Optimizing Pool Size for Pooled Testing of SARS-CoV-2

Jerry Li

Personal Section:

One evening in the summer of 2020, well after the severity and endurance of the COVID-19 pandemic had become evident, I was having a chat with my father at the dinner table. Both STEM people, our talks often leaned towards the topic of science, especially in the realm of current events. This time, it was the matter of COVID testing that made its appearance. Testing, so essential to managing an outbreak, yet so scarce when it was needed. That night, I learned about a mostly unemployed method in the pandemic called "pooled testing," where, rather than testing polymerase chain reaction (PCR) samples individually, multiple are combined and tested together. In theory, if a pooled sample were to test negative, it would indicate that all individual samples of that pool are also negative, meaning many tests can be saved. So why not squeeze as many samples as possible (without jeopardizing accuracy) into each pool? The issue here is that with so many samples per pool, more pools are likely to test positive, and *all* individuals of positive pools must be retested to identify those with the disease. This posed an interesting problem. If some balance between too many and too few individuals per pool could be found, then the amount of tests and thus resources/money saved could be greatly boosted.

In the fall, junior year began, and I started taking calculus in school. Combined with the outside-of-the-box and analytical thinking skills I had developed from learning competition math, I realized that my pooled testing puzzle was one that I could realistically examine. Deriving formulas to represent pooled testing would take some creativity and understanding all of its nuances would take time, but it felt doable, and more importantly, intriguing. As I began to play with equations and research the mechanisms of pooled testing, I found deeper aspects that I incorporated into my study, namely the issue of testing accuracy.

Over the course of my project, which spanned from late fall of 2020 to summer of 2021 (with large pauses in between), I received advice from a mentor on the process of reviewing literature, producing novel contributions in the field, and writing a paper. However, my research and derivations were performed almost exclusively from the desk in my room. But even without an extravagant lab or experiments to run, this intersection of science and mathematics and public health was enough to fully entertain me.

To any high schoolers interested in research, particularly that which combines math and science, I would say explore the questions that actually interest you, especially the ones that you find lingering in your mind time after time. And once your curiosity is truly hooked by an idea, stay motivated in pursuing it. Whether this leads to a research project or new skills, it will be well worth your time.

Research Section:

Abstract:

The spread of COVID-19, kindled by a lack of mass testing in early stages, has affected hundreds of millions of lives. Even with recent vaccination developments, such testing is still critical. Thus, a timely, cost-efficient method for extensive testing is imperative for fighting the pandemic. One solution is pooling multiple samples into a single PCR test. This study aims to determine the optimal size of these pools based on prevalence rate and testing accuracy.

The R programming language was used to simulate pooling in a population with a prevalence rate of 0.05. This revealed a binomial distribution, where each positive case in a pool represents a "success." The following formula (1) was derived: $F = (1 - p)^n - \frac{1}{n}$, where F is the reduction factor, n is the pool size, and p is the prevalence. Accounting for testing sensitivity and specificity, the modified reduction factor formula (2) becomes:

$$F_m = 1 - \frac{1}{n} - S_n(1 - (1 - p)^n) - (1 - S_p)(1 - p)^n,$$

where F_m is the modified reduction factor, S_n is the testing sensitivity, and S_p is the testing specificity. Plotting equation (2) revealed a single relative maximum in the first quadrant. Taking the derivative and setting it equal to zero, the optimal pool size formula (3) was found:

$$0 = n^2(1 - p)^n \ln(1 - p)\,(S_n + S_p - 1) + 1$$

When S_n and S_p are held constant, as prevalence rate increases, pooling efficiency decreases. For instance, when testing accuracy is perfect, $p = 0.005$ corresponds to an optimal pool size of 15 and an 86% reduction in tests needed. At $p = 0.2$, however, the optimal pool size of 5 can only save 13%, and at $p = 0.3$, the reduction is negligible.

These findings allow for public health officials to perfect their design of a testing plan. Finally, increasing specificity and decreasing sensitivity both result in increasing maximum reduction. However, decreasing sensitivity leads to a higher risk of disease spread. Sufficient testing sensitivity should not be sacrificed for higher pooled testing efficiency.

