Variation of quantified infection rates of mixed samples to enhance rapid testing during an epidemic

Usama Kadri

School of Mathematics, Cardiff University, Cardiff CF24 4AG, UK

ABSTRACT

Rapid testing of appropriate samples from patients suspected for a disease during an epidemic, such as the current Coronavirus outbreak, is of a great importance for disease management and control. We propose a method to enhance processing large amounts of collected samples. The method is based on mixing samples in testing tubes (pooling) in a specific configuration, as opposed to testing single samples in each tube, and recognise infected samples from variations of the total infection rates in each tube. To illustrate the efficiency of the suggested method we carry out numerical tests for actual scenarios under various test conditions. Applying the proposed method allows testing many more patients using the same number of testing tubes, where all positives are identified with no false negatives, and no need for independent testing, and the effective testing time can be reduced drastically even when the uncertainty in the test is relatively high.

Keywords: rapid testing; quantified infections; Covid-19;

1 Introduction

The World Health Organization has declared the growing epidemic of novel coronavirus infectious disease (COVID-19) a global pandemic. The virus emerged in Wuhan, China, at the end of 2019, and as of August 19, 2020 over 22 million cases were identified in 213 countries and territories, with over three-quarters of a million deaths being reported (WHO, 2020). In most countries around the world the actual number of cases is believed to be larger than reported. The relatively low reported number is attributed to a number of factors, including but not limited to mismanagement of the epidemic at the political level (Ruiu, 2020), high ER visit costs (Konrad, 2020), and a lack of resources that limit the number of tests drastically. For example in the UK 25,000 tests were carried out in the period since January 2020 and up until March 11, 2020, which is equivalent to the number of tests carried out in South Korea in two and half days according to the World Health Organization. Thus, while in some countries there is a (front-end) problem of sample collection, in other countries the main concern is in processing the collected samples (back-end problem). Here we are concerned with the back-end problem, namely processing a large amount of collected samples.

Rapid testing of appropriate specimens from suspected patients during an epidemic in general, and COVID-19 in particular, is of great importance for clinical management and outbreak control. The current outbreak has evoked researchers and experts from various fields to reevaluate the feasibility of multi-sample pools, e.g. Kadri (2020*a,b*) and Yelin (2020), where samples from a number of patients are mixed together, as opposed to testing individual samples. In standard pooling (e.g. see Bilder & Tebbs (2012), Dorfman (1943) and references within), samples from a number of patients are mixed together in a single tube. If the test result is negative, all samples within must be negative (with some uncertainty). However, if the result is positive, it indicates that at least one patient is infected; in this case each patient is then tested individually. In this work, we propose an advanced testing method where samples from each patient are mixed in multiple tubes in a unique configuration, then variation of quantified infection rates in the tubes are employed in order to calculate all possible positives, i.e. without a need for a repetition. The first part of the method is by itself powerful when the percentage of infected patients is extremely low, as long as proper mixing is done – keeping in mind the dilution threshold (due to mixing) required for identifying the disease. However, as the percentage of the infected patients increases it becomes much more challenging to determine positives without performing new tests. To this end, an accurate quantitative approach can be employed to determine all positives without risking having false negatives. For this approach to be effective, an accurate method for quantification of PCR is required (Boulter, 2016). The proposed method takes into account the uncertainty in the test, which includes all technical, personal and conceptual issues, e.g. sensitivity of the test is one important factor. Even when the uncertainty increases (i.e. for less accurate tests), all positives are still obtained with no false negatives, though false positives start to arise as well. A qualitative description of the method followed by the mathematical model, and algorithms for samples distribution and test results analysis are presented in section 2. Numerical tests for actual scenarios under various test conditions are presented in section 3, which is followed by a discussion in section 4.
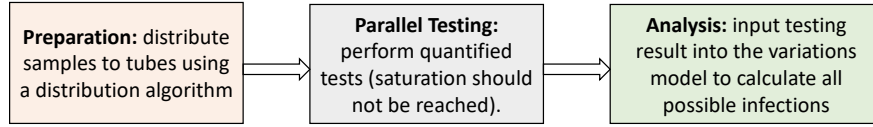
Figure 1. Processing collected samples. Main phases to process collected samples: distribution of samples, parallel testing, and analysis of test results.
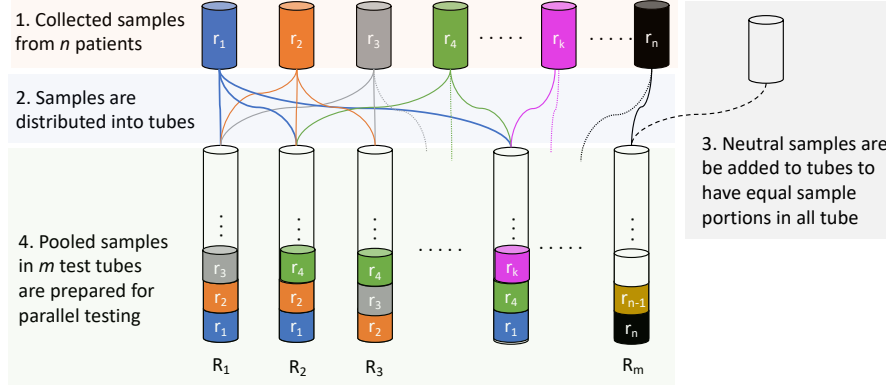
Figure 2. Tubes preparation. Distribution of collected samples into test tubes, in preparation for parallel pooling test.

