COVID-19 prevalence estimation by random sampling in population - Optimal sample pooling under varying assumptions about true prevalence

Ola Brynildsrud [1,2]

[1] Norwegian Institute of Public Health, Oslo, Norway

[2] Norwegian University of Life Science, Ås, Norway

Contact: olbb@fhi.no

ABSTRACT

Background: The number of confirmed COVID-19 cases divided by population size is used as a coarse measurement for the burden of disease in a population. However, this fraction depends heavily on the sampling intensity and the various test criteria used in different jurisdictions, and many sources indicate that a large fraction of cases tend to go undetected.

Methods: Estimates of the true prevalence of COVID-19 in a population can be made by random sampling. Here I use simulations to explore confidence intervals of prevalence estimates under different sampling strategies, exploring optimal sample sizes and degrees of sample pooling at a range of true prevalence levels.

Results: Sample pooling can greatly reduce the total number of tests required for prevalence estimation. In low-prevalence populations, it is theoretically possible to pool hundreds of samples with only marginal loss of precision. Even when the true prevalence is as high as 10% it can be appropriate to pool up to 15 samples, although this comes with the cost of not knowing which patients were positive. Sample pooling can be particularly beneficial when the test has

imperfect specificity can provide more accurate estimates of the prevalence than an equal number of individual-level tests.

Conclusion: Sample pooling should be considered in COVID-19 prevalence estimation efforts.

BACKGROUND

It is widely accepted that a large fraction of COVID-19 cases go undetected. A crude measure of population prevalence is the fraction of positive tests at any given date. However, this is subject to large ascertainment bias since tests are typically only ordered from symptomatic cases, whereas a large proportion of infected might show little to no symptoms [1,2]. Non-symptomatic infections can still shed the Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) virus and are therefore detectable by reverse transcriptase polymerase chain reaction (RT-PCR)-based tests. It is therefore possible to test randomly selected individuals to estimate the true disease prevalence in a population. However, if the disease prevalence is low, very little information is garnered from each individual test. Under such situations it can be advantageous to pool individual patient samples into a single pool [3]. Pooling strategies can be efficient to increase the test capacity and are less wasteful with ingredients required for a RT-PCR-based test.

METHODS

I simulated the effect sample pooling had on prevalence estimates under five different settings for true prevalence, p. I started by generating a population of 500,000 individuals and then let each individual have p probability of being infected at sampling time. The number of patient

52 samples collected from the population is denoted by n, and the number of patient samples that
53 are pooled into a single well is denoted by k. The total number of pools are thus $\frac{n}{k}$, hereby called
54 m. The number of positive pools in an experiment is termed x. I calculated the estimated
55 prevalence $\hat{p}$ at each parameter combination by replicating the experiment 10,000 times and
56 report here the 2.5% and 97.5% quantiles of the distribution of $\hat{p}$.
57
58 Explored parameter options:
59 $p \in \{0.001, 0.003, 0.01, 0.03, 1.0\}$
60 $n \in \{200, 500, 1000, 1500, 2000, 3000, 5000\}$
61 $k \in \{1, 3, 5, 7, 10, 15, 20, 25, 30, 40, 50, 70, 100, 200\}$
62
63 I considered the specificity (θ) of a PCR-based test to be 1.0 but include simulations with the
64 value set to 0.99. Test sensitivity (η) depends on a range of uncontrollable factors such as virus
65 quantity, sample type, time from sampling, laboratory standard and the skill of personnel [4].
66 There have also been reports of it varying with pooling level [5]. For the purposes of this study, I
67 fixed the sensitivity at 0.95, irrespective of the level of pooling. This estimate is rather low, which
68 would suggest that I am somewhat overestimating the uncertainty of $\hat{p}$. However, since it is
69 possible that tests will be carried out under suboptimal and non-standardized conditions I prefer
70 to err on the side of caution.
71
72 The formula of Cowling *et al.*, 1999 [6] was used to calculate $\hat{p}$ from a single sample:

$$\hat{p} = 1 - \left(\frac{\eta - \frac{x}{m}}{\theta + \eta - 1} \right)^{\frac{1}{k}} \quad (1)$$