1. Introduction

Numerous clustered cases of pneumonia were reported from Wuhan, China, in December of 2019 (Centers for Disease Control and Prevention). In January, 2020, the cause of this was determined to be a novel coronavirus, and was named severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) (Centers for Disease Control and Prevention).

The polymerase chain reaction (PCR) test, which amplifies a DNA sample by making countless copies, has been the principal method for testing for SARS-CoV-2. On February 5th of 2020, the U.S. Centers for Disease Control and Prevention began to send out PCR testing kits to state testing labs (Patel). However, many of these kits had faulty negative controls that erroneously caused false positives, so samples often had to be shipped to the CDC for testing (Patel). For many crucial weeks, tracking of the initial COVID-19 outbreak was hampered.

This lack of efficient mass testing of SARS-CoV-2 was thus a large reason for the development of the COVID-19 pandemic. As the pandemic progressed, testing capacity has remained an issue. To properly monitor not only individuals but communities and larger trends as well, mass testing is essential. Even with vaccinations recently becoming widespread, testing is still as necessary as ever. However, testing can be costly, both in terms of time and resources.

One solution to this is pooled testing, where groups of individuals are pooled into a single PCR test. Studies have shown that a positive sample can still be detected in a pool of 32 samples (dan Yelin et. al, 2020). After the initial pooled stage, positive pools would still require testing of individuals' samples to identify specific cases. Pools that test negative, however, have no potential cases to be detected, thus vast amounts of further tests can be saved. Because of the large fluctuations in pooled testing efficiency, it is necessary to use the most ideal pool size. Thus, the primary aim of this study is to derive an optimization equation. In addition, testing accuracy may also affect the efficacy of pooled testing. This study will account for testing sensitivity and specificity in the optimization to produce a more accurate equation.

2. Methods and results

2.1. R coding procedure

In order to better understand the mathematics of pooled testing, a simulation was first coded with R. Some hypothetical parameters were used in this process; prevalence rate (p), pool size (n), and population size (N) were initially set equal to 0.05, 5, and 1,000 respectively. A vector of size 1000 was filled (1 to 1000) to represent the population and a vector of size 200 was created to represent the number of pools (or $\frac{population\ (N)}{pool\ size\ (n)}$), where each five number interval from 1 to 1000 represents a pool (i.e., 1–5, 6–10, …, 996–1000). Then, fifty numbers from 1 to 1000 were randomly selected to represent positive cases, and the number of positive cases in each interval was tallied. Once the code was functional, this simulation was repeated 2,000 times, effectively resulting in 400,000 pools and a total population of 2,000,000. The distribution was graphed and appeared to be binomial (Section 3.1, figure 5).

2.2. Reduction factor formula

Reduction factor (F) is defined as $F = \frac{t_o - t_p}{t_o}$, where t_o is the number of tests needed without using pools (standard testing) and t_p is the number of tests needed with the usage of pools. Because standard testing requires one test per person, t_o is equal to N. Pooled testing on the other hand, requires two rounds of testing: one test per pool initially, followed by individual testing of the members of positive pools.

This first round requires $\frac{N}{n}$ tests, equal to the number of pools, because each pool receives an initial test. In the second round, all individuals in positive pools are re-tested, while negative pools are finished. The true proportion of negative pools is equal to the probability that *no individuals* of a pool test positive, and the true proportion of positive pools is equal to the probability that *at least one* individual tests positive in a pool. The former is equal to $(1-p)^n$, and the latter is the complement of this, or $(1-(1-p)^n)$. Multiplying $(1-p)^n$, the proportion of negative pools, by $\frac{N}{n}$ results in the number of negative pools, and multiplying this by *n,* the number of people per pool, results in the total number of people in negative pools:

$(1-p)^n \cdot \frac{N}{n} \cdot n$. The same can be done with the the proportion of positive pools to get $(1-(1-p)^n) \cdot \frac{N}{n} \cdot n$ as the total number of people in positive pools. Because testing accuracy is assumed to be perfect, this expression represents the number of tests needed in the second round. Figure 1 below visualizes the process of the derivation.

