Biostatistics (2021), **0**, 0, *pp.* 1–30
doi:10.1093/biostatistics/output

BAYESIAN GROUP TESTING WITH DILUTION EFFECTS

CURTIS TATSUOKA*, WEICONG CHEN, XIAOYI LU,

Department of Population and Quantitative Health Sciences, Case Western Reserve University, Cleveland, Ohio, Department of Computer and Data Science, Case Western Reserve University, Cleveland, Ohio, Department of Computer Science and Engineering, University of California Merced, Merced, California, U.S.A.

cmt66@case.edu

Summary

A Bayesian framework for group testing under dilution effects has been developed, using lattice-based models. This work has particular relevance given the pressing public health need to enhance testing capacity for COVID-19 and future pandemics, and the need for wide-scale and repeated testing for surveillance under constantly varying conditions. The proposed Bayesian approach allows for dilution effects in group testing and for general test response distributions beyond just binary outcomes. It is shown that even under strong dilution effects, an intuitive and simple-to-implement group testing selection rule that relies on the model order structure, referred to as the Bayesian halving algorithm, has attractive optimal convergence properties. Analogous look-ahead rules that can reduce the number of stages in classification by selecting several pooled tests at a time are proposed and evaluated as well. Group testing is demonstrated to provide great

*To whom correspondence should be addressed.

© The Author 2021. Published by Oxford University Press. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

savings over individual testing in number of tests needed even for moderately high prevalence levels, although with a trade-off of higher number of testing stages, and increased variability. A web-based calculator is introduced to assist in weighing these factors and to guide decisions on when and how to pool under various conditions. High performance distributed computing methods have also been implemented for considering larger pool sizes, when savings from group testing can be even more dramatic.

Key words: Group testing, surveillance, Bayesian, dilution effects, lattices, optimal rates of convergence, COVID-19.

1. Introduction

The group testing formulation originated by Dorfman (Dorfman (1943)) has found use in a diverse array of applications, including COVID-19 testing. The motivating idea of group testing for a disease like COVID-19 is that, say if biomarker samples from $N, N > 1$ subjects are pooled, and if the prevalence is low, most likely the test result for the pooled sample will be negative, indicating that all N subjects in the pool are negative, using only in one test. On the other hand, if the result indicates a positive test, and hence that there is at least one positive sample present among the pool, then further testing can be conducted to identify the positive subjects. Dorfman suggested that all samples which comprise a positively-tested pool should subsequently be tested individually. This approach has been adopted broadly. Nonetheless, in the presence of testing error, this approach can be problematic, in that classification error thresholds may not be reached, and it may not be comparatively efficient, as pooling subsets can still be useful. An important potential source of testing error is through dilution, which may occur when pooled samples contain few positives relative to a larger number of negatives. Described here is a Bayesian framework for systematically addressing these important issues surrounding pooled testing such

as for COVID-19.

Recently, the Food and Drug Administration (FDA) has released guidelines for allowing group testing of patients suspected of having COVID-19. There is thus a renewed flurry of interest in group testing for COVID-19 (Lohse *and others* (2020); Majid *and others* (2020); Hogan *and others* (2020); Shental *and others* (2020); Donoho *and others* (2020)), as it can play a fundamental role in efficient disease surveillance. As there is a large percentage of asymptomatic cases, widespread testing is important for understanding and controlling spread. There also is a need for monitoring through repeated testing of large swaths of the population, such as for frontline and essential workers, and those at high risk. A potential drawback to such type of surveillance is the reliance on imprecise testing. PCR-based assays based on samples taken with nasal swabs can be highly accurate, although issues with sample collection or mis-priming can lead to errors. These error levels also arise with pooling, and increase with pool size. We refer to this phenomenon as a *dilution effect*. For instance, false negative rates from pooling with PCR-based assays with COVID-19 have respectively been estimated to range from 0.07, 0.09, and 0.19 for pools of size 5, 10 and 50 Bateman *and others* (2021).

Fast, less expensive, and non-intrusive approaches for testing, such as those based on saliva or antigens, can broaden the scope of testing, but may be even less accurate. Emergence of variants also may affect testing error in the future.

A basis for the proposed group testing methods are models known as lattices, in which the classification states that represent the profiles of positivity among test subjects follow a partial ordering. This order structure is used to guide next stage pool selection, and importantly, it provides fundamental insight into the convergence properties of Bayesian group testing with dilution effects. The proposed Bayesian group testing approach explicitly acknowledges testing error and heterogeneity in individual risk levels through prior distribution specification, and can be applied with general response distributions, i.e. either quantitative or categorical test

responses. For COVID-19, these outcomes could be viral load level or type of variants detected. Importantly, individual-level classification error can be systematically be reduced to any error threshold, as our proposed approach has attractive optimality properties even in light of dilution effects in which correct classification can efficiently be made for all test subjects. A framework for sequential Bayesian classification on lattices has been described earlier (Tatsuoka and Ferguson (2003)). This current work differs from the previous formulation as we now consider test response distributions that depend on pool size and the number of positives in the pool. In terms of theoretical work involving partially ordered classification models, rules that rely on Kullback-Leibler information of response distributions that are specific to the experiment, such as in cognitive and educational adaptive testing, has been studied (Tatsuoka (2014)), as well as in group testing and partially ordered classification models (Ferguson and Tatsuoka (2004); Tatsuoka (2005)).

Group testing has a rich statistical literature that has expanded upon Dorfman's seminal work. An important result in group testing by Ungar (Ungar (1960)) gives a demarcation for when it is optimal to group test when responses are binary and there is no testing error. A Bayesian approach was first described in Sobel and Groll (Sobel and Groll (1966)), where binary responses without testing error are assumed. In addition to screening for disease, group testing has been used in genetics testing (Gastwirth (2000); Sham *and others* (2002)), and to identify promising drug compounds (Xie *and others* (2001); Remlinger *and others* (2006)). The phenomenon of testing error in group testing has been widely recognized (Graff and Roeloffs (1972); Hwang (1976); Litvak *and others* (1994); Thierry-Mieg (2006); Kim *and others* (2007)), including dilution effects with specific functional forms (Hwang (1976); Zenios and Wein (1998); Hung and Swallow (1999)). The possibility that objects have different probabilities of being positive but with no testing error has been considered (Hwang (1975); Black *and others* (2012)). Beyond binary, responses have been assumed to be normally distributed (Zenios and Wein (1998)), and to reflect viral load

levels (Ghosh *and others* (2020)). Group testing has been studied in the context of multiplex and high throughput testing (Bilder *and others* (2019); Donoho *and others* (2020). In (Bilder *and others* (2019)), it is shown that taking into account heterogeneous prior probabilities can improve efficiency. Recently, methods for non-adaptive single-stage COVID-19 group testing have been suggested which are primarily based on the idea of compressed sensing (Shental *and others* (2020); Ghosh *and others* (2020)). These methods allow large pooling sizes and potentially a single test stage to identify all positive samples. However, they can only work well when prevalence rate is very low ($< 1.3\%$), and are not necessarily convergent, as the confounding of positive cases is still possible. There also was no prior information included on samples and no dilution effect. Estimation of proportion such as for disease prevalence is an important application in group testing. Previous work has been based on binary outcomes (Sobel and Groll (1966); Gastwirth and Hammick (1989); Gastwirth and Johnson (1994); Hughes-Oliver and Swallow (1994); Brookmeyer (1999); Tebbs (2003)).

