Harvard University
Harvard University Biostatistics Working Paper Series

Year 2008 *Paper* 57

A Matrix Pooling Algorithm for Disease Detection

Bethany L. Hedt* Marcello Pagano†

*Harvard School of Public Health, bhedt@hsph.harvard.edu

†Harvard University, pagano@hsph.harvard.edu

A Matrix Pooling Algorithm for Disease Detection

B. L. Hedt and M. Pagano

April 24, 2008

Abstract

In this paper, we introduce a new pooling algorithm for testing for a disease, matrix pooling, which accommodates imperfect tests. We find that matrix pooling belongs to the class of pooling methods that have greater accuracy than individual testing, under reasonable levels of test kit sensitivity and specificity. Additionally, this increase in accuracy is achieved with fewer tests than individual testing for low prevalences of disease. Indeed, the savings can be considerable when dealing with very low prevalence situations, making screening for some diseases more realistic via our proposed technique.

1 INTRODUCTION

Diagnostic testing is the cornerstone of patient treatment and care. By suitably identifying individuals infected with a disease, especially in the early stages of infection, physicians can often offer a timely and comprehensive response. This is especially critical if early detection would prove beneficial to the patient or, as in the case of infectious diseases, might delay or slow down the spread of a disease to others. The desiderata of any testing procedure are high accuracy and low cost. The absence of the latter is one reason why despite the obvious benefits, universal screening for many diseases is not realistic due to limited resources, and often times infected individuals are not identified until the disease has progressed, sometimes too far for intervention.

Dorfman introduced pooled testing as an ingenious method of identifying individuals infected with syphilis at a reduced cost [1]. With the Dorfman pooling procedure, n samples are grouped into a pool and the pool is tested. If the result is negative, then all samples in the pool are classified as disease negative. If the test result is positive, then the individuals in the pool are tested separately. This testing algorithm is less expensive than individual testing in low prevalence settings because, in most cases, only one test will be performed instead of n tests. The larger the pool, the fewer expected

number of tests per result; however, once the pool size is so large that one or more samples within the pool are expected to be infected, the benefit of fewer tests is lost. Therefore the selection of n depends on the prevalence. Additionally, one is also constrained by the desire to avoid a pool which is so large that the test loses the ability to detect an infected individual. Since Dorfman's initial work, pooling methods have evolved to incorporate a number of testing properties and algorithms, with two primary goals: classification, or identifying individuals with a disease, and estimation of prevalence of disease in the general population. For a review of pooling methods, see Lancaster and Keller-McNulty, Venette, Moon and Hutchison, Zenios and Wein, and references therein [2, 3, 4].

While pooling samples reduces the cost of testing for disease in a low prevalence setting, most of the original pooling algorithms proposed, including the Dorfman pooling method, have the major disadvantage of a higher false negative predictive value than individual testing because the overall sensitivity is compromised by the procedure. This decline in accuracy has limited the popularity of pooling methods in practice. In order to address these concerns, Litvak, Tu and Pagano propose a pooling algorithm, T_2^+, that increases sensitivity by retesting pools that test negative at all stages [5],[6]. The first step of T_2^+ begins with testing a pool of n samples. If the pool tests negative, then the entire pool is retested. If the pool tests negative a second time, then the entire pool is declared uninfected. If the pool tests positive either at the first test or the retest, then the pool is divided into two subpools of equal size if the original pool size, n, is even, and into sizes $(n-1)/2$ and $(n+1)/2$ if n is odd, and the subpools are then treated like the original pool. Henceforth this splitting process will be referred to as the "halving step". To identify infected individuals, this testing procedure continues until each sample is determined to be positive or negative for disease. By retesting negative pools, the overall sensitivity is not only preserved but is improved compared to individual testing while maintaining the benefit of fewer tests in low prevalence situations.

Pooling algorithms with retesting form pools by grouping the same individuals or a subset of the same individuals. However, if there is a risk of sympathetic or antithetic paired behavior, then there are potential benefits in testing pools with distinct compositions. One design that achieves this, at least at the first stage, arranges samples in a matrix-like design and forms column and row pools, so that at the first stage of testing, a pair of samples will occur together in at most one pool. Additionally, a matrix pooling testing algorithm can often be executed faster than other retesting algorithms because the initial double testing is conducted in parallel instead of in sequence. Noting these benefits, and recognizing the potential for additional advantages, we explore the properties of one matrix pooling design. Our algorithm introduced here, MT_2^+, naturally builds on the T_2^+ method

and is described in Section 2 together with the outline of the derivations of the sensitivity, specificity and expected number of tests. Finally, in Section 3 we compare the properties of the matrix pooling algorithm to individual testing, and show the superiority of matrix pooling with respect to increased accuracy and a decrease in the expected number of tests per result at low prevalence settings.

2 MATRIX POOLING

Consider a disease where an individual is either infected or uninfected, and a test that can identify a sample from an infected individual as positive with probability S_e and correctly identify a sample from an uninfected individual with probability S_p, the sensitivity and specificity of the test respectively. We assume that for a pool of n samples, that if all of the samples are uninfected, then the pooled sample will be uninfected but if any of the samples are infected, then the pooled sample will be infected. Additionally, we assume that the sensitivity and specificity of the test is preserved under pooling; in other words, if an infected sample would test positive individually with probability S_e, then upon combining this sample with $n-1$ other samples in a pool, the pool will still test positive with probability S_e. Also, a pool of samples from n uninfected individuals will test negative with probability S_p. One further assumption made throughout the calculations is that repeated tests are independent.

The method we propose, called matrix pooling, or MT_2^+, extends the T_2^+ algorithm and is a matrix testing method that accommodates imperfect tests. With MT_2^+, n^2 samples are randomly placed in an $n \times n$ matrix in order to form two sets of pools: the n pooled rows, $r_1, \ldots, r_n$, and the n pooled columns, $c_1, \ldots, c_n$.

The MT_2^+ testing algorithm is defined as follows. Test all the row pools and all the column pools. Then,

1. if all $2n$ test negative, stop the testing and declare all n^2 samples to be disease free.
2. if r rows (r=1,...,n) and no columns test positive, then subject each of these r pools to further testing using T_2^+ applying a halving step first, since they have already tested positive. The remaining $(n-r) \times n$ samples are declared disease free.
3. if c columns (c=1,...,n) and no rows test positive, then similar to step 2 above, subject each of these c pools to further testing using T_2^+ applying a halving step first; since they have already tested positive. The remaining $(n-c) \times n$ samples are declared disease free.

4. if n rows and n columns test positive, then submit the n row pools to further testing using T_2^+ applying a halving step first, since they have already tested positive.

5. otherwise, if r rows (r=1,...,n) and c columns (c=1,...,n), but $rc \neq n^2$, test positive, then declare those samples that appear in rows and columns that have both tested negative to be disease free. Subject each of the r pools of size $n-c$, made up of the samples in the r rows that test positive and in the $n-c$ columns that test negative; and each of the c pools of size $n-r$ in the c columns that test positive and in the $n-r$ rows that test negative to further testing using T_2^+. Finally, individually test the remaining rc samples, retesting if the initial result is negative.

2.1 Sensitivity of Matrix Pooling

Let $pos_n^{(1)}(r, c|p)$ be the probability that r of the first $n-1$ row pools and c of the first $n-1$ column pools test positive, given p is the prevalence of disease (derivation shown in Appendix A), and let $\psi_{T_2^+(k)}$ denote the sensitivity of T_2^+ with a pool of size k (derived by Litvak et al [5]).