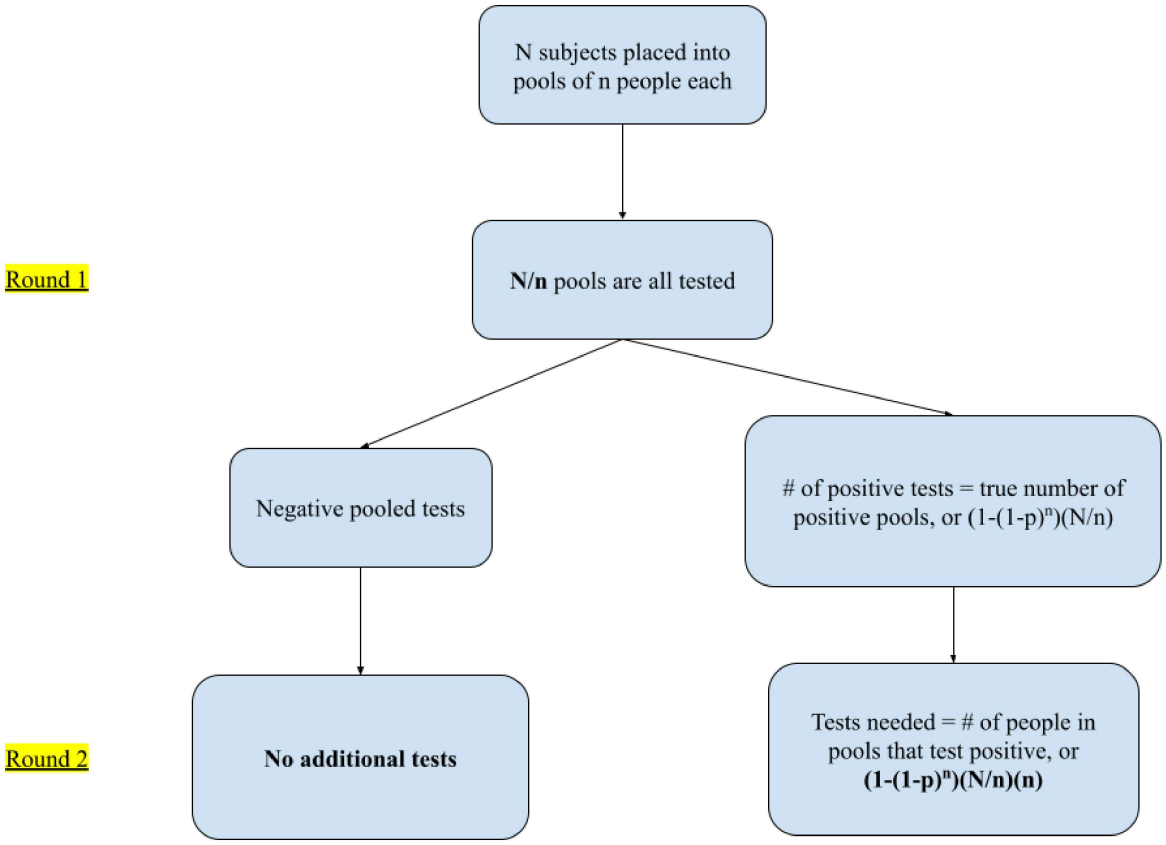

Figure 1: Flowchart of a pooled testing approach with perfect testing accuracy. **Bold** indicates number of tests needed in each round

Thus, $t_p = \frac{N}{n} + (1-(1-p)^n) \cdot \frac{N}{n} \cdot n$. After substituting t_o and t_p into the formula for F and simplifying, the derivation is complete (and terms involving N cancel out, indicating that population size theoretically does not have an effect):

$$F = (1-p)^n - \frac{1}{n} \quad (1)$$

This equation is consistent with those produced in other studies, although a metric other than reduction factor is typically used (discussed later in section 3.2) (Aragón-Caqueo D et al, 2020).

2.2.1. Incorporating Testing Accuracy

The table below (Figure 2) illustrates the classifications of testing results. Let S_n represent testing sensitivity and S_p represent specificity. By definition $S_n = \frac{\#\ of\ true\ positives}{\#\ of\ true\ positives + \#\ of\ false\ negatives}$ and $S_p = \frac{\#\ of\ true\ negatives}{\#\ of\ false\ positives + \#\ of\ true\ negatives}$. In the derivation above (section 2.2), perfect testing accuracy is assumed. This means that sensitivity and specificity are both equal to 1. In reality, however, this is not the case, so the number of positive test results would include both true positives and false positives.

