NBER WORKING PAPER SERIES

GROUP TESTING IN A PANDEMIC: THE ROLE OF FREQUENT TESTING, CORRELATED RISK, AND MACHINE LEARNING

Ned Augenblick
Jonathan T. Kolstad
Ziad Obermeyer
Ao Wang

Working Paper 27457
http://www.nber.org/papers/w27457

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
July 2020

Authors in alphabetical order. We are grateful to Katrina Abuabara, Sylvia Barmack, Kate Kolstad, Maya Petersen, Annika Todd, Johannes Spinnewijn, and Nicholas Swanson for helpful comments. All opinions and errors are our own. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

At least one co-author has disclosed a financial relationship of potential relevance for this research. Further information is available online at http://www.nber.org/papers/w27457.ack

Group Testing in a Pandemic: The Role of Frequent Testing, Correlated Risk, and Machine Learning

Ned Augenblick, Jonathan T. Kolstad, Ziad Obermeyer, and Ao Wang

NBER Working Paper No. 27457
July 2020
JEL No. I1,I18

ABSTRACT

Group testing increases efficiency by pooling patient specimens and clearing the entire group with one negative test. Optimal grouping strategy is well studied in one-off testing scenarios with reasonably well-known prevalence rates and no correlations in risk. We discuss how the strategy changes in a pandemic environment with repeated testing, rapid local infection spread, and highly uncertain risk. First, repeated testing mechanically lowers prevalence at the time of the next test. This increases testing efficiency, such that increasing frequency by x times only increases expected tests by around $\sqrt{x}$ rather than x. However, this calculation omits a further benefit of frequent testing: infected people are quickly removed from the population, which lowers prevalence and generates further efficiency. Accounting for this decline in intra-group spread, we show that increasing frequency can paradoxically reduce the total testing cost. Second, we show that group size and efficiency increases with intra-group risk correlation, which is expected in natural test groupings based on proximity. Third, because optimal groupings depend on uncertain risk and correlation, we show how better estimates from machine learning can drive large efficiency gains. We conclude that frequent group testing, aided by machine learning, is a promising and inexpensive surveillance strategy.

Ned Augenblick
University of California, Berkeley
Haas School of Business
EAP Group
545 Student Services Building, 1900
Berkeley, CA 94720-1900
ned@haas.berkeley.edu

Jonathan T. Kolstad
Haas School of Business
University of California, Berkeley
Berkeley, CA 94720
and NBER
jkolstad@berkeley.edu

Ziad Obermeyer
School of Public Health
University of California at Berkeley
2121 Berkeley Way
Berkeley, CA 94704
zobermeyer@berkeley.edu

Ao Wang
University of California, Berkeley
Evans Hall
Berkeley, CA 94705
ao.wang@berkeley.edu

1 Introduction

The current costs and supply constraints of testing make frequent, mass testing for SARS-CoV-2 infeasible. The old idea of group testing (Dorfman (1943)) has been proposed as a solution to this problem (Lakdawalla et al. (2020); Shental et al. (2020)): to increase testing efficiency, samples are combined and tested together, potentially clearing many people with one negative test. Given the complicated tradeoff between the benefits of increasing group size and the cost of follow-up testing for a positive result, a large literature has emerged on optimal group testing strategies. This literature uses a set of assumptions grounded in the common use cases to date: one-time testing of a set of samples, e.g., screening donated blood for infectious disease,[1] with independent risk of infection (Dorfman (1943); Sobel and Groll (1959); Hwang (1975); Du et al. (2000); Saraniti (2006); Feng et al. (2010); Li et al. (2014); Aprahamian et al. (2018, 2019)).

Many of these environmental assumptions are violated when dealing with a novel pandemic with rapid spread. In this case, people may need to be tested multiple times before they are infected, testing groups will be formed from populations with correlated infection risk, and risk levels at any time are very uncertain. This paper notes how these different factors change the optimal testing strategy, and open up ways to dramatically increase testing efficiency. This would allow data-driven, frequent testing – even daily – in workplaces and communities as a cost-effective way to contain infection spread.

We start with the well-known observation that group testing is more efficient when the population prevalence is lower, because the likelihood of a negative group test is increased. We then show how increased testing frequency lowers the prevalence and therefore increases efficiency. For example, in a simple model with a reasonable level of independent risk, testing twice as often cuts the prevalence at the time of testing by (about) half, which lowers the expected number of tests at each testing round to about 70% of the original number. The savings are akin to a "quantity discount" of 30% in the price of testing. Therefore, rather than requiring two times the numbers of tests, doubling the frequency only costs around 40% more tests. More generally, we demonstrate that testing more frequently requires fewer tests than might be naively be expected: increasing the frequency by x times only uses about $\sqrt{x}$ as many tests, implying a quantity discount of $(1 - 1/\sqrt{x})\%$.

The benefits to frequency are even greater when there is intra-group spread, as would be expected in a pandemic. In this case, testing more frequently has the additional benefit of quickly removing infected individuals, which stops others from being infected. This in turn lowers prevalence, and provides yet another driver of efficiency. We show that in this case – somewhat paradoxically – the quantity discount is so great that more frequent testing can actually *lower* the total number of tests. Given that current testing for COVID-19 is done

[1] This is the most common real-world use of group testing today (Cahoon-Young et al. (1989); Behets et al. (1990); Quinn et al. (2000); Dodd et al. (2002); Gaydos (2005); Hourfar et al. (2007). More recently for COVID-19, Nebraska has adopted group testing as have some national governments such as Ghana and India.

relatively infrequently, we therefore believe the optimal frequency is likely much higher.[2]

Next, we note that grouping samples from people who are likely to spread the infection to each other – such as those that work or live in close proximity – increases the benefits of group testing. Intuitively, increased correlation of infection in a group with a fixed risk lowers the likelihood of a positive group test result, which increases efficiency. Consequently, we conclude that groups should be formed from people who are likely to infect each other, such as those in a work or living space.

Finally, we note that, while there is a literature that estimates risk profiles to drive groupings (Hwang (1975); Bilder et al. (2010); Bilder and Tebbs (2012); Black et al. (2012); McMahan et al. (2012); Tebbs et al. (2013); Black et al. (2015); Aprahamian et al. (2019)), the methods are often appropriately simplistic and designed for situations with stable risk rates, no correlation, and large amounts of previous outcome data. These methods are likely less appropriate for risk estimation in a quickly-changing pandemic with highly uncertain and correlated risk, such as COVID-19. We discuss the impacts and solutions to these issues in designing groups. We start by quantifying the significant efficiency losses from choosing groups based on incomplete or incorrect risk and correlation information. We then advocate for a strategy using machine learning to continually estimate these parameters using observable data such as location, demographics, age, job type, living situation, along with past infection data.

