Optimal Pool Size for COVID-19 Group Testing

Jeffrey Y. Chen and Andrew S. C. Chen

EduTron Corporation

stoshi715@gmail.com

April 20, 2020

Abstract

This paper presents an analytical formulation for determining optimal pool size in the initial pooling stage and the subsequent retests for COVID-19. A generalized constant compaction approach confirms the efficiency of "halving" targeted population between retest stages. An analytical gain formula is derived to aid future test designs. It is observed that optimal gain relies on the proper choice of the initial pool size. This optimal compaction scheme outperforms the conventional algorithms in most cases and may provide a mathematically-native road map for us to operate beyond the standard super-even-number-based (64, 32, 16, 8…, 1) group testing algorithms.

Introduction

Group (or pooled) testing for a disease (or other attribute) involves testing "pools" of individual samples. If a pool tests negative, then all samples therein are assumed negative; if it tests positive, then further pooled or individual tests ensue. Group testing was first studied in the combinatorial context in 1962, with the introduction of *a multi-stage algorithm*[1,2]. The game of improving group testing efficiency is largely driven by the prevalence[1*] rate. The strong appeal of pooled testing is that it can significantly reduce the number of tests and associated costs when the prevalence for a disease is small. The 2020 COVID-19 pandemic calls for timely extensive testing for a large populace. The lack of available tests aggravates the situation around the world, and hence this study. At the time of writing, Germany and Israel have reported group testing efforts. This study aims to develop a strategy for choosing the optimal pool size and compaction factor for each stage to further improve the testing efficiency.

Formulation

The main parameters we will use in the model include: p_i: prevalence at stage i, b_i: pool size for stage i, n_i: the number of individual samples to be tested in stage i, and N: total number of people in the target population to be tested. It is obvious that $n_1 = N$—the full population is targeted in stage 1. The probability of a particular pool tested negative in stage i is $(1-p_i)^{b_i}$. The number of pools tested negative in stage 1 is $\frac{n_1}{b_1}(1-p_1)^{b_1}$. The number of pools tested positive in stage 1 is $\frac{n_1}{b_1}[1-(1-p_1)^{b_1}]$. The number of people tested positive after stage i is

[1] * Prevalence is the percentage of persons in a population who have a particular disease (or attribute) at a specified point in time (or over a specified period of time).

therefore $n_1[1-(1-p_1)^{b_1}]$. Recall that n_1 $becomes$ $n_1[1-(1-p_1)^{b_1}]$ in stage 1. It is productive to define a *Compaction Factor* $\epsilon_1$for stage 1:

$$\epsilon_1 = \frac{n_1[1-(1-p_1)^{b_1}]}{n_1}$$

$$\epsilon_1 = 1-(1-p_1)^{b_1}$$

In general, the compaction factor in stage *i* is

$$\epsilon_i = [1-(1-p_i)^{b_i}], \qquad 0 \le \epsilon_i \le 1. \tag{1}$$

As a result of a compaction ϵ_i in stage *i*

$$n_{i+1} = n_i \cdot \epsilon_i \tag{2}$$

$$p_{i+1} = p_i/\epsilon_i \tag{3}$$

Pooling in a given stage reduces the size of the population to be tested and increases the prevalence in the next stage. The task here is that given p_i, one must choose a good combination of pool size b_i and an effective compaction factor $\epsilon_i$to drive next stage compaction.

The critical parameters ϵ_i depends on b_i as shown in (1). To see how b_i depends on ϵ_i, solve for b_i using (1).

$$(1-p_i)^{b_i} = [1-\epsilon_i]$$

$$log(1-p_i)^{b_i} = log(1-\epsilon_i)$$

$$b_i = \frac{log(1-\epsilon_i)}{log(1-p_i)} \tag{4}$$

For our region of interest, $0 \le p_i \ll 1$, Eq (4) can be approximated with

$$b_i \sim \frac{log(1-\epsilon_i)}{-p_i} \tag{5}$$

Recall that the goal of group testing is to identify all positive individuals with as few total tests as possible. We need to examine the number of tests required. The number of tests required in stage 1 is $t_1 = \frac{n_1}{b_1}$. The number of tests required in stage 2 is $t_2 = \frac{n_2}{b_2}$, and so on. Denote the total number of stages as *S*. The compaction continues thorough *S*-1 stages until the prevalence rate p_S is so large that it is more effective to just test each sample individually, i.e., set the pool size $b_S = 1$, and $t_S = \frac{n_S}{b_S} = n_S$.

