The Effect of Correlation and False Negatives in Pool Testing Strategies for COVID-19

Leonardo J. Basso, Vicente Salinas, Denis Sauré, Charles Thraves
University of Chile, lbasso@ing.uchile.cl, vicentesalinas@ing.uchile.cl, dsaure@dii.uchile.cl, chtraves@dii.uchile.cl

Natalia Yankovic
University of Los Andes, nyankovic.ese@uandes.cl

During the current COVID-19 pandemic, active testing has risen as a key component of many response strategies around the globe. Such strategies have a common denominator: the limited availability of diagnostic tests. In this context, pool testing strategies have emerged as a means to increase testing capacity. The efficiency gains obtained by using pool testing, derived from testing combined samples simultaneously, vary according to the spread of the SARS-CoV-2 virus in the population being tested. Motivated by the need for testing closed populations, such as long-term care facilities (LTCFs), where significant correlation in infections is expected, we develop a probabilistic model for settings where the test results are correlated, which we use to compute optimal pool sizes in the context of two-stage pool testing schemes. The proposed model incorporates the specificity and sensitivity of the test, which makes it possible to study the impact of these measures on both the expected number of tests required for diagnosing a population and the expected number and variance of false negatives. We use our experience implementing pool testing in LTCFs managed by SENAMA (Chile's National Service for the Elderly) to develop a simulation model of contagion dynamics inside LTCFs, which incorporates testing and quarantine policies implemented by SENAMA. We use this simulation to estimate the correlation of test results among collected samples when following SENAMA's testing guidelines. Our results show that correlation estimates are high in settings representative of LTCFs, which validates the use of the proposed model for incorporating correlation in determining optimal pool sizes for pool testing strategies. Generally, our results show that settings in which pool testing achieves efficiency gains, relative to individual testing, are likely to be found in practice. Moreover, the results show that incorporating correlation in the analysis of pool testing strategies both improves the expected efficiency and broadens the settings in which the technique is preferred over individual testing.

Key words: Pool Testing, COVID-19, Simulation, Beta-Binomial

1. Introduction

Motivation and Background. SARS-CoV-2, the virus that causes COVID-19, has put virtually all of the world's health systems under stress. While multiple strategies have been adopted by different governments to contain the spread of the virus, contact tracing appears as a common factor among them. In this context, the containment policies proposed to date typically consider tracing and testing all *close contacts* of confirmed or suspected COVID-19 patients (CDC 2020). Close contact refers to people who have been "within 6 feet of an infected person for at least 15 minutes starting from 2 days before illness onset (or, for asymptomatic patients, 2 days prior to specimen collection) until the time the patient is isolated" CDC (2020). In Chile, tracing policies require identifying all people who have been in close contact with a confirmed COVID-19 patient, from two days before the onset of the patient's symptoms to 14 days after the onset of symptoms. For asymptomatic patients, the tracing period extends to 14 days after testing positive for COVID-19 (Diario Oficial 2020).[1] The close contact definition naturally includes people living under the same roof on a daily basis, which is the case of those living in LTCFs.

Because infectiousness in patients arises before the onset of symptoms and because of the large fraction of asymptomatic cases, when following contact tracing guidelines, it is reasonable to expect a high degree of **correlation** of test results within certain populations (such as LTCFs) at the moment of sample collection.

Another common factor among the strategies adopted by various governments is the need to manage scarce resources. In particular, many countries either have struggled or continue to struggle to ramp up testing capacity, and many patients in need of testing have to wait days to be diagnosed. In these situations, where test availability is limited, *pool testing* arises as a strategy to significantly increase testing capacity. In a pool testing strategy, individual samples are pooled into a single sample, which is tested as if it belongs to a single patient: if the test result is negative, it is assumed that all patients in the pool would have obtained a negative result when tested individually; on the other hand, if the test result is positive, it is concluded that at least one of the patients would have obtained a positive result when tested individually. If the goal of the testing strategy is to diagnose every patient, then only a positive pool result requires additional tests in order attain such goal. To illustrate this approach, consider a pool of 10 patients whose samples are combined into a single sample: a negative result implies a saving of 90% of tests (relative to individual testing), including savings in associated reagents and process times. To the extent that negative results are frequent (we provide a characterization in Section 3), pool testing appears as an effective method to expand testing capacity. The technique has in fact been used in the implementation of massive testing initiatives in several countries (WSJ 2020) and has been widely adopted by Chile's Ministry of Health (Ministerio de Salud 2020).

The question of how to pool samples to diagnose a population has been addressed extensively in the last 80 years, with works differing mainly in the degree of sophistication allowed to the pooling strategy. In general, optimal strategies often involve multiple adaptive stages, which are very difficult to find (its calculation is extremely expensive in terms of computational resources)

[1] At the timing of writing, confirmation of COVID-19 cases is conducted exclusively using polymerase chain reaction (PCR) tests, which detect unique sequences of the virus RNA via nucleic acid amplification.

and extremely complex to implement due to their adaptive nature. As a result, simple two-stage schemes have been widely adopted in practice; see Section 2.

Implementing pool testing strategies for fighting the pandemic requires considering a series of practical issues. First, collecting nasopharyngeal samples (necessary for conducting the most commonly used PCR testing technique) involves a rather unpleasant procedure that requires qualified health personnel, so repeated sampling should be avoided. Samples provide enough material for two tests, hence the importance of narrowing down to two-stage pool testing strategies.

Second, correlation of test results has been observed when testing populations such as LTCFs (which have exhibited a large number of deadly outbreaks of COVID-19, particularly in Spain and Italy as stated in Kluge, Hans H. (2020)). However, traditional pool-testing models typically assume independence of test results and often ignore the risk of false negatives (Woloshin et al. 2020). The effect of correlation of test results in such recommendations is not well understood, nor is it its effect on the variance of false negatives when testing is conducted using two-stage pool testing strategies.

Objectives. In this work we explore the implications of correlation of test results for the implementation of pool testing strategies in the context of active testing of populations. For this purpose, we propose a probabilistic model for the *prevalence*[2] of COVID-19 in a closed population that allows for correlation of the test results among individuals. The model includes as input the operating parameters (*specificity* and the *sensitivity*) of the testing technique (see Section 3 for definitions).

In this regard, the proposed model builds upon traditional models, which assume independence of test results between individuals in the population (leading to a binomial distribution for the total number of infections), and introduces a correlation structure by incorporating randomness in the patients' probability of testing positive. The resulting model allows to study, for example, the effect of correlation in optimal pool sizes and in the variance of false negatives.

Complementing the above, we develop a simulation model for the contagion dynamics within a closed population. The model is capable of incorporating different interaction dynamics between subgroups of the population (for example, residents and staff of an LTCF) as well as replicating different policies of care and quarantine. The simulation model allows us to produce correlation estimates under different care/testing strategies, thus making it possible to quantify the degree of correlation in practical settings so that optimal pool sizes can be computed.

Contributions. Our first contribution lies in extending the traditional model of prevalence within a population to settings with correlation of test results, thus enabling the computation of optimal pool sizes in practical settings as well as other analyses not covered by the literature in pool testing.

[2] In this context, prevalence is understood as the number of positive test results that follow from testing a particular population.

This is the case, for example, of the risk of false negatives, an analysis that is of interest even under the independence assumption of traditional models. In this regard, we show the existence of a trade-off between the pool size used and the variance in the number of false negatives, a relationship that is exacerbated with higher degrees of correlation in test results. Our analysis indicates that setting the pool size to minimize the expected number of tests used might lead to undesirable outcomes: depending on the sensitivity of the test, the "optimality" of the pool size might rely too strongly on an initial false negative result. This is undoubtedly a factor to consider when deciding on a pool testing strategy, making the analysis that we present not only novel and sensitive but also of quite relevant to decision makers.

In addition to the above, our work contributes to showing that it is plausible, in practice, to find scenarios with high correlation of test results when testing relatively closed populations, as in the case of LTCFs. In particular, our analysis shows how the prevalence and correlation observed at the time of testing LTCFs depend on the risky interactions between members of the community, quarantine and testing policies. In a prescriptive way, our analysis can be used to design/enable active testing campaigns, aimed to prevent outbreaks while using scarce resources efficiently.

Article structure. The rest of the paper is organized as follows. In Section 2, we review the relevant literature. Then, in Section 3, we present the probabilistic model for the prevalence, which we use to compute performance measures associated with different pool sizes under a two-stage pool testing strategy. Section 4 presents our results, including potential savings in the number of tests and the risk in the number of false negatives. Section 5 illustrates the application of our model in the context of the LTCFs managed by SENAMA, and Section 6 presents our conclusions. The details of our analytical results are relegated to Appendix A.

2. Literature Review

The work on pool testing can be classified according to its treatment of the prevalence, which can be probabilistic or combinatorial: in the probabilistic case, an underlying probabilistic model on the number of positive test results is set; and in the combinatorial case, it is assumed that there is a (possibly unknown) set of individuals that would test positive. Considering the application area at hand, in this section, we review the literature on the probabilistic model. See Knill (1995) and the references therein for a guide on the combinatorial case.

The seminal work by Dorfman (1943) presents the original two-stage strategy for detecting syphilis-infected blood samples in the US military. The underlying (probabilistic) model there considers independence of the test results across individuals and perfect sensitivity of the test, so false negatives do not exist. The analysis of such a model provides theoretical support for the technique by showing that the expected number of tests necessary to diagnose a population is much lower than that associated with individual testing when the prevalence is low.

Building upon Dorfman (1943), numerous works have explored alternative multiple-stage strategies. For example, Sterrett (1957) and Gill and Gottlieb (1974) propose using additional stages where subpools are formed and tested whenever a pool tests positive. Generally, pool testing strategies differ on how the pool tested in each stage is formed, and in that regard, they can be classified as adaptive or nonadaptive. In the adaptive strategies, successive pools are dependent on the results of previously performed tests (thus, the strategy proposed by Dorfman (1943) is adaptive, with two stages). As a problem of sequential decision-making under uncertainty, finding the best adaptive strategy can be achieve, theoretically, using dynamic programming; see, for example, Wein and Zenios (1996). In the nonadaptive strategies, the series of pools to be tested in each stage is defined prior to knowing any intermediate test results. In this context, Hwang (1975) considers settings with heterogeneous population. Recently, Aprahamian et al. (2019) extend the latter work to settings with imperfect test parameters, while Aprahamian et al. (2020*b*) analyze such a setting with emphasis on the implementation of a strategy. For a more detailed literature review on nonadaptive strategies see Balding et al. (1996) and Aprahamian et al. (2020*a*).

The novel coronavirus pandemic has drawn significant attention to the analysis of pool testing strategies, as authorities struggle to increase testing capacity. For example, the work by Mentus et al. (2020) revisits the analysis of multistage pool strategies for diagnosing COVID-19 patients, and Noriega and Samore (2020) propose a Bayesian inference scheme to estimate optimal pool sizes. Closer to our analysis, Cherif et al. (2020) uses simulation to evaluate the efficiency of the method, using the model by Dorfman (1943) but incorporating the test operating parameters; however, neither correlation of test results nor the risk of false negatives are quantified. More recently Mutesa et al. (2020) present a non-adaptive testing scheme, and put special attention to dilution effects.

Because the work on the subject is dynamic and growing, we do not attempt to provide a comprehensive summary here.

Regarding the evidence of the validity of using pool testing for diagnosing COVID-19 patients, to the best of our knowledge, Yelin et al. (2020) is the first study to validate the procedure using PCR-based testing: they showed that it is possible to pool up to 32 samples without modifying the testing protocol. In Chile, the method has been validated for sizes of up to 10 samples by Farfan et al. (2020), whose results were independently replicated by various laboratories. Currently, Chile's Ministry of Health guidelines call for using pool testing broadly.

3. Mathematical Model

Consider the problem of diagnosing a population of N patients using a two-stage pool testing strategy. Most literature on pool testing assumes that each patient tests positive independently with probability $p \in (0,1)$ (which also denotes the prevalence in the population) and use a binomial distribution to model the total number of individual positive test results. Instead, we assume that there is correlation in the test results of any two individuals, which we denote by $\rho \in [0,1)$.