		Disease Condition	
		Positive	**Negative**
Test Result	**Positive**	True positive	False positive
	Negative	False negative	True negative

Figure 2: Testing classification table

S_n and S_p do not affect t_o, as all subjects will be tested regardless of test accuracy when there is no pooled testing. This also does not affect the first round of testing, since all pools still must receive an initial test.

For the second round, $(1 - (1 - p)^n) \cdot \frac{N}{n} \cdot n$, the theoretical actual number of people in positive pools, would no longer equal the total number of tests needed. This is because the imperfect testing accuracy means that some positive pools may be undetected. Therefore, $(1 - (1 - p)^n) \cdot \frac{N}{n}$ must be multiplied by S_n to obtain the number of positive pools that also *test* positive. Multiplying by n yields the number of people in these pools:

$$S_n(1 - (1 - p)^n) \cdot \frac{N}{n} \cdot n.$$

False positives, negative pools that *test* positive, would also contribute to the number of tests needed. The number of false positive pools is equal to the false positive rate multiplied by the number of negative pools. So, $(1 - S_p)$ is multiplied by $(1 - p)^n \cdot \frac{N}{n}$. Multiplying this then by n results in the number of people in false positive pools: $(1 - S_p) \cdot (1 - p)^n \cdot \frac{N}{n} \cdot n$. The figure below helps to illustrate this process.