The total number of tests required is $T = \sum_{i=1}^{S} \frac{n_i}{b_i}$. The figure of merit we used to guide the study is the gain g defined as

$$g \equiv \frac{N}{T}$$

More specifically,

$$g(\epsilon_i, b_i) = \frac{n_1}{\sum_{i=1}^{S} \frac{n_i}{b_i}} \tag{6}$$

Computer simulations were used to scan through parameter space $\{p_1, \epsilon_i, b_i\}$ to hunt for effective compacting algorithms. The main results are summarized in Figure 1.

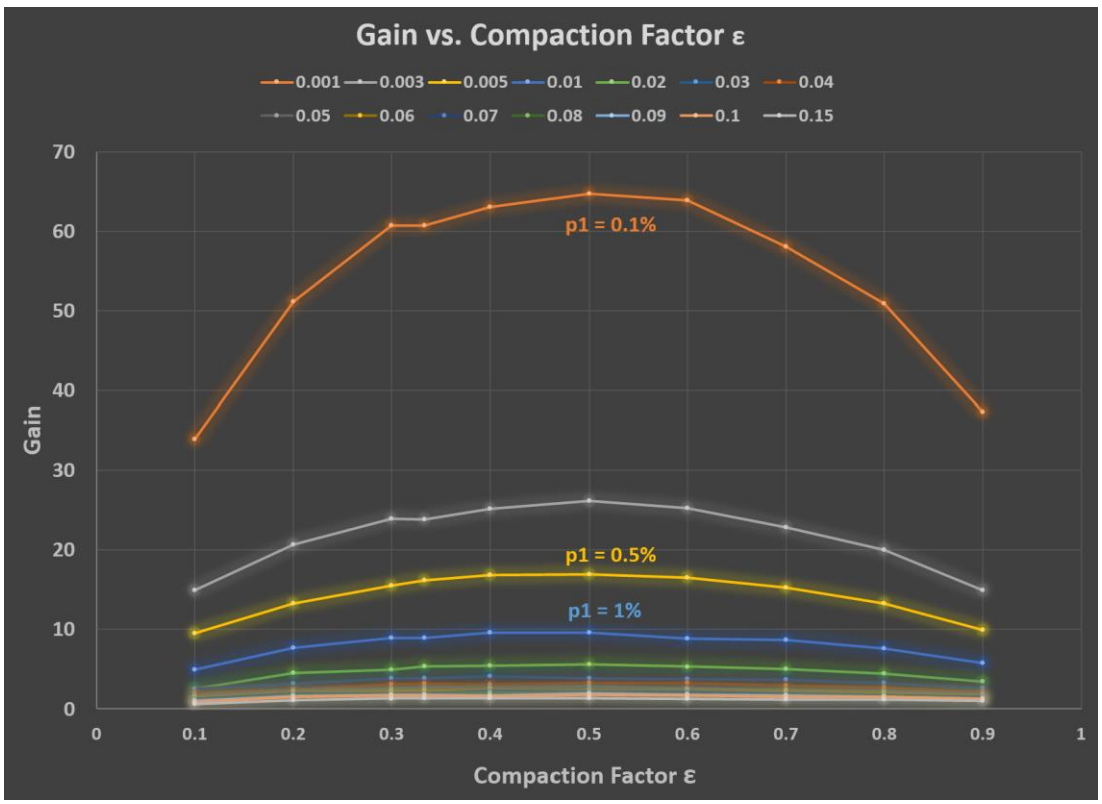

Figure 1. Computer simulation was used to conduct a parameter scan through the "pool size"-"compaction" space $\{b_i, \epsilon_i\}$ to look for high gains--given various values of prevalence p_1. The results revealed a consistent pattern. The best gains are obtained with a constant compaction ϵ_i of 0.5 for all stages, except for the last stage where there is no pooling. The smoothness of the curve, in reality, is modulated by the ceiling function in the gain formula (Eq 16).