2 Variation of quantified infection rates

Testing patients for infection can be carried out using various methods and often requires a number of stages, from collecting samples to providing test results. Here, attention is focused merely on processing collected samples that can be divided into three main phases: preparation, testing, and analysis (Fig. 1). In particular, the proposed method involves the preparation and analysis phases. For clarity purposes, we initially present a qualitative description of the proposed method, and later provide a detailed description of the mathematical models involved.

2.1 Phase 1: preparation

The preparation of samples prior to testing is a subtle task, which requires the following steps (Fig. 2): (a) test tubes are ordered and labelled; (b) all patients are allocated an equal number of test tubes that they share with other patients - thus each tube will comprise samples from different patients (pooling); (c) each patient is allocated a unique combination of tubes; (d) equal volumes from patients' samples are added to the allocated tubes; (e) all tubes are filled to the same level - if required a neutral sample (say water) is added. These steps can be automated using a robot where a *distribution algorithm* can be integrated (section 2.5).

2.2 Phase 2: parallel testing

The proposed method does not require a specific testing technique. If infection rates are relatively low, any standard testing technique that returns positive or negative results may suffice. However, quantitative results become essential as the infection rates rise. Thus, multiple testing techniques can be employed as the pandemic progresses. The role of the quantitative results will become clear below.

2.3 Phase 3: data analysis

Tubes that return negative results indicate that all samples within are negative and thus all associated patients test negative. At low infection rates, the combination of test tubes that return positive results will directly indicate the patient whose samples were distributed to these tubes. However, at higher infection rates a positive test result might be associated with more than one patient. In order to identify all infected patients there is a need for quantified results. Here the quantification technique used is very important. It is essential that we apply a technique that preserves the uniqueness of the infection rate of each patient. For example, if a PCR technique is used calculations must be made at identical time frames and before a threshold has been reached (alternatively one can make use of the time duration required for a threshold to be reached). Even though we only know the total infection rate in each tube, since each patient has a unique

(though unknown) infection rate we can calculate all possible combinations of infected patients that lead to the observed test results by solving a set of simple linear algebraic equations.