Though our results are theoretical, they have a clear application in dealing with the COVID-19 pandemic. Therefore, we end by discussing potential solutions to important practical challenges, such as implementation barriers, imperfect tests, and fixed testing costs.

The paper proceeds as follows: Section 2 reviews a main finding in the group testing literature that efficiency rises as prevalence falls; Section 3 discusses the relationship between testing frequency and efficiency; Section 4 demonstrates how correlated infection leads to larger group sizes and greater efficiency; Section 5 discusses the usefulness of machine learning to estimate risk and correlation; Section 6 discusses important implementation challenges and caveats; and Section 7 concludes.

2 Group Testing: Benefits rise as prevalence falls

2.1 Background on Group Testing

To understand the basic benefits of group testing, consider a simple example of 100 people, each having an independent likelihood of being positive of 1% and a test that (perfectly) determines if a sample is positive. To test each person individually, the conventional approach, requires 100 tests.

[2]It is important to note that these results do not rely on intense optimization of group size or sophisticated cross-pooling or multi-stage group testing. Furthermore, the results are qualitatively similar when using other reasonable group sizes.

Suppose instead that the individuals' samples are combined into five equally-sized groups of 20. Each of these combines samples are then tested with one test. If any one of the 20 individuals in a combined sample is positive then everyone in that group is individually tested, requiring 20 more tests (21 in total). The probability that this occurs is $1-(1-.01)^{20} \approx 18\%$. However, if no one in the group is positive – which occurs with probability $\approx 82\%$ – no more testing is required. Because the majority of tests require no testing in the second case, the expected number of tests for this simple grouping method is only around 23, a significant improvement over the 100 tests required in the non-grouped method.

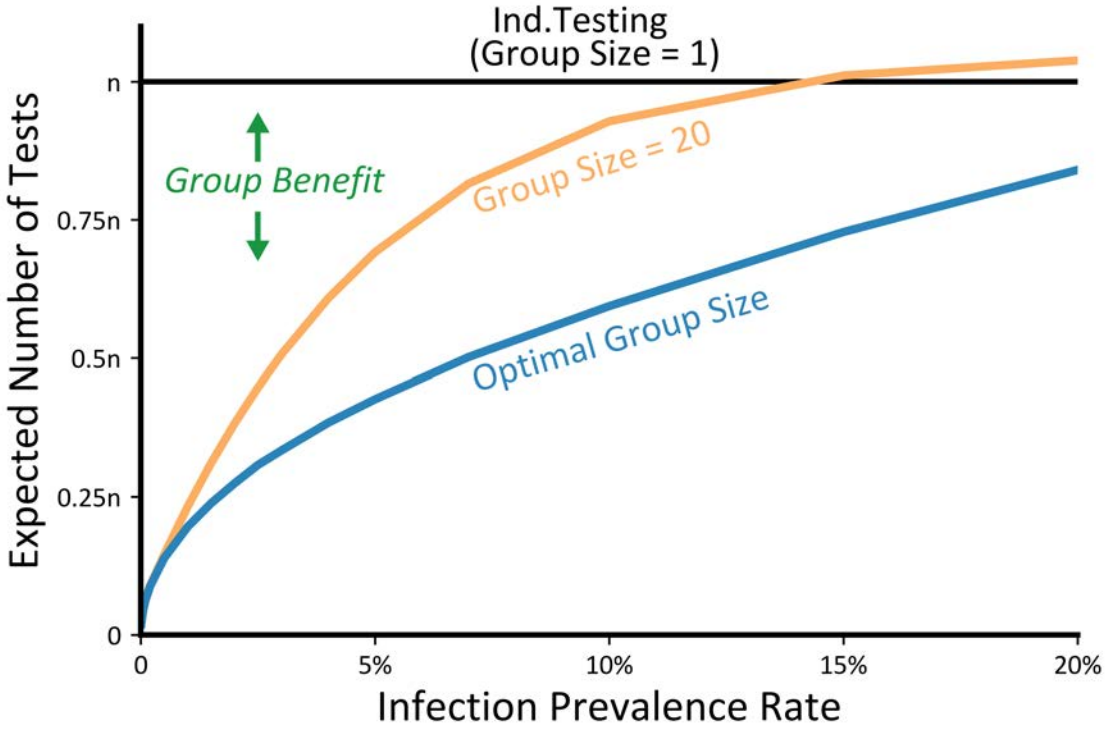
The approach is well studied with a large literature focused on improving the efficiency of group testing. These include using optimal group size (e.g. in this example the optimal group size of 10 would lower the expected number of tests to around 20), placing people into multiple groups (Phatarfod and Sudbury (1994)), and allowing for multiple stages of group testing (Sterrett (1957); Sobel and Groll (1959); Litvak et al. (1994); Kim et al. (2007); Aprahamian et al. (2018)). Particularly noteworthy is the strategy of Litvak et al. (1994), who show that limited retesting of group specimens (including negatives) can increase sensitivity and specificity at lost cost (if test results are conditionally independent of the true result). There are also methods to deal with complications, such as incorporating continuous outcomes (Wang et al. (2018)) or optimizing given imperfect tests and divisibility problems (which is largely solved in Aprahamian et al. (2019) for the standard case). Any of these modifications can be incorporated in our group testing strategy. For simplicity of exposition, in the paper, we present results for simple two-stage "Dorfman" testing – in which every person in a positive group is tested individually – to make calculations transparent, and demonstrate that our conclusions are not driven by particular assumptions and do not require highly complex groupings.[3]

2.1.1 Laboratory Evidence on Group Testing for SARS-CoV-2

Fortunately, several studies have looked specifically at the feasibility of group testing for detection of SARS-CoV-2 via PCR. A particular focus of these studies is false negatives, driven by sample dilution. Yelin et al. (2020) found that group sizes up to 32 were feasible at usual amplification levels, and Shental et al. (2020) use pools of 48, albeit with a more complex combinatorial pooling strategy. Hogan et al. (2020) evaluate the false positive rate and find it to be very low. Overall, these studies paint a reassuring picture of the test characteristics of grouping for SARS-CoV-2 via PCR specifically; these could be further improved with the method of Litvak et al. (1994).

[3]In general, we advocate for these more sophisticated strategies when feasible as they further increase efficiency.

Figure 1: Efficiency of group testing rises with prevalence



Notes: This figure plots the expected number of tests (y-axis) from group testing given a population of n people as the population infection prevalence rate (x-axis) changes. The black flat line shows the number of tests from individual testing (equivalent to a group size of 1), which always requires n tests regardless of prevalence. The results from using a group size of 20 is orange, while the blue line represents the number of tests given the optimal group size for a given prevalence. Finally, the green text notes that benefit from group testing is the distance between the black individual-testing line and those from group testing. For example, as noted in the text, using a group size of 20 for a prevalence of 1% leads to $.23 \cdot n$ tests rather than n tests, while the optimal group size (10) leads to $.20 \cdot n$ tests.