Simulation results conclusively indicated that the best strategy for choosing $\{\epsilon_i, b_i\}$ to maximize the gain g is to use a constant compaction of 0.5 in stages 1 through S-1.

$$\epsilon_i = 0.5, \quad i = 1 \; to \; S - 1$$

Deploying all ϵ_i with an identical value of 0.5 is a clear indication that this algorithm is structurally a close cousin of the standard 64-32-16-8 "halving" algorithms. The distinctive feature of the approach in this study is the continuous optimization of pool size b_i. It determines an optimal pool size at every stage that deviates freely from the super even sequence of "2-4-8-16…" Such an approach regularly outperforms the standard group testing algorithms based on the conventional super-even sequence.

The results of two sample runs are summarized in Table 1: (1) p = 1% and (2) p = 2%. A population size of 1,000,000 is used. These are the only two input parameters: p and N.

Table 1. Comparing Gains of Optimal Pool Size Approach with Conventional Approach.

Prevalence	Optimal Pool Size Approach	Conventional Approach
2%	5.48	5.24
1%	9.451	9.449

Table1shows the 2%-prevalence simulation results of a gain of 5.48 using the Optimized Pool Size approach developed here. In contrast, a conventional approach achieved a gain of 5.24. The 1%-prevalence simulation shows a gain of 9.451 using the Optimized Pool Size approach and the conventional approach has a gain of 9.449.

Tables 2 and 3 compare details of the 2%-prevalence simulation runs. Tables 4 and 5 compare details of the 1%-prevalence simulation runs.

Table 2. A gain of 5.48 is achieved for a sample run using the Optimized Pool Size algorithm developed in this study. Input: 2% prevalence, N = 1,000,000. Total Tests Required: 182,619

stage	population	prevalence	pool size	compaction	population for the next stage	prevalence for the next stage	tests required
i	n_i	p_i	b_i	ε	n_{i+1}	p_{i+1}	n_i/b_i
1	1,000,000	2.0%	**34**	0.497	496863	4.0%	**29,412**
2	496863	4.0%	**17**	0.503	249745	8.0%	**29,227**
3	249745	8.0%	**8**	0.487	121662	16.4%	**31,218**
4	121662	16.4%	**4**	0.512	62346	32.1%	**30,416**
5	62346	32.1%	**1**	1.000	0	NA	**62,346**

Table 3. A gain of 5.24 is achieved for a sample run using the conventional algorithm. Input: 2% prevalence, N = 1,000,000. Total Tests Required: 190,975

stage	population	prevalence	pool size	compaction	population for the next stage	prevalence for the next stage	tests required
i	n_i	p_i	b_i	ε	n_{i+1}	p_{i+1}	n_i/b_i
1	1000000	2.0%	16	0.276	276202	7.2%	62,500
2	276202	7.2%	8	0.452	124821	16.0%	34,525
3	124821	16.0%	4	0.503	62744	31.9%	31,205
4	62744	31.9%	1	1.000	0	NA	62,744

Table 4. A gain of 9.451 is achieved for a sample run using the Optimized Pool Size algorithm developed in this study. Input: 1% prevalence, N = 1,000,000. Total Tests Required: 105,809

stage	population	prevalence	pool size	compaction	population for the next stage	prevalence for the next stage	tests required
i	n_i	p_i	b_i	ε	n_{i+1}	p_{i+1}	n_i/b_i
1	**1000000**	**1.0%**	69	0.500	500163	2.0%	14,493
2	500163	2.0%	34	0.497	248455	4.0%	14,711
3	248455	4.0%	17	0.503	124876	8.0%	14,615
4	124876	8.0%	8	0.487	60832	16.4%	15,609
5	60832	16.4%	4	0.512	31173	32.1%	15,208
6	31173	32.1%	1	1.000	0	NA	31,173

Table 5. A gain of 9.449 is achieved for a sample run using the conventional algorithm. Input: 1% prevalence, N = 1,000,000. Total Tests Required: 105,831.

stage	population	prevalence	pool size	compaction	population for the next stage	prevalence for the next stage	tests required
i	n_i	p_i	b_i	ε	n_{i+1}	p_{i+1}	n_i/b_i
1	1000000	1.0%	64	0.474	474404	2.1%	15,625
2	474404	2.1%	32	0.494	234481	4.3%	14,825
3	234481	4.3%	16	0.502	117731	8.5%	14,655
4	117731	8.5%	8	0.508	59856	16.7%	14,716
5	59856	16.7%	4	0.519	31046	32.2%	14,964
6	31046	32.2%	1	1.000	0	NA	31,046

In the following, we will construct an analytical model for the gains in testing efficiency based on a constant compaction factor ϵ.

