Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Modeling the optimization of COVID-19 pooled testing: How many samples can be included in a single test?

Lu Liu *

School of Economics, Southwestern University of Finance and Economics, China

Keywords: Pooled testing COVID-19 Resource cost function Time function Machine learning

ABSTRACT

Objectives: This study tries to answer the crucial question of how many biological samples can be optimally included in a single test for COVID-19 pooled testing. Methods: It builds a novel theoretical model which links the local population to be tested in a region, the number of biological samples included in a single test, the “attitude” toward resource cost saving and time taken in a single test, as well as the corresponding resource cost function and time function, together. The numerical simulation results are then used to formulate the resource cost function as well as the time function. Finally, a loss function to be minimized is constructed and the optimal number of samples included is calculated. Results: In a numerical example, we consider a region of 1 million population which needs to be tested for the infection of COVID-19. The solution calculates the optimal number of biological samples included in a single test as 4.254 when the time taken is given the weight of 50% under the infection probability of 10%. Other combinations of numerical results are also presented. Conclusions: As we can see in our simulation results, given the infection probability at 10%, setting the number of biological samples included in a single test (in the integer level) at [4,6] is reasonable for a wide range of the subjective attitude between time and resource costs. Therefore, in the current practice, 5-mixed samples would sound better than the commonly used 10-mixed samples.

1. Introduction

Along with the rapid development of the COVID-19 pandemic around the globe, the need for testing the virus of SARS-CoV-2 at a massive level increases dramatically. The detection of SARS-CoV-2 RNA by Polymerase Chain Reaction (PCR) technology in testing is a common and sensitive method of COVID-19 diagnosis [1,2], even for reinfection [3]. In fact, whether such diagnostic tests are rapid and accurate enough or not can be the key to containing the pandemic [4,5].

After the outbreak in early 2020, to implement the full scale of work resumption in Wuhan, China, from May 15 to May 24, 2020, 9 million biological samples were collected, where >6.5 million were laboratory-tested [6,7]. In mid-August 2020, as school reopening is imminent, a massive level of COVID-19 tests with fast results was expected in the US [8]. On October 11, 2020, a small wave of outbreaks took place in Qingdao, China [9]. As a rapid response, the massive level population was quickly tested for COVID-19. A similar situation emerged in Chengdu, Shijiazhuang, Shenyang, and some other cities in China at the end of 2020. Mass testing is proven to be an efficient way to act as the urgent response to disease control [10–14], especially at the screening stage [15–18], where frequency and turnaround time are more important than the test sensitivity [19]. In practice, this would reduce the shock to the economy due to the COVID-19 pandemic to a large extent [20,21].

In addition, the widespread COVID-19 testing in both clinical and non-clinical settings is also paid much attention [22–25]. Unfortunately, the failure to test the possible COVID-19 cases early enough would cause tragic results to contain outbreaks [26]. In fact, the limited availability of testing COVID-19 may be one important reason for the disastrous outcome of the pandemic in many places [27].

In the meantime, more and more studies emerge to examine the practical performance and testing strategies for SARS-CoV-2 [28,29], where the optimal testing strategies and maths-based strategies have begun to be discussed [30,31]. Now, the pooled testing technique became more and more popular since July 2020 in both the US and many other regions [32]. As self-explained by its name, COVID-19 pooled testing mixes several biological samples in a “batch”, and then tests the pooled samples with a diagnostic test. This approach increases
the number of samples to be tested using the same (or similar) amount of resources, hence reducing the medical cost and saving the time for a single test.

The advantage of pooled testing is obvious in the sense that we can save substantial time by examining a single test for mixed biological samples from multiple individuals. In addition, it can be quite cost-effective as well since fewer quantities of scarce resources, such as test kits, other testing supplies, and medical workers, are used. In fact, cost-effectiveness plays a central role in the strategies of SARS-CoV-2 testing from both medical and management perspectives [33,34]. Indeed, “fast and cheap” are essential to obtain and manage the tests [35–37]. Most importantly, it can boost the test capacity of a region dozens of times. For a simple example, if 10 biological samples are mixed for a single test, then theoretically the local test capacity for identifying the COVID-19 cases would be increased by about 10 times.

