Fast and Efficient Data Science Techniques for COVID-19 Group Testing[*]

Varlam Kutateladze[†] and Ekaterina Seregina[‡]

October 20, 2020

Abstract

Researchers and public officials tend to agree that until a vaccine is developed, stopping SARS-CoV-2 transmission is the name of the game. Testing is the key to preventing the spread, especially by asymptomatic individuals. With testing capacity restricted, group testing is an appealing alternative for comprehensive screening and has recently received FDA emergency authorization. This technique tests pools of individual samples, thereby often requiring fewer testing resources while potentially providing multiple folds of speedup. We approach group testing from a data science perspective and offer two contributions. First, we provide an extensive empirical comparison of modern group testing techniques based on simulated data. Second, we propose a simple one-round method based on ℓ_1-norm sparse recovery, which outperforms current state-of-the-art approaches at certain disease prevalence rates.

JEL Classification: C020
Keywords: Pooled Testing; Compressed Sensing; Sparse Recovery; Lasso; Sensing Matrix; Coronavirus; SARS-CoV-2

[*] This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. Declarations of interest: none.

[†] Department of Economics, University of California, Riverside, CA 92521. Email: varlam.kutateladze@email.ucr.edu.

[‡] Department of Economics, University of California, Riverside, CA 92521. Email: ekaterina.seregina@email.ucr.edu.

1 Introduction

There is broad consensus among economists and epidemiologists that massive testing is one of they key ingredients to preventing the spread of Covid-19. However, large-scale testing is not realistic due to substantial restrictions in testing kits, chemical reagents, skilled personnel and time. Group testing, also known as pooled testing or specimen pooling, is an appealing alternative to individual testing that suggests to combine a set of individual specimens into a common pool, and test the pool rather than each individual sample. As long as the disease prevalence is not too large, testing pooled samples permits to considerably reduce the total number of tests required for diagnosing the population.

First experiments with pooling samples trace back to dilution studies in 1915 (Hughes-Oliver [2006]), which attempted to determine the presence or absence of organisms in a fluid based on pooled information. Researchers cultured samples of the fluid to let the bacteria, if they were present, grow, which served as a test. The results were then gathered across the samples to infer the bacterial density in the original fluid.

Many academics, however, attribute the invention of group testing to a Harvard economist Robert Dorfman, whose influential work (Dorfman [1943]) proposed a simple pooling method for weeding out syphilitic men called up for induction. Instead of analyzing individual blood samples for the presence or absence of ”syphilitic antigen”, it is suggested to examine pooled samples combining the individual blood sera into groups of five. If the corresponding men are healthy, the pooled test should be negative. On the other hand, if at least one of the patients is syphilitic, the pool will contain antigen, which the test is supposed to reveal. In that case, all associated patients need to be retested individually. Putting aside possible dilution concerns, it is clear that such strategy leads to savings of chemical reagents and higher overall testing capacity in a population with low disease prevalence. The idea is illustrated in Figure 1.

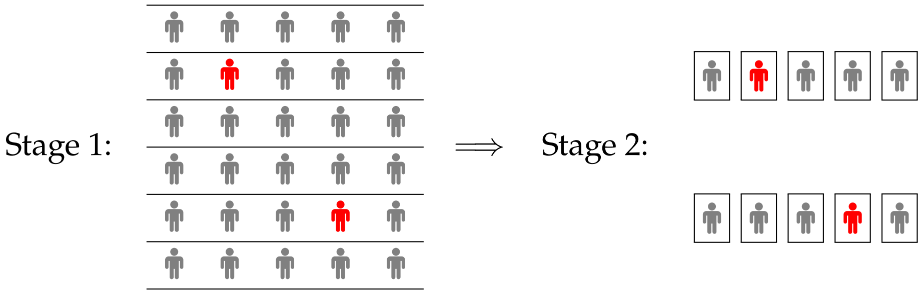


Figure 1: Dorfman pooling illustration. "User" icons represent individuals, red are infected and grey are healthy. In the first stage, all $N = 30$ specimen are pooled into $N/n = 6$ groups (rows) of $n = 5$, which are then tested. In the second stage, everyone in infected groups (rows two and four) is tested individually. As a result, it is possible to detect $k = 2$ positives with $6 + 10 < N$ tests.

Due do its simplicity, Dorfman's two-stage approach has found widespread use in medicine. Many of its properties are also readily available. Suppose we collect N individual samples and pool them into N/n groups of size n. Given the disease prevalence rate per hundred p (which is also the probability of a randomly selected individual being positive), the expected number of tests is

$$\mathbb{E}(T) = N/n + \underbrace{(1-(1-p)^n)}_{\mathbb{P}(\text{at least one positive})} n\frac{N}{n}. \tag{1}$$

The first term on the right-hand side corresponds to the number of tests in the first stage, the second term is n times the expectation of a random variable distributed as $Bi(N/n, (1-(1-p)^n))$, characterizing the number of positive groups in the second stage. Clearly, this method is more beneficial at lower prevalence rates p. For fixed $\mathbb{E}(T)$ and p, one could also optimize over the pool size n to get the largest possible coverage N,

$$n^* = \frac{2W(-\frac{1}{2}\sqrt{-\ln(1-p)})}{\ln(1-p)}, \tag{2}$$

where $W(\cdot)$ is the Lambert W function. Notice that this also is decreasing in p and, interestingly, is independent of $\mathbb{E}(T)$. The expected number of tests per person is approximately minimized when the group size is $n = 1/\sqrt{p}$ and hence the expected number of tests per person is $2\sqrt{p}$. Graphs illustrating the above relationships are provided in Appendix A.1.

Sterrett [1957] proposed a modification to Dorfman's second stage: instead of testing every individual, one should only do so until a positive sample is found, after which continue with group testing. In that case, if the prevalence is low, it is likely that the new sub-pool will be negative. This leads to an increase in savings of tests from Dorfman's 80% (compared to individual testing) to 86% at 1% rate. There have been other alternatives as well, e.g. Sobel and Groll [1959], halving techniques in Litvak et al. [1994] . These methods trade off higher efficiency with more complexity and longer wait times.

In group testing literature, Dorfman's approach and its modifications are classified as adaptive (or hierarchical), in a sense that they "adapt" to the results of preceding stages. An alternative approach, known as non-adaptive (or non-hierarchical), designs a single-stage experiment, results of which should allow to infer (often in a probabilistic manner) the original assignment of positives and negatives. This should generally, although not necessarily, come at the cost of having to run more tests overall since the sequential approach possesses more information. The distinctive feature of the non-adaptive approach is in assigning a single individual to multiple groups, i.e. groups are overlapping. How to design such assignment is a crucial question that is considered later. In a scenario such as the current SARS-CoV-2 pandemic, with the disease spreading fast and classical testing kits not showing results immediately, such non-adaptive approaches would have a clear advantage.

Furthermore, with multiple stages of testing, adaptive techniques also bear the risk of running out of tests before learning the outcomes. Given a limited number of tests, Dorfman's approach may require more test than are available. In contrast, single-round designs do not suffer from such indeterminacy.

Hence, we focus on such "fast" single-stage techniques. We consider several recent combinatorial and probabilistic algorithms. Importantly, we propose a simple method based on ℓ_1-norm sparse recovery, which outperforms the above algorithms.