Constant Compaction Model

The Constant Compaction Factor will be denoted as ϵ whose value is chosen to be 0.5 based on the simulation results. A common compaction ϵ in all compaction stages leads to the following.

$$n_{i+1} = n_i \cdot \epsilon \tag{7}$$

$$p_{i+1} = p_i / \epsilon \tag{8}$$

$$b_i = \frac{log(1-\epsilon)}{log(1-p_i)} \tag{9}$$

$$b_i \sim \frac{log(1-\epsilon)}{-p_i} = \frac{log(0.5)}{-p_i} \tag{10}$$

The optimal pool size at every stage depends on the prevalence at that stage. The dependence is shown in Figure 2.

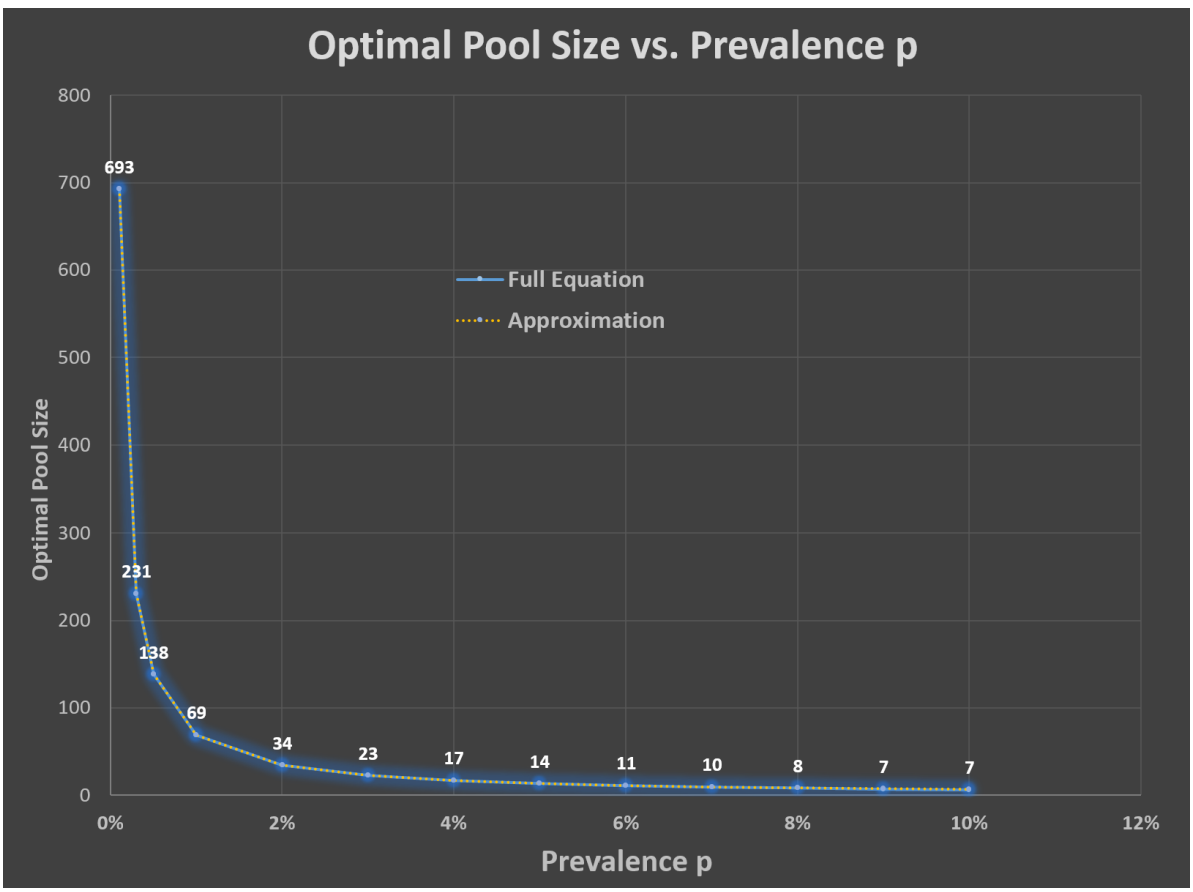

Figure 2. Optimal pool size is plotted against prevalence. The approximated values and the exact values agree well. See Eqs (9) and (10).