2.2 Prevalence and Group Testing

For all of these different incarnations of group testing, the benefits of group testing rise as the prevalence rate falls in the population. Lower prevalence reduces the chance of a positive group test, thereby reducing the likelihood the entire pool must be retested individually. For example, if the prevalence drops from .1% to .01%, the likelihood of a positive result in the first stage drops significantly, such that the expected number of tests using a group size of 20 drops to 6.9 tests (or 6.3 tests given the optimal group size). Similarly, if the prevalence rises from .1% to 1%, the expected number of tests using a group size of 20 rises to 93 (or 59 given the optimal grouping).

The full relationship is shown in Figure 1, which plots the expected number of tests in a population of n people given different group sizes and visually highlights the results based on (i) individual testing – which always leads to n tests, (ii) using groups of 20, and (iii) using optimal grouping given two stages. For simplicity, we construct

these figures by assuming that n is large to remove rounding issues that arise from breaking n people into groups sizes that are not divisible by n.[4] There are large gains from group testing at any prevalence level, though they are appreciably larger at low prevalence rates.

3 Increasing Test Frequency

3.1 Interaction Between Frequent Testing and Group Testing

Our first insight is the important complementarity between group testing and testing frequency. Intuitively, the benefits of group testing rise as prevalence falls and frequent testing keeps the prevalence at each testing period low.

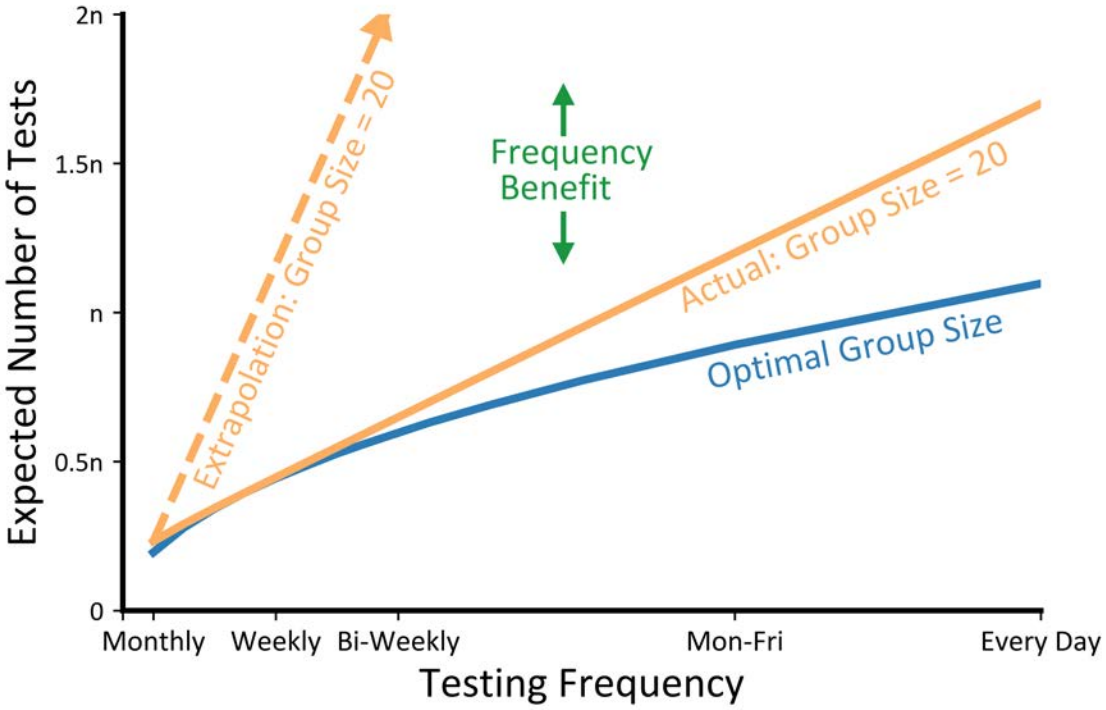
Continuing with our example, suppose that 100 people have a 1% independent chance of being positive over the course a given time period. As discussed above, one could either sample everyone (requiring 100 tests), use group testing with a group size of 20 (requiring 23 expected tests), or use group testing with an optimal group size (requiring 20 expected tests).

Suppose instead that people are tested ten times as frequently. Testing individually at this frequency requires ten times the number of tests, for 1000 total tests. It is therefore natural think that group testing also requires ten times the number of tests, for more than 200 total tests. However, this estimation ignores the fact that testing ten times as frequently reduces the probability of infection at the point of each test (conditional on not being positive at previous test) from 1% to only around .1%. This drop in prevalence reduces the number of expected tests – given groups of 20 – to 6.9 at each of the ten testing points, such that the total number is only 69. That is, testing people 10 times as frequently only requires slightly more than three times the number of tests. Or, put in a different way, there is a "quantity discount" of around 65% by increasing frequency. The same conclusion holds using optimal group sizes: the one-time group test would require 20 expected tests, while testing ten times as frequently requires 6.3 tests at each testing point for a total of 63. The savings relative to the 1000 tests using individual testing are dramatic, with only approximately 6% of the total tests required.

Figure 2 represents this effect more generally for different levels of test frequency given a prevalence of 1% over the course of a month. Note that, at a frequency of a once a month, the numbers match those in Figure 1, which was based on one test at a prevalence of 1%. Unlike in Figure 1, we do include the results for individual testing in this graph as testing individually everyday requires 20-30 times more tests than group testing, which renders the graph unreadable. The dotted orange line represents the naive (and incorrect) calculation for group testing by extrapolating the cost of testing multiple times by using the number of tests required for one test. That is, as above, one might naively think that testing x times using a group size of 20 in a population of n

[4] We note that this figure replicates many similar figures already in the literature going back to Dorfman (1943).

Figure 2: Efficiency of group testing rises with frequency



Notes: This graph presents the effect of testing frequency (x-axis) on the expected number of tests (y-axis), given a prevalence in the population at 1% over a month. When the frequency is once a month, the points correspond to those in Figure 1 given prevalence of 1%: n for individual testing, $.23 \cdot n$ when using a group size of 20 and $.20 \cdot n$ tests when using the optimal group size. The dotted orange line represents the (incorrect) extrapolation that if a group size of 20 leads to $.23 \cdot n$ tests when frequency is once a month, it should equal $x \cdot .23 \cdot n$ if frequency is x times a month. In reality, the expected tests are much lower, due to a "quantity discount" or "frequency benefit," highlighted by the green text. Finally, the blue line highlights tests given the optimally-chosen group size.

would require $x \cdot .23 \cdot n$ tests given that testing once requires $.23 \cdot n$ tests. Group testing is in fact much cheaper due to the reduction in prevalence — the central contribution of this section. We therefore denote the savings between the extrapolation line and the actual requirements of group testing as the "frequency benefit."