Theorem 1 (Sensitivity)**.** *The sensitivity of the matrix pooling algorithm, denoted $\psi_{MT_2^+(n|p)}$, is the probability that an infected sample is identified as disease positive at the end of the testing process. The sensitivity of MT_2^+ is a function of the matrix size and is dependent on the prevalence of disease and can be written as*

$$\begin{aligned}\psi_{MT_2^+(n|p)} &= 2S_e\left(1-S_e\right)\left\{\sum_{r=0}^{n-1} pos_n^{(1)}(r,0|p)\left(\frac{a}{n}\psi_{T_2^+(a)}+\frac{b}{n}\psi_{T_2^+(b)}\right)+\right.\\ &\qquad\left.\sum_{r=0}^{n-1}\sum_{c=1}^{n-1} pos_n^{(1)}(r,c|p)\,\psi_{T_2^+(n-c)}\right\}+\\ &\qquad S_e^2\left[pos_n^{(1)}(n-1,n-1|p)\left(\frac{a}{n}\psi_{T_2^+(a)}+\frac{b}{n}\psi_{T_2^+(b)}\right)+\right.\\ &\qquad\left.\left\{\sum_{r=0}^{n-1}\sum_{c=0}^{n-1} pos_n^{(1)}(r,c|p)-pos_n^{(1)}(n-1,n-1|p)\right\}\left(2S_e-S_e^2\right)\right].\end{aligned}$$

Proof Without loss of generality, fix the (n,n) sample of the initial testing matrix to be infected. In order for this sample to be identified as positive, then it is necessary that at the end of the first

step either the n^{th} row pool tests positive and the n^{th} column pool tests negative, or the n^{th} row pool tests negative and the n^{th} column pool tests positive, both of which happen with probability $S_e(1-S_e)$; or both the n^{th} row and column pools test positive, which happens with probability S_e^2.

First suppose that the n^{th} row pool tests positive and the n^{th} column pool tests negative. If no other column pools test positive, then for any combination of positive row pools, $r = 0, \ldots n-1$, the n^{th} row is submitted to T_2^+ with a halving step for further testing, with the probability of the infected sample being classified as positive equal to $(a/n)\psi_{T_2^+(a)} + (b/n)\psi_{T_2^+(b)}$. Here, $a = \lceil \frac{n}{2} \rceil$ and $b = \lfloor \frac{n}{2} \rfloor$. If one or more of the other column pools test positive at the initial step, $c = 1, \ldots, n-1$, for any combination of positive row pools, $r = 0, \ldots n-1$, the n^{th} row, excepting the c samples at the intersection of positive column tests, is submitted to T_2^+ for further testing. In this case, the probability of the infected sample being identified as positive is equal to $\psi_{T_2^+(n-c)}$. Using this logic, the probability that the infected sample is identified as positive for disease in the case when the n^{th} row pool tests positive and the n^{th} column pool tests negative is

$$\sum_{r=0}^{n-1} pos_n^{(1)}(r, 0|p) \left(\frac{a}{n}\psi_{T_2^+(a)} + \frac{b}{n}\psi_{T_2^+(b)} \right) + \sum_{r=0}^{n-1}\sum_{c=1}^{n-1} pos_n^{(1)}(r, c|p)\, \psi_{T_2^+(n-c)}. \tag{1}$$

Note that the same reasoning applies to the situation when the n^{th} row pool tests negative and the n^{th} column pool tests positive. Since the initial testing matrix has equal numbers of samples in the rows and columns, the probabilities that the infected sample is identified as positive in these two situations are equal, hence each term in Equation (1) above is multiplied by two.

Now suppose that both the n^{th} row pool and the n^{th} column pool test positive. Then if all of the other row pools and column pools test positive, so $r = n-1$ and $c = n-1$, then the n^{th} row is submitted to T_2^+ with a halving step for further testing, with the probability of the infected sample being classified as positive equal to $(a/n)\psi_{T_2^+(a)} + (b/n)\psi_{T_2^+(b)}$. If some combination of the other rows and columns test positive, $r, c = 0, \ldots, n-1, rc \neq (n-1)^2$, then the (n, n) sample is tested individually with the probability of being identified as positive equal to $2S_e - S_e^2$. Combining these four possible ways of identifying an infected sample as positive leads to the sensitivity as stated in Theorem 1.

□

2.2 Specificity of Matrix Pooling

Let r_n^+ and c_n^+ be the events that the n^{th} row and n^{th} column test positive given that the (n,n) cell is uninfected, respectively, and r_n^- and c_n^- be the events that the n^{th} row and column test negative given that the (n,n) cell is uninfected. Let p_B^* be the prevalence of disease in samples at the intersection of negative column tests and the n^{th} row, given then n^{th} row tests positive; and let p_A^* be the prevalence of disease in samples at the intersection of positive column tests and the last row, given the last row tests positive, or the intersection of positive row tests and the last column, given the last column tests positive (see Appendix B). Also, let $\phi_{T_2^+(n|p)}$ be the specificity of the T_2^+ when applied to a pool of size n given a prevalence p (see [5]).

Theorem 2 (Specificity)**.** *The specificity of matrix pooling, $\phi_{MT_2^+(n|p)}$, is the probability that an uninfected sample is identified as disease free at the end of the testing algorithm, and is dependent on both the matrix size and prevalence of disease. The specificity can be written as*

$$\begin{aligned}
\phi_{MT_2^+(n|p)} \;=\; 1 - \Bigg(& 2P(r_n^+)P(c_n^-)\left[\sum_{c=1}^{n-1}\sum_{r=0}^{n-1} pos_n^{(1)}(r,c|p)\left(1-\phi_{T_2^+(n-c|p_B^*)}\right) + \right. \\
& \left. \sum_{r=0}^{n-1} pos_n^{(1)}(r,0|p)\left\{\frac{a}{n}\left(1-\phi_{T_2^+(a|p_B^*)}\right)+\frac{b}{n}\left(1-\phi_{T_2^+(b|p_B^*)}\right)\right\}\right] + \\
& P(r_n^+)\,P(c_n^+)\left[pos_n^{(1)}(n-1,n-1|p)\left\{\frac{a}{n}\left(1-\phi_{T_2^+(a|p_A^*)}\right)+\frac{b}{n}\left(1-\phi_{T_2^+(b|p_A^*)}\right)\right\} + \right. \\
& \left. (1-S_p^2)\left\{\sum_{r=0}^{n-1}\sum_{c=0}^{n-1} pos_n^{(1)}(r,c|p) - pos_n^{(1)}(n-1,n-1|p)\right\}\right]\Bigg). \qquad (2)
\end{aligned}$$

Proof Without loss of generality, fix the (n,n) sample of the initial testing matrix to be uninfected. The derivation below shows the probability that the uninfected sample will be identified as positive using matrix pooling. One minus this probability gives the specificity of MT_2^+. In order for this uninfected sample to be identified as positive then at the end of the first step, either the n^{th} row pool tests positive and the n^{th} column pool tests negative, or the n^{th} row pool tests negative and the n^{th} column pool tests positive, both of which happen with probability $P(r_n^+)P(c_n^-) = P(r_n^-)P(c_n^+)$; or both the n^{th} row and column pools must test positive, which happens with probability $P(r_n^+)P(c_n^+)$.

