Identifying patients from a large number of people by a small number of tests

Yaohua Xie. Yaohua.Xie@hotmail.com

Abstract For many diseases, patients need to be screened to determine whether they are ill or not. An approach, termed "Merged Testing for Recursive Groups (MTRG)", is proposed for screening patients from more people with less tests. It can be used for various diseases, and one of the direct application is the suppression of pandemic (e.g., coronavirus pneumonia). In this approach, people are divided into several groups, and each group has multiple people. For each group, the samples (e.g., respiratory secretions) of all the people are mixed to generate a merged sample. All the people can be excluded if the merged sample indicates no illness. If the merged sample indicates illness, the group will be divided into smaller groups, and then handled recursively in the same way. By doing so, negative groups are excluded gradually, and the group size becomes smaller and smaller. Finally, each group has only one person. This approach is especially suitable for the case when only small percentage of people are ill (e.g., infected).

In many cases, patients need to be identified by testing their samples. Each test requires costs such as time, people, apparatus and money. This is an especially severe problem for pandemic diseases (or infectious diseases). In human history, pandemic diseases have brought huge disasters to the world many times[1]. For example, coronavirus has caused large pandemic in different countries in the last two decades[2]. In order to prevent pandemic diseases, it is very important to identify infected patients as efficient as possible. However, the resource requirements for test may rise dramatically when a pandemic disease breaks out, especially when there are asymptomatic patients[3]. The lack of resources may delay the identity and therapy of large amount of patients, and thus put them in great danger. The idea of Group Testing has already discussed in some existing publications, mainly as a mathematical problem[4,5]. In this study, we further propose an approach termed "Merged Testing for Recursive Groups". It is optimized using computer algorithm, and can be used to improve the efficiency of patient screening. It is applicable for all the diseases that can be tested with "merged samples" (defined below), but may be especially suitable for the prevention and control of pandemic diseases. It may help save large amount of time, people, apparatus, money and other possible costs.

In this approach, people are divided into several groups, and each group has several people. For each group, the samples of all the people are mixed (maybe also proliferated and concentrated) to generate a merged sample. Then it is tested using a suitable method. All the people will be excluded if the merged sample is negative (e.g., not infected). If the merged sample is positive (e.g., infected), the group will be further divided into smaller groups, and then handled recursively in the same way as mentioned before. By doing so, negative groups (large or small) are excluded gradually. The group size becomes smaller and smaller during the whole procedure. Finally, each remained group has only one person. After these groups are tested, every person has been identified.

For example, there are 1000 people need to be screened, and there are probably only a few real patients. We can first divide the 1000 people into 10 groups, and there are 100 people in each group. Then, a sample (e.g., respiratory secretion) is taken from each person. After that, the samples of all the 100 people in each group are handled (e.g., mixed, proliferated and then concentrated) to get a merged sample. If the merged sample indicates no illness, all the 100 people in the group can be excluded. Otherwise, the original group is further divided into 10 groups, and there are 10 people in each group. Repeat the same procedures recursively until finally there is only 1 person in each unscreened groups. After these unscreened people are tested one by one, all the patients are identified out of the 1000 people.

The above procedure was verified with a computer program (coded with Python). It was run in Windows 10 -64bit system, and Matlab R2017b environment. In this program, we verify the performance of various number of people and patients. The number of people to be tested (denoted as P) varies from 10 to 100, while the number of patients (denoted as Q) varies from 1 to P. For a certain pair of P and Q, the whole group of people can be divided into M sub-groups recursively. Also, different values of M are tested in this program, which vary from 2 to P. For example, when P = 100 and Q = 5, the value of M is tested from M = 2 to M = 100. Different values of M may lead to different efficiency. By comparing the results for all these M values, we can find the most efficient M for a pair of P and Q. If a value of M leads to the least time of tests (denoted as “TotalChk_min”), we define an index “OptChecking” as the result of “TotalChk_min” divided by P. For example, different values of M can be used when there are totally 100 people (i.e., P = 100) and 1 patients (i.e., Q = 1). Assume that it requires only 13 times of tests (including both merged samples and individual samples) when M = 3. If 13 times is less than the tests required for any other values of M, then TotalChk_min = 13. Thereby, OptChecking = 13 / 100 = 0.13. Given the tests of a merged sample and an individual sample may require different resources (such as time, people, apparatus and money). So we also define another index “OptResource”, where different weighing factors are assigned for merged samples and individual samples. For example, the test number of merged sample is multiplied with 0.6, and that of individual sample is multiplied with 0.4. Or, multiply the former with a “zooming factor” (e.g., 1.5), and keep the latter unchanged. By doing so, we integrate the potential fact that testing a merged sample may require more resources than testing an individual sample. Since these weighing factors are unknown in this study, we will mainly discuss the first index (OptChecking).

