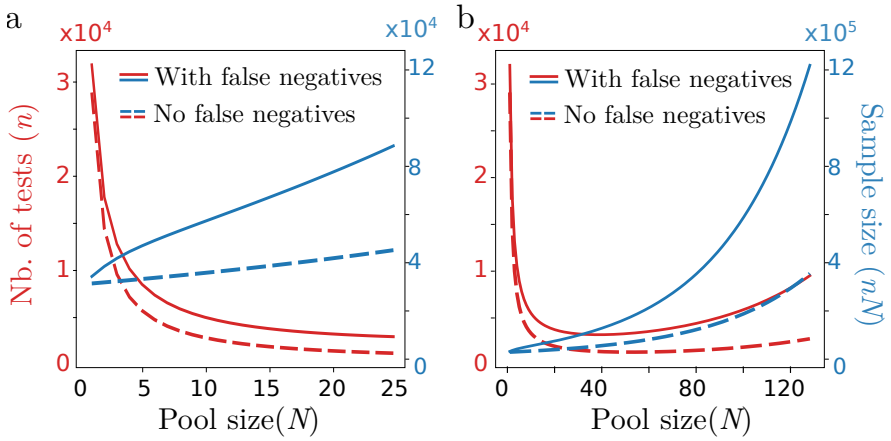


Figure S12. (a,b) Total number of tests (red) and total number of sampled individuals (blue) in order to estimate a prevalence of $p = 3\%$ with a $\pm 0.2\%$ precision with 95% confidence interval as a function of the pool size N for the perfect case with no false negative (dashed lines) versus the case with false negatives (solid lines) estimated according to the Jones et al. dataset [2]. In (a) N ranges from 0 to 25; in (b) N ranges from 0 to 128; as visible in (b), the valley around the optimal pool size $N_{\text{opt}}^{(\text{perf})} \approx 50$ is large: near optimal savings in tests are achieved even for moderately large pool sizes that require smaller number of individuals to sample.

the measured prevalence. Plots of $N \mapsto f(p, N)$ are provided for several values of p in Figure S13.

We observe that the quantity N^{opt} will approach the quantity x^* which minimizes $x \mapsto \log f(p, x)$. Observe that x^* then satisfies

$$0 = -\frac{1}{x^*} + \frac{1}{2}\frac{-\log(1-p)}{1 - e^{x^* \log(1-p)}} \iff x^*(-\log(1-p)) = 2(1 - e^{x^* \log(1-p)})$$
$$\iff (x^*(-\log(1-p)) - 2)\, e^{x^*(-\log(1-p))} = -2$$

Therefore, the minimum of $x \mapsto f(p, x)$ is attained at point

$$x^* = \frac{2 + W(-2e^{-2})}{-\log(1-p)},$$

with W the Lambert W function (the inverse function of $x \mapsto xe^x$).

Observe that different optimization could be considered, for example choosing values of n and N that minimize the width of the interval of confidence on the measure of the prevalence for a given cost C, measured as $aN + nN$, with a representing the cost of a test, the cost of sampling an individual being normalized at 1. In this situation, the optimization problem becomes very similar, using that $n = C/N - a$. In this situation, the width of the asymptotic confidence interval decays as $2(1-p)q_\alpha g(p, C, N)$ with

$$g(p, C, N) = \frac{1}{\sqrt{CN - aN^2}} \sqrt{\frac{1 - (1-p)^N}{(1-p)^N}}.$$

The optimal value of N in this situation interpolates between $N = 1$ when $a \to 0$ and $N = x^*$ when $a \to \infty$.

III.3 Number of tests and sample size as function of the population prevalence

We trace here, for various values of the prevalence, the number of tests and total number of samples needed to archive a given precision for the confidence interval. We observe that over a large range of prevalences, the number of tests needed to reach a given precision on the measure of the prevalence remains small for a large range of pool sizes. On the other hand, the total number of individuals to sample grows quadratically, and the test sensibility decreases with the size of the pools. Hence, it might be interesting to consider a suboptimal choices $N < N_{\text{opt}}$ for the pool sizes when measuring the prevalence.