2.4 Mathematical model

Let n be the number of patients (sample size), m the size of the test tube set, and l the size of a subset of the tube test set. Each patient sample is distributed to a different configuration of l tubes. Thus, the relation between the maximum number of patients, the total number of tubes, and the number of tubes allocated to each patient is given by

$$n = \frac{m!}{l!(m-l)!}. \tag{1}$$

Relation (1) can be used to calculate any of the three parameters n, m, and l, given the other two. However, the current format is useful particularly as we anticipate l to be determined from dilution considerations (i.e. the maximum number of samples allowed together), whereas m could be a technical limitation, e.g. of resources. Thus, given such limitations, relation (1) provides the maximum possible number of patients that can be tested simultaneously. If maximum number of mixed samples allowed in a tube due to dilution is N_{max}, then any combination of l, m, and n should satisfy the condition

$$\frac{(m-1)!}{(l-1)!(m-l)!} \leq N_{max}. \tag{2}$$

The test results in each tube $j = 1..l$ can be described by

$$R_j = \sum_{i=1}^{n} r_{ij}\delta_i, \qquad (j = 1..l), \tag{3}$$

where r_{ij} is the contribution of the i-th patient to the infection rate of the j-th tube, and δ_i is the delta function accounting for the true statuses of the individuals, being zero or unity for healthy or infected patients, respectively. We require all samples to be of the same volume, and to avoid unbalanced dilution we also require that all tubes contain the identical volumes, i.e. neutral samples might need be added to some tubes. Thus, we can now construct a set of n algebraic equations

$$\begin{pmatrix} r_1 & r_1 & r_1 & \cdots & \cdots & 0 & 0 \\ r_2 & r_2 & 0 & r_2 & \cdots & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & \cdots & \cdots & r_k & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & \cdots & \cdots & r_n & r_n & r_n \end{pmatrix}^T \begin{pmatrix} \delta_1 \\ \delta_2 \\ \vdots \\ \delta_k \\ \vdots \\ \delta_n \end{pmatrix} = \begin{pmatrix} R_1 \\ R_2 \\ \vdots \\ R_k \\ \vdots \\ R_n, \end{pmatrix}, \tag{4}$$

or for simplicity we write $\mathbf{r}^T\boldsymbol{\delta} = \mathbf{R}$, where subscript T is the transpose operator. The solution vector $\mathbf{R}$ represents the test results, and thus known with some degree of test uncertainty, ΔR, which we shall account for. On the other hand, while the matrix $\mathbf{r}$ is unknown, the distributions of the samples in the tubes is our choice, and that is simply the matrix $\mathbf{r}$ with nonzero elements replaced by ones,

$$\mathbf{I}_{i,j} = \begin{pmatrix} 1 & 1 & 1 & \cdots & \cdots & 0 & 0 \\ 1 & 1 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & \cdots & \cdots & 1 & 1 & 1 \end{pmatrix}, \tag{5}$$

which is the distribution matrix $(n \times l)$ that shows how samples from each patient (rows) are added to the tubes (columns). Finally, the vector $\boldsymbol{\delta}$ contains the information we seek, on positive and negative patients. Thus, our objective is find $\boldsymbol{\delta}$.

2.4.1 Step 1: identifying immediate negatives.

The test results vector $\mathbf{R}$ may contain zero elements (i.e. tests that return negative). For each test tube j that returns zero, i.e. $\mathbf{R}_j = 0$, all patients i that have $\mathbf{I}_{i,j} = 1$ should test negative, otherwise the test tube would not have returned zero ($\mathbf{R}_j \neq 0$). Thus, at this phase we are able to identify q patients who test negative, with $0 \leq q \leq n$.

2.4.2 Step 2: identifying all positives.

We rewrite equation (4), excluding the negatives identified in the previous phase as $\tilde{\mathbf{r}}^T \tilde{\boldsymbol{\delta}} = \tilde{\mathbf{R}}$. Thus, all elements of $\tilde{\mathbf{R}}$ are now non zero. In order to identify all positives we seek all possible solutions for $\tilde{\boldsymbol{\delta}}$, using $(n-q)$ algebraic equations, for each possible solution $\tilde{\boldsymbol{\delta}}$. Now we carry out the following programme: (i) Consider all combinations that have a single positive, i.e. exactly $(n-q)$ combinations. Note that this can be the case only when elements in $\tilde{\mathbf{R}}$ are identical, within the test uncertainty ΔR. (ii) Consider all combinations with two positives, which is also easy to identify/exclude as it requires $\tilde{\mathbf{R}}$ to have either two or three different elements; two numbers that each corresponds to one of the infected patients, and possibly a third in case they are combined in a third tube. (iii) In general, we keep adding an additional possible positive (i.e. increase the number of 1's in $\tilde{\mathbf{R}}$), and seek for a solution.