Note that we restrict our attention to nonnegative correlation under the logic that in the applications of interest, a positive test for one individual increases the possibility of testing positive for the other individuals.

Formally, we define the results from testing the population using the vector $X := \{X_1, \ldots, X_N\}$, where we define

$$X_i := \begin{cases} 1 & \text{if patient } i \text{ tests positive when tested individually,} \\ 0 & \sim, \end{cases} \qquad i \leq N.$$

We assume that, given a value $q \in (0,1)$, X_i is distributed *Bernoulli*(q) (that is, $\mathbb{P}\{X_i = 1\} = q$) for $i \leq N$, and that the sequence $\{X_i,\, i \leq N\}$ is independent and identically distributed. Note that when $\rho = 0$, we have that $q = p$, and the number of positive (individual) results in any group of n patients follows a *Binomial*(n,p) distribution, which is the model of proposed by Dorfman (1943). In the sequel, when $\rho > 0$, we assume that q is a random variable distributed *Beta*(α,β) where

$$\alpha = p\,(1/\rho - 1)\,, \quad \beta = (1-p)\,(1/\rho - 1)\,. \tag{1}$$

(While both α and β depend on the probability p and the correlation ρ, we omit such dependency to streamline the exposition.) Our modeling choice has two important consequences: first, the randomness in q introduces a correlation in the test results of the population; and second, the distribution of the number of positive (individual) results in a group of size n follows a *BetaBinomial*(n,α,β) distribution. The following lemma formalizes these properties. While these results are well known, we include their proof in Appendix A for sake of completeness.

Lemma 1. *Consider a set of patients $M \subseteq \{1,\ldots,N\}$ and define $X(M) := \sum_{i \in M} X_i$, the number of patients in M that would test positive if tested individually. We have that $X(M) \sim$ BetaBinomial$(|M|,\alpha,\beta)$, i.e.,*

$$\mathbb{P}\{X(M) = k\} = \binom{|M|}{k} \frac{B(k+\alpha, |M| - k + \beta)}{B(\alpha,\beta)}, \quad k \leq |M|,$$

where $|M|$ denotes the cardinality of the set M, and $B(\cdot,\cdot)$ is the Beta function. Additionally, for $i \neq j$, we have that

$$Corr(X_i, X_j) := \rho, \quad \mathbb{E}\{X_i\} = p, \quad \textit{and} \quad \mathrm{Var}(X_i) = p\,(1-p).$$

It is worth noticing that if a Beta-Binomial distribution is fitted on uncorrelated data, it will result on low correlation values very close to zero. See Appendix F for more details.

Consider the case $\rho > 0$, and let n denote the pool size used in a two-stage pool testing strategy. To compute the expected number of tests to be used under this strategy, we consider the *specificity* and *sensitivity* of the testing technique. Let $S_e \in [0,1]$ denote the probability that a sample from an infected patient indeed tests positive (the sensitivity of the test) and let $S_p \in [0,1]$ denote the probability that a sample from a patient who is not infected indeed tests negative (the specificity of the test). Like in prior work under the independence assumption, we assume that these operating parameters are not affected by the size of the pool and that each test fails the diagnosis independently, even if the same sample is used (in successive tests).

Remark 1. The assumption above on the specificity of PCR-based tests is rather mild since false positives mostly occur due to problems in the handling of the samples. The assumption on the sensitivity of the test is slightly stronger since false negatives occur when, for example, one of the samples included in the pool is very close to but below the detection threshold[3] (in which case the sample, tested individually, tests positive). In this case, sample dilution may occur, which can place the pooled sample slightly above the detection threshold, resulting in the sample being incorrectly labeled as pathogen-free. In practice, however, evidence suggests that it is difficult to find samples close to the detection threshold (Farfan et al. 2020). □

In the sequel, we let T denote the number of tests used to diagnose the entire population and n denote the pool size used. The following Lemma, whose proof is a direct consequence of Lemma 1, provides an expression for the expected value of T.

Lemma 2. *Suppose that N is a multiple of n; then,*

$$\mathbb{E}\{T\} = N\left(\frac{1}{n} + S_e + (1 - S_e - S_p)\frac{B(\alpha, n+\beta)}{B(\alpha,\beta)}\right).$$

When there is no correlation ($\rho = 0$), we recover the result presented in Cherif et al. (2020), that is based on the work presented by Sobel and Groll (1959), which extends the model in Dorfman (1943) to include the specificity and sensitivity of the test. Note that the expression above is very easy to evaluate (the Beta function is built into most statistical software), thus considering that in practice, pool sizes are bounded by above,[4] this expression can be used directly to find optimal pool sizes via enumeration.

[3] The results from PCR tests are based on how many cycles (heating/cooling the sample) are required to amplify the presence of the pathogen to make it detectable; therefore, if such a time (in cycles) is less than a certain threshold, then it is concluded that the result is positive.

[4] In addition to the fact that the technique has so far been validated for pools no larger than 32, consider that in the absence of automated processing technologies, laboratory personnel can handle relatively small pool sizes.

From Lemma 2 we see that for a given pool size, the operating parameters directly affect the expected number of tests (used to diagnose the population). In particular, we note that the higher the specificity of the test is, the lower the expected number of tests. (The specificity of PCR tests is close to 100%.) On the other hand, the effect of sensitivity is the opposite, and the higher the sensitivity is, the greater the expected number of tests. However, we see below that this occurs at the expense of an increase in the risk of obtaining a false negative on the pooled sample. To the best of our knowledge, the effect of a strategy on the variance of false negatives has been omitted in the analysis presented in the extant literature.

Let F_- denote the number of false negatives associated with the diagnosis of the population. The following lemma, whose proof follows from Lemma 1 and can be found in Appendix A, characterizes the expected value and variance of F_-, depending on the operating parameters and the pool size.

Lemma 3. *Suppose that N is a multiple of n; then, $\mathbb{E}\{F_-\} = N(1-S_e^2)$ and*

$$\text{Var}(F_-) = N(1-S_e^2)p - N(1-S_e)(1+S_e-S_e^2-nS_e^3)(p^2+p(1-p)\rho) + N^2(1-S_e^2)^2p(1-p)\rho.$$

Note that while the expected number of false negatives is independent of the pool size, the second part of Lemma 3 shows that as the pool size increases, the variance of F_- also increases. Hence, the larger the pool size is, the greater the risk of false negatives (consider, for example, the extreme case where $n = N$). We explore these issues in the next section.

Let F_+ denote the number of false positive tests in a population of N individuals. The following result gives a closed-form expression for the expectation of F_+.

Lemma 4. *Suppose that N is a multiple of n; then*

$$\mathbb{E}\{F_+\} = N(1-S_p)\left(S_e(1-p) + (1-S_e-S_p)\frac{B(\alpha,\beta+n)}{B(\alpha,\beta)}\right).$$

At an intuitive level, at equal prevalence, (positive) correlation in test results should lead to a reduction in the number of tests necessary to diagnose the population, as an individual negative test result is likely to be accompanied by similar negative tests results for other patients in the pool, which is the favorable scenario for pool testing; on the other hand, an (individual) positive result is likely to be accompanied by more positive results in the pool, however having one or more positive test results requires the same number of tests on a potential second stage. The next result formalizes this intuition.

Proposition 1. *For any pool-size n and prevalence p fixed, if $S_e + S_p \geq (\leq) 1$, then the expected number of tests of a Beta-Binomial model is less (greater) or equal than the one of using a Binomial model.*

Note that for most (if not all) testing techniques available for the case of SARS-CoV-2, one has that $S_e + S_p > 1$, thus one expects to have less tests used under the Beta-Binomial model. From this result, one can conclude that the expected number of tests used under the optimal pool-size for the Beta-Binomial model will be lower or equal than that for the Binomial model. The result, however, does not say much about the relative size of the optimal pool sizes under these models. The next result states that, for the case of ideal operating parameters, the optimal pool size is larger under the Beta-Binomial model.

PROPOSITION 2. *If $S_e = S_p = 1$, then the optimal pool size under the Beta-Binomial model is greater or equal than the one under the Binomial model.*

Note that the operating parameters found in practice, while not perfect, are quite high, thus one would still expect optimal pool sizes when considering correlation in test results to be greater in the case of correlated tests results. We explore this point in our numerical experiments.

4. Results

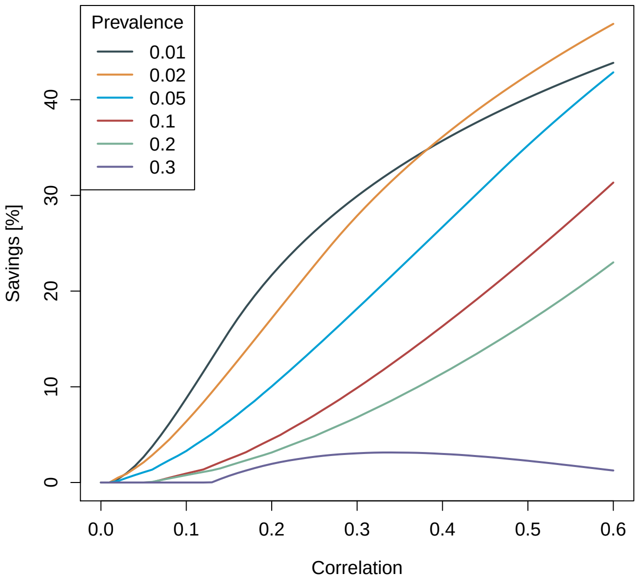
Table 1 presents the results of using the model in a population of one hundred patients ($N = 100$) for prevalence that varies from 0.01% to 40%, considering 4 levels of correlation (0.2, 0.4, 0.6, and 0.8). The optimal pool size, the expected number of tests and the savings in the number of tests compared to the individual testing strategy are included. We also include the savings in the number of tests compared to the pool testing strategy using the group size obtained without the correlation. Additionally, in the event that the test is not perfect and may yield false negatives, the expected value of false negatives and their standard deviation are included. To evaluate the impact of pool testing on the risk of false negatives, we consider $S_e = 0.7$, $S_p = 1$ following the scenarios analyzed in Cherif et al. (2020). We present the expected value and standard deviation of false negatives on the right part of Table 1. In order to include practical implementation issues we limit the pool size to be up to 32 samples ($n \leq 32$).

The savings in the number of tests can be as great as 97% compared to performing individual tests and up to 36% compared to performing pool testing when using the pool size calculated from a model that does not consider correlation. Additionally, it is observed that for low prevalence, the pool sizes are large, that is, equal to the size of the population. The optimal pool size decreases for higher prevalence values when $S_e = 1$. In the case of having imperfect tests ($S_e < 1$), the relation between the optimal pool size and the prevalence might not be decreasing; indeed, we can observe that for low correlations (less than or equal to 0.2), the optimal pool size increases by taking the upper limit value for high prevalence.

Figure 1 shows the savings in the number of tests when using the optimal pool size of the model (which explicitly includes the correlation) versus the case where the correlation is ignored—as a function of the correlation for different levels of prevalence.

Figure 2 shows the optimal pool size as a function of the prevalence for different levels of correlation. It can be seen that optimal pool testing strategy in high prevalence and low correlation scenarios is using the larger possible pool size (i.e., $n = 32$) showing a discontinuity. In these scenarios is unlikely not to have a positive sample within the pool, but if the test has a sensitivity lower than one (in this example $S_e = 0.7$), we may have a false negative result that would assign a negative (wrong) status to all the individual samples

Figure 1 **Savings in the expected number of tests considering the correlation explicitly vs pool testing ignoring the correlation for different levels of prevalence** ($N = 100$, $n \leq 32$, $S_e = 0.7$, $S_p = 1$)



We also explore the impact of pool testing in the expected false positives and false negatives when using imperfect tests ($S_p < 1$, $S_e < 1$). Although the expected number of false negatives does not depend on the pool size, its standard deviation is increasing on n (see Lemma 3). Figure 3 shows how the standard deviation of false negatives decreases as the pool size decreases, while the expected number of tests increases. We can see that by using $n = 6$ instead of the optimal pool size ($n^* = 32$), the expected number of tests is increased by 3.7% (from 72.5 to 75.2), but the standard deviation of the false negatives is decreased by 16.5% (from 10.3 to 8.6).