Pooling strategies help resolve two kinds of problems, namely, estimation and classification. The first seeks to estimate the prevalence of positive individuals in a population. The second, which may or may not rely on the information on estimated prevalence, aims to identify the infected individuals. The performance is typically gauged by the expected number of tests required for a given specificity or sensitivity, or vice versa, predictive accuracy for a given number of tests. We focus on classification.

Notation. For a vector $\mathrm{v} \in \mathbb{R}^d$, we write its i-th element as v_i. The corresponding ℓ_p norm is $\|\mathrm{v}\|_p = \left(\sum_{i=1}^d |v_i|^p\right)^{1/p}$, which is a norm for $1 \leq p \leq \infty$. For a matrix $A \in \mathbb{R}^{m \times d}$, we write its (i,j)-th entry as $\{A\}_{ij} = a_{ij}$ and denote its i-th row (transposed) and j-th column as column vectors $A_{i\cdot}$ and $A_{\cdot j}$ respectively. Its singular values are $\sigma_1(A) \geq \sigma_2(A) \geq \ldots \geq \sigma_q(A)$, where $q = \min(m, d)$. The spectral norm is $\|A\|_2 = \max\limits_{\mathrm{v} \neq 0} \frac{\|A\mathrm{v}\|_2}{\|\mathrm{v}\|_2} = \sigma_1(A)$. The ℓ_1 norm is $\|A\|_1 = \max\limits_{1 \leq j \leq d} \sum_{i=1}^m |u_{ij}|$. Finally, for a sequence of random variables $\{X_n\}_{n=1}^\infty$ and a sequence of real nonnegative numbers $\{a_n\}_{n=1}^\infty$, denote $X_n = O_{\mathbb{P}}(a_n)$ if $\forall \epsilon > 0, \exists M, N > 0$ such that $\forall n > N,\ \mathbb{P}(|X_n/a_n| \geq M) < \epsilon$; and denote $X_n = o_{\mathbb{P}}(a_n)$ if $\forall \epsilon > 0,\ \lim\limits_{n \to \infty} \mathbb{P}(|X_n/a_n| \geq \epsilon) = 0$.

1.1 RT-qPCR test

The two main ways of determining whether an individual has a SARS-CoV-2 virus are (1) to check for the presence of antibodies to the virus, (2) to check for the presence of the virus RNA itself. The former, although capable of uncovering whether a recovered individual had the virus in the past, is less widespread; the latter includes so-called reverse transcription quantitative polymerase chain reaction (RT-qPCR), is the gold standard for Covid-19 testing recommended by the Centers for Disease Control and Prevention. Though popular, massive individual PCR testing is not possible due serious constraints in equipment, chemical reagents and skilled personnel. The resulting readouts of RT-qPCR are of key interest to this study so we very briefly describe the testing process.

To perform this test, nasopharyngeal swabs from subjects are collected and diluted in a fluid medium. The first stage, reverse transcription, then transforms the virus RNA to complementary DNA (cDNA). This eventually allows to start the next stage, polymerase chain reaction, which aims to exponentially amplify the viral cDNA molecules through the process that involves up to nearly 40 cycles of heating and cooling. To trace this increase, the virus-specific sequences are marked by fluorescent. The testing machine measures the amount of fluorescent signal in real time and displays it as a function of cycles. This information we are interested it is when (if at all) the fluorescence exceeds the critical level associated with a positive subject. This is given by a cycle threshold (C_t), the number of cycles completed before crossing the threshold. The subject is then declared positive if the threshold is passed before about 40 cycles. The C_t indicator is (negatively) correlated with the original viral load, with larger initial viral loads leading to sooner crossing of the threshold and thus shorter cycle thresholds.

The entire process takes up to approximately 4 hours.

The information on cycle thresholds of pooled samples is the key input to group testing algorithms. Rather surprisingly, many known group testing algorithms do not take this quantitative information into account and instead work with degenerate binary transformations. The proposed algorithm described in this study is capable of not only incorporating the quantitative information, but also producing corresponding quantitative predictions measuring the original viral load in each individual.

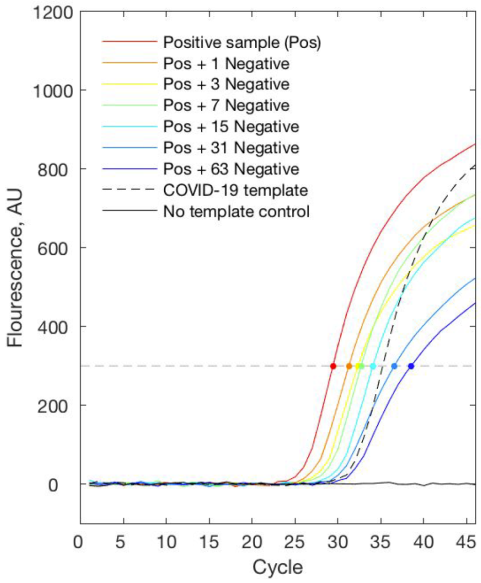


Figure 2: Source: Yelin et al. [2020]

1.2 Biomedical considerations

One of the major concerns with any type of pooling approaches is that of dilution. Fortunately, there is growing evidence that pooling of SARS-CoV-2 with negative samples does not lead to substantial dilution of the virus DNA. In a recent study (Yelin et al. [2020]), Israeli researchers discovered that it is possible to detect a single positive SARS-CoV-2 sample in pools of up to 64 samples with reasonably high accuracy. Other investigations ((Abdalhamid et al. [2020], Hogan et al. [2020], Mutesa et al. [2020])

tend to agree with such claims.

Pooled testing for SARS-CoV-2 has been conducted in a number of countries, including US (Stanford Health Care Clinical Virology Laboratory and Nebraska's Public Health Laboratory), Germany (University Hospital Frankfurt at Goethe University), China and Israel (Rambam Health Care Campus).

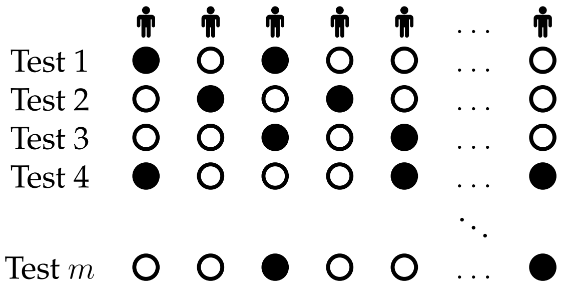
Group testing has been used for detecting the HIV (Emmanuel et al. [1988]); in fact, it is now a routine option in blood screening. Pooling not only decreases the cost but also the probability of making an error in low disease prevalence populations. Pooling has also been deployed against malaria (Taylor et al. [2010]), influenza (Van et al. [2012]) and a few other diseases.

2 Non-adaptive group testing

2.1 Traditional group testing

We first briefly review some of the recent non-adaptive algorithms introduced in the literature (Chan et al. [2011]). As all are one-round approaches, it is required to construct an $m \times N$ pooling matrix A, which would assign each of the N individuals to the appropriate group. As opposed to adaptive testing, we may have the same subject sample split among several groups. Figure 3 illustrates the idea. The corresponding pooling matrix has its (i, j)th entry equal to one if an ith individual is assigned to group j, $i \leq N, j \leq m$, and zero otherwise. One would typically also normalizes the matrix, but this does not raise any substantial challenges so we omit this issue for now.