2.4.3 Step 3: identifying more negatives.

All patients that were not identified to test positive in phase 2, nor negative in phase 1, have to be negative, as all positives have been identified. Thus all patients with $i(\tilde{\boldsymbol{\delta}}_k = 0)$ are identified as negative.

2.5 Distribution & analysis algorithms

In order to process the collected samples effectively, we introduce distribution and analysis algorithms, that can be employed to automate the process. A flow chart of the distribution and analysis algorithms are given in Fig. 3, which are also listed in the following steps:

- Each patient is allocated a binary number with l one's and $(m-l)$ zero's, e.g. (0 0 1 0 1 1 0 ...). The location of each digit of the binary number corresponds to a single tube (m tubes in total), e.g. the first digit from left (0) corresponds to the first tube, the second digit (0) corresponds to the second tube, the third digit (1) corresponds to the third tube, etc.

- Samples from each patient are added to all tubes that correspond to digits with number '1'. For example, samples from patient (0 0 0 1 1 1) are added to tubes number 4, 5, and 6 (from left).

- All m tubes are sent for testing. It is important that testing results are evaluated at a cycle where no saturation has been reached, so that the value of each positive test represents a summation of unique values for each positive sample in that tube. A quantitative result ($\mathbf{R}^T$) is obtained in the following form:
  0   34.0972   159.3790   125.2818   159.3790   0   0   159.3790   ...

- All tubes that return negative indicate that all patients that have samples in that tubes are not infected (negatives).

- Given the test results and uncertainty within, we then calculate all possible combination of solutions $\tilde{\boldsymbol{\delta}}$ that return $\tilde{\mathbf{R}}$, which when found, return all positives, and possibly additional negatives.

- In the unlikely scenario of no solution, a repeated test for all unidentified samples is required. This can be done by reshuffling samples, or splitting them between other test groups.

2.6 Numerical example

An analysis code has been developed in Matlab R18a. In order to run numerical examples, the user is required to input the number of patients n, number of tubes m, and the number of tubes related to each patient, l. The code randomly selects the infected patients, where a range of the infection percentage can be adjusted by the user, say between 5%-15% of patients. An example of the output for a single test is presented in Fig. 4. In this example we considered 70 patients and 8 tubes. The code randomly selects infected patients as an input of the test, then evaluates all positives and negatives using the programme above. In the example, the two infected patients were successfully identified.