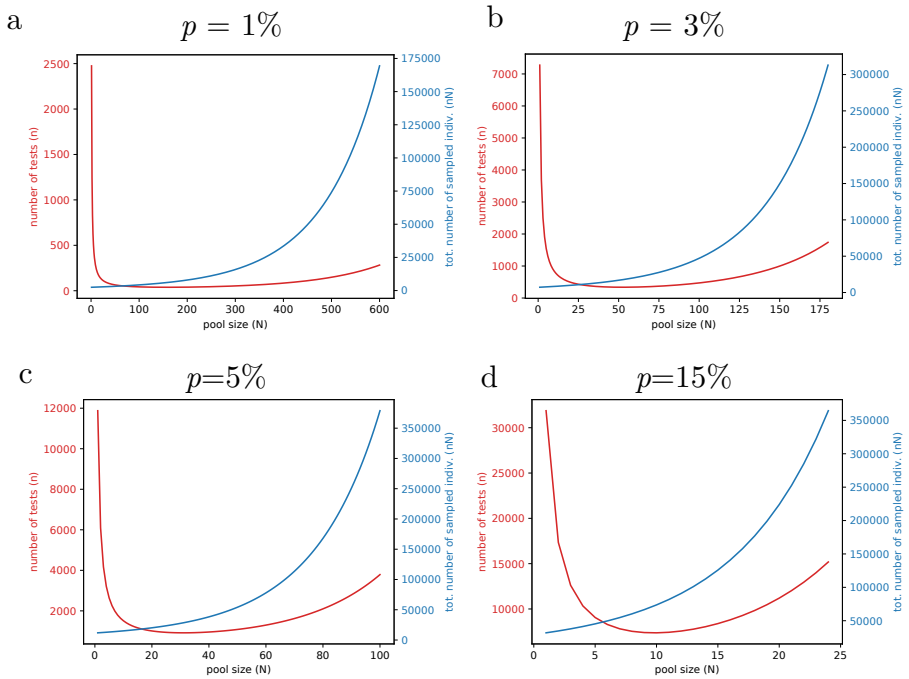


Figure S13. Total number of tests and sampled individuals so that the width of the 95% confidence interval is smaller than 0.4% as a function of the pool size N chosen for a perfect test, for a prevalence p equal to $p = 1\%$ (a), $p = 3\%$ (b), $p = 5\%$ (c), $p = 15\%$ (d).

III.4 Bayesian inference

We are now interested in a Bayesian approach to the measure of prevalence. We started with an initial prior distribution with density $f_0(p) = 6p(1-p)\mathbf{1}_{\{0\leq p\leq 1\}}$ for the prevalence, and for each new test j we do the following:

1. take the the mean value $\overline{p}_{j-1} = \int_0^1 p f_{j-1}(p)\mathrm{d}p$ of the prior;
2. choose the size N_j of the pool of the jth test computed as (cf. Eq. (18)):

$$N_j = \left\lfloor -\frac{c_\star}{\log(1-\overline{p}_{j-1})} \right\rfloor; \tag{S13}$$

3. choose N_j individuals at random and test them in a group:
   - if the test is positive, then $f_j(p) = C_j^+(1-(1-p)^{N_j})f_{j-1}(p)$;

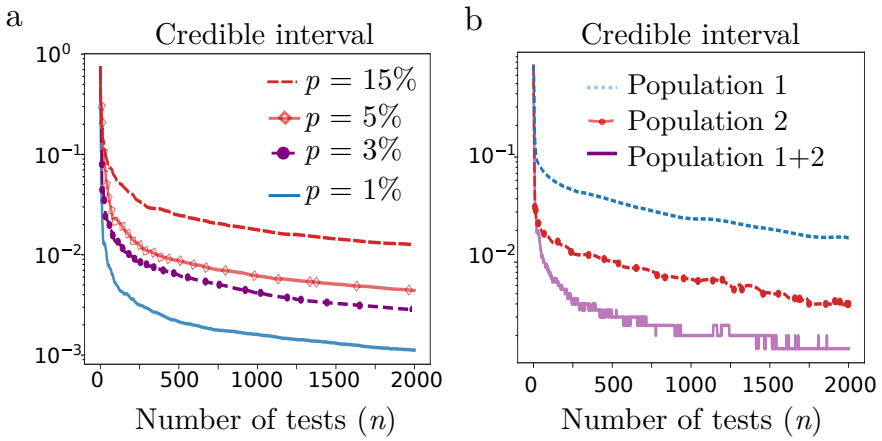


Figure S14. (a) Width of the 95% credible interval on the prevalance p with adaptative Bayesian sampling as a function of the number of tests n for a set of values in the prevalence ranging from $p = 15\%$ (top, magenta dashed line) to $p = 1\%$ (bottom, blue solid line). (b) Width of the credible intervals in a two-category mixed population for the prevalence either: in the general population (magenta solid line); for the less exposed population 1 with a prevalence of 0.5%, representing 80% of the general population (blue dashed line); for the more at-risk population 2 with a prevalence of 5% representing 20% of the general population (red dotted line with circles).

