Mathematics of Pooled testing

A. V. Sreejith[1] and Somenath Biswas[1]

[1]School of Mathematics and Computer Science, IIT Goa
{sb,sreejithav}@iitgoa.ac.in

Abstract

Within a few months of the first incident, the COVID-19 pandemic is threatening everyone across the globe. Until vaccines and medicines for treatment come to market the best method for curtailing the spread of the disease is to *test, trace and isolate the positive*.

Unfortunately, testing the entire Indian population is difficult given the scarcity of resources. One way to mitigate this difficult scenario is to mix the swab samples of many individuals and test this mixture for the virus (called *pooled testing*). If the test returns positive for virus, we individually check each person to detect who is infected. On the other hand if the test returned negative, then everyone in the pool will be considered negative. This strategy is called *onetime pooling*.

In this abstract we identify when pooled strategy is better and when it is not. We also identify how many testing kits should we buy and store. Thus pooled testing has the potential to develop into a mass testing strategy where large number of people can be tested, identified and positive cases isolated.

1 Introduction and overview

Within a few months of the first incident, the COVID-19 pandemic is threatening almost everyone across the globe to such an extent that the United Nations has declared a public health emergency, calling the situation the world's "most challenging crisis" since the World Wars. Although India has acted early with strong efforts to contain the pandemic, the number of cases has been increasing. In the absence of a vaccine, the only possible containment measure is the implementation of a lock-down. Given the shortage of testing kits, the present ICMR policy is to test the symptomatic patients with minor exceptions like people who are in direct contact with positive cases. On the other hand, it is possible to open the economy and go to a normal life only if we test a large percentage of the population. *Pooled* testing is suggested as a mass testing strategy in this difficult scenario.

An usual test consists of collecting mouth swabs from an individual. This sample is then tested for the presence of the Corona virus through polymerase chain reaction (RT-PCR) testing. The idea of pooled testing is to collect mouth swabs from multiple individuals and running RT-PCR test on this mixture. If the test result is negative we can declare all the individuals to be negative. On the other hand, if the test is positive all individuals in the pool is tested. This pooled strategy is called *onetime pooling*. This is the strategy we look in this writeup. Indian Council of Medical Research (ICMR) has suggested the same testing strategy. There are also other possible strategies. See Bignert et al. (2013); Narkiss et al. (2020); Schisterman and Vexler (2008) for various pooled testing strategies. Pooled testing was introduced by Dorfman (1943) for large scale testing of Syphillis in US army. Pooled testing can increase the speed of testing when the infection in the population is low. It also reduces the number of testing kits used. Thus pooled testing has the potential to develop into a mass testing strategy where large number of people can be tested quickly. Itzkovich (2020) has suggested that pooled testing can be extended to serological tests to detect the presence of antibodies.

We use the notations in Table 1. The pool size and population are denoted by m and N respectively. The rate of infection and non-infection respectively by p and $q = 1 - p$. That is,

$$p = \frac{\text{number of infected}}{\text{population size}}$$

pool size	m
population size	N
rate of infection	p
rate of non-infection	q

Table 1: Notations used

2 Onetime pooling

The onetime pooling consists of two steps of tests. In the first round, a pool of individuals are formed and tested. If the result is negative all individuals in the pool are declared negative. If the result is positive, each person in the pool is individually tested. This is the pooled strategy followed in many states of India, Israel, US for the detection of the covid virus (see Gupta et al. (2020); Yelin et al. (2020)). This is the simplest of pooled testing strategy. It has an additional advantage that even in the worst case scenario, we can get the results in two rounds of testing.