3 Results

To gain more quantitative understanding of the proposed method, we performed numerical tests with actual scenarios under various conditions. We considered group sizes of 28, 56, and 70 patients that were tested using 8 tubes only (Fig. 5), and group sizes of 120, 210 and 252 patients using 10 tubes - see Fig. 6. Infected patients were selected randomly, with a percentage that ranged from 0.8 to 21.43. Each result point is an average of a hundred test repetitions, which is important when discussing uncertainties in the tests that were allowed to be between $\Delta R = 0.05\%...30\%$. Without loss of generality, each of the positively tested patients were allocated a random number between 0-220 nano grams (ng), that mimics possible reading, e.g. using a PCR technique, though any other range could be equally implemented. Note that
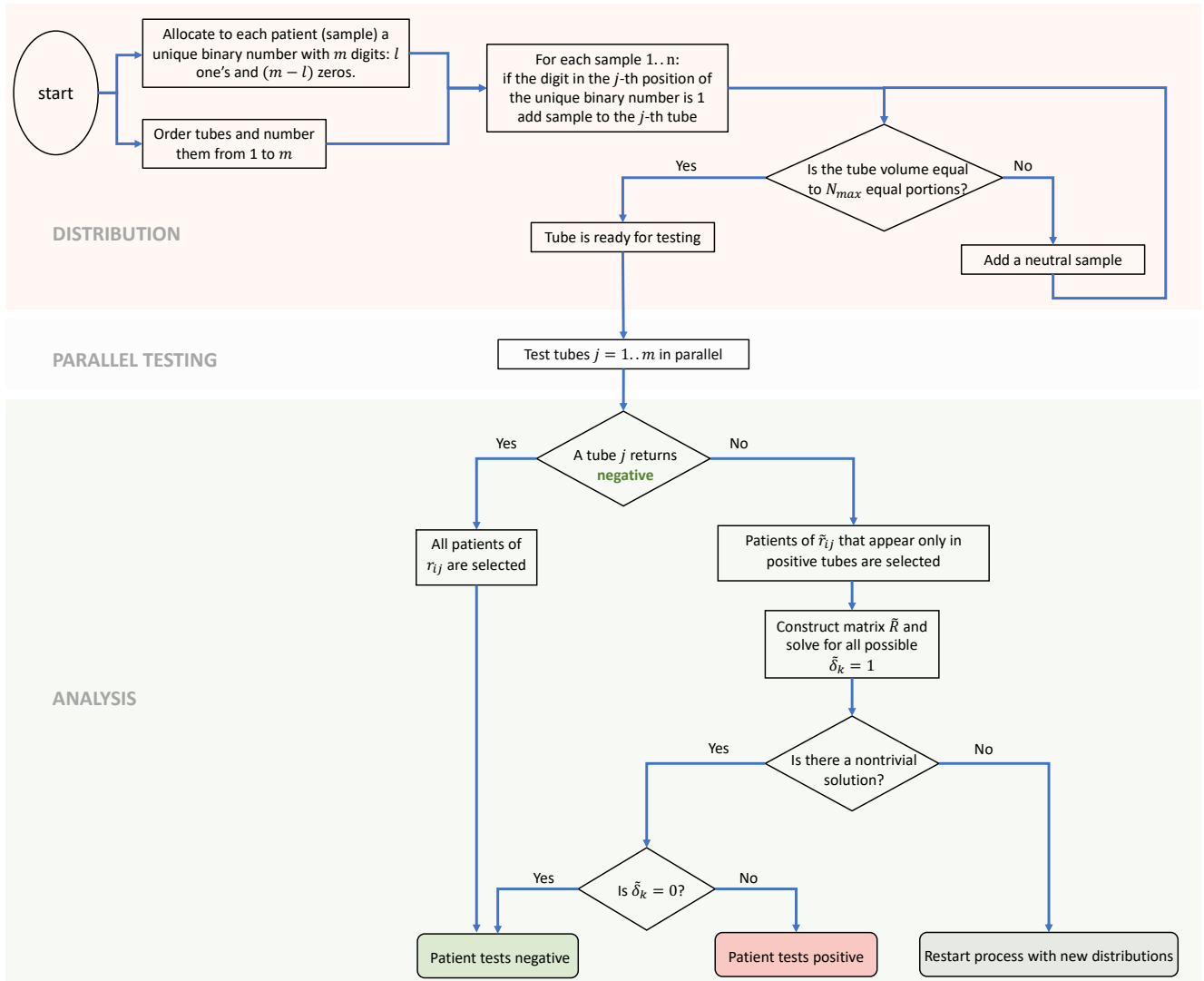
Figure 3. Distribution and analysis flow chart.

```
**************************************************************
**************************************************************
Number of patients:
   70

Input positive patients (random) #:
   42   58

Test Results:
    0   34.0972  159.3790  125.2818  159.3790      0      0  159.3790

**************************************************************
Immediate negatives are (100% certainty) #:
    1    2    3    4    5    6    7    8    9   10   11   12   13   14
   15   16   17   18   19   20   21   22   23   24   25   26   27   28
   29   30   31   32   33   34   35   37   38   40   41   43   44   45
   46   47   49   50   51   52   53   54   55   56   57   59   60   61
   62   63   64   65   66   67   68   69   70

Calculated negatives are (100% certainty) #:
   36   39   48

**************************************************************
Calculated positives are #:
   42   58

**************************************************************
Positives successfully identified %:
   100

Negatives successfully identified %:
   100
```