Below, we will first introduce the proposed method and related notions, with an example. We then summarize theoretical properties for optimal designs in Bayesian group testing, and for an intuitive pooled test selection rule. k-test look ahead analogues are presented as well, which select k tests at a time, and can reduce the back and forth in preparation and testing. The proposed approaches can be employed with general response distributions, such as for viral load outcomes, or multinomial outcomes, such as for different variants of COVID-19. Finally, we introduce a web-based tool that implements the proposed algorithms and graphically represents results. Example scenarios are presented, which demonstrate the efficiency gains of group testing, and the trade-offs among the rules. As well, a brief note in the Appendix is given on how Bayesian estimation of prevalence can be conducted within the proposed framework through mixtures of posteriors. This is relevant to surveillance of COVID-19. Proofs of theorems and a more thorough technical discussion are given in the Appendix.

2. Methods

2.1 *Lattice Models for Bayesian Group Testing*

We illustrate how lattice-based classification models can be applied in Bayesian group testing with dilution effects.

Example 1. Consider all the possible subsets generated by subjects A and B, as in Figure 1. Suppose each subset represents a profile of the subjects that are negative for COVID-19. Hence, one of these subsets represents the true state. The collection of these subsets thus form the classification model, and we henceforth refer to them as states: State AB denotes that both A and B are negative, state A denotes that only A is negative, and hence B is positive. Similarly, state B denotes that only B is negative. State $\hat{0}$ represents that both subjects are positive.

Denote the collection of these states as S. In the above example, $S = \{AB, A, B, \hat{0}\}$. These states (i.e. all the possible subsets of subjects that are negative) actually have a specific order structure in that they can be ordered by inclusion, and form what is known as a powerset lattice. When ordering by inclusion, note for instance that state AB is greater than states A, B, and $\hat{0}$ because the set comprised of both A and B contains A alone, B alone, and $\hat{0}$, respectively. On the other hand, states A and B are incomparable, as there is no inclusion relationship between the respective subsets. In this example, there are two subjects, and 4 possible profiles of them being positive or negative. In general, for N subjects, there are 2^N such possibilities, corresponding to the powerset of the N subjects. Further, note that there is essentially a one-to one correspondence between states in lattice model and potential pooled tests. There are 2^N - 1 possible way to pool tests, since the empty set, represented as $\hat{0}$, is not a possible test. So, in this example, for state AB, there is also a pooled test with samples from both A and B; for state B, there is a test consisting of a sample only from subject B, etc. As the size of classification model and experiment pool grow exponentially in N, significant computational challenges arise. We have implemented distributed computing and numerical algorithms that allow models with $N = 20$ (> 1 million

classification states and experiments) to feasibly be analyzed when comparing performance across algorithms.

We adopt a Bayesian approach. An advantage is the allowance for individual-level prior information about positivity into the classification process, as well as systematic characterization of uncertainty from potential testing error. In a Bayesian framework, the posterior probability values embody the combined empirical and prior evidence as to which state in the classification model is the true one. Larger posterior probability values for the true state are clearly desirable, with a probability value of 1 indicating certainty in correctly identifying the true status of positivity for all the subjects. A *stage* of testing is an iteration of preparing a collection of (pooled) samples for testing and observing the results of the test(s). The posterior distribution after n stages of testing on the collection of classification states S will be denoted as π_n, and for a state $j \in S$, the posterior probability value that state j is the true state is $\pi_n(j)$. Corresponding prior distribution values are denoted as π_0 and $\pi_0(j)$.

For instance, prior to testing for COVID-19, each subject could be assigned a prior probability of positivity based on logistic regression modeling, with explanatory variables such as age, gender, location of residence, travel history, status of household contacts, self-reported exposure, community infection rates, etc. For a given profile of the N subjects in terms of their positivity status, to obtain its prior probability, we take the product of the respective individual prior probabilities for their respective status in the profile. These values form a prior probability distribution across the possible states. After a (pooled) test result is observed, using Bayes rule, prior probabilities are updated to posterior probabilities. We illustrate these computations in the example below.

Example 1, continued. Suppose that for a subject A the prior probability that he/she is a positive (e.g. has COVID-19) is 0.05, and independently for a higher at-risk subject B it is 0.10. The prior probabilities for state membership follows: for state $AB = \hat{1}$, $\pi_0(AB) = (0.95)(0.90) = 0.855$, for state A it is $\pi_0(A) = (0.95)(0.10) = 0.095$, $\pi_0(B) = (0.05)(0.90) = 0.045$,

and $\pi_0(\hat{0}) = 0.005$. Possible tests include pooling A and B, and testing each subject individually. Suppose also that a test has two possible outcomes, indicating that either at least one positive is present in the pool, or that only negatives are in the pool. Assume that the specificity, the probability of observing a negative outcome given that the pool does in fact consist only of negative samples, is 0.99. Also, assume that the sensitivity, the probability of observing a positive outcome given that the pool contains at least one positive sample, also is 0.99 when all samples are positive, and reduces slightly to 0.98 when only one of two samples is positive.

Let the first test be comprised of pooled samples from A and B. Then, if a negative outcome is observed, it can be seen by Bayes rule that $\pi_1(AB) = 0.9966$. Hence, combined with the prior information, this test result would lead to strong indication that both of the samples are negative. If instead a positive test outcome is seen, then $\pi_1 = \{\pi_1(AB) = 0.0567, \pi_1(A) = 0.6177, \pi_1(B) = 0.2926, \pi_1(\hat{0}) = 0.0328\}$. Suppose next that samples from A and B are tested individually in the next stage as in Dorfman, and the outcome for A is negative, and for B is positive. If we update the posterior probabilities for the two observed outcomes simultaneously, then, $\pi_2 = \{\pi_2(AB) = 0.0009, \pi_2(A) = 0.9985, \pi_2(B) = 0.0001, \pi_2(\hat{0}) = 0.0005\}$. In particular, the posterior probability that B is positive is the sum of respective state membership probabilities that indicate that B is positive: $\pi_2(A) + \pi_2(\hat{0}) = 0.9990$. This is clear evidence that indeed B is positive. Similarly, note that the probability that A is positive is $\pi_2(B) + \pi_2(\hat{0}) = 0.0006$, which indicates that A is clearly negative. □

Making "optimal" pooled test selections becomes more interesting and complex as the number of subjects N being considered for pooling increases, the individual risks of positivity vary, and as testing error increases through dilution effects. We consider the following criterion as a basis for optimality in pool selection, from a theoretical perspective. First, let $s \in S$ be the true state (i.e. representing the correct identification of negatives and positives). Typically, $1 - \pi_n(s)$ converges exponentially to 0 at the order $e^{-\alpha n}$ for some α, and α is called the *rate of convergence*. The

mathematical definition of the rate of convergence is taken to be

$$\alpha = \liminf_{n\to\infty} -(1/n)\log(1-\pi_n(s)).$$

We establish group testing designs and selection rules that attain the optimal rates of convergence (largest possible alpha) with probability one under general dilution effects. Note that rates are in terms of number of tests administered, with one test per stage. We show that they are also characterized in terms of Kullback-Leibler information values, which are discrepancy measures for probability distributions.