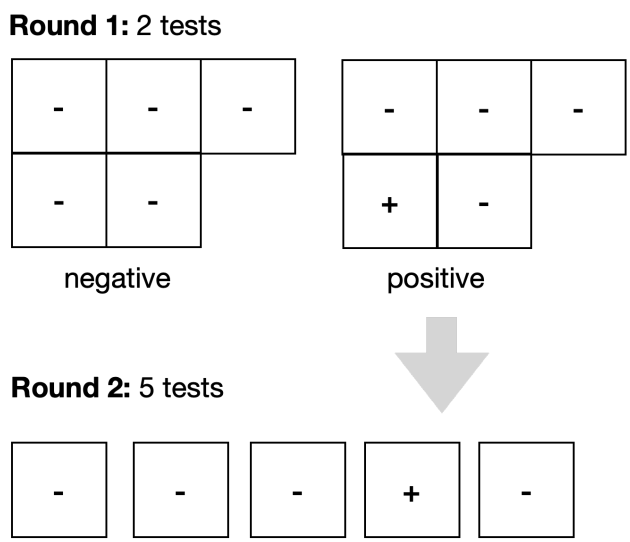

Figure 1: Onetime pooled testing

2.1 Expected number of tests

Our aim is to find out the expected number of tests required in this method. For a pool size of m, the probability that no one has infection is q^m and the probability that at least one person is infected in the pool is $1 - q^m$. Therefore, the expected number of tests is

$$\begin{aligned}\mathbf{Exp}(m, q) &= q^m + (m+1) \times (1 - q^m) \\ &= m(1 - q^m) + 1\end{aligned}$$

It follows that the expected number of tests required for a population size of N is

$$\lceil N(1 - q^m) \rceil + \lceil \frac{N}{m} \rceil$$

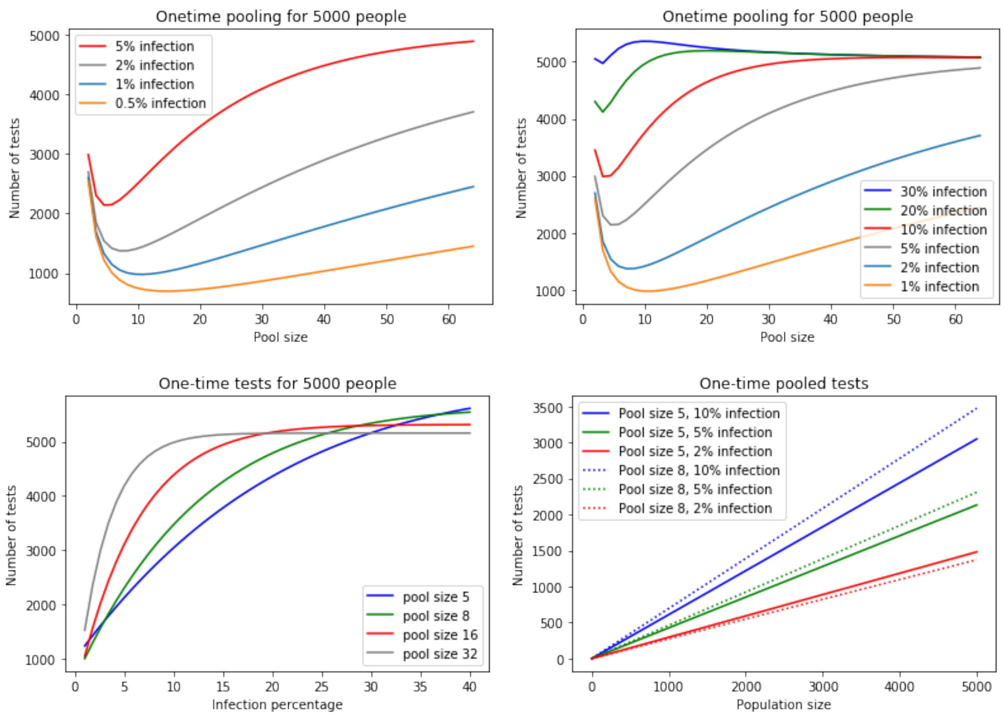

Figure 2: Expected number of tests in onetime pooled strategy.

Figure 2 shows the relationship between pool size and expected number of tests required. The key observation is that pooled testing loses its efficiency if the rate of infection is high.

Figure 2. shows the expected number of tests required for varying population size. We can observe that for a population of 5000 and an infection rate of 2%, we require around 1250 and 1000 tests when pool size is 5 and 8 respectively. This is a 5 fold decrease in the number of testing kits required. If the infection rate is 5% of the population, then the expected number of tests is around 2000, giving us a 2.5 fold decrease in the number of tests. As the infection increases to 10%, efficiency of pooled test decreases. Yet 3500 tests as shown in the figure is better than 5000 individual tests. The expected number of tests required for various pool sizes and infection rate is also given in Table 2 and Table 3. The advantages of pool sizes decreases with higher infection rate. Traditional individual test is better once the infection rate becomes $20\% - 22\%$ and bigger.