The level of savings of the frequency benefit changes depending on the prevalence p given one test and the frequency x. Interestingly, numerical simulations suggest that expected cost of group testing x times as frequently is always around $\sqrt{x}$ when p is not too high (p<.05) when using optimal-group-sized two-stage group testing, and asymptotes to this exact amount as p falls to zero. In other words, the quantity discount of increased frequency is close to $(1 - 1/\sqrt{x})\%$. So, for example, group testing using optimally-sized groups every week (about 5 times total) costs around $\sqrt{x} \approx 2.24$ times the number of tests from group testing every month, implying a quantity discount of 55%. Or, in an extreme example, testing 3 times a day (around 100 times a month) costs about $\sqrt{100} = 10$ times the tests, implying a quantity discount of 90%. We have found that this formula provides a

very good approximation for reasonable prevalence rates and fixed group sizes.

3.2 Avoiding Exponential Spread Through Frequent Testing

The logic above ignores a major benefit of frequent testing: identifying infected people earlier and removing them from the population. Beyond the obvious benefits, removing people from the testing population earlier stops them from infecting others, which reduces the prevalence, and therefore increases the benefit of group testing.

In the previous section, we shut down this channel by assuming that every person in the testing population had an independent probability of becoming infected. If the testing population includes people that interact, such as people who work or live in the same space, infections will instead be correlated.

To model this correlation, we construct a simple network of infection for the testing population.[5] Specifically, we suppose that the 100 people continue to face a 1% chance of being infected over the course of a time period due to contact with someone outside of the test population. However, once a person is infected, this person can spread the infection to other people. We then chose an infection transmission rate such that if a person was infected at the start of the time period and was not removed from the population, the infected person would have a total independent probability x% over the time period of personally spreading the infection to each other person. [6]

Now, consider the spread over the entire time period. With 37% probability, no one in the population is infected. However, with a 63% chance, a person is infected from the outside of the group, setting off exponential-like growth within the rest of the test population (which ends only after testing and infected people are removed).[7] Given this growth, for example, if x=1%, 5%, 10%, approximately 2, 8, and 26 people infected at the end of the time period. However, by testing ten times in the time period, infected people are removed more quickly and there is less time for a person to infect others, such that only around 1, 2, or 4 people are infected by the end of the time period.

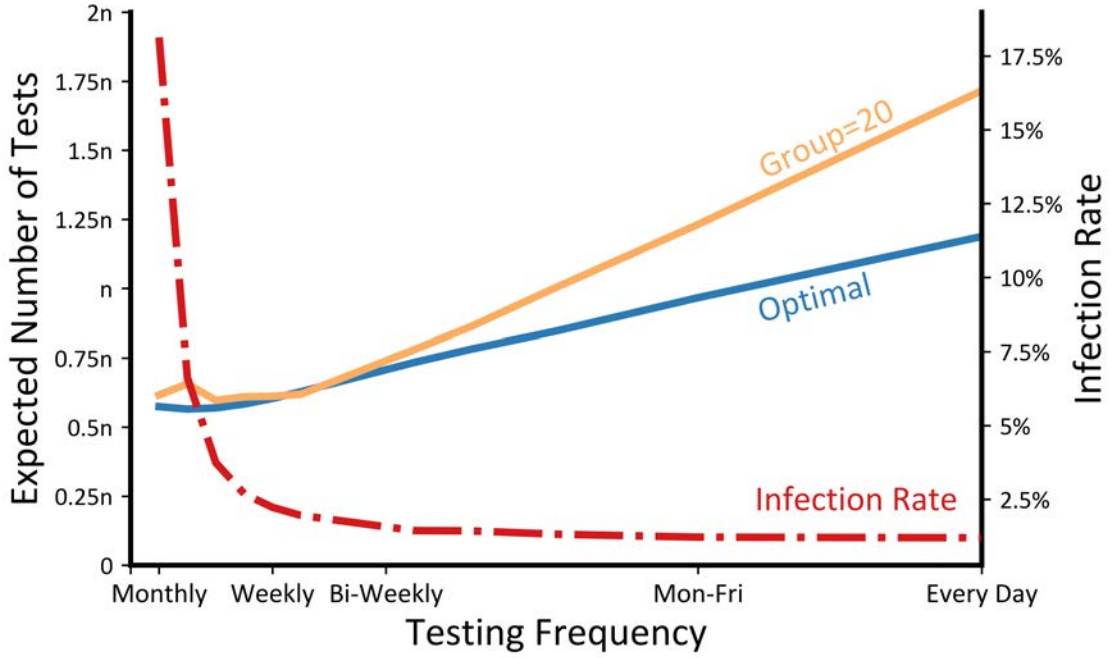
Not only are infections rates reduced, but the reduction in prevalence reduces the relative cost of more frequent group testing. For example, using a group size of 20 and testing once over the time period leads to 23 tests in expectation given no spread. However, with spread given x=1%, 5%, 10%, this number rises to 30, 50, and 58, respectively. On the other hand, the expected number of tests when testing ten times as frequently does not grow at the same rate due to controlling the spread and therefore prevalence. Specifically, the number of tests needed

[5] There are different infection networks in different testing population. Numerical simulations using a variety of potential structures are qualitatively similar.

[6] For example, if x=5% and the time period is a month, then infected person i who has not been removed has a independent $1-(1-.05)^{(1/30)} \approx .17\%$ daily chance of infecting non-infected person j, $\approx .17\%$ daily chance of infecting non-infected person k, etc.

[7] We consistently use the term "exponential-like" spread as we are considering smaller testing populations in which the infection rate slows as infected people are removed or become immune, such that the spread is actually logistic. However, in our simulations, we focus on situations in which the spread is caught early and therefore still close to exponential.

Figure 3: Increased frequency lowers infections with few additional tests given intra-group spread



Notes: This graph presents the effect of testing frequency (x-axis) on the expected number of tests (y-axis 1) and infection rate (y-axis 2) given a model with intra-group spread over the course of a month. As shown in red dot-dashed line of infection rates, increased frequency reduces the exponential-like spread because infected people are removed from the population. The number of expected tests required is shown for group size 20 in orange and in blue for the optimal group size. There is not much increase (and even an early decrease) in the number of tests required as frequency increases because the increased frequency reduces the spread and therefore prevalence.

rises from 69 (with no spread) to 71, 84, and 115, respectively. That is, for example, at x=5%, the number of needed tests rises by 150% when testing once due to spread, but only rises by 22% when testing ten times as frequently.