- if the test is negative, then $f_j(p) = C_j^-(1-p)^{N_j} f_{j-1}(p)$;

with $C_j^{\pm}$ normalizing constants, chosen such that $\int_0^1 f_j(p)\mathrm{d}p = 1$.

We trace in Fig. S14 the result in blue of this experiment, the 95% credible interval being $[a_j, b_j]$, with a_j being the 2.5%th quantile of f_j and b_j its 97.5% quantile.

Simultaneously to this statistical experiment, one can follow the prevalence in sub-populations of interest. For example, if we assume the population consists of two sub-populations 1 and 2 with different prevalences p_1 and p_2. Starting with a prior distribution $f_j(p_1, p_2)\mathrm{d}p_1\mathrm{d}p_2$ for these prevalences, if a group consisting of a individuals of the first sub-population and b individuals of the second population is sampled positive, then Bayes rules gives $C_{j+1}^+(1-(1-p_1)^a(1-p_2)^b)$ for the updated law of (p_1, p_2). A similar update is made if the test is negative. As a result, we get estimates for the prevalence in each sub-population at the same time as we are measuring the prevalence in the overall population.

We test the above statistical experiment on a population which is composed of two sub-populations, one large subpopulation of sparsely exposed individuals (prevalence 0.5%, representing 4/5th of the whole population), and a smaller subpopulation of very exposed individuals (prevalence 5%). At each step, we choose the size of the pool according to the available estimate for the prevalence in the complete population. The composition of the pool in terms of individuals of each sub-population is chosen at random (at the jth step, there are $\mathrm{Ber}(N_j, 0.8)$ individuals of the first sub-population). We also update our estimation of the prevalences (p_1, p_2) in each of the two sub-populations.

The results are traced in Fig. S14 in orange and green curves. One can see that the width of the credibility intervals of the sub-populations decay much slower than for the whole population. The reason is that the size of the groups are optimized to measure as precisely as possible the mean value p.

However, observe that even with a naive group construction (without segregating individuals according to their sub-population), one can extract information on the