2.2 *Dilution Effects and Optimal Designs*

Note that as pool size increases, it is possible that testing outcomes can become less reliable. For instance, for a given number of positive samples, it may be increasingly harder to detect their presence as pool size increases. Another dilution effect can occur in relation to the number of positive samples in a pool. For instance, for a given pool size, it may be increasingly easier to detect the presence of positive subjects as their number increases. Such dilution effects and their effect on the optimality of pooled test selection in terms of attaining the optimal rate of convergence is characterized by Theorem 1 in the Appendix.

We assume the following conditions that characterize dilution effects, which are formally represented in terms of Kullback-Leibler information values (see Appendix). We assume that i) for a given number of positive samples, as pool size increases, pooled tests become less discriminatory in identifying the presence of positive samples; ii) as the difference in number of positive samples increases for a given pool size, it is relatively easier to discriminate the presence of the positives; (iii) for a given pool size and a fixed difference in the number of positive samples present, as the total number of positive samples increases, the respective test response distributions become harder to discriminate. Finally, we assume a non-restrictive technical condition, (A4) in the Appendix. Theoretical demarcations when it is preferable to group test depend on which state is

true. We illustrate examples of demarcations in the Appendix, which show that group testing is attractive asymptotically even under strong dilution effects.

Under these conditions, optimal designs can be considered for three case: 1) all subjects are positive (the bottom state in the lattice), 2) all subjects being tested are negative (the top state in the lattice), and 3) there is a mix of negative and positive subjects (a state in the lattice"in-between" the top and bottom states). In Theorem 1, we see that optimal strategies for attaining the fastest possible rates of convergence are: 1) test subjects individually, and 2) to group test samples from all subjects. For the third case, optimal strategies are more complex. The covers and anti-covers of the true state (states directly above and below it) must be considered, as these states pose the most difficulty in discriminating the true state, and hence have posterior probability values that converge to zero the slowest. Under general conditions, the optimal strategy is to group test the negatives and individually test the positives. Optimal rates depend on the statistical discriminatory efficiency of the pooled tests, as measured by Kullback-Leibler information. They also depend on the complexity of the lattice around the true state, as measured by the number of states directly surrounding the true state. Respective tests are selected proportionally so that the rate at which the slowest of the non-true state posterior probability value converges to zero is maximized. In other words, it is desirable that the posterior probabilities of the covers and anti-covers converge to zero at the same rate. More technical details are discussed in the Appendix. In practice, the true state identity is not known. Next, we propose a testing selection rule that will eventually identify the true state and adopt the corresponding optimal strategy, hence attaining the optimal rate of convergence, no matter the true state.

2.3 *Bayesian Halving Algorithm and Analogous k-Test Look Ahead Rules*

A key to obtaining systematic efficiency gains is to base the sequential selection of pooled tests on the Bayesian information provided by the posterior probability values for the states. In Theorem

2 of the Appendix, we will establish that under general conditions for dilution effects, a simple and intuitive selection rule for pooling, referred to as the *Bayesian halving algorithm (BHA)*, attains the optimal rate for the true state posterior probability to converge to the value 1. While this is an asymptotic property, given that the rates are exponential, attaining optimal rates is practically attractive for finite testing horizons as well, and insures that any error thresholds will be attained eventually with sufficient testing.

A key concept from order theory is the idea of an up-set for a state in the lattice. This concept is useful for motivating the proposed test selection rules. For a state j, denote its up-set as j. This is defined as the subset of states within the lattice that are at least as great as (i.e. contain) the state in question. In Figure 1, $AB = \{AB\}$; for $B, B = \{B, AB\}$, for $A, A = \{A, AB\}$, and for $\hat{0}$ it is all 4 of the states. Practical interpretations of these up-sets are as follows: the up-set of AB is all the states for which a pool of A and B would contain only negative subjects, the up-set of B is all the states such that B is negative, etc. Conversely, the complement of the up-set of $AB = \{A, B, \hat{0}\}$, and these are the states for which a pooled test AB would contain at least one positive sample. Similarly, the complement of the up-set of $B = \{A, \hat{0}\}$ is comprised of the states for which an individual test of a sample from subject B would be positive, etc. Hence, the up-set of a state and its complement generate a partition of the classification states according to whether the corresponding pooled test would contain all negatives or not.

Now let us consider rules for pooled test selection based on π_n. It is desirable to have a rule that adopts optimal strategies and attains optimal rates of convergence of the true state posterior probability value to 1, no matter which state is true, and under dilution effects, which are likely to arise especially for large pool sizes. The following intuitive pooled test selection rule possesses these properties, as we establish in Theorem 2 in the Appendix.

Consider the Bayesian halving algorithm, which selects the pool e that minimizes

$$h(\pi_n, e) = |m_n(e) - 0.5|, \text{ where } m_n(e) = \sum_{j \in upe} \pi_n(j). \tag{1}$$

The Bayesian halving algorithm systematically partitions the lattice model, so that regardless of the actual observed outcome, through Bayes rule, one of the two partitions of states will get increased posterior probability values, while the states in the other partition will get decreased values. This systematically encourages posterior mass to eventually accumulate to one state. This property supports that purposeful and systematic pool choices will be made as test results are observed that will lead to correct classification. From Example 1, we compute $h(\pi_0, AB) = 0.355, h(\pi_0, A) = 0.45, h(\pi_0, B) = 0.40$. Hence, it would be attractive to first select $e = AB$, and to group test.

We also extend the halving algorithm to an analogous k-test look ahead rule, $k > 1$, so that k pools are selected for simultaneous testing, as opposed to the one at a time testing implicit with the proposed halving algorithm. Selecting k pooled samples at a time may be more practical, as this can reduce the back and forth involved in preparing pooled samples for testing based on observed results. We define a *stage* of testing to be the simultaneous submission of one or more pooled samples for testing. Given π_n at a stage n, the idea behind our proposed k-test version of the halving algorithm is to partition the posterior mass on the lattice model into 2^k partitions according to sets of the following form. We choose k pooled samples $e_1, \ldots, e_k$ that minimize

$$|\sum_{j \in e_1 \cap e_2 \ldots e_k} \pi_n(j) - 1/2^k| + |\sum_{j \in e_1 \cap e_2 \ldots e_k^c} \pi_n(j) - 1/2^k| + \ldots$$

$$|\sum_{j \in e_1^c \cap e_2^c \ldots e_k^c} \pi_n(j) - 1/2^k|.$$

For example, when $k = 2$, this criterion involves selecting e_1, e_2 that splits the posterior mass on the lattice model into $2^k = 4$ parts as equally as possible, minimizing

$$|\sum_{j\in e_1\cap e_2}\pi_n(j)-1/4|+|\sum_{j\in e_1\cap e_2^c}\pi_n(j)-1/4|+$$
$$|\sum_{j\in e_1^c\cap e_2}\pi_n(j)-1/4|+|\sum_{j\in e_1^c\cap e_2^c}\pi_n(j)-1/4|. \tag{2}$$

We can approximate the minimization in (2) by sequentially selecting the k experiments one at a time. First, given current posterior probability π_n, we choose e_1 with the halving algorithm, e_2 that minimizes (2) given π_n and e_1, etc. This approximation is computationally much faster.