Figure 2 **Optimal pool size based on prevalence for different correlation levels ($N = 100$, $n \le 32$, $S_e = 0.7$, $S_p = 1$)**

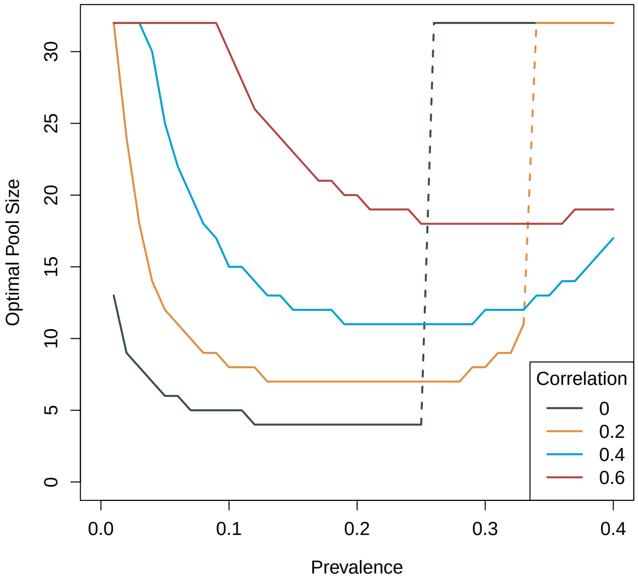


Figure 4 shows the optimal pool size when considering an alternative optimization model taking into account, in addition to the expected number of tests, the expected number of false positives constraining on the standard deviation of false negatives, using the same setting than in Figure 3. Namely,

$$\min_{n\in\mathbb{Z}_+} \lambda\mathbb{E}[T] + (1-\lambda)\mathbb{E}[F_+] \tag{2}$$

$$\text{s.t.}\, \mathrm{Var}(F_-) \le u \tag{3}$$

where $\lambda \in [0,1]$ denotes the relative weight in the objective function between the expected number of tests and the expected number of false positives, and $u > 0$ denotes the upper bound on the variance of false negatives. We can see that considering the expected false positives or the standard deviation of false negatives in the optimization model lead to smaller pool sizes and larger expected number of tests.

5. Case Study: Application of pool testing in a LTCF in Chile

The correlation model of infections presented in Section 3 is motivated by the reality of the LTCFs managed by SENAMA. In these facilities, a group of older adults lives under the care of a team of health professionals. We use a dataset that has test results for a set of LTCFs for 3 months. Because

Figure 3 **Standard deviation of false negatives and expected number of tests for different pool sizes ($N = 100$, $n \leq 32$, $p = 0.3$, $\rho = 0.1$, $S_e = 0.7$, $S_p = 0.9$).**

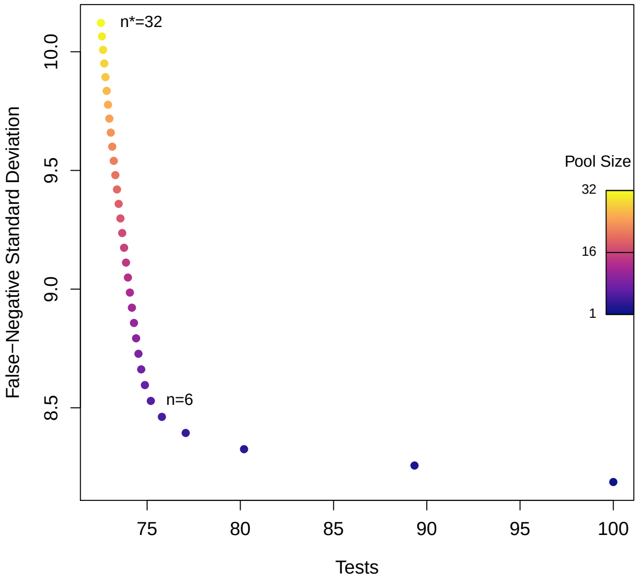


Figure 4 **Optimal pool size for each parameter λ when minimizing the convex combination of the expected number of tests and the number of false positives, i.e. $\lambda\mathbb{E}[T] + (1-\lambda)\mathbb{E}[F_+]$ subject to constraining by above the standard deviation of false negatives ($N = 100$, $n \leq 32$, $p = 0.3$, $\rho = 0.1$, $S_e = 0.7$, $S_p = 0.9$).**

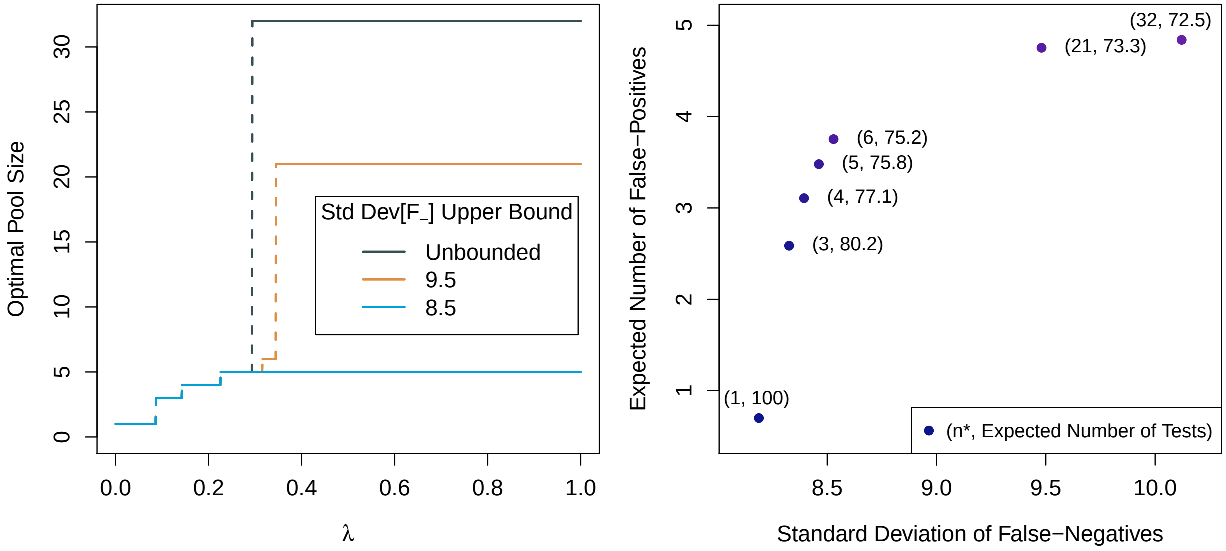


Table 1 Optimal pool sizes, expected number of tests and percentage of savings in relation to the individual testing strategy and pool testing strategy without considering the correlation ($N = 100$, $n \leq 32$). On the left, $S_e = 1$, and on the right, $S_e = 0.7$. The expected number of false negatives and their standard deviation are also reported. $S_p = 1$ in all cases.

			Savings [%]				Savings [%]		False Negatives
Prevalence	n*	Exp. tests	Ind. testing	$\rho = 0$	n*	Exp. tests	Ind. testing	$\rho = 0$	Exp. (Std.)
Correlation = 0									
0.0001	32	3	97	0	32	3	97	0	0 (0)
0.001	32	6	94	0	32	5	95	0	0 (0)
0.01	11	20	80	0	13	16	84	0	1 (1)
0.02	8	27	73	0	9	23	77	0	1 (1)
0.05	5	43	57	0	6	35	65	0	3 (3)
0.1	4	59	41	0	5	49	51	0	5 (5)
0.2	3	82	18	0	4	66	34	0	10 (10)
0.3	3	99	1	0	32	73	27	0	15 (15)
0.4	1	100	0	0	32	73	27	0	20 (20)
Correlation = 0.2									
0.0001	32	3	97	0	32	3	97	0	0 (0)
0.001	32	4	96	0	32	4	96	0	0 (0)
0.01	31	12	88	18	32	9	91	22	1 (1)
0.02	17	19	81	12	24	15	85	17	1 (1)
0.05	9	34	66	7	12	27	73	10	3 (3)
0.1	6	51	49	3	8	40	60	5	5 (5)
0.2	4	73	27	2	7	57	43	3	10 (10)
0.3	4	90	10	1	8	68	32	2	15 (15)
0.4	1	100	0	0	32	72	28	0	20 (20)
Correlation = 0.4									
0.0001	32	3	97	0	32	3	97	0	0 (0)
0.001	32	4	96	0	32	3	97	0	0 (0)
0.01	32	8	92	36	32	7	93	36	1 (1)
0.02	32	13	87	31	32	10	90	36	1 (1)
0.05	18	25	75	22	25	19	81	27	3 (3)
0.1	11	41	59	13	15	31	69	16	5 (5)
0.2	7	62	38	10	11	47	53	11	10 (10)
0.3	7	78	22	6	12	58	42	3	15 (15)
0.4	8	90	10	10	17	66	34	1	20 (20)
Correlation = 0.6									
0.0001	32	3	97	0	32	3	97	0	0 (0)
0.001	32	3	97	0	32	3	97	0	0 (0)
0.01	32	6	94	46	32	5	95	44	1 (1)
0.02	32	9	91	45	32	7	93	48	1 (1)
0.05	32	18	82	39	32	14	86	43	3 (3)
0.1	20	31	69	28	30	23	77	31	5 (5)
0.2	13	50	50	21	20	37	63	23	10 (10)
0.3	12	65	35	15	18	48	52	1	15 (15)
0.4	12	77	23	23	19	56	44	1	20 (20)

people in each facility are tested only a few times during that time lapse, we fit a beta-binomial distribution by aggregating all facilities for each of the three months. This is performed by finding the distribution parameters that maximize the log-likelihood; see Appendix C for more details. Table 2 shows the fitted parameters for each month, the optimal log-likelihood obtained, and the resulting correlations. These results confirm the intuition behind the definition of close contact and the testing recommendations established by the Ministry of Health (Ministerio de Salud 2020) since relevant levels of correlation in infections are observed.

Table 2 For each month, the fitted beta-binomial parameters α and β, the log-likelihood ll, and the prevalence p and correlation ρ.

Month	α	β	ll	p	ρ
June	0.47	11.61	-89.71	0.04	0.08
July	0.20	4.34	-123.84	0.04	0.18
August	0.14	19.29	-48.78	0.01	0.05

To compute the optimal pool testing size, it is necessary to know the correlation of the population at the precise time when performing the testing. For this purpose, we developed a tool that allows simulating the evolution of the infection in an LTCF, in which we can track the number of infected patients on each day in every simulated scenario. Appendix D presents the details of the tool.

In the simulation, we consider two groups of individuals: residents and staff. The simulation begins with the entire population of the LTCF (residents and staff) free of the virus, in which the staff potentially introduces the infection into the facility with a probability that depends on the external prevalence.

A matrix of interactions between the population is defined that specifies the probability that two individuals (staff or residents) come into contact during a shift. The greater the probability of interactions is, the faster the expected spread of the infection is. The probability of daily interaction between any two members of the population is assumed to be fixed (simulations are performed with different values for this probability). In this way, residents can only be infected by interactions with the staff or other residents of the LTCF, while the staff can be infected in their interactions at the facility or exogenously outside of work. The probability of infection given an interaction with an infected individual will depend on the intensity of the interaction and the contagion capacity of the infected patient.

