Application of Pooled Testing in Screening and Estimating the Prevalence of COVID-19

Anjali Susan Oommen

April, 2022

I will be reviewing the article: Application of pooled testing in screening and estimating the prevalence of COVID-19 by Pritha Guha, Apratim Guha and Tathagata Bandyopadhyay. I attempted to understand the article and recreate all the tables in the paper on my own.

Introduction

As of 30^{th} April, 8:00 IST, there have been over 43 million Covid cases in India. This data was found on the website: Worldometers which is a verified source. However 40 - 45 percent of SARS-CoV-2 infections are asymptomatic. So the actual number of cases may be much more. This leads to two issues.
The first issue is that these asymptomatic cases spread the virus faster. In order to combat this, the World Health Organisation has said that the most effective way to control the spread of the disease is to test as many people as possible. However that poses an issue in a developing country like India where high costs and short supply of testing kits are problems. Some experts proposed the use of Dorfman's pooled testing as they felt that it would reduce the number of tests and subsequently reduce costs.
The second issue is that due to the presence of asymptomatic undetected cases of SARS - Cov - 2, the prevalence of the disease was miscalculated which initially led to scandalously high fatality rates as the WHO based it on fraction of deaths among symptomatic cases instead of fraction of deaths among actual number of cases(symptomatic and asymptomatic). Most experts recommend pooled testing only for screening purposes but later developments in pooled research has shown that it can be used for estimation of prevalence.
This paper will be divided into two parts: statistical theory behind the Dorfman's pooled testing procedure for screening (given that the test is either perfect or imperfect) and and the statistical theory behind finding the prevalence.
The tables constructed for this paper can be found on this link : Tables for Dorfman's pooled testing

1 Dorfman's pooling technique

In 1943,Robert Dorfman published an article, "The Detection of Defective Members of Large Populations". In this, he detailed a new way of testing which could reduce the cost of testing. His motivation behind this was that United States Public Health Service needed an inexpensive and efficient way to test men for syphilis during the Second World War. The technique for Dorfman's pooled testing is at follows - Suppose there are n samples to be tested for a specific disease. Instead of doing n tests to test each sample, we test a pool with all n samples. If the test comes back negative, we know that all n samples are negative so we have done 1 test instead of n tests. If the test comes back positive, we will test all the n samples individually. In this case, we would have done $n+1$ tests instead of n tests.
We can attach a Bernoulli variable, Y to each sample . If the sample is positive, $Y = 1$. If the sample is negative, $Y = 0$. Thus given n samples, we can see that

$$Y_n^* = max(Y_1, Y_2, Y_3.....Y_n)$$

where Y_n^* is the Bernoulli variable attached to the pooled group which contains n individual samples and Y_i where $i \in [1, n]$ is the Bernoulli variable attached to each sample.

It can be seen that if even one sample is positive i.e $\exists Y_i = 1$ where $i \in [1, n]$, the pooled group will give a positive result ($Y_n^* = 1$) which will lead to more tests. However the pooled sample will give a negative result ($Y_n^* = 0$) only if all samples are negative i.e. $\forall i \in [1, n]$, $Y_i = 0$.
The probability of the pooled group giving a positive result will depend on multiple parameters such as prevalence and group size when test being used is perfect. When the test is imperfect, in addition to this, there are other parameters such as sensitivity and specificity.
Given a sample of size n, if n is large we do not generally test the group as a whole. Instead we divide the sample into j groups of size k each.
$n = jk$

1.1 Definitions

.
Prevalence : Prevalence is the proportion of a particular population found to be affected by a medical condition at a specific time. It is a measure of disease that allows us to determine a person's likelihood of having that disease. It is denoted by p.

$$p = \frac{\text{number of people with the disease}}{\text{population}}$$

Prevalence is the probability of having the disease.

True Positive : The result of test is said to be a true positive if the test result is positive and the person is infected.

False Positive : The result of test is said to be a false positive if the test result is positive and the person is actually not infected.

True Negative : The result of test is said to be a true negative if the test result is negative and the person is not infected.

False Negative : The result of test is said to be a false negative if the test result is negative and the person is actually infected.