A halving algorithm that uses prior probabilities is previously described in Black et al.(Black *and others* (2012)), where samples are successively split in half for positive pooled tests while also accounting for heterogeneity in risk. This approach is related to what we propose here, but our approach is fully Bayesian, with posterior probability updating serving as the basis for selection and stopping. Latvik et al.(Litvak *and others* (1994)) present a set of several rules that are related to the halving algorithm. However, the user has to select a rule from this set. These rules differ in the amount of re-testing of pools that is conducted. In contrast, the proposed Bayesian halving algorithm is an automated rule. It also can be used with general test response distributions.

3. Results

3.1 *Simulations*

For the Bayesian halving algorithm and its k-test look ahead counterparts, $k > 1$, we have created a web-based calculator and code for high performance computing (HPC) distributed computing that generates important performance statistics relating to expected values and standard deviations of number of tests and stages, as well as classification rates, and false positive and false negative rates. The website address is `https://bayesgrouptest.case.edu`, and the code for the HPC implementation can be found at `https://github.com/ben072292/BGTF-artifact`. These performance statistics can be used as the basis for comparison and selection of group testing

strategies depending on a specific set of conditions. Being able to flexibility assess the trade-offs between different group testing strategies is important for COVID-19 surveillance, as prevalence levels and individual risks, and testing approaches and lab capacities can greatly vary over time and location. Overall performance is derived by averaging state-level expected performance with corresponding prior probability values in π_0. Required inputs are a vector of prior probabilities of positivity for each subject, a matrix with the probabilities indicating presence of a positive that depend on pool size and the number of positives in the pool, and a target posterior probability error threshold for individual classification. Pooled test sequences are also generated. Note that this tool currently only employs binary outcomes for pooled tests ("positive is present", or not), and henceforth only that case is considered.

Example 2 We next consider examples with $N = 12$ as an illustration, with varying individual prior probability risk levels: 1) all prior probability values of positivity p_0 for each subject is 0.02, 2) all such values are 0.20, 3) mixed prior values, with one subject having the value 0.20, while the rest having the value 0.02. We also assume a dilution effect model modified from Hung and Swallow (Hung and Swallow (1999)), which depends on constants $0 < \alpha < 1$ and $h > 0$, the number of positives in a pool e, denoted as r, and pool size $|e|$. Suppose responses are binary, with specificity $p_{r=0,|e|} = \alpha$, and one minus the sensitivity value

$$q_{r,|e|} = \alpha \cdot r/((|e| - r) \cdot h + r), 1 \leqslant r \leqslant |e|.$$

For our examples, we assume $\alpha = 0.99$, and $h = 0.005$. As an illustration, when pool size is 12, sensitivity ranges from 0.9384 when only one subject is positive, to 0.99 when all of the 12 subjects in the pool are positive. These dilution effects are comparable for those found with COVID-19 PCR-based pooling Bateman *and others* (2021). $5 : 0.9759, 10 : 0.9529, 50 : 0.7942$ The webtool can generate matrices of dilution effects for varying α and h. The dilution effect conditions required in Theorem 2, which relate to optimality of the Bayesian halving algorithm, are satisfied with this example. In these settings, we consider performance of individual testing,

k-test look ahead rules, and the Bayesian halving algorithm (denoted as $k = 1$).

Finally, stopping of testing is assessed at the individual level. Note that the posterior probability that a subject is negative (positive) is equal to the sum of the posterior probability values of all the states for which the subject is represented as negative (positive). In our example, an individual subject is classified and no longer considered for further testing once this posterior classification error of being negative or positive is at or below a given error threshold. Unless stated otherwise, for our simulations, the error threshold for positive classification is 0.01, while for negative classification it is 0.005. The negative threshold is stricter to reduce the false negative rate. After subject-level classification, the lattice model is reduced accordingly, mainly to ease computational burden associated with the larger lattice size. Posterior probabilities are renormalized among remaining states. For instance, in Figure 1, if A is removed, then states AB and B are merged, A and $\hat{0}$ are merged, etc., and posterior probabilities are correspondingly aggregated.

Simulations are based on exhaustive analyses of possible test response sequences, up to a number stages that varies by algorithm. This variation was mainly due to numerical challenges as k increases, and computational time limitations. Hence, the reported averages and error levels are approximations. Note for $k = 1$: 24 stages; $k = 2$: 12 stages; $k = 3 : 12$ stages; $k = 4$: 10 stages for $p_0 = 0.02$ and for the case with mixed p_0 values, 6 stages for $p_0 = 0.2$; Individual: 5 stages. Overall rates of subjects that are classification by these respective stages, including are generally around 98%, except when $k = 4$, which are a bit lower.

Figure 2a shows, as expected, that in terms of average number of tests needed to classify all subjects, group testing with the Bayesian halving algorithm ($k = 1$) has the smallest average number of tests and individual testing the largest, across all the scenarios and selection rules considered here. Indeed, the reduction achieved through group testing can be dramatic. Testing efficiency is particularly striking when the prior probability values of positivity for subjects are small. This gain in testing efficiency reduces as k increases among k-test look ahead rules.

Conversely, as seen in Figure 2b, in terms of the number of stages, individual testing has the smallest average while the $k = 1$ case has the largest. This isn't surprising given that more tests are submitted per stage as k increases. There is thus a trade off between minimizing number of tests versus number of stages, with the intermediate values of $k > 1$ serving as a compromise between the flexibility of single test at-a-time ($k = 1$) group testing versus the relative logistical ease of individual testing. We only consider rules with $k \leqslant 4$ as this still allows for adaptability with respect to test responses, as well as numerical feasibility. Larger values of k require greater computational resources in generating and analyzing possible binary testing sequences, as the number of branches grows exponentially in k and stage number.

In terms of average false positive and false negative rates, note in Figure 3 that false negative rates are higher, particularly as k decreases, and as p_0-values increase. However, as shown in Figure 4, it is possible to adjust error thresholds for negative versus positive classification, making the threshold for negative classification relatively more stringent. As an example, for $k = 1, p_0 = 0.2$ and negative classification threshold of 0.001, average false negative rate drops to 0.273%. However, the corresponding average number of tests does increase, to 11.04. Table 1 lists the first stage test selections. Multi-stage lists can be generated with the webtool. In Table 1, note for instance when k=2 and $p_0 = 0.02$ for all subjects, the first two tests selected for the first stage are to pool all subjects and to individually test subject A. For all prior probability scenarios, group testing is initially selected.

Variability in the number of tests and stages per rule also is an important consideration. See Figure 5. Note that the standard deviation in the number of stages is approximately the standard deviation for the number of tests divided by k, the number of tests conducted per stage. Hence, as k increases, the variability in number of stages can be even smaller than individual testing. Individual testing has lower variability in number of tests. Overall, the variability in number of tests and stages per algorithm appears to slightly increase in N.