Once the pooling matrix is specified, one then observes the results of m pooled tests via $m \times 1$ vector y and the goal is to identify which of the $N \gg m$ individuals are truly



	🚹	🚹	🚹	🚹	🚹	...	🚹
Test 1	●	○	●	○	○	...	○
Test 2	○	●	○	●	○	...	○
Test 3	○	○	●	○	●	...	○
Test 4	●	○	○	○	●	...	●
						⋱	
Test m	○	○	●	○	○	...	●

Figure 3: Illustration of a pooling matrix assigning N individuals to m pooled tests. A black circle indicates that the corresponding individual (column) has been assigned to the given test (row). If an individual is assigned to several tests, his sample is split accordingly.

positive. We briefly discuss novel algorithms and provide an intuitive explanations behind their principles. Our code at the end contains exact algorithmic formulations.

Combinatorial Basis Pursuit (CBP) is a simple algorithm that is based on the following idea: declare all individuals that were included in negative pools as negative (since if at least one sample was positive, the entire group would have been positive), and declare the remaining individuals as positives. Since this strategy would identify healthy individuals for certain, it will not produce any false negatives.

Instead of looking at each test (row), one may instead try to decode the matrix "column"-wise, i.e. by going over the individuals. This is what the combinatorial orthogonal matching pursuit (COMP) algorithm does: If all tests an individual participated in turn out to be positive, then the individual is considered to be infected; otherwise he is negative. This deciphering method never produces false negatives, only false positives. A false positive would only occur if a healthy individual happened to always participate in tests that contained at least one infected sample. Of course the probability of this happening decreases with m.

Definite defectives (DD) algorithm starts with COMP to leverage its ability of identifying true negatives. Once COMP is over, DD switches to "row"-wise search by looking at positive tests and seeks to determine individuals that are "definite defectives" (positives). All remaining subjects are declared negative. This reversal in the algorithm leads to DD producing only false negatives, and no false positives. This leads to greater accuracy in sparse settings: because there are a lot more healthy subjects, one should by default assume that an individual is not infected, all else equal.

Finally, sequential COMP (SCOMP) further attempts to improve DD by modifying its last step of labeling remaining subjects as negatives. The key is to observe that if the current set of individuals that are declared positive cannot explain all of the positive pooled tests, one can do better by sequentially declaring potential positives as positives until the set of positives accounts for all positive tests. From a list of potential candidates, the algorithm picks the one that would account for the largest number of unexplained tests. SCOMP has been shown to perform close to the information-theoretic bound.

3 Problem formulation

We now turn to our algorithm that leverages recent advancements in the field of compressed sensing in engineering and statistics literature (Candes et al. [2006], Donoho [2006]). Our main goal is to efficiently infer x, the N-dimensional sparse vector of individual viral loads, from m available group test results stacked in a vector y. In general, we have $\mathrm{y} = g(\mathrm{Ax}) + \epsilon$, however we will restrict our attention to the simplest case when $\mathrm{y} = \mathrm{Ax} + \epsilon$. Hence, $\mathrm{A} \in \mathbb{R}^{m \times N}$ represents a set of linear measurements on the variable of interest x. This formulation has a crucial difference with a regression type of problems: in our setting one gets to choose how to design the pooling matrix A, while in regression problems A is pre-determined by the data. Thus, there are two steps to solving such problems.

The first step is to conveniently encode the sparse signal, by designing a proper pooling matrix A. This matrix provides the assignments for each individual specimen to the corresponding groups and must satisfy certain desirable conditions pertaining to the group testing problem.

The second step attempts to decipher the first step with fewest test measurements. For a large-dimensional vector x finding the corresponding sparsest vector that would be consistent with m pooled observations is an NP-hard problem that would not be possible to solve. However, recent advancements in engineering allow to transition this problem to convex domain where exact decoding is feasible with high probability.

While x is generally a vector of quantitative measurements of all individual, one can equally think of it as a sparse vector, i.e. with most entries equal to 0 (associated with healthy individuals) and very few 1's, without loss of generality. What matters is that the vector needs to be sparse in some transformed coordinate system.

3.1 Pooling matrix design

We first focus on the design of a pooling (also known as sensing or measurement) matrix A. Due to the nature of our primary application, we only consider sparse pooling matrices with few nonzero elements. When properly formed, this should ensure there is not too much dilution: a single sample is not split into too many subsamples and any one group sample does not contain too many specimen. The simplest approach would be to generate a random Bernoulli matrix with entries, which together with a normally distributed random matrix, has been shown to satisfy desirable properties, mainly the null space condition (NSC) and the restricted isometry property (RIP) discussed later, that guarantee a precise recovery of the original vector of interest with high probability.

One of the recent studies (Yi et al. [2020]) examined Bernoulli design and documented its superior performance. Another study (Ghosh et al. [2020]) proposed to use deterministic Kirkman triple designs, but these are have are not very flexible as the matrix dimension ratio is given and the theoretical properties are not well-established.

We instead use a pooling matrix that was proposed in a different branch of group testing literature. Known as a constant column weight design (Aldridge et al. [2016]), it has shown to outperform simple Bernoulli matrices in terms of its encoding capabilities. The initial approach outlined in Aldridge et al. [2016] constructs A by inserting up to an L of ones into each column. Concretely, L indices of each column are sampled with replacement and ones are inserted in the unique positions. This complication seems to be necessary for their proofs, however the real performance does not depend on whether one bootstraps or simply permutes a fixed number of ones. Hence, we focus on a simpler, permutation version. One example of such matrix with $N = 6$, $L = 2$ and $m = 4$ is

$$\begin{bmatrix} 1 & 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 \end{bmatrix}$$

As discussed earlier, the ith individual is assigned to group j only if the (i, j)th entry is 1. This design avoids too much dilution as long as $L < m \ll N$, and outperforms Bernoulli design. Importantly, we prove that this design is RIP which has immediate theoretical implications, which are discussed in the next section.

Theorem 1. *A random matrix* $\mathrm{A} \in \mathbb{R}^{m \times N}$ *of constant column weight design with* L *ones in each column satisfies the restricted isometry property with high probability, specifically there* *exists* $\delta \in (0,1)$ *such that*

$$(1-\delta)\left\|\mathrm{x}\right\|_2^2 \leq \left\|\mathrm{Ax}\right\|_2^2 \leq (1+\delta)\left\|\mathrm{x}\right\|_2^2$$

holds with high probability for any $\mathrm{x} \in \mathbb{R}^N$ *and* $0 < L < m$.