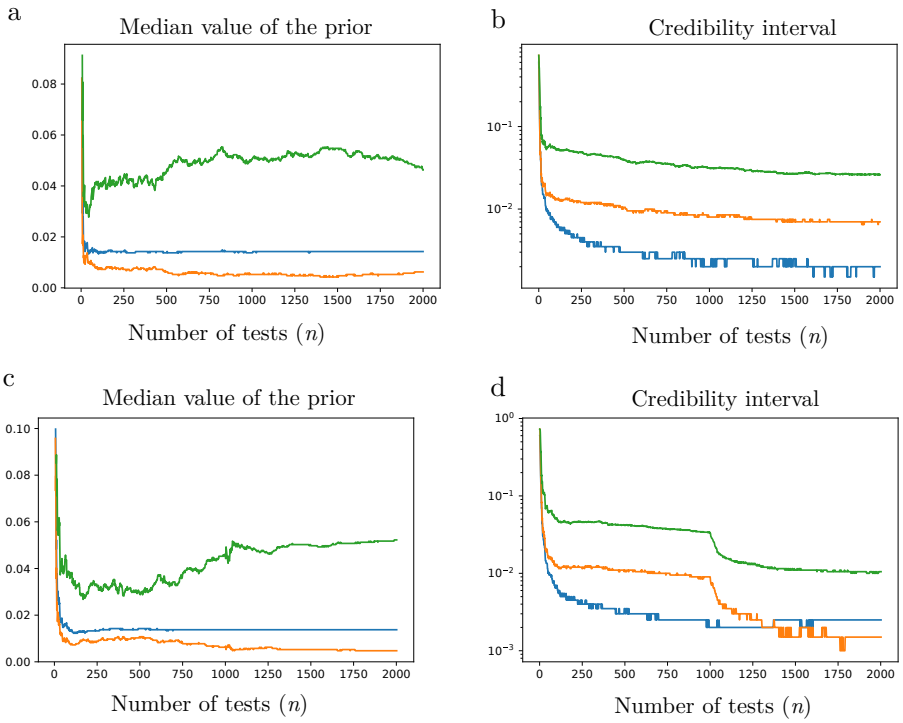


Figure S15. (a-b) Bayesian estimation of the parameters of a mixed population, consisting of 80% individuals of type 1 with a prevalence of 0.5% and 20% individuals of type 2 with a prevalence of 5%. Pooled samples are constituted by sampling randomly individuals from the two sub-populations, with a size optimized for the speed of convergence of the overall prevalence of 1.4%. (a-c) Median value of the priors, overall population in blue, first resp second population in orange resp green. (b-d) Width of the 95% credible intervals. In (c,d), the first 1000 tests are made on groups whose size is optimized to estimate the prevalence of the overall population, the next 1000 tests are divided into two groups that are used on homogeneous sets of the sub-populations, in groups optimized to estimate the prevalences within these sub-populations. This has the effect of drastically improving the speed of convergence of the estimator of the prevalence in the subpopulations.

prevalence of the sub-populations of interest. Therefore, a design for the measure of the prevalence in a stratified population could be the following: in a first time, pool testing is implemented on randomly constructed group of individuals from the general population. Data is then analysed to detect sub-populations with different prevalences (e.g. according to geography, age, occupation, ...). In a second time, once sub-populations of interest are identified, pool testing is applied to each of the sub-populations independently. We implemented this method if Fig. S15, with the same number of tests a much more detailed estimate of the prevalence is obtained.

IV Optimization of the frequency of test

We study here the impact of the regularity of tests on the rate of detection of a infection occurring in a closed community of A individuals. We assume a fixed budget of tests per unit of time, which are made on pools of fixed size N of individuals chosen at random in the community. We compare different strategies for the detection of outbreaks in the community depending on the frequency of the tests. At one extreme, a single test is made on one pool of N individuals every T units of time, at the other extreme, every individual in the community is tested every TA/N unit of time. These two strategies both use on average one test every T units of time, the first one emphasizing the regularity of testing, while the second one exhaustively tests every member of the community.

More generally, for all $1 \leq f \leq A/N$, we can consider the detection strategy with period f in which every fT units of time, a number fN of individuals in the community get tested in pooled samples. The aim of this strategy is to detect as soon as possible the infection of the community in order to deploy additional aseptic measures and prepare for a potential influx of hospitalized patients from this community. Perhaps unsurprisingly, we show that the smaller the frequency is, the lower the number of infected individuals is at the first time of detection of the outbreak. However, note that in many closed communities (e.g. professional athletes in Germany football league, US baseball league, etc.) the opposite strategy is put in practice with testing of the whole team at regular intervals rather than randomly selected members every day.

We model the initial outbreak in a community as a Crump-Mode-Jagers (CMJ) process [8]. Individuals are infected from the outside at a Poisson rate of small parameter η. Every infected individual then goes through several stages of the disease. After an incubation period t_O, the individual starts excreting the virus up to the time t_f. During that phase, the individual will infect members of its community at a Poisson rate of parameter λ. The probability that an individual eventually becomes symptomatic is denoted r; in this case, at a random time denoted S between t_O and t_f, the individual will start showing symptomatic.

For our purpose, the detection of the outbreak corresponds to the first time at which either:

- an individual becomes symptomatic in the population,
- or a pooled test turns positive.

We place ourselves in a stationary regime, with a screening strategy of period f. In this situation, as exterior infections happen according to a Poisson process, the first exterior infection will occur at a time chosen uniformly at random between two screening times. In other words, the first screening which might detect the outbreak will happen UfT units of time aft after the exterior infection. Comparing with the screening strategy with period 1, we see that the latter strategy will on average use $fT/2$ tests on the population between the first screening time of the strategy f. Additionally, the longer the time between the infection and the test, the higher the probability that a first symptomatic individual will appear, making useless the screening.

To quantify the above heuristic, we considered a simpler model in which $t_O = 0$ (the time of incubation is neglected), $t_f = \infty$ (the time of recovery is neglected) and the time of apparition of symptom S is chosen as an exponential random variable with parameter ρ. In this situation, we obtain analytic values for the number of infected individuals at the first detection time.