In Figure 6, note that as N increases, the ratio of expected number of tests for individual testing relative to the proposed group testing rules also increases. For instance, for the case when $p_0 = 0.02$, as N increases from 6 to 15, note that for the Bayesian halving algorithm, this ratio grows from 4.75 to 8.17. The growth rate appears slightly less as k increases, with the ratio increasing from 1.47 to 3.34 when $k = 4$. As prior risk increases, these ratios and the advantage of group testing lessens, as noted in the mixed and $\pi_0 = 0.20$ scenarios in Figures 6b and 6c. For instance, for the Bayesian halving algorithm, the ratios range from 1.77 to 2.32, and for $k = 4$, 0.99 to 1.64. Conversely, the ratio of the expected number of stages with individual testing versus group testing decreases in N. For instance, for the case when $\pi_0 = 0.02$, as N increases from 6 to 15, note that for the Bayesian halving algorithm, this ratio decreases from 0.79 to 0.54. Interestingly, the rate of decrease appears slower as k increases, with a ratio equal to 0.77 for $N = 15$ when $k = 2$. As prior risk increases, these ratios are smaller, and the relative "cost" of group testing increases. Note when $\pi_0 = 0.20$, for the Bayesian halving algorithm, the ratios range from 0.3 to 0.15. For $k > 1$, the ratios of expected stages are larger. For $k = 4$, ratio values range from 0.66 to 0.44. Roughly, this trade-off appears more attractive for lower prior risk, and less attractive as prior risk levels increase. Variation also should be considered. At $\pi_0 = .20$, unless the number of stages is relative inexpensive and easy to process, individual testing becomes more attractive. We stress that a final decision depends on "local" factors.

Computation of performance statistics presents combinatorial challenges as N and k increases. The lattice model has 2^N states, so it grows exponentially in size, as does the number of pooled tests to be considered. Similarly, the number of branches at each stage is 2^k, so that k-test look ahead rules also present challenges. To address these challenges and systematically evaluate simulations under a wider range of scenarios that are beyond the practical scope of the webtool in terms of computational time requirements, we utilize state-of-the-art big-data techniques. We have designed and developed an Apache Spark Zaharia *and others* (2010) based simulation

framework.

We implemented several novel algorithmic optimizations to our proposed simulation framework, which reduce general computational complexity of simulations, such as 1) Removing subjects as soon as their individual classification threshold is reached, which reduces lattice size; 2) In computing classification statistics, as in the Tables, we first generate a tree comprised of branches reflecting the possible test response sequences up to a specified stage. Branches with probabilities of less than some threshold, such as 0.001 for a given state, are "pruned" in assessing performance for that state. Spark executors are used to generate respective branch probabilities for the different states being true; 3) When the individual prior probability values p_0 are all equal, symmetries within the powerset lattice are utilized, and redundant computations involving states with same number of negatives can be avoided. For instance, in such cases, when $N = 12$, computationally only 13 states need be considered, as opposed to all 2^{12} states. 4) We skip the evaluation of states with low prior probabilities, such as removing states with the lowest prior probability values that sum up to 1%. This has only a small effect on the derived summary statistic values.

These optimization techniques were systematically evaluated on an AMD-based HPC cluster with up to 272 CPU cores. We report that our proposed simulation framework can achieve near-perfect scalability, with simulation time under 1 hour for $N = 20, p = 0.02$. Computational feasibility for $N = 20$ supports the practicality of the proposed methods. Table 2 illustrates the derived summary statistic values, note we use a more stringent negative classification threshold for these simulations of 0.001.

4. Conclusion

In a Bayesian group testing framework, dilution effects are considered that depend on pool size and the number of positives in a pool. It is seen that even when dilution effects are quite significant, asymptotically optimal strategies rely on group testing. Moreover, a simple and intuitive

Bayesian halving algorithm utilizes the natural order structure in the group testing problem when viewed as a lattice model, and it is shown to attain optimal rates of convergence. k-test look ahead analogues of this halving algorithm also are proposed, which still can take advantage of pooling and adaptability, while reducing the number of stages in testing at the expense of testing efficiency. Computational tools have been developed that allow for assessment of the proposed group testing approaches, given inputted prior probabilities of positivity per subject, dilution effects and classification error level thresholds. These tools include the use of distributed computing to address the Big Data lattice representations with larger N, and they establish the feasibility and practicality of simulated group testing designs that involve posterior updating up to $N = 20$. Future work will focus on extending the computational feasibility for even larger N. The respective expected number of tests and stages (and standard deviations), classification error, and comparisons to individual testing are generated. Balancing the expected number of tests versus stages, and the respective variances, can be weighed across different k-test look ahead rules, to compare against individual testing. Deciding when to group test depends on the relative expense of tests versus stages, and could vary across testing sites.

Detailed discussion on the estimation of dilution effects is deferred here, as approaches will depend on the test response distribution modeling. It is possible to consider estimating response probabilities with uncertain disease status (reflected by posterior probabilities in classification) using Bayesian latent class estimation approaches that can include nonparametric density estimation (Tatsuoka (2002); Tatsuoka *and others* (2013)). Varied and well-characterized pooled test data should become more accessible as group testing becomes more widespread through population-level surveillance. Also, individual-level prior probabilities of positivity can for instance be estimated through generalized linear regression models based on demographic, geographic, clinical and exposure information. Finally, we conjecture that it will be possible to embed Bayesian halving algorithm-based testing sequences in high-volume, automated workflows,

which will reduce the practical burden of requiring higher number of stages.

5. Software

Source code available at

`https://github.com/ben072292/bayesgrouptestbackend`

`https://github.com/ben072292/bayesgrouptestfrontend`

The web tool referenced in Section 3 is available at

`https://bayesgrouptest.case.edu`.

6. Supplementary Material

Appendix referenced in Section 1 and Section 2 is available with this paper at the Biostatistics website on Wiley Online Library. Data sharing not applicable to this article as no datasets were generated or analysed during the current study.

Acknowledgments

This work was supported by grants NSF DRL1561716, R01 MH65538, NSF CCF-1822987 and UL1TR002548.

Conflict of Interest: None declared.

References

Bateman, Allen C., Mueller, Shanna, Guenther, Kyley and Shult, Peter. (2021). Assessing the dilution effect of specimen pooling on the sensitivity of sars-cov-2 pcr tests. *Journal of Medical Virology* **93**(3), 1568–1572.

Bilder, Christopher R., Tebbs, Joshua M. and McMahan, Christopher S. (2019). Informative group testing for multiplex assays. *Biometrics* **75**(1), 278–288.

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

Brookmeyer, Ron. (1999). Analysis of multistage pooling studies of biological specimens for estimating disease incidence and prevalence. *Biometrics* **55**(2), 608–612.

Donoho, David, Lofti, Masha and Ozturkler, Batu. (2020). The mathematics of mass testing for covid-19.

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

Ferguson, Thomas S. and Tatsuoka, Curtis. (2004). An optimal strategy for sequential classification on partially ordered sets. *Statistics Probability Letters* **68**(2), 161–168.

Gastwirth, Joseph L. (2000). The efficiency of pooling in the detection of rare mutations. *The American Journal of Human Genetics* **67**(4), 1036–1039.

