# Group testing as a strategy for the epidemiologic monitoring of COVID-19

Vincent Brault, Bastien Mallein, Jean-François Rupprecht

# Group testing as a strategy for the epidemiologic monitoring of COVID-19

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

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

(Dated: May 27, 2020)

In this article, we propose an analysis and applications of sample pooling to the epidemiologic monitoring of COVID-19 with an economy of tests. We first show that group testing is a theoretically efficient and economic tool to provide a direct measure of the prevalence of the disease. We then introduce a precise model of the RT-qPCR process, used to test for the presence of virus in a sample. We construct a statistical model for the viral load in a typical infected individual based on clinical data from Jones et. al. (2020). Using these models, we then propose a method for the measure of the prevalence in a population, based on group testing, taking into account the increased number of false negatives associated to this method. Finally, we present an application of sample pooling for the prevention of epidemic outbreak in relatively closed connected communities (e.g. care homes for the elderly).



Regularly monitoring the *prevalence* of a disease , i.e. the proportion of infected individuals within the general population at a given time, is a key element to prevent the onset of an epidemic wave, to estimate the effect of social distancing policies and to anticipate a potential increase in the demand for hospitalization in intensive care units [1].

In the context of the COVID-19 epidemics, contagious individuals are generally assumed to bear a viral load of SARS-COV-2 in their respiratory tract [2, 3]. Such viral load can be detected and quantified within swab samples using a technique called *reverse transcription quantitative polymerase chain reaction* (RT-qPCR) [4]. With tests performed in priority on symptomatic patients, the proportion of positive tests (which corresponds to an *apparent* prevalence [5]) is larger than the prevalence among the whole population, which we call *overall* prevalence. In principle, the overall prevalence could be deduced from the apparent prevalence based on inferred model estimates for the proportion of tested individuals among the infected population. However, reliable models of COVID-19 are challenging to construct given (a) the current large uncertainty regarding the proportion of asymptomatic carriers (estimated to be in $20-50\%$ range [6–9]) and (b) the delay between the contamination and first symptoms, which varies from 1 to 5 days [10, 11].

Testing a large portions of the population at random would allow for a direct measure of the overall prevalence, including the proportion of asymptomatic individuals. Unfortunately, it appears that the production of reactants used in RT-qPCR diagnostic would not meet a demand in regular large-scale population testing [12, 13].

In such context of a shortage in reactants, group testing has received renewed interest. The principle of group testing consists in combining samples from multiple individuals into a single pool that is then tested using a single test-kit. The pool sample is considered to be positive if and only if at least one individual in the group is contaminated. The idea of group testing is not new, with a long history that dates back to 1943 [14] in the context of syphilis detection (see [15] for a review). Optimal diagnostic strategies include smart-pooling, whereby pools are organised according to lines and columns on a grid –or hypercube– with overlaps enabling the identification of positive individuals [16–18].

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 [19]. A recent publication from Saarland University, Germany, shows that positive sample with a relatively mild viral load from asymptomatic patients could still be detected within pools of 30 [20]. Further works suggest that RT-qPCR viral detection can been achieved in pools of up to 64 individuals [21–24].

In parallel, the theoretical literature on group testing for SARS-COV-2 diagnostic is growing at a fast pace [12, 25–27]. 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 [4]. 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 rarely taken into account [18].

In this communication, we do not address any diagnostic problems, such as the question of determining optimal strategies to provide individual positive diagnostic to a large population using a minimal number of tests. Rather, we propose to evaluate pooling strategies as a

*Electronic address: vincent.brault@univ-grenoble-alpes.fr; mallein@math.univ-paris13.fr; rupprecht@cpt.univ-mrs.fr

tool for the study of epidemiologic questions. In the pooling strategies we discuss below, no individual will be part of two different pools at the same time, so no information on the infection status of any distinguished individual is obtained. We instead focus on (i) the measure of the overall prevalence and (ii) on the early detection of contamination in a closed community.

The rest of the article is constructed as follows. We first present a simple protocol for the measure of the prevalence in the population by the use of group testing; we make in Section I the assumption of *perfect test*, i.e. that the test used is not subject to any false positive or negative, no matter the size of the pool to which the test is applied. In Section II, we provide a short description of the RT-qPCR and propose a statistical model for its study, that underlines its limit of detection at small concentrations. Then in Section III, we analyse part of the information recovered from the quantified viral charge in patients from the clinical dataset of Ref. [28]; this gives a sense of the viral load infected patients should carry in the general population, therefore the effect of dilution on the rate of false negatives. Finally, we show in Section IV how, using the statistical model and the measure of the error rate discussed above, one can measure the viral prevalence in the general population, or design a protocol allowing an early detection of an epidemic outbreak in a closed vulnerable community (e.g. schools, retirement homes, detention centers).

## I. MEASURING PREVALENCE WITH PERFECT TESTS

We investigate in this section the measure of the prevalence of the disease in a population using a group testing strategy, under the assumption of *perfect* tests, i.e. with no risks of false negative (or false positive). Our derivation is similar to [29].

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 $i$th test is positive (i.e. if and only if at least one of the $N$ individuals in the $i$th 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 I.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*

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

*where* $q_\alpha$ *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* I.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 I.1 satisfies

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

In other words, 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.

A classical calculation shows that the prefactor in Eq. (2) is minimal when the number of mixed samples per pool is equal to:

$$N_{\mathrm{opt}}^{(\mathrm{perf})} = -\frac{c_\star}{\log(1-p)} \iff (1-p)^{N_{\mathrm{opt}}^{(\mathrm{perf})}} \approx 0.20, \quad (3)$$

where $c_\star = 2 + W(-2e^{-2}) \approx 1.59$ and $W$ is the Lambert $W$ function [12]. Specifically, the size of the pools is optimal when approximately 80% of the tests made on the groups turn positive.

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 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. 1, 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

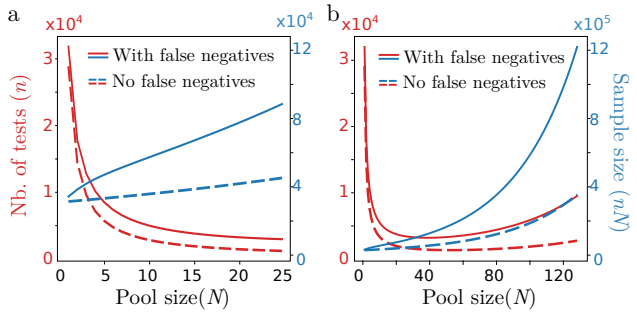



Figure 1: (a,b) Total number of tests (red) and total number of sampled individuals (blue) in order to estimate a prevalence of $p = 3\%$ with a $\pm 0.2\%$ precision with 95% confidence interval as a function of the pool size $N$ for the perfect case (dashed lines) considered in Sec. I with no false negative, and the more realistic case (solid lines) considered in Sec. IV (with false negatives; parameters are defined in Table II). 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 $N_{\rm opt}^{\rm (perf)} \approx 50$ is large: near optimal savings in tests are achieved even for moderately large pool sizes, which requires smaller number of individuals to sample.

of quasi-optimal pool sizes that can be chosen with minimal diminution of the precision in the measure of the prevalence.

Taking the exemple of a prevalence at $p = 3\%$, we expect the pool size minimizing the number of tests needed to read $N_{\rm opt}^{\rm (perf)} = 50$, which divides by 20 the number of tests needed (see Fig. 1). But to reach similar level of precision than in single testing, the total number of individuals that need to be sampled is more than doubled. Choosing instead a pool size of $N = 20$ requires almost the same number of tests, yet at a cost of a30% in the total number $Nn$ of sampled individuals (c.f. Fig. 1). The same observation holds for different values of the prevalence, see for example the graphs in Fig. S1.

## II. STATISTICAL MODELLING OF THE PCR

The RT-qPCR technique has been extensively used to estimate the concentration of viral material in samples [30], by making it fluorescent and grow at a controlled exponential rate. We briefly sketch the main steps of an RT-qPCR diagnostic protocol in Box 1. The qPCR 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. 2).

Combined measures of two viral RNA strands with a control of a human RNA strand are recommended in order to detect defective sampling that could induce false negatives, but also to improve precision of the measure as well as to normalize the number of virus copies by the quantity of human DNA [4]. Such combined measure can also improve precision of the measure as well as to normalize the number of virus copies by the quantity of human. We do not intend to include such features in our model. Furthermore, we do not model here the possible errors at the reverse transcription stage, which could lead to some biased measure of the viral load distribution

PCR tests are prone to amplify non-specific DNA sequences [30, 31] together with the target DNA of interest. This can creates fluorescence in a negative samples (in which there is no viral contamination), hence a negative sample might show up as positive after a certain number of cycles; such false positives typically occur after a critical number of cycles which we denoted $d_{\rm cens}$ (detection threshold), with is usually in the 35 to 40 range. We refer to these false positive outcomes within negative samples (in which the virus is absent) as *artefacts* [31] that we model as if triggered by a vanishingly small artificial viral concentration, denoted $\epsilon_1$. In classical qPCR testing, samples becoming fluorescent after cycle $d_{\rm cens}$ are treated as negative, and this threshold is chosen so that artefact very rarely show below this level; fake negative samples are samples with an initial concentration so low that the fluorescence will show after $d_{\rm cens}$, or will be confused with the artefact concentration.