In the fourth quarter of 2020, the global daily increase of confirmed COVID-19 cases keeps at the one-million level. In many places all over the world, million-scale biological samples need to be tested every day. Although the multi-sample pools have already been academically studied [38–41], to our knowledge, the optimization of COVID-19 pooled testing is still not paid enough attention. In practice, it appears that the “rule of the thumb” prevails. Sometimes, 5 biological samples are mixed, while sometimes 10 samples or even more are combined. Then, how many samples can be optimally included in a single test? Easy as it appears to be, however, the answer to such a simple question is yet unknown. This study, therefore, tries to answer this crucial question and contribute to both the literature and practice. This paper attempted to obtain the optimal number of samples to be included in a single test for COVID-19 pooled testing, to balance the trade-offs between first-attempt costs and re-examine costs. Especially from a practical perspective, it studies an important problem to deal with the COVID-19.

2. Methods

2.1. The basic model

Let us start from a simple setting. Considering that a local population to be tested in a region is $N_{total}$ and the number of biological samples included in a single test is $G_{test}$, hence the number of tests can be denoted as $N_{total}/G_{test}$. In addition, we consider both the physical or monetary resource cost and the time taken in each test. Therefore, the optimization of COVID-19 pooled testing requires a tradeoff between the two aspects. Consequently, we introduce a parameter $\lambda$ to show the “attitude” toward resource cost saving. The corresponding “attitude” toward time-saving is thus $1-\lambda$. Moreover, we consider that the resource cost in monetary value for a single test is $C_{test}(G_{test})$, and the time taken in a single test is $T_{test}(G_{test})$. As we can see, we treat both $C_{test}$ and $T_{test}$ as functions of $G_{test}$. This assumption is for a more general purpose, which can be simplified if both $C_{test}$ and $T_{test}$ are constants. However, we do see the nonlinear relationship in reality. Therefore, such a non-linear relationship may be more attractive.

Now, the total resource cost in the monetary value of the COVID-19 test in a region, $C_{total}$, can be expressed as follows.

$$C_{total} = \frac{N_{total}}{G_{test}} C_{test}(G_{test})$$

(1)

The corresponding total time taken, $T_{total}$, can be expressed as follows as well.

$$T_{total} = \frac{N_{total}}{G_{test}} T_{test}(G_{test})$$

(2)

Thus, we form a loss function to be minimized.

$$Loss = \lambda C_{total} + (1-\lambda) T_{total}$$

(3)

Substituting Eq. (1) and Eq. (2) into Eq. (3), and after simplification, we have:

$$min\text{Loss} = \frac{N_{total}}{G_{test}} [\lambda C_{test}(G_{test}) + (1-\lambda) T_{test}(G_{test})]$$

(4)

Although the construction of the above loss function appears to be subjective, it is a reasonable representation of the trade-off between resource cost and time cost in COVID-19 testing. Solving the first-order condition from the above expression, we can obtain the optimal value of the number of samples in a single test below.

$$G_{test}^{*} = \frac{C_{test}(G_{test})}{\frac{\partial C_{test}(G_{test})}{\partial G_{test}}} + (1-\lambda) \frac{\partial T_{test}(G_{test})}{\partial G_{test}}$$

(5)

2.2. The extended model

Here, we introduce an extended model, which fits reality better. According to the standard protocol of the pooled testing of COVID-19, the positive or negative outcome of the pooled test would result in different operational treatments. Therefore, we need to consider the outcome of a single test, which would be “Yes” (i.e., positive) or “No” (i.e., negative). If “No”, then this test is passed. If “Yes”, then everybody in this group of samples would be tested again individually. In addition, we need to consider the “background” of the test, which is the infection intensity or severity of COVID-19 in a region. We thus introduce a probability of local COVID-19 infection, $P_{infection}$. Now, the probability of the positive result in a single test, i.e., $P_{Yes}$, can be expressed as follows.

$$P_{Yes} = \frac{G_{test} P_{infection}}{N_{total}}$$

(6)

This is a multiplication of probabilities. First, there is a “background” infection rate in the local population. Second, it is the ratio of the sampled group to be positively tested. Therefore, when we link them together, we obtain Eq. (6). Of course, such notation relies on the assumption of uniform distribution of the COVID-19 infection in a target group of the population. Then, the corresponding probability of the negative result in a single test, $P_{No}$, becomes $1-P_{Yes}$.