Gastwirth, Joseph L. and Hammick, Patricia A. (1989). Estimation of the prevalence of a rare disease, preserving the anonymity of the subjects by group testing: application to estimating the prevalence of aids antibodies in blood donors. *Journal of Statistical Planning and Inference* **22**(1), 15–27.

Gastwirth, J. L. and Johnson, Wesley O. (1994). Screening with cost-effective quality control: Potential applications to hiv and drug testing. *Journal of the American Statistical Association* **89**(427), 972–981.

Ghosh, Sabyasachi, Rajwade, Ajit, Krishna, Srikar, Gopalkrishnan, Nikhil, Schaus, Thomas E., Chakravarthy, Anirudh, Varahan, Sriram, Appu, Vidhya, Ramakrishnan, Raunak, Ch, Shashank, Jindal, Mohit, Bhupathi, Vadhir, Gupta, Aditya, Jain, Abhinav, Agarwal, Rishi, Pathak, Shreya, Rehan, Mohammed Ali, Consul, Sarthak, Gupta, Yash, Gupta, Nimay, Agarwal, Pratyush, Goyal, Ritika, Sagar, Vinay, Ramakrishnan, Uma, Krishna, Sandeep, Yin, Peng, Palakodeti, Dasaradhi *and others*. (2020). Tapestry: A single-round smart pooling technique for covid-19 testing. *medRxiv*.

Graff, Lois E. and Roeloffs, Robert. (1972). Group testing in the presence of test error; an extension of the dorfman procedure. *Technometrics* **14**(1), 113–122.

Hogan, Catherine A., Sahoo, Malaya K. and Pinsky, Benjamin A. (2020). Sample pooling as a strategy to detect community transmission of sars-cov-2. *Jama* **323**(19), 1967.

Hughes-Oliver, Jacqueline M. and Swallow, William H. (1994). A two-stage adaptive group-testing procedure for estimating small proportions. *Journal of the American Statistical Association* **89**(427), 982–993.

Hung, M. and Swallow, William H. (1999). Robustness of group testing in the estimation of proportions. *Biometrics* **55**(1), 231–237.

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

Hwang, F. K. (1976). Group testing with a dilution effect. *Biometrika* **63**(3), 671–673.

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

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

Lohse, Stefan, Pfuhl, Thorsten, Berkó-Göttel, Barbara, Rissland, Jürgen, Geissler, Tobias, Gärtner, Barbara, Becker, Sören L, Schneitler, Sophie and Smola, Sigrun. (2020). Pooling of samples for testing for sars-cov-2 in asymptomatic people. *The Lancet Infectious Diseases* **20**(11), 1231–1232.

Majid, Farhan, Omer, Saad B and Khwaja, Asim Ijaz. (2020). Optimising sars-cov-2 pooled testing for low-resource settings. *The Lancet Microbe* **1**(3).

Remlinger, Katja S, Hughes-Oliver, Jacqueline M, Young, S. Stanley and Lam, Raymond L. (2006). Statistical design of pools using optimal coverage and minimal collision. *Technometrics* **48**(1), 133–143.

Sham, Pak, Bader, Joel, Craig, Ian, O'Donovan, Michael and Owen, Michael. (2002, 12). Dna pooling: A tool for large-scale association studies. *Nature reviews. Genetics* **3**, 862–71.

Shental, Noam, Levy, Shlomia, Wuvshet, Vered, Skorniakov, Shosh, Shalem, Bar, Ottolenghi, Aner, Greenshpan, Yariv, Steinberg, Rachel, Edri, Avishay, Gillis, Roni, Goldhirsh, Michal, Moscovici, Khen, Sachren, Sinai, Friedman, Lilach M., Nesher, Lior, Shemer-Avni, Yonat, Porgador, Angel *and others*. (2020). Efficient high-throughput sars-cov-2 testing to detect asymptomatic carriers. *Science Advances* **6**(37).

Sobel, Milton and Groll, Phyllis A. (1966). Binomial group-testing with an unknown proportion of defectives. *Technometrics* **8**(4), 631.

Tatsuoka, Curtis. (2002). Data analytic methods for latent partially ordered classification models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **51**(3), 337–350.

Tatsuoka, Curtis. (2005). Optimal sequencing of experiments in bayesian group testing. *Journal of Statistical Planning and Inference* **133**(2), 479 – 488.

Tatsuoka, Curtis. (2014). Sequential classification on lattices with experiment-specific response distributions. *Sequential Analysis* **33**(3), 400–420.

Tatsuoka, Curtis and Ferguson, Thomas. (2003). Sequential classification on partially ordered sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **65**(1), 143–157.

Tatsuoka, Curtis, Varadi, Ferenc and Jaeger, Judith. (2013). Latent partially ordered classification models and normal mixtures. *Journal of Educational and Behavioral Statistics* **38**(3), 267–294.

Tebbs, J. M. (2003). Estimating ordered binomial proportions with the use of group testing. *Biometrika* **90**(2), 471–477.

Thierry-Mieg, Nicolas. (2006, 02). A new pooling strategy for high-throughput screening: The shifted transversal design. *BMC bioinformatics* **7**, 28.

Ungar, Peter. (1960). The cutoff point for group testing. *Communications on Pure and Applied Mathematics* **13**(1), 49–54.

Xie, Minge, Tatsuoka, Kay, Sacks, Jerome and Young, S Stanley. (2001). Group testing with blockers and synergism. *Journal of the American Statistical Association* **96**(453), 92–102.

Zaharia, Matei, Chowdhury, Mosharaf, Franklin, Michael J., Shenker, Scott and Stoica, Ion. (2010, June). Spark: Cluster computing with working sets. In: *2nd USENIX Workshop on Hot Topics in Cloud Computing (HotCloud 10)*. Boston, MA: USENIX Association.

Zenios, Stefanos A. and Wein, Lawrence M. (1998). Pooled testing for hiv prevalence estimation: exploiting the dilution effect. *Statistics in Medicine* **17**(13), 1447–1467.

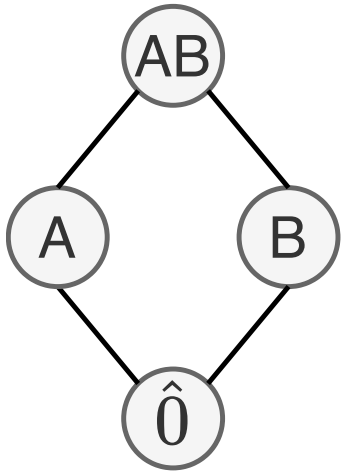


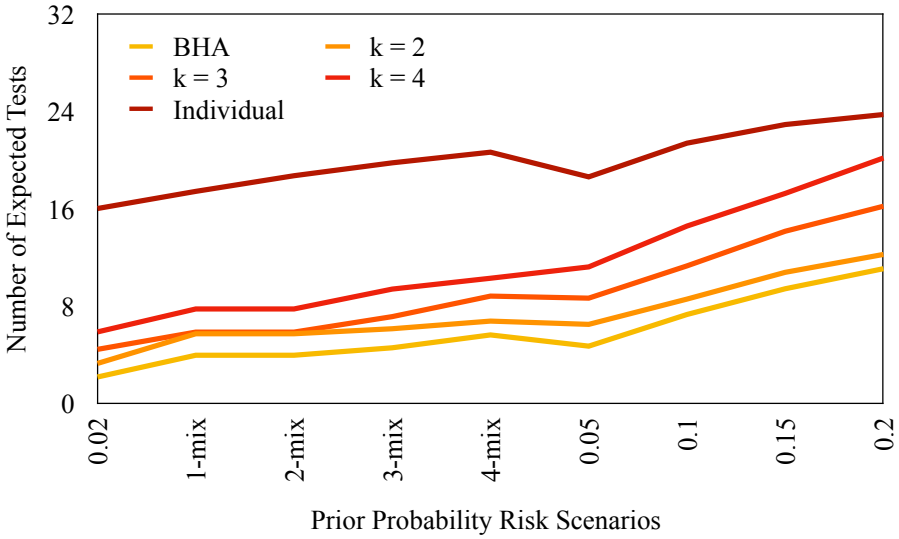
Fig. 1: Hasse diagram of lattice model with 2 subjects, A and B

Table 1: First stage test selections, $N = 12$.