It is worth noting that Eq (10) is an almost perfect approximation for Eq (9) in our operating range $0 \leq p_i \ll 1$. The error is $O({p_i}^2)$. It could be used as a simple rule of thumb when one wants to determine the optimal pool size quickly.

$$\boldsymbol{Optimal\ Pool\ Size = \frac{0.693}{Prevalence}}$$

$$\boxed{b_i \sim \frac{0.693}{p_i}}$$

Furthermore, the constant compaction leads to some desirable simplifications. Consider the number of tests required in each stage. In stage 1, $t_1 = \frac{n_1}{b_1}$, in subsequent retest stages,

$$t_i = \frac{n_i}{b_i} = \frac{n_1 . \epsilon^{i-1}}{b_1 \cdot \epsilon^{i-1}} = \frac{n_1}{b_1} = \frac{N}{b_1} = t_1, \qquad 2 \leq i \leq S-1 \tag{11}$$

The fact that ***each retest stage requires the same number of tests as the initial stage*** (except for the last non-pooling stage, *S*) simplifies the derivation of the gain in closed form. To calculate the gain, we will need to know the number of compaction stages *S-1*. The very last ("test-every-remaining-sample") stage is triggered by a *threshold in prevalence.*

At the end of each stage *i* near the end, with a population of n_i remaining to be tested and an expected prevalence of p_i, a decision will be made regarding whether to (1) test everyone individually and conclude the test, or (2) keep pooling in the next stage and perhaps wrap up in the stage after that. (1) ***Stop pooling and test everyone***. The required number of tests to conclude the whole test will be n_i. (2) ***Keep pooling***. The required number of tests to conclude the test will be the *number of pooled tests for the next stage + population remained after that stage.* Define $a \equiv 0.693$

$$\frac{n_i}{\frac{a}{p_i}} + n_i \cdot \left(1 - (1-p_i)^{\frac{a}{p_i}}\right)$$

We will stop pooling and test everyone if

$$n_i < \frac{n_i}{\frac{a}{p_i}} + n_i \cdot \left(1 - (1-p_i)^{\frac{a}{p_i}}\right)$$

The condition simplifies to

$$1 < \frac{1}{\frac{a}{p_i}} + \left(1 - (1-p_i)^{\frac{a}{p_i}}\right)$$

$$(1-p_i)^{\frac{a}{p_i}} < \frac{1}{\frac{a}{p_i}} \tag{12}$$

The inequality can be solved graphically for p_i, yielding $0.3034 < p_i$.

One can also use binomial expansion to solve the inequality analytically with a good approximation.

$$1 - p_i \cdot \frac{a}{p_i} + \frac{\frac{a}{p_i} \cdot \left(\frac{a}{p_i} - 1\right)}{2!} \cdot {p_i}^2 + \cdots < \frac{1}{\frac{a}{p_i}}$$

$$1 - a + \frac{a^2}{2} - \frac{a}{2} \cdot p_i < \frac{p_i}{a}$$

$$1 - a + \frac{a^2}{2} < (\frac{1}{a} + \frac{a}{2}) \cdot p_i$$

$$\frac{1 - a + \frac{a^2}{2}}{\frac{1}{a} + \frac{a}{2}} < p_i$$

Plugin $a = 0.693$,

$$0.305 < p_i$$

This analytical approximation is very close to the *threshold in prevalence* value 0.3034 obtained from graphical solution of Eq (12). In the following, when the prevalence of the next stage is more than 0.3, or 30%, then we test everyone and conclude the test. Now we are ready to calculate the number of stages required.

$$\frac{p_1}{\epsilon^{i-1}} \geq 0.3$$

$$p_1 \geq 0.3 \cdot \epsilon^{i-1}$$

$$\log\left(\frac{p_1}{0.3}\right) \geq (i-1) \cdot \log(\epsilon)$$

$$\frac{\log\left(\frac{p_1}{0.3}\right)}{\log(\epsilon)} \geq i - 1$$

So, the total number of stages with pooling, not counting the very last stage which every sample is tested, is $S-1$

$$S-1=\left[\frac{\log\left(\frac{p_1}{0.3}\right)}{\log(\epsilon)}\right]_{ceiling} \tag{13}$$

The ceiling function in the equation means rounding up to an integer for obtaining an integer number of stages. As a result, the smoothness of the overall gain formula, in reality, is modulated by the ceiling function. The effect can be seen in Figures 1 and 3.This is also true for the rounding required in Eq (4) for obtaining integer optimum pool size.