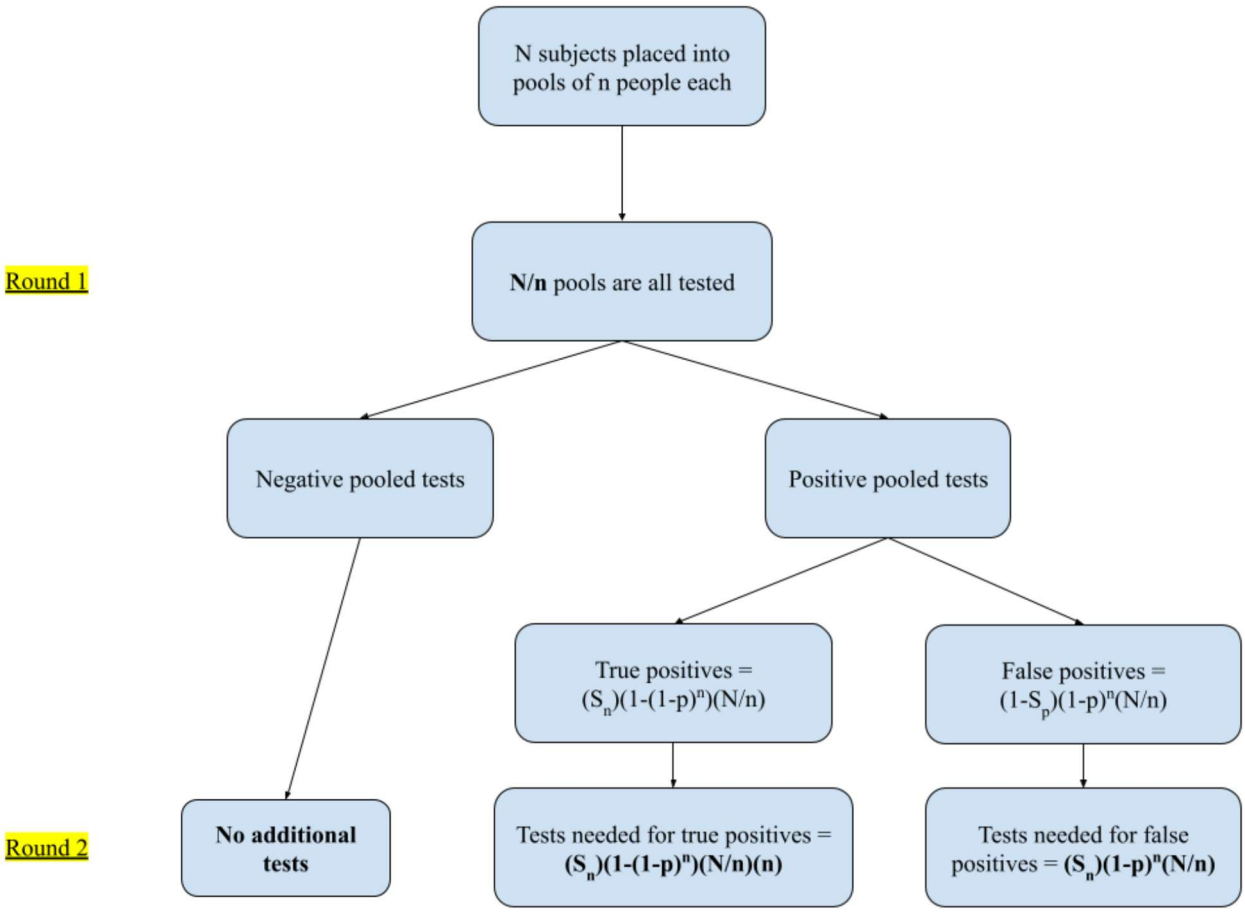

Figure 3: Flowchart of a pooled testing approach with testing sensitivity S_n and specificity S_p. **Bold** indicates number of tests needed in each round

The sum of the number of people in true positive pools and the number of people in false positive pools is the total number of people who receive a test in the second round. Hence,

$t_p = \frac{N}{n} + \left[S_n(1 - (1 - p)^n) \cdot \frac{N}{n} \cdot n\right] + \left[(1 - S_p)(1 - p)^n \cdot \frac{N}{n} \cdot n\right]$, and the modified reduction factor formula is:

$$F_m = 1 - \frac{1}{n} - S_n(1 - (1 - p)^n) - (1 - S_p)(1 - p)^n \quad (2)$$

2.3. Formula for optimal pool size

Equation (2) was graphed with F_m as a function of n to determine a procedure for optimization. Sensitivity, specificity, and prevalence rate are constants, so S_n, S_p, and p were arbitrarily assigned values of 1, 1, and 0.05, respectively, for visualization purposes.

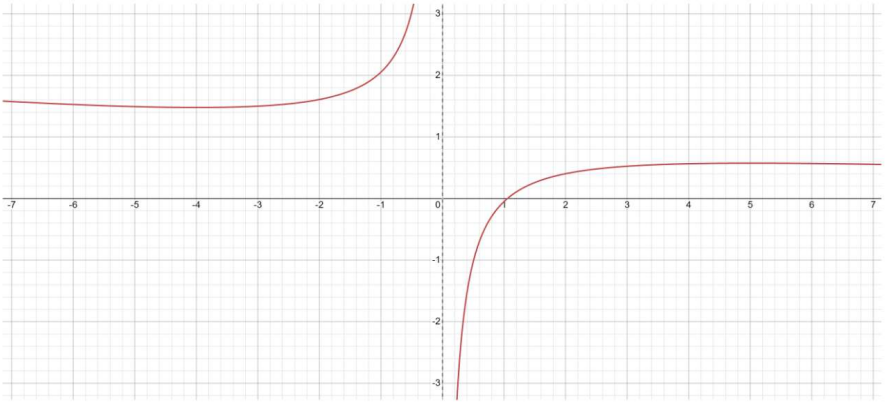

Figure 4: the graph of equation (2) has a relative maximum in the first quadrant

The graph appears quite complex, but it can be simplified. Pool size must be greater than or equal to 1, so negative values of n can be ignored. Further, the intended reduction factor will always be a positive value so negative values of F_m can be ignored as well. This leaves only the first quadrant, in which there is a single relative maximum. The value of n that corresponds to this maximum is the pool size that causes the greatest possible reduction factor, i.e., the optimal pool size. This inspires the use of calculus, as the maximum point must have an instantaneous slope of zero. Thus, by taking the first derivative of equation (2) and setting it equal to zero, the final formula for optimal pool size can be obtained. The derivative of equation (2) is $\frac{dF_m}{dn} = \frac{n^2(1-p)^n ln(1-p)(S_n+S_p-1)}{n^2}$. When the numerator equals zero, the derivative equals zero, so the final optimization equation is:

$$0 = n^2(1-p)^n \ln(1-p)\,(S_n + S_p - 1) + 1 \tag{3}$$

3. Discussion

3.1. Binomial distribution

One major result is the observation that the distribution of the number of positive cases in pools is binomial. This was shown through the R simulation, where n=5, p=0.05.