Given the (n, n) cell is fixed to be uninfected,

$$\begin{aligned} P(r_n^+) &= (1-p)^{n-1}(1-S_p) + (1-(1-p)^{n-1})S_e = P(c_n^+) \\ P(r_n^-) &= (1-p)^{n-1}S_p + (1-(1-p)^{n-1})(1-S_e) = P(c_n^-). \end{aligned}$$

First suppose that the n^{th} row pool tests positive and the n^{th} column pool tests negative. If no other column pools test positive, then for any combination of positive row pools, $r = 0, \ldots n-1$, the n^{th} row is submitted to T_2^+ with a halving step for further testing, with the probability of the uninfected sample being classified as positive equal to $(a/n)\left(1-\phi_{T_2^+(a|p_B^*)}\right)+(b/n)\left(1-\phi_{T_2^+(b|p_B^*)}\right)$. Here, the last term is one minus the specificity of T_2^+ when applied to samples of sizes a and b with prevalence p_B^*. If one or more of the other column pools test positive at the initial step, $c = 1, \ldots, n-1$, for any combination of positive row pools, $r = 0, \ldots n-1$, the n^{th} row, excepting the c samples at the intersection of positive column tests, is submitted to T_2^+ for further testing, with the probability of the uninfected sample being identified as positive equal to $1-\phi_{T_2^+(n-c|p_B^*)}$. Thus, the probability that the uninfected sample is identified as positive for disease in the case when the n^{th} row pool tests positive and the n^{th} column pool tests negative is

$$\sum_{r=0}^{n-1} pos_n^{(1)}(r,0|p)\left\{\frac{a}{n}\left(1-\phi_{T_2^+(a|p_B^*)}\right)+\frac{b}{n}\left(1-\phi_{T_2^+(b|p_B^*)}\right)\right\}+ \sum_{c=1}^{n-1}\sum_{r=0}^{n-1} pos_n^{(1)}(r,c|p)\left(1-\phi_{T_2^+(n-c|p_B^*)}\right). \tag{3}$$

Note that the same reasoning applies to the situation when the n^{th} row pool tests negative and the n^{th} column pool tests positive, and since the initial testing matrix has equal numbers of samples in the rows and columns, the probabilities that the uninfected sample is identified as positive in either of these situations are equal, hence each term in Equation (3) is multiplied by two.

Now suppose that both the n^{th} row pool and the n^{th} column pool test positive. Then if all of the other row pools and column pools test positive, $r = c = n-1$, then the n^{th} row is submitted to T_2^+ with a halving step for further testing, with the probability of the uninfected sample being classified as positive equal to

$$\frac{a}{n}\left(1-\phi_{T_2^+(a|p_A^*)}\right)+\frac{b}{n}\left(1-\phi_{T_2^+(b|p_A^*)}\right),$$

where given the (n, n) cell is uninfected and p_A^*. If some combination of the other rows and columns

test positive, $rc \neq (n-1)^2$, then the (n,n) sample is tested individually with the probability of being identified as positive equal to $1-S_p^2$. Combining these four scenarios yields the proof of the theorem.

□

2.3 Expected Number of Tests for Matrix Pooling

Let $pos_n(r,c|p)$ denote the probability that r out of n row pools and c out of n column pools test positive when the disease prevalence is p (Appendix C). Let p_B denote the prevalence of disease in samples that are at the intersection discordant row pool and column pool tests and p_A the prevalence of disease in samples that are at the intersection of concordant positive row and column pools (Appendix D). Finally, $\mathcal{E}(N)_{T_2^+(k|p)}$ is the expected number of tests for T_2^+ with a pool size k and prevalence p (see Litvak et al [5]).

Theorem 3 (Expected Number of Tests)**.** *$E(N)_{MT_2^+(n|p)}$ denotes the expected number of tests for matrix pooling with an $n \times n$ testing matrix and disease prevalence, p, then*

$$\begin{aligned}
\mathcal{E}(N)_{MT_2^+(n|p)} \;=\; & 2n\,[pos_n(0,0|p)+ \\
& \sum_{r=1}^{n} r\left\{\mathcal{E}(N)_{T_2^+(a|p_B)}+\mathcal{E}(N)_{T_2^+(b|p_B)}\right\} pos_n(r,0|p)+ \\
& \sum_{c=1}^{n} c\left\{\mathcal{E}(N)_{T_2^+(a|p_B)}+\mathcal{E}(N)_{T_2^+(b|p_B)}\right\} pos_n(0,c|p)+ \\
& n\left\{\mathcal{E}(N)_{T_2^+(a|p_A)}+\mathcal{E}(N)_{T_2^+(b|p_A)}\right\} pos_n(n,n|p)+ \\
& \sum_{r=1}^{n}\sum_{c=1}^{n}(1-\delta_{rc,n^2})\left\{r\,c\,\mathcal{E}(N)_{T_2^+(1|p_A)}+r\,\mathcal{E}(N)_{T_2^+(n-c|p_B)}+\right. \\
& \left. c\,\mathcal{E}(N)_{T_2^+(n-r|p_B)}\right\} pos_n(r,c|p)\Big],
\end{aligned} \tag{4}$$

with $\delta_{y,z}=0$ if $y=z$.

Proof Initially n row pools and n column pools are tested, resulting in $2n$ total tests. Now we calculate the additional expected number of tests for the matrix pooling algorithm by enumerating over each possible combination of r row and c column pools testing positive times the probability that r row pools and c column pools test positive.

- If all the row and column pools test negative, with probability $pos_n(0,0|p)$, then no further

tests are conducted.

- If r $(= 1, \ldots, n)$ row pools test positive but no column pools test positive, which occurs with probability $pos_n(r, 0|p)$, then each row corresponding to a positive row pool test is submitted to T_2^+ with a halving step so that an additional $r\left\{\mathcal{E}(N)_{T_2^+(a|p_B)} + \mathcal{E}(N)_{T_2^+(b|p_B)}\right\}$ tests are expected. Similarly, if c $(= 1, \ldots, n)$ column pools test positive but no row pools test positive, then each column corresponding to a positive column pool test is submitted to T_2^+ with a halving step so that an additional $c\left\{\mathcal{E}(N)_{T_2^+(a|p_B)} + \mathcal{E}(N)_{T_2^+(b|p_B)}\right\}$ tests are expected.
- If all row pools and all column pools test positive, then each of the n rows is submitted to T_2^+ with a halving step, increasing the expected number of tests by $n\left\{\mathcal{E}(N)_{T_2^+(a|p_A)} + \mathcal{E}(N)_{T_2^+(b|p_A)}\right\}$.
- Finally, if some combination of r rows and c columns test positive $(rc \neq n^2)$, then the rc samples at the intersection of the positive row and column tests are retested individually, confirming negative tests with a retest. This requires, on average, $rc\left\{\mathcal{E}(N)_{T_2^+(1|p_A)}\right\}$ additional tests. Each positive row, minus the samples that intersect with positive column tests, are submitted to T_2^+ adding $r\left\{\mathcal{E}(N)_{T_2^+(n-c|p_B)}\right\}$ tests, on average. And likewise, each positive column, minus the samples that intersect with positive row tests, are submitted to T_2^+ adding $c\left\{\mathcal{E}(N)_{T_2^+(n-r|p_B)}\right\}$ tests, on average.

Aggregating across all combinations of r and c completes the proof.

□

3 COMPARISON OF MATRIX POOLING TO INDIVIDUAL TESTING

For the purpose of comparison, individual testing is defined as performing one test on each sample and classifying the individual's disease status based on the results of this one test. For many diagnostic settings, individual testing is the standard. The sensitivity and specificity of individual testing are the same as for the test, S_e and S_p, and individual testing requires one test per result.

Figures 1–3 compare the sensitivity, specificity and expected number of tests of the matrix pooling algorithm, with matrix sizes 4×4, 8×8, 12×12 and 16×16, to individual testing. The overall algorithm sensitivity for 5% disease prevalence and 95% specificity, is presented in Figure 1 for a range of single test sensitivity of 80–100%. The 45^o line represents the sensitivity of

individual testing. It is clear from this figure that the sensitivities of matrix pooling for the different matrix sizes are improved over the sensitivity of individual testing, for this fixed prevalence level and specificity. Additionally, for most low prevalence levels and reasonable range of specificities, improvement in the sensitivity of MT_2^+ is still observed and often times even more pronounced (not shown). Figure 2 shows the matrix pooling specificity, for 5% prevalence and individual test sensitivity of 95%, for a range of specificities between 80–100%. Again, there is clear improvement of the matrix pooling algorithm, at the fixed prevalence and sensitivity, with respect to specificity as compared to individual testing, and this improvement continues, and in some settings increases, for low prevalences and reasonable sensitivities.