In terms of the epidemiological model, the incubation time is supposed to follow a lognormal distribution (He et al. (2020)), while the infectiousness follows the (scaled) curve of pathogen-detection via PCR testing , which we model after Sethuraman et al. (2020). Specifically, the incubation period t_{inc} follows a lognormal(1.621, 0.418) distribution, the patient's infectiousness starts at $t_{inf} = \min(\text{Uniform}[t_{inc}/3, t_{inc}], t_{inc} - 1)$ (regardless of the symptoms showed), and the recovery time follows a uniform distribution between 2 and 4 weeks $t_{rec} = t_{inc} + \text{Uniform}[14, 28]$; a patient is contagious between t_{inf} and t_{rec}. The infectiousness, incubation period and whether the patient shows symptoms or not are independent variables across individuals. For the simulation, we considered that 30% of the patients are asymptomatic regardless of the group to which they belong (in line with international evidence (Stephen A. Lauer et al. 2020)).

We assume that every symptomatic patient is isolated in preventive quarantine for 14 days from the second day after the onset of the symptoms and does not have the possibility of infecting others while in quarantine. The latter assumption seeks to replicate the reactive testing strategy that has been applied in general by SENAMA, in which those who present symptoms of the disease are selectively tested.

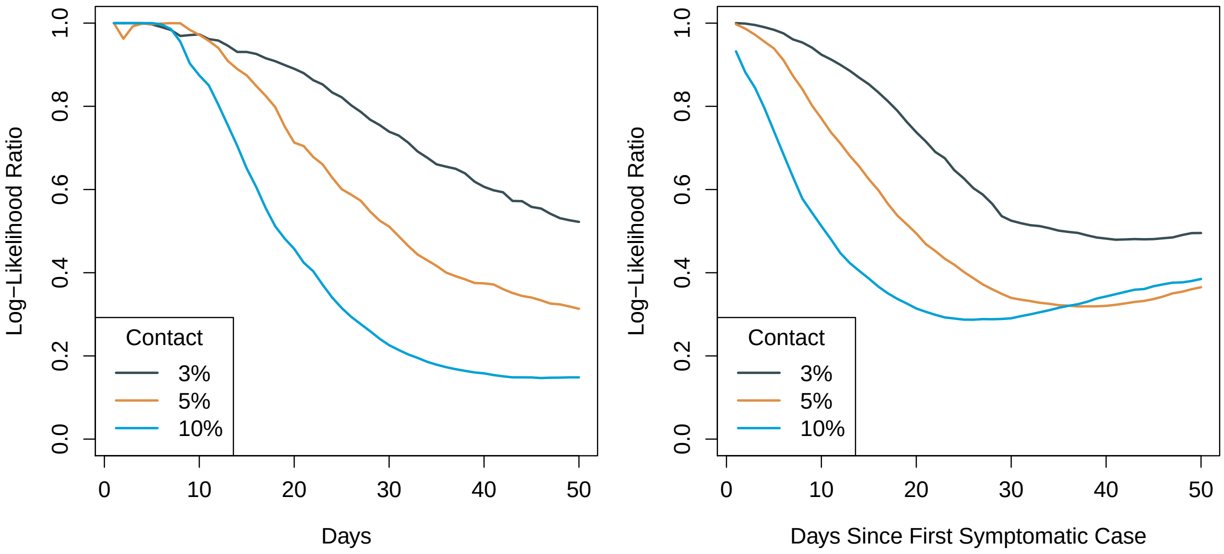
In our case study, we consider an LTCF with 30 residents and 20 employees in two shifts of 10 people each; thus, $N = 50$. The external prevalence considered is 0.1%. The total number of simulated scenarios is 1000.

For each day from the start of the simulation, we fit a beta-binomial distribution by maximizing the log-likelihood considering the infected cases in each simulated scenario. Once the parameters are estimated for each day t, namely, α_t and β_t, we solve Equations (1) to obtain the estimate correlation for each day. This procedure is also performed considering the days with respect to the first symptomatic case for each simulation. In addition to the beta-binomial model, we fitted a binomial distribution by maximizing the log-likelihood function. (Recall that the latter distribution does not allow for correlations.) See Appendix E for more details. The ratio between the log-likelihood values obtained for each day with the beta-binomial model with respect to the binomial model are shown in Figure 5. It can be seen from Figure 5 that the likelihood is significantly better in the beta-binomial probability model, i.e., in the case that incorporates correlation.

Figure 6 presents the evolution of the prevalence and correlation for the total population of the LTCF as a function of the simulation days, while Figure 7 considers the time horizon with respect to the first day an individual shows symptoms. Both graphs are constructed for three levels in the probability of daily interaction (5, 10, and 20%).

In Figure 6, both prevalence and correlation increase with time and with the probabilities of daily interaction. However, when shifting the horizon to the first day on which a patient shows symptoms, the prevalence and correlation increase with time and the probability of daily interaction for several days before decreasing, a feature that is exacerbated for the scenario with a probability of daily interaction at 20%, since most of the population will be infected (or recovered from the infection) after 30 days of the first symptomatic case, as shown in Figure 7.

Figure 5 Log-likelihood ratio between the beta-binomial and binomial models, for probabilities of daily interactions of 5, 10 and 20% ($N = 50$). Left-panel: days since the first day of the simulation; right-panel: days since the first symptomatic case.



Given the time evolution of prevalence and correlation, the epidemiological situation of the population under study may be very different from one day to the next. This fact implies that the recommended pool sizes, if a pool testing strategy is used, will be different depending on the stage in which the population is found.

Tables 3 and 4 present the prevalence, correlation and the pool size recommended by the model presented in Section 3 and the pool size recommended by the model that ignores the correlation. The last column includes the savings in the expected number of tests if including the correlation when defining the pool size, for probabilities of daily interactions of 3, 5 and 10%. This analysis assumes that $S_e = 1$, to prevent the undesirable effect of “betting” on a negative result of the whole group due to the sensitivity of the test, as discussed in Section 4. In the case of Table 3, the days refer to the start of the simulation, while in Table 4, the first day is considered to be the day on which we identify the first symptomatic patient.

It can be seen from Table 3 that the recommended pool size, both considering and not considering correlation, decreases as the days progress. This result is due to the sustained increase in prevalence and correlation. Furthermore, it can be seen that the cases with the highest correlation occur when assuming a higher degree of interaction (10% interaction probability). The consequence of the presented results is that omitting the correlation implies a loss of efficiency in the expected number of tests to be used. For example, it can be seen that after a month of the simulation, using

Figure 6 **Prevalence and correlation in the population for probabilities of daily interactions of 5, 10 and 20% considering the first day of the simulation ($N = 50$).**

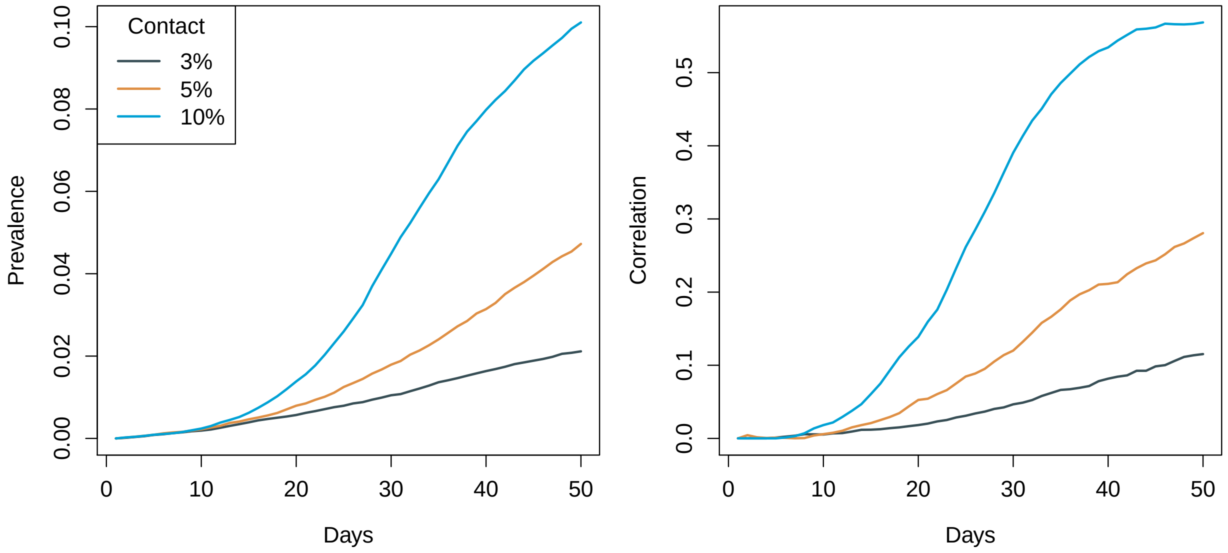
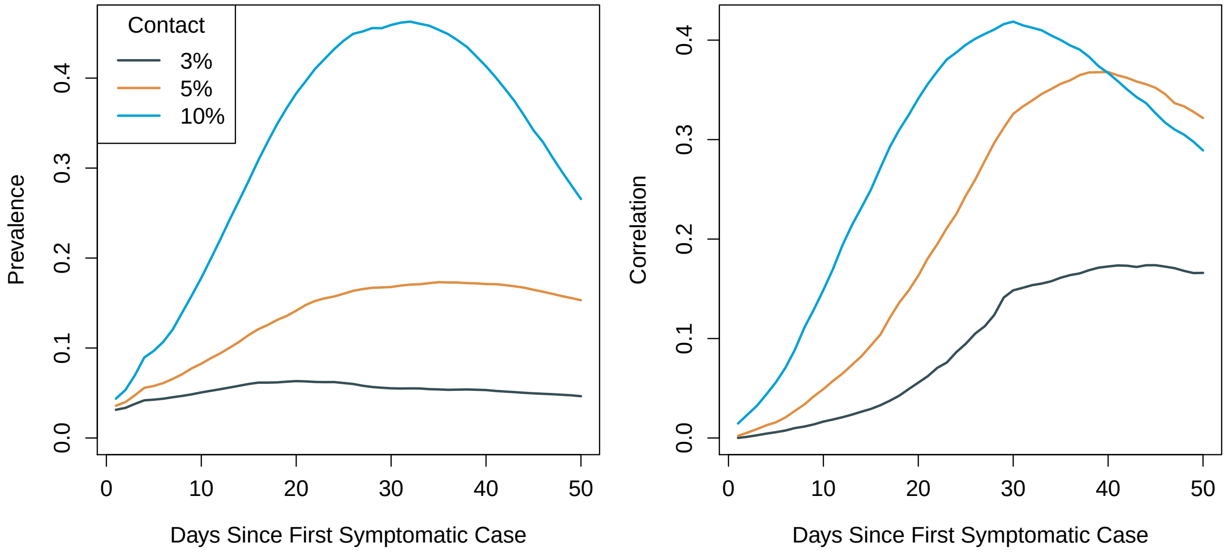


Figure 7 **Prevalence and correlation in the population for probabilities of daily interactions of 3, 5 and 10% considering the first day on which there are symptomatic patients ($N = 50$).**



a testing strategy that considers correlation can contribute savings of 28.4% in the number of tests (versus pool testing ignoring the correlation). When performing the same exercise on the results of Table 4, it is observed that the savings in the expected number of tests are of lesser magnitude than in the previous case. Still, these results should be observed with caution, as the specific degree of

Table 3 For every simulated day: prevalence, correlation, optimal pool size (n^*), optimal pool size without considering the correlation ($n^\diamond$), and savings in the expected number of tests. Probabilities of daily interactions of 3, 5 and 10% from left to right.

