Pooling of Samples to Increase Testing Capacity for COVID-19

Navid Izady*

Cass Business School, City, University of London, UK

Abstract

Test, trace and isolate are the main pillars of the containment strategies promoted by epidemiologists in the COVID-19 pandemic. Equipment, material and labour required for testing is, however, limited, making it a challenge to adopt testing at a large scale. Pooling of samples has the potential to reduce the number of tests required for screening a population with a low infection prevalence. We provide a detailed analysis of a well-known pooling strategy called two-stage pooling which involves testing pools of a fixed size. We show that, while this approach can potentially reduce the number of tests, evaluating its cost effectiveness and configuring it optimally require existence of a reliable estimate of prevalence in the population. In the absence of such information, we propose inferring a prior distribution of the underlying prevalence using a combination of expert opinion and a limited exploratory testing of the population, and applying it with either a two-stage fixed pooling strategy, or a multi-stage adaptive pooling strategy. We explain how each of these strategies can be applied, propose algorithms for finding their corresponding optimal pool size, and identify the situations under which each of these strategies is preferred.

1. Introduction

The coronavirus disease 2019 (COVID-19) pandemic has put an immense pressure on health systems worldwide. Widespread testing is often proposed as a fundamental measure in the response against the pandemic (WHO Director General 2020). It helps identify infected individuals and subsequently isolate them and their contacts, hence prevent the spread of the disease. It also informs policy planning by identifying infection clusters in a population over time. Equipment, material and labour required for testing is, however, limited, making it a challenge for even high-

*Corresponding author
Email address: navid.izady@city.ac.uk (Navid Izady)

income countries to adopt testing at a large scale. In particular, the polymerase chain reaction, performed in the most common testing method to identify the presence of viral RNA in a sample, appears to be the major bottleneck in ramping up the testing capacity (Shani-Narkiss et al. 2020). It is a time-consuming and costly process, requiring chemical reagents that are reported to be in short supply due to a surge in global demand.

Pooling of samples, dating back to 1943 (Dorfman 1943), has proved beneficial in expanding testing capacity for other infectious diseases, see, e.g., Litvak et al. (1994) and Truscott et al. (2018). In this approach, pools of samples are tested instead of individual samples. A negative outcome for a pool implies a negative outcome for all samples in the pool, while a positive outcome indicates that there is at least one infected sample in the pool. Further tests may then be conducted on positive pools if needed. Since negative pools do not need further testing, this approach has the potential to substantially reduce the number of tests required for screening a population with a low infection prevalence. It is particularly attractive as it does not typically require any additional material, training or equipment.

Yelin et al. (2020) recently conducted an investigation into the application of sample pooling for COVID-19 tests. Their experiments indicate that a single sample with the virus RNA can be detected in pools of up to 32 samples, with a false negative rate of about 10%. Another research conducted in Germany reports 100% sensitivity and specificity in a field study with 50 samples, including 5 positive and 45 negative samples, divided into 10 pools of size 5 (Goethe University 2020). Lohse et al. (2020) report similar results for pools of up to 30 samples but they warn that border line positive samples, typically observed in convalescent patients 14–21 days after symptomatic infection, may escape detection in large pools. Noriega and Samore (2020) argue that specificity would remain at the same level as individual testing with pool testing for COVID-19 (around 99%). There will, however, be a loss of sensitivity as a result of pooling, but their analysis suggests that this would results in only a small increase in the probability of an individual being infected given a negative pool test.

Gollier and Gossner (2020) propose testing pools of asymptomatic individuals so that all members of pools with a negative outcome can be released from a lockdown situation. They derive the optimal pool size maximizing the expected number of people with a negative outcome. A pooling method is also proposed by Narayanan et al. (2020) for classifying a population as low or high risk

using a small number of tests. Our focus here is on using the pooling approach to identify the test status of all samples in a population. The most common strategy for this involves testing pools of a fixed size in the first stage, followed by testing all constituent samples of positive pools in the second stage. The optimal pool size and cost-effectiveness of this two-stage method is investigated numerically in a stream of papers, see, e.g. Dorfman (1943), Truscott et al. (2018) and Shani-Narkiss et al. (2020). Another pooling strategy involves starting with an initial pool of a given size, and iteratively re-pooling any subsequent pools that is found positive at half-size. The sensitivity and specificity of this multi-stage strategy is analyzed in Litvak et al. (1994), and a methodology for selecting its initial pool size is proposed in Shani-Narkiss et al. (2020). A non-adaptive pooling strategy, which requires only one stage of pool testing, is proposed by Täufer (2020). They argue that it reduces the processing time and simplifies the work flow in the lab as compared to two- or multi-stage pooling, but may increase the probability of a false positive outcome.

The main assumption of the studies outlined above is the presence of a reliable estimate for prevalence in the target population. In the absence of such information, it would neither be possible to decide whether pooling is beneficial, nor to choose the size of the pools optimally. Truscott et al. (2018) propose a Bayesian framework for using the prior information available on prevalence to inform decision making on pooling. They conclude that, assuming a prevalence of less than 10%, running an exploratory survey on 10-15 samples would be sufficient to decide whether pooling is beneficial or not.

In this paper, assuming 100% sensitivity and specificity for individual as well as pool testing, we consider two different situations. One is where a reliable estimate of infection prevalence in the population is available, and the other is where such information is not available. For the first situation, we provide a rigourous mathematical analysis the two-stage pooling strategy. This includes an analytical characterization of the optimal pool size minimizing the expected number of tests. We also show that the optimal pool size and the resulting savings in expected number of tests (as compared to individual testing) would decrease in prevalence level and increase in population size. For populations larger than 100, however, both measures would only depend on prevalence as long as it is above 0.01%. We further illustrate the situations wherein errors in estimating prevalence would make pooling non-beneficial, or reduce its benefits.

For the second situation, we propose inferring a probability distribution of prevalence using

a combination of expert opinion and testing a limited number of samples in the population. The resulting distribution can then be applied in the two-stage pooling strategy. We develop a numerical method for evaluating the cost effectiveness of this strategy as well as its optimal pool size given the underlying prevalence distribution. We also illustrate through examples that the optimal pool size as well as the corresponding saving would increase with the variability of prevalence distribution.

We further propose a multi-stage adaptive pooling strategy as an alternative to two-stage pooling for the situation where a reliable estimate of infection is not available. In this strategy, the optimal size of the pool in each stage is decided based on the prior distribution of prevalence combined with the outcomes of the tests conducted in previous stages. We formulate this strategy as a dynamic programming model, and solve it numerically for given prior distributions. Our numerical experiments show that when the variability of prevalence is high, this approach could reduce the expected number of tests as compared to two-stage pooling by as much as 12%.