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

$$Y = -\log_2(c + \epsilon_1) + \epsilon_2, \tag{4}$$

where $\epsilon_1$ is the law of the artefact, modelled as a log-normal distribution with parameters $(\nu, \tau^2)$; $\epsilon_2$ is an intrinsic measurement error on the threshold value $C_t$ measurement, modelled as a centred Gaussian random variable with variance $\rho^2$. In the idealized no false positive artefact limit ($\epsilon_1 \to 0$), the PCR threshold intensity of a negative patient ($c = 0$) is never reached ($Y \to \infty$).

As mentioned above, tests are considered to be reliably positive when the number of amplification cycles is below $d_{\rm cens}$. In other words, to avoid false positives, the threshold $d_{\rm cens}$ is chosen such that $\mathbb{P}(\epsilon_1 > 2^{-d_{\rm cens}}) \ll 1$. 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

$$Y(c) \approx \max(-\log_2(c), d_{\rm cens}) + \epsilon_2, \tag{5}$$

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

We now consider what happens when constructing a pooled sample of $N$ individuals. For each $i \leq N$, we write $Z_i = 1$ if $i$ is contaminated, $Z_i = 0$ otherwise, and $C_i$ for the concentration in viral RNA of each sample. A combined sample is created from a homogeneous mixing of the individual samples, the viral concentration in

**Box 1: A brief description of RT-qPCR tests**

We very briefly review some of the steps implemented during an RT-qPCR diagnostic procedure [4]:

1. The sample is treated so that a target RNA sequence (characteristic of the virus) is transcribed into DNA (reverse transcription);
2. The sample is placed in a PCR machine, which can measure the concentration of DNA of interest in the sample by making it fluorescent;
3. A reactive is added which approximatively doubles the number of DNA of interest at every cycle, driven by temperature changes;
4. The time series of the concentration in DNA over time is recorded; on a linear regression of of the logarithm of the fluorescent signal over time, one deduces an estimate of the viral concentration in the sample from the linear regression value at the origin.

the mixed sample is then $\frac{1}{N}\sum Z_i C_i$ (at least under the assumption that contaminated individuals have a reasonably high number of viral copies per sample, so that taking a portion $1/N$ of a contaminated sample brings a fraction $1/N$ of its viral charge). The result of the RT-qPCR measure of a grouped test with $N$ individuals then reads

$$Y^{(N)} = \max\left(\log_2 N - \log_2\left(\sum_{j=1}^{N} Z_j C_j\right), d_{\text{cens}}\right) + \epsilon_2. \tag{6}$$

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.

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.

## III. STATISTICAL ANALYSIS OF THE VIRAL LOAD MEASURED IN POSITIVE SAMPLES

In [28], the authors propose an analysis of SARS-CoV-2 viral load by patient age. Their analysis is based on the clinically measured viral load of a series of 3,712 infected patients. In particular, they presented in their Fig. 1 an histogram showing the frequency distribution of the (logarithm base 10 of the) viral load, estimated using RT-qPCR testing. We present in this section a method to analyse these data as a mixture of Gaussian random variables. As RT-qPCR does not allow the measure of viral load below a certain value, we take this fact into account

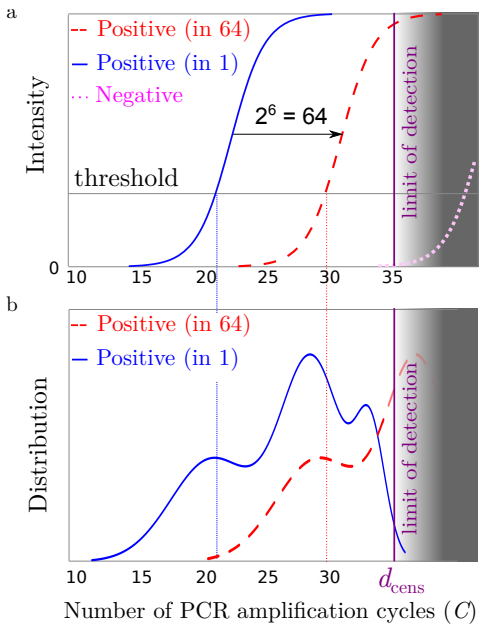



Figure 2: (a) Sketch of an RT-qPCR fluorescence intensity signal for a positive patient without pooling (sold red line) a single positive patient in a pool of 64 patients (dashed red curve) and for a negative sample representing the response of an artefact (dotted magenta curve); as pooling dilutes the 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 (sold 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 $d_{\text{cens}}$.

in our statistical modelling by considering the Gaussian in the mixture to be censored after a given threshold.

Unfortunately, as the clinical dataset of [28] is not available, we created a simulated set of measures which reproduces the distribution described by the histogram in Fig. 1 [28]. As the precise distribution of data points within each class of the histogram of Fig. 1 [28] is unknown, we assumed that points were distributed uniformly in their class. To check for dependency with respect to this simulated data, we made several independent reconstructions so that we can test the sensibility of the statistical analysis to the lack of detailed knowledge of the empirical distribution. The measure error $\epsilon_2$ of the qPCR being expected to be small with respect to the width of the histogram classes, we neglect it by setting $\rho = 0$ in the rest of the section.

The histogram of our simulated data is plotted in Fig. 3. It is similar to the histogram [28, Fig. 1] but with the $x$-axis graduated as $-\log_2$ of the viral concentration (to obtain an estimation of the attended $C_t$ value). We observe the presence of a wide and rather low first bump, centred around the value of 20. Additionally, two taller but less wide bell shapes seem to be present, centred

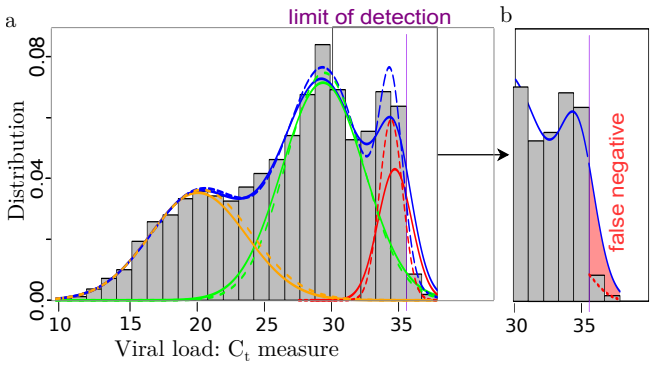



Figure 3: (a) Representation of the density for the classical mixing Gaussian model (dashed lines) and the partially censored model (solid lines) each composed as a sum of 3 components for the Gaussian model (orange/green/red dashed lines) and the partially censored model (orange/green/red solid lines); (purple vertical line) location of the threshold $d_{\text{cens}} \approx 35.6$. (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 $d_{\text{cens}}$ (red color filled area).

around the values 29 and 34 respectively. After the value of about 35.5, there is a significant drop in the number of registered values: this is probably due to the fact that one reaches the limit of detection by the machines used for the qPCR process, which can fluctuate depending on their tuning.

The presence of three well-marked distinct bell shapes in the histogram suggest to model the density of $-\log_2$ of the viral load of contaminated patients as a mixture of Gaussian distributions. Heuristically, the patients could be decomposed into several groups (that we call clusters), such that patients in cluster $i$ have a $C_t$ value distributed according to a normal distribution with parameters $(\mu_i, \sigma_i)$.

In the next section, we use a classical statistical tool to estimate the number of clusters and their parameters (frequencies, means, variances). We validate in particular that a decomposition in three clusters seems to be optimal, based on the data.

However, the proposed Gaussian modelling does not take into account the sudden drop observed for values around $C_t \sim 35.5$, hence the estimated density does not fit well to such drop. This seems to be due to the sudden loss of sensitivity of qPCR measure for low viral load. To better represent the variables, we thus introduce a (partially) censored Gaussian model in Section III B, which, to our knowledge, was not previously proposed.

Finally in Section III C, we construct a model consisting of a mixing of censored Gaussian variables, that we apply to the data of [28]. This model provides a better fit to the data with stable estimation of the parameters, and provides further validation of the previous decomposition of the patients into three clusters.

### A. Mixture model

The shape of the histogram Fig. 3 suggests that the law of the viral load of infected individuals is distributed according to a mixture of three or more Gaussian distributions. To estimate the parameters of this mixture, we use the R package *Rmixmod* (see [32]) using the *Expectation Maximisation* algorithm developed by [33]. To choose the correct number of clusters, we select the best model according to the *Bayesian Information Criterion* (see [34]).

Recall that we did not have access to raw data, but rather reconstructed similar data with same histogram of the distribution, randomizing the 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). The typical decomposition into three clusters is represented in dashed lines in Fig. 3(a). We also plotted in Fig. S4 the histogram with an example of the estimated density with 3 and 4 clusters.

We observe that the cluster associated to the lowest viral load, i.e. the highest $C_t$ value (in red) has a small variance. However, as recalled by [28], there is a loss of sensibility of the measure for very small viral loads, which can explain the drop in the number of detected cases around $d_{\text{cens}} = 35.6$. This motivates the introduction of the censored Gaussian model in the next section.