Note that the number of positive pools, x, can be approximated in infinite populations as a stochastic variable subject to a binomial distribution with parameters m and P, where the latter is the probability that a single pool will test positive. A positive pool can arise from two different processes: There can be one or more true positive samples in the pool, and they are detected, or there can be no true positive samples in the pool, but the test gives a false positive result. These two possibilities are represented by the first and second part of the following equation, respectively:

$$P(p,k) \;=\; (1-(1-p)^k)\eta \;+\; (1-p)^k(1-\theta) \tag{2}$$

Closer inspection of the above formula reveals something disheartening: When p approaches zero, P converges towards $1-\theta$. Thus, in low-prevalence scenarios, and for typical values of test sensitivity and specificity, most positive test results will be false positives. Nevertheless, with appropriate levels of sample pooling it is possible to get decent estimates of the true prevalence because the probability of having no positive samples in a pool decreases with k.

Freedom from disease

Pooled sampling can also be used to efficiently assert freedom from disease with a certain probability. If the population is free from the disease, then we find no truly positive sample. The question then becomes how many samples we need to take from a population with prevalence p to ensure that the probability of sampling at least one single positive patient is α or higher. I calculated this number using the formula of Christensen and Gardner, 2000 [7]:

$$n \geq \frac{log(1-(1-\alpha))}{log(\theta(1-p) \;+\; (1-\eta)p)} \tag{3}$$

The formula of Christensen and Gardner can be modified to accommodate pooled sampling:

$$m \geq \frac{log(1-(1-\alpha))}{log((1-p)^k\theta \ + \ (1-(1-p)^k)(1-\eta)} \tag{4}$$

When the prevalence is zero, the formula becomes much simpler:

$$m \geq \frac{\log{(\alpha)})}{log(\theta)} \ , \theta \neq 1 \tag{5}$$

When tests have imperfect specificity $(\theta \neq 1)$, it generally requires a high number of samples to distinguish between a truly disease-free population and one with low prevalence. One way to calculate how many samples are needed to distinguish between the two is to see it as a classical statistical hypothesis testing problem. If we let X be the number of positive samples from a disease-free population, and Y the number of positive samples from a population with low prevalence, we can analyze the probability mass function of the difference between these two binomial variables to see what sample numbers are required to reject the null hypothesis that $X = Y$.

$$X \sim BIN(n, P_1), \qquad P_1 = (1-\theta) \tag{6}$$

$$Y \sim BIN(n, P_2), \qquad P_2 = p\mu + (1-\theta)(1-p) \tag{7}$$

$$Z = X - Y \tag{8}$$

Mathematically, the probability mass function of Z can be calculated as follows:

$$p(z) = \begin{cases} \sum_{i=0}^{n} \binom{n}{i+z} {P_1}^{i+z} (1-P_1)^{n-(i+z)} \binom{n}{i} {P_2}^{i} (1-P_2)^{n-i}, & if\ z \geq 0 \\ \sum_{i=0}^{n} \binom{n}{i} {P_1}^{i} (1-P_1)^{n-i} \binom{n}{i+z} {P_2}^{i+z} (1-P_2)^{n-(i+z)}, & otherwise \end{cases} \tag{9}$$

RESULTS

Estimates of prevalence

In the following, I use simulations to calculate the central 95% estimates of $\hat{p}$ using a test with perfect specificity (Fig. 1). As expected, higher number of total samples are associated with a distribution of $\hat{p}$ more narrowly centered around the true value, while higher levels of pooling are generally associated with higher variance in the $\hat{p}$ estimates. The latter effect is less pronounced in populations with low prevalence. For example, if the true population prevalence is 0.001 and a total of 500 samples are taken from the population (Fig. 1, panel A), the expected distribution of $\hat{p}$ is nearly identical whether samples are run individually (k=1) or whether they are run in pools of 25. Thus, it is possible to economize lab efforts by reducing the required number of pools to be run from 500 to 20 (500 divided by 25) without any significant alteration to the expected distribution of $\hat{p}$. With 5000 total samples, the central estimates of $\hat{p}$ vary little between individual samples (95% interval 0.00021-0.0021) and a pooling level of 200 (95% interval 0.0022-0.0021), making it possible to reduce the number of separate RT-PCR setups by a factor of 200. On the other hand, when the true population prevalence is 0.1, (Fig.1, panel E),