2. Known Prevalence

Assuming a reliable estimate of prevalence is available for a population of samples, our objective here is to provide a comprehensive analysis of the two-stage pooling strategy. Let $N \geq 2$ be the size of the population to be tested. Suppose the population is divided into k pools of size n, i.e. $k = N/n$, with $n \geq 2$. Following the literature on pooling of samples (e.g., Truscott et al. 2018), each sample in the population is assumed to be infected with a probability $\theta \in (0, 1)$, independent of other samples. So, the number of tests for the two-stage pooling strategy would be $T = k + nN^+$, where N^+ is the number of pools with a positive outcome. Since $N^+ \sim Binomial(k, 1 - (1 - \theta)^n)$, the expected number of tests is obtained as

$$E(N, \theta, n) \triangleq \mathbb{E}[T] = k + n\,k\,(1 - (1 - \theta)^n) = N f(\theta, n), \tag{1}$$

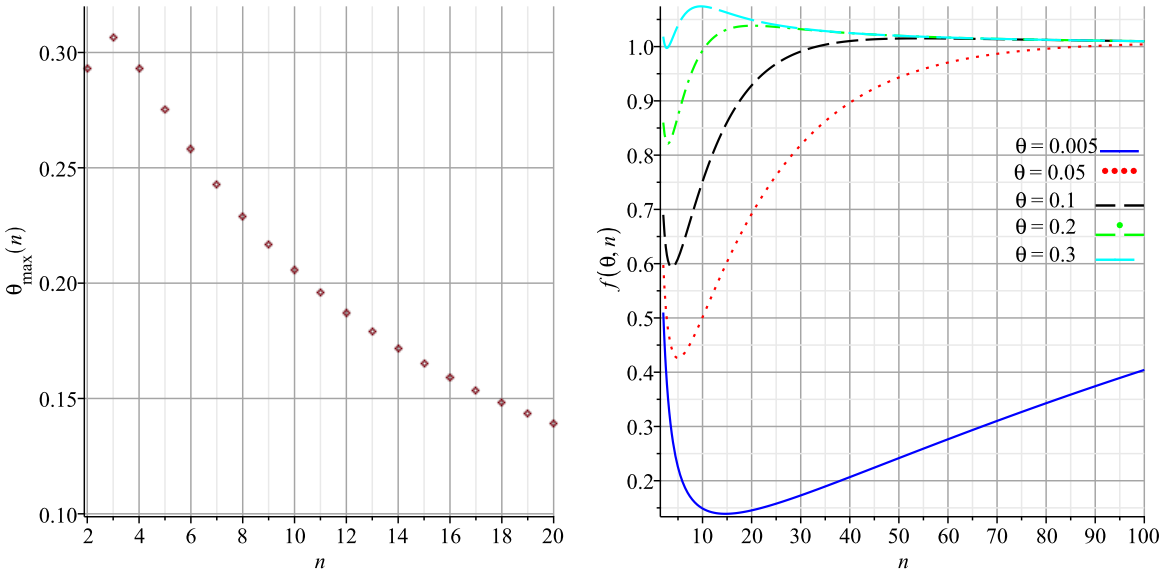
where

$$f(\theta, n) = \frac{1}{n} + 1 - (1 - \theta)^n, \tag{2}$$

for $N = 2, \ldots, \infty$, $n = 2, \ldots, N$ and $\theta \in (0, 1)$. We first investigate when pool testing is beneficial as compared to individual testing. We then characterize the optimal pool size minimizing the expected number of tests. This is followed by an investigation into the impact of error in estimating θ.

2.1. Pool Testing vs. Individual Testing

We assume pooling is not beneficial when $E(N, \theta, n) \geq N$. Solving this inequality for θ, we obtain the largest value of θ for which pooling is beneficial for a given n as $\theta_{max}(n) = 1 - \exp(-ln(n)/n)$. Thus, pooling is never beneficial for $\theta > \alpha_N = \max_{n=2,\ldots,N}\{\theta_{max}(n)\}$ for a given N. Using differentiation, it is easy to verify that $\theta_{max}(n)$ is increasing for $n \in [0, e)$ and decreasing for $n \in (e, \infty)$. Hence, $\alpha_2 = \theta_{max}(2) \simeq 0.2929$, and $\alpha_N = \theta_{max}(3) \simeq 0.3066$ for $N \geq 3$ (see the plot for $\theta_{max}(n)$ in Figure 1, panel (a)).



(a) (b)

Figure 1: $\theta_{max}(n)$ as a function of n (a), and $f(\theta, n)$ as a function of n (b).

2.2. Optimal Pool Size

We assume $\theta \in (0, \alpha_N)$ in this section. It is then trivial that a pool of size 2 will be optimal for $N = 2$, and the corresponding expected number of tests will be $2f(\theta, 2)$. We therefore assume that $N > 2$.

The minimum of $E(N, \theta, n)$ and $f(\theta, n)$ clearly happen at the same point, so we focus on $f(\theta, n)$ for simplicity. To gain some insight into the behaviour of $f(\theta, n)$, we assume it is a continuous function of $n \in [2, \infty)$, and plot it in Figure 1, panel (b), for different values of θ. These plots suggest that for values of θ less than a threshold, $f(\theta, n)$ decreases first then increases to $\lim_{n\to\infty} f(\theta, n) = 1$ asymptotically. For θ larger than the threshold, on the other hand, the function decreases first, then

increases to a maximum before decreasing to one asymptotically. In order to find the minimum of $f(\theta, n)$, we thus need to compare its value at the first point where derivative is equal to zero with one. Differentiating $f(\theta, n)$ with respect to n, we arrive at

$$\frac{d}{dn} f(\theta, n) = -\frac{1}{n^2} - (1-\theta)^n ln(1-\theta).$$

Setting the above equal to zero and solving for n, we get

$$n^*(\theta) = \frac{2L_w(-\sqrt{-ln(1-\theta)}/2)}{ln(1-\theta)}, \tag{3}$$

where $L_w(x)$ is the principal branch of the Lambert W function (Corless et al. 1996). Inserting $n^*(\theta)$ from (3) in Equation (2), we obtain

$$f(\theta, n^*(\theta)) = \frac{ln(1-\theta)}{2L_w(-\sqrt{-ln(1-\theta)}/2)} + 1 - (1-\theta)^{\frac{2L_w(-\sqrt{-ln(1-\theta)}/2)}{ln(1-\theta)}}.$$

Using the properties of the Lambert W function, it can be verified that $f(\theta, n^*(\theta)) < 1$ for $\theta \in (0, 1-\exp(-\exp(-1)) \simeq 0.3078)$. So, $\min_{n\in[2,\infty)}\{f(\theta, n)\}$ occurs at $n = n^*(\theta)$ for $\theta \in (0, \alpha_N)$.