Proof. See appendix A.2. □

3.2 ℓ_1 sparse recovery

Once m measurements in y are formed, one can employ several strategies for decoding the original signal. A direct, brute force approach to tackle the problem would be to find the sparsest vector of viral loads x that is consistent with the linear measurements, that is

$$\min_{\mathrm{x}\in\mathbb{R}^N} \quad \left\|\mathrm{x}\right\|_0 \quad \text{s.t.} \quad \left\|\mathrm{Ax}-\mathrm{y}\right\|_2 \leq \epsilon.$$

Unfortunately, this problem is NP-hard as its solution requires an exhaustively search over all possible combinations in x, although this may still be feasible for low-dimensional problems. Luckily, a convenient convex relaxation is available, which has been proven to yield accurate solutions as long as the sensing matrix A satisfies RIP (Candes et al. [2006],Donoho [2006]). This is a sufficient condition, which we show holds under high probability. In practice one could generate such random matrix and attempt to verify whether RIP holds for a given deterministic matrix, although this by itself is also NP-hard (Bandeira et al. [2013]). The corresponding convex alternative is

$$\min_{\mathrm{x}\in\mathbb{R}^N} \quad \left\|\mathrm{x}\right\|_1 \quad \text{s.t.} \quad \left\|\mathrm{Ax}-\mathrm{y}\right\|_2 \leq \epsilon.$$

This is known as Basis Pursuit Denoising (Shaobing Chen and Donoho [1994]), although many statisticians are more familiar with its equivalent formulation, LASSO,

$$\min_{\mathrm{x}\in\mathbb{R}^N} \quad \|\mathrm{Ax}-\mathrm{y}\|_2^2 + \lambda \|\mathrm{x}\|_1$$

The two problems are identical for certain choices of ϵ and λ. We simply add a non-negativity constraint $\mathrm{x} \geq 0$ which improves the empirical performance.

This type of ℓ_1-norm recovery uses $m = O(k \log(N))$ tests while standard group testing algorithms require $m = O(k^2 \log(N))$ test. Another advantage of this approach is in the ability to handle real-value quantitative readouts; many group testing algorithms are only capable of dealing with binary measurements. Furthermore, the output is also a real-valued number estimating individual's viral load.

4 Application

This section compares the performance of the above algorithms in simple numerical experiments with no noise. For clarity of exposition, the vector of interest x is generated to be a binary 0-1 vector instead of a real-numbered qPCR-like measurements. More general treatments can be found in the code provided at the end.

We consider a case with $N = 100$ specimens where there are $k = 2$ true positive cases. This is a conservative estimate in a sense that this share of positives is about twice the share of active cases in the United States as of October 19, 2020 (Worldometer [2020]). In Appendix A.3 we additionally report less favorable (from group testing perspective) cases with $k = 4$ and $k = 6$.

To illustrate, we generate a 100-dimensional binary vector with 2 ones and the sensing matrices as described above to obtain $m = 20$ linear measurements. We then apply the decoding algorithms to try to infer the original binary vector (both the number of positive k and their positions) with only 20 measurements. Figure 4 demostrates a particular realization where only the proposed algorithm, denoted as SR (for sparse recovery), is capable of correctly identifying the positions. Other algorithms produce either false positives, false negatives or both.

Specifically, to obtain SR estimates we first solve

$$\tilde{\mathrm{x}} = \underset{\mathrm{x}\in\mathbb{R}^N}{\arg\min} \quad \|\mathrm{A}\mathrm{x} - \mathrm{y}\|_2^2 + \lambda \|\mathrm{x}\|_1, \quad \mathrm{x} \geq 0, \tag{3}$$

which generally would not produce 0-1 estimates. Hence we simply round the estimates at the threshold value of $\tau = .5$, that is

$$\hat{x}_i = \begin{cases} 1 & \text{if } \tilde{x}_i \geq \tau \\ 0 & \text{if } \tilde{x}_i < \tau. \end{cases}$$

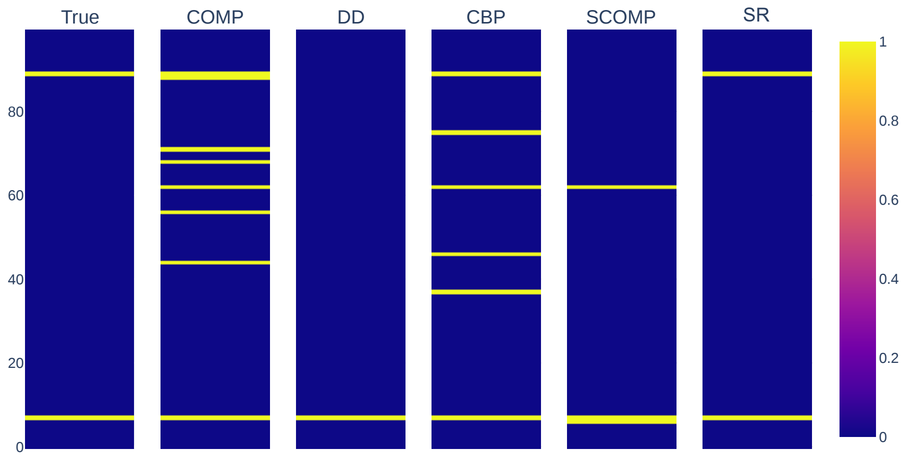


Figure 4: Identification of negative and positive positions. Only the proposed (SR) algorithm is able to successfully determine which samples correspond to positive/negative specimens.

Next, we repeat this process for 1000 iterations across different group sizes m and report the average root mean square error (RMSE) plotted against the number of test measurements m in Figure 5. RMSE is defined as $\frac{\|\mathrm{x}-\hat{\mathrm{x}}\|_2}{\|\hat{\mathrm{x}}\|_2}$, where x is a true binary vector and $\hat{\mathrm{x}}$ is one of the estimates. We still keep $N = 100$ and $k = 2$, but Appendix A.3 reports cases for $k = 4$ and $k = 6$.

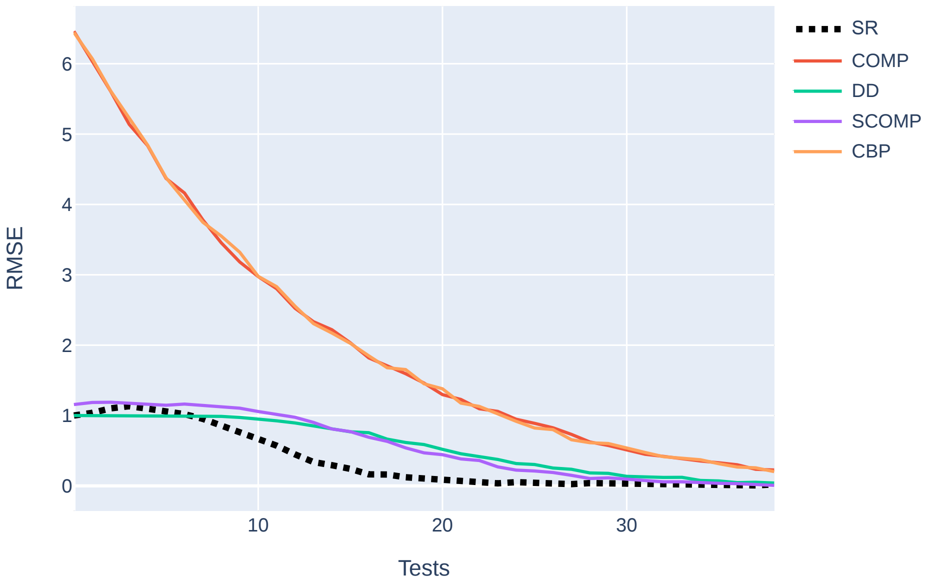


Figure 5: RMSE of each approach as a function of test measurements m. SR outperforms standard non-adaptive group testing algorithms.