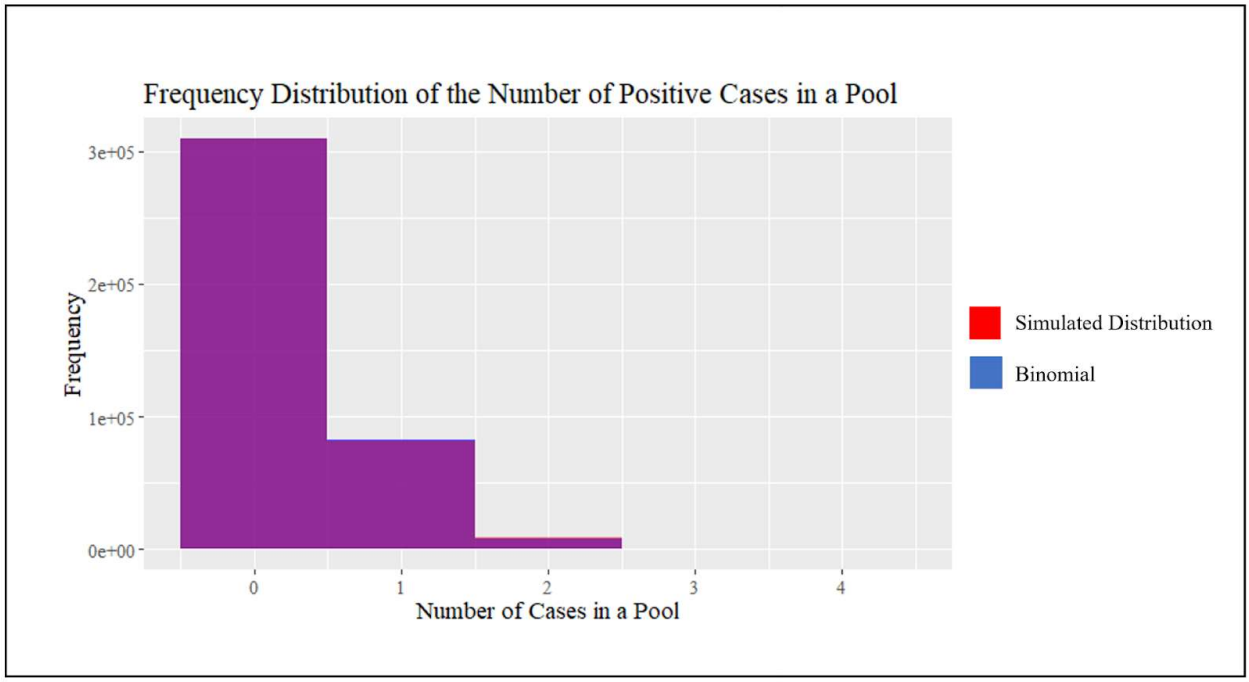

Figure 5: overlaid frequency histograms of the simulated distribution and binomial distribution

In figure 5, because the simulated distribution (red) matches the expected binomial distribution (blue), the overlaid histograms form what appears to be a single purple distribution. This provides evidence that the distribution is in fact binomial.

Further, in pooled testing, the pool size is fixed, each person can only test positive or negative, and the prevalence rate is constant. Thus, each pool is just a series of *n* Bernoulli trials, so it is logical that the distribution is binomial when there is a large number of pools. This allows for calculation of the proportion of pools with some specific number of positive cases using the binomial probability mass function:

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x} \tag{4}$$

In this formula, n is the number of trials, p is the probability of success, and x is the number of successes. Under the context of pooled testing, f(x) represents the probability that a pool of n people has x positive cases given a prevalence rate p.

It should also be noted that in reality, each trial may not be independent. Creation of pools would likely be done via cluster sampling, as the grouping of potential positive individuals into the same pool may boost the reduction factor. Or, this could provide more convenience (e.g., members of the same family are placed in the same pool). Overall, the distribution will not be perfectly binomial due to the fact that for any contagious disease it is almost impossible for individuals to be independent.