Turning back to the discrete range $n = 2, \ldots, N$, it is now clear that the minimum of $f(\theta, n)$ over this range would occur at $n = N$ if $n^*(\theta) \geq N$, and at either $n = \lfloor n^*(\theta) \rfloor$ or $n = \lfloor n^*(\theta) \rfloor + 1$, depending on which yields a smaller value, when $n^*(\theta) < N$ ($\lfloor x \rfloor$ is the largest integer less than or equal to x.) Solving $n^*(\theta) = N$ for θ yields

$$\theta_s(N) = 1 - \exp(L_w(-1/N)/N), \tag{4}$$

as the threshold for θ, below (above) which $n^*(\theta) > N$ ($n^*(\theta) < N$). This is because the derivative of $n^*(\theta)$ is negative over the interval $(0, 1-\exp(-\exp(-1)) \simeq 0.3078)$, and so it is a decreasing function for $\theta \in (0, \alpha_N)$. The proposition below summarizes our findings in this section.

Proposition 1. *When pooling is beneficial, the minimum expected number of tests is evaluated as*

$$E^*(N, \theta) \triangleq \min_{n=2,\cdots,N} E(N, \theta, n) = \begin{cases} Nf(\theta, N), & \text{if } \theta \leq \theta_s(N), \\ N \min\{f(\theta, \lfloor n^*(\theta) \rfloor), f(\theta, \lfloor n^*(\theta) \rfloor + 1)\}, & \text{otherwise}, \end{cases} \tag{5}$$

for $N \geq 3$, with $f(\theta, n)$, $n^*(\theta)$ and $\theta_s(N)$ given in Equations (2), (3) and (4), respectively. Accordingly, the optimal pool size, denoted by $n^*(N, \theta)$, is obtained as

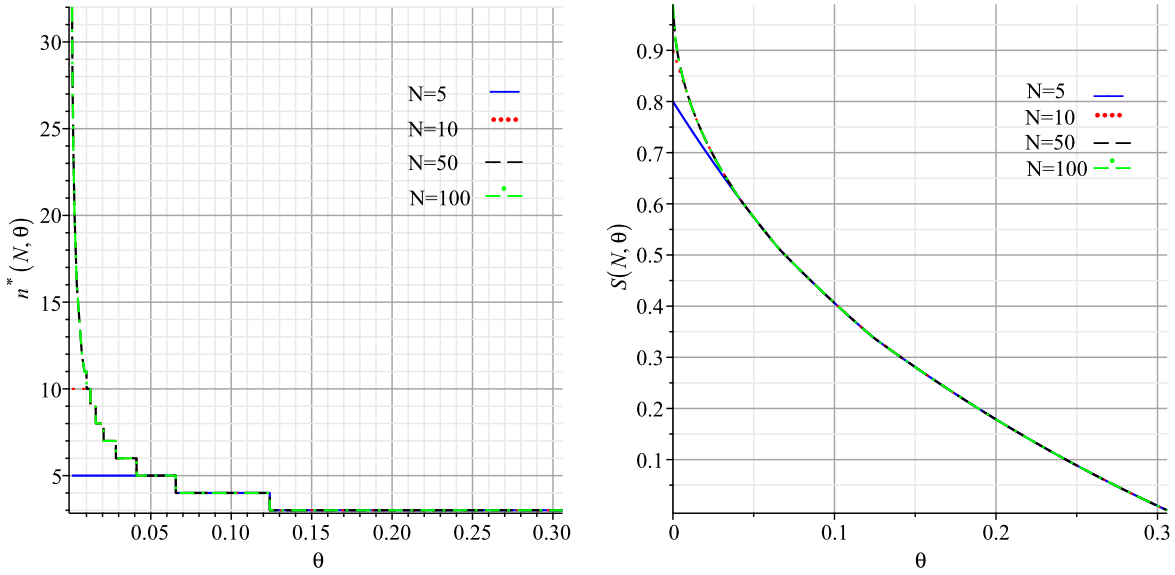
$$n^*(N, \theta) = \begin{cases} N, & \text{if } \theta \leq \theta_s(N), \\ \lfloor n^*(\theta) \rfloor, & \text{if } \theta > \theta_s(N) \text{ and } f(\theta, \lfloor n^*(\theta) \rfloor) < f(\theta, \lfloor n^*(\theta) \rfloor + 1), \\ \lfloor n^*(\theta) \rfloor + 1, & \text{otherwise.} \end{cases} \tag{6}$$

for $N \geq 3$. *Finally, the saving in expected number of tests with optimal pool sampling as compared to individual sampling is evaluated by*

$$S(N,\theta) \triangleq 1 - \frac{E^*(N,\theta)}{N} = \begin{cases} f(\theta, N), & \text{if } \theta \leq \theta_s(N), \\ \min\{f(\theta, \lfloor n^*(\theta) \rfloor), f(\theta, \lfloor n^*(\theta) \rfloor + 1)\}, & \text{otherwise.} \end{cases} \tag{7}$$

for $N \geq 3$.

Equations (6) and (7) imply that $n^*(N,\theta)$ and $S(N,\theta)$ depend only on θ when θ is larger than $\theta_s(N)$. Since $\theta_s(N)$ is decreasing in N, we conclude that for populations of 100 or more samples, both measures would depend only on θ when $\theta > \theta_s(100) \simeq 0.0001$. As an example, in Figure 2 we plot $n^*(N,\theta)$ and $S(N,\theta)$ as functions of θ for $N = 5, 10, 50, 100$. These plots reveal that $n^*(N,\theta)$ and $S(N,\theta)$ decrease with θ and increase with N, but the impact of N disappears when $\theta > \theta_s(5) \simeq 0.051$.



(a) (b)

Figure 2: $n^*(N,\theta)$ as a function of θ (a), and $S(N,\theta)$ as a function of θ (b).

2.3. Prevalence Estimation Error

The assumption of analysis presented in Sections 2.1 and 2.2 is the existence of a reliable estimate of θ. Error in estimating θ may render pooling non-beneficial in two extreme situations. One is when prevalence is estimated to be smaller than α_N, hence pooling is deemed beneficial,

while the true value of prevalence is larger than α_N. Hence, for prevalence estimates close to border value α_N, two-stage pooling should not be applied.