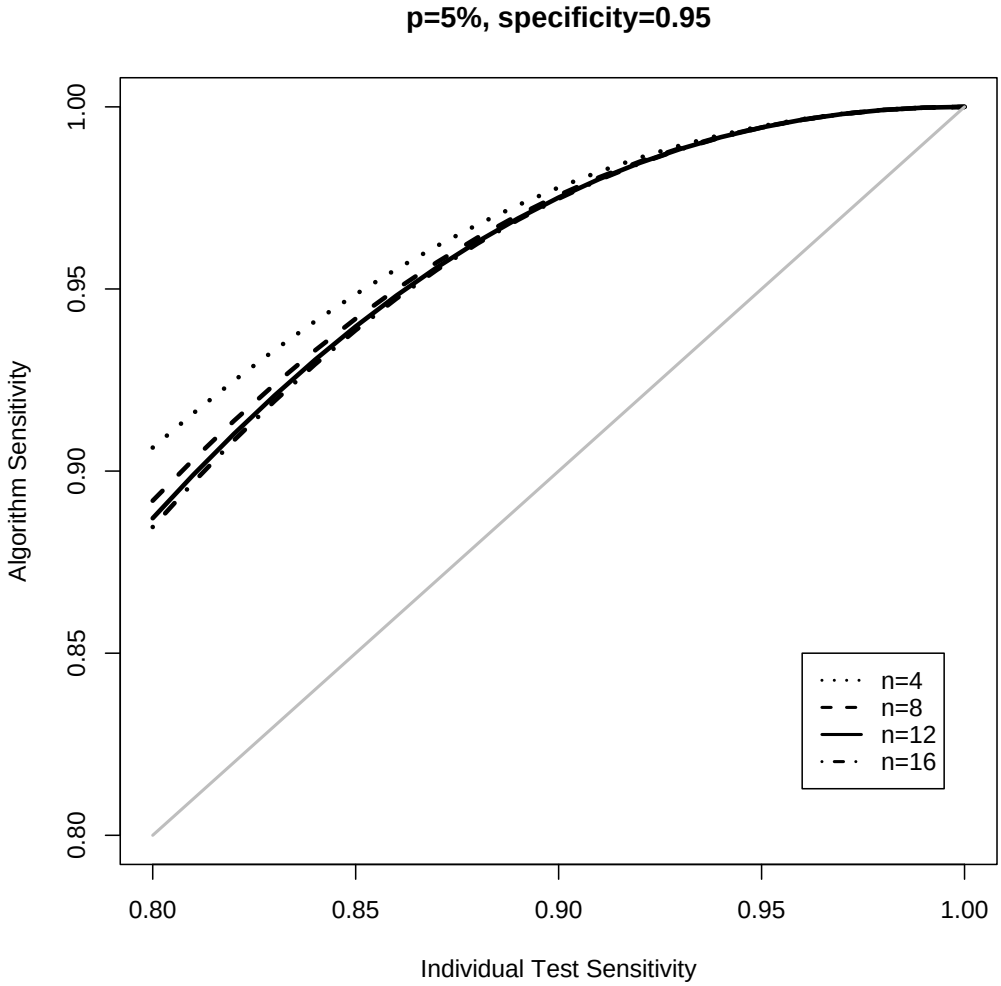

Figure 1: The Sensitivity of Matrix Pooling Compared to Individual Testing.

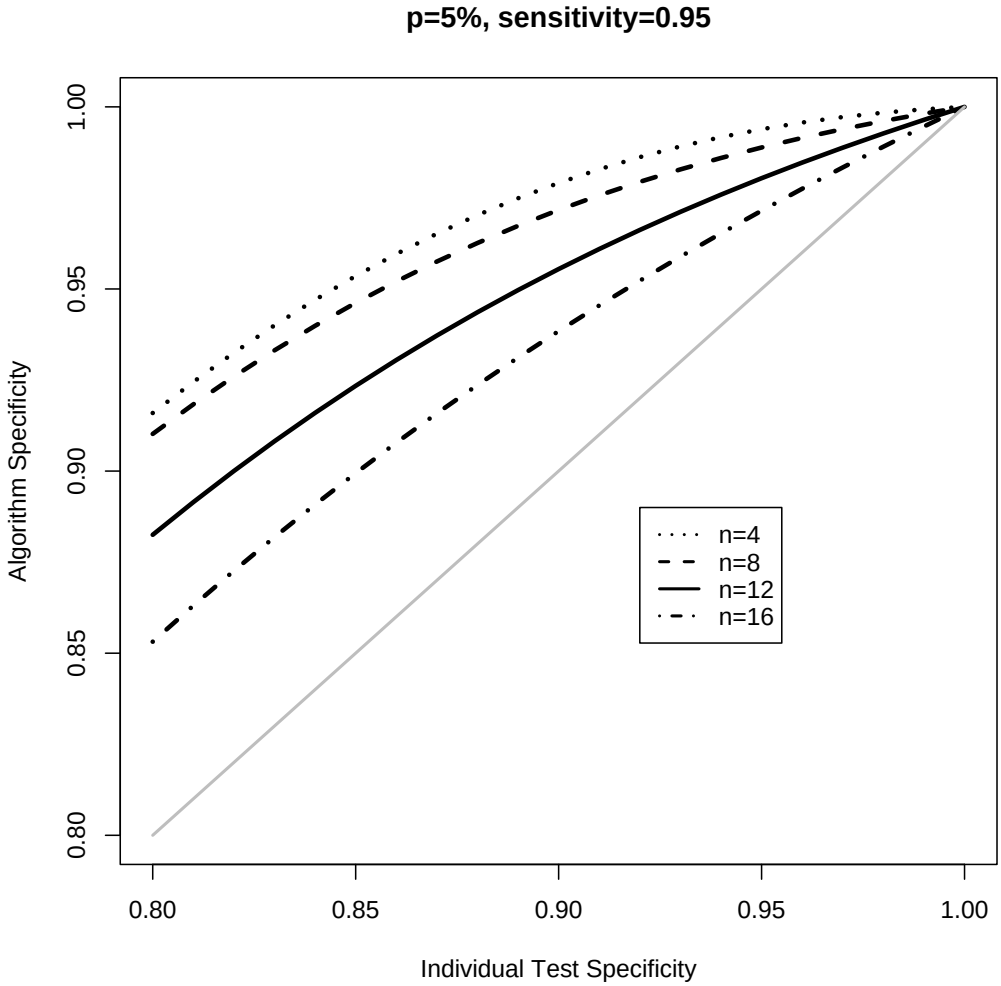

Figure 2: The Specificity of Matrix Pooling Compared to Individual Testing.

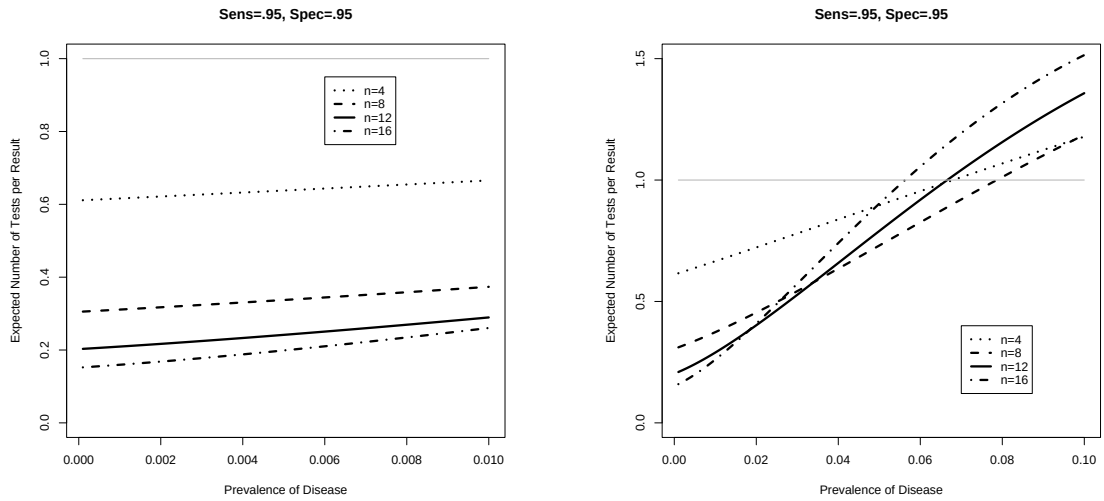

Figure 3: The Expected Number of Tests of Matrix Pooling Compared to Individual Testing. The left graph shows the expected number of tests for the very low prevalence levels (0.01–1.0%) and the right for moderate prevalence levels (0.1–10.0%).