The number of tests in the last stage is

$$n_S=N\cdot\epsilon^{S-1} \tag{14}$$

Combining all the required tests for all stages using Eqs (11) and (14), we get the total number of tests

$$T=\text{Number of tests per stage}\ \cdot\text{Number of stages}+\text{Number of tests in the last stage}$$

$$T=\left(\frac{N}{b_1}\right)\cdot(S-1)+\ N\cdot\epsilon^{S-1}$$

Recalling the gain

$$g\equiv\frac{N}{T}$$

$$g=\frac{N}{\left(\frac{N}{b_1}\right)\cdot(S-1)+\ N\cdot\epsilon^{S-1}}$$

$$\boxed{\boldsymbol{g}=\frac{1}{\left(\frac{1}{\boldsymbol{b_1}}\right)\cdot(\boldsymbol{S}-\mathbf{1})+\ \boldsymbol{\epsilon}^{\boldsymbol{S}-\mathbf{1}}}} \tag{15}$$

This simple-looking formula for gain elegantly embodies the group-testing mechanism. It captures the dependence of gain $\boldsymbol{g}$ on the choice of the initial pool size $\boldsymbol{b_1}$**.** Note: $\frac{1}{b_1}$ sets the number of tests per stage in subsequent pooled stages. Further note: $b_1$is determined by p_1**.** The number of pooled-test stages $\boldsymbol{S}-\mathbf{1}$ is solely determined by the prevalence p_1(See Eq 13). Above all, all the elegance and efficiency comes from a constant compaction factor $\boldsymbol{\epsilon}$.

A full expression for the gain is

$$g(p_1) = \frac{1}{\left(\frac{log(1-p_1)}{log(1-\epsilon)}\right) \cdot \left[\frac{\log\left(\frac{p_1}{0.3}\right)}{\log(\epsilon)}\right]_{ceiling} + \epsilon^{\left[\frac{\log\left(\frac{p_1}{0.3}\right)}{\log(\epsilon)}\right]_{ceiling}}} \tag{16}$$

The ϵ in the equation is a placeholder for 0.5 in the model. It is clearly seen in the equation that the whole multi-stage group testing optimization through constant compaction is driven by only one critical parameter—the initial expected prevalence p_1. High gains in group testing efficiency can be expected for $p_1 \ll 1$. The dependence of gain on prevalence is shown in Figure 3.

