Adaptive Group Testing Models for Infection Detection of COVID-19

Zizhan Tang
11th grade, Beijing 101 High school
Beijing, China
932381136@qq.com

***Abstract*—Two adaptive group testing models are studied based on zero-error criterion in this paper. Firstly, for the single-stage group testing model, the analytical expression of optimal number of grouping members without integer constraints is given. Secondly, the optimal number of grouping members with integer constraints are given by numerical calculation. Finally, a grouping coefficient-based multistage model and the principles for selecting optimal grouping coefficient are proposed. The results of this paper can help medical institutions to improve group testing efficiency of infection detection of COVID-19.**

Keywords—COVID-19, Infection detection, Group testing

I. INTRODUCTION

Since January 2020, a pandemic caused by COVID-19 has been spreading across the globe. More than thirteen million people have been infected by early July. How to control the pandemic is a great challenge for all mankind, and an important issue for the building of smart cities[1].

The first step to control the spread of the virus is to detect those who are already infected. The detection test commonly used now is the PCR- nucleic acid test. If the test result is positive, the individual tested will be considered infected; if negative, not infected. Currently, the PCR test takes about two hours to get the test result of a single case. However, the number of people need to take the test in big cities often reaches thousands to even millions. Individual PCR test is not efficient enough. Therefore, finding a more efficient detection method is of vital importance for blocking the spreading of COVID-19.

Comparing with individual testing, group testing is a more efficient method which tests mixed samples of a group of people. If the result of the test is negative, all the people in this group are considered not infected. If the result of the test is positive, there is at least one carrier of COVID-19 in the group. Then, individual tests need to be carried out to find the carrier(s) in this group.

Some factors are crucial to the detection efficiency of group testing, such as the size of the sample pool, the number of group members, and the design of the detection algorithms. When the number of individuals in each group is 1, the detection efficiency will be the same as that of individual test. When the number of individuals in each group is large, the infected individuals are uniformly distributed in the sample pool, and the probability of infection is high, group testing may take longer time, which is less efficient.

Therefore, to ensure that the efficiency of group testing is higher than that of individual testing, the group testing algorithm should be carefully designed, taking into consideration variables such as the size of the sample pool, the probability of infection and the number of group samples. From a strict mathematical perspective, some variables such as the size of the sample pool and the number of groups are integers. Usually it is difficult to find an analytical solution for optimization problems with integer constraints.

Dorfman [2] first proposed the idea of group testing and a single-stage grouping model. To solve the problem of finding the optimal numbers of grouping members, he used numerical calculation method, ignoring integer constraints. He also pointed out that no simple general solution to this problem can be found. Subsequently, Sobel et al.[3-4] proposed that binary splitting method could be used to improve the efficiency of group testing. The group testing models, such as those of Dorfman's and Sobel's, which continuously adapt to the change of test dataset in the detection process, are called adaptive group testing models, which have been further studied and improved by other researchers [5-6]. Another important type of models are non-adaptive. These models focus on the design of a pre-defined test array [7] and decoding algorithms. Such models are often based on permutation and combinatorial methods, so whether the number of infected individuals in the sample pool is known in advance will influence the design of the test array and decoding algorithms.

The study in this paper is mainly based on adaptive group testing model. Firstly, for the single-stage adaptive group testing, this paper gives an analytical expression of the optimal number of individuals in a group based on the probability of infection. Secondly, this paper presents a multistage grouping model with integer constraints, which is more general than binary splitting model. Based on the results of this paper, the optimal number of grouping members can be directly found according to the probability of infection, and the efficiency improvement of group testing can be estimated.

The rest of the paper is organized as follows: in the second section, the basic principles of group testing and notations are introduced; in the third section, the analysis and simulation results of two different models, the single-stage model and the multistage model, are presented; the conclusion is given in the fourth section.

II. BASIC PRINCIPLES

A. Characteristic Analysis of COVID-19

COVID-19 is a highly contagious virus. Accurate detection of infection is crucial to the control of the spread of this virus. The probability of infection varies in different regions and the number of infected people in a pool is usually unknown in advance. In view of the above characteristics, this paper adopts the following design principles:

First, an adaptive group testing model is used, which means the number of groups for the current stage of testing is determined by the results of the previous stage of testing.