### B. Censored model and partially censored model

To model a partial lack of detection of low viral load ($C_t$ higher than a threshold $d_{\text{cens}}$), we introduce the partially censored Gaussian variable as a building block for the representation of the density of the viral load in infected patients. Recall that to avoid the risk of false positive, qPCR procedure is stopped after a certain number of cycles. However, depending on the machine tuning, the $C_t$ value of the sample can shift by an additive constant, which is computed by measuring the $C_t$ value of a standard solution of viral DNA to tare the measure. Then, some tests might allow the detection of lower viral loads than others.

In view of the shape of the histogram Fig. 3, it is reasonable to assume that the $C_t$ value of patients in a cluster follows a Gaussian distribution. To model this partial censorship phenomenon, we assume that if the sampled $C_t$ value is lower than the detection threshold $d_{\text{cens}}$, the measure is always made in the individual detected as contaminated. If the value is higher than $d_{\text{cens}}$, the sample will be detected with probability $q$, and its measure will

be registered. Otherwise, it will be discarded as a (false) negative, with probability $1-q$. The parameter $q$ represents the probability of detection of a viral load that falls below the detection threshold of some PCR measures.

*Remark* III.1. 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 $d_{\text{cens}}$.

We call this statistical model a *partially censored Gaussian model*, denoted by $\mathcal{CN}_{d_{\text{cens}}}(\mu,\sigma,q)$, with $\mu$ and $\sigma$ 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:

$$f_X(x)=\frac{f_{\mu,\sigma}(x)}{q+(1-q)F_{\mu,\sigma}(d_{\text{cens}})}\times\begin{cases}1 \text{ if } x\leq d_{\text{cens}},\\ q \text{ otherwise.}\end{cases} \tag{7}$$

*Remark* III.2. In the absence of censorship (i.e. in the limits $q\to 1$ or $d_{\text{cens}}\to+\infty$), we check that Eq. (7) converges to a Gaussian density distribution.

To avoid the problem of modelling of the partial censorship described in Remark III.1, a solution that we implement here as a comparison tool, is *to forget* the values after the threshold and to fit a *completely censored* model (i.e. with $q=0$) to the remaining data, that we denote by $\mathcal{CN}_{d_{\text{cens}}}(\mu,\sigma)=\mathcal{CN}_{d_{\text{cens}}}(\mu,\sigma,0)$, with density defined for every $x\in\mathbb{R}$ by

$$f_X(x)=\frac{f_{\mu,\sigma}(x)}{F_{\mu,\sigma}(d_{\text{cens}})}\mathbb{1}_{\{x\leq d_{\text{cens}}\}} \tag{8}$$

where $\mathbb{1}_{\{x\leq d_{\text{cens}}\}}$ is the indicator function equal to 1 if $x\leq d_{\text{cens}}$, and 0 otherwise.

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 [35]) which implements a Newton-type algorithm. The following Theorem III.3 guarantees the quality of the maximal likelihood estimators, hence of the estimations if the maximization procedure is done correctly.

**Theorem III.3.** *The estimators $(\widehat{\mu},\widehat{\sigma},\widehat{q})$ of $(\mu,\sigma,q)$ 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. [36, Chapter 9] and SI B 2). 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.

### C. Censored mixture model

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 [28] with a mixture model and a censoring threshold $d_{\text{cens}}\approx 35.6$ (so the two rightmost bars in the histogram of Fig. 3, 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 that tested individual. In Fig. 3, 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" Gaussian mixture model, constructed without taking into account the detection threshold. The principal difference between the naive model and the censored model is that the viral load of the population with the lowest load has a larger variance in the censored model, 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 Tab. III) and the model, we can calculate a theoretical false negative rate with the formula Eq. (S3): in this case, the value is approximately 3.8% (represented by the red area on the Fig. 3 (b)); it mostly belongs to the third cluster. Besides being based on simulated data, hence subject to caution, this value should be treated as a lower bound, as it is possible that a fourth cluster of infected individuals exists but with very small viral load, below the detection threshold. It allow us to predict for example that at least 150 clinical tests in the series of [28] might have resulted in a fake negative due to their low viral load.

To validate the censored model, we can verify that if one erases the data to the right of a certain value then use the totally censored model on the remaining data, a similar estimate should be obtained for the parameters. We display in Fig. S7 the density obtained using the censored mixture estimation with $d_{\text{cens}}\approx 35.6$, 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 means and standard deviations of the last component are almost

the same for $d_{\text{cens}} \approx 34.4$ and $d_{\text{cens}} \approx 35.6$ (see table IV); only the proportions naturally decrease with the threshold. On the other hand, the mean moves slightly to the left for $d_{\text{cens}} \approx 33.2$; this is due to the fact that we lose 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 (recall that we assume that it is an uniform distribution).

Note that if we were to treat the threshold $d_{\text{cens}} \approx 34.4$ or 33.2 as threshold for the partially censored model without erasing data, the optimization procedure `nlm` does not converge. This is further indication that a detection drop happens in the neighbourhood of 35.6.

The current analysis is consistent with a qualitative analysis of two other datasets [22, 37], whereby we typically find that the law of $C$ can be modelled as a log-normal distribution with a standard deviation $\sigma$ in the 5 to 6 range. But this large quantity of precise data allow us to extract three sub-populations with different measured viral concentration. It would be interesting to link these observed categories to characteristics of the individual (age, stage of the disease, physical condition, etc.).

## IV. GROUP TESTING: APPLICATION TO EPIDEMIOLOGICAL PROBLEMS

We now show how the previous analysis of the tests used to detect COVID-19 and the viral load to be expected in patients can be used to precise the epidemiological monitoring of the disease in the general population.

### A. Pooled sample viral concentration cannot be used to infer the precise number of infected individuals

The measured value of the pooled sample viral concentration cannot be used to estimate the number of infected individual within the pool. Indeed, in the previous section we found that the viral concentration in randomly selected infected individuals spans several order of magnitudes. Therefore, the PCR test will typically only return the concentration of the largest sample, with a drift of $\log_2(N)$ (cf. Fig. 2), using the approximation $\log_2(\frac{1}{N}\sum a_j) \approx \log_2(\max a_j) - \log_2(N)$.

As a result, the measure obtained from the test on the pool cannot be expect to give information on the number of contaminated individuals in that pool (which would have been possible otherwise in the context of a more concentrated distribution, e.g. with smaller $\sigma$). We point out that the RT-qPCR measure in the viral load of samples 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.

### B. Group testing and the measure of viral prevalence

We take in this section another look at the use of group testing for the measure of the prevalence of COVID-19 in a population. Here, in contrast with Sec. I, we no longer consider that the RT-qPCR tests used to detect viral charge in samples is perfect. As discussed in Eq. (6), 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.

#### 1. *Estimation of the false negative rate induced by pooling*

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. 2. A pooled sample returns positive only if the average concentration is smaller than $d_{\text{cens}}$; thus using the observation of Sec. III C, contamination will be detected in a group of $N$ individuals typically if at least one individual in the group has a viral charge larger than $N2^{-d_{\text{cens}}}$. Therefore, there is a risk that low viral charge samples (that would have been tested positive using individual tests) would no longer in pool tests.

Using (5), the increased rate of false negative due to pooling is given by

$$\mathbb{P}(-\log_2(C) + \epsilon_2 \leq d_{\text{cens}} - \log_2(N)), \tag{9}$$

Table I: 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. I) and imperfect tests (with false negatives; model parameters defined in Table II); computed using Eqs. (1) and (11).

| 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 |

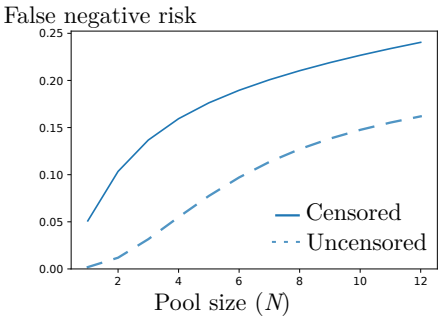



Figure 4: Estimated false negative risk probability of a sample pooling test containing a single infected individual (with a viral load distributed according to the simulated statistics presented in Sec. III) as a function of the number of pooled individuals $N$, using the estimated density obtained for the uncensored (dashed line) and censored (solid line). The censored and uncensored model give different estimates of (and probably both underestimate) the false negative rate in group pooling, has an impact on the estimation of the prevalence.

where $\log_2(C)$ is the viral concentration of the positive individual. For simplicity we ne the measurement error of the qPCR, i.e. considering that $\rho = 0$, thus approximation the increased rate of false negative by $1 - \Phi(d^{(N)}_{\text{cens}})$, where

$$\Phi(d^{(N)}_{\text{cens}}) = \mathbb{P}(-\log_2(C) \leq d_{\text{cens}} - \log_2(N)), \qquad (10)$$

Observe 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 partly due to the fact that the censored model makes the assumption that $d_{\text{cens}} \approx 35.6$, whereas $d_{\text{cens}} \approx 37.3$ in the uncensored model. Choosing a correct statistical model for the repartition of $C_t$ values has a critical impact on the estimation of the false negative risk, therefore on the estimate of the prevalence, as is discussed next.