Day	Prev.	ρ	n^*	$n^\diamond$	Savings [%]	Prev.	ρ	n^*	$n^\diamond$	Savings [%]	Prev.	ρ	n^*	$n^\diamond$	Savings [%]
5	0.00	0.00	32	32	0.0	0.00	0.00	32	32	0.0	0.00	0.00	32	32	0.0
10	0.00	0.01	23	23	0.0	0.00	0.01	22	22	0.0	0.00	0.02	25	21	1.3
15	0.00	0.01	18	17	0.2	0.01	0.02	18	15	1.1	0.01	0.06	20	13	5.1
20	0.01	0.02	16	14	0.4	0.01	0.05	16	12	2.6	0.01	0.14	17	9	9.8
25	0.01	0.03	14	12	0.9	0.01	0.09	14	9	6.0	0.03	0.26	18	7	15.5
30	0.01	0.05	13	10	2.4	0.02	0.12	13	8	6.4	0.05	0.39	18	5	24.2
35	0.01	0.07	12	9	3.0	0.02	0.18	14	7	9.5	0.06	0.49	20	5	23.0
40	0.02	0.08	12	8	4.5	0.03	0.21	13	6	11.6	0.08	0.53	19	4	28.4
45	0.02	0.10	12	8	3.8	0.04	0.24	12	6	9.5	0.09	0.56	19	4	26.9
50	0.02	0.12	12	7	6.7	0.05	0.28	12	5	14.0	0.10	0.57	18	4	25.0

Table 4 For every simulated day starting with the first symptomatic case: prevalence, correlation, optimal pool size (n^*), optimal pool size without considering the correlation ($n^\diamond$), and savings in the expected number of tests. Probabilities of daily interactions of 3, 5 and 10% from left to right.

Day	Prev.	ρ	n^*	$n^\diamond$	Savings [%]	Prev.	ρ	n^*	$n^\diamond$	Savings [%]	Prev.	ρ	n^*	$n^\diamond$	Savings [%]
5	0.04	0.01	5	5	0.0	0.06	0.02	5	5	0.0	0.10	0.06	4	4	0.0
10	0.05	0.02	5	5	0.0	0.08	0.05	5	4	0.1	0.18	0.15	4	3	1.7
15	0.06	0.03	5	5	0.0	0.12	0.09	4	4	0.0	0.29	0.25	5	3	2.1
20	0.06	0.06	5	5	0.0	0.14	0.16	5	3	3.6	0.38	0.34	7	1	8.3
25	0.06	0.10	6	5	0.6	0.16	0.24	5	3	5.4	0.44	0.40	10	1	6.2
30	0.06	0.15	7	5	3.1	0.17	0.33	6	3	8.4	0.46	0.42	11	1	6.1
35	0.05	0.16	8	5	3.9	0.17	0.36	7	3	9.5	0.45	0.40	10	1	5.5
40	0.05	0.17	8	5	4.7	0.17	0.37	7	3	10.2	0.41	0.37	8	1	6.9
45	0.05	0.17	8	5	5.7	0.17	0.35	7	3	10.0	0.34	0.33	6	1	12.0
50	0.05	0.17	8	5	6.3	0.15	0.32	7	3	9.4	0.27	0.29	5	3	3.3

interaction among individuals of the population under study will imply different correlations and therefore savings.

6. Discussion and Conclusions

This work presents a model for two-stage pool testing that explicitly incorporates correlation in test results and can be used to minimize the expected number of tests. The model is inspired by the progression of the COVID-19 infection in (partially) closed communities, such as LTCFs, where correlation in test results is likely.

In the case of tests with sensitivities less than one, an explicit formula is presented to evaluate the risk in false negatives, which increases with the pool size. This highlights the trade-off between minimizing the expected number of tests versus the risk in the number of false negatives.

To estimate the prevalence and the correlation present in an LTCF, we built a simulation model that allows following the evolution of infected patients by using parameters from the literature and from the policies implemented locally in Chile by the SENAMA. Adjusting a beta-binomial distribution, we estimated the prevalence and correlation and obtained the optimal pool size using the presented model. In this way, we can advise an optimal pool size that considers both the number of days since the start of the simulation (the entire population is healthy) and the number of days after the first patient shows symptoms.

Our analysis characterizes the savings in the number of expected tests needed to diagnose the population when using the optimal pool size recommended by the model versus that recommended by the model that ignores correlation. The savings are significantly more pronounced when observing the simulation data from the beginning of the simulation because of the evolution of the correlation. In addition, our results highlight the importance of testing the LTCF promptly once symptomatic cases have been detected, due to the rapid growth in prevalence in the days immediately following, as illustrated in Figure 7. In the same sense, the timeliness in test results truly makes it possible to manage a preventive quarantine since it is of little utility to test a population that is unable to minimize the risks of contagion while waiting for the results.

These results highlight the importance of having a mechanism that prevents a large outbreak, for example, by frequently performing tests on all members of the population. Indeed, periodic pool testing in closed groups could be recommended every two weeks in our case study: in this way, it would be possible to identify any potential outbreak in time by using a very limited number of tests since this would “restart” the dynamics of the evolution of the infection (as illustrated in Figure 6). For the simulation, we have considered preventive quarantine of all symptomatic cases from the second day of the beginning of symptoms, showing that not even strict quarantines can prevent large outbreaks from occurring if the incubation period is long and there is a significant proportion of asymptomatic cases.

In terms of future research directions, the natural next step is to validate the dynamics of contagions in our simulation model. Once validated, the model can serve as the basis for the evaluation of testing and preventive quarantine strategies, which could include the frequent pool testing of the entire population under study. On the other hand, our model makes a number of assumptions regarding the temporal evolution of the infection and the dynamics of contagion, based on partial evidence collected to date regarding the pandemic. As the knowledge about the virus improves, new and better models of the infection dynamics can be considered and used in our simulation model.

References

Aprahamian, H., Bish, D. R. and Bish, E. K. (2019), 'Optimal risk-based group testing', *Management Science* **65**(9), 4365–4384.

Aprahamian, H., Bish, D. R. and Bish, E. K. (2020*a*), 'Optimal group testing: Structural properties and robust solutions, with application to public health screening', *INFORMS Journal on Computing* **32**(4), 895–911.

Aprahamian, H., Bish, E. K. and Bish, D. R. (2020*b*), 'Static risk-based group testing schemes under imperfectly observable risk', *Stochastic Systems* **10**(4), 361–390.

Balding, D., Bruno, W., Torney, D. and Knill, E. (1996), A comparative survey of non-adaptive pooling designs, *in* 'Genetic mapping and DNA sequencing', Springer, pp. 133–154.

CDC (2020), Contract tracing plan, covid-19.
URL: *https://www.cdc.gov/coronavirus/2019-ncov/php/contact-tracing/contact-tracing-plan/contact-tracing.html*

Cherif, A., Grobe, N., Wang, X. and Kotanko, P. (2020), 'Simulation of pool testing to identify patients with coronavirus disease 2019 under conditions of limited test availability', *JAMA Network Open* **6**(3).

Diario Oficial (2020), Resolución 424 del 9 de junio de 2020, ministerio de salud, gobierno de chile.
URL: *https://www.diariooficial.interior.gob.cl/publicaciones/2020/06/09/42676/01/1771191.pdf*

Dorfman, R. (1943), 'The detection of defective numbers of large populations', *Annals of Mathematical Statistics* **1**(14), 436–440.

Farfan, M., Torres, J., O'Ryan, M., Olivares, M., Gallardo, P. and C., S. (2020), 'Optimizing rt-pcr detection of sars-cov-2 for developing countries using pool testing', *Rev. chil. infectol.* **37**(3), 276–280.

Gill, A. and Gottlieb, D. (1974), 'The identification of a set by successive intersections', *Information and Control* **24**(1), 20–35.

He, X., Lau, E. H., Wu, P., Deng, X., Wang, J., Hao, X., Lau, Y. C., Wong, J. Y., Guan, Y., Tan, X. et al. (2020), 'Temporal dynamics in viral shedding and transmissibility of covid-19', *Nature medicine* **26**(5), 672–675.

Hwang, F. K. (1975), 'A generalized binomial group testing problem', *Journal of the American Statistical Association* **70**(352), 923–926.

Kluge, Hans H. (2020), Statement – older people are at highest risk from covid-19, but all must act to prevent community spread.
URL: *https://www.euro.who.int/en/about-us/regional-director/statements-and-speeches/2020/statement-older-people-are-at-highest-risk-from-covid-19,-but-all-must-act-to-prevent-community-spread*

Knill, E. (1995), Lower bounds for identifying subset members with subset queries, *in* 'Proceedings of the Sixth Annual ACM-SIAM Symposium on Discrete Algorithms', SODA '95, Society for Industrial and Applied Mathematics, USA, p. 369–377.

Mentus, C., Romeo, M. and DiPaola, C. (2020), 'Analysis and applications of adaptive group testing methods for covid-19', *medRxiv* .

Ministerio de Salud (2020), Protocolo de coordinación para acciones de vigilancia epidemiológico durante la pandemia covid-10 en chile: Estrategia nacional de testeo, trazabilidad y aislamiento.
URL: *https://www.minsal.cl/wp-content/uploads/2020/07/Estrategia-Testeo-Trazabilidad-y-Aislamiento.pdf*

Mutesa, L., Ndishimye, P., Butera, Y., Souopgui, J., Uwineza, A., Rutayisire, R., Ndoricimpaye, E., Musoni, E., Rujeni, N., Nyatanyi, T., Ntagwabira, E., Semakula, M., Musanabaganwa, C., Nyamwasa, D., Ndashimye, M., Ujeneza, E., Mwikarago, I., Muvunyi, C., Mazarati, J., Nsanzimana, S., Turok, N. and Ndifon, W. (2020), 'A pooled testing strategy for identifying sars-cov-2 at low prevalence', *Nature* **589**, 276 – 280.

Noriega, R. and Samore, M. H. (2020), 'Increasing testing throughput and case detection with a pooled-sample bayesian approach in the context of covid-19', *bioRxiv* .
URL: *https://www.biorxiv.org/content/early/2020/04/05/2020.04.03.024216*

Sethuraman, N., Jeremiah, S. S. and Ryo, A. (2020), 'Interpreting Diagnostic Tests for SARS-CoV-2', *JAMA* **323**(22), 2249–2251.

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

Stephen A. Lauer, Kyra H. Grantz, Q., Bi et al. (2020), 'The Incubation Period of Coronavirus Disease 2019 (COVID-19) From Publicly Reported Confirmed Cases: Estimation and Application', *Annals of Internal Medicine* **172**(9), 577–582.

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

Wein, L. M. and Zenios, S. A. (1996), 'Pooled testing for hiv screening: capturing the dilution effect', *Operations Research* **44**(4), 543–569.

Woloshin, S., Patel, N. and Kesselheim, A. (2020), 'False negative tests for sars-cov-2 infection - challenges and implications', *New England Journal of Medicine* **38**(383).

WSJ (2020), Wuhan tests nine million people for coronavirus in 10 days.
URL: *https://www.wsj.com/articles/wuhan-tests-nine-million-people-for-coronavirus-in-10-days-11590408910*

Yelin, I., Aharony, N., Shaer-Tamar, E., Argoetti, A., Messer, E., Berenbaum, D., Shafran, E., Kuzli, A., Gandali, N., Hashimshony, T. et al. (2020), 'Evaluation of covid-19 rt-qpcr test in multi-sample pools', *MedRxiv* .

Appendix A: Analytic results

Preliminaries. Before starting, let us consider the case $\rho > 0$ and note that, using (1), it can be shown that

$$p = \frac{\alpha}{(\alpha+\beta)}, \quad \rho = \frac{1}{(\alpha+\beta+1)}.$$

We will use these relationships repeatedly in the remainder of this appendix. □

Proof of Lemma 1. The first part of the lemma is a direct consequence of the definition of a beta-binomial distribution. We have that

$$\begin{aligned}
\mathbb{P}\{X(M)=k\} &= \int_0^1 \mathbb{P}\{X(M)=k|q=x\}\,\frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}dx \\
&= \int_0^1 \binom{|M|}{k} x^k(1-x)^{|M|-k}\frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}dx \\
&= \binom{|N|}{k}\frac{B(\alpha+k,\beta+|M|-k)}{B(\alpha,\beta)}\int_0^1 \frac{x^{\alpha+k-1}(1-x)^{\beta+|M|-k-1}}{B(\alpha+k,\beta+|M|-k)}dx \\
&= \binom{|M|}{k}\frac{B(\alpha+k,\beta+|M|-k)}{B(\alpha,\beta)}.
\end{aligned}$$

In this development, we first condition on the value of q (we use the density of a random variable $Beta(\alpha,\beta)$), and then, we use the fact that, conditional on the value of q, $X(M)$ follows a binomial distribution. We note that the last equality above follows from recognizing the integral of the density of a random variable $Beta(\alpha+k,\beta+|M|-k)$ over its domain.