As can be seen, the proposed method makes approximately the same error with $m = 20$ tests as the best alternative (SCOMP) with $m = 30$ tests. For comparison, Dorfman approach would require approximately $m = 30$ tests and two testing stages.

However, RMSE does not tell the whole story. One is also interested in sensitivity (or true positive rate) and specificity (or true negative rate). These are defined as the ratio of identified positives to all true positives and the ratio of identified negatives to all true negatives respectively. These are reported in Figures 6 and 7.

Notice that CBP and COMP report perfect sensitivity. This is a sanity check since these algorithms should not produce any false negatives. Among the other algorithm

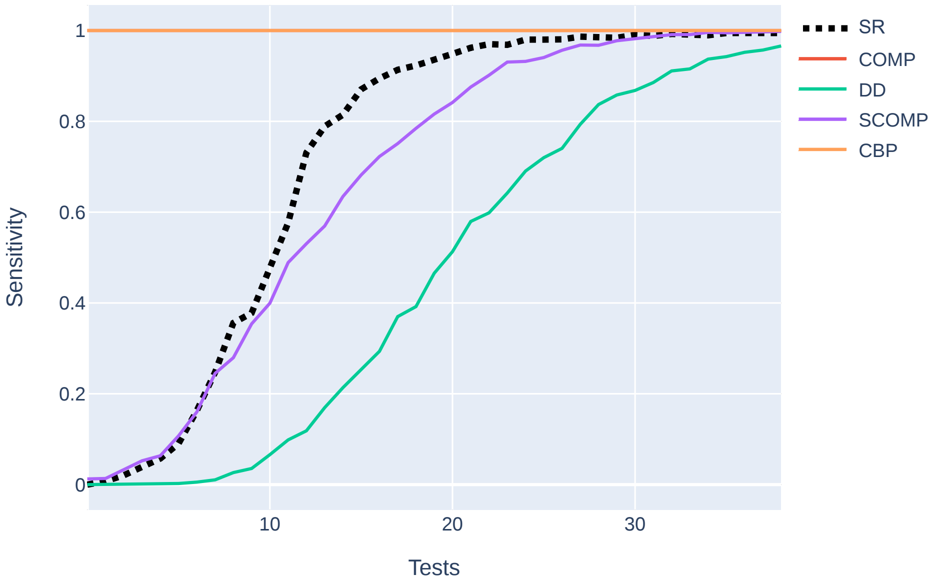


Figure 6: Sensitivity of each approach as a function of test measurements m

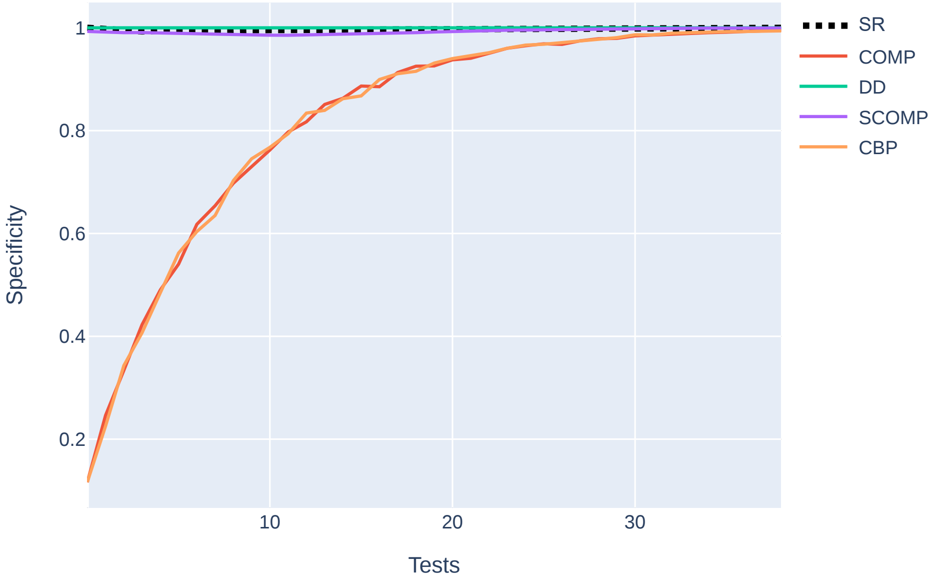


Figure 7: Specificity of each approach as a function of test measurements m

SR is again a clear winner. Naturally, the relationship between the two groups reverses for specificity: we have SR, COMP and DD achieving ideal (or almost ideal) specificity with a minimum number of tests, while COMP and CBP slowly catch up.

Additionally, we plot the receiver operating characteristic curve (ROC) and area under the curve (AUC) for SR in Figure 8, where we keep the same parameter values $N = 100$, $m = 20$, $k = 2$. This curve traces the true and false positive rates for different values of the threshold τ. The figure is indicative of a strong classification ability of the proposed method.

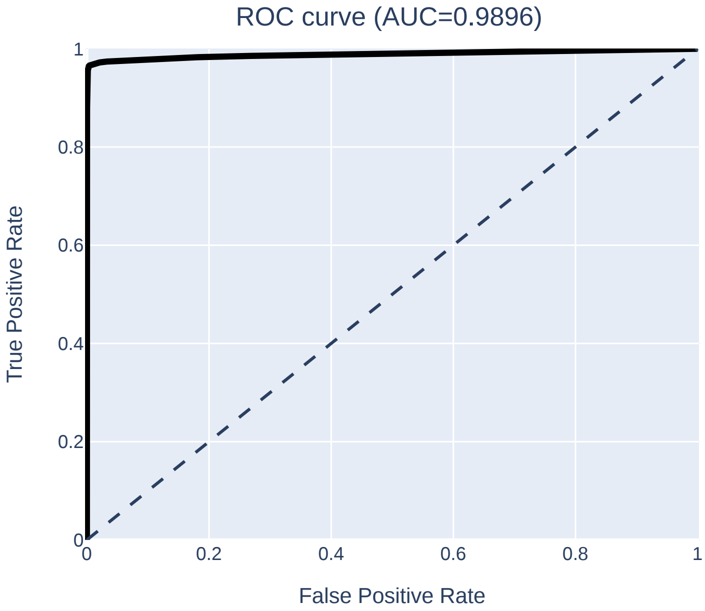


Figure 8: ROC, thresholding SR estimates.

Finally, we report so-called improvement factors in Table 1. These are given as the ratio of the number of specimens N to the expected number of tests required for achieving at least 95% in specificity & sensitivity. When needed, the expected number is computed through Monte Carlo averaging. An improvement factor measures the effectiveness of a given method by computing how many more tests a standard individual testing would need compared to a group testing algorithm. It essentially provides an estimate of how many individuals can one test effectively "cover". For the three prevalence ratios, SR dominates both the the non-adaptive algorithms and Dorfman approach.

	$\frac{k}{N}=2\%$	$\frac{k}{N}=4\%$	$\frac{k}{N}=6\%$
Dorfman	3.37	2.60	2.15
COMP	4.53	2.80	1.96
DD	2.80	1.99	1.49
CBP	4.60	2.81	1.93
SCOMP	3.81	2.48	1.78
SR	5.11	4.01	3.42

Table 1: Improvement factors, $\frac{N}{\mathbb{E}(\text{\# of tests})}$, for three different prevalence rates

5 Reproducible research

The code supplement Kutateladze and Seregina [2020] is available in Google Colab environment. It is written in Python and readily allows to replicate all the graphs provided, as well as produce additional exercises.