In this simplified model, the number of infected individuals t units of time after the first infection, denoted by $N(t)$ follows a standard Yule process [9], therefore $N(t)$ is distributed as a geometric random variable with parameter $1 - e^{-\lambda t}$. Moreover, given $N(t)$, a new infection will occur at rate $\lambda N(t)$ while an individual will become

symptomatic at rate $\kappa r N(t)$, using that a fraction r of the population is symptomatic. In the absence of screening, the apparition time of first symptom can be expressed as

$$T_s = \sum_{j=1}^{G} e_j, \tag{S14}$$

where G is the number of infections up to the first symptomatic one and $e_j$is time interval between the $j-1$th and jth infection events; we consider G to be geometrically distributed ($\mathbb{P}(G=k) = p^{k-1}(1-p)$) with parameter $p = (r\kappa)/(\lambda + r\kappa)$ and that the e_j are independent exponential random variables with parameter $j(\lambda + r\kappa)$ (since each newly infected individual contribute to the intra-community attack rate by a multiplicative factor $\lambda + r\kappa$).

Averaging Eq. (S14), we find that the average apparition time of first symptom $\langle T_s \rangle$ reads

$$\langle T_s \rangle = \frac{-\log(1 - \frac{\lambda}{\lambda + r\kappa})}{(\lambda + r\kappa)}. \tag{S15}$$

Next, we observe that at the first screening time, there is a number $N(fTU)$ of infected individuals with U an independent uniform random variable. Based on Eq. (14), we find that, as long as $e^{\lambda f T} \ll A$, the first screening test will detect the outbreak with a probability approximately equal to

$$\mathbb{P}[+] = \langle \Phi_0(d_{\text{cens}}^{(N)}) \left(1 - (1 - N(fTU)/A)^{fN}\right) \rangle, \tag{S16}$$

$$\approx \Phi_0(d_{\text{cens}}^{(N)}) f N \langle N(fTU) \rangle / A, \tag{S17}$$

$$\approx \frac{\Phi_0(d_{\text{cens}}^{(N)}) N}{AT} (e^{\lambda f T} - 1). \tag{S18}$$

where $1 - \Phi_0(d_{\text{cens}}^{(N)})$ is the group test false-negative rate. From Eq. (S18), the screening detection probability quantity is an increasing of the sampling period period T and infection rate f; the corresponding detection time will then exceeding the onset of symptom time $\langle T_s \rangle$ for large infection rates f or sampling periods T.

Factoring in the fact that the number of infected individuals grows exponentially fast, and that more frequent screening implies several chances of detecting the infection before first symptoms show up, these computations show that frequent testing is key to a successful screening strategy, much more than exhaustive testing of the community.

As a validation of the previous computations, we estimate by Monte-Carlo method the average number of infected individuals $\langle N(T_d) \rangle$, we find that a screening strategy consisting in sampling a random subgroup of the community as frequently as possible is more efficient than the one consisting in testing larger portion of the community at less frequent time intervals. In Fig. S16, we compare different screening scenarios for a large community composed of $A = 1000$ individuals. We vary the value of the screening time interval τ while keeping fixed (1) the average number of tests per unit of time and (2) the size of the pools on which each test is used. Our simulation range from checking N individuals every day (with one test) to checking $12 \times N$ individuals in 12 pools every 12 days.