### 2. *Correction of the prevalence estimate by the false negative rate*

Assuming a false negative rate of $1 - \Phi(d^{(N)}_{\text{cens}})$ 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 I) is a consistent estimator of $p\Phi(d^{(N)}_{\text{cens}})$ (c.f. Lemma I.1). As a result, the confidence interval constructed for the prevalence $p$ now reads

**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, denoted $\hat{p}_0$.
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 Lemma I.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.

$$\text{CI}_{1-\alpha}(p) = \left[ \frac{1 - (1 - \overline{X}^{(N)}_n)^{1/N}}{\Phi(d^{(N)}_{\text{cens}})} \right. \\ \left. \pm \frac{q_\alpha}{\sqrt{n}} \frac{(1 - \overline{X}^{(N)}_n)^{1/N-1} \sqrt{\overline{X}^{(N)}_n (1 - \overline{X}^{(N)}_n)}}{N\Phi(d^{(N)}_{\text{cens}})} \right]. \qquad (11)$$

For the numerical applications presented in Fig. 1 and Table I, we consider a viral load $C$ that is distributed according to Eq. (8). As expected, due to false negatives, we find that the number of tests needed to reach a given precision on the prevalence is increased, but relatively moderately. In particular, the optimal pool size value, $N^{(\text{imper})}_{\text{opt}}$, that minimizes the number of tests needed to reach a given precision level, is close to the value $N^{(\text{perf})}_{\text{opt}}$, defined in Eq. (3).

Similarly, one can observe that using a different distribution with similar mean and variance for $-\log_2 C$ as Eq. (8) would lead to moderate changes to the values estimated in Table I. 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 in the viral load distribution.

Based on Eq. (11), 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.

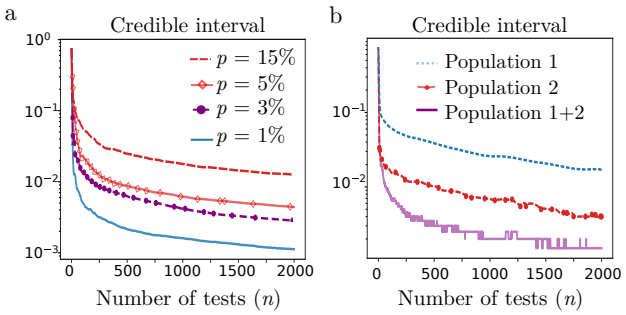



Figure 5: (a) Width of the 95% credible interval on the prevalance $p$ with adaptative Bayesian sampling as a function of the number of tests $n$ for a set of values in the prevalence ranging from $p = 15\%$ (top, magenta dashed line) to $p = 1\%$ (bottom, blue solid line). (b) Width of the credible 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).

### C. 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 [39], which we expect to be significantly higher than the estimate prevalence within the general population.

Here we show that we do not need 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 Fig. 5, we consider the problem of estimating the prevalence within two categories of the population: one at a prevalence $p_1 = 5\%$ representing 20% of the total population [39], the other being at a prevalence $p_2 = 0.5\%$ [38]. More information on this adaptative protocol can be found in Section A 2.

*Remark* IV.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 of the other, to attain similar levels of precisions 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.

### D. Group testing for regular monitoring in communities

We now consider some applications of group testing to the early detection of an epidemic outbreak within a community, that is interconnected and reasonably closed to the outside (work offices, schools, retirement homes, detention centres). Our aim is to emphasize some properties of the screening procedures that could be put in place to detect an outbreak before it has the time to spread. We first consider the impact of using multiple tests and the size of the group tested to detect the existence of a contaminated individual in a group that is not contaminated. We then study the effect of the regularity of testing, comparing e.g. the strategy of testing the whole community every $\tau$ days with the method consisting in testing at random a portion $\tau^{-1}$ of this community every day.

#### 1. *Optimization of the size of pools*

Here we study a working model to choose the optimal size of pool to make in order to detect the existence of an infected individual in a community. We consider a community of $A$ individuals in which we assume the presence of a unique contaminated individual (patient 0), that carries a viral load $C$. According to the modelling in Eq. (6), if we make $k$ pools of $N$ individuals in the community, the probability to detect the patient 0 reads

$$\mathbb{P}\left[+|k \text{ tests}\right] = kN\Phi_0(d_{\text{cens}}^{(N)})/A, \quad \text{with } kN \leq A, \quad (12)$$

where $\Phi_0(d_{\text{cens}}^{(N)})$ is defined according to Eq. (10), 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. (8) based on clinical data. For simplicity, we will assume in the following that $\Phi_0$ is the cumulative distribution of a log-normal distribution $\log\mathrm{N}(\mu_0, \sigma_0)$ of mean $\mu_0$ and variance $\sigma_0$.

In Fig. 6, we represent the evolution of 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 $\mu_0$ is close enough to $d_{\text{cens}}$, i.e. if the viral charge of the patient 0 is close to being undetectable, then there will exist an optimal size for the pools to be sampled. When $N$ becomes too large the risk of false negative overcomes the potential benefits of testing larger portions of the community (see Fig. 6a). 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.

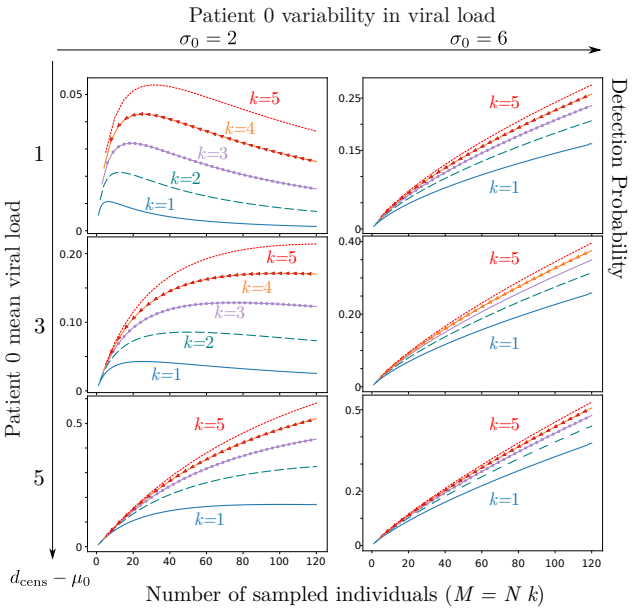



Figure 6: Detection probability of a single patient 0 with low viral load within a community of 120 as a function of the total number of sampled individuals $M = k \times N$, where $k$ is the total number of tests used and $N$ the number of samples pooled together in a test, with $k = 5$ (red dotted 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 $\sigma_0$ and a mean $\mu_0$ and a threshold at a value denoted $d_{\text{cens}}$ 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) representing an patient 0 with typical concentration; $\sigma_0 = 2$ (left column); $\sigma_0 = 6$ (right column).

### 2. *Optimization of the regularity of tests*

We now put ourselves in the context of the prevention of an epidemic outbreak in a closed community. Our aim is to control that no contamination occurred within that community at any time, and to detect as soon as possible the presence of the contamination. We consider a protocol consisting in testing, at regular intervals, members of the community at random to detect if any of them gets contaminated. We show that for a constant budget of tests per unit of time, the most efficient strategy is to spread the tests as uniformly as possible.

We consider a continuous-time steady-state version of this situation. A community of $A$ individuals is being monitored by regular testing every $\tau$ units of time, and we observe the impact of this parameter $\tau$. Every $\tau$ units of time, $k \times N$ individuals are sampled at random in the population to be tested in $k$ pools. To compare the impact of the frequency of testing for a constant budget, we assume that $k/\tau$ is equal to a fixed constant, representing the amount of tests spent per unit of time by the community to detect infections.

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

| Symbol | Meaning | Value |
|---|---|---|
| $d_{\text{cens}}$ | Maximal cycle number | 35.6 |
| $\mu_i, \sigma_i, p_i$ | Viral load distribution fits | Table IV |
| $\rho$ | PCR measurement error (std.) | neglected |
| $\phi$ | Delay before onset of symptoms | 5 days |
| $\lambda$ | Intra-community contamination rate | 0.5 days$^{-1}$ |
| $r$ | Asymptomatic probability | 40 % |
| $\tau$ | Time interval between grouped tests | $1 - 10$ days |
| $A$ | Total number in the community | 120 or 1000 |
| $N$ | Pool size | 1–128 |
| $\mu_0$ ; $\sigma_0$ | Viral load of patient 0 (mean, std.) | Variable |

At a random date, which we choose to be time $t = 0$, the patient 0 in the population gets contaminated, and immediately starts infecting members of its community at rate $\lambda$. Each newly infected individual then contributes to spreading the disease at the same rate $\lambda$. In this simplified model, we do not model the variability of the infectivity of individuals over time.

We denote by $r$ the proportion of asymptomatic individuals. Symptomatic individuals start showing signs of being contaminated a given number of days after infection, denoted by $\phi$ (that we assume to be constant in this simplified model). The outbreak is detected either if one of the screening tests finds a contaminated individual, or if one of the contaminated individuals shows symptoms. We denote by $T_s$ the time at which the first individual in the population shows symptoms and $T_d$ the detection time of contaminated individuals within the community.