Regarding the second part of the lemma, we have that

$$\begin{aligned}
\mathbb{E}\{X_i\} &= \int_0^1 \mathbb{E}\{X_i|q=x\}\,\frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}dx \\
&= \int_0^1 x\,\frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}dx = \frac{\alpha}{\alpha+\beta} = p,
\end{aligned}$$

where in the last equality we recognize the expectation of a random variable of distribution $Beta(\alpha,\beta)$ and use the definition of α and β in terms of p and ρ. On the other hand, we have that

$$\begin{aligned}
\mathbb{E}\{X_j\,X_i\} &= \int_0^1 x^2\,x^{\alpha-1}(1-x)^{\beta-1}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\,dx \\
&= \frac{\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)}\int_0^1 x^{\alpha+1}(1-x)^{\beta-1}\frac{\Gamma(\alpha+\beta+2)}{\Gamma(\alpha+2)\Gamma(\beta)}\,dx \\
&= \frac{\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)},
\end{aligned}$$

where we have written $Beta(\cdot,\cdot)$ in terms of the function (gamma) $\Gamma(\cdot)$. With this, we have that

$$\operatorname{Cov}\{X_i,X_j\} = \mathbb{E}\{X_i\,X_j\} - \mathbb{E}\{X_i\}\,\mathbb{E}\{X_j\}$$

$$= \frac{\alpha}{\alpha+\beta}\left(\frac{\alpha+1}{\alpha+\beta+1} - \frac{\alpha}{\alpha+\beta}\right) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} = p(1-p)\rho.$$

Because of the binary nature of X_i, we have that $\mathbb{E}\{X_i^2\} = \mathbb{E}\{X_i\} = \alpha/(\alpha+\beta)$. This implies that

$$\mathrm{Var}(X_i) = \frac{\alpha}{\alpha+\beta}\left(1 - \frac{\alpha}{\alpha+\beta}\right) = \frac{\alpha\beta}{(\alpha+\beta)^2} = p(1-p).$$

Finally, we conclude that for $i \neq j$,

$$\mathrm{Corr}(X_i, X_j) = \frac{\mathrm{Cov}(X_i, X_j)}{\sqrt{\mathrm{Var}(X_i)\,\mathrm{Var}(X_j)}} = \frac{\mathrm{Cov}(X_i, X_j)}{\mathrm{Var}(X_j)} = \rho.$$

This concludes the proof of the lemma. □

Proof of Lemma 2. Let M_k be the set of patients included in pool k to be tested, formed so that $\{M_k,\, k = 1 \ldots N/n\}$ forms a partition of the population, and let T_k denote the number of tests necessary to diagnose patients in pool k. We have that

$$\begin{aligned}
\mathbb{E}\{T\} &= \sum_{k=1}^{N/n} \mathbb{E}\{T_k\} \\
&= \sum_{k=1}^{N/n} \left((1 + n(1-S_p))\,\mathbb{P}\{X(M_k) = 0\} + (1 + n\,S_e)\,(1 - \mathbb{P}\{X(M_k) = 0\})\right) \\
&= \sum_{k=1}^{N/n} 1 + nS_e + n\,(1 - S_e - S_p)\,\mathbb{P}\{X(M_k) = 0\} \\
&= N\left(\frac{1}{n} + S_e + (1 - S_p - S_e)\,\frac{B(\alpha, n+\beta)}{B(\alpha,\beta)}\right).
\end{aligned}$$

The first equality above comes from the linearity of the expectation, the second is from conditioning on the number of patients with the pathogen in pool k, and the last equality is from Lemma 1 and the fact that the number of infections in a pool is distributed equally in each group. □.

Proof of Lemma 3. Following the proof of Lemma 2, let F_{-k} be the number of false negatives obtained when testing the group k. We have that

$$\begin{aligned}
\mathbb{E}\{F_-\} &= \sum_{k=1}^{N/n} \mathbb{E}\{F_{-k}\} \\
&= \sum_{k=1}^{N/n} \sum_{i=1}^{n} \mathbb{E}\{F_{-k} | X(M_k) = i\}\,\mathbb{P}\{X(M_k) = i\}
\end{aligned}$$

$$
\begin{aligned}
&\overset{(a)}{=} \sum_{k=1}^{N/n}\sum_{i=1}^{n}\left(i\,(1-S_e)+i(1-S_e)S_e\right)\mathbb{P}\left\{X(M_k)=i\right\} \\
&= \sum_{k=1}^{N/n}(1-S_e^2)\mathbb{E}\left\{X(M_k)\right\} = \frac{N}{n}(1-S_e^2)\,n\frac{\alpha}{\alpha+\beta} = N(1-S_e^2)\,p.
\end{aligned}
$$

We observe that in (a) above, we use the fact that when $X(M_k)=i$, if the pool test results in a false negative (which occurs with probability $(1-S_e)$), this results in i false negatives, and when the pool test gives the correct result (which happens with probability S_e), this results, on average, in $(1-S_e)i$ false negatives coming from the individual tests. The last equality above uses the fact that the expectation of a distributed random variable $BetaBinomial(k,\alpha,\beta)$ is $k(\alpha/(\alpha+\beta))$.

Now, consider calculating the variance of F_-. First, let us note that conditional on q, the false negatives in each pool are independent random variables, so we have that

$$
\mathbb{E}\left\{F_-^2|q\right\} = \sum_{k=1}^{N/n} E\left\{F_{-k}^2|q\right\} + \frac{N}{n}\left(\frac{N}{n}-1\right)\left(n\,q\,(1-S_e^2)\right)^2.
$$

To develop the term associated with each pool, we remember that if $X\sim Binomial(n,q)$, then

$$
E\left\{X^2\right\} = \mathrm{Var}(X) + \mathbb{E}\left\{X\right\}^2 = nq(1-q)+(nq)^2.
$$

We proceed using the fact that, conditional on $X(M_k)=i$ and that the pool test did not fail, the number of false negatives obtained in the group k follows a $Binomial(i,(1-S_e))$ distribution. Let G_k denote the event that the test of group k does not fail; we have that

$$
\begin{aligned}
E\left\{F_{-k}^2|q\right\} &= \sum_{i=1}^{n}\left((1-S_e)\mathbb{E}\left\{F_{-k}^2|X(M_k)=i,\,\bar{G}_k\right\} + S_e\mathbb{E}\left\{F_{-k}^2|X(M_k)=i,\,G_k\right\}\right)\mathbb{P}\left\{X(M_k)=i|q\right\} \\
&= \sum_{i=1}^{n}\left((1-S_e)\,i^2 + S_e\left(i\,S_e(1-S_e)+i^2(1-S_e)^2\right)\right)\mathbb{P}\left\{X(M_k)=i|q\right\} \\
&= S_e^2(1-S_e)\,n\,q + (1-S_e)(1+S_e(1-S_e))\left(q(1-q)n+n^2q^2\right) \\
&= \left(S_e^2(1-S_e)+(1-S_e)(1+S_e(1-S_e)\right)\,n\,q - ((1-S_e)(1+S_e(1-S_e))\,q^2n(1-n) \\
&= (1-S_e^2)\,n\,q - ((1-S_e)(1+(1-S_e)\,S_e)\,q^2n(1-n).
\end{aligned}
$$

Then, we have that

$$
\mathbb{E}\left\{F_-^2|q\right\} = N\,(1-S_e^2)\,q - ((1-S_e)(1+(1-S_e)\,S_e)\,q^2(1-n) + N^2q^2\left((1-S_e^2)\right)^2 - Nnq^2\left((1-S_e^2)\right)^2.
$$

Then, we note that

$$\mathbb{E}\left\{q^2\right\} = \mathrm{Var}(q) + \mathbb{E}\left\{q\right\}^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} + \frac{\alpha^2}{(\alpha+\beta)^2} = p(1-p)\rho + p^2.$$

Finally, taking the expectation (with respect to q) over $\mathbb{E}\left\{F_-^2|q\right\}$, and subtracting $\mathbb{E}\left\{F_-\right\}^2$, we have

$$\begin{aligned}
\mathrm{Var}\left\{F_-\right\} &= N\,(1-S_e^2)\,p - N^2(1-S_e^2)^2p^2 \\
&\quad - \Big(((1-S_e)(1+(1-S_e)\,S_e)\,(1-n) + N^2\left((1-S_e^2)\right)^2 - Nn\left((1-S_e^2)\right)^2\Big)\left(p(1-p)\rho + p^2\right) \\
&= N(1-S_e^2)p - N(1-S_e)(1+S_e-S_e^2-nS_e^3)(p^2+p(1-p)\rho) + N^2(1-S_e^2)^2p(1-p)\rho,
\end{aligned}$$

where the last equality comes from grouping terms according to their dependencies, after some algebra. □

Proof of Lemma 4. Following the proofs of the previous Lemmas, let F_{+k} denote the number of false positive results in pool k. We have that

$$\begin{aligned}
\mathbb{E}\left\{F_+\right\} &= \sum_{k=1}^{N/n} E\left\{F_{+k}\right\} \\
&= \sum_{k=1}^{N/n}\sum_{i=0}^{n} E\left\{F_{+k}|X(M_k)=i\right\}\mathbb{P}\left\{X(M_k)=i\right\}
\end{aligned}$$

Note that, if $X(M_k)=i$, then $F_{+k}=0$ if the first pool test returns negative; otherwise, if the result is positive, then the expected number of false positives is $(n-i)(1-S_p)$. Using this, separating the case when $X(M_k)=0$, we have that

$$\begin{aligned}
\mathbb{E}\left\{F_{+k}\right\} &= (1-S_p)^2\,n\,\mathbb{P}\left\{X(M_k)=0\right\} + \sum_{i=1}^{n} E\left\{F_{+k}|X(M_k)=i\right\}\mathbb{P}\left\{X(M_k)=i\right\} \\
&= (1-S_p)^2\,n\,\frac{B(\alpha,\beta+n)}{B(\alpha,\beta)} + \sum_{i=1}^{n} S_e(n-i)(1-S_p)\mathbb{P}\left\{X(M_k)=i\right\} \\
&= (1-S_p)(1-S_p-S_e)\,n\,\frac{B(\alpha,\beta+n)}{B(\alpha,\beta)} + S_e(1-S_p)n(1-p).
\end{aligned}$$

Combining the above, we obtain that

$$\mathbb{E}\left\{F_+\right\} = N(1-S_p)\left((1-S_p-S_e)\frac{B(\alpha,\beta+n)}{B(\alpha,\beta)} + S_e(1-p)\right).$$

This concludes the proof. □

Proof of Proposition 1. Consider a setting with prevalence p and correlation $\rho > 0$, i.e. such that $\alpha = p\theta$ and $\beta = (1-p)\theta$, where we define $\theta := (1/\rho - 1)$. Under the Beta-Binomial model, the expected number of tests to be used per person T^{BB} is

$$\mathbb{E}\left\{T^{BB}\right\} = \frac{1}{n} + S_e + (1 - S_e - S_p)\frac{B(p\theta, n + (1-p)\theta)}{B(p\theta, (1-p)\theta)},$$

whereas in a Binomial model the expected number of tests T^B is

$$\mathbb{E}\left\{T^{B}\right\} = \frac{1}{n} + S_e + (1 - S_e - S_p)(1-p)^n.$$

We have then that

$$\mathbb{E}\left\{T^{BB}\right\} - \mathbb{E}\left\{T^{B}\right\} = (1 - S_e - S_p)\left(\frac{B(p\theta, n + (1-p)\theta)}{B(p\theta, (1-p)\theta)} - (1-p)^n\right).$$

We conclude that the result holds true if the second factor on the right-hand-side (rhs.) is non-negative, or alternatively, that

$$\ln\left(\frac{B(p\,\theta, n + (1-p)\,\theta)}{B(p\,\theta, (1-p)\,\theta)}\right) \geq n\ln((1-p)) \tag{A-1}$$

Using the definition of the Beta function B in terms of the Gamma function Γ, we have that

$$\begin{aligned}
\ln\left(\frac{B(p\,\theta, n + (1-p)\,\theta)}{B(p\,\theta, (1-p)\,\theta)}\right) &= \ln(\Gamma(p\,\theta)) + \ln(\Gamma(n + (1-p)\,\theta)) - \ln(\Gamma(\theta + n)) \\
&\quad - \ln(\Gamma(p\,\theta)) - \ln(\Gamma((1-p)\,\theta)) + \ln(\Gamma(\theta)) \\
&= \ln(\Gamma(n + (1-p)\,\theta)) + \ln(\Gamma(\theta)) - \ln(\Gamma(\theta + n)) - \ln(\Gamma((1-p)\,\theta)) \\
&= \sum_{h=1}^{n} \ln(n + (1-p)\,\theta - h) - \sum_{h=1}^{n} \ln(n + \theta - h) \\
&= \sum_{h=1}^{n} \ln\left(1 - p\frac{\theta}{n + \theta - h}\right) \\
&> \sum_{h=1}^{n} \ln(1-p) \\
&= n\ln(1-p),
\end{aligned}$$

where in the third equality we have used the fact that for any $x > 0$, $y \in \mathbb{Z}_+$ such that $x - y > 0$, $\ln(\Gamma(x)) - \ln(\Gamma(x-y)) = \sum_{h=1}^{y} \ln(x-h)$. We conclude that (A-1) holds true. This concludes the proof. □

Proof of Proposition 2. Define $\bar{p} := 1-(1/3)^{1/3} \approx 0.306$ and consider the following technical lemma, whose proof can be found in Appendix B.