The final comparison is between the pool size and expected number of tests. The figure shows that number of tests decreases as the pool size increases, reaches a minimum and then start increasing. For a small rate of infection (say less than 10%), the optimal pool size is less than 10. Thus the pool size, suggested by ICMR of 5 seems to be near to the optimal pool size. A pool size of 8 is also good (in fact better in very low infection of 2%). We also observe that lower the infection rate, the bigger the pool size should be. A higher pool size of 32 and 64 is good if the infection rate is very low (less than 1%).

2.2 Number of testing kits required in the worst case

The expected number of testing need not happen all the time. Therefore, we cannot use this to determine the number of testing kits required. In this subsection, we identify the number of tests required in the worst case. Let us assume we do pooled testing on a population of size N and pool size m. Let p be the infection rate. Then $\lceil pN \rceil$ many individuals are infected. In a worst case partitioning of the pool, we will have all the $\lceil pN \rceil$ individuals to be in different pools. Therefore $\lceil \frac{N}{m} \rceil - \lceil pN \rceil$ many pools are not infected. Therefore

Population size (N)	Pool size (m) and infection rate (p)					
	m=5			m=8		
	p = 1%	2%	5%	p=1%	2%	5%
N = 25	7	8	11	6	7	12
100	25	30	43	21	28	47
1000	250	297	427	203	275	462
5000	1246	1481	2132	1012	1372	2308
10000	2491	2961	4263	2023	2743	4616

Table 2: Expected number of tests in pooled testing

Population size (N)	Pool size (m=5) and Infection rate (p)									
	p=8%	10%	12%	15%	18%	20%	22%	25%	27%	30%
N=25	14	16	17	19	21	22	23	25	25	26
100	55	61	68	76	83	88	92	97	100	104
1000	541	610	673	757	830	873	912	963	993	1032
5000	2705	3048	3362	3782	4147	4362	4557	4814	4964	5160
10000	5410	6096	6723	7563	8293	8724	9113	9627	9927	10320

Table 3: Expected number of tests in pooled testing, with pool of size 5

the maximum number of tests required is

$$\lceil\frac{N}{m}\rceil + \lceil pN\rceil * m$$

From Figure 3, Table 4 and Table 5 we observe that pooled testing is better than traditional testing when the infection is low. For low infection of 2% to 5% the maximum number of tests required is significantly low. It is comparable to the expected number of tests. The Table shows that for a pool size of 5 and infection rate of 2%, we need at most 300 tests kits to test 1000 and 3000 test kits for testing 10000. The rates are slightly lesser when we have a pool size of 8. The Figure shows that as that pooled testing loses its advantages once the rate of infection reaches 10%. It also shows that lower the infection rate, a bigger pool size is better. In short the worst case and expected case are almost similar for low infection rate.

When the infection rate becomes greater than 18% we observe that the number of tests required in the worst case is not better than the traditional individual test. Note that, the expected number of tests in pooled testing is better than individual testing even at an infection rate of 22%. The worst case situation occurs when each pool has exactly one infected individual. In the next section we will observe that this is a very low probability event.

2.3 Removing the low probability events

The worst case scenario happens when positive individuals are send to different pools. How often can this happen? It turns out this is a very low probability event when population size is large.

Let us assume S number of tests were required on a population of size N and pool of size m. Let p be the infection rate. How often do we have to do S tests? Let us denote by E the expected number of tests.