To explain the other situation in which error in estimating prevalence makes pooling non-beneficial, consider the graph in Figure 3, panel (a). The shaded area in this graph illustrates all combinations of θ and n values for which pooling is not beneficial, i.e. $f(\theta, n) \geq 1$. For example, the graph shows that pooling is not beneficial for $n \geq 11$ when $\theta = 0.2$. Any estimation of prevalence, denoted by θ_{est}, that leads to optimal pool size larger than or equal to 11 would therefore make pooling non-beneficial. Since $n^*(\theta)$ is a decreasing function of θ, and solving $n^*(\theta) = 11$ yields $\theta = 0.0099$, pooling will not be beneficial for $\theta_{est} \in (0, 0.0099)$ when the true value of θ is 0.2. The same analysis can be conducted for different values of θ as presented in Figure 3, panel (b). The shaded area in this graph illustrates the set of (θ, θ_{est}) values for which pooling is not beneficial. The graph shows that a substantial under-estimation of prevalence could make pooling non-beneficial. For example, when $\theta = 0.15$, any values of $\theta_{est} \in (0, 0.0034)$, i.e. an under-estimation error of at least 97%, would make pooling non-beneficial.

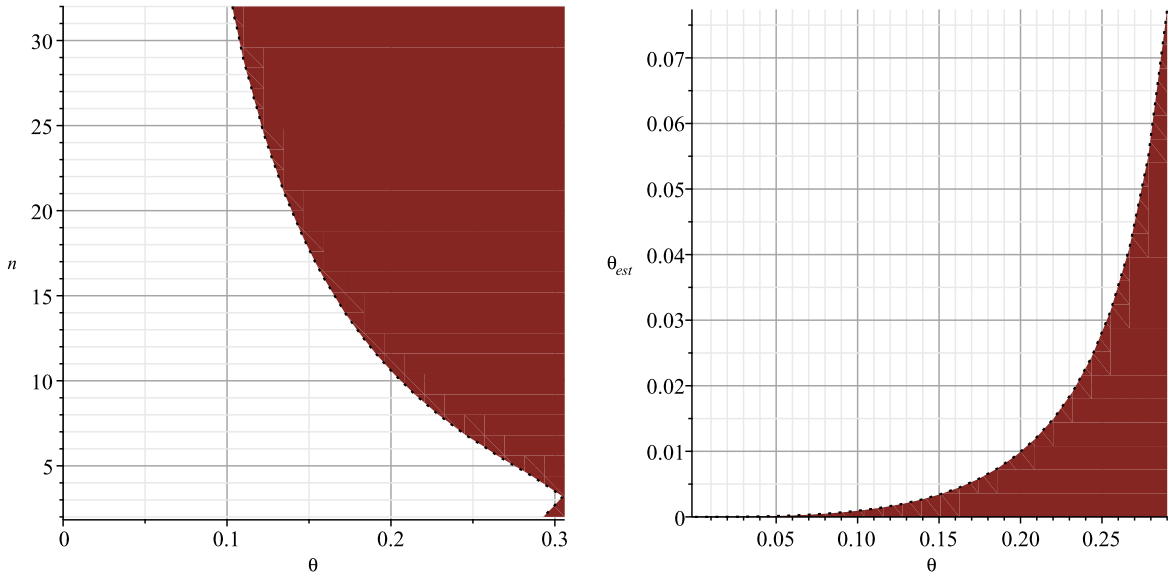


Figure 3: Shaded area illustrates the set of values of (θ, n) (panel a) and (θ, θ_{est}) (panel b) making pooling non-beneficial.

When the two extreme situations described above do not happen, pooling remains beneficial even if there is an error in estimating prevalence. However, the full benefits of pooling may not

be realized. For example, the plots in Figure 4 illustrate the additional number of tests performed for a population of size 100 with $\theta_{est} = 0.005, 0.01, 0.05$, as a function of true value of θ. This is obtained by finding the difference between $E(N, \theta, n)$ obtained from Equation (1) with pool size set to the optimal value with θ_{est}, and the minimum expected number of tests obtained from Equation (5) with true value of θ. This figure suggests, for example, a 30% increase in test numbers when a 1% prevalence is over-estimated at 5%, or when a 3% prevalence is under-estimated at 0.5%.

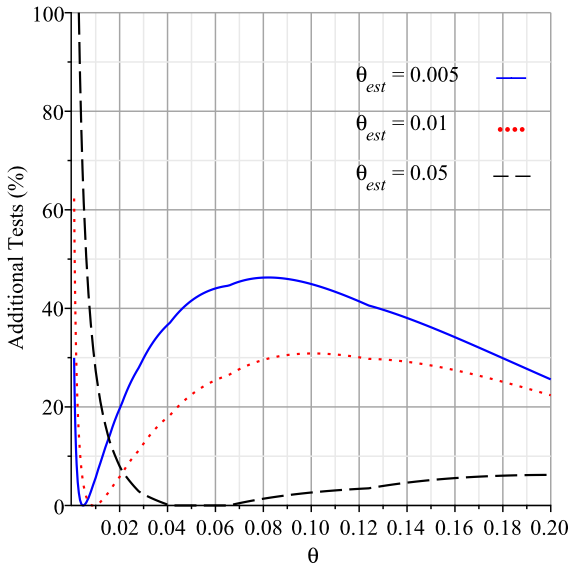


Figure 4: Additional tests due to error in estimating θ as a function of θ .

3. Unknown Prevalence

In this section, we assume a reliable estimate of prevalence, θ, is not available. For this situation, we propose inferring a probability distribution for θ, by using a combination of expert opinion and performing a limited set of exploratory tests on the population. This could be represented as a Uniform (using the maximum and minimum prevalence) or Triangular (using the minimum, most likely and maximum prevalence) or more complicated distributions such as Beta, see Truscott et al. (2018) for more detail. This distribution can be applied in the two-stage pooling strategy as we explain in Section 3.1. Alternatively, a multi-stage adaptive pooling can be applied as we propose in Section 3.2.