The situation changes when the test specificity (θ) is set to 0.99, that is, allowing for false positive test results (Fig. 2). This could theoretically occur from PCR cross-reactivity between COVID-19 and other viruses, or from human errors in the lab. A problem with imperfect specificity tests are that false positives typically outnumber true positives when the true prevalence is low. This creates a seemingly paradoxical situation in which higher levels of sample pooling often leads to prevalence estimates that are more accurate. This is because

many pools test positive without containing a single true positive sample, leading to inflated estimates of the prevalence. When the level of pooling goes up, the probability that a positive pool contains at least one true positive sample increases, which increases the total precision. The trends about appropriate levels of pooling for different sample numbers and levels of true population prevalence are similar as for the perfect specificity scenario, but with imperfect specificity, we have an added incentive for sample pooling in that prevalence estimates are closer to the true value with higher levels of pooling.

Freedom from disease

Sample pooling can also be of great benefit in order to establish freedom from disease. True freedom in a population is not possible to assert without sampling every individual. However, we can establish how many samples we need in order to have at least $(1-\alpha)\%$ probability of getting a positive sample if the true prevalence was p. For example, from Fig. 3, panel A, we can see that if by sampling 60 patients from a population with a true prevalence of 0.06, we would be 95% certain that at least one of our samples came out positive. That is, if the true prevalence in the source population was exactly 0.06, we would only expect to get 60 negative samples by chance 5% of the time. A common interpretation of this is that if all pools test negative, we can be 95% certain that the true prevalence in the source population is 0.06 or lower.

Note that when the specificity is 1.0, we will never get a positive test from a completely disease-free population no matter how many samples we take. However, if the test specificity is less than 1.0, the sample size needed to ensure $(1-\alpha)\%$ probability of getting at least one positive sample has an upper bound, even when the population is free of disease. This is the sample

size for which we would expect at least a $\alpha\%$ probability of getting a false positive result (Fig. 3, panel B). From formula (5) this number is 299 for an α of 5% and a specificity of 0.99.

Clearly, it becomes very hard to discriminate truly disease-free populations from populations with low prevalence when a test is imperfect. Using formula (9), we find that with a test specificity of 0.99, taking 2743 samples from a disease-free population and as many from a population with a prevalence of 0.005, we will have a 5% probability of getting *more* positive samples from the disease-free population than from the low-prevalence one (Fig. 4). If the true prevalence drops to 0.001, the required number of samples to reliable differentiate from a disease-freeness 95% of the time jumps to nearly 70000. With this is mind, it is hard to imagine a situation in which random testing with an imperfect specificity test could be used to demonstrate true freedom rather than a very low prevalence.

DISCUSSION

The relationship between true prevalence, total sample number and level of pooling is not always intuitive. Note in particular that for certain combinations of parameters, the wavy shape of some curves indicates that the precision can actually *increase* with higher levels of pooling (Figs. 1 and 2). This is particularly true for the lower number of sample counts. This phenomenon is mainly caused by the discrete nature of each estimate of $\hat{p}$. That is, $\hat{p}$ is not continuous and for limited pool sizes miniscule changes in the number of positive pools can affect the estimate quite a bit.

For example, if we take 200 samples and go with a pool size of 100, there are only three potential outcomes: First, both pools are negative, in which case we believe the prevalence is 0. Second, one pool is positive and the other negative, in which case we estimate $\hat{p}$ as

approximately 0.007 if the test sensitivity is 0.95. Finally, both pools are positive, in which case the formula of Cowling *et al.* does not provide an answer because the fraction of positive pools is higher than the test sensitivity. This formula is only intended to be used when the fraction of positive pools is much lower than the test sensitivity.