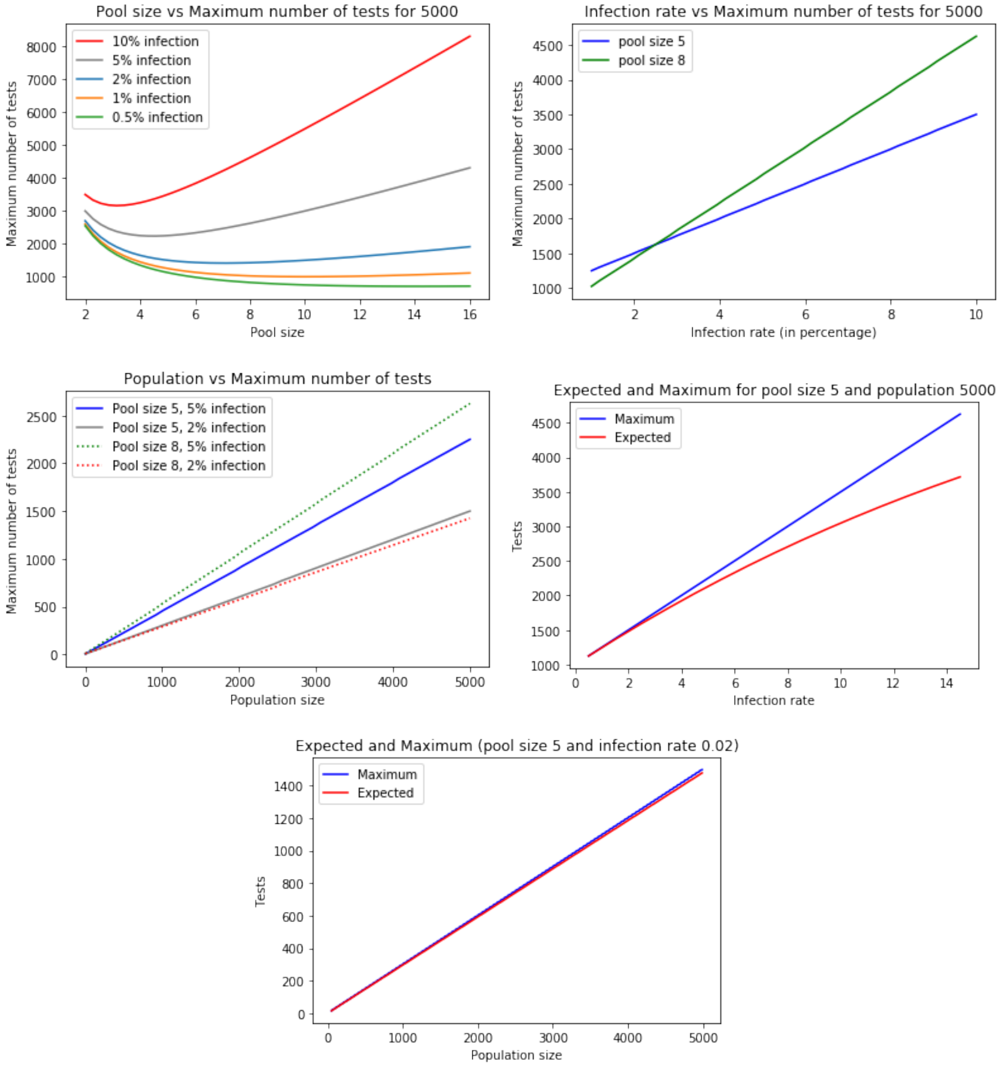

Figure 3: Maximum number of tests required.

Population size (N)	Pool size (m) and infection rate (p)					
	m=5			m=8		
	p = 1%	2%	5%	p=1%	2%	5%
N = 25	10	10	15	12	12	20
100	25	30	45	21	29	53
1000	250	300	450	205	285	525
5000	1250	1500	2250	1025	1425	2625
10000	2500	3000	4500	2050	2850	5250

Table 4: Maximum number of tests.

Let us partition the N samples randomly into $M = \frac{N}{m}$ pools. Let $X_1, X_2, \ldots, X_M$ be random variables,

$$X_i = \begin{cases} 1 \text{ if } i^{th} \text{ pool tests positive} \\ 0 \text{ otherwise} \end{cases}$$

Population size (N)	Pool size (m=5) and Infection rate (p)					
	p=8%	10%	12%	15%	18%	20%
N=25	15	20	20	25	30	30
100	60	70	80	95	110	120
1000	600	700	800	950	1100	1200
5000	3000	3500	4000	4750	5500	6000
10000	6000	7000	8000	9500	11000	12000

Table 5: Maximum number of tests, pool size $m = 5$

Population size (N)	Pool size (m=5) and Infection rate (p)							
	8%	10%	12%	15%	18%	20%	22%	23%
1000	694	777	852	952	1037	1087	1132	1153
5000	3047	3423	3765	4219	4612	4843	5052	5148
10000	5894	6626	7293	8182	8951	9404	9813	10003