These effects are shown in Figure 3. We plot the expected number of tests (left y-axis) and infection rate (right y-axis) for different testing frequencies assuming x=5%.[8] The infection rate rises in an exponential-like manner as frequency decreases and the infection is allowed to spread. The expected number of tests given different frequencies uses the same colors to represent group sizes of 20 (orange) and optimal size (blue). Comparing Figures 2 and 3 is instructive. In Figure 2, we see a consistent increase in the tests required as the frequency of testing is increased. In Figure 3, though, the tests required are relatively flat and even decrease for early frequency increases. The difference is due to the fact that, with intra-group infections, testing more frequently has the additional benefit of lowering the prevalence rate by containing infections. For example, the quantity

[8]There is one minor difference between this figure and the example in the text. In the figure, we use a population of 120 rather than 100, as it allows for easier division into many group sizes. This has a very slight impact as it slightly increases the time for the exponential spread to be curbed by increasing population immunity.

discount from a frequency of 2 is higher than 50% in the case of optimal group sizes, such that the total cost from doubling the frequency actually falls.

3.3 Optimal Testing Frequency

The main benefit of increased frequency is reducing the exponential rise in infections. As shown in Figure 3, the marginal benefit from reduced infections due to increasing frequency is high at low frequencies and drops as frequency rises, eventually to zero. We can trade-off the marginal benefit with the increased cost associated with higher frequency testing that requires more tests. This is not straightforward, however. As shown in Figure 3, the number of tests can actually fall as frequency rises when starting at low frequencies. Therefore, for low frequencies, there is, in fact, no trade-off of raising frequency: it both reduces infections and reduces tests.

As testing frequency rises, the number of expected tests will inevitably rise leading to a trade-off between marginal benefit and cost.[9] Consequently, at very higher frequencies, there is an increased cost without a large benefit. The optimal frequency lies between these extremes, but depends on the value of reducing infections versus the cost of tests, which is an issue beyond the scope of this paper. However, our strong suspicion given the economic costs of testing versus the economic (and human) costs of higher infections is that the optimal frequency is higher than the seemingly-common policy of testing on the order of monthly or only when an employee is symptomatic.[10]

4 Correlated Infection

In this Section, we discuss and isolate an additional factor implicitly included in the above example, which includes correlation between people whose samples are grouped together. We use our simple example to separate the insight that this correlation alone (i) is complementary with group testing in that it reduces the number of expected tests, (ii) leads to larger optimal group sizes, and (iii) has greater reduction if the structure is known and used to determine the composition of the groups.

To first understand the benefit of correlation given group testing, it is useful to broadly outline the forces that determine the expected number of tests with simple two stage testing with a group size of g and a large testing population n. In the first stage, a test will be run for every n/g group, while in the second stage, every n/g group faces a probability q that at least one sample will be positive, such that all g people in the group will need to be individually tested. Combining and simplifying these factors leads to a simple formula of the expected

[9] As an extreme example, if testing is so frequent that the prevalence rate at each test is effectively zero, then increasing the frequency by 1 will lead to an additional test for each group without reducing the prevalence rate at each testing period. This can be seen in Figure 3 for group size of 20 where, at a frequency of around bi-weekly, the number of expected tests rises close to linearly with a slope of $1/20 = .05 \cdot n$.

[10] We also note a an important comparative static: it is more valuable to more frequently test a person who is more likely to catch and spread the infection (such as a person who meets with many other people versus a person who works alone in the back room).

number of tests given a group size: $n \cdot (1/g + q)$. As noted above, in the case of infections with independent probability p, $q = 1 - (1-p)^g$. However, as infections become more positively correlated, q falls for every group size $g > 1$. For example, with two people in a group whose infections have correlation r, q can be shown to be $1 - (1-p)^2 - r \cdot p \cdot (1-p)$. That is, when $r = 0$, we recover the original formula $1 - (1-p)^2$, while raising r linearly drops the probability until it is p when $r = 1$. Intuitively, the group has a positive result if either person 1 *or* person 2 is infected, which – holding p constant – is less likely when infections are correlated and therefore more likely to occur simultaneously.

As an example of how q falls with more people and consequently reduces the number of tests, suppose that $p = .01$: when infections are uncorrelated, q is around 9.6%, 18.2%, 26.0%, and 33.1% given respective group sizes 10, 20, 30, and 40, while q respectively drops to around 3.1%, 3.9%, 4.4%, and 4.8% when every person is pairwise-correlated with $r = 0.5$. Therefore, the respective expected number of tests given these group sizes falls from $.196 \cdot n$, $.232 \cdot n$, $.294 \cdot n$, and $.356 \cdot n$ when uncorrelated to $.131 \cdot n$, $.089 \cdot n$, $.077 \cdot n$, and $.073 \cdot n$ when $r = 0.5$. First, note that the number of expected tests is universally lower at every group size given correlation (and the savings are very significant). Second, note that while the group size with the lowest number of expected tests given these potential group sizes is 10 when there is no correlation, larger group sizes are better given correlation. This statement is more general: higher correlation raises the optimal group size. The intuition is that the marginal benefit of higher group size (reducing the 1/g first-stage tests) is the same with or without correlation, but the marginal cost (increasing the probability of second stage testing) is reduced with higher correlation, thus leading to a higher optimum. As an example, while the optimal group size given $p = .01$ is 10 given no correlation, the optimal group sizes given r of 0, 0.2, 0,4, 0.6, 0.8 are 11, 22, 44, 107, and 385, respectively. Finally, note that when $r = 1$, the optimal group size is unbounded, because adding an extra person to a group adds benefit by reducing first-stage tests, but has no cost because the probability of a positive in the group remains constant at p. Obviously, extremely large group sizes are potentially technologically infeasible, but the intuition remains that correlation within testing groups should raise group size.

5 Machine Learning Predictions of Correlated Risk

To this point, we have assumed or imposed prevalence and infection risk. In practice, optimal group size, composition and frequency are based on (presumably unknown) prevalence and infection structure. Machine learning can be used to estimate these key parameters, enabling the application of our strategy. In this section, we explore how combining machine learning techniques with information on the testing population – in aggregate and at the individual level – and indirect information about network structure can significantly improve group testing efficiency by allowing us to predict empirical measures of prevalence and infection. We focus on demonstrating

the value of accurate prediction in terms of efficiency and present some general approaches to prediction that lend themselves to the COVID-19 testing environment. We do not, however, develop a specific algorithm as the prediction methods themselves are beyond the scope of this paper and likely to be quite specific to each testing setting.