3.1. Two-Stage Pooling

Let $\boldsymbol{\theta}$ be the random variable representing prevalence with support $(0, 1)$, and denote its probability density function (pdf) by $\pi(\theta)$. Given a population of size $N \geq 2$ and k pools of size $n \geq 2$, the expected number of tests will be

$$E_\pi(N, n) = N f_\pi(n), \tag{8}$$

where

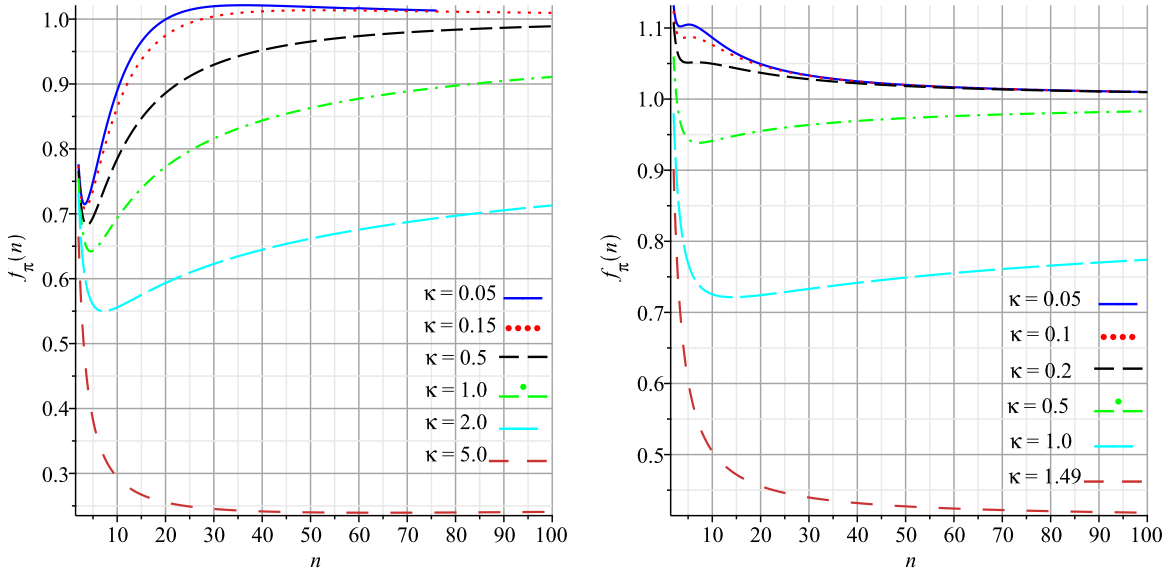
$$f_\pi(n) = \frac{1}{n} + 1 - \mathbb{E}_\pi[(1 - \boldsymbol{\theta})^n]. \tag{9}$$

To gain some insight into the behaviour of $f_\pi(n)$, we assume $\boldsymbol{\theta} \sim Beta(a, b)$, with $a, b > 0$, the shape parameters of the distribution. We set these parameters as

$$a = -\frac{\kappa\mu + \mu - 1}{\kappa}, \quad \text{and} \quad b = \frac{\kappa\mu^2 - \kappa\mu + \mu^2 - 2\mu + 1}{\kappa\mu},$$

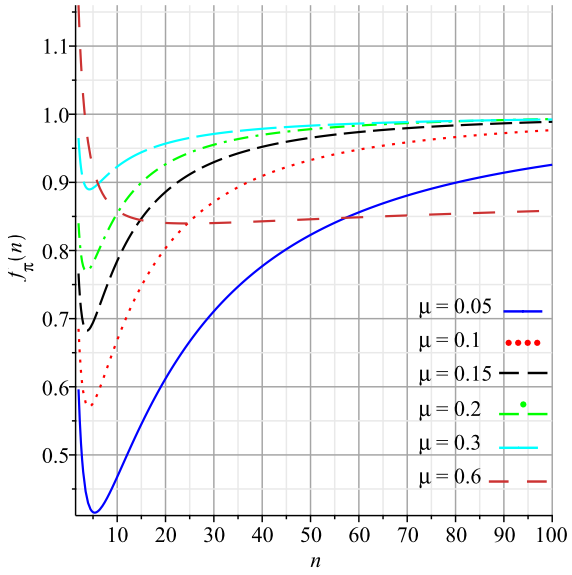
where $\mu \in (0, 1)$ is the mean, and κ is the squared coefficient of variation (SCV) of the distribution with $0 < \kappa < 1/\mu - 1$. In Figure 5, we plot $f_\pi(n)$ for $n \in [2, \infty)$ with different values of μ and κ. The first insight from these plots is that for κ below a threshold, $f_\pi(n)$ first decreases to a minimum, then increases to a maximum before decreasing to its asymptotic value. For κ above that threshold, the function first decreases then increases to its asymptotic value. For example, compare the plots with $\kappa = 0.05, 0.15$ with those with $\kappa = 0.5, 1.0, 2.0, 5.0$ in panel (a) of Figure 5. The second insight is that when it is beneficial to pool samples, i.e. when $f_\pi(n) < 1$ for some values of n, the minimum of $f_\pi(n)$ over the interval $[2, \infty)$ occurs at the first point where derivative is zero. For example, for distributions of $\boldsymbol{\theta}$ represented in panels (a), (c) and (d) of Figure 5 $f_\pi(n)$, the minimum of $f_\pi(n)$ occurs at the point where derivative is equal to zero. For distributions with $\mu = 0.4$ and $\kappa = 0.05, 0.1, 0.2$ represented in panel (b), on the other hand, the minimum does not occur at the point with zero derivative, but pooling is not beneficial for any value of n for these distributions. The last insight is that $f_\pi(n)$ is decreasing in variability of $\boldsymbol{\theta}$ (as measured by its SCV). This implies that even for large values of μ, pooling may be beneficial if variability is sufficiently large. For example, the plots in panel (b) illustrate that pooling is beneficial for $\mu = 0.4$ when $\kappa \geq 0.5$.