Figure 4. Numerical example. Matlab output example of a single test. With $n = 70$, $m = 8$, $l = 4$. There are two infected patients (randomly selected), that are identified successfully.
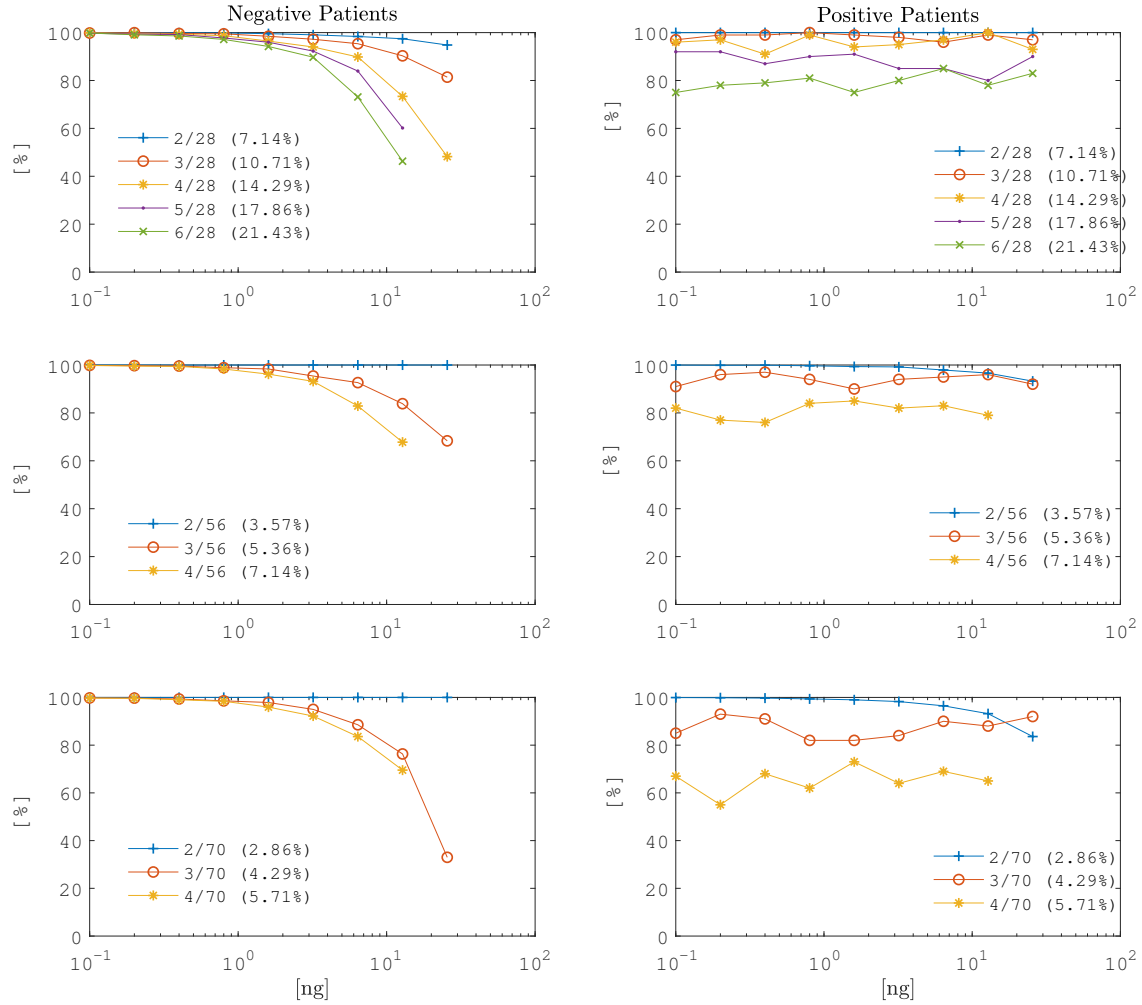
Figure 5. Fixed number of positives. Uncertainty analysis in tests using 8 tubes ($m = 8$), for fixed number of randomly selected positives, ranging from 2.86% to 21.43% of total patients. Each calculation point is an average of 100 repetitions. Left column: percentage of negative patients successfully identified. Right column: solutions found for positive patients (note that once a solution is found 100% of