Notably, when the result in a single test is positive and everyone in the group needs to be re-examined, both the resource cost and time required in that group would increase dramatically. We hence denote the resource cost of re-examining a group to be $G_{test} C_{test}(G_{test})$, where “($G_{test}=1$)” means that the function “$C_{test}(G_{test})$” is evaluated at the particular value of $G_{test}=1$. Please consider “$C_{test}(•)$” as a generalized function, then it would be easier to understand here. The corresponding time associated with the re-examination would be $T_{test}$, which is $G_{test}$ as well. As we can see, now both $C_{test}$ and $T_{test}$ are not functions of $G_{test}$ anymore as $G_{test}$ takes the specific numerical value of 1.

Finally, we can set up the mathematical expectation of the loss function as below.

$$E[Loss] = P_{Yes} \frac{N_{total}}{G_{test}} [\lambda C_{test}(G_{test}) + (1-\lambda) T_{test}(G_{test})]$$

(7)

$$+ (1-P_{Yes}) \frac{N_{total}}{G_{test}} [\lambda C_{test}(G_{test}) + (1-\lambda) T_{test}(G_{test})]$$

Simplifying the above equation along with Eq. (6), we have:

$$E[Loss] = P_{infection} \frac{G_{test}}{N_{total}} [\lambda C_{test}(G_{test}) + (1-\lambda) T_{test}(G_{test})]$$

(8)

In order to find the optimal solution, now we differentiate the above equation with respect to $G_{test}$ and we obtain the following expression.

$$\frac{\partial E[Loss]}{\partial G_{test}} = P_{infection} [\lambda C_{test}(G_{test}) + (1-\lambda) T_{test}(G_{test})]$$

$$- \frac{N_{total} G_{test}^{2}}{G_{test}^{2}} [\lambda C_{test}(G_{test}) + (1-\lambda) T_{test}(G_{test})]$$

$$+ \frac{N_{total}}{G_{test}} [\lambda C_{test}(G_{test}) + (1-\lambda) T_{test}(G_{test})]$$

(9)
Multiplying the term $G_{T test}^2$ on both sides of the above equation and after rearrangement, we have:

$$P_{infected}(\xi(C_{T test} = 1) = \frac{\partial C_{T test}(G_{T test})}{\partial G_{T test}} + (1 - \xi)|T_{T test}(G_{T test} = 1) = \frac{\partial T_{T test}(G_{T test})}{\partial G_{T test}}|G_{T test}^2$$

Now, for the more general situation, without any form of simplification, function as Eq. (7) has shown. This strand of methods has already intelligence and especially machine learning, which starts from a loss for its overall resource cost. The trouble is the number of valid observations. Using the latest information on the prices of single or pooled tests in China, we construct a data set of only 21 observations (Table 1).

As shown above, the method used here follows the idea of artificial intelligence and especially machine learning, which starts from a loss function as Eq. (7) has shown. This strand of methods has already emerged in the study of COVID-19 pandemic-related issues [42–44]. Now, for the more general situation, without any form of simplification, we assume the following functional form.

$$C_{T test} = a + bG_{T test}$$

We therefore have:

$$C_{T test}(G_{T test} = 1) = a + b$$

$$\frac{\partial C_{T test}(G_{T test})}{\partial G_{T test}} = 2bG_{T test}$$

Similarly, we assume:

$$T_{T test} = d + eG_{T test}$$

And we have:

$$T_{T test}(G_{T test} = 1) = d + e$$

$$\frac{\partial T_{T test}(G_{T test})}{\partial G_{T test}} = 2eG_{T test}$$

Substituting the above expressions into Eq. (10), we have:

$$P_{infected}(\xi(a + b - 2bG_{T test}) + (1 - \xi)(d + e - 2eG_{T test}))G_{T test}^2$$

$$+ N_{infected}(\xi(bG_{T test} + (1 - \xi)2eG_{T test})G_{T test}^2)$$

$$- N_{infected}(\xi(a + bG_{T test}) + (1 - \xi)(d + eG_{T test})) = 0$$

After rearrangement, we finally have:

$$G_{T test}^2$$

After rearrangement, we finally have:

$$2P_{infected}(\xi(a + b - 2bG_{T test}) + (1 - \xi)(d + e - 2eG_{T test}))G_{T test}^2$$

$$- (P_{infected}(\xi(a + b) + (1 - \xi)(d + e)) + N_{infected}(\xi(b + (1 - \xi)e))G_{T test}^2$$

$$+ (N_{infected}(a + N_{infected}(1 - \xi)d) = 0$$

As we know, the cubic equation is “notoriously” difficult to find a closed-form solution. Unfortunately, we cannot find a closed expression for the solution of the above equation. However, we can still use numerical methods to find an answer.