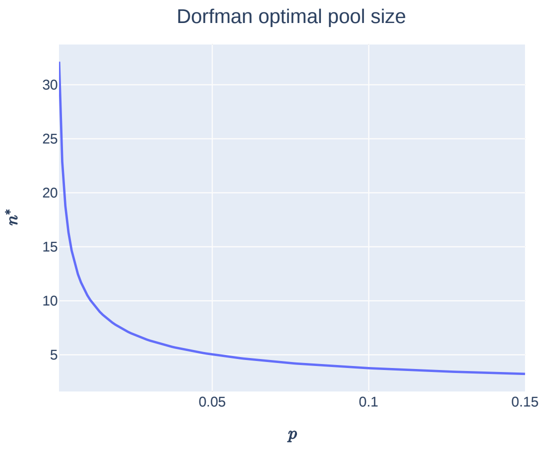
6 Concluding remarks

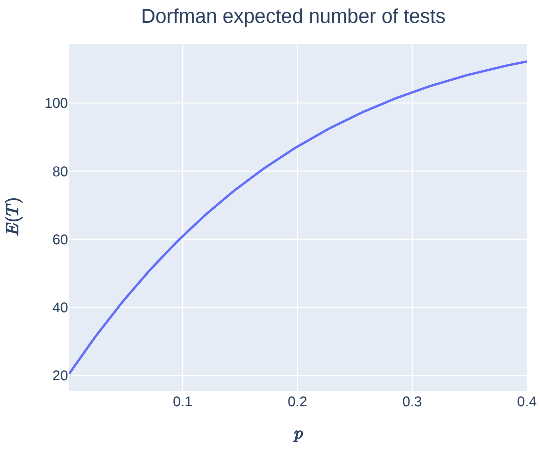
Pooled testing has been around for more than 70 years and has been successfully employed against a number of diseases. There are reasons to believe that pooling can also be effective against SARS-CoV-2. First, low prevalence of the virus is crucial to making group testing effective. Second, the recent evidence with dilution experiments suggests that pooling can be a viable method. Third, pooling is also compatible with the widely used testing kits such as RT-qPCR. Finally, group testing has been authorized by FDA (Food and Administration [2020]), which claimed it to be "especially important as infection rates decline and we begin testing larger portions of the population."

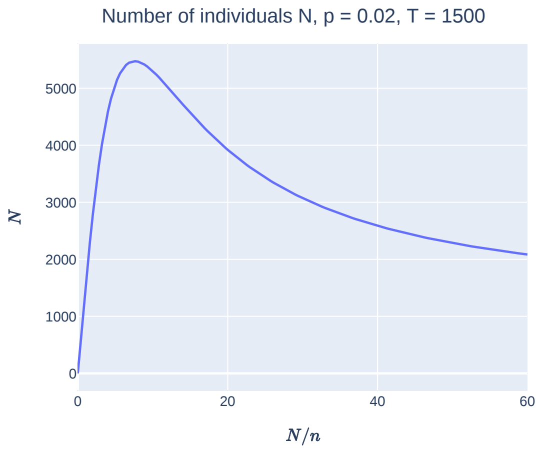
To this end, we considered a simple one-stage group testing method that is able to diagnose a large number of specimens with the fewest number of tests and thus substantially increase the through-out of testing. Our approach does not require to know the number of positive samples in population to run and compares favorably based on the experiments on synthetic data. It is able to perform classification with very few false positives and false negatives, and also capable of predicting viral loads. Compared to widely used adaptive strategies it minimizes latency in delivering test results, while compared with non-adaptive strategies it only requires $m \sim O(k \log n)$ tests.

Appendix

A.1 Dorfman group testing figures







A.2 Proof of Theorem (1)

$\mathrm{A} = \{a\}_{ij}$ is an $m \times N$ matrix where each column randomly permutes $0 < L < m$ ones among zeros. Without loss of generality, let us assume that each column has been demeaned and normalized to be of unit length, i.e. divided by $\sqrt{L\left(1-\frac{L}{m}\right)^2 + (m-L)\left(\frac{L}{m}\right)^2} = \sqrt{L\left(1-\frac{L}{m}\right)}$. It is then evident that $\mathbb{E}(a_{ij}) = 0$ and $\mathbb{E}(a_{ij}^2) = \frac{1}{m}$.

First, we want to show that for any fixed $\mathrm{x} \in \mathbb{R}^N$, the random variable $\|\mathrm{Ax}\|_2^2$ concentrates around its mean, i.e.

$$\Pr\left(\left|\|\mathrm{Ax}\|_2^2 - \|\mathrm{x}\|_2^2\right| \geq \epsilon \|\mathrm{x}\|_2^2\right) \leq 2e^{-m(\epsilon^2/4-\epsilon^3/6)}. \tag{4}$$

For $i = 1, \ldots, m$, denote c_i the ith entry of Ax, i.e. $c_i = \sum_{j=1}^N a_{ij}x_j$, then

$$\mathbb{E}\, c_i = \mathbb{E}\left(\sum_{j=1}^N a_{ij}x_j\right) = \sum_{j=1}^N \mathbb{E}(a_{ij})x_j = 0,$$

$$\begin{aligned}\mathbb{E}(c_i^2) =& \mathbb{E}\left(\left(\sum_{j=1}^N a_{ij}x_j\right)^2\right) = \mathbb{E}\left(\sum_{j=1}^N (a_{ij}x_j)^2 + 2\sum_{l=1}^N\sum_{m=1}^N a_{lj}a_{mj}x_lx_m\right)\\ =& \sum_{j=1}^N \mathbb{E}(a_{ij}^2)x_j^2 + 2\sum_{l=1}^N\sum_{m=1}^N \mathbb{E}(a_{lj})\,\mathbb{E}(a_{mj})x_lx_m = \frac{1}{m}\|\mathrm{x}\|_2^2,\end{aligned}$$

and hence $\mathbb{E}(\|\mathrm{Ax}\|_2^2) = \mathbb{E}\left(\sum_{i=1}^m c_i^2\right) = \sum_{i=1}^m \mathbb{E}(c_i^2) = \|\mathrm{x}\|_2^2$.