3.2. Reduction factor

In most of the studies regarding this matter, an explicit definition of reduction factor is not used; instead, a variable for the average minimum number of tests needed to diagnose a subject is used (Aragón-Caqueo D et al., 2020). In a simple scenario where all individuals are tested, the average number of tests needed per individual is just 1. If pooled testing allows this number to fall to 0.4, for instance, then the other 0.6 is the amount saved. Thus, the reduction factor is the complement of the average minimum number of tests needed per individual. Reduction factor was used in this study because it makes for an easier interpretation. It also allows for a convenient conversion to the cost saved: (reduction factor)*(cost of test).

3.2.1. Modified reduction factor

The modified reduction factor formula (2) shows the effect of specificity and sensitivity. When S_n and S_p both equal one, formula (2) becomes equal to formula (1). This aligns with the fact that formula (1) assumes perfect testing accuracy.

When sensitivity (S_n) increases, $S_n(1 - (1 - p)^n)$ also increases, so the reduction factor decreases. This is reasonable because sensitivity measures the ability of a test to accurately identify patients with a disease. Higher sensitivity means positive samples will be more likely to be detected, which leads to more tests being needed to re-test the individuals of those positive

pools. Conversely, lower sensitivity means more positive samples go unnoticed. Negative pools are not re-tested, so the reduction factor is increased artificially.

When specificity (S_p) increases, $(1 - S_p)$decreases, so $(1 - S_p)(1 - p)^n$also decreases, which leads to an increase in the reduction factor. Logically, this is because specificity is a measure of a test's ability to accurately identify patients who do not have a disease. Thus, higher specificity means a higher proportion of actual negative samples will be correctly classified by the test as negative. As a result, there are less false positives, so more tests can be saved.

It is important to note that an increase in reduction factor through a decrease in sensitivity is not desirable. An increase in the reduction factor is contingent on having sufficient testing accuracy first. An increase in reduction factor through increasing specificity, on the other hand, is ideal.

3.3. Usage of formula (3) for optimal pool size

Formula (3) provides the ability to calculate precise optimal pool sizes based on prevalence rate, sensitivity, and specificity, but there are a number of nuances to be aware of.

A slight drawback of equation (3) is that since it is in the exact form, the formula makes it too difficult to solve for *n.* Thus, one must substitute the constant values (p, S_n, and S_p) and input the resulting equation into a computation engine, such as WolframAlpha. The engine can then solve for the value of n through the iterative method.

Additionally, there are domain restrictions of equation (3) to be aware of. Group size *n* must be an integer greater than 1, as pools must contain an integer number of two or more people. This means that negative solutions of equation (3) can be disregarded. Further, equation (3) may return two positive solutions for *n*. This is because the graph of formula (2) may have a relative minimum in the fourth quadrant, meaning a second positive value of *n* that has a first derivative of 0. However, this second solution corresponds to a negative reduction factor, meaning *more* tests would be used than in a simple testing approach. The purpose of pooled testing is to achieve a positive reduction, so the smaller of the two positive solutions for *n* is the desired one. To find the true optimal pool size, the value should then be rounded as necessary to the next highest or lowest integer, depending on which yields a higher reduction factor.

3.4. Impact of prevalence rate on pooled testing strategy

To understand the impact of prevalence rate on the reduction factor, equation (2) was graphed for varying prevalence rates with S_n and S_p equal to one.