The first two insights given above are similar to those we proved analytically for the situation with a known prevalence in Section 2.2. Based on these two insights, in order to find the minimum

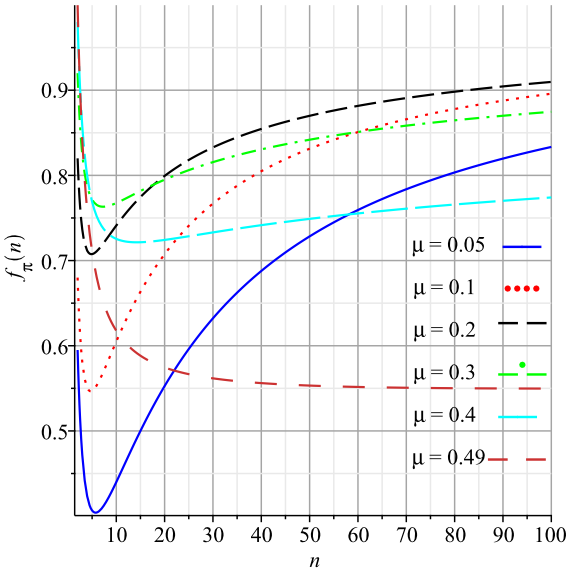


(a)

(b)



(c)



(s)

Figure 5: $f_\pi(n)$ as a function of n for θ following a Beta distribution with $\mu = 0.15$ (a), $\mu = 0.4$ (b), $\kappa = 0.5$ (c), and $\kappa = 1.0$ (d).

of $f_\pi(n)$ over the discrete range $n = 2, \ldots, N$, we first find the point where the derivative of the function is equal to zero. This can be done numerically using, for example, the Newton's method (see, e.g., Abramowitz 1965). We denote this point by n_π. If this point is larger than or equal to N, the minimum occurs at N. Otherwise, we need to compare the value of $f_\pi(n)$ at $\lfloor n_\pi \rfloor$ and $\lfloor n_\pi \rfloor + 1$ to find the optimal n. This is summarized in the proposition below.

Proposition 2. *Suppose $\boldsymbol{\theta}$ is the prevalence random variable with pdf $\pi(\theta)$. The minimum expected number of tests for two-stage pooling is*

$$E_\pi^*(N) \triangleq \min_{n=2,\cdots,N} E_\pi(N,n) = \begin{cases} N f_\pi(N), & \text{if } N \le n_\pi, \\ N \min\{f_\pi(\lfloor n_\pi \rfloor), f_\pi(\lfloor n_\pi \rfloor + 1)\}, & \text{otherwise,} \end{cases} \tag{10}$$

with $f_\pi(n)$ given in Equation (9), and n_π obtained numerically by setting the derivative of $f_\pi(n)$ equal to zero. Pooling will not then be beneficial if $E_\pi^*(N) \ge 1$. Accordingly, the optimal pool size, denoted by $n_\pi^*(N)$, is obtained as

$$n_\pi^*(N) = \begin{cases} N, & \text{if } N \le n_\pi, \\ \lfloor n_\pi \rfloor, & \text{if } N > n_\pi \text{ and } f_\pi(\lfloor n_\pi \rfloor) < f_\pi(\lfloor n_\pi \rfloor + 1), \\ \lfloor n_\pi \rfloor + 1, & \text{otherwise.} \end{cases} \tag{11}$$

We plot $n_\pi^*(N)$ for θ following a Beta distribution with different values of μ and κ in Figure 6, panels (a) and (b). The corresponding savings, evaluated as $S_\pi(N) \triangleq 1 - E_\pi^*(N)/N$, are illustrated in panels (c) and (d) of the same figure. Panels (a) and (c) show that for a fixed value of μ, the optimal n and the corresponding saving increase with variability. They also shows that the optimal n, and thus the corresponding saving, is the same for different values of N up to a certain level of variability. For example, $n_\pi^*(N)$ is the same for $N = 10, 50, 100$ for $\kappa \le 2.63$, and for $N = 50, 100$ for $\kappa \le 4.83$. Panels (c) and (d) show that for a fixed κ, the optimal n and the corresponding saving first decrease then increase with μ. In all cases, the saving is increasing in N but the difference seems negligible for $N \ge 50$.

3.2. Multi-Stage Adaptive Pooling

We now propose a multi-stage adaptive pooling strategy as an alternative to two-stage pooling strategy when a reliable estimate of pooling is not available. Similar to Section 3.1, we assume a