5.1 The Role of Accurate Prediction

To quantify the value of accurate predictions, we introduce heterogeneity in risk to our simple simulations. We consider a large population in which half of the population is high risk (5% chance of being infected at testing) and half is low risk (.1% chance of infection). Within these risk groups, there are clusters of 40 people whose infections are pairwise correlated with $r = 0.25$. These clusters could represent floors of an office building or apartment building, in which people can easily spread the infection to each other in a cluster but don't spread across clusters.

Machine learning has two objectives in this framework. First, it can make explicit predictions on individual risk in each group. Second, it can be used to recover the (unobserved) structure of risk correlations that are produced by person-to-person transmission of infection.[11]

To understand the value of additional information machine learning could provide we consider a variety of scenarios for testing in our simulation. Table 1 presents these scenarios; each differs based on a measure of the quality of prediction or beliefs of the testing strategy designer. The first line captures the baseline case with no group testing. The 2nd through 4th lines capture plausible scenarios for ad hoc estimates of risk. The 5th and 6th lines the kinds of information amenable to machine learning predictions.

Implementing a group testing strategy requires an estimate for the underlying risk to determine group size. How such a risk level is chosen in practice likely varies but it is plausible that simple heuristics are used. If we were to assume the whole population had the relatively low risk (.1%) the optimal group size is large: 34. However, in practice there are too many high risk individuals who end up in such large groups, requiring more testing in the second stage, thereby lower efficiency. In fact, group testing in this case only lowers the total expected tests to $.58n$. Alternatively, one could assume the whole population is relatively high risk (5%). That has the opposite problem: smaller groups than optimal given that many people are low risk. The efficiency gains are larger under this assumption, in part because the returns to group testing at any level are high, the central issue being avoiding re-testing. We see a reduction in tests to $.32n$. Suppose instead we simply took the mean infection rate. The 4th line of Table 1 shows this scenario (population risk of 2.5%). The implied group size is 8 and we generate $.3n$ tests.

[11] We note that, even when machine learning is impossible due to data constraints, ex ante knowledge of transmission dynamics – e.g., working the same shift, residing on the same building floor – can also help approximate this structure. Furthermore, despite being a simplified representation, the broad lessons hold with more complicated risk heterogeneity and infection networks.

Table 1: Efficiency Gains From Information

Belief	Implied grouping	Exp Tests	Issue
No group testing	1	n	Individual Testing Inefficient
Everyone .1%	34	$.58n$	Believe all low risk $\Longrightarrow$ groups too big
Everyone 5%	5	$.32n$	Believe all high risk $\Longrightarrow$ groups too small
Everyone ˜2.5%	8	$.30n$	Believe homogenous $\Longrightarrow$ don't split
Half .1%, Half 5%	.1%→34,5%→5	$.24n$	Believe $\rho = 0 \Longrightarrow$ groups too small
Half .1%, Half 5% (corr)	.1%→40,5%→10	$.18n$	Optimal

Notes: This table outlines the effect of different beliefs on grouping strategy given a situation in which half of the n people have .1% correlated risk with $r = 0.25$ and the other half have 5% correlated risk with $r = 0.25$. For example, in the second row, if the group designer (falsely) believes that everyone is a low risk type with .1% uncorrelated risk, everyone will be placed into believed-to-be-optimal groups of 34, leading the expected number of tests to be $.58n$, which is sub-optimal because the groups for the high-risk types are too big. In the 6th line, the designer correctly estimates the correct variables and therefore optimally designs the groupings.

The gains from machine learning can be seen clearly in the last two lines of Table 1. The 5th line presents a scenario in which we can accurately separate individual into the two risk types. Based on this prior information, we then form optimal groups based on risk. Low risk (.1%) populations are put in groups of 34 and high risk (5%) are in groups of 5. Here machine learning has lowered the expected number of tests to $.24n$. Finally, the 6th line captures the ability to not only estimate infection risk (from outside of the setting) but also an accurate estimate of the infection network. This scenario is the equivalent of a machine learning model that can recover the structure of the data generating process in our model. In this case, group sizes are slightly larger, because we are able to account for correlated risk using the strategy outlined in Section 4, and the expected number of tests falls to $.18n$. Both scenarios capturing the information produced by machine learning represent a large gain relative to individual testing as well as relative to group testing with a single prior. The combined model (line 6) reflects an improvement of 67% to more than 300% over conventional group testing with a single prior, depending on which prior we are considering.

5.2 Machine Learning to Predict COVID-19 Risk

Risk prediction is a specific case of a 'prediction policy problem' (Kleinberg et al. (2015)), in the sense that the likelihood of a negative (positive) group test is an important driver of the value of group testing. We do not need causal estimates of why someone is at high or low risk – just accurate predictions. In general, predicting risk has two central components: estimate individual risk as a function of observable characteristics and the underlying transmission network for infections in the population. We address each in turn.[12]

To estimate an individual risk profile we must develop a prediction of the probability they are infected (i.e. a positive test result) at time t based on available data, captured by a matrix X. This problem lends itself

[12]The discussion is general and we do not provide a specific algorithm. Depending on the particular data and setting in which a strategy is implemented different off-the-shelf approaches can be applied productively.

naturally to the core focus of a supervised machine learning problem: predict the outcome as a function of the rich observable data without over-fitting. In practice, many settings where a testing policy is being developed will be able to collect data such as a population demographics, home locations, or even geo-coded travel data from a mobile device or known contacts where technology enabled contact tracing tools are implemented. The power of machine learning tools for this kind of problem are the ability to estimate non-parametric models that can generate a risk estimate for an individual at a specific point in time as a function of the joint density of independent variables. In the case of time specific COVID-19 infection risk, such non-linearities seem likely to be important. Take, for example, data on the home zip code of an individual. While zip code alone is certainly informative, interactions with other observable factors such as age, travel or consumption patterns are likely to improve model performance in terms of prediction for day-specific risk (e.g. living in a location with high prevalence increases ones risk but more so if a person frequently eats out and socializes and even more so if they did so in the last two days).

The second empirical artifact we are interested in is the infection network. This problem can be approached in two ways, depending critically on data and on prevalence in the population. An infection network could be estimated directly as a supervised problem by trying to predict actual infections observed. However, in many cases this will present a challenge. With relatively low infection rates and networks that are inherently high dimensional due to the the large number of ways in which individuals interact we are particularly concerned about sparsity. Therefore, we propose a second strategy that relies on an unsupervised approach to measure proximity in terms of contact between individuals as a function of high-dimensional distance measure. If we have data on n individuals on the z ways in which they might interact, the goal is to reflect proximity by estimating a z dimensional distance between them. Individuals who are closer on this measure are more likely to interact and, therefore, infect one another. We define an infection structure matrix (N) for a population where each (i, j) element is the distance between individual i and j.