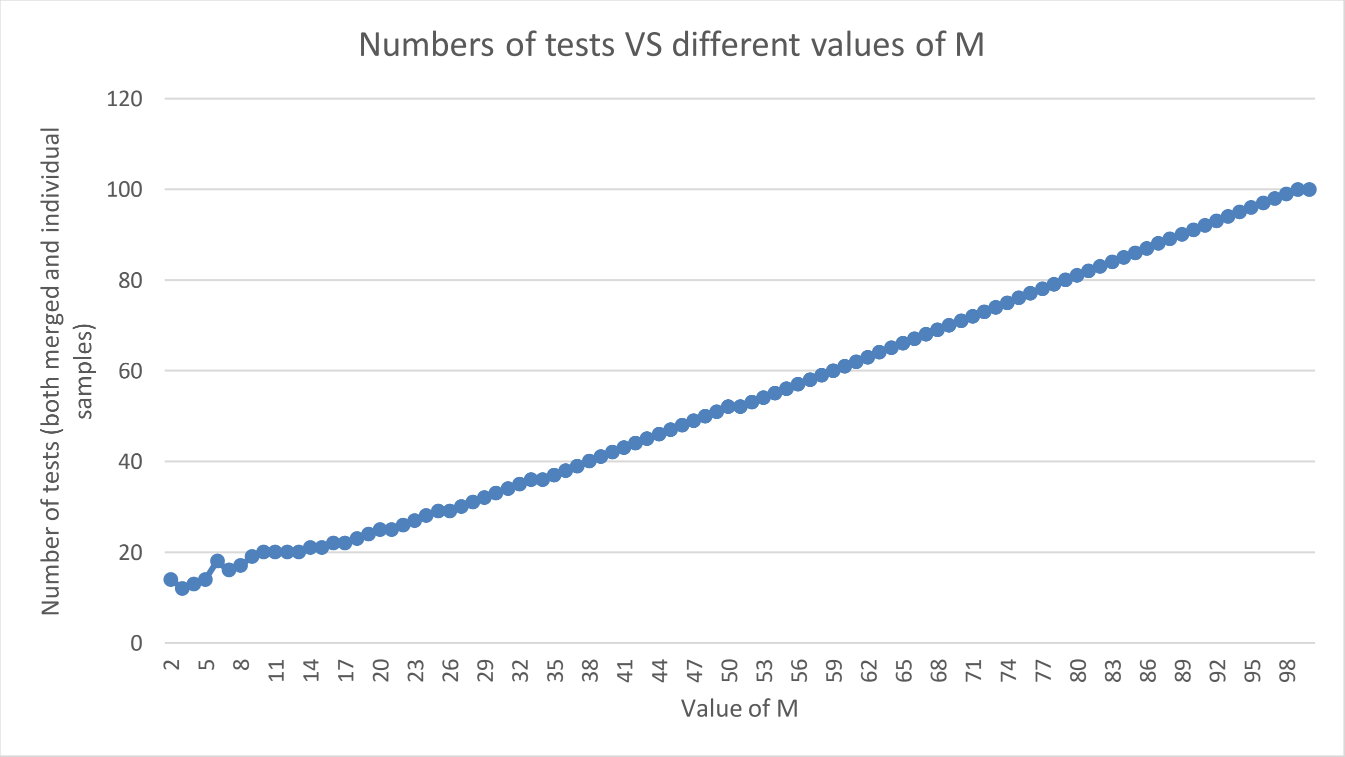

Fig.1. All the values of TotalChk for the pair (P = 100 and Q = 1)

First, let's take a look at a certain pair of P and Q, i.e., P = 100 and Q = 1. It means that there are totally 100 people, and 1 patient. Since the number of patient is unknown before identification, here we run the proposed approach with various values of M, i.e., from M = 2 to M = 100. For each value of M, we calculate the corresponding OptChecking according to the aforementioned steps. For example, OptChecking = 0.13 when M = 3. It means that we can identify all the patients (only one in this example) out the 100 people with as few as 13 tests. For the pair (P = 100 and Q = 1), there are totally 99 values of TotalChk, among which OptChecking (equals 0.13) is the optimal one. Each value of TotalChk corresponds to a value of M, from 2 to 100. All these values are shown in Figure 1.