In this model, the number of infected individuals at date $t$, denoted by $N(t)$, is distributed according to a geometric random variable with parameter $1 - e^{-\lambda t}$, as expected for such a Yule process [40]. In the absence of screening tests, the average detection time is $T_s$, the time at which first symptoms appear, which by straightforward computations satisfy $\langle T_s \rangle = \phi - \log(1 - r)/\lambda$. The average number of contaminated individuals at that time then reads: $\langle N(T_s) \rangle = e^{\lambda\phi}/(1 - r)$. In comparison, based on Eq. (12), we find that the first screening test after contamination will detect the outbreak with a probability

$$\mathbb{P}\left[+|k \text{ tests}\right] = \frac{k(e^{\lambda\tau} - 1)\Phi(d_{\text{cens}}^{(N)})}{\lambda\tau A}. \tag{13}$$

By the time of detection $t = \tau$, a number of $\langle N(\tau) \rangle = (e^{\lambda\tau} - 1)/\lambda\tau$ individuals has been contaminated.

In Fig. 7, we compare different screening scenarios for a large community composed of $A = 1000$ individuals. We vary the value of the screening time interval $\tau$ 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.

Based on estimations in the average number of contaminated 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 the whole community at less frequent time intervals. Unless the viral load of presymptomatic patients is extremely small, making a pool with as large a subgroup of the community as possible always improves the outcome. However, the gain made by increasing the size of the sample decreases after a certain value (e.g. around 40 individuals in Fig. 7, representing approximatively 4% of the population of $A = 1000$).

We consider here 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 charge for patients after their contamination, 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, Ref. [41] also concludes on the efficiency of group testing in preventing epidemic outbreaks in health care structures.

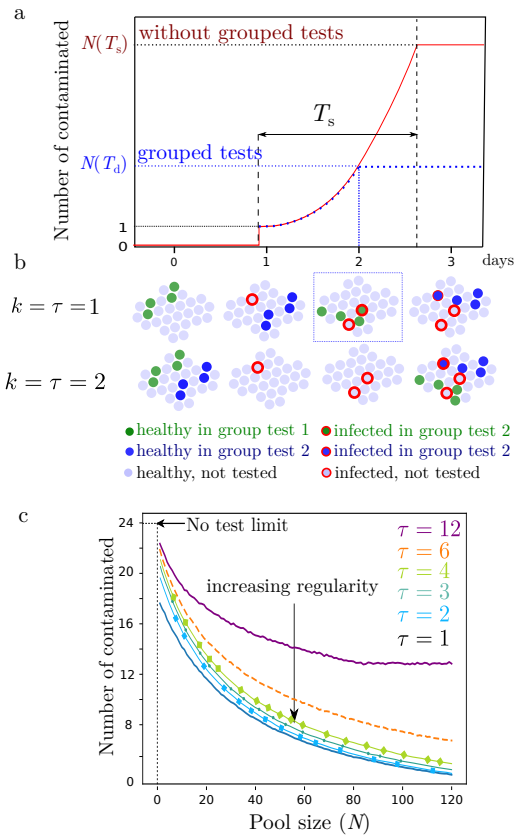



Figure 7: (a) Sketch of the time evolution of the number of contaminated individuals in a community. The patient 0 is contaminated from the outside of the community 0.8 units of time after a test date. In the absence of screening tests, the contamination is detected at the time $T = T_s$ (after appearance 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 contamination. (c) Number of infected individuals at the detection of the outbreak as a function of the pool size, using $k = \tau$ tests performed at $\tau$-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 follows a Gaussian distribution with mean $\mu_0 = 30$ and standard deviation $\sigma_0 = 2\log_2(2)$; all others parameters can be found in Table II.

## Conclusion

We would to emphasize here that, due to the lack of detailed datasets to analyse and the theoretical nature of the considerations of that article, our results should not be considered as conclusive, or used as a baseline for clinical practice/healthcare related behaviour.

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 charge of patients. Inspired by the statistical results presented in [28], we expect the statistics of the number of amplification cycles to be well described as a mixture of 3 Gaussian variables censored at the RT-qPCR sensibility limit. Our analysis may hint at the existence of 3 sub-classes of interest that could each be interpreted in terms of medical or physiological criteria [37].

We point out that the viral load distribution that we determine here based on the clinical sampling presented in [28] is possibly biased as compared to the viral load that would be measured within a population targeted by a group testing strategy. Indeed, asymptomatic individuals may exhibit a different viral load distribution than the population that got tested. Furthermore, the measures of [28] were based on nasal swabs samples, and their statistical distribution is likely to vary according to the sampling method chosen for group testing.

Tests based on saliva samples appear to show promising results in terms of false negative risks while improving the acceptance of tests among the population and reducing the time needed for sample collection and the exposure of health care practitioners [42–46]. In this context, we are looking forward to seeing whether group testing on saliva samples would provide reliable results; statistics on the distribution in viral load in saliva samples would then be of interest to evaluate the efficiency of a group testing strategy based on such samples.

We believe 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 and potential eradication of the epidemic [10, 47, 48].

**Code accessibility** The codes used in this paper can be found at the following address: `https://github.com/Lionning/CovidPooling`

**Acknowledgements.** We wish to thank members of the MODCOV initiative and in particular Françoise Praz and Florence Debarre who gave us numerous helpful comments on the first version of this manuscript. We also thank Marie-Claude Potier and Marc Sanson for insightful discussions on RT-qPCR tests as well as on the interest of saliva samples.

[1] H. Salje, C. Tran Kiem, N. Lefrancq, N. Courtejoie, P. Bosetti, J. Paireau, A. Andronico, N. Hozé, J. Richet, C.-l. Dubost, Y. Le Strat, J. Lessler, D. Levy-Bruhl, A. Fontanet, L. Opatowski, P.-Y. Boelle, and S. Cauchemez, Science **3517**, 3517 (2020).
[2] R. Wölfel, V. M. Corman, W. Guggemos, M. Seilmaier, S. Zange, M. A. Müller, D. Niemeyer, T. C. Jones, P. Vollmar, C. Rothe, M. Hoelscher, T. Bleicker, S. Brünink, J. Schneider, R. Ehmann, K. Zwirglmaier, C. Drosten, and C. Wendtner, Nature (2020), 10.1038/s41586-020-2196-x.
[3] World Health Organization, WHO - Interim guidance **2019**, 1 (2020).
[4] V. M. Corman, O. Landt, M. Kaiser, R. Molenkamp, A. Meijer, D. K. Chu, T. Bleicker, S. Brünink, J. Schneider, M. L. Schmidt, D. G. Mulders, B. L. Haagmans, B. van der Veer, S. van den Brink, L. Wijsman, G. Goderski, J.-L. Romette, J. Ellis, M. Zambon, M. Peiris, H. Goossens, C. Reusken, M. P. Koopmans, and C. Drosten, Eurosurveillance **25**, 1 (2020).
[5] N. Speybroeck, B. Devleesschauwer, L. Joseph, and D. Berkvens, International Journal of Public Health **58**, 791 (2013).
[6] K. Mizumoto, K. Kagaya, A. Zarebski, and G. Chowell, Eurosurveillance **25**, 1 (2020).
[7] Q. Bi, Y. Wu, S. Mei, C. Ye, X. Zou, Z. Zhang, X. Liu, L. Wei, S. A. Truelove, T. Zhang, W. Gao, C. Cheng, X. Tang, X. Wu, Y. Wu, B. Sun, S. Huang, Y. Sun, J. Zhang, T. Ma, J. Lessler, and T. Feng, The Lancet Infectious Diseases **3099**, 1 (2020).
[8] Y. Bai, L. Yao, T. Wei, F. Tian, D.-Y. Jin, L. Chen, and M. Wang, JAMA **323**, 1406 (2020).
[9] E. Lavezzo, E. Franchin, C. Ciavarella, G. Cuomo-dannenburg, L. Barzon, M. Sciro, S. Merigliano, E. Decanale, M. C. Vanuzzo, F. Onelia, M. Pacenti, S. Parisi, G. Carretta, D. Donato, K. A. M. Gaythorpe, I. College, L. C. Response, and R. Alessandra, medRxiv , 1 (2020).
[10] L. Ferretti, C. Wymant, M. Kendall, L. Zhao, A. Nurtay, L. Abeler-Dörner, M. Parker, D. Bonsall, and C. Fraser, Science **368**, 6936 (2020).
[11] N. Sethuraman, S. S. Jeremiah, and A. Ryo, JAMA (2020).
[12] C. Gollier and O. Gossner, Covid Economics , 32 (2020).
[13] K. B. Pouwels, L. S. J. Roope, A. Barnett, D. J. Hunter, T. M. Nolan, and P. M. Clarke, PharmacoEconomics - Open , 2 (2020).
[14] R. Dorfman, The Annals of Mathematical Statistics (1943).
[15] M. Aldridge, O. Johnson, and J. Scarlett, Foundations and Trends in Communications and Information Theory **15**, 196 (2019).
[16] E. Barillot, B. Lacroix, and D. Cohen, Nucleic Acids Research **19**, 6241 (1991).
[17] N. Thierry-Mieg, Nature Methods **3**, 161 (2006).
[18] T. Furon, [Research Report] RR-9164, Inria Rennes Bretagne Atlantique , 1 (2018).
[19] C. A. Hogan, M. K. Sahoo, and B. A. Pinsky, JAMA **323**, 1967 (2020).
[20] S. Lohse, T. Pfuhl, B. Berkó-Göttel, J. Rissland, T. Geißler, B. Gärtner, S. L. Becker, S. Schneitler, and S. Smola, The Lancet Infectious Diseases **3099** (2020).
[21] I. Torres, E. Albert, and D. Navarro, Journal of Medical Virology , 25971 (2020).
[22] I. Yelin, N. Aharony, E. Shaer-Tamar, A. Argoetti, E. Messer, D. Berenbaum, E. Shafran, A. Kuzli, N. Gandali, T. Hashimshony, Y. Mandel-Gutfreund, M. Halberthal, Y. Geffen, M. Szwarcwort-Cohen, and R. Kishony, medRxiv (2020).
[23] R. Ben-Ami, A. Klochendler, M. Seidel, T. Sido, O. Gurel-Gurevich, M. Yassour, E. Meshorer, G. Benedek, I. Fogel, E. Oiknine-Djian, A. Gertler, Z. Rotstein, B. Lavi, Y. Dor, D. G. Wolf, M. Salton, and Y. Drier, medRxiv , 2020.04.17.20069062 (2020).
[24] N. Shental, S. Levy, S. Skorniakov, V. Wuvshet, Y. Shemer-Avni, A. Porgador, and T. Hertz, medRxiv , 2020.04.14.20064618 (2020).
[25] A. Deckert, T. Bärnighausen, and N. Kyei, (2020), 10.2471/BLT.20.257188.
[26] N. Sinnott-Armstrong, D. Klein, and B. Hickey, medRxiv (2020), 10.1101/2020.03.27.20043968.
[27] K. Narayanan, I. Frost, A. Heidarzadeh, K. K. Tseng, S. Banerjee, J. John, and R. Laxminarayan, medRxiv , 2020.04.03.20051995 (2020).
[28] T. C. Jones, B. Mühlemann, V. Talitha, Z. Marta, J. Hofmann, A. Stein, A. Edelmann, V. M. Corman, and C. Drosten, Preprint Charité Hospital (2020).
[29] K. H. Thompson, Biometrics **18**, 568 (1962).
[30] A. Forootan, R. Sjöback, J. Björkman, B. Sjögreen, L. Linz, and M. Kubista, Biomolecular Detection and Quantification **12**, 1 (2017).
[31] A. Ruiz-Villalba, E. van Pelt-Verkuil, Q. D. Gunst, J. M. Ruijter, and M. J. van den Hoff, Biomolecular Detection and Quantification **14**, 7 (2017).
[32] R. Lebret, S. Iovleff, F. Langrognet, C. Biernacki, G. Celeux, and G. Govaert, Journal of Statistical Software **67** (2015).
[33] A. P. Dempster, N. M. Laird, and D. B. Rubin, Journal of the Royal Statistical Society: Series B (Methodologi-