positives are identified). Rows top to bottom: $l = 3$, 4, and 5, with $n = 28$, 56, and 70, respectively.
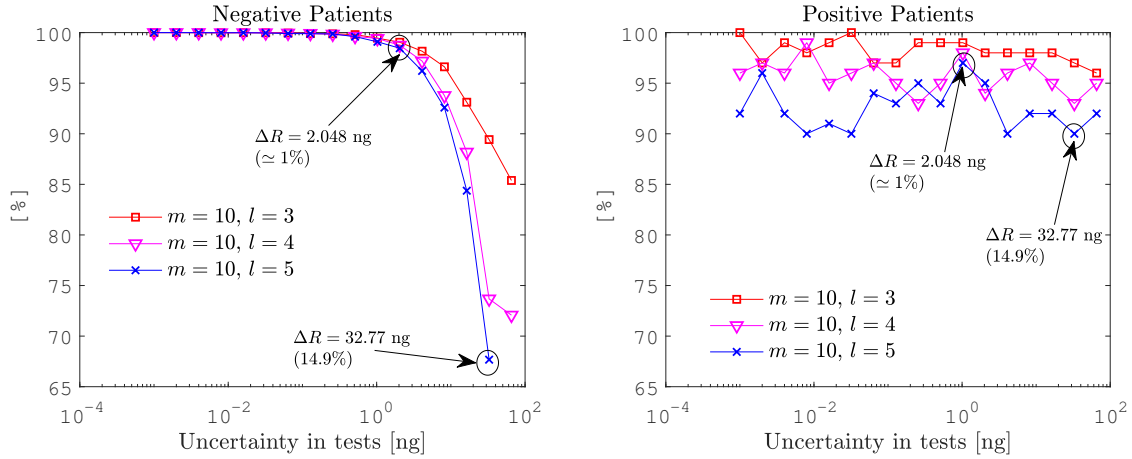
Figure 6. Random number of positives. Uncertainty analysis in tests for random number of positives (randomly selected), ranging from 0.8% to 5% of total patients. Using 10 tubes ($m = 10$), and $l = 3$, 4, and 5, corresponding to a total number of patients $n = 120$, 210, and 252, respectively. Left plot: percentage of negative patients successfully identified. Right column: solutions found for positive patients (note that once a solution is found 100% of positives are identified).