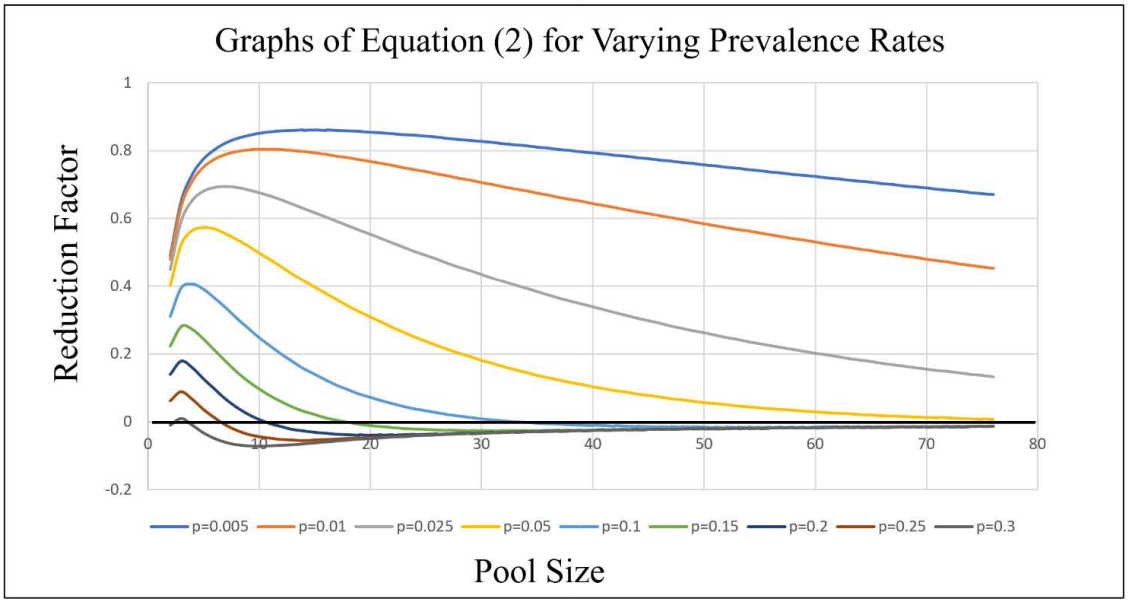

Figure 6: First quadrant graphs of equation (2) for a number of common prevalence rates

Figure 6 demonstrates that the reduction factor curves experience a downward shift as prevalence rate increases. Observing the maximum of each curve, there are also significant decreases in the maximum reduction factor.

The curves also suggest that there is a prevalence rate at which pooled testing is no longer a logical strategy. On the curve representing p=0.3, the peak is barely above the x axis; for greater prevalence rates, even the maximum reduction factor will become negative, indicating a waste of resources if pooled testing is used. For a reduction factor that is negative or 0, the pooled testing strategy is clearly not ideal. For a reduction factor that is small but positive, researchers must decide whether the potential inconvenience associated with pooling samples is worth the reduction in resources.

4. Conclusion

Pooled PCR testing is a strategy at the forefront of fighting the pandemic. In order to effectively execute this strategy, it is necessary to have a means of calculating and predicting related statistics. This study related pooled testing to the binomial distribution, derived formula (2) for the modified reduction factor, and derived formula (3) for the optimal pool size while accounting for testing accuracy. These findings allow for public health officials to perfect their design of a testing plan. Finally, increasing specificity and decreasing sensitivity both result in increasing maximum reduction. However, decreasing sensitivity leads to a higher risk of disease spread. Sufficient testing sensitivity should not be sacrificed for higher pooled testing efficiency.

5. Bibliography

1. Aragón-Caqueo D, Fernández-Salinas J, Laroze D. Optimization of group size in pool testing strategy for SARS-CoV-2: A simple mathematical model. J Med Virol. 2020;10.1002

2. Centers for Disease Control and Prevention. (2021, August 24). *Coronavirus disease 2019 (COVID-19) 2021 case definition*. Centers for Disease Control and Prevention. Retrieved November 6, 2021, from https://ndc.services.cdc.gov/case-definitions/coronavirus-disease-2019-2021/.

3. dan Yelin, Noga Aharony et al, Evaluation of COVID-19 RT-qPCR Test in Multi sample Pools, *Clinical Infectious Diseases*, Volume 71, Issue 16, 15 October 2020

4. Francesca Regen, Neriman Eren, Isabella Heuser, Julian Hellmann-Regen, A simple approach to optimum pool size for pooled SARS-CoV-2 testing, International Journal of Infectious Diseases, Volume 100, 2020, Pages 324-326

5. Interim guidance for use of Pooling procedures in SARS-COV-2 Diagnostic, screening, and Surveillance Testing. (n.d.). Retrieved March 06, 2021, from https://www.cdc.gov/coronavirus/2019-ncov/lab/pooling-procedures.html

6. Patel, N. V. (2020, April 10). *Why the CDC botched its coronavirus testing*. MIT Technology Review. Retrieved November 6, 2021, from https://www.technologyreview.com/2020/03/05/905484/why-the-cdc-botched-its-coronavirus-testing/.

7. Watloms AE et al. Increasing SARS-CoV-2 Testing Capacity with pooled Saliva Samples. Emerging Infectious Diseases. 27(4), April 2021.