### 3. Data

As shown previously, the functional form of the resource cost for a single test $C_{T test}(G_{T test})$, and the time taken $T_{T test}(G_{T test})$ would be crucial. However, the available information on such functional form is very limited. Therefore, we have to derive them from both real-world data and simulation-based techniques.

For the resource cost of the test, we use the price of the test to stand for its overall resource cost. The trouble is the number of valid observations. Using the latest information on the prices of single or pooled tests in China, we construct a data set of only 21 observations (Table 1).

| $C_{T test}$ | $G_{T test}$ |
|-------------|-------------|
| 80          | 1           |
| 150         | 5           |
| 200         | 10          |
| 150         | 10          |
| 60          | 1           |
| 58          | 1           |
| 90          | 5           |
| 130         | 10          |
| 40          | 1           |
| 40          | 5           |
| 80          | 10          |
| 35          | 1           |
| 100         | 5           |
| 50          | 5           |
| 150         | 10          |
| 16          | 1           |
| 25          | 5           |
| 50          | 10          |
| 6           | 1           |
| 17.5        | 5           |
| 35          | 10          |

### Table 1

| $T_{T test}$ | $G_{T test}$ |
|--------------|-------------|
| 2            | 1           |
| 3            | 1           |
| 4            | 1           |
| 8            | 500         |
| 9            | 500         |
| 10           | 500         |

### Table 2

| $T_{T test}$ | $G_{T test}$ |
|--------------|-------------|
| 2            | 1           |
| 3            | 1           |
| 4            | 1           |
| 8            | 500         |
| 9            | 500         |
| 10           | 500         |
We consider a function mentioned in the above models. However, when the exogenous conditions change, such calculated results would change accordingly. In Table 5, we show the new results when \( N_{\text{Total}} \) and \( \zeta \) change their numerical values. As we can see, when \( N_{\text{Total}} \) becomes smaller, the number of samples included in a single test would be slightly larger. But it does not really make a difference in the practice level. On the contrary, the smaller \( \zeta \) leads to a significantly larger value in \( G_{\text{Total}} \). This result confirms our intuition that when the time taken in the test is given more weight in the decision (i.e., smaller value in \( \zeta \)), there must be more samples included in a single test.

However, the discussion of the function is much more complex. Although we do not have a closed-form solution for \( G_{\text{Total}} \) now, we can still use the Implicit Function Theorem to derive the partial derivative of \( G_{\text{Total}} \) with respect to \( P_{\text{Infection}} \) as below:

\[
\frac{\partial G_{\text{Total}}}{\partial P_{\text{Infection}}} = -\frac{\zeta(a + b - 2G_{\text{Total}}) + (1 - \zeta)(d + e - 2G_{\text{Total}}))}{\partial P_{\text{Infection}}} G_{\text{Total}} - P_{\text{Infection}}(\zeta - 2b) + (1 - \zeta)(-2e)G_{\text{Total}} \tag{19}
\]

By simplifying the above equation and rearranging terms, we have:

\[
\frac{\partial G_{\text{Total}}}{\partial P_{\text{Infection}}} = \frac{\zeta(a + b - 2G_{\text{Total}}) + (1 - \zeta)(d + e - 2G_{\text{Total}}))}{\partial P_{\text{Infection}}} G_{\text{Total}} - 2P_{\text{Infection}}(\zeta - 2b) + (1 - \zeta)(d + e - 2G_{\text{Total}})) \tag{20}
\]

We can easily locate a particular value where \( \frac{\partial G_{\text{Total}}}{\partial P_{\text{Infection}}} = 0 \) as follows:

\[
G_{\text{Total}}^{*} = \frac{\zeta(a + b) + (1 - \zeta)(d + e)}{2\zeta + 2(1 - \zeta)e} \tag{21}
\]

Please note that such a particular value only means that \( P_{\text{Infection}} \) does not influence \( G_{\text{Total}} \). However, the expression of Eq. (20) also reveals the complex nature of such a relationship. It can be either positive or negative in a more general situation. Therefore, the sign of the above expression may be ambiguous.