Sensitivity/True positive rate : The sensitivity of a test is its ability to correctly identify people with the disease. It is the probability of a positive test, conditioned on truly being positive. It is denoted by S_e.

$$S_e = \frac{\text{True positive}}{\text{True positive} + \text{False negative}}$$

$$S_e = \frac{\text{True positive} + \text{False negative} - \text{False negative}}{\text{True positive} + \text{False negative}}$$

$$S_e = 1 - \frac{\text{false negative}}{\text{Actual positives}}$$

Specificity/True negative rate : The sensitivity of the test is its ability to correctly identify people without the disease. It is probability of a negative test, conditioned on truly being negative. It is denoted by Sp

$$S_p = \frac{\text{True negative}}{\text{True negative} + \text{False positive}}$$

$$S_p = \frac{\text{True negative} + \text{False positive} - \text{False positive}}{\text{True negative} + \text{False positive}}$$

$$S_p = 1 - \frac{\text{false positive}}{\text{Actual negatives}}$$

Efficiency : Percentage reduction in the expected number of tests achieved by using the pooled testing over the individual testing.

$$E(N) = \frac{E(N_p) - E(N_i)}{E(N_i)}$$

1.2 When the test is perfect

$$P(Y_t^* = 0) = (1-p)^k \tag{1}$$

where $P(Y_t^* = 0)$ where $t \in [1, j]$ is the probability that the ith group of the given sample tests negative. This is because as seen earlier $P(Y_t^* = 0)$ only if all samples in the pooled group are negative.
P(negative sample) = 1 - P(positive sample) = 1 - prevalence
P(k negative samples) $= (1-p)^k$
$\implies P(Y_t^* = 0) = (1-p)^k$
No of tests required = 1

$$P(Y_t^* = 1) = 1 - (1-p)^k \tag{2}$$

Either the pooled group returns a positive or a negative result.
$P(Y_t^* = 1) = 1 - P(Y_t^* = 0)$
$\implies P(Y_t^* = 1) = 1 - (1-p)^k$
No of tests required $= k + 1$

Therefore the expected number of tests, N_t is

$$E(N_t) = k \times P(Y_t^* = 1) + 1 \times P(Y_t^* = 0)$$

$$E(N_t) = k + 1 - k(1-p)^k \tag{3}$$

It can easily be seen that if the prevalence is low, the expected number of tests for each group would be close to 1 even for a sufficiently large group size k(Using Equation (3)). So when pooled testing is used around 1 test would have to be done per group(k samples) vs when individual testing is used k tests would have to be done per group. Thus in a group of n samples where $n = jk$. If the prevalence is low and we used pooled testing, the expected number of tests for the whole sample would be j (for k samples we perform 1 test). Whereas for individual testing, the expected number of tests for the whole sample would be $j \times k$. It is clear to see that if the prevalence is low, it would save money and time if pooled testing was used.

1.3 When the test is imperfect

Dorfman suggested this pooling procedure assuming that test being used was perfect(i.e no false positives and no false negatives). However in real life, tests are rarely ever perfect. There are chances of false positives and negatives. Thus to talk about a test, we need to know how accurate this test is. This accuracy is generally measured by two numbers - sensitivity and specificity(definitions given in subsection 1.1). The higher the sensitivity and specificity, the better the test is. Even if a test has a high sensitivity and a high specificity in laboratory settings (controlled temperature, pressure and so on), outside the laboratory the test will have lower sensitivity and specificity values due to testing conditions, methods of sample preservation and methods of sample collection.
Let observed pooled test result of group t be $\hat{Y}_t^*$ and actual pooled test result of group t be Y_t^*.

$$P(\hat{Y}_t^* = 1 | Y_t^* = 1) = S_e$$

$\because S_e = \frac{\text{True positive}}{\text{Actual positive}}$

Similarly

$$P(\hat{Y}_t^* = 0 | Y_t^* = 0) = S_p$$

An observed positive can happen in two ways - observed positive and sample is truly positive and observed positive and sample is truly negative

$$P(\hat{Y}_t^* = 1) = P(Y_j^* = 1)S_e + P(Y_j^* = 0)(1 - S_p)$$

$$P(\hat{Y}_t^* = 1) = (1 - (1-p)^k)S_e + (1-p)^k(1 - S_p)$$

$$P(\hat{Y}_t^* = 1) = S_e - S_e(1-p)^k + (1-p)^k(1 - S_p)$$

$$\therefore P(\hat{Y}_t^* = 1) = S_e + (1-p)^k(1 - S_p - S_e) \tag{4}$$

Similarly an observed negative can happen in two ways - observed negative and sample is truly positive and observed negative and sample is truly negative