Numerous data sources could be used to inform the network but the central goal is to identify variables that are plausible measures of physical contact. For example, physical layout data combined with time and individual observables can be used to identify how people who live in a building or work together at a factor may come into contact close enough to transmit an infection (e.g. work on the same shift, sharing a lunch break and being in close proximity on an assembly line).

To predict risk we propose combining the overall prediction model that relies on structured individual observables (X) and the infection network matrix (N). Importantly, N need not be estimated as a supervised problem. Instead, the output of the unsupervised network model can simply be included in the overall prediction problem.

5.3 Using Test Collection to Recover Transmission Networks

As we have discussed, the specifics of the empirical approach are likely to be setting specific and will require different refinements based on data availability and the underlying interactions in a job setting (e.g. a factory is potentially very different from a job in closed offices). We do want to note a specific opportunity to recover the infection network. Regardless of data availability, and particularly in settings in which data are limited, the act of testing itself may allow us to estimate the infection network both cheaply and with potentially high predictive performance.

The simple idea is that the order and location of a test being administered can shed light on proximity. One could, in fact, design the testing approach to accomplish this (e.g. instead of having a testing center a nurse walks around gathering tests and geo-coding and time stamping them).

Consider, for example, the person doing testing simply walked down a hallway to collect samples from people in the same workplace or apartment, the ordering of samples naturally encodes the physical proximity of the members of the testing population, a plausible proxy for interaction and infection correlation. Alternatively, consider a setting in which factory workers are tested as they enter a factory through a specific entrance at a specific time. If samples were collected at the door the particularly shift structure of the factory is encoded by location and time stamps. Consequently, combining people who were sampled around the same time into the same group will likely increase the within-group correlation and therefore the efficiency of group testing.

6 Discussion and Practical Considerations

To construct transparent results, we study a stylized environment with a number of important assumptions. We assume, amongst other things, that testing populations are large, group sizes are unconstrained, tests have perfect sensitivity and specificity, test results are instantaneous, group testing does not increase testing costs, there is no fixed cost to collect samples and there is known correlation within the testing population. While removing these constraints further complicates the problem, we do not believe that they change our main conclusions.

For example, there is a fear that the sample dilution inherent in group testing will lead to a loss of test sensitivity. We first note that the sensitivity loss of group testing given reasonable group sizes has been shown to be negligible in other domains (Shipitsyna et al. (2007); McMahan et al. (2012)). However, even if there is a loss of sensitivity on a single test, this is counteracted by the large increase in overall sensitivity coming from running a larger number of tests given increased frequency (a similar observation is made for repeated group testing by Litvak et al. (1994)). For example, if group testing leads the sensitivity to drop from 99% to 90% on a single test, sampling x times as frequently will increase overall sensitivity to $1-(0.10)^x$, which is higher than 99% as long as

$x \geq 2$.[13] This conclusion about specificity potentially informs another implementation decisions. For example, mass testing is presumably more feasible given non-nurse-administered saliva testing, which has recently been shown to be effective (Wyllie et al. (2020)). However, as noted above, even if there is a loss of sensitivity on a single test from using saliva, this is counteracted by the large increase in overall sensitivity coming from running a larger number of tests given increased frequency.[14]

We note that our results are not dependant on infeasible group sizes or complicated grouping algorithms. For example, although we allow arbitrarily large group sizes when we calculate the optimal group sizes, we note that our qualitative results continue to hold given presumably feasible (Shental et al. (2020)) group sizes, such as 10 or 20. In our examples and simulations, we use a very simple two-stage Dorfman testing scheme in which people are partitioned into one group and then individually tested if the group sample is positive. This simplicity implies that our efficiency results are in fact underestimated. As discussed above, there are further possible efficiency gains if the testing strategy uses overlapping groups or multi-stage testing. Using these methods would therefore strengthen efficiency, further our conclusion in the benefits of using machine learning to automate this more complicated process, and strengthen the feasibility of and argument for frequent testing.

Of course, the binding constraint to group testing may not be statistical efficiency or cost-effectiveness, but rather logistical factors related to the mechanics of testing. Because grouping specimens is in a gray area with respect to laboratory best practices – and goes against the usual priority to keep samples strictly separated to avoid contamination – laboratories may be reluctant to implement it. This issue can be mitigated to some extent by considering this application of group testing to be about screening or surveillance as inputs to public health decision making – as with the common practice of screening blood bank samples for HIV or HCV – rather than as a replacement for gold-standard clinical testing. Other problems related to the physical challenges of grouping in the lab could be solved with user-oriented front-end software solutions or better sample containers. But overall, the practical challenges should not be trivialized, and could be a major obstacle to implementation. Nevertheless, the magnitude of the gains in efficiency and reduced infection suggest it is worth overcoming these barriers in fighting a pandemic such as COVID-19.

7 Conclusions

The combination of high-frequency testing with machine learning predictions and knowledge of risk correlation (e.g., in workplaces or facilities) is more plausible and cost-efficient than previously thought. While a formal

[13] Of course, this particular formula depends on the independence assumption of tests. Suppose instead that there is extremely high correlation, such that the probability of a false positive of a group given a previous false positive of that group is 50%. Even then, group testing 4-5 times as frequently will recover the same false positive rate as individual testing.

[14] We also note that testing parameters must include induced behavior responses: if tests are physically unpleasant such that people will potentially avoid getting tested, the test sensitivity is effectively 0%.

cost-effectiveness analysis is beyond the scope of this paper, we estimate that, given marginal test cost of $50, daily group testing could be offered for between $3-5 per person per day, depending on infection prevalence. This makes high-frequency, intelligent group testing a powerful new tool in the fight against COVID-19, and potentially other infectious diseases.