Finally, Figure 3 shows the expected number of tests per result for matrix pooling at a fixed sensitivity and specificity of 95%. The graph on the left shows the expected number of tests for a range of disease prevalence between 0.01–1.0%, and the graph on the right for prevalences between 0.1–10%. In both graphs the gray horizontal line at one represents the expected number of tests for individual testing, since this does not depend on the prevalence of disease. As with all pooling algorithms, if the prevalence of disease is too high, the benefits of matrix pooling with regards to expected number of tests disappear. At this fixed sensitivity and specificity, matrix pooling requires fewer tests for all four matrix sizes until the prevalence of disease reaches 5%. Since the expected number of tests is a function of the sensitivity and specificity, it is possible at these higher prevalence levels, that different combinations of individual test sensitivity and specificity will lead to different optimal matrix sizes or for a different prevalence cutoffs when individual testing requires fewer tests than MT_2^+. However, for the lowest prevalences, matrix pooling always requires fewer tests when compared to individual testing, and the larger the matrix size, the greater the savings. Also, at these low levels of disease, the savings are observed for varying levels of sensitivity and specificity.

4 CONCLUSION

In 1994, Phatarfod and Sudbury described two testing algorithms, SA1 and SA2, that arrange mn samples into a $m \times n$ rectangular array [7]. Xie et al describe the properties of a similar testing

algorithm under the conditions of blocker and synergistic effects between samples [8]. It is important to note these testing algorithms because the initial steps of the matrix pooling algorithm are very similar, but MT_2^+ deviates after the first step. Additionally, both papers only describe their methods under the assumption of perfect sensitivity and specificity, while the MT_2^+ accommodates the more realistic imperfect sensitivity and specificity.

Matrix pooling offers benefits over many of the traditional pooling methods because while pooling keeps costs of testing low, retesting all negative tests preserves and often improves the overall sensitivity of the testing procedure, lowering the false negative predictive value and often times the false positive predictive value as well. Other authors have explored pooling with repeat testing and found similar improved results ([5],[6],[9]). However, MT_2^+ offers two marked benefits over these methods. First, the repeat testing at the first pooling stage occurs simultaneously instead of sequentially, leading to faster test results. Secondly, while each sample is tested twice, the sample is tested in two distinct pools with different companion samples in the first testing stage of matrix pooling, which might help if the assumption that test sensitivity is preserved under pooling is compromised.

One limitation of the matrix pooling method is that these methods are complicated and more difficult to implement in practice. However, it would be possible to automate this procedure to facilitate the testing and minimize the introduction of human error. These methods are also limited because of the strong, yet testable, assumption that diluting an infected sample by pooling will not change the ability of a test to detect the infected sample. Additionally, the methods above can be extended to reflect a varying sensitivity, S_e, if this assumption is truly violated for a particular disease.

The matrix pooling algorithm described here provides the desired benefit of reducing the expected number of tests when compared to individual testing at low prevalence settings whilst improving the overall sensitivity and specificity of the testing process. While it is impossible to compare the properties of our method to every diagnostic testing algorithm, we show elsewhere that although matrix pooling has a higher expected number of tests at the lowest prevalence settings compared to the Dorfman and two-stage Dorfman testing procedures, there is a great improvement in the sensitivity using MT_2^+, which leads to a far lower false negative predictive value [10]. Because of the improved accuracy and decreased expected number of tests when compared to individual testing, matrix pooling has the potential to make screening for some rare diseases a reality, even in the case when the tests are expensive.

Appendix A

In order to derive $pos_n^{(1)}(r, c|p)$, the probability that r of the first $n-1$ rows and c of the first $n-1$ columns test positive, given prevalence of disease p, we will show:

1. $p_n(j, k|m)$, the probability that j of the n rows and k of the n columns contain at least one infected sample, given that the $n \times n$ testing matrix contains m infected samples,
2. $p_n^{(1)}(j, k|y)$, the probability that j of the first $n-1$ rows and k of the first $n-1$ columns contain at least one infected sample, given that the $n \times n$ testing matrix, excepting the (n, n) cell, contains y infected samples,
3. $p_n^{(1)}(j, k|p)$, the probability that j of the first $n-1$ rows and k of the first $n-1$ columns contain at least one infected sample, given the prevalence of disease p in the $n \times n$ testing matrix, excepting the (n, n) cell,
4. and finally, $pos_n^{(1)}(r, c|p)$.

Lemma A.1. *Let $p_n(j, k|m)$ denote the probability that j rows and k columns of an $n \times n$ matrix have at least one infected sample given m out of n^2 samples are infected, $m = 0, \ldots, n^2$. Assuming that the n^2 samples are randomly placed in the testing matrix,*

$$p_n(j, k \,|\, m) = \sum_{m_1=\lambda_1}^{\upsilon_1} \sum_{m_2=\lambda_2}^{\upsilon_2} \sum_{m_3=\lambda_3}^{\upsilon_3} \sum_{j_1=0}^{j} \sum_{k_1=0}^{k} \sum_{m_{21}=\lambda_{21}}^{\upsilon_{21}} \sum_{m_{31}=\lambda_{31}}^{\upsilon_{31}} p_{n-1}(j_1, k_1 \,|\, m_1) \frac{\binom{(n-1)^2}{m_1}\binom{j_1}{m_{21}}\binom{n-1-j_1}{m_2-m_{21}}\binom{k_1}{m_{31}}\binom{n-1-k_1}{m_3-m_{31}}}{\binom{n^2}{m}} \tag{5}$$

where $j = j_1 + m_2 - m_{21} + 1 - \delta_{0,m_3+m_4}$ and $k = k_1 + m_3 - m_{31} + 1 - \delta_{0,m_2+m_4}$, and where the lower limits of the summations, the lambdas, and the upper limits, the upsilons, are defined in the proof of the Lemma.

Proof Consider the m infected samples and how they are distributed within the matrix. First define m_1 to be the number amongst the m that fall within the top-left $(n-1) \times (n-1)$ submatrix; m_2 to be those amongst the m that fall in the rightmost column, except for the bottom cell of that column; m_3 the number that fall into the bottom row, except for the rightmost cell of that row; and m_4 to be the number that fall into the corner cell at the bottom of the last column, and at the right-end of the last row. We do not lose any samples, so $m = m_1 + m_2 + m_3 + m_4$. Decomposing the m in

this fashion,

$$p_n(j,k\,|\,m) = \sum_{m_1=\lambda_1}^{\upsilon_1}\sum_{m_2=\lambda_2}^{\upsilon_2}\sum_{m_3=\lambda_3}^{\upsilon_3} p_n(j,k|\,m=(m_1,m_2,m_3,m_4))\frac{\binom{(n-1)^2}{m_1}\binom{n-1}{m_2}\binom{n-1}{m_3}\binom{1}{m_4}}{\binom{n^2}{m}} \quad (6)$$

where

$$\lambda_1 = max(0, m-2n+1) \qquad \upsilon_1 = min(m,(n-1)^2)$$

$$\lambda_2 = max(0, m-m_1-n) \qquad \upsilon_2 = min(m-m_1, n-1)$$

$$\lambda_3 = max(0, m-m_1-m_2-1) \qquad \upsilon_3 = min(m-m_1-m_2, n-1).$$

We have extended the notation for $p_n(j,k\,|\,m)$ to show the decomposition of m into four components, where each component is known. Equation (6) follows from the use of the multivariate hypergeometric distribution.