### 6. Conclusion and future work

This study has proposed a theoretical model for the number of biological samples needed in a single test. It will help in resource and time minimization. As we can see in our simulation results, given the infection probability at 10%, setting the number of biological samples included in a single test (in the integer level) at \([4,6]\) is reasonable for a wide range of the subjective attitude between time and resource costs. If we relax this criterion for a little bit, the range of \([3,7]\) would also be OK.
for the mixed biological samples from multiple individuals. Therefore, in the current practice, 5-mixed samples would sound better than the commonly used 10-mixed samples.

The method proposed in this study is of general purpose and it can be applied elsewhere other than in China. It can be used to analyze issues of other infectious diseases as well. One just needs to calibrate the functional parameters again with the local data, and the model would be still working. Although we may have made some novel contributions in this paper, this study still has some shortcomings.

On the one hand, concerns regarding the pooled testing arise in the sense that the virus concentration would decrease after being mixed with other samples. Indeed, sensitivity matters in the test [47,48]. Since samples are diluted, there is a higher likelihood to obtain false-negative results. Therefore, the chance of error in test results may be higher in mixed samples than that in a pure sample when other things are equal, which may result in higher infection risk as well as re-examine cost [49]. It is true in reality in China. In some less developed regions (or cities), 20 samples are pooled together in a single test, which is reported to be less accurate. In fact, some cases show that it may be one of the reasons for the delay in the detection of infected residents in the local area. Therefore, the expression of concern here is meaningful.

On the other hand, the data used in this study are still very limited. More information is needed to obtain the resource cost function and time function in the COVID-19 testing with higher accuracy. These issues among others are left for future research.

Funding

Supported by Guanghua Talent Project of Southwestern University of Finance and Economics.

Data availability statement

All data used in this study are publicly available.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

I thank the editors and anonymous referees for their valuable comments. All remaining mistakes are my own.

References

[1] Gandhi RT, Lynch JB, del Rio C. Mild or moderate Covid-19. N Engl J Med 2020;383:1757–66.
[2] Zhang Z, Bi Q, Fang S, et al. Insight into the practical performance of RT-PCR testing for SARS-CoV-2 using serological data: a cohort study. Lancet Microbe 2021;2(2):e279–87.
[3] Wyllie AL, Fournier J, Casanovas-Massana A, et al. Saliva or nasopharyngeal swab specimens for detection of SARS-CoV-2. N Engl J Med 2020;383:1283–8.
[4] Hannen CH, Michlmaier D, Gubbels SM, et al. Assessment of protection against reinfection with SARS-CoV-2 among 4 million PCR-tested individuals in Denmark in 2020: a population-level observational study. Lancet 2021;397(10280):1264–12.
[5] Van den Berg O, Martiny D, Rochas O, et al. Considerations for diagnostic COVID-19 tests. Nat Rev Microbiol 2021;19:171–83.
[6] China News, http://www.chinanews.com/sp2020/05-25/9194183.shtml.
[7] Liu L. Emerging study on the transmission of the novel coronavirus (COVID-19) from urban perspective: evidence from China. Cities 2020;103:102759.
[8] Rafiei Y, Mello MM. The missing piece — SARS-CoV-2 testing and school reopening. N Engl J Med 2020;383:e126.
[9] King YH, Ni W, Hu XW, et al. Rapid response to an outbreak in Qingdao. China. N Engl J Med 2020;383:e126.
[10] Studdert DM, Hall MA. Disease control, civil liberties, and mass testing — calibrating restrictions during the covid-19 pandemic. N Engl J Med 2020;383:102–4.

5
[45] The paper, https://www.thepaper.cn/newsDetail_forward_11322518.
[46] Heilongjiang Daily. http://epaper.hljnews.cn/hljrb/20200510/470583.html.
[47] Mina MJ, Parker R, Larremore DB. Rethinking covid-19 test sensitivity — a strategy for containment. N Engl J Med 2020;383:e120.

[48] Congrave-Wilson Z, Lee Y, Jumarang J, et al. Change in saliva RT-PCR sensitivity over the course of SARS-CoV-2 infection. JAMA 2021;326(11):1065–7.
[49] Coffey KC, Diekema DJ, Morgan DJ. Interpreting SARS-CoV-2 test results. JAMA 2021;326(15):1528–9.