References

Aprahamian, Hrayer, Douglas R Bish, and Ebru K Bish, "Optimal risk-based group testing," *Management Science*, 2019, *65* (9), 4365–4384. 1, 2, 3

_ , Ebru K Bish, and Douglas R Bish, "Adaptive risk-based pooling in public health screening," *IISE Transactions*, 2018, *50* (9), 753–766. 1, 3

Behets, Frieda, Stefano Bertozzi, Mwamba Kasali, Mwandagalirwa Kashamuka, L Atikala, Christopher Brown, Robert W Ryder, and Thomas C Quinn, "Successful use of pooled sera to determine HIV-1 seroprevalence in Zaire with development of cost-efficiency models.," *AIDS (London, England)*, 1990, *4* (8), 737–741. 1

Bilder, Christopher R and Joshua M Tebbs, "Pooled-testing procedures for screening high volume clinical specimens in heterogeneous populations," *Statistics in medicine*, 2012, *31* (27), 3261–3268. 2

_ , _ , and Peng Chen, "Informative retesting," *Journal of the American Statistical Association*, 2010, *105* (491), 942–955. 2

Black, Michael S, Christopher R Bilder, and Joshua M Tebbs, "Group testing in heterogeneous populations by using halving algorithms," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 2012, *61* (2), 277–290. 2

_ , _ , and _ , "Optimal retesting configurations for hierarchical group testing," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 2015, *64* (4), 693–710. 2

Cahoon-Young, B, A Chandler, T Livermore, J Gaudino, and R Benjamin, "Sensitivity and specificity of pooled versus individual sera in a human immunodeficiency virus antibody prevalence study.," *Journal of Clinical Microbiology*, 1989, *27* (8), 1893–1895. 1

Dodd, RY, EP Notari IV, and SL Stramer, "Current prevalence and incidence of infectious disease markers and estimated window-period risk in the American Red Cross blood donor population," *Transfusion*, 2002, *42* (8), 975–979. 1

Dorfman, Robert, "The detection of defective members of large populations," *The Annals of Mathematical Statistics*, 1943, *14* (4), 436–440. 1, 5

Du, Dingzhu, Frank K Hwang, and Frank Hwang, *Combinatorial group testing and its applications*, Vol. 12, World Scientific, 2000. 1

Feng, Jiejian, Liming Liu, and Mahmut Parlar, "An efficient dynamic optimization method for sequential identification of group-testable items," *IIE Transactions*, 2010, *43* (2), 69–83. 1

Gaydos, Charlotte A, "Nucleic acid amplification tests for gonorrhea and chlamydia: practice and applications," *Infectious Disease Clinics*, 2005, *19* (2), 367–386. 1

Hogan, Catherine A, Malaya K Sahoo, and Benjamin A Pinsky, "Sample pooling as a strategy to detect community transmission of SARS-CoV-2," *Jama*, 2020, *323* (19), 1967–1969. 3

Hourfar, Michael Kai, Anna Themann, Markus Eickmann, Pilaipan Puthavathana, Thomas Laue, Erhard Seifried, and Michael Schmidt, "Blood screening for influenza," *Emerging infectious diseases*, 2007, *13* (7), 1081. 1

Hwang, FK, "A generalized binomial group testing problem," *Journal of the American Statistical Association*, 1975, *70* (352), 923–926. 1, 2

Kim, Hae-Young, Michael G Hudgens, Jonathan M Dreyfuss, Daniel J Westreich, and Christopher D Pilcher, "Comparison of group testing algorithms for case identification in the presence of test error," *Biometrics*, 2007, *63* (4), 1152–1163. 3

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer, "Prediction policy problems," *American Economic Review*, 2015, *105* (5), 491–95. 12

Lakdawalla, Darius, Emmet Keeler, Dana Goldman, and Erin Trish, "Getting Americans back to work (and school) with pooled testing," *USC Schaeffer Center White Paper*, 2020. 1

Li, Tongxin, Chun Lam Chan, Wenhao Huang, Tarik Kaced, and Sidharth Jaggi, "Group testing with prior statistics," in "2014 IEEE International Symposium on Information Theory" IEEE 2014, pp. 2346–2350. 1

Litvak, Eugene, Xin M Tu, and Marcello Pagano, "Screening for the presence of a disease by pooling sera samples," *Journal of the American Statistical Association*, 1994, *89* (426), 424–434. 3, 14

McMahan, Christopher S, Joshua M Tebbs, and Christopher R Bilder, "Informative dorfman screening," *Biometrics*, 2012, *68* (1), 287–296. 2, 14

Phatarfod, RM and Aidan Sudbury, "The use of a square array scheme in blood testing," *Statistics in Medicine*, 1994, *13* (22), 2337–2343. 3

Quinn, Thomas C, Ron Brookmeyer, Richard Kline, Mary Shepherd, Ramesh Paranjape, Sanjay Mehendale, Deepak A Gadkari, and Robert Bollinger, "Feasibility of pooling sera for HIV-1 viral RNA to diagnose acute primary HIV-1 infection and estimate HIV incidence," *Aids*, 2000, *14* (17), 2751–2757. 1

Saraniti, Brett A, "Optimal pooled testing," *Health care management science*, 2006, *9* (2), 143–149. 1

Shental, Noam, Shlomia Levy, Shosh Skorniakov, Vered Wuvshet, Yonat Shemer-Avni, Angel Porgador, and Tomer Hertz, "Efficient high throughput SARS-CoV-2 testing to detect asymptomatic carriers," *medRxiv*, 2020. 1, 3, 15

Shipitsyna, Elena, Kira Shalepo, Alevtina Savicheva, Magnus Unemo, and Marius Domeika, "Pooling samples: the key to sensitive, specific and cost-effective genetic diagnosis of Chlamydia trachomatis in low-resource countries," *Acta dermato-venereologica*, 2007, *87* (2), 140–143. 14

Sobel, Milton and Phyllis A Groll, "Group testing to eliminate efficiently all defectives in a binomial sample," *Bell System Technical Journal*, 1959, *38* (5), 1179–1252. 1, 3

Sterrett, Andrew, "On the detection of defective members of large populations," *The Annals of Mathematical Statistics*, 1957, *28* (4), 1033–1036. 3

Tebbs, Joshua M, Christopher S McMahan, and Christopher R Bilder, "Two-stage hierarchical group testing for multiple infections with application to the infertility prevention project," *Biometrics*, 2013, *69* (4), 1064–1073. 2

Wang, Dewei, Christopher S McMahan, Joshua M Tebbs, and Christopher R Bilder, "Group testing case identification with biomarker information," *Computational statistics & data analysis*, 2018, *122*, 156–166. 3

Wyllie, Anne Louise, John Fournier, Arnau Casanovas-Massana, Melissa Campbell, Maria Tokuyama, Pavithra Vijayakumar, Bertie Geng, M Catherine Muenker, Adam J Moore, Chantal BF Vogels et al., "Saliva is more sensitive for SARS-CoV-2 detection in COVID-19 patients than nasopharyngeal swabs," *Medrxiv*, 2020. 15

Yelin, Idan, Noga Aharony, Einat Shaer-Tamar, Amir Argoetti, Esther Messer, Dina Berenbaum, Einat Shafran, Areen Kuzli, Nagam Gandali, Tamar Hashimshony et al., "Evaluation of COVID-19 RT-qPCR test in multi-sample pools," *MedRxiv*, 2020. 3