Now consider how many rows and columns in the top-left $(n-1)\times(n-1)$ matrix contain an infected sample (denoted by j_1 and k_1, respectively):

$$p_n(j,k\,|\,m) = \sum_{m_1=\lambda_1}^{\upsilon_1}\sum_{m_2=\lambda_2}^{\upsilon_2}\sum_{m_3=\lambda_3}^{\upsilon_3}\sum_{j_1=0}^{j}\sum_{k_1=0}^{k} p_n(j,k|\,m=(m_1,m_2,m_3,m_4),(j_1,k_1))\times p_{n-1}(j_1,k_1|m_1)\frac{\binom{(n-1)^2}{m_1}\binom{n-1}{m_2}\binom{n-1}{m_3}\binom{1}{m_4}}{\binom{n^2}{m}}. \quad (7)$$

Focus now on the top-left $(n-1)\times(n-1)$ matrix, and suppose that the m_1 infecteds are distributed in such a manner that j_1 rows and k_1 columns are infected, for some non-negative integers, j_1 and k_1. Without loss of generality, suppose these are the first j_1 rows and the first k_1 columns.

Now consider the whole $n\times n$ matrix and determine what impact the addition of $m-m_1$ infected samples into the last column and row has on the number of infected rows and columns. Suppose that of the m_2 infected samples that fall into the last column (excepting the last cell), m_{21} fall into the first j_1 rows. These m_{21} would not impact on the number of rows infected when going from the $(n-1)\times(n-1)$ matrix to the $n\times n$ matrix. Whereas the m_2-m_{21} that fall into the last, but one, $n-j_1-1$ rows would *each* increase the number of rows infected by one. The last row could also

increase the number of infected rows by one, if $m_3 + m_4 > 0$.

Now, to measure the impact of the m_3 infecteds that fall into the last row, except for the last element. Suppose that m_{31} of these fall into the first k_1 columns, and the remaining $m_3 - m_{31}$ fall into the next $n - k_1 - 1$ columns. This will increase the number of infected columns by $m_3 - m_{31}$. The last column could also increase the number of infected columns by one, if $m_2 + m_4 > 0$.

Thus the total number of rows infected in the $n \times n$ matrix is $j = j_1 + m_2 - m_{21} + 1 - \delta_{0,m_3+m_4}$, and the total number of columns infected is $k = k_1 + m_3 - m_{31} + 1 - \delta_{0,m_2+m_4}$, with $\delta_{a,b}$ denoting Kronecker's delta that is equal to one if $a = b$ and zero otherwise.

Combining this reasoning with Equation (7) we have that,

$$p_n(j,k\,|\,m) = \sum_{m_1=\lambda_1}^{\upsilon_1}\sum_{m_2=\lambda_2}^{\upsilon_2}\sum_{m_3=\lambda_3}^{\upsilon_3}\sum_{j_1=0}^{j}\sum_{k_1=0}^{k}\sum_{m_{21}=\lambda_{21}}^{\upsilon_{21}}\sum_{m_{31}=\lambda_{31}}^{\upsilon_{31}} p_{n-1}(j_1,k_1\,|\,m_1)\times \tag{8}$$

$$p_n(j,k|\,m=(m_1,m_2,m_3,m_4),(j_1,k_1),m_2=(m_{21},m_2-m_{21}),m_3=(m_{31},m_3-m_{31}))\times$$

$$\frac{\binom{(n-1)^2}{m_1}\binom{n-1}{m_2}\binom{n-1}{m_3}\binom{1}{m_4}}{\binom{n^2}{m}} \times \frac{\binom{j_1}{m_{21}}\binom{n-1-j_1}{m_2-m_{21}}}{\binom{n-1}{m_2}}\,\frac{\binom{k_1}{m_{31}}\binom{n-1-k_1}{m_3-m_{31}}}{\binom{n-1}{m_3}}$$

where

$$\lambda_{21} = max(0, m_2 - n + 1 + j_1) \quad \upsilon_{21} = min(j_1, m_2)$$

$$\lambda_{31} = max(0, m_3 - n + 1 + k_1) \quad \upsilon_{31} = min(k_1, m_3).$$

But note that the probability

$$p_n(j,k|\,m=(m_1,m_2,m_3,m_4),(j_1,k_1),m_2=(m_{21},m_2-m_{21}),m_3=(m_{31},m_3-m_{31}))$$

is either one or zero. It is one if both $j_1 = j + m_2 - m_{21} - 1 + \delta_{0,m_3+m_4}$ and $k_1 = k + m_3 - m_{31} - 1 + \delta_{0,m_2+m_4}$, and zero otherwise. Canceling terms in equation (8) then proves the Lemma, and to start the recursion we have that $p_1(0,0) = 1 - p_1(1,1) = 1 - p$.

□

Lemma A.2. *Let $p_n^{(1)}(j,k|y)$ be the probability that j of the first $n-1$ rows and k of the first $n-1$ columns have at least one infected sample, given that the (n,n) cell is fixed and y of the remaining n^2-1 samples are infected. Then,*

$$p_n^{(1)}(j,k|y) = \sum_{y_1=\lambda_1}^{\upsilon_1}\sum_{y_2=\lambda_2}^{\upsilon_2}\sum_{j_1=0}^{j}\sum_{k_1=0}^{k} \sum_{y_{21}=\lambda_{21}}^{\upsilon_{21}}\sum_{y_{31}=\lambda_{31}}^{\upsilon_{31}} p_{n-1}(j_1,k_1|y_1)\frac{\binom{(n-1)^2}{y_1}\binom{j_1}{y_{21}}\binom{n-1-j_1}{y_2-y_{21}}\binom{k_1}{y_{31}}\binom{n-1-k_1}{y_3-y_{31}}}{\binom{n^2-1}{y}}, \tag{9}$$

where

$\lambda_1 = max(0, y-2(n-1))$ $\quad$ $\upsilon_1 = min(y,(n-1)^2)$

$\lambda_2 = max(0, y-y_1-(n-1))$ $\quad$ $\upsilon_2 = min(y-y_1, n-1)$

$\lambda_{21} = max(0, y_2-(n-1-j_1))$ $\quad$ $\upsilon_{21} = min(j_1, y_2)$

$\lambda_{31} = max(0, y_3-(n-1-k_1))$ $\quad$ $\upsilon_{31} = min(k_1, y_3)$

and $j = j_1 + y_2 - y_{21}$ *and* $k = k_1 + y_3 - y_{31}$.

Proof This proof follows similar logic to that outlined for Lemma A.1 where now of the y infected samples in the n^2-1 cells in the matrix, y_1 fall in the top-left $(n-1)\times(n-1)$ cells, y_2 fall into the n^{th} column (excluding the (n,n) cell), and y_3 fall into the n^{th} row (excluding the (n,n) cell), with $y = y_1 + y_2 + y_3$. Thus,

$$p_n^{(1)}(j,k|y) = \sum_{y_1=\lambda_1}^{\upsilon_1}\sum_{y_2=\lambda_2}^{\upsilon_2} p_n^{(1)}(j,k|y=(y_1,y_2,y_3))\frac{\binom{(n-1)^2}{y_1}\binom{n-1}{y_2}\binom{n-1}{y_3}}{\binom{n^2-1}{y}}, \tag{10}$$

with lambdas and upsilons as described above.