Since $\|\mathrm{Ax}\|_2^2$ is proportional to $\|\mathrm{x}\|_2^2$, it is sufficient to demonstrate the concentration for arbitrary unit vectors. For all fixed unit vectors $\mathrm{x} \in \mathbb{R}^N$,

$$\begin{aligned}\mathbb{P}\left(\|\mathrm{Ax}\|_2^2 > 1+\epsilon\right) &= \mathbb{P}\left(e^{t\|\mathrm{Ax}\|_2^2} > e^{t(1+\epsilon)}\right) &(5)\\ &< \mathbb{E}\left(e^{t\|\mathrm{Ax}\|_2^2}\right)e^{-t(1+\epsilon)}, &(6)\end{aligned}$$

where (5) and (6) simply apply the Chernoff technique for $t > 0$. Now, because the columns c_i are i.i.d. we have $\mathbb{E}\left(e^{t\|\mathrm{Ax}\|_2^2}\right) = \mathbb{E}\left(e^{t\sum_{i=1}^m c_i^2}\right) = \left(\mathbb{E}\left(e^{tc_1^2}\right)\right)^m$, leading to

$$\begin{aligned}\mathbb{P}\left(\|\mathrm{Ax}\|_2^2 > 1+\epsilon\right) &< \left(\mathbb{E}\left(e^{tc_1^2}\right)\right)^m e^{-t(1+\epsilon)} &(7)\\ &\leq (1-2t/m)^{-m/2}e^{-t(1+\epsilon)}, &(8)\end{aligned}$$

where (8) follows from Lemma 1. Optimizing this bound with respect to t, $t^* = \frac{m\epsilon}{2(1+\epsilon)}$, we can write

$$\mathbb{P}\left(\|\mathrm{Ax}\|_2^2 > 1+\epsilon\right) < ((1+\epsilon)e^{-\epsilon})^{m/2} \tag{9}$$

$$< e^{-m(\epsilon^2/4-\epsilon^2/6)}, \tag{10}$$

where the last inequality comes from truncating the Taylor approximation of (9). Similarly, for the other bound,

$$\mathbb{P}\left(\|\mathrm{Ax}\|_2^2 < 1-\epsilon\right) < \left(\mathbb{E}\left(e^{-tc_1^2}\right)\right)^m e^{t(1-\epsilon)} \tag{11}$$

$$< \left(\mathbb{E}\left(1-tc_1^2+t^2c_1^4/2\right)\right)^m e^{t(1-\epsilon)} \tag{12}$$

$$\leq \left(1-\frac{t}{m}+\frac{3t^2}{2m^2}\right)^m e^{t(1-\epsilon)} \tag{13}$$

$$= \left(1-\frac{\epsilon}{2(1+\epsilon)}+\frac{3\epsilon^2}{8(1+\epsilon)^2}\right)^m e^{\frac{m\epsilon(1-\epsilon)}{2(1+\epsilon)}} \tag{14}$$

$$< e^{-m(\epsilon^2/4-\epsilon^3/6)}, \tag{15}$$

where (12) and (15) is a Taylor approximation, (13) uses the fact $\mathbb{E}(c_1^4) = \frac{1}{m^2} \leq \frac{3}{m^2}$ and (14) plugs in the earlier value of t^*.

Lemma 1. *For* $m \geq 1$ *and all* $\mathrm{x} \in \mathbb{R}^N$ *s.t.* $\|\mathrm{x}\|_2^2 = 1$, $\mathbb{E}\left(e^{tc_1^2}\right) \leq (1-2t/m), \quad \forall t \in [0, m/2]$.

Proof. Let $W \sim \mathcal{N}(0, \frac{1}{m})$, then

$$\begin{aligned} \mathbb{E}\left(e^{tc_1^2}\right) =& \sum_{i=1}^{\infty} \frac{t^i}{i!} \mathbb{E}\left(c_1^{2i}\right) & (16) \\ \leq& \sum_{i=1}^{\infty} \frac{t^i}{i!} \mathbb{E}\left(W^{2i}\right) & (17) \\ =& \mathbb{E}\left(e^{tW^2}\right) & (18) \\ =& (1-2t/m)^{-1/2}. & (19) \end{aligned}$$

Observe that for $t \in [0, m/2]$ the expectations in (16) and (18) are bounded, allowing to push the expectation inside the limiting sums in (16) and (17). Inequality in (17) holds since $\mathbb{E}(c_1^{2i}) = m^{-i} \leq \mathbb{E}\left(W^{2i}\right) = m^{-i}\frac{(2i)!}{i!2^i}$ holds for each $i = 0, 1, 2, \ldots$. □

Given the concentration of $\|\mathrm{Ax}\|_2^2$ around its mean, RIP follows from Lemma 5.1 in Baraniuk et al. [2008], which is adapted and reiterated below for completeness.

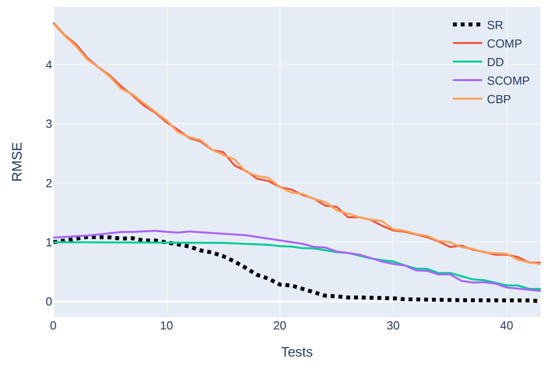
Lemma 2. *Let a random matrix $A \in \mathbb{R}^{m\times N}$ satisfy the concentration inequality in* (4)*. Then, for any set T with $q = \#(T) < m$ and any $0 < \delta < 1$, we have*

$$\Pr\left((1-\delta)\|\mathrm{x}\|_2^2 \leq \|\mathrm{Ax}\|_2^2 \leq (1+\delta)\|\mathrm{x}\|_2^2\right) \geq 1 - 2(12/\delta)^q e^{-(\delta/2)m(\epsilon^2/4-\epsilon^3/6)}$$

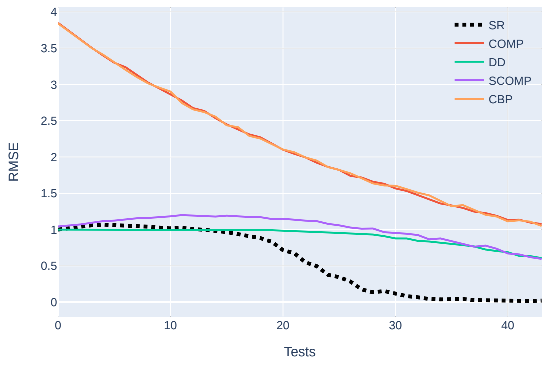
Proof. See Lemma 5.1 in Baraniuk et al. [2008]. □

A.3 Additional experiments for $k = 4$ and $k = 6$

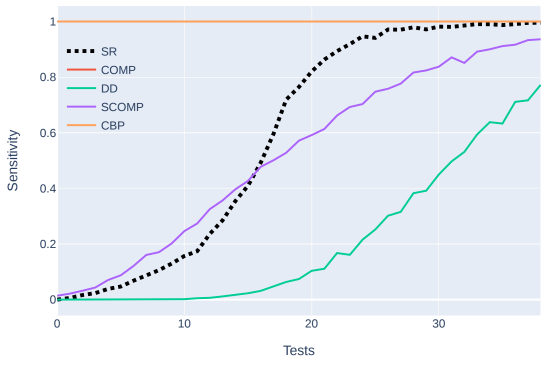
RMSE as a function of group tests, N=100, k=4



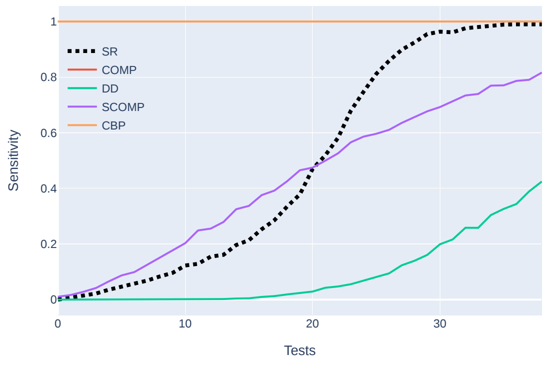
RMSE as a function of group tests, N=100, k=6