In general, very high levels of pooling are not appropriate since, depending on the true prevalence, the probability that every single pool has at least one positive sample approaches 1. In low prevalence settings however, it can be appropriate to pool hundreds of samples, but the total number of samples required to get a precise estimate of the prevalence is much higher. Thus, decisions about the level of pooling need to be informed by the prior assumptions about prevalence in the population, and there is a prevalence-dependent sweet spot to be found in the tradeoff between precision and workload.

CONCLUSION

Attempts to estimate the true prevalence of COVID-19 in presumptive low-prevalence or disease-free populations can benefit from sample pooling strategies. Such strategies have the potential to greatly reduce sampling-associated costs with only slight decreases in the precision of prevalence estimates. If the prevalence is low, it is generally appropriate to pool even hundreds of samples, but the total sample count needs to be high in order to get reasonably precise estimates of the true prevalence. On the other hand, if the prevalence is high there is little to be gained by pooling more than 15 samples.

LIST OF ABBREVIATIONS

RT-PCR = Reverse transcriptase polymerase chain reaction

SARS-CoV-2 = Severe acute respiratory syndrome coronavirus 2

DECLARATIONS

Ethics approval and consent to participate – Not applicable

Consent for publication – Not applicable

Availability of data and materials - Code written for this project is available at https://github.com/admiralenola/pooledsampling-covid-simulation. All simulations and plots were created in R version 3.2.3 [8].

Competing interests – Not applicable

Funding – Not applicable

Authors' contributions – All work was done by OB

Acknowledgements – Not applicable

FIGURE LEGENDS

Fig. 1 – Central 95% estimates of $\hat{p}$ with a test with perfect specificity under different combinations of total number of samples and level of sample pooling. Panel A: $p = 0.001$; Panel B: $p = 0.003$; Panel C: $p = 0.01$; Panel D: $p = 0.03$; Panel E: $p = 0.10$

Fig. 2 – Central 95% estimates of $\hat{p}$ with a test with a specificity of 0.99 under different combinations of total number of samples and level of sample pooling. Panel A: $p = 0.001$; Panel B: $p = 0.003$; Panel C: $p = 0.01$; Panel D: $p = 0.03$; Panel E: $p = 0.10$

Fig. 3 – Testing for freedom of disease with a test with perfect specificity. The x-axis represents different true levels of p, and the colored lines represent the number of samples associated with 95% probability of having at least one positive sample at that prevalence level. For perfect specificity tests this is commonly interpreted as meaning that we can be 95% certain that the true prevalence is lower. The effects of sample pooling are explored with different color lines. Panel A: Test specificity = 1.0; Panel B: Test specificity = 0.99.

Fig. 4 – Using a test with specificity of 0.99 to discriminate a disease-free population from a population with $p = 0.005$ with 2743 samples from both populations. Panel A: The expected number of positive samples from the disease-free and the low-prevalence populations; Panel B: The probability mass function of the difference in the number of positive samples between the

low-prevalence and the disease-free population. With 2743 samples from both populations, there is a 5% probability of getting more positive tests from the disease-free population.

REFERENCES

1. 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. doi:10.2807/1560-7917.es.2020.25.10.2000180

2. Q&A: Similarities and differences – COVID-19 and influenza. [cited 17 Apr 2020]. Available: https://www.who.int/news-room/q-a-detail/q-a-similarities-and-differences-covid-19-and-influenza

3. Hogan CA, Sahoo MK, Pinsky BA. Sample Pooling as a Strategy to Detect Community Transmission of SARS-CoV-2. JAMA. 2020. doi:10.1001/jama.2020.5445

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

5. 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. medRxiv. 2020; 2020.03.26.20039438.

6. Comparison of methods for estimation of individual-level prevalence based on pooled samples. Prev Vet Med. 1999;39: 211–225.

7. Christensen J, Gardner IA. Herd-level interpretation of test results for epidemiologic studies of animal diseases. Preventive Veterinary Medicine. 2000. pp. 83–106. doi:10.1016/s0167-5877(00)00118-5

8. R Core Team. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2013.