the chosen range is arbitrary and was introduced for illustration purposes, with 0 and 220 representing the lowest and highest possible readings, respectively.

Uncertainty in the test results, ΔR, depends on a number of factors among which are the accuracy and precision of the test method. The less sensitive the test is the larger ΔR becomes, e.g. an uncertainty of 10% is equivalent to 20 ng so that we are unable to distinguish between two readings with difference that is less than 20 ng. Therefore, as ΔR increases the size of the solved matrix increases where false solutions may appear, and thus false positive results are expected to rise. However, the elements that comprise the true size of the matrix (which contains all positive solutions), would forever be part of any increased matrix size, and thus regardless to how large the uncertainty is, false negatives are not expected, which is extremely important for disease management and control. If the uncertainty is very small, we are always able to obtain a single solution δ that determines accurately all positively and negatively tested patients. Unfortunately, current technology is associated with a degree of uncertainty that requires optimising the solution using other factors such as tube set size m, and subset l. The smaller the percentage of infected patients, the higher the percentage of negatives identified (Fig. 5). As the uncertainty increases, the size of subset l becomes important. Having a smaller l size results in a higher percentage of negatives, though the size of the group n becomes smaller (Fig. 6 and eq.(1)). When m and l increase the length of $\tilde{\delta}$ becomes larger, which requires solving for a greater combination of possible solutions. If the uncertainty in the tests is large, finding a solution and determining the positives is not always possible, though if solutions are found we always determine all positives, even if there are false positives. For example, for 25,200 patients, we performed 100 independent tests with 10 tubes each, and allowed each patient samples to be distributed to 5 tubes (see right subplot of Fig. 6, blue-cross curve), i.e. in each test there are $n = 252$ patients; even at high uncertainty of 32.77 ng (14.9% error), we were able to determine all positives in 90 tests out of 100 tests that were carried out. Thus, all positives from 22,680 patients were determined. In the remaining 8 tests, no solution was found and thus there is no risk for false negatives, in any of the tests. However, in this case we are only able to determine, on average 67.67% of all negatives in all 100 tests (see left subplot of Fig. 6, blue-cross curve). If a more accurate test is applicable, say with an uncertainty of 2.048 ng ($\simeq 1\%$ error), 99.6% of all negatives are identified in all 100 tests. Note that in Fig. 6 the percentage of positives is variable and selected randomly to mimic an actual scenario where we don't know the exact percentage of infected patients a priori.

4 Discussion

We propose an advanced quantified pooling method for determining infected samples. The proposed method does not require implementing the traditional two-stage Dorfman approach (Dorfman, 1943) with a second stage of individual testing. Though, until lab testings are carried out to increase confidence in the proposed method, a second stage of testing might be considered, which would also enhance the overall sensitivity and specificity of the results. Moreover, for clinical use, additional factors need to be considered, including sample types, viral load, prevalence, and inhibitor substances.

Although attention was focused here on enhancing lab tests for COVID-19, in particular, and epidemics in general, we

believe that the proposed method can be equally implemented in a variety of lab tests. In fact, once the solutions vector $\boldsymbol{\delta}$ is obtained one could structure back quantities in each of the infected samples, i.e. $\mathbf{r} = \mathbf{R}^T \boldsymbol{\delta}$. Such a quantitative method could be complementary to a wide range of environmental, healthcare and safety, and engineering applications, e.g. for testing contaminated water or food, spread of diseases in population and sewage, and proper concentration of substances in chemical products.

References

BILDER, C. R. & TEBBS, J. M. 2012 Pooled-testing procedures for screening high volume clinical specimens in heterogeneous populations. *Statistics in medicine* **31** (27), 3261–3268.

BOULTER, N., SUAREZ F.G. SCHIBECI S. ET AL. A 2016 A simple, accurate and universal method for quantification of pcr. *BMC Biotechnology* **16** (27).

DORFMAN, R. 1943 The detection of defective members of large populations. *Annals of Mathematical Statistics* **1943** (14), 436–440.

KADRI, U. 2020*a* Enhancing the number of lab tests with a "poisoned wine" approach. *Preprint at https://vixra.org/abs/2004.0198* .

KADRI, U. 2020*b* Variation of positiveness to enhance testing of specimens during an epidemic. *medRxiv* .

KONRAD, W. 2020 After battling covid-19, survivors may have to fight big medical bills. *CBS News May 15, 2020* .

RUIU, M. L. 2020 Mismanagement of covid-19: lessons learned from italy. *Journal of Risk Research* pp. 1–14.

WHO 2020 Coronavirus disease 2019 (covid-19) situation report-93. *World Health Organization* .

YELIN, I., ET AL. 2020 Evaluation of covid-19 rt-qpcr test in multi-sample pools. *Preprint at medRxiv 2020.03.26.20039438* .

Additional information

No human or human data or human samples were used in the study.

Competing interests

The author declares that he has no competing financial interests.