$$P(\hat{Y}_t^* = 0) = P(Y_j^* = 1)(1 - S_e) + P(Y_j^* = 0)S_p$$

$$P(\hat{Y}_t^* = 0) = (1 - (1-p)^k)(1 - S_e) + (1-p)^k S_p$$

$$P(\hat{Y}_t^* = 0) = 1 - S_e - (1-p)^k(1 - S_e) + (1-p)^k S_p)$$

$$\therefore P(\hat{Y}_t^* = 0) = 1 - S_e - (1-p)^k(1 - S_e - S_p) \tag{5}$$

Now we can see that the expected number of tests required for a group of k people when the test is imperfect with sensitivity = S_e and specificity = S_p

$$E(N_t) = 1 \times P(\hat{Y}_t^* = 0) + (k+1) \times P(\hat{Y}_t^* = 1)$$

$$E(N_t) = 1 \times \big(1 - S_e - (1-p)^k(1 - S_e - S_p)\big) + (k+1) \times \big(S_e + (1-p)^k(1 - S_p - S_e)\big)$$

$$E(N_t) = \big(1 - S_e - (1-p)^k(1 - S_e - S_p)\big) + k \times \big(S_e + (1-p)^k(1 - S_p - S_e)\big) + 1 \times \big(S_e + (1-p)^k(1 - S_p - S_e)\big)$$

$$\therefore E(N_t) = 1 + k\big(S_e + (1-p)^k(1 - S_p - S_e)\big) \tag{6}$$

Given that there are j groups with equal samples(k) , the total number of expected tests is $j \times E(N_t)$
Above we were simply looking at the result of the pooled group of samples (one result) which is why we used S_e and S_p.

However when we are looking at pooled testing as an approach, we can only fully understand its use if we look at the accuracy of pooled testing which we can measure with the sensitivity and specificity of Dorfman's procedure. Let us denote this with S_e^D and S_p^D.
Let us use a few symbols so that we can easily read the next part
T^+ = Test positive
T^- = Test negative
I = infected
I^c = uninfected

$S_e^D = P(T^+|I)$
$S_p^D = P(T^-|I^c)$

S_e^D
When using Dorfmans procedure, we know that a person turns out positive only after 2 outcomes :

pooled test comes back positive and individual test also comes out positive
$S_e^D = S_e \times S_e$

$$S_e^D = S_e^2 \tag{7}$$

S_p^D
A person who is negative can be present in 2 groups:
Group 1 : Pool consists entirely of uninfected people
Group 2 : Pool contains some infected individuals

P(person being part of Group 1) $= (1-p)^k$
P(person being part of Group 2) = P(group with size k -1 being infected) $\times$ P(kth person not being infected)
P(person being part of Group 2) $= \big(1-(1-p)^{k-1}\big) \times (1-p)$
P(person being part of Group 2) $= (1-p) - (1-p)^k$
Specificity can be written as 1 - false positive rate. Using that going forward.
In Group 1, for a person to falsely test positive both the individual and the pool test must give false positive,
P(falsely testing positive in group 1) $= (1-S_p)^2$
In Group 1, for a person to falsely test positive, the individual test will give a false positive but the pooled test will give a true positive (as there are other infected people)
P(falsely testing positive in group 2) $= (1-S_p)S_e$

Therefore the probability of a person falsely testing positive is

$$\text{False positive rate} = \frac{(1-S_p)^2(1-p)^k + (1-S_p)S_e\big((1-p)-(1-p)^k\big)}{1-p}$$

($\because$ false positive rate = false positives / actual negatives)

$$\text{False positive rate} = (1-S_p)\big(S_e + (1-S_p-S_e)(1-p)^{k-1}\big)$$

$$\therefore S_p^D = 1-(1-S_p)\big(S_e + (1-S_p-S_e)(1-p)^{k-1}\big) \tag{8}$$

$**$ S_e^D depends only on S_e
$**$ S_p^D depends on S_e, S_p, p, k

There are two more numbers which are often used to assess the performance of the test in everyday life. They are measures of diagnostic error and are called posterior or inverse probabilities.
False Positive Predictive Value (FPPV) $= P(I^c|T^+)$ (represents the proportion of misclassified individuals among those tested positive).
False Negative Predictive Value (FNPV) $= P(I|T^-)$ (represents the proportion of misclassified individuals among those tested negative)