Lemma 5. *If $S_e = S_p = 1$, the optimal pool size of the Binomial model is greater than one if and only if $p < \bar{p}$.*

Considering Lemma 5 above, we only consider the case of $p < \bar{p}$ and $\rho > 0$ (otherwise, the result follows trivially). Define $\theta := (1/\rho - 1)$, so that $\alpha = p\,\theta$ and $\beta = (1-p)\,\theta$, and let $g^{BB}(n)$ and $g^{B}(n)$ denote the expected number of tests used under the Beta-Binomial and Binomial models, respectively, as a function of the pool size n. That is,

$$g^{BB}(n) = \frac{1}{n} + 1 - \frac{B(p\theta, n+(1-p)\theta)}{B(p\theta,(1-p)\theta)}, \quad \text{and} \quad g^{B}(n) = \frac{1}{n} + 1 - (1-p)^n, \quad n \in \mathbb{N}.$$

Let us first examine $g^B(\cdot)$. The following Lemma, whose proof can be found in Appendix B, shows that $g^B(\cdot)$ is unimodal for the relevant range for n.

Lemma 6. *If $S_e = S_p = 1$ and $p < \bar{p}$, then $g^B(n)$ unimodal for $n \in [1, p^{-1}]$. Moreover, the optimal pool size of the Binomial model is bounded above by p^{-1}*

As a consequence of the result above, the optimal pool size for the Binomial model is the smallest n for which $\Delta g^B(n) > 0$. In what follows, we show that

$$\Delta g^{B}(n) := g^{BB}(n+1) - g^{BB}(n) \le \Delta g^{B}(n) := g^{B}(n+1) - g^{BB}(n), \quad n \le 1/p$$

thus implying that the optimal pool size for the Beta-Binomial is greater than or equal to that for the Binomial model. With some algebra, and applying the properties of the Gamma function (see proof of Proposition 1), we have that

$$\Delta g^{B}(n) - \Delta g^{BB}(n) = \frac{B(p\theta, n+(1-p)\theta)}{B(p\theta,(1-p)\theta)} \frac{p\theta}{n+\theta} - p(1-p)^n \tag{A-2}$$

Taking logarithm on the first term on the rhs. above, we have that

$$\begin{aligned}
\ln\left(\frac{B(p\theta, n+(1-p)\theta)}{B(p\theta,(1-p)\theta)} \frac{p\theta}{n+\theta}\right) &= \sum_{h=1}^{n} \ln\left(1 - p\frac{\theta}{n+\theta-h}\right) + \ln\left(\frac{\theta}{n+\theta}\right) + \ln(p) \\
&= \sum_{h=0}^{n-1} \ln\left(1 - p\frac{\theta}{h+\theta}\right) + \ln\left(\frac{\theta}{n+\theta}\right) + \ln(p) \\
&= \sum_{h=0}^{n-1} \ln\left(\frac{h\theta^{-1}+1-p}{(h\theta^{-1}+1)(1-p)}\right) + \ln\left(\frac{\theta}{n+\theta}\right) + \ln(p) + n\ln(1-p) \\
&= \sum_{h=0}^{n-1} \ln\left(\frac{h\theta^{-1}+1-p}{(h\theta^{-1}+1)(1-p)}\right) + \ln\left(\frac{1}{n\theta^{-1}+1}\right) + \ln\left(p(1-p)^n\right)
\end{aligned} \tag{A-3}$$

Let us examine the summation above. We have that

$$\begin{aligned}\frac{\partial}{\partial p}\left(\sum_{h=0}^{n-1}\ln\left(\frac{j\theta^{-1}+1-p}{(h\theta^{-1}+1)(1-p)}\right)\right) &= \sum_{h=0}^{n-1}\frac{(h\theta^{-1}+1)(1-p)}{(h\theta^{-1}+1-p)}\frac{(-(h\theta^{-1}+1)(1-p)+(h\theta^{-1}+1-p)(h\theta^{-1}+1))}{(h\theta^{-1}+1)^2(1-p)^2}\\ &= \sum_{h=0}^{n-1}\frac{h\theta^{-1}}{(h\theta^{-1}+1-p)(1-p)}\geq 0.\end{aligned}$$

Thus, we conclude that for $p<n^{-1}$ the following inequality holds true.

$$\sum_{h=0}^{n-1}\ln\left(\frac{h\theta^{-1}+1-p}{(h\theta^{-1}+1)(1-p)}\right)+\ln\left(\frac{1}{n\theta^{-1}+1}\right)\leq\sum_{h=0}^{n-1}\ln\left(\frac{nh\theta^{-1}+n-1}{(h\theta^{-1}+1)(n-1)}\right)+\ln\left(\frac{1}{n\theta^{-1}+1}\right)\quad\text{(A-4)}$$

Consider now the rhs. of the equation above; using the change of variable $x=\theta^{-1}$, we have that

$$\begin{aligned}\frac{\partial}{\partial x}\left(\sum_{h=0}^{n-1}\ln\left(\frac{nhx+n-1}{(xh+1)(n-1)}\right)+\ln\left(\frac{1}{nx+1}\right)\right) &= \sum_{h=0}^{n-1}\frac{h}{(nhx+n-1)(xh+1)}+\frac{-n}{nx+1}\\ &= \sum_{h=0}^{n-1}\left(\frac{h}{(nhx+n-1)(xh+1)}-\frac{1}{nx+1}\right)\\ &= \sum_{h=0}^{n-1}\frac{-nh^2x^2-(n-1)hx-(n-(h+1))}{(nhx+n-1)(xh+1)(nx+1)}\leq 0.\end{aligned}$$

Thus, we conclude that

$$\sum_{h=0}^{n-1}\ln\left(\frac{nh\theta^{-1}+n-1}{(h\theta^{-1}+1)(n-1)}\right)+\ln\left(\frac{1}{n\theta^{-1}+1}\right)\leq\sum_{h=0}^{n-1}\ln\left(\frac{n-1}{(n-1)}\right)+\ln\left(\frac{1}{1}\right)=0.$$

Combining the above with (A-3) we have that

$$\ln\left(\frac{B(p\theta,n+(1-p)\theta)}{B(p\theta,(1-p)\theta)}\frac{p\theta}{n+\theta}\right)\leq\ln(p(1-p^n)),$$

which in turn implies that the rhs. of (A-2) is non-positive, thus proving the result. □

Appendix B: Proof of auxiliary results

Proof of Lemma 5. For $x\in\mathbb{R}_+$ and $p\in(0,1)$, define

$$g(x,p)=\frac{1}{x}+1-(1-p)^x,\quad x\in\mathbb{R}_+,\,p\in(0,1),$$

and note that when $x\in\mathbb{Z}$, $g(x,p)$ coincides with the expected number of tests used under the Binomial model when the pool size is x and the prevalence is p. We begin analyzing the derivative of g with respect to x,

$$\frac{\partial g(x,p)}{\partial x} = -\frac{1}{x^2} - (1-p)^x \ln(1-p), \quad x \in \mathbb{R}_+, p \in (0,1).$$

Note that for any x such that $\frac{\partial g(x,p)}{\partial x} = 0$ one has that

$$g(x,p) = \frac{1}{x} + 1 - (1-p)^x = \frac{1}{x} + 1 + \frac{1}{\ln(1-p)x^2} = 1 + \frac{1}{x^2}\left(x + \frac{1}{\ln(1-p)}\right).$$

Thus, for such an x, we have that $g(x,p) > 1$ if and only if $x + \frac{1}{\ln(1-p)} > 0$, or equivalently, $p > 1 - e^{-x^{-1}}$. In particular, if $p > 1 - e^{-1/2}$, then $g(x,p) > 1$ for all $x \geq 2$. This implies that, when $p > 1 - e^{-1/2}$, there is no $x \geq 2$ such that $\frac{\partial g(x,p)}{\partial x} = 0$ and $g(x,p) \leq 1$, implying that the optimal pool size is $n = 1$, i.e. pooling is not optimal.

Consider now the case of $p \leq 1 - e^{-1/2}$. Define $p(x)$ to be such that $g(x, p(x)) = 1$, i.e.

$$p(x) := 1 - x^{-1/x},$$

and note that

$$\left.\frac{\partial g(x,p)}{\partial x}\right|_{(x,p(x))} = -\frac{1}{x^2}\left(1 - \ln x\right), \quad x \in \mathbb{R}_+.$$

We conclude that $\frac{\partial g(x,p)}{\partial x} = 0$ and $g(x,p) = 1$ when $x = e$ and $p = 1 - e^{-e^{-1}}$. Now, from the unimodality of g w.r.t. x (see Lemma 6), this also implies that pooling is not optimal for $p = 1 - e^{-e^{-1}}$. Moreover, by the continuity of g, reducing the value of p results in an optimal pool-size of either 2 or 3 (note that $g(\cdot)$ is non-decreasing in p). In particular, because $p(2) < p(3)$, pooling is not optimal for $p \geq p(3)$. Note that $p(3) = \bar{p}$. This concludes the proof of the Lemma. □

Proof of Lemma 6.