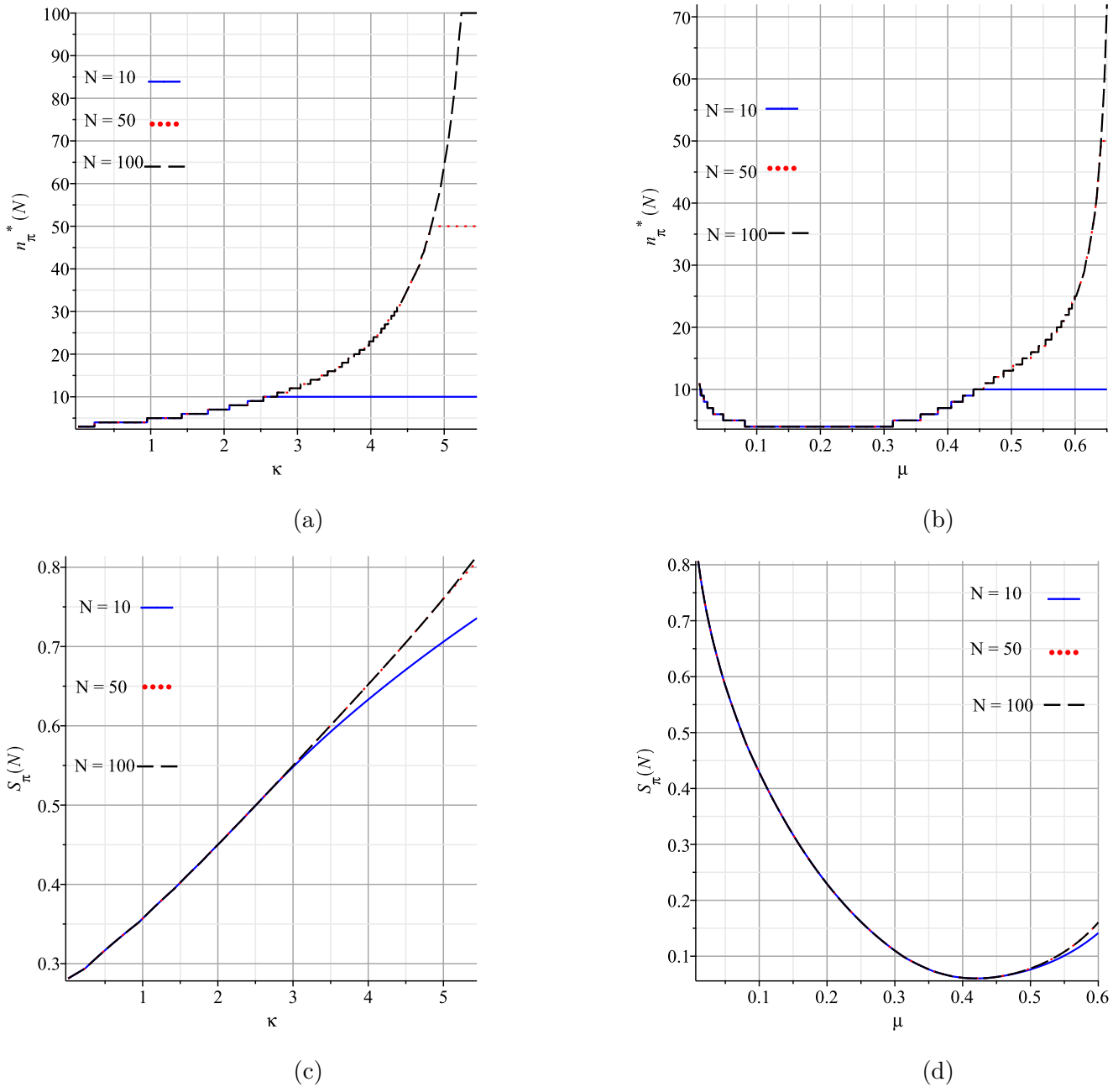


Figure 6: $n_\pi^*(N)$ (a and b) and $S_\pi(N)$ (c and d) with θ following a Beta distribution with $\mu = 0.15$ (a and c), and $\kappa = 0.5$ (b and d) .

prior distribution of prevalence is available. This distribution is updated in each stage of pooling based on the latest test results, informing the optimal pool size for the next stage. Let $\pi(\theta)$ be the prior probability distribution of $\boldsymbol{\theta}$. Given p positive results out of t tests, we use the Bayes' theorem to obtain the following posterior distribution

$$\tilde{\pi}_{(t,p)}(\theta) \triangleq f(\theta|t,p) = C\binom{t}{p}\theta^p(1-\theta)^{t-p}\pi(\theta), \tag{12}$$

where C is a normalizing constant. We formulate the problem as a dynamic programming (DP) model. Let $\Delta_N \triangleq \{(l,p) : l = 0,1,\ldots,N; p = 0,1,\ldots,N-l\}$ be the state space of the DP formulation for a population of size N, where l is the number of samples not tested yet, and p is the number of positive samples so far. For state $\delta \triangleq (l,p) \in \Delta_N$, the set of actions is $\mathcal{A}(\delta) \triangleq \{n : n = 1,\ldots,\min\{l, n_{max}\}\}$, where n is the pool size for the next stage, and n_{max} is the largest feasible size of a pool. Taking a pool of size $n \in \mathcal{A}(\delta)$ when the system is in state $\delta \in \Delta_N$ yields the following expected number of tests for the pool

$$T(\delta,n) = 1 + I_2(n)\left(1 - \mathbb{E}_{\tilde{\pi}_{(N-l,p)}}\left[(1-\boldsymbol{\theta})^n\right]\right), \tag{13}$$

where $I_2(x)$ is equal to x when $x \geq 2$, and 0 otherwise, and expected value is over random variable $\boldsymbol{\theta}$ with its updated pdf $\tilde{\pi}_{(N-l,p)}(\theta)$. $I_2(x)$ is included in Equation (13) to allow for individual sampling as a permissible action in each stage of pooling. The transitions into state $\delta' \triangleq (l-n, p+i)$, with $i = 0,\ldots,n$, representing the number of positive samples in the pool, would then occur with probability

$$\xi(\delta,\delta') = \mathbb{E}_{\tilde{\pi}_{(N-l,p)}}\left[\binom{n}{i}\boldsymbol{\theta}^i(1-\boldsymbol{\theta})^{n-i}\right]. \tag{14}$$

Let $V_N(\delta)$ be the minimum expected number of tests for a population of size N when testing starts in state $\delta \in \Delta_N$. The minimum expected number of tests starting with the entire population is therefore given by $V_N(N,0)$, and can be obtained by recursively solving the optimality equations

$$V_N(\delta) = \begin{cases} 0, & \text{for } \delta = (0,.), \\ \min_{n\in\mathcal{A}(\delta)}\{T(\delta,n) + \sum_{i=0}^{n} V_N(\delta')\xi(\delta,\delta')\}, & \text{otherwise.} \end{cases} \tag{15}$$

The size of the state space for this DP is $O(N^2)$, with each state having $O(\min\{n_{max}, N\})$ actions associated with it. The total number of actions is therefore $O(\min\{n_{max}, N\}N^2)$. Hence, the DP can be solved in linear time in the number of actions. The posterior probability distributions given

in (12) can be evaluated once and saved to memory to speed up computations. We denote the optimal actions obtained by solving Equation (15) for state δ as $n^*(\delta)$.

To see how the model results can be used, consider a population of $N = 10$ samples with $n_{max} = 32$. Assuming that the prior probability distribution of θ is a uniform distribution between 0 and 0.3, we solve the optimality equations in (15) to obtain $n^*(\delta)$. For simplicity, we represent the optimal pool size as the following matrix

$$\boldsymbol{n}^* = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 \\ 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 0 & 0 \\ 3 & 3 & 3 & 3 & 3 & 3 & 3 & 3 & 0 & 0 & 0 \\ 4 & 4 & 4 & 4 & 4 & 4 & 1 & 0 & 0 & 0 & 0 \\ 5 & 2 & 2 & 2 & 2 & 2 & 0 & 0 & 0 & 0 & 0 \\ 3 & 3 & 3 & 3 & 3 & 0 & 0 & 0 & 0 & 0 & 0 \\ 4 & 3 & 3 & 3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 4 & 4 & 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 4 & 3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \tag{16}$$

where the (i, j)th element is $n^*(i-1, j-1)$ for $i, j = 1, \ldots, 11$. Starting with $\delta = (10, 0)$, the matrix above indicates a pool of size 3. If the pool test is negative, we will have $\delta = (7, 0)$ for the next stage, and so the next pool size will be 4. If the pool test is positive, individual tests will be performed on each constituent sample in the pool. Depending on having 1, 2, or 3 positive samples in the pool, we will have $(7, 1)$, $(7, 2)$, or $(7, 3)$, respectively, for the next stage, all of which yields a pool size 3 according to the matrix above. This process continues until $\delta = (0, .)$ is reached. The model gives 6.982 for the average number of tests performed, yielding a saving of 30.18% as compared to individual testing. When the population size increases to $N = 200$, the saving increases to 48.9%.

To compare the multi-stage adaptive pooling with two-stage pooling, suppose the initial distribution of $\boldsymbol{\theta}$ is a Beta distribution with $\mu = 0.15$ and $\kappa = 0.5, 2.5, 5.0$. Table 3.2 shows the expected number of tests with adaptive as well as fixed size pooling for $N = 10, 50, 100, 200$. The results in the table show that the benefit of using adaptive strategy increases with population size and prevalence variability. In particular, we observe 12% reduction in average test numbers with

adaptive pooling as compared to fixed size pooling with $N = 200$ and $\kappa \geq 2.5$.