Table 6: Number of tests required with confidence of 99%

The expectation $\mathbf{Exp}(X_i) = (1 - q^m)$. The expected number of positive pools, $X = \sum_i X_i$ and $\mu = \mathbf{Exp}(X) = M\mathbf{Exp}(X_i)$. By Chernoff bound

$$\mathbf{Prob}\ \{X \leq (1+\delta)\mu\} \geq 1 - \big(\frac{1}{e}\big)^{\frac{\delta^2\mu}{3}}$$

Since we had to conduct S tests when there were X positive pools, we get $S = m * X + \frac{N}{m}$. Multiplying by m and adding by $\frac{N}{m}$ on the equation given above, we get

$$\mathbf{Prob}\ \{S \leq m(1+\delta)\mu + \frac{N}{m}\} \geq 1 - \big(\frac{1}{e}\big)^{\frac{\delta^2\mu}{3}} = 1 - \big(\frac{1}{e}\big)^{\frac{N\delta^2(1-q^m)}{3m}}$$

We want S to happen with probability 99%. We get δ as

$$\delta = \sqrt{-\frac{3m\ln(1-0.99)}{N(1-q^m)}} = \sqrt{\frac{3m\ln 100}{N(1-q^m)}}$$

We can therefore say that for 99% of times, the number of tests will be less than

$$m\mu\Big(1 + \sqrt{\frac{3m\ln 100}{N(1-q^m)}}\Big) + \frac{N}{m}$$

Table 6 finds this stock limit for a pool size of 5. It shows that for a population size of 10000, we can use one-time testing even for a 23% infection rate. For a population size of 1000, pooled testing is advantageous even at 18% infection.

Note that some of the numbers are worse than the worst case scenario. Obviously, we do not require to do 952 tests when infection is 15% and population size is 1000 when the maximum number of tests is only 950. We get this number 952 because we calculate an upper bound and not an exact bound. One should read it as saying, 952 tests gives a confidence of greater than 99% (in fact it gives 100% confidence). Thus, this worse than worse case scenario numbers do not violate our claim.

2.4 Pool family together or not

In this subsection we answer the following question: Should we pool family members together or not? Let the probability of a person being positive for Corona be p. We say that x and y are in close contact if they come into contact often (either they live nearby or they are in the same family etc). In this case, we have a tendency for x to infect y and vice versa. Therefore,

$$\begin{aligned}\mathbf{Prob}\,\{y \text{ is positive} \mid\ x \text{ is positive}\} &> \mathbf{Prob}\,\{y \text{ is positive}\} = p\\ \mathbf{Prob}\,\{y \text{ is negative} \mid\ x \text{ is negative}\} &> \mathbf{Prob}\,\{y \text{ is negative}\} = q\end{aligned}$$

The updated pooled testing strategy is to pool individuals who are in close contact. Our aim is to find out the expected number of tests required in this method. For a pool size of m, the probability that no one has infection is $q^m + \epsilon$ where $\epsilon \in [0, 1]$ and the probability that at least one person is infected in the pool is $1 - q^m - \epsilon$. Therefore, the expected number of tests in this "pooling family together" method is equal to

$$\begin{aligned}&q^m + \epsilon + (m+1) \times (1 - q^m - \epsilon)\\ &= m(1 - q^m - \epsilon) + 1\\ &\leq m(1 - q^m) + 1 = \mathbf{Exp}(m, q)\end{aligned}$$

In other words, the expected number of tests required is less than random selection of pools. We have the following takeaway.

Pooling family members and those living close together is a better strategy than taking a random pool.

2.5 Criticism of pooled testing

The pooled testing brings in a layer of book keeping on testing. One needs to keep track of the individuals and the pools they are assigned to. It also requires that the swab samples of individuals are stored. This is required since if a pool turns out to be positive, the stored swab samples will have to be retrieved and tested. This brings in a cost overhead. Similarly, the sample collectors need to collect multiple samples from individuals.

The one-time pooled testing is time consuming from an individuals perspective. A positive individual will be identified only after two rounds of test. The first test detects the individuals pool as positive followed by the second test identifying the individual in the pool.