Fix $p \leq \bar{p}$ and $x \in \mathbb{R}_+$ define

$$g(x,p) = \frac{1}{x} + 1 - (1-p)^x, \quad x \in \mathbb{R}_+,$$

and note that when $x \in \mathbb{Z}$, $g(x,p)$ coincides with the expected number of tests used under the Binomial model when the pool size is x and the prevalence p. Note that,

$$\frac{\partial g(x,p)}{\partial x} = -\frac{1}{x^2} - \ln(1-p)(1-p)^x \quad \text{and} \quad \frac{\partial^2 g(x,p)}{\partial^2 x} = \frac{2}{x^3} - (\ln(1-p))^2(1-p)^x.$$

Let $\mathcal{X}'(p)$ denote the set of values for $x \geq 0$ for which $\frac{\partial g(x,p)}{\partial x} = 0$, and define $a := -\frac{1}{2}\ln(1-p) > 0$. From the above, for $x \in \mathcal{X}'(p)$,

$$x^{-2} = 2\,a\,e^{-2\,a\,x} \iff -\left(\frac{a}{2}\right)^{1/2} = -a\,x\,e^{-a\,x} \iff \mathcal{X}(p) = \left\{-\frac{1}{a}W_0\left(-\left(\frac{a}{2}\right)^{1/2}\right), -\frac{1}{a}W_{-1}\left(-\left(\frac{a}{2}\right)^{1/2}\right)\right\},$$

where $W_0(\cdot)$ and $W_{-1}(\cdot)$ denote the two real branches of the Lambert W function (these solutions exists when $-(a/2)^{1/2} \geq -e^{-1} \iff p \leq 1 - e^{-4e^{-2}} \approx 0.418$). Let $\mathcal{X}''(p)$ denote the value of $x \geq 0$ for which $\frac{\partial^2 g(x,p)}{\partial^2 x} = 0$. From the above, we have that

$$\frac{2}{x^3} = (2a)^2 e^{-2ax} \iff -\frac{(4a)^{1/3}}{3} = -\frac{2}{3}a\,xe^{-\frac{2}{3}a\,x} \iff \mathcal{X}''(p) = \left\{-\frac{3}{2a}W_0\left(-\frac{(4a)^{1/3}}{3}\right), -\frac{3}{2a}W_{-1}\left(-\frac{(4a)^{1/3}}{3}\right)\right\}.$$

(These solutions exists when $-(4a)^{1/3}/3 \geq -e^{-1} \iff p \leq 1 - e^{-27e^{-3}/2} \approx 0.489$). Note now that $\frac{\partial g(x,p)}{\partial x} > 0$ for x in the proximity of $x = 0$, and that $\lim_{x\to\infty} g(x,p) = 1$. Because of the continuity of the first two derivatives of $g(\cdot,p)$, we conclude that $g(x,p)$ is initially decreasing and convex in x, then increasing and concave, and then approaches (asymptotically) 1 by above. Thus, the result follows from showing that

$$\frac{\partial g(x,p)}{\partial x}|_{(1/p,p)} = -\frac{1}{p^2} - \ln(1-p)(1-p)^{1-p} \geq 0.$$

We show this result next. For $p \leq \bar{p}$, define $f(p) = -\frac{1}{p^2} - \ln(1-p)(1-p)^{1-p}$. One can check that $f(\bar{p}) \approx 0.017 > 0$ and that $\lim_{p\to 0+} f(p) = 0$. Additionally, one has that

$$f''(p) = (1-p)^{\frac{1}{p}-2}\underbrace{(2(1-p)b(p)+1)}_{A} - 2 - \ln(1-p)(1-p)^{\frac{1}{p}}\underbrace{\left(b(p)^2 - \frac{2b(p)}{p} - \frac{1}{p(1-p)^2}\right)}_{B}$$

where $b(p) := \frac{\ln(1-p)}{-p^2} - \frac{1}{p(1-p)}$. We have that

$$A = 2(1-p)b(p)+1 = 2(1-p)\left(\frac{\ln(1-p)}{-p^2} - \frac{1}{p(1-p)}\right) + 1 \leq 2(1-p)\left(\frac{2-p}{2p(1-p)} - \frac{1}{p(1-p)}\right) + 1 = 0,$$

where we have used the fact that $\frac{\ln(1+x)}{x} \geq \frac{2}{2+x}$ for all $x > -1$. Note now that, for $p \in (0,1)$,

$$b(p) = \frac{\ln(1-p)}{-p^2} - \frac{1}{p(1-p)} \leq \frac{1}{p(1-p)^{1/2}} - \frac{1}{p(1-p)} \leq 0,$$

where we have used the fact that $\frac{\ln(1+x)}{x} \le \frac{1}{\sqrt{1+x}}$ for all $x > -1$. Also, we have that

$$\begin{aligned}
B &= \left(\frac{\ln(1-p)}{-p^2} - \frac{1}{p(1-p)}\right)^2 - \frac{2}{p}\left(\frac{\ln(1-p)}{-p^2} - \frac{1}{p(1-p)}\right) - \frac{1}{p(1-p)^2}\\
&= \left(\frac{\ln(1-p)}{-p^2} - \frac{1}{p(1-p)}\right)\left(\frac{\ln(1-p)}{-p^2} - \frac{1}{p(1-p)} - \frac{2}{p}\right) - \frac{1}{p(1-p)^2}\\
&\le \left(\frac{2}{p(2-p)} - \frac{1}{p(1-p)}\right)\left(\frac{2}{p(2-p)} - \frac{1}{p(1-p)} - \frac{2}{p}\right) - \frac{1}{p(1-p)^2}\\
&= \frac{-1}{(1-p)(2-p)} \cdot \frac{-2p^2+5p-4}{p(1-p)(2-p)} - \frac{1}{p(1-p)^2}\\
&= \frac{-1}{(1-p)(2-p)^2}\\
&\le 0
\end{aligned}$$

where in the first inequality we have used the fact that $\frac{\ln(1+x)}{x} \ge \frac{2}{2+x}$ for all $x > -1$, and the negativity of $b(p)$ (shown above). Putting the above together, we conclude that $f''(p) \le 0$. i.e. $f(\cdot)$ is concave $p \in (0,1)$. Because $\lim_{p \to 0^+} f(p) = 0$ and $f(\bar{p}) > 0$, it must be the case that $f(p) > 0$ for all $p \le \bar{p}$. This concludes the proof of the Lemma. □

Appendix C: Log-likelihood for LTCFs

For a fixed month, consider a set of M LTCFs, where each has a population of N_m. Let X_m be the random variable that denotes the number of infected people in the LCTF m and let x_m its realization. Then, the log-likelihood of a beta-binomial distribution is given by the following expression:

$$\begin{aligned}
ll(\alpha,\beta) &= \log\left(\prod_{m=1}^{M} \mathbb{P}(X_m = x_m)\right)\\
&= \sum_{\substack{M\\m=1}} \log\left(\mathbb{P}(X_m = x_m)\right)\\
&= \sum_{\substack{M\\m=1}} \log\left(\binom{N_m}{x_m} \frac{B(x_m+\alpha, N_m - x_m + \beta)}{B(\alpha,\beta)}\right)
\end{aligned}$$

where the beta-binomial distribution parameters α and β are used as decision variables to maximize the log-likelihood.

Appendix D: Details of the simulation

D.1. Assumptions

- People can be infected only once (no reinfections are considered).
- Staff in quarantine is covered by the rest of the workers in the shift.

- The infectiousness, distribution of the incubation period and whether the patient shows symptoms or not are independent variables across individuals.
- The time of incubation t_{inc} follows a LogNormal(1.621, 0.418) distribution . The infectiousness starts at $t_{inf} = \min(\text{Uniform}[t_{inc}/3, t_{inc}], t_{inc} - 1)$, and the recovery time follows a uniform distribution between 2 and 4 weeks $t_{rec} = t_{inc} + \text{Uniform}[14, 28]$, (He et al. (2020)).
- The probability of a positive test result is based on the positivity curve presented in Sethuraman et al. (2020), considering the evolution of the patient's infection.
- Infectiousness is a scaled version of the positivity curve, and patients have a contagion potential during t_{inf} and t_{rec}, that has a peak of 0.2.

D.2. Parameters

- **Shift matrix:** Hours of the day the shift works at the facility. Residents comprise a matrix of ones, and for the staff, we have ones for the night shift (and zeros for the rest of the day); in addition, the day shift is the opposite.
- **Interaction matrix:** (i, j) is the probability of interaction between an individual of group i and another from group j. For the simulation, we use a fixed probability for all groups.
- **Interaction intensity:** is associated with the fraction of a the day the interaction can occur. For residents is 100% and for residents from each shift is 50%.
- **Groups:** We have two groups: residents (30 people) and staff (20 people).
- **Asymptomatic patients:** Each infected patient does not show symptoms with a fixed probability. We use a 30% probability (Stephen A. Lauer et al. 2020).
- **Exogenous rate of infection for the staff:** A daily probability of 0.1%.
- **Preventive quarantine for symptomatic patients:** All symptomatic patients start preventive quarantine after a number of days from the onset of the symptoms. For the simulation we use two days.

D.3. Simulation Process

We perform a Monte Carlo simulation for the daily evolution of the infection. We keep track of the number of people infected (either symptomatic or not) and whether they are in preventive quarantine or not. Every simulation starts with all of the population noninfected.

- At $t = 0$, we randomly assign each individual the condition of asymptomatic, so in case of becoming infected at any moment in the simulation horizon, these individuals will not show symptoms.
- At the beginning of $t = j$, we compute the probability that a patient will acquire the infection during the day.

- We generate the contagions. For each newly infected patient, we generate the incubation, infection and recovery time.
- We move to the next day $t = j + 1$.

Appendix E: Log-likelihood

Consider a simulation with N people in total, and a given day t (this day t can be considered with respect to the first day of the simulation, or, alternatively, from the first day of a symptomatic case). Denote M the number of simulations performed. Let X_t be the random variable of the number of infected cases on that day t and let x_{ts} its realized value in simulation $s \in \{1, \ldots, M\}$.

E.1. Beta-Binomial

Under the beta-binomial model, the log-likelihood for a given day t can be written as

$$\begin{aligned} ll(\alpha_t, \beta_t) &= \log\left(\prod_{s=1}^{M} \mathbb{P}(X_t = x_{ts})\right) \\ &= \sum_{s=1}^{M} \log\left(\mathbb{P}(X_t = x_{ts})\right) \\ &= \sum_{s=1}^{M} \log\left(\binom{N}{x_{ts}} \frac{B(x_{ts} + \alpha_t, N - x_{ts} + \beta_t)}{B(\alpha_t, \beta_t)}\right) \end{aligned}$$

where the beta-binomial distribution parameters α_t and β_t are used as decision variables to maximize the log-likelihood.

E.2. Binomial

For the binomial probability model, the log-likelihood for a given day t can be written as

$$\begin{aligned} ll(p_t) &= \log\left(\prod_{s=1}^{M} \mathbb{P}(X_t = x_{ts})\right) \\ &= \sum_{s=1}^{M} \log\left(\mathbb{P}(X_t = x_{ts})\right) \\ &= \sum_{s=1}^{M} \log\left(\binom{N}{x_{ts}} p^{x_{ts}(1-p)^{N-x_{ts}}}\right) \end{aligned}$$

where in this case, the binomial parameter p_t is the decision variable to maximize the log-likelihood expression.

Appendix F: Beta-Binomial with uncorrelated data

The used Beta-Binomial assumes a correlation term which can be expressed as $\rho = 1/(1 + \alpha + \beta)$, while always will be a positive number. However, if there is the case in which we do not have

Table 5 Estimate prevalence and correlation of a Beta-Binomial on uncorrelated sampled data.

		Prevalence			Correlation		
Real prevalence	Population	Avg.	Min.	Max.	Avg.	Min.	Max.
0.05	50	0.0502	0.0474	0.0524	0.000001	0.000001	0.000004
0.10	50	0.1000	0.0974	0.1030	0.000001	0.000001	0.000001
0.20	50	0.2001	0.1954	0.2050	0.000001	0.000001	0.000001
0.05	100	0.0500	0.0479	0.0518	0.000001	0.000001	0.000001
0.10	100	0.1000	0.0978	0.1027	0.000001	0.000001	0.000001
0.20	100	0.2001	0.1965	0.2031	0.000001	0.000001	0.000001
0.05	500	0.0500	0.0493	0.0507	0.000008	0.000001	0.000188
0.10	500	0.1000	0.0988	0.1012	0.000006	0.000001	0.000165
0.20	500	0.2001	0.1986	0.2017	0.000011	0.000001	0.000204

correlation in reality, and therefore we would prefer to not have a probability model that results in a non-zero correlation. We proceed to fit a Beta-Binomial distribution via maximum likelihood over sampled instances for different values of prevalence and population. For each of the cases, we evaluate 100 scenarios, where on each scenario there are 1000 sample points from which a Beta-Binomial distribution is fitted.

Table 5 shows for different prevalences and populations the average, minimum and maximum estimated prevalence and correlation according to the fitted Beta-Binomial distribution. We can observe that the correlation terms are very close to zero in all cases. As a result, using the Beta-Binomial probability distribution model even if the underlying data has no correlation will not impact the pool size selection.