Using Bayes theorem for conditional probability we know that $P(A|B) = \frac{P(A\cap B)}{P(B)}$

$$FPPV = \frac{(1-p)(1-S_p^D)}{(1-p)(1-S_p^D) + (pS_e^D)} \tag{9}$$

$$FNPV = \frac{p(1-S_e^D)}{p(1-S_e^D) + (1-p)S_p^D} \tag{10}$$

where $(1-p)$: probability of person not being infected
$(1-S_p^D)$: probability that pooled test gave a false positive
(pS_e^D) : probability that pooled test correctly identified an infected person
$p(1-S_e^D)$: probability that an infected person falsely tested positive in the pool test

$(1-p)S_p^D$: probability that a person who is not infected did not test positive in the pool test

In individual testing, $FPPV_I$ and $FNPV_I$ can be calculated by replacing S_e^D and S_p^D by S_e and S_p
In 1994, Litvak (in the context of HIV testing) (Guha et al, 2020, Litvak et al, 1994) proposed that five key elements should be used to capture the overall performance of a pooled testing procedure. There are as follows : $E(N_t), S_e^D, S_p^D, FPPV$ and $FNPV$.

1.3.1 Looking as the performance of the pooled test (Table 1)

Table 1 shows the values of the 5 key elements as well as $FPPV_I, FNPV_I, E_i(N_t)$ and efficiency given some values of p, S_e and S_p.

Conclusions drawn

1. With increase in p, and decrease in S_e and S_p, the efficiency reduces.
2. $S_e^D \leq S_e$ but for given set of p, S_e, S_p, $S_p^D > S_p$
3. For small values of S_p , the difference between S_p^D and S_p is substantial. This implies that the pooled testing procedure may have a much higher specificity, especially when the specificity of the individual test is low
4. There is a contrast of effects of S_e and S_p on FPPV. Even a slight decrease in S_p leads to a major increase in the value of $FPPV$ for any given values of p and S_e but a change in the value of S_e has barely any effect on the value of $FPPV$ for any given values of p and S_p.
5. With an increase in p, FPPV shows a decrease for any given values of Se and Sp.
6. Compared to the individual testing, the pooled testing leads to a substantial reduction in the value of FPPV.
7. The impact of changes in the values of the parameters has little effect on FNPV.
8. Pooled testing leads to a higher rate of false negative cases compared to individual testing. However this change is negligible

When testing for the SARS-CoV-2 virus, Table 1 suggests that when the prevalence is low, (less than 10 percent) pooled testing leads to a much lower rate of false positive cases than individual testing, especially if the specificity of the test used is low. Pooled testing leads to a slightly higher rate of false negative cases when compared to individual testing but as said in point 7 above, the effect is negligible.

Finally,the impact of these diagnostic errors are very different. High FPPV means many uninfected people are quarantined which causes physical and mental distress to many and also impacts the economy. However a high FNPV means many infected people are going on with their daily lives causing the disease to spread faster.

2 Derive for prevalence

This section will be further divided into subsections where we look at 2 questions - What is the theory of estimation of prevalence of a disease from the test data using basic probability theory? and What is the impact of misspecification of sensitivity and specificity on the estimate of prevalence.

2.1 Theory of estimation of prevalence of a disease

The reported percentage of asymptomatic patients vary from place to place but what most scientists and epidemiologists around the world can agree on is that there is a large percentage of undetected covid cases. This has led to overestimation of fatality rates(as discussed above). If, for example, 75 percent of the cases are asymptomatic, then the estimate of fatality rate would be 4 times the true fatality rate. This makes us believe that the virus is deadlier than it actually is. Ioannidis, a Stanford medicine professor claims that no country has "reliable data on the prevalence of the virus in a representative random

sample of the general population". He argued that had we had such data, the estimation of fatality rate would not be the gross overestimation it is now. Ioannidis has said that if the virus is not as deadly as we assume, we could be destroying the economy (as we have been doing to fight the virus) over nothing. A team of Stanford medical scientist published a research paper where they estimated the prevalence of SARS-CoV-2 virus by selecting a sample of over 3000 people. They estimated the prevalence to be much higher than it was in official data. The theory they used for estimation is given below

Let the probability of a positive test result for an arbitrarily chosen individual be denoted by π.

$$\pi = P(\text{positive result}) = (1-p)(1-S_p) + pS_e$$

Now let us denote proportion of positive results from the sample tested by $\hat{\pi}$. We can rearrange the terms of the above equation to get

$$p = \frac{\hat{\pi} + S_p - 1}{S_e + S_p - 1} \tag{11}$$

If we intend on using Dorfman's pooled testing procedure, we have to use S_e^D and S_p^D instead of S_e and S_p

$$\pi = (1-p)(1-S_p^D) + pS_e^D$$

$$\pi = (1-p)S_e(1-S_p) + (1-S_p)(1-S_p-S_e)(1-p)^k + pS_e^2 \tag{12}$$

We can put in the estimate of π got from the data and solve for p. This answer is consistent for p i.e it is close to p as the sample size gets larger.