Second, the criterion of zero-error detection is used. Because COVID-19 is extremely contagious and has a high fatality rate, missing one infected individual may have serious consequences. The zero-error criterion requires that the final testing result should be accurate. False detection result is not allowed.

Meanwhile, for the convenience of analysis, the following assumptions are taken in this paper: First, the i.i.d. prior is taken in this paper, i.e., each individual is infected independently with identical distribution. Second, the noiseless assumption is taken to simplify the problem, which means the PCR test results of the infected samples must be positive and those of the non-infected must be negative.

B. Expression of the Results of Group Testing

In this paper, the positive test result is denoted as 1 and the negative test result as 0. The principle of group testing is similar to the operation of Boolean inclusive OR, which means, as long as there is one or more than one positive samples in a group, the testing outcome of this group is 1. The testing outcome is 0 only when all the individuals in the group are negative.

C. Notations and Remarks

The following notations will be used in this paper:

- M: the number of samples in the testing pool.
- t: the time required for a single test.
- p: the probability of each individual being **not** infected.
- μ: the grouping coefficient, where $0 < \mu \leq 0.5$.
- M_i: the number of individuals corresponding to the i[th] group, where $M_i = \lceil M\mu^i \rceil, i = 1, \dots, D$.
- D: the maximum number of grouping stages, where $M\mu^D \leq 1$ and $D \geq \frac{log\frac{1}{M}}{log\mu}$.
- G_i: the number of groups in the i[th] stage of grouping.
- EIR: the group testing efficiency improvement ratio.

Remarks:

- The final testing time of the adaptive group testing is a random variable, and only the mean value of this random variable is considered in this paper.
- The group testing efficiency improvement ratio (EIR) is defined as the ratio of the grouping testing time to the one-by-one serially individual testing time.
- Testing for different groups is assumed to be processed serially, i.e., the parallel detection process is not considered in this paper.

III. Problem Models

In this section, the following two grouping models are studied:

- Model 1: Single-stage group testing, i.e., the sample pool is divided into testing groups only once, and then the members of groups with positive results are tested one by one to detect the infected individuals. Model 1 is further divided into two cases, with case (a) as an analytical mathematical model without integer constraints, and case (b) as a numerical model with integer constraints.
- Model 2: Multistage group testing, i.e., the sample pool is grouped in multiple stages until the number of members in the final suspicious group is 1, to ensure the accuracy of the test.

A. Model 1 (a) : Single-Stage Group Testing (without integer constraints)

The testing process of this model is as follows:

- In the grouping stage, the sample pool is grouped according to the grouping coefficient μ ($\mu \in (0,0.5]$), i.e., the number of individuals in each group is $M\mu$, where $M\mu \geq 1$. Then the number of groups is $G = \frac{1}{\mu}$, and the time for testing all groups after the first grouping is $\frac{1}{\mu}t$.
- After the grouping stage, the individuals in the positive groups are tested one by one. Since the probability of not being infected for an individual is p and the individuals are independent of each other, the mean value of the number of groups with positive testing results is $\frac{1}{\mu}(1-p^{M\mu})$. Therefore, the time for this step is $\frac{1}{\mu}(1-p^{M\mu})M\mu t = (1-p^{M\mu})Mt$.

If the group testing method is not adopted and each individual is tested one by one, the testing time is Mt. Compared with the one-by-one detection method, the testing efficiency improvement Ra of Model 1 (a) is

$$R_{1a} = \frac{(\frac{1}{\mu}+(1-p^{M\mu})M)t}{Mt} = \frac{1}{M\mu} + (1-p^{M\mu}) . \quad (1)$$

Let $x = M\mu$, then the optimization problem can be defined as

$$\min_{x \geq 1} f(x) = \frac{1}{x} - p^x + 1 . \quad (2)$$

By taking the first derivative of the objective function, we can find the solution of this optimization problem, where x satisfies

$$x^2 p^x \ln p = -1 . \quad (3)$$

When the actual probability of infection is small, i.e., $p \approx 1$, the approximate analytical solution can be obtained, which is

$$x_0 \approx \sqrt{-\frac{1}{lnp}} . \quad (4)$$

Fig. 1 illustrates the curves of (2) with $p = 0.9,\ 0.99,\ 0.999, 0.9999$, where the abscissa is the number of group members, and the ordinate is the efficiency improvement ratio. From Fig. 1, it can be found that, when $p = 0.9,\ 0.99,\ 0.999, 0.9999$, the corresponding optimal numbers of group member are 4,11,32 and 101.