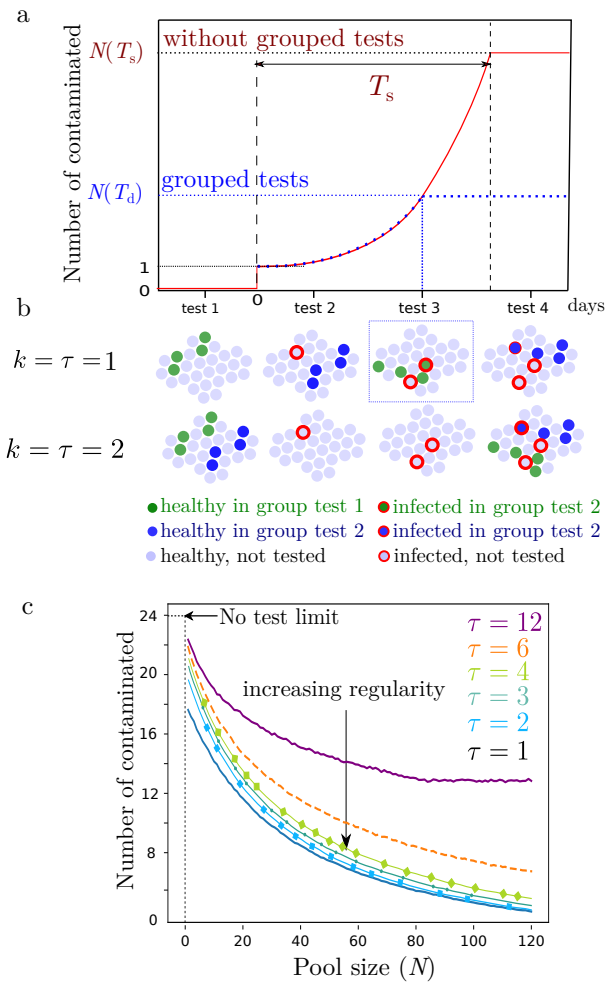


Figure S16. (a) Sketch of the time evolution of the number of infected individuals in a community. The patient 0 is infected from the outside of the community 0.8 units of time after a test date. In the absence of screening tests, the infection is detected at the time $T = T_s$ (after appearance of the first symptoms); with grouped tests, an infected individual is detected at a time $T = T_d$. (b) Sketch of two group testing strategies, here with pools of size $N = 4$, one with a single ($k = 1$) grouped tests every day ($\tau = 1$); the other with $k = 2$ grouped tests every second day ($\tau = 2$); the second strategy (least frequent testing) fails to detect the outbreak early and results in more infections. (c) Number of infected individuals at the detection of the outbreak as a function of the pool size, using $k = \tau$ tests performed at τ-day intervals, with $\tau = 12$ (solid purple line), $\tau = 6$ (dashed orange line), $\tau = 4$ (dark green solid line with square), $\tau = 3$ (light green solid line with circles), $\tau = 2$ (cyan line with squares) and $\tau = 1$ (solid blue line). Here we consider a large community composed of $A = 1000$ individuals. The patient 0 has a viral load concentration distributed according to a log-normal distribution with mean $\mu_0 = 30$ and standard deviation $\sigma_0 = 2\log_2(2)$; all others parameters can be found in Table II.

Table S16. Table with standard parameter values considered in SI. Sec. IV

Symbol	Meaning	Value
t_O	Incubation time (as defined in Eq. (9))	0
t_f	End of symptom time (as defined in Eq. (9))	∞
S	Random time of onset of symptoms (mean κ)	
κ	Mean onset of symptoms time	5 days
η	External attack rate on the community	$\eta \ll \lambda$
λ	Intra-community infection rate	$0.5\,\text{days}^{-1}$
r	Probability that an individual remains asymptomatic	40 %
τ	Time interval between grouped tests	$1 - 12$ days
A	Total number in the community	1000
N	Pool size	1–128

References

1. Nielsen OE. Information and exponential families : in statistical theory. Chichester U.K. New York: John Wiley & Sons; 2014.

2. Jones TC, Mühlemann B, Talitha V, Marta Z, Hofmann J, Stein A, et al. An analysis of SARS-CoV-2 viral load by patient age. Preprint Charité Hospital. 2020;.

3. Cabrera JJ, Rey S, Perez S, Martinez-Lamas L, Cores-Calvo O, Torres J, et al. Pooling for SARS-CoV-2 control in care institutions. medRxiv. 2020;doi:10.1101/2020.05.30.20108597.

4. Lennon NJ, Bhattacharyya RP, Mina MJ, Rehm HL, Hung DT, Smole S, et al. Comparison of viral levels in individuals with or without symptoms at time of COVID-19 testing among 32,480 residents and staff of nursing homes and assisted living facilities in Massachusetts. medRxiv. 2020; p. 2020.07.20.20157792.

5. Watkins AE, Fenichel EP, Weinberger DM, Vogels CBF, Brackney DE, Casanovas-Massana A, et al. Pooling saliva to increase SARS-CoV-2 testing capacity. medRxiv. 2020;doi:10.1101/2020.09.02.20183830.

6. Verwilt J, Mestdagh P, Vandesompele J. Evaluation of efficiency and sensitivity of 1D and 2D sample 1 pooling strategies for diagnostic screening purposes 2 3. medRxiv. 2020; p. 2020.07.17.20152702.

7. Thompson KH. Estimation of the Proportion of Vectors in a Natural Population of Insects. Biometrics. 1962;18(4):568. doi:10.2307/2527902.

8. Schertzer E, Simatos F. Height and contour processes of Crump-Mode-Jagers forests (I): general distribution and scaling limits in the case of short edges. Electron J Probab. 2018;23:43 pp. doi:10.1214/18-EJP151.

9. Meleard S. Modèles aléatoires en Ecologie et Evolution. CMAP; 2016. Available from: `http://www.cmap.polytechnique.fr/IMG/pdf/LIVRE07102013.pdf`.