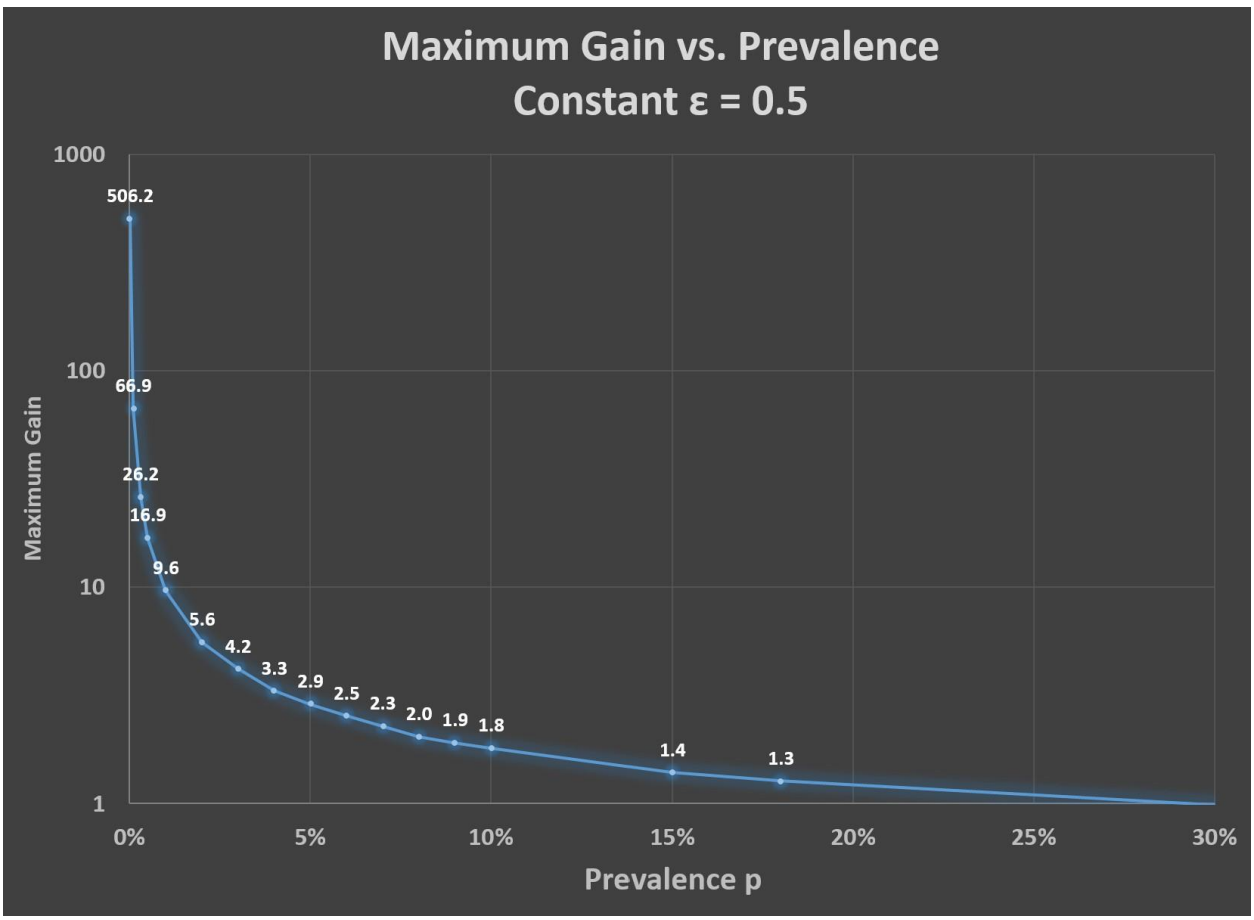

Figure 3. Dependence of maximum gain in group testing efficiency on Prevalence. The smoothness of the curve is perturbed by the ceiling functions in the gain formula. See Eq 16. The high gain regime of operation requires a prevalence lower than 1%. Smart screening and heterogeneous grouping can be used to accomplish high gains. See the example in the next section.

Shifting of Best Compaction (ϵ) in High Prevalence (p_i) Regime

The ceiling function in the gain formula introduces important modifications. Figure 4 shows the gain dependence on compaction for these scenarios. The cliff-like drop on the peak gain can be as large as 11% when p = 2%. The ϵ also appears to shift towards 0.4 from 0.5for larger prevalences. These modifications need to be addressed in actual implementation of the optimum pool size approach when the prevalence is ≥ 2%.