2.1.1 Looking at the probability of testing positive for individual testing as well and Dorfman's pooled testing (Table 2)

In Table 2, the value of π is found for both individual testing (π_i) (Using Equation (11)) and Dorfman's pooled testing(π_D) for different values of p, S_e and S_p (Using Equation (12)) . The optimal pool size was taken from Dorfman's paper. He used the relative cost curve to find the optimal size for multiple values of p. (Dorfman,1943)

Conclusions drawn

1. S_e has a negligible impact on π but S_p has a substantial impact on π .
2. So S_e and S_p will also have a similar impact on estimation of prevalence.

So for accurate estimation of π, it is important to specify S_p correctly but specifying S_e wrong has very little effect.

2.2 The impact of misspecification of sensitivity and specificity on the estimate of prevalence

As discussed in Section 1.3, the sensitivity and specificity of a test is not the same in the field and in the laboratory. Let us denote the sensitivity and specificity of a test in the lab as S_e^P and S_p^P and the sensitivity and specificity of a test in real world situations as S_e^T and S_p^T. A lot of the time, S_e^P and S_p^P are substantially higher than S_e^T and S_p^T. Scientists generally use S_e^P and S_p^P as the sensitivity and specificity values as it is influenced by laboratory values. However this would lead to a deviation of the estimated prevalencep_p from the true prevalence, p_t. If S_e^T and S_p^T are used in place of S_e and S_p, we would get the true prevalence, p_t.

We study the effect of the misspecification of sensitivity and specificity values on the estimate of prevalence by evaluating the bias $(p_t - p_p)$.

2.2.1 Understanding effects of misspecification of S_e and S_p(Table 3)

In Table 3, we report the bias for both individual and Dorfman's testing for different values of prevalence, perceived sensitivity and specificity and true sensitivity and specificity. Here as well the optimal pool size was taken from Dorfman's paper. (Dorfman,1943) A look up table was used to find π_D.

Conclusions drawn

1. With increase in prevalence p, the bias decreases.

2. The misspecification of S_e and S_p does not have the same effect on prevalence . Misspecification of S_e has a little effect on the bias. However, the misspecification of S_p has a significantly large effect on the bias. Also, more is the deviation from the true values more is the effect on bias.

3. When compared to individual testing, it is important to note that pooled testing shows less of a bias due to misspecification. This is particularly interesting as misspecification is quite common.

Conclusion

We have discussed the statistical theory behind Dorfman's pooled testing which is used for screening. We have also discussed the theoretical method for estimation of prevalence from both individual and pooled testing data. In Section 1.3.1, we have also discussed what this theoretical data means in practical problems.

The theoretical data shows that pooled testing is preferable to individual testing as it reduces time and cost of screening. Pooled testing also gives a much lower value of FPPV than individual testing. This means that the number of people getting a false positive is much lower with pooled testing. This could lead to less disruption in economy and day to day life. However it is important to note here that all our observations have been made considering low values of prevalence (less than or equal to 10 percent).

The tables constructed also show that for prevalence estimation it is better to use data from pooled testing than individual testing as the bias is less. Here as well it is important to note here that all our observations have been made considering low values of prevalence (less than or equal to 10 percent).

References:

- Pritha Guha, Apratim Guha, Tathagata Bandyopadhyay; 'Application of pooled testing in screening and estimating the prevalence of COVID-19', (May 27, 2020).
DOI : https://doi.org/10.1101/2020.05.26.20113696
Available at: https://www.medrxiv.org/content/10.1101/2020.05.26.20113696v1.full#disp-formula-3

- Dorfman, Robert. "The Detection of Defective Members of Large Populations." The Annals of Mathematical Statistics 14, no. 4 (1943): 436–40.
Available at : http://www.jstor.org/stable/2235930

- Eugene Litvak, Xin M. Tu Marcello Pagano Screening for the Presence of a Disease by Pooling Sera Samples, Journal of the American Statistical Association, 89:426, 424-434, (1994) DOI: 10.1080/01621459.1994.10476764

- Worldometers.info
Available at : https://www.worldometers.info/coronavirus/country/india/
Accessed on: 30th April,2021