It can be seen from Figure 1 that M = 3 is optimal for the pair (P = 100 and Q = 1). As mentioned before, the corresponding OptChecking is 0.13. But the optimal value of M and OptChecking may be different for other values of P and Q. Thereby, we further calculate the optimal OptChecking for each pair of P and Q, where P varies from 10 to 100 and Q varies from 1 to P. Please refer to the excel sheet in Supplementary Material 1 for the data. In the sheet, each row represents a certain value of P, and each column represents a certain value of Q. The value in each cell represents the optimal (least) number of tests for the corresponding P and Q. It can be seen from each row in the sheet, the values of OptChecking are usually smaller than 1 when Q is less than 30% of P. It means that the proposed approach usually improves efficiency when the patients are less than 30%. Then, the values of OptChecking are usually about or smaller than 20% of P when Q is less than 2% of P. In other words, the numbers of required tests are usually about or less than 20% of the total number of people when the patients are less than 2%. Furthermore, the required tests can be less than 15% of the total number of people when the patients are less than 1%. For example, if there are one patient among 100 people, it is possible to identify the patient with only 13 tests.

The program in the experiment can not only be used for verification. It can also be used to find relatively optimal "scheme of division", i.e., how many sub-groups should a large group be divided into. Based on the Matlab program, we further built a simple tool using Python3.7.0. After a user inputs the number of the total people and estimated patients, it gives a recommended (approximately optimal) number of sub-groups. Before testing, the number of patients is actually unknown among all the people. But better scheme will be found with the program if the number of patients estimated more accurately.

The proposed approach can be treated as the extension of the regular approach (i.e., testing one by one). When there is a large percentage of patients, the regular approach may be the most efficient way, and should be used in prior. But the proposed approach can be much more efficient when there is only a small percentage of patients. Sometimes we need to identify small ranges (but not every patient). In that case, the proposed procedure can be stopped halfway. One of the more simplified applications: perform a group-wise test for all the people in a certain range (such as room, building, family, institution, community or vehicle). If the result indicates no illness, all these people can be excluded. Otherwise, we will know these people should be given extra concern, even when the patients do not need to be identified exactly. Sometimes repeated and/or multipole types of tests may be required to guarantee the accurate of results. Even in that case, the resource requirements is still affordable for excluding a large group of people. This is especially suitable for the monitoring and prevention of pandemic.

It can be seen that the efficiency and Effectiveness-Cost Ratio tends to be higher when the percentage of patients is smaller. Such a feature is helpful for giving consideration to both the coverage and accuracy of isolation. First of all, a relatively large group of people can be selected to be isolated at the beginning. Such a group should be large enough to cover all the people that in the risk of infection. For example, that may include people in the same building, vehicle, community, village, etc., with any known patients. These people should be taken good care at home, hospital, restaurant or somewhere else, and must be isolated from uninfected people. After that, a smaller group is selected from the whole group. The people in the small group can be excluded if its merged testing indicates negative (without disease). More and more small groups are excluded from the rest people until finally every patients are identified.

Such a “cover-shrink-identify” procedure has some benefits. On the one hand, all the possible patients are isolated from other people as soon as possible. By doing so, the tendency of disease’s spread is prevented as comprehensively and efficiently as possible. On the other hand, the total resources required is relatively more affordable. It is highly possible that there are no patients in many small groups. Thereby, a lot of people can excluded (and then resume normal work and life) with only a little resources. In addition, this is a versatile approach that can be applied to various epidemic. It can also be implemented even when little knowledge are known about the disease (but the safety of implementation should be guaranteed). For example, it is estimated (e.g., according to R0) that about 2 people have been infected in a certain case. Using regular way, maybe 100 people is then isolated. But there is still relatively higher risk that not all the patients are covered, and that may cause the continued spread of the disease. Using the ”cover-shrink-identify” procedure, probably 1000 people can be isolated at first. Then, merged testing is performed on smaller groups (e.g., 100 people per group), and most of the groups (people) can be excluded soon. By doing so, the number of isolated people can be soon shrink to an affordable range.

In applications, it is very important to make sure that samples can be handled (e.g., merged, concentrated, tested) safely, and do not result in false or missed detection. This study focuses on the arrangement of testing procedures, but not medical/biological operations (testing samples, etc.). In fact, different diseases may require different ways of testing. But we hope suitable ways can be found for each diseases by experts in corresponding fields. Then those ways can be used in combination with the proposed approach to form a practical technique. Particularly, we hope a large number of people can be excluded with just one test or a few tests in this way.

References

1 Kaye, P. M. Infectious diseases of humans: Dynamics and control: Roy M. Anderson and Robert M. May, Oxford University Press, 1992. &#xa3;22.50 (viii + 757 pages) ISBN 0 19 854040 X. *Immunology Today* **14**, 616, doi:10.1016/0167-5699(93)90204-X (1993).

2 Zhou, P. *et al.* A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature*, doi:10.1038/s41586-020-2012-7 (2020).

3 Chan, J. F. W. *et al.* A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. *The Lancet* (2020).

4 Li, L.-a. On the Feasibility of Group Testing and the Calculation of Group Dividing. *Journal of Hubei Automotive Industries Institute* (2002).

5 Tu, B. Discussion on Grouped Testing of Blood Samples. *Journal of Chongqing Institute of Technology* (2005).