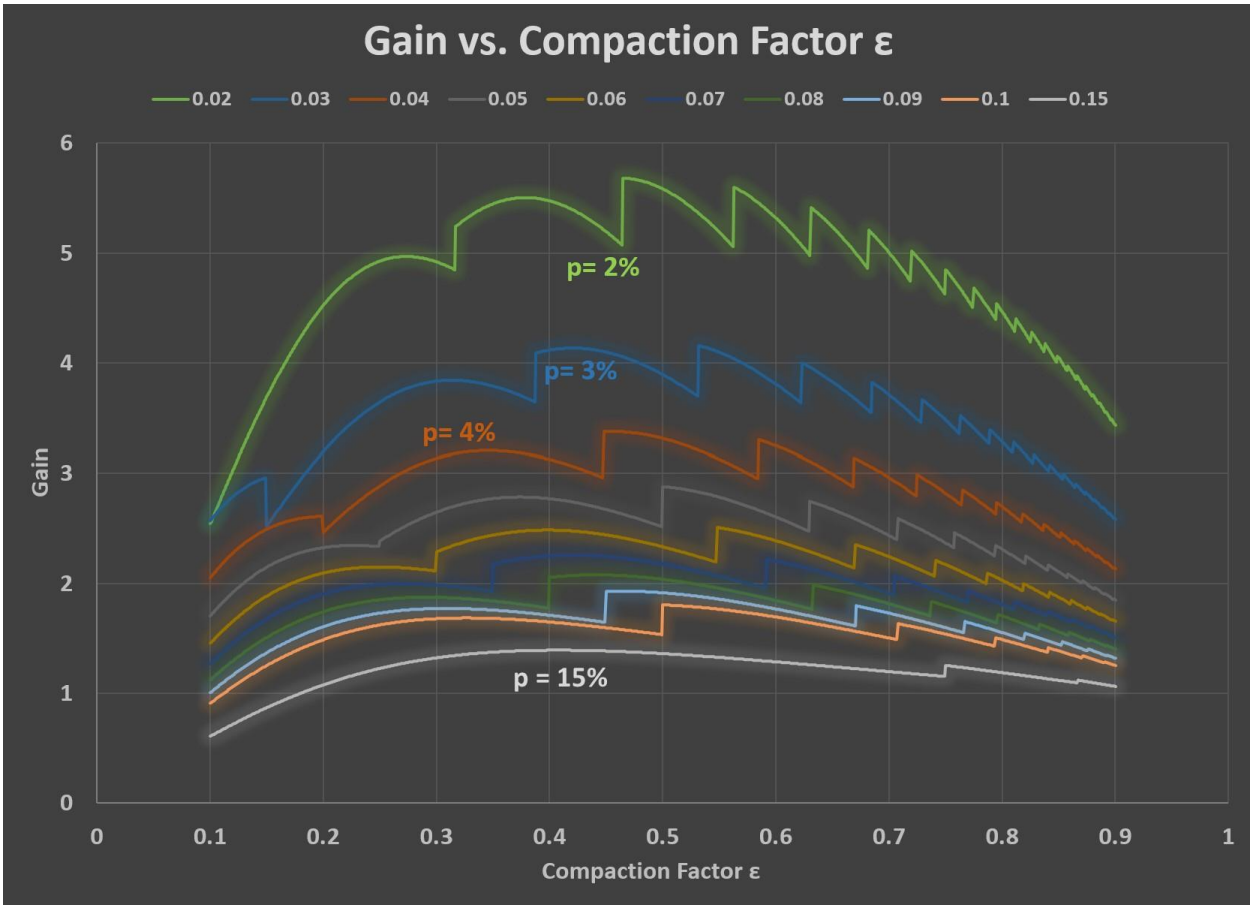

Figure 4. The ceiling function in the gain formula induces a peak gain shift from 0.5 to 0.4 when prevalence is higher than 2%. Application of the optimal pool size approach will need to take this effect into consideration.

Discussions and Policy Implications

The potential high gain region of operation can reduce overall testing costs and time. In order to take full advantage of low prevalence, it is essential to conduct smart screening to accomplish a hybrid group testing by forming heterogeneous subgroups. Smart grouping calls for risk assessments. The risks can be measured in a number of ways traditionally. Some recent techniques involve using artificial intelligence, in which a training data set of individual diagnoses and corresponding risk factors are used to establish a binary regression model. This model can then be applied to the current individuals being screened in order to estimate their risk of having the disease.

As an example, when the whole target population is divided into low risk (p = 0.1%), medium risk (p = 1%), and high risk (p = 20%) groups, a gain of 64.7 can be expected for the low risk group and a gain of 9.6 can be expected for the medium risk group. A composite gain of 30 is accomplished in this example. See Table 6.

Table 6. An example illustrating a composite gain of 30. With a population of 1,000,000, the number of test required is 33,578.

	Group %	Prevalence	# of people	Tests Required
Low Risk	90%	0.10%	900,000	14,055
Medium Risk	9%	1%	90,000	9,523
High Risk	1%	20%	10,000	10,000

This idealized mathematical study focuses narrowly on improving the group testing efficiencies. Other important and compelling issues related to testing—such as sensitivity, specificity and technical logistics, have not been discussed here. They are very relevant and deserve to be studied in detail. In addition, the unconventional choices of pool size in this approach dictate that future robotic pipetting will need major development to accommodate cross-pool mixing arrangement during retest stages. These advanced robots also have to handle rows and columns of non-super-even numbers as well as super-accurate low-dose dispensing.

The approach studied here, deep down, is isomorphic to the conventional multi-stage “halving” algorithms based on super even numbers. It is believed that the initial choice of pool size is very relevant in retaining a high gain in testing efficiency. Using the usual 16 (or 64) to start group testing, though pleasingly simple, may cost us in efficiency.

References

1. Li, Chou Hsiung (June 1962). "A sequential method for screening experimental variables". Journal of the American Statistical Association. **57** (298): 455–477. *doi*:*10.1080/01621459.1962.10480672*
2. Ding-Zhu, Du; Hwang, Frank K. (1993). Combinatorial group testing and its applications. Singapore: World Scientific.