cal) **39**, 1 (1977).
[34] G. Schwarz, The Annals of Statistics (1978), 10.1214/aos/1176344136.
[35] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria (2020).
[36] O. E. Nielsen, *Information and exponential families : in statistical theory* (John Wiley & Sons, Chichester U.K. New York, 2014).
[37] Y. Liu, L.-M. Yan, L. Wan, T.-X. Xiang, A. Le, J.-M. Liu, M. Peiris, L. L. M. Poon, and W. Zhang, The Lancet. Infectious diseases (2020).
[38] D. F. Gudbjartsson, A. Helgason, H. Jonsson, O. T. Magnusson, P. Melsted, G. L. Norddahl, J. Saemundsdottir, A. Sigurdsson, P. Sulem, A. B. Agustsdottir, B. Eiriksdottir, R. Fridriksdottir, E. E. Gardarsdottir, G. Georgsson, O. S. Gretarsdottir, K. R. Gudmundsson, T. R. Gunnarsdottir, A. Gylfason, H. Holm, B. O. Jensson, A. Jonasdottir, F. Jonsson, K. S. Josefsdottir, T. Kristjansson, D. N. Magnusdottir, L. le Roux, G. Sigmundsdottir, G. Sveinbjornsson, K. E. Sveinsdottir, M. Sveinsdottir, E. A. Thorarensen, B. Thorbjornsson, A. Löve, G. Masson, I. Jonsdottir, A. D. Möller, T. Gudnason, K. G. Kristinsson, U. Thorsteinsdottir, and K. Stefansson, New England Journal of Medicine , NEJMoa2006100 (2020).
[39] E. S. Barrett, D. B. Horton, J. Roy, M. L. Gennaro, A. Brooks, J. Tischfield, P. Greenberg, T. Andrews, S. Jagpal, N. Reilly, M. J. Blaser, J. Carson, and R. A. Panettieri, medRxiv , 2020.04.20.20072470 (2020).
[40] S. Meleard, *Modèles aléatoires en Ecologie et Evolution*, edited by CMAP (2016).
[41] D. R. Smith, A. Duval, K. B. Pouwels, D. Guillemot, J. Fernandes, B.-T. Huynh, L. Temime, and L. Opatowski, medRxiv (2020).
[42] A. L. Wyllie, J. Fournier, A. Casanovas-Massana, M. Campbell, M. Tokuyama, P. Vijayakumar, B. Geng, M. C. Muenker, A. J. Moore, C. B. F. Vogels, M. E. Petrone, I. M. Ott, P. Lu, A. Lu-Culligan, J. Klein, A. Venkataraman, R. Earnest, M. Simonov, R. Datta, R. Handoko, N. Naushad, L. R. Sewanan, J. Valdez, E. B. White, S. Lapidus, C. C. Kalinich, X. Jiang, D. J. Kim, E. Kudo, M. Linehan, T. Mao, M. Moriyama, J. E. Oh, A. Park, J. Silva, E. Song, T. Takahashi, M. Taura, O.-E. Weizman, P. Wong, Y. Yang, S. Bermejo, C. Odio, S. B. Omer, C. S. D. Cruz, S. Farhadian, R. A. Martinello, A. Iwasaki, N. D. Grubaugh, and A. I. Ko, medRxiv , 2020.04.16.20067835 (2020).
[43] L. Azzi, G. Carcano, F. Gianfagna, P. Grossi, D. D. Gasperina, A. Genoni, M. Fasano, F. Sessa, L. Tettamanti, F. Carinci, V. Maurino, A. Rossi, A. Tagliabue, and A. Baj, Journal of Infection , 1 (2020).
[44] K. K. W. To, O. T. Y. Tsang, W. S. Leung, A. R. Tam, T. C. Wu, D. C. Lung, C. C. Y. Yip, J. P. Cai, J. M. C. Chan, T. S. H. Chik, D. P. L. Lau, C. Y. C. Choi, L. L. Chen, W. M. Chan, K. H. Chan, J. D. Ip, A. C. K. Ng, R. W. S. Poon, C. T. Luo, V. C. C. Cheng, J. F. W. Chan, I. F. N. Hung, Z. Chen, H. Chen, and K. Y. Yuen, The Lancet Infectious Diseases **20**, 565 (2020).
[45] E. Williams, K. Bond, B. Zhang, M. Putland, and D. A. Williamson, Journal of Clinical Microbiology **50** (2020), 10.1128/JCM.00776-20.
[46] Z. Khurshid, S. Zohaib, C. Joshi, S. F. Moin, M. Sohail, D. J. Speicher, D. Implantology, K. Faisal, M. Sciences, M. Al, B. Sciences, and B. Sciences, (2020).
[47] T. A. Treibel, C. Manisty, M. Burton, Á. McKnight, J. Lambourne, J. B. Augusto, X. Couto-Parada, T. Cutino-Moguel, M. Noursadeghi, and J. C. Moon, The Lancet **6736**, 19 (2020).
[48] S. M. Kissler, C. Tedijanto, E. Goldstein, Y. H. Grad, and M. Lipsitch, Science **5793**, eabb5793 (2020).