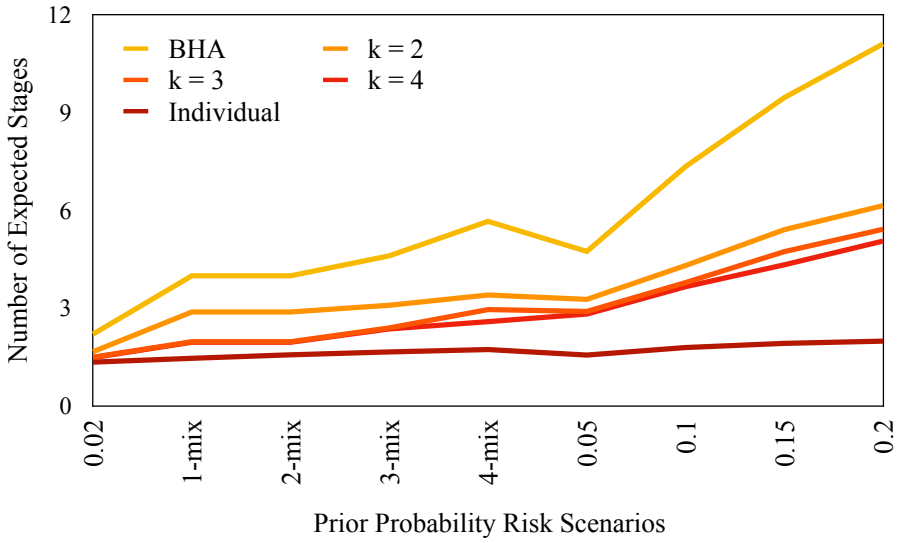
Simulation	$p_0 = 0.02$	$p_0 = 0.20$	mixed p_0
k=1	ABCDEFGHIJKL	ABC	ABCDEFGHIJKL
k=2 (add to k=1)	A	DEH	A
k=3 (add to k=1,2)	ABCDE	ABCDE	ABCDEFGH
k=4 (add to k=1,2,3)	ABCDEFGHI	ABCDH	ABCDEFGH

Table 2: Statistic comparison between BHA and individual testing. $N = 20$, $p_0 = 0.02$. Note the negative classification threshold = 0.001.

	BHA	Individual
False Positive Rate	0.041%	0.0%
False Negative Rate	0.42%	0.33%
Expected Number (SD) of Tests	4.40 (3.13)	30 (0.55)
Expected Number (SD) of Stages	4.40 (3.13)	1.5 (0.0275)



(a) Expected number of tests



(b) Expected number of stages

Fig. 2: Comparison of expected number of stages and expected number of tests among five test selection rules. $N = 12$.

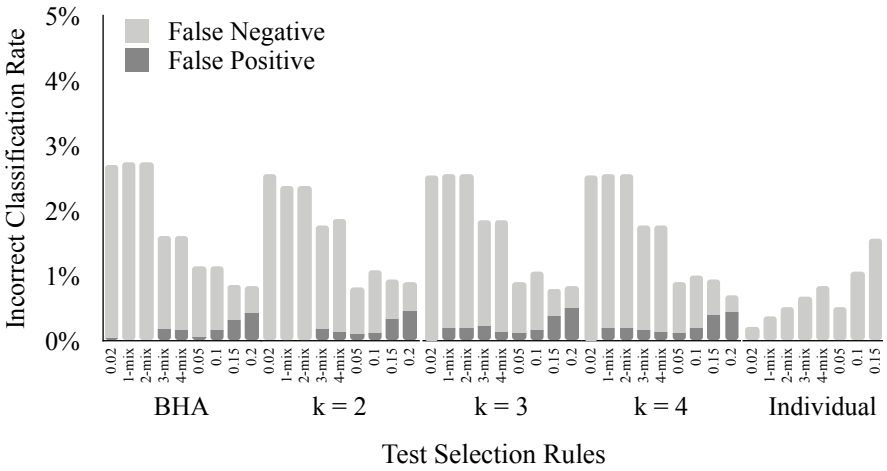


Fig. 3: False positive and false negative rates. $N = 12$.

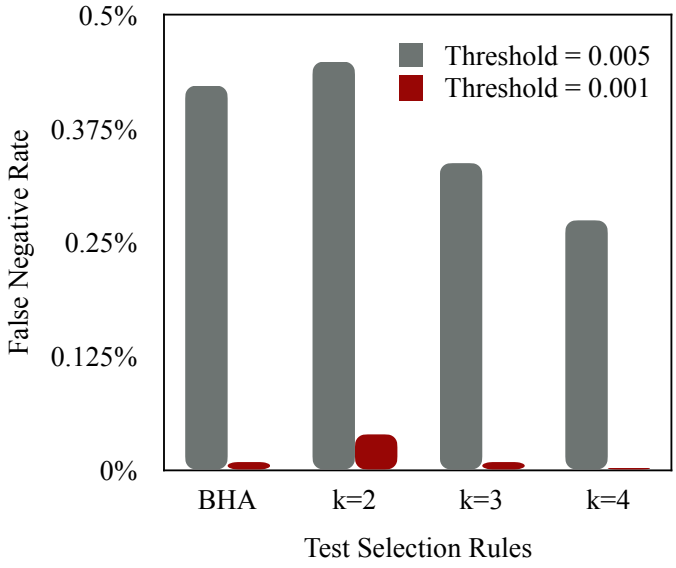
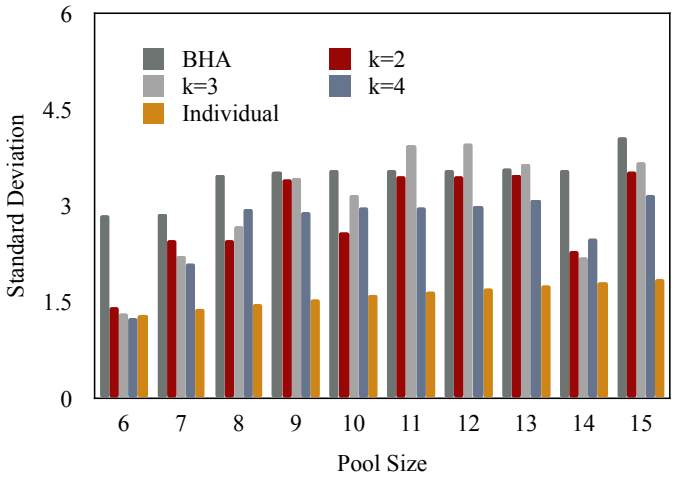
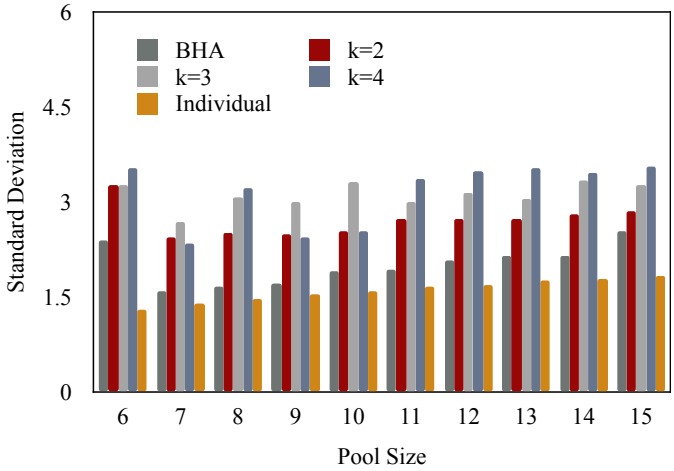