According to (4), it is interesting that the optimal result is determined only by the individual probability of infection, i.e. $1 - p$. In Fig. 2, we compare the approximate results of (4) to the ideal optimal values of (2).

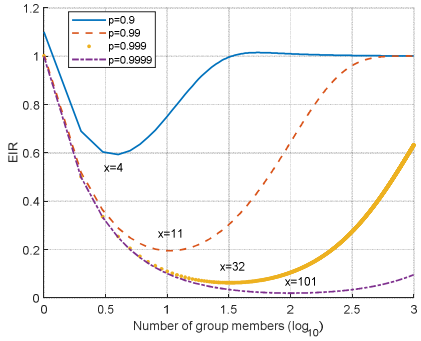

Fig. 1. EIR of Model 1(a) for different values of p

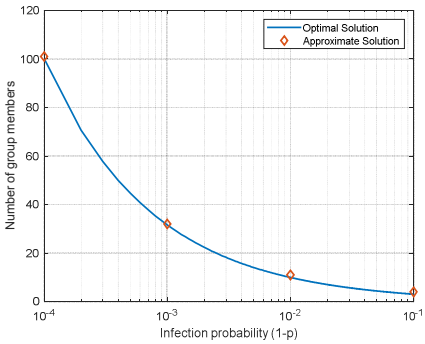

Fig. 2. The relation between optimal number of testing group member and the probability of infection. The blue line corresponds to numerical solution of (2), and the Diamonds correspond to the approximate solutions of (4).

It can be seen from Fig. 2 that the approximate value of (4) is nearly the same as the ideal optimal value of (2) when $p = 0.9,\ 0.99,\ 0.999, 0.9999$. Therefore, we believe that $x_0 \approx \sqrt{-\frac{1}{lnp}}$ can be used as a valid analytical solution for the case of Model 1(a).

B. *Model 1 (b) : Single-Stage Group Testing (with integer constraints)*

For Model 1 (b), which refers to the single-stage group testing method with integer constraints, the testing procedure is as follows:

- In the grouping stage, the sample pool is grouped according to the grouping coefficient μ, and the total number of groups in the grouping stage is G=$\left\lceil \frac{M}{\lceil M\mu \rceil} \right\rceil$, where the number of individuals in the 1st to the (G-1)th group is $\lceil M\mu \rceil$, and the number of individuals in the Gth group is $N = M - (G-1)\lceil M\mu \rceil,\ 0 < N \leq \lceil M\mu \rceil$. It is obvious that $\left\lceil \frac{M}{\lceil M\mu \rceil} \right\rceil$ is the number of tests for this stage of group testing.
- After the grouping stage, the individuals in the positive groups are tested one by one. The mean value of the number of groups with positive results is $\frac{1}{\mu}(1-p^{M\mu})$. Therefore, the time of this step is $\frac{1}{\mu}(1-p^{M\mu})M\mu t = (1-p^{M\mu})Mt$.