Now, again consider the number of rows, j_1, and columns, k_1, in the top-left $(n-1)\times(n-1)$ matrix that contain a diseased sample. It follows that,

$$p_n^{(1)}(j,k|y) = \sum_{y_1=\lambda_1}^{\upsilon_1}\sum_{y_2=\lambda_2}^{\upsilon_2}\sum_{j_1=0}^{j}\sum_{k_1=0}^{k} p_n^{(1)}(j,k|y=(y_1,y_2,y_3),(j_1,k_1)) \times p_{n-1}(j_1,k_1|y_1)\frac{\binom{(n-1)^2}{y_1}\binom{n-1}{y_2}\binom{n-1}{y_3}}{\binom{n^2-1}{y}}. \tag{11}$$

Note that $p_{n-1}(j_1,k_1|y_1)$ is the same term used in the expression of $p_n(j,k|m)$, since a fixed sample in the (n,n) cell, whether infected or uninfected, is independent of the samples in the top-left $(n-1)\times(n-1)$ matrix.

Finally, we explore the placement of the y_2 and y_3 diseased samples in the n^{th} column and row, excluding the last cell which is fixed. Using similar notation to the proof of Lemma A.1, let y_{21} be the number of diseased samples in the last column that overlap with the j_1 rows that have at least one diseased sample. These y_{21} samples do not contribute to the number of rows with diseased samples, but the remaining $y_2 - y_{21}$ do contribute, making $j = j_1 + (y_2 - y_{21})$. Similarly, let y_{31} be the number of diseased samples in the last row that overlap with the k_1 columns from the top-left $(n-1) \times (n-1)$ matrix that have diseased samples. It follows that $k = k_1 + (y_3 - y_{31})$. Therefore,

$$p_n^{(1)}(j,k|y) = \sum_{y_1=\lambda_1}^{\upsilon_1} \sum_{y_2=\lambda_2}^{\upsilon_2} \sum_{j_1=0}^{j} \sum_{k_1=0}^{k} \sum_{y_{21}=\lambda_{21}}^{\upsilon_{21}} \sum_{y_{31}=\lambda_{31}}^{\upsilon_{31}} p_{n-1}(j_1,k_1|y_1) \times \frac{\binom{(n-1)^2}{y_1}\binom{n-1}{y_2}\binom{n-1}{y_3}}{\binom{n^2-1}{y}} \times \frac{\binom{j_1}{y_{21}}\binom{n-1-j_1}{y_2-y_{21}}}{\binom{n-1}{y_2}} \times \frac{\binom{k_1}{y_{31}}\binom{n-1-k_1}{y_3-y_{31}}}{\binom{n-1}{y_3}}, \quad (12)$$

with lambdas and upsilon's are described above.

Equation 12 immediately reduces to $p_n^{(1)}(j,k|y)$ defined in the Lemma, and the recursion will start with the same term as in Lemma A.1 with $p_1(0,0) = 1 - p_1(1,1) = 1 - p$.

□

Lemma A.3. *The expression of the number of columns and rows that have a diseased sample, conditional on the (n,n) element being fixed and the prevalence of disease, p, is*

$$p_n^{(1)}(j,k|p) = \sum_{y=0}^{n^2-1} p_n^{(1)}(j,k|y) \binom{n^2-1}{y} p^y (1-p)^{n^2-1-y}. \quad (13)$$

Proof

Since it is assumed that the n^2 samples of the testing matrix, excluding the fixed sample in the (n,n) cell, are selected at random from the population with p prevalence of disease, the $P(Y = y)$ follows a binomial distribution. This, together with the formula for the total probability decomposition, proves the lemma.

□

Lemma A.4. *The probability that r rows and c columns of the $n \times n$ matrix excluding the last row and column test positive, conditional on the (n,n) cell as fixed and prevalence of disease, p, is*

expressed as

$$
\begin{aligned}
pos_n^{(1)}(r,c|p) &= \sum_{j=0}^{n-1}\sum_{k=0}^{n-1} p_n^{(1)}(j,k|p) \times \\
&\left\{ \sum_{i=\lambda_1}^{\upsilon_1} \binom{j}{i} S_e^i (1-S_e)^{j-i} \binom{n-1-j}{r-i} (1-S_p)^{r-i} S_p^{n-1-r+i-j} \times \right. \\
&\left. \sum_{l=\lambda_2}^{\upsilon_2} \binom{k}{l} S_e^l (1-S_e)^{k-l} \binom{n-1-k}{c-l} (1-S_p)^{c-l} S_p^{n-1-c+l-k} \right\},
\end{aligned} \tag{14}
$$

with $\lambda_1 = max(0, r-(n-1-j))$, $\upsilon_1 = min(j,r)$, $\lambda_2 = max(0, c-(n-1-k))$, *and* $\upsilon_2 = min(k,c)$.

Proof The proof follows from a complete enumeration of the number of rows and columns that have at least one infected sample, $j,k = 1,\ldots,n-1$. Note that the number of row pools that test positive is not dependent on k and the number of column pools that test positive is not dependent on j, so that

$$
\begin{aligned}
pos_n^{(1)}(r,c|p) &= \sum_{j=0}^{n-1}\sum_{k=0}^{n-1} p_n^{(1)}(j,k|p)\, pos_n^{(1)}(r,c|j,k\bigcap p) \\
&= \sum_{j=0}^{n-1}\sum_{k=0}^{n-1} p_n^{(1)}(j,k|p)\, pos_n^{(1)}(r|j\bigcap p)\, pos_n^{(1)}(c|k\bigcap p).
\end{aligned} \tag{15}
$$

Suppose r of the first $n-1$ rows test positive, if i of these occur in rows with at least one infected sample, then $r-i$ occur in rows with no infected samples. Because of the independence of the samples, and therefore the independence of the row pools, for a fixed j, the probability of this event follows the product of two binomial distributions, one the probability that i out of j infected row pools test positive and the other the probability that $r-i$ out of $n-1-j$ uninfected subpools test positive. Thus,

$$
pos_n^{(1)}(r|j\bigcap p) = \sum_{i=\lambda_1}^{\upsilon_1} \binom{j}{i} S_e^i (1-S_e)^{j-i} \binom{n-1-j}{r-i} (1-S_p)^{r-i} S_p^{n-1-j-(r-i)} \tag{16}
$$

with λ_1 and υ_1 defined above.

Using similar arguments for the number of column pools that test positive, and combining Equations (15) and (16) completes the proof.

□

Appendix B

Lemma B.1. *Let $S^+_{i,j}$ and $S^-_{i,j}$ be the events that the sample in the (i,j) cell of the testing matrix is infected or not infected, respectively. Note that since the samples are randomly placed, then for all $i,j \in [1,\ldots,n]$, $P(S^+_{i,j}) = p$ and $P(S^-_{i,j}) = 1-p$. Also, let r^+_i and r^-_i be the events that the i^{th} row pool tests positive or negative, respectively, and similarly, c^+_j and c^-_j be the events that the j^{th} column pool tests positive or negative, respectively.*