## Supplementary Information

Group testing as a strategy for the epidemiologic monitoring of COVID-19

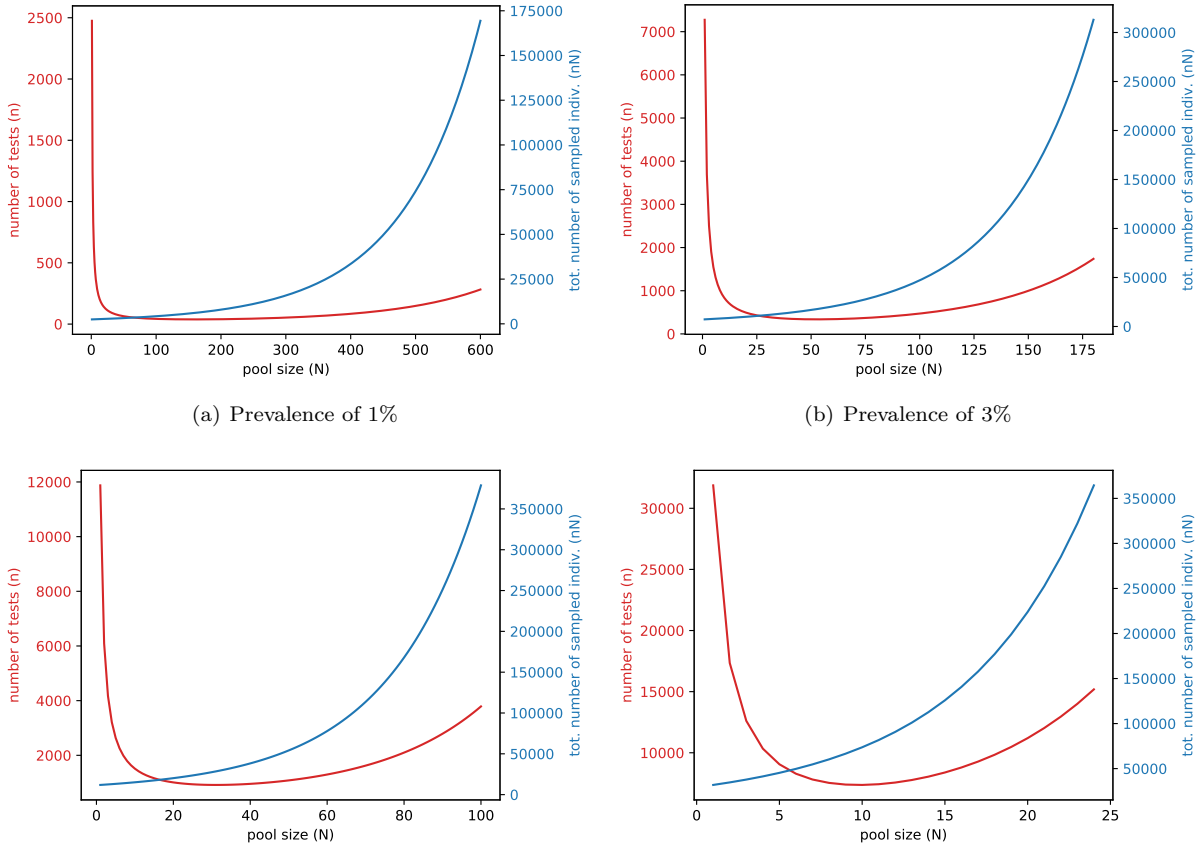
Vincent Brault, Bastien Mallein, Jean-Francois Rupprecht

### Appendix A: Ideal tests

We present here some of results obtained from the computations made in Sec. I, 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.

#### 1. 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.




(a) Prevalence of 1%

(b) Prevalence of 3%

(c) Prevalence of 5%

(d) Prevalence of 15%

Figure S1: 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.

### 2. Bayesian inference

We are now interested in a Bayesian approach to the measure of prevalence. We started with an initial prior distribution with density $f_0(p) = 6p(1-p)\mathbf{1}_{\{0\le p\le 1\}}$ 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 $j$th test computed as $\left\lfloor -\frac{c_\star}{\log(1-\overline{p}_{j-1})} \right\rfloor$ (cf. Eq. (3));
3. choose $N_j$ individuals at random and test them in a group:
   - if the test is positive, then $f_j(p) = C_j^+(1-(1-p)^{N_j})f_{j-1}(p)$;
   - if the test is negative, then $f_j(p) = C_j^-(1-p)^{N_j} f_{j-1}(p)$;

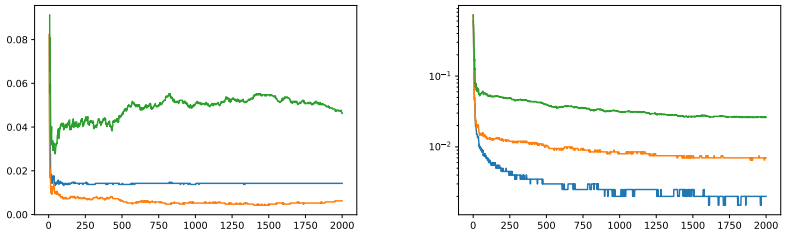
   with $C_j^\pm$ normalizing constants, chosen such that $\int_0^1 f_j(p)\mathrm{d}p = 1$.

We trace in Fig. S2 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 $j$th 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. S2 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 $p$.



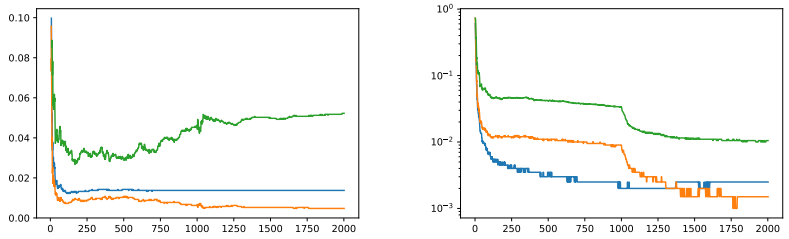

(a) Median value of the priors, overall population in blue, first resp second population in orange resp green.

(b) Width of the 95% credible intervals

Figure S2: 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%.

However, observe that even with a naive group construction (without segregating individuals according to their sub-population), one can extract information on the 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. S3, with the same number of tests a much more detailed estimate of the prevalence is obtained.



(a) Median value of the priors, overall population in blue, first resp second population in orange resp green.

(b) Width of the 95% credible intervals

Figure S3: 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%. 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.

## Appendix B: Censored Gaussians

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

### 1. Naive method based on mixing models

In this section, we trace the density estimated by a simple mixture of Gaussian variable presented in Sec. III A. An estimation of the parameters of this mixture are given in Table III.

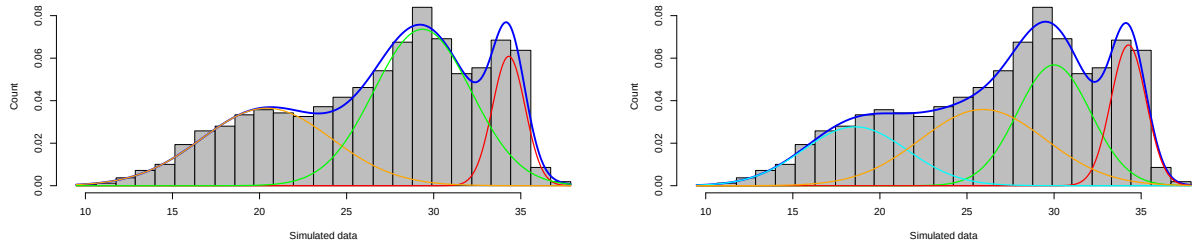



Figure S4: Representation of the histogram with the densities estimated with 3 classes (on the left) and 4 classes (on the right): the color lines (other than blue) represent the density of each component and the blue line the density of the mixture.

### 2. Proof of the Theorem III.3

To prove the lemma III.3, 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, $\eta = \left(\frac{\mu}{\sigma^2},\ -\frac{1}{2\sigma^2},\ \ln q\right)$ the natural parameters, $T(x) = \left(x,\ x^2,\ \mathbb{1}_{\{x>d_{\text{cens}}\}}\right)$ 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 (S1), the (partially) censured model belongs to the family of exponential laws and the maximum likelihood estimators are strongly consistent and asymptotically normal.

### 3. Simulations for Sec. III B

To study the quality of the estimators defined in Sec. III B, we simulated $10^4$ samples of size $n \in \{10^2, 10^3, 10^4, 10^5\}$ of variables following the model $\mathcal{CN}_{d_{\text{cens}}}(0, 1, p)$ 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. S5 and a zoom on significant part in Fig. S6. 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. S5 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 $d_{\text{cens}}$ 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 $d_{\text{cens}}$ 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 $p$.

### 4. Censored mixture model

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

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

The completely censured mixture model has the same density than the Eq (S2) with $q_k = 0$.

With the model (S2), we can estimate the theoretical false negative rate by the following formula:

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

Table III: Estimated parameters for the naive Gaussian mixture fit and the censored Gaussian mixture fits defined in Eq (S2).

| Model | $q_1$ | $\mu_1$ | $\sigma_1$ | $\pi_1$ | $q_2$ | $\mu_2$ | $\sigma_2$ | $\pi_2$ | $q_3$ | $\mu_3$ | $\sigma_3$ | $\pi_3$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Naive | | 20.41 | 3.74 | 0.34 | | 29.43 | 2.81 | 0.52 | | 34.32 | 0.89 | 0.14 |
| Partially | 0.21 | 20.14 | 3.60 | 0.32 | 0.19 | 29.35 | 2.96 | 0.53 | 0.20 | 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 |