3 Multi-pooling

The multi-pooling strategy consists of multiple pooling tests. At first level we use a pool size of m. In the next round all positive pools are split into two. That is, we consider pools of size $\frac{m}{2}$. In the third round, we split the positive pools from the second round into two (size of $\frac{m}{4}$). The process is continued until the pool size becomes one or all individuals are tested separately. This is the best strategy to reduce the number of tests. Unfortunately, it takes multiple pooling tests taking too much time for a positive individual to know the result of the test. Moreover, multi-pooling strategy works only if the rate of infection is low (around 1% or less).

3.1 Expected number of tests

In this subsection we find the expected number of tests required. Let N_1 be the initial population size, q_1 the rate of non-infection and m be the starting pool size. The total number of infected is $q_1 N_1$ and the total

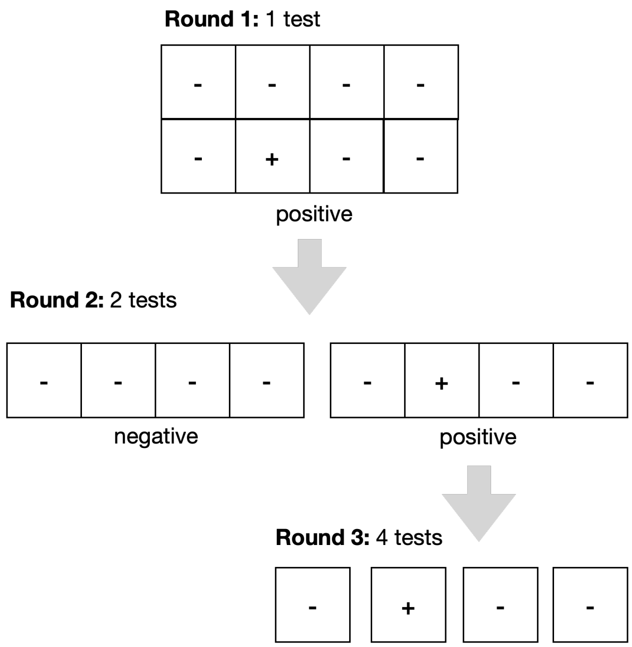

Figure 4: Multi-pooling testing

number of tests in the first round (denoted by $\mathbf{ET}_1$) is

$$\mathbf{ET}_1 = \lceil \frac{N_1}{m} \rceil$$

The probability that a pool is infected in the first round is $1 - q_1^m$. Hence the expected number of positive pools in the first round is is

$$\mathbf{PositivePools}_1 = \frac{N_1}{m}(1 - q_1^m)$$

and the population to be tested in the second round is $N_2 = N_1(1 - q_1^m)$. Let us now calculate the number of tests in the second round. The probability of non-infection in the second round is $q_2 = \frac{q_1 N_1}{N_2}$. We now partition the N_2 individuals into pools of size $\frac{m}{2}$ and tests them. This gives us, the total number of tests in this round to be $\lceil \frac{2N_2}{m} \rceil$. But we can go one step further. Look at a pool from the first round which contains exactly one positive individual. The second round splits the pool into two. Testing the non-positive split will immediately tell us that the other split is positive an information we gauged without testing that pool. Note that the non-positive split will be tested first with probability half. The expected number of pools with exactly one infection is $\mathbf{SingletonPools}_1 = N_1(1 - q_1)q_1^{m-1}$. The expected number of tests in the second round is

$$\mathbf{ET}_2 = 2 \times (\mathbf{PositivePools}_1 - \mathbf{SingletonPools}_1) + \frac{3}{2} \times \mathbf{SingletonPools}_1$$

We progress by considering further reduction in the pool size. Thus the expected number of tests in this multi-pooling strategy comes out to be

$$\sum_k \mathbf{ET}_k$$

for k rounds of pooling ($k = \log m$). In the next section we use this multi-pooling strategy to test all individuals of Goa. See Table 7.

The above analysis can be approximated by assuming that the a positive pool in the first round will contain only one infected individual. If the infection is low, this is almost surely going to happen. The expected number of tests required if a pool of size m contains one infected individual is approximately

$$\frac{3}{2}\log m + 1$$

Assuming a population of size N and a probability of infection of p we get the expected number of tests required to be

$$\mathbf{Exp}(m,p) \sim \frac{3pN}{2}\log m + \frac{N}{m}$$

For a population of 16 lakh and an infection ratio of 1% we get that around 1 lakh seventy thousand tests are required with an initial pool size m of 32. The table 7 shows that we can test the entire population with a little more than 1 lakh sixty thousand.