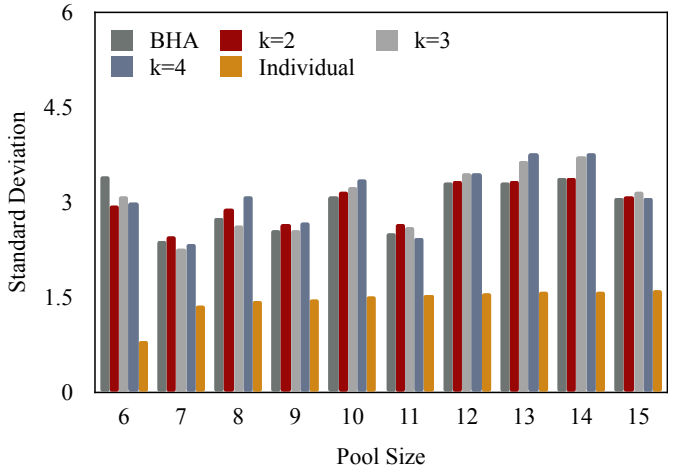
Fig. 4: False negative rate comparison between using negative threshold of 0.005 versus 0.001. $N = 12$, $p_0 = 0.2$



(a) 0.02

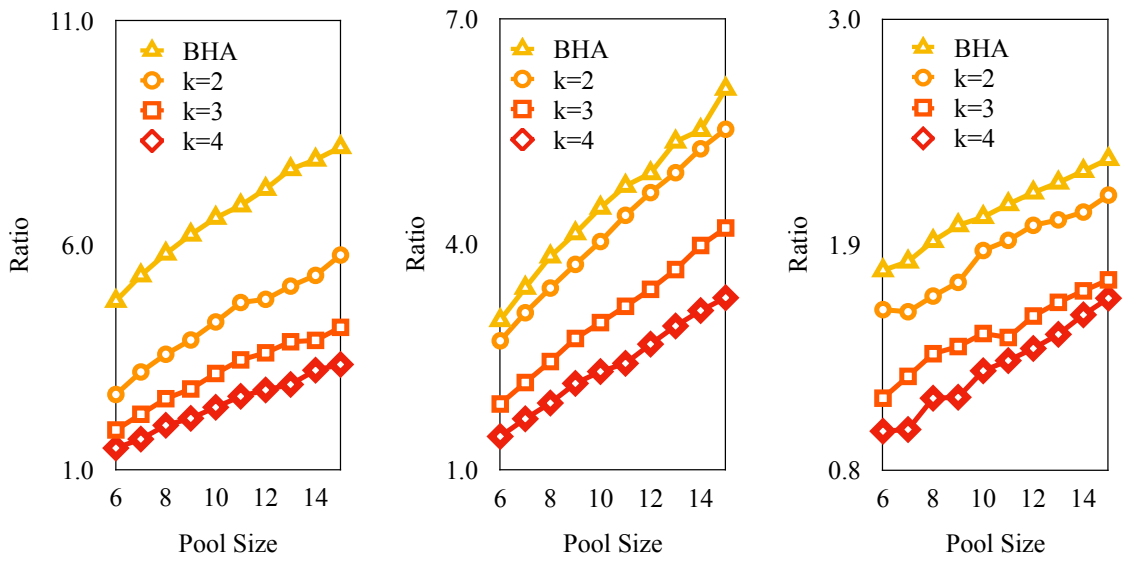


(b) std-1-mix



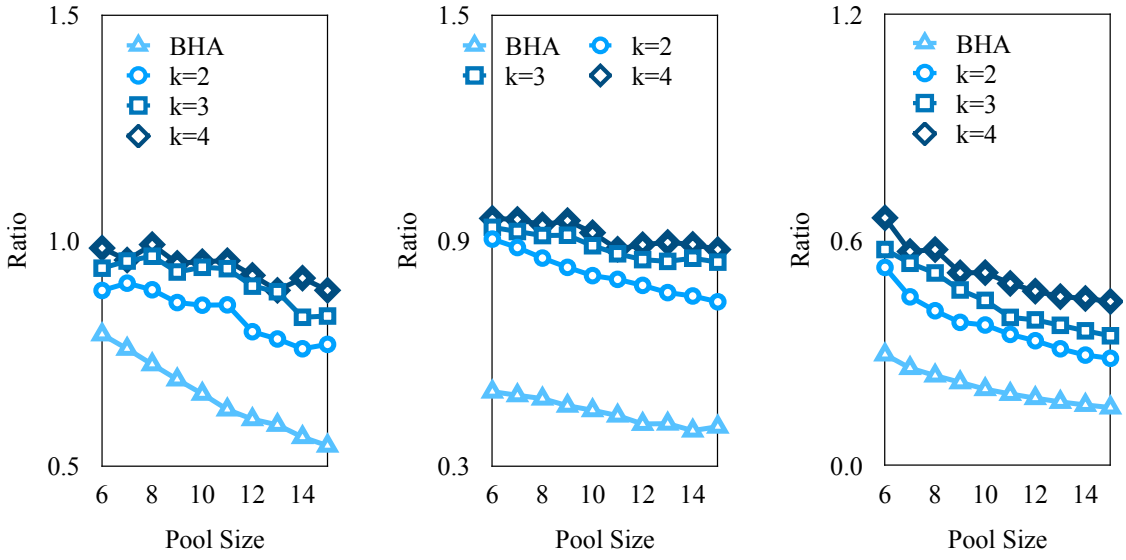
(c) 0.2

Fig. 5: Standard deviations of number of tests. $N = 12$.



(a) 0.02 (b) 1-mix (c) 0.2

Fig. 6: Ratio of expected number of tests compared to individual testing



(a) 0.02 (b) 1-mix (c) 0.2

Fig. 7: Ratio of expected number of stages compared to individual testing