Therefore, the EIR of Model 1 (b) is

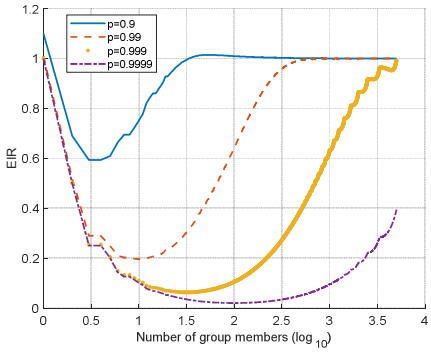

Fig. 3 Relationship between EIR and the number of group members (M=10000)

$$R_{1b} = \frac{\left(\left\lceil \frac{M}{\lceil M\mu \rceil} \right\rceil + \left(\left\lceil \frac{M}{\lceil M\mu \rceil} \right\rceil - 1\right)\left(1-p^{\lceil M\mu \rceil}\right)\lceil M\mu \rceil + (1-p^{N})N\right)t}{Mt} \quad (5)$$

where,

- $\left\lceil \frac{M}{\lceil M\mu \rceil} \right\rceil$ is the number of tests in the grouping stage.
- $\left(\left\lceil \frac{M}{\lceil M\mu \rceil} \right\rceil - 1\right)\left(1-p^{\lceil M\mu \rceil}\right)\lceil M\mu \rceil$ is the number of further tests on individuals of positive groups.
- $(1-p^{N})N$ is the number of further tests if the result of the last group with N individuals is positive.

In Fig. 3, the relationships between R_{1b} and the number of group members of different p values are illustrated, where M=10000. It is found that although the curves with integer constraints show some jitters, the main shapes are almost the same as those in Fig. 1, where the integer constraints are ignored.

By comparing the curves in Figs. 1, 2 and 3, we can reach the following conclusions:

1) Model 1 has similar optimization results in the cases of both (a) and (b).

2) $x_0 \approx \sqrt{-\frac{1}{lnp}}$ can be used as a valid approximation for the optimal number of grouping members for Model 1.

C. *Model 2: Multistage Group Testing*

Considering that the number of individuals in the sample pool is usually large, for simplicity, we assume that the number of individuals in the last group is the same as that in other groups. The multistage testing process is as follows:

1) In the first stage, the total number of individuals to be tested is M, the number of individuals in each group is $M_1 = \lceil M\mu \rceil$, the number of groups is $G_1 = \lceil \frac{M}{\lceil M\mu \rceil} \rceil$, the testing time is $t_1 = G_1 t$, and the mean value of the number of positive groups is $a_1 = G_1(1-p^{M1})$.

2) In the second stage, the positive groups are further grouped according to the grouping coefficient μ. The number of individuals in each group is $M_2 = \lceil M\mu^2 \rceil$, so there are $G_2 = \lceil \frac{M_1 a_1}{M_2} \rceil$ groups. The testing time is $t_2 = G_2 t$, and the mean value of the number of positive groups is $a_2 = G_2(1-p^{M_2})$.

3) The positive groups are further grouped in the subsequent stages until the final D^{th} stage;

4) The number of total stages is at most $D = \lceil \log_{\mu} \frac{1}{M} \rceil$, where the number of individuals in the group in the final stage should satisfy $\lceil M\mu^{D} \rceil = 1$, which will guarantee that the infected individuals can be accurately detected.

According to the above process, the total testing time is

$$T = t \sum_{n=1}^{D} G_n \ . \qquad (6)$$

Therefore, the EIR of Model 2 is

$$R_2 = \frac{\sum_{n=1}^{D} G_n t}{Mt} = \frac{\sum_{n=1}^{D} G_n}{M} \ . \qquad (7)$$

It is difficult to get analytical result for this model, so we give numerical simulation results directly in Figs. 4 and 5.

Figs. 4 and 5 show the relationship between EIR and grouping coefficient μ under different probabilities of infection with M as 1000 and 10000 respectively. From these two figures, we can find that the performance behaviors of grouping coefficients μ can be discussed in two different regions, which is summarized as follows:

1) In Fig. 4, where the total number of individuals is 1000, Region A corresponds to $\mu \leq 0.05$, while Region B corresponds to $\mu > 0.05$.

2) In Fig. 5, where the total number of individuals is 10000, Region A corresponds to $\mu \leq 0.01$, while Region B corresponds to $\mu > 0.01$.

In Region A, the shapes of the curves are similar to those in Figs. 1 and 3. At the same time, it is observed that in this region, if the probability of infection is high, there is a unique optimal grouping coefficient that corresponds to the optimal EIR. If the probability of infection is small, the testing efficiency will continuously improve with the increase of μ.