4 Goa case study

The objective is to test all citizens of Goa. The population of Goa is around 16 lakh.

Our assumption is that the rate of infection of Goa to be less than 1%, or $16,000$ individuals are positive. This is a very low rate of infection. This suggests we can use a higher pool size. Experimental data shows that RT-PCR can be used even for a pool size of 32 (see Itzkovich (2020); Mallapaty (2020)).

The next modification is to use multiple levels of pooling. At first level we will have the a pool size of 32. In the next round all individuals in the positive pools will be split into pools of size 16, that is half of the previous pool size. In the third round, we split the positive pools from the second round into pools of size 8 and in the final round we use pools of size 4. Finally all individuals in the positive pools are tested. Table 7 gives the total number of tests required to test everyone in Goa. The table shows that around one lakh sixty thousand tests required. The required number of tests is $\frac{1}{10}^{th}$ the population size.

	N	pN	m	p	pools	+ pools	1-pools	Tests	1-pool	total
Round 1	16,00,000	16,000	32	0.01	50,000	13,751	11,717	50,000	0	50,000
Round 2	65,088	4,283	16	0.07	4,068	2,700	1,543	4,068	87,878	91,946
Round 3	18,512	2,740	8	0.15	2,314	1,672	893	2,314	9,258	11,572
Round 4	6,232	1,847	4	0.30	1,558	1,177	644	1,558	4,019	5,577
Individual test	4,708	–	1	–	–	–	–			4,708
Total tests										1,63,803

Table 7: Testing the entire population of Goa: expected number of tests required

5 Conclusion

We conclude by noting down some of the observation.

1. Pooled testing is better than traditional testing when the rate of infection is less than 20%
2. Pooled testing can be done even for an infection rate of $20-22\%$ (with pool size 5 or lesser).
3. Pooled testing becomes significantly better when the rate of infection is very low (less than 10%).
4. A pool size of $5-8$ seems to be optimum.

5. At very low infection (less than 2%) higher pool sizes (like 16) is better than pool size 5.
6. The expected number of pool tests and the maximum number of pool tests are 'almost' similar in low rate of infection. In other words, the number of test kits to stock is close to the expected number.
7. Pooling family members and close neighbourhood together is a better strategy than random pools.
8. Multi-pooling can further reduce the number of tests.

References

Bignert, A., Eriksson, U., Nyberg, E., Miller, A., and Danielsson, S. (2013). Consequences of using pooled versus individual samples for designing environmental monitoring sampling strategies. *Chemosphere*, 94:177–182.

Dorfman, R. (1943). The detection of defective members of large populations. *The Annals of Mathematical Statistics*, 14:436–440.

Gupta, N., Praharaj, I., Bhatnagar, T., Vivian Thangaraj, J., Giri, S., Chauhan, H., Kulkarni, S., Murhekar, M., Singh, S., Gangakhedkar, R., Bhargava, B., and ICMR COVID Team, . (2020). Severe acute respiratory illness surveillance for coronavirus disease 2019, India, 2020. *Indian Journal of Medical Research*, 151(2):236–240.

Itzkovich, J. S. (2020). https://www.tabletmag.com/sections/science/articles/pooling-covid-19-israel.

Mallapaty, S. (2020). https://www.nature.com/articles/d41586-020-02053-6.

Narkiss, H. S., Gilday, O. D., Yayon, N., and Landau, I. D. (2020). Efficient and practical sample pooling for high-throughput PCR diagnosis of covid-19. *medRxiv*.

Schisterman, E. F. and Vexler, A. (2008). To pool or not to pool, from whether to when: applications of pooling to biospecimens subject to a limit of detection. *Paediatr Perinat Epidemiol*, 22:486–496.

Yelin, I., Aharony, N., Shaer-Tamar, E., Argoetti, A., Messer, E., Berenbaum, D., Shafran, E., Kuzli, A., Gandali, N., Hashimshony, T., Mandel-Gutfreund, Y., Halberthal, M., Geffen, Y., Szwarcwort-Cohen, M., and Kishony, R. (2020). Evaluation of covid-19 RT-qPCR test in multi-sample pools. *medRxiv*.