	$\kappa = 0.5$			$\kappa = 2.5$			$\kappa = 5.0$		
	F	A	Diff	F	A	Diff	F	A	Diff
$N = 10$	6.830	6.878	-1	5.006	5.008	0	2.940	2.940	0
$N = 50$	34.152	33.703	1	25.032	23.065	8	11.999	11.906	1
$N = 100$	68.304	66.966	2	50.065	44.969	10	23.946	22.614	6
$N = 200$	136.609	133.338	2	100.129	88.121	12	47.892	42.01	12

Table 1: Comparison between expected test numbers of Adaptive (A) and Fixed (F) pooling strategies. The Diff column represents 1-A/P in percentage.

4. Discussion

Pooling of samples, if used correctly, has the potential to reduce the number of tests required for identifying the infection status of a population of samples. If a reliable estimate of infection prevalence in the population is available, the two-stage pooling strategy can be applied using the closed-form expression we provided for the optimal size of pools. The amount of expected savings, if any, can also be evaluated in advance to be traded off against the potential implementation challenges.

If a reliable estimate of infection rate is not available, taking a limited random set of samples from the population and testing them individually, combined with expert judgement, would provide a probability distribution of prevalence. If the resulting distribution is not highly variable, a two-stage pooling strategy with a fixed pool size can still be applied. We provided a numerical method for evaluating the optimal pool size for this strategy given prevalence distribution.

If the initial distribution is highly variable, our proposed adaptive pooling strategy would provide additional savings. In this multi-stage strategy, the optimal pool size in each stage is decided based on the most up to date distribution of prevalence, combining the information from initial distribution and the test outcomes in previous stages. We developed a DP model for obtaining the optimal pool size as a function of the number of tests conducted and the number of positive outcomes. With a maximum pool size of 32, the model can be solved for populations of around 200 samples in a reasonably short time. For larger populations, it would be more efficient to first apply

adaptive pooling for a subset of the population for which the optimal pool size can be found in a short time. The updated prevalence distribution obtained at the end of this can then be applied with the two-stage pooling for the rest of the population.

A practical issue with adaptive pooling is that, since the pools need to be tested in a sequential manner, the overall processing time would be longer compared to two-stage pooling. To address this issue, the population of samples can be divided into a number of sub-populations depending on the parallel processing capacity of the lab. These sub-populations can then be tested in parallel using the adaptive approach, hence utilizing the full capacity of the lab and providing the results in a shorter time.

References

Abramowitz, Irene A., Milton abd Stegun, ed. 1965. *Handbook of Mathematical Functions: with Formulas, Graphs, and Mathematical Tables*. Dover Publications.

Corless, R. M., G. H. Gonnet, D. E.G. Hare, D. J. Jeffrey, D. E. Knuth. 1996. On the Lambert W function. *Advances in Computational Mathematics* **5**(1) 329–359.

Dorfman, Robert. 1943. The Detection of Defective Members of Large Populations. *Annals of Mathematical Statistics* **14**(4) 436–440.

Goethe University, Frankfurt. 2020. Corona 'pool testing' increases worldwide capacities many times over. Retrieved May 03, 2020. URL `https://healthcare-in-europe.com/en/news/corona-pool-testing-increases-worldwide-capacities-many-times-over.html`.

Gollier, Christian, Olivier Gossner. 2020. Group testing against Covid-19 1. Tech. rep. URL `http://eprints.lse.ac.uk/104228/`.

Litvak, Eugene, Xin M Tu, Marcello Pagano. 1994. Screening for the Presence of a Disease by Pooling Sera Samples. Tech. Rep. 426.

Lohse, Stefan, Thorsten Pfuhl, Barbara Berkó-Göttel, Jürgen Rissland, Tobias Geißler, Barbara Gärtner, Sören L Becker, Sophie Schneitler, Sigrun Smola. 2020. Pooling of samples for testing for SARS-CoV-2 in asymptomatic people. *The Lancet Infectious Diseases* **3099**(20) 2019–2020.

Narayanan, Krishna, Isabel Frost, Anoosheh Heidarzadeh, Katie K Tseng, Sayantan Banerjee, Jacob John, Ramanan Laxminarayan. 2020. Pooling RT-PCR or NGS samples has the potential to cost-effectively generate estimates of COVID-19 prevalence in resource limited environments. *medRxiv* 2020.04.03.20051995.

Noriega, Rodrigo, Matthew H. Samore. 2020. Increasing testing throughput and case detection with a pooled-sample Bayesian approach in the context of COVID-19. *bioRxiv* 2020.04.03.024216.

Shani-Narkiss, Haran, Omri David Gilday, Nadav Yayon, Itamar Daniel Landau. 2020. Efficient and Practical Sample Pooling High-Throughput PCR Diagnosis of COVID-19. *medRxiv* 2020.04.06.20052159URL `http://medrxiv.org/content/early/2020/04/07/2020.04.06.20052159.abstract`.

Täufer, Matthias. 2020. Rapid, large-scale, and effective detection of COVID-19 via non-adaptive testing. *bioRxiv* 2020.04.06.028431URL `https://www.biorxiv.org/content/10.1101/2020.04.06.028431v1`.

Truscott, James E., Julia C. Dunn, Marina Papaiakovou, Fabian Schaer, Marleen Werkman, D. Timothy J. Littlewood, Judd L. Walson, Roy M. Anderson. 2018. Calculating the prevalence of soil-transmitted helminth infection through pooling of stool samples: Choosing and optimizing the pooling strategy. *PLoS Neglected Tropical Diseases* **13**(3).

WHO Director General. 2020. WHO Director-General's opening remarks at the media briefing on COVID-19 - 16 March 2020. Retrieved April 03, 2020. URL `{https://www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---16-march-2020}`.

Yelin, Idan, Noga Aharony, Einat Shaer Tamar, Amir Argoetti, Esther Messer, Dina Berenbaum, Einat Shafran, Areen Kuzli, Nagham Gandali, Omer Shkedi, Tamar Hashimshony, Yael Mandel-Gutfreund, Michael Halberthal, Yuval Geffen, Moran Szwarcwort-Cohen, Roy Kishony. 2020. Evaluation of COVID-19 RT-qPCR test in multi-sample pools. *Clinical Infectious Diseases* URL `https://doi.org/10.1093/cid/ciaa531`.