In Region B, in the case of high probability of infection, the testing efficiency ratio fluctuates dramatically. Sometimes EIR is even greater than 1, indicating that group testing is not a good choice for this situation. In the case of small probability of infection, group testing is still helpful to the improvement of detection efficiency.

Based on the above analysis, when the probability of infection is high, to ensure a robust performance, we should find the optimal grouping coefficient in Region A. When the probability of infection is small, we should use the maximum possible grouping coefficient in Region B.

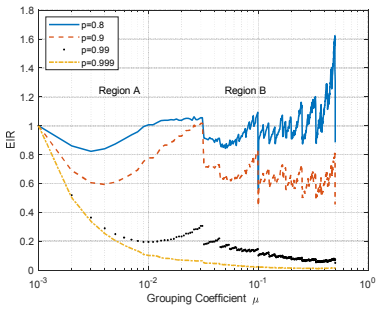

Fig. 4 EIR of multistage group testing method versus μ (M=1000)

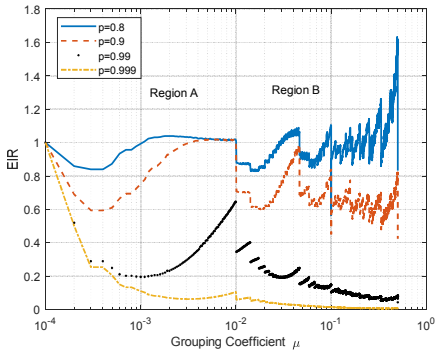

Fig. 5 EIR of multistage group testing method versus μ (M= 10,000)

IV. Conclusion

Aiming at increasing the testing efficiency for the infection detection of COVID-19, two adaptive group testing models are studied based on zero-error criterion. The models can help medical institutions to determine the critical parameters of single-stage and multistage group testing methods.

In real application, the uncertainties of the models proposed in this paper might occur due to the lack of a priori knowledge of the individual infection probability and the dilution of positive samples in a group which leads to the existence of false test outcomes. Future studies will focus on using observed infection probability to adaptively refine the models and taking the false detection probability as an additional parameter in the models.

Acknowledgment

As a high school student, I wanted to make my contribution in fighting against the COVID-19. Doing research on group testing was a challenging task for me. I couldn't have completed this paper without the help of professor Lizhong Zheng of MIT, who provided me with valuable resources related to group testing, and my statistics course instructor Qi He, who inspired my interest in this research topic.

References

[1] Lai, C.S.; Jia, Y.; Dong, Z.; Wang, D.; Tao, Y.; Lai, Q.H.; Wong, R.T.K.; Zobaa, A.F.; Wu, R.; Lai, L.L. A Review of Technical Standards for Smart Cities. Clean Technol. 2020, *2*, 290-310.

[2] R. Dorfman, The detection of defective members of large populations. The Annals of Mathematical Statistics 14 (4) : 436 {440, 1943, doi: 10.1214 / aoms / 1177731363

[3] Milton Sobel, Phyllis A. Groll. Group testing to eliminate all defectives in A binomial sample. Bell Labs Technical Journal, 38(5):1179{1252, 1959, doi: 10.1002/j.1538-7305.1959.

[4] M. Sobel, p. a. Groll, Binomial group - testing with an unknown proportion of defectives. Technometrics, 8 (4) : 631 {656, 1966, doi: 10.1080/00401706.1966.10490408

[5] C. h. Li, A sequential method for screening experimental variables. The Journal of the American Statistical Association, 57 (298) : 455 {477, 1962, doi: 10.1080/01621459.1962.10480672

[6] Christopher r. Bilder, Group Testing for Identification. Wiley StatsRef: Statistics Reference Online. 09 May 2019, https://doi.org/10.1002/9781118445112.stat08227

[7] A. Allemann, An efficient algorithm for Combinatorial Group testing. In H. Aydinian, F. Cicalese, and C. Deppe, editors, Information Theory, Combinatorics, and Search Theory: In Memory of Rudolf Ahlswede, pp. 569{596. Springer, 2013, Doi: 10.1007/978-3-642-36899-8_29.