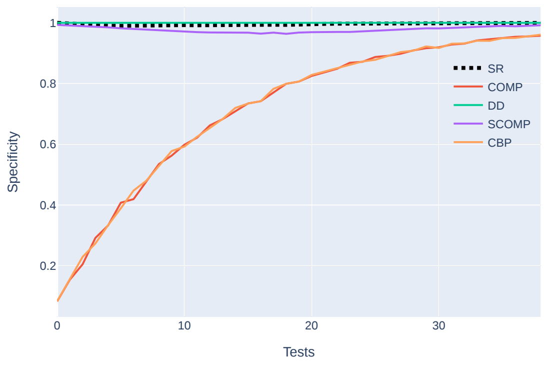
Sensitivity as a function of group tests, N=100, k=4



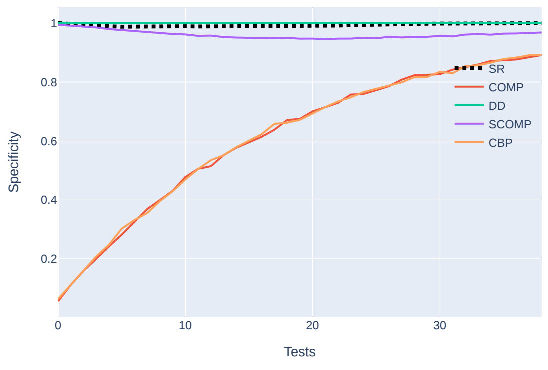
Sensitivity as a function of group tests, N=100, k=6



Specificity as a function of group tests, N=100, k=4



Specificity as a function of group tests, N=100, k=6



References

Abdalhamid, B., Bilder, C. R., McCutchen, E. L., Hinrichs, S. H., Koepsell, S. A., and Iwen, P. C. (2020). Assessment of Specimen Pooling to Conserve SARS CoV-2 Testing Resources. *American Journal of Clinical Pathology*, 153(6):715–718.

Aldridge, M., Johnson, O., and Scarlett, J. (2016). Improved group testing rates with constant column weight designs. In *2016 IEEE International Symposium on Information Theory (ISIT)*, pages 1381–1385.

Bandeira, A. S., Dobriban, E., Mixon, D. G., and Sawin, W. F. (2013). Certifying the restricted isometry property is hard. *IEEE Transactions on Information Theory*, 59(6):3448–3450.

Baraniuk, R., Davenport, M., DeVore, R., and Wakin, M. (2008). A simple proof of the restricted isometry property for random matrices. *Constructive Approximation*, 28(3):253–263.

Candes, E. J., Romberg, J., and Tao, T. (2006). Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52(2):489–509.

Chan, C. L., Che, P. H., Jaggi, S., and Saligrama, V. (2011). Non-adaptive probabilistic group testing with noisy measurements: Near-optimal bounds with efficient algorithms. In *2011 49th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1832–1839.

Donoho, D. L. (2006). Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306.

Dorfman, R. (1943). The detection of defective members of large populations. *Ann. Math. Statist.*, 14(4):436–440.

Emmanuel, J. C., Bassett, M. T., Smith, H. J., and Jacobs, J. A. (1988). Pooling of sera for human immunodeficiency virus (hiv) testing: an economical method for use in developing countries. *Journal of Clinical Pathology*, 41(5):582–585.

Food, U. and Administration, D. (2020). Emergency Authorization for Sample Pooling. `https://www.fda.gov/news-events/press-announcements/coronavirus-covid-19-update-fda-issues-first-emergency-authorization\-sample-pooling-diagnostic`.

Ghosh, S., Agarwal, R., Rehan, M., Pathak, S., Agarwal, P., Gupta, Y., Consul, S., Gupta, N., Goyal, R., Rajwade, A., and Gopalkrishnan, M. (2020). A compressed sensing approach to group-testing for covid-19 detection.

Hogan, C. A., Sahoo, M. K., and Pinsky, B. A. (2020). Sample Pooling as a Strategy to Detect Community Transmission of SARS-CoV-2. *JAMA*, 323(19):1967–1969.

Hughes-Oliver, J. M. (2006). *Pooling Experiments for Blood Screening and Drug Discovery*, pages 48–68. Springer New York, New York, NY.

Kutateladze, V. and Seregina, E. (2020). Code supplement to "Fast and Efficient Data Science Techniques for Covid-19 Group Testing. `https://tinyurl.com/y4vo86sb`.

Litvak, E., Tu, X. M., and Pagano, M. (1994). Screening for the presence of a disease by pooling sera samples.

Mutesa, L., Ndishimye, P., Butera, Y., Souopgui, J., Uwineza, A., Rutayisire, R., Musoni, E., Rujeni, N., Nyatanyi, T., Ntagwabira, E., Semakula, M., Musanabaganwa, C., Nyamwasa, D., Ndashimye, M., Ujeneza, E., Mwikarago, I. E., Muvunyi, C. M., Mazarati, J. B., Nsanzimana, S., Turok, N., and Ndifon, W. (2020). A strategy for finding people infected with sars-cov-2: optimizing pooled testing at low prevalence. *medRxiv*.

Shaobing Chen and Donoho, D. (1994). Basis pursuit. In *Proceedings of 1994 28th Asilomar Conference on Signals, Systems and Computers*, volume 1, pages 41–44 vol.1.

Sobel, M. and Groll, P. A. (1959). Group testing to eliminate efficiently all defectives in a binomial sample. *Bell System Technical Journal*, 38(5):1179–1252.

Sterrett, A. (1957). On the detection of defective members of large populations. *The Annals of Mathematical Statistics*, 28(4):1033–1036.

Taylor, S. M., Juliano, J. J., Trottman, P. A., Griffin, J. B., Landis, S. H., Kitsa, P., Tshefu, A. K., and Meshnick, S. R. (2010). High-throughput pooling and real-time pcr-based strategy for malaria detection. *Journal of Clinical Microbiology*, 48(2):512–519.

Van, T. T., Miller, J., Warshauer, D. M., Reisdorf, E., Jernigan, D., Humes, R., and Shult, P. A. (2012). Pooling nasopharyngeal/throat swab specimens to increase testing capacity for influenza viruses by pcr. *Journal of Clinical Microbiology*, 50(3):891–896.

Worldometer (2020). US SARS-CoV-2 cases. `https://www.worldometers.info/coronavirus/country/us/`.

Yelin, I., Aharony, N., Shaer-Tamar, E., Argoetti, A., Messer, E., Berenbaum, D., Shafran, E., Kuzli, A., Gandali, N., Hashimshony, T., Mandel-Gutfreund, Y., Halberthal, M., Geffen, Y., Szwarcwort-Cohen, M., and Kishony, R. (2020). Evaluation of covid-19 rt-qpcr test in multi-sample pools. *medRxiv*.

Yi, J., Mudumbai, R., and Xu, W. (2020). Low-cost and high-throughput testing of covid-19 viruses and antibodies via compressed sensing: System concepts and computational experiments.