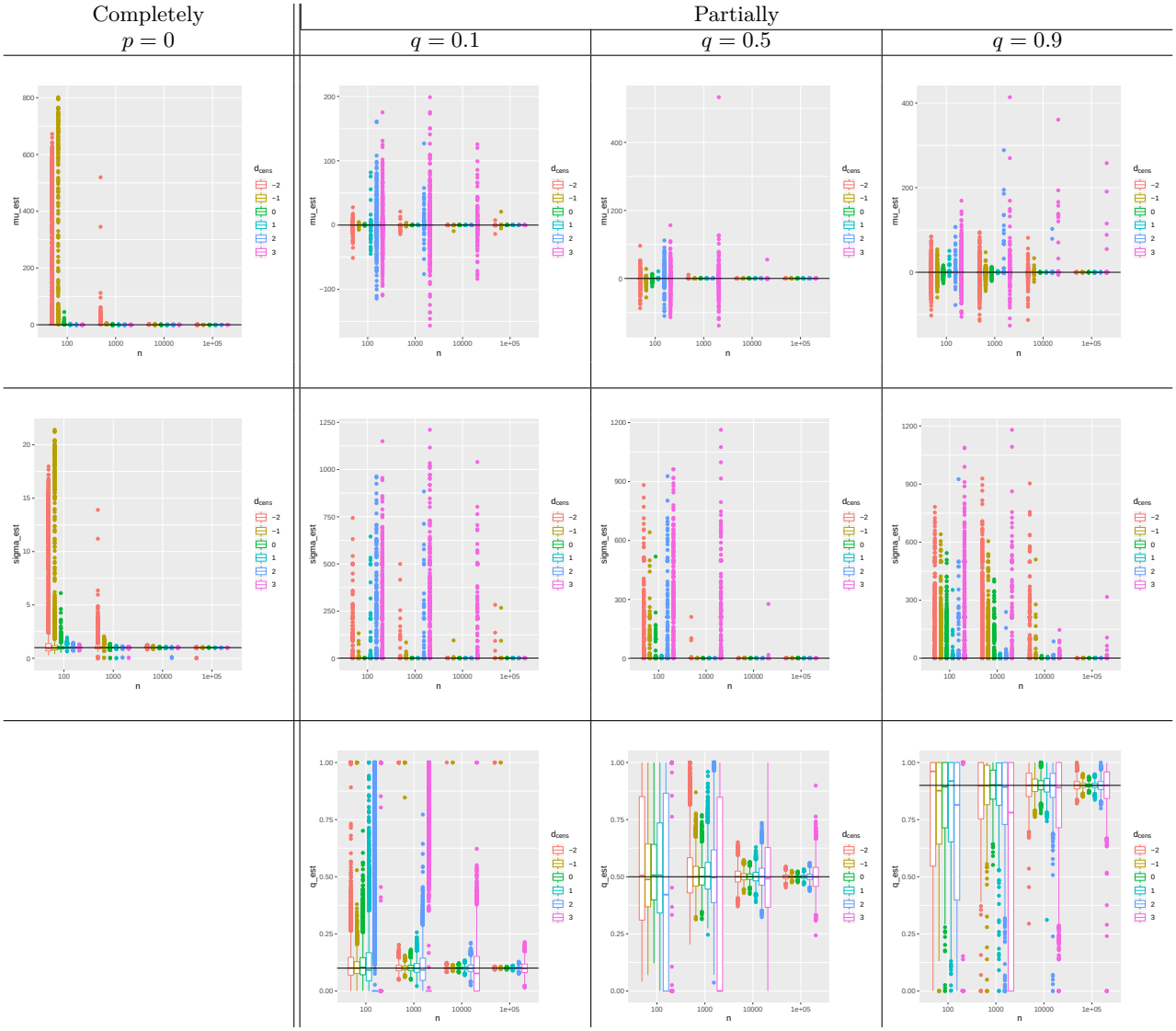



Figure S5: Boxplots of the estimations of $\mu$ (first row), $\sigma$ (second row) and $p$ (last row ; only for partially censored model) in function of model (columns), the size $n$ of sample (x-axis) and the value of the threshold $s$ (color). The true value is symbolised by the horizontal black line.

Table IV: Estimated parameters for the censored Gaussian mixture fit define in Eq (S2) for different values of the threshold $d_{\text{cens}}$, applied to reconstructed data data with same distribution as in [28] erased above $d_{\text{cens}}$.

| $d_{\text{cens}}$ | $\mu_1$ | $\sigma_1$ | $\pi_1$ | $\mu_2$ | $\sigma_2$ | $\pi_2$ | $\mu_3$ | $\sigma_3$ | $\pi_3$ |
|---|---|---|---|---|---|---|---|---|---|
| 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 |

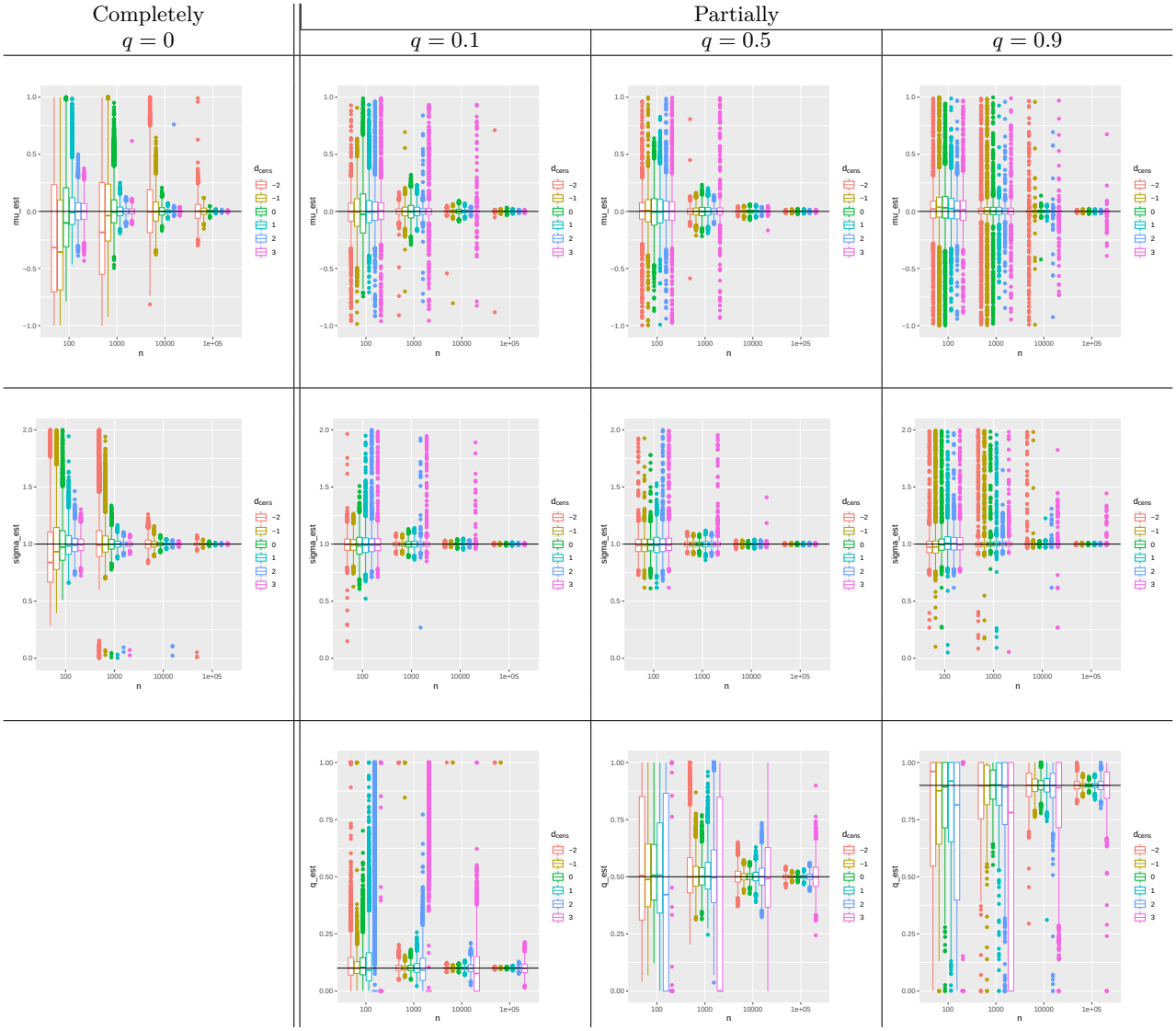



Figure S6: Zoom on boxplots of the estimations of $\mu$ (first row), $\sigma$ (second row) and $q$ (last row ; only for partially censored model) in function of model (columns), the size $n$ of sample (x-axis) and the value of the threshold $s$ (color). The true value is symbolised by the horizontal black line.

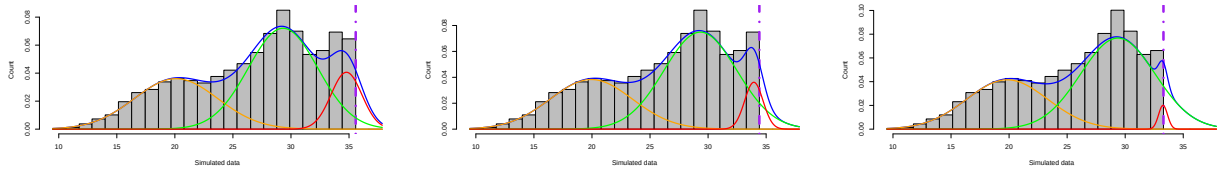



Figure S7: Density of the fits of the censored model with three components (obtained when erasing data to the right of the threshold) with a threshold at 35.6 (left), 34.4 (middle) and 33.2 (right): the orange, green and red lines represent the density of each component and the blue line the density of the mixture.