Now, fix the (n,n) sample to be uninfected and let r^+_n and c^+_n be the events that the n^{th} row and column, respectively, test positive given the sample in the last cell is uninfected. Let r^-_n and c^-_n be the events that the n^{th} row and column, respectively, test negative given the sample in the last cell is uninfected. Let p^*_A denote the prevalence of infection in samples at the intersection of the last row and column pools that test positive, given that the n^{th} row pool tests positive and the (n,n) cell is uninfected. Note, this is also the prevalence of infection in samples at the intersection of the last column and row pools that test positive, given that the n^{th} column pool tests positive and the (n,n) cell is uninfected. Let p^*_B denote the prevalence of infection in samples at the intersection of the last row and column pools that test negative, given that the n^{th} row pool tests positive and the (n,n) cell is uninfected. Also note, this is the prevalence of infection in samples at the intersection of the last column and row pools that test negative, given that the n^{th} column pool tests positive and the (n,n) cell is uninfected. Therefore, we have

$$\begin{aligned} p^*_A &= \frac{P(r^+_n, c^+_j | S^+_{n,j}, S^-_{n,n})\, P(S^+_{n,j})}{P(r^+_n, c^+_j | S^+_{n,j}, S^-_{n,n})\, P(S^+_{n,j}) + P(r^+_n, c^+_j | S^-_{n,j}, S^-_{n,n})\, P(S^-_{n,j})} \qquad (17)\\ &= \frac{P(r^+_i, c^+_n | S^+_{i,n}, S^-_{n,n})\, P(S^+_{i,n})}{P(r^+_i, c^+_n | S^+_{i,n}, S^-_{n,n})\, P(S^+_{i,n}) + P(r^+_i, c^+_n | S^-_{i,n}, S^-_{n,n})\, P(S^-_{i,n})}, \end{aligned}$$

$$\begin{aligned} p^*_B &= \frac{P(r^+_n, c^-_j | S^+_{n,j}, S^-_{n,n})\, P(S^+_{n,j})}{P(r^+_n, c^-_j | S^+_{n,j}, S^-_{n,n})\, P(S^+_{n,j}) + P(r^+_n, c^-_j | S^-_{n,j}, S^-_{n,n})\, P(S^-_{n,j})}\\ &= \frac{P(r^+_i, c^-_n | S^+_{i,n}, S^-_{n,n})\, P(S^+_{i,n})}{P(r^+_i, c^-_n | S^+_{i,n}, S^-_{n,n})\, P(S^+_{i,n}) + P(r^+_i, c^-_n | S^-_{i,n}, S^-_{n,n})\, P(S^-_{i,n})}, \end{aligned}$$

with $P(r_i^+, c_j^+|S_{i,j}^+, S_{n,n}^-)$, $P(r_i^+, c_j^+|S_{i,j}^-, S_{n,n}^-)$, $P(r_i^+, c_j^-|S_{i,j}^+, S_{n,n}^-)$, $P(r_i^+, c_j^-|S_{i,j}^-, S_{n,n}^-)$, $P(r_i^-, c_j^-|S_{i,j}^+, S_{n,n}^-)$, *and* $P(r_i^-, c_j^-|S_{i,j}^-, S_{n,n}^-)$ *defined in the proof below.*

Proof First note, because of the independence of samples,

$$P(r_n^*, c_j^*|S_{n,j}^*, S_{n,n}^-) = P(r_n^*|S_{n,j}^*, S_{n,n}^-)\, P(c_j^*|S_{n,j}^*, S_{n,n}^-)$$

and

$$P(r_i^*, c_n^*|S_{i,n}^*, S_{n,n}^-) = P(r_i^*|S_{i,n}^*, S_{n,n}^-)\, P(c_n^*|S_{i,n}^*, S_{n,n}^-).$$

And also note that due to independence, $P(r_i^*|S_{i,n}^*, S_{n,n}^-) = P(r_i^*|S_{i,n}^*)$ and $P(c_j^*|S_{n,j}^*, S_{n,n}^-) = P(c_j^*|S_{n,j}^*)$. Now, if sample (n, j) is infected, the n^{th} pooled row sample will also be infected and test positive with probability S_e, and likewise, if sample (i, n) is infected, then the n^{th} pooled column sample will test positive with probability S_e. Hence,

$$P(r_n^+|S_{n,j}^+, S_{n,n}^-) = P(c_n^+|S_{i,n}^+, S_{n,n}^-) = S_e.$$

However, if the (n, j) sample is uninfected, then suppose all the other samples in the n^{th} row pool are uninfected, excepting the (n, n) cell which is fixed to be uninfected, then the n^{th} row pool positive with probability $1 - S_p$. This event occurs with probability $(1-p)^{n-2}$. But, if any of the other samples in the n^{th} row pool are infected, which occurs with probability $1-(1-p)^{n-2}$, then the n^{th} row pool tests positive with probability S_e. The same reasoning applies to the n^{th} column pool so that

$$P(r_n^+|S_{i,j}^-, S_{n,n}^-) = P(c_n^+|S_{i,j}^-, S_{n,n}^-) = (1-S_p)(1-p)^{n-2} + S_e\{1-(1-p)^{n-2}\}.$$

Finally, the results for Lemma B.1 follow directly from conditional probabilities, so that

$$\begin{aligned} p_A^* &= P(S_{n,j}^+|r_n^+, c_j^+, S_{n,n}^-) = \frac{P(S_{n,j}^+ \bigcap r_n^+, c_j^+|S_{n,n}^-)}{P(r_i^+, c_j^+|S_{n,n}^-)} \qquad (18) \\ &= \frac{P(r_n^+, c_j^+|S_{n,j}^+, S_{n,n}^-)\, P(S_{n,j}^+)}{P(r_n^+, c_j^+|S_{n,j}^+, S_{n,n}^-)\, P(S_{n,j}^+) + P(r_n^+, c_j^+|S_{n,j}^-, S_{n,n}^-)\, P(S_{n,j}^-)}, \\ &= P(S_{i,n}^+|r_i^+, c_n^+, S_{n,n}^-) \end{aligned}$$

$$
\begin{aligned}
p_B^* &= P(S_{n,j}^+ | r_n^+, c_j^-, S_{n,n}^-) = \frac{P(S_{n,j}^+ \bigcap r_n^+, c_j^- | S_{n,n}^-)}{P(r_n^+, c_j^-, S_{n,n}^-)} \\
&= \frac{P(r_n^+, c_j^- | S_{n,j}^+, S_{n,n}^-)\, P(S_{n,j}^+)}{P(r_n^+, c_j^- | S_{n,j}^+, S_{n,n}^-)\, P(S_{n,j}^+) + P(r_n^+, c_j^- | S_{n,j}^-, S_{n,n}^-)\, P(S_{n,j}^-)} \\
&= P(S_{i,n}^+ | r_i^-, c_n^+, S_{n,n}^-).
\end{aligned}
$$

□

Appendix C

In order to derive $pos_n(r, c|p)$, the probability that r of the n rows and c of the n columns test positive, given prevalence of disease p, we will show:

1. $p_n(j, k|p)$, the probability that j of the n rows and k of the n columns contain at least one infected sample, given prevalence of disease p in the $n \times n$ testing matrix,
2. and then, $pos_n(r, c|p)$.

Lemma C.1. *Assume that the samples are placed at random in an $n \times n$ matrix and they represent a random sample from a population where p is the prevalence of disease. Then for $j, k = 0, \ldots, n$ for any positive integer $n \geq 1$,*

$$
p_n(j, k|p) = \sum_{m=0}^{n^2} \binom{n^2}{m} p^m (1-p)^{n^2-m}\, p_n(j, k \,|\, m), \tag{19}
$$

with $p_n(j, k \,|\, m)$ as shown in Equation (5).

Proof This proof follows immediately from the proof of Lemma A.3, where we now extend to all n rows and n columns.

□

Lemma C.2. *Let $pos_n(r, c|p)$ denote the probability of r rows and c columns testing positive in an $n \times n$ matrix of randomly chosen samples. Then,*

$$\begin{aligned} pos_n(r,c|p) &= \sum_{j=0}^{n}\sum_{k=0}^{n} p_n(j,k|p) \times \\ &\quad \left\{ \sum_{i=\lambda_1}^{\upsilon_1} \binom{j}{i} S_e^i (1-S_e)^{j-i} \binom{n-j}{r-i} (1-S_p)^{r-i} S_p^{n-r+i-j} \times \right. \\ &\quad \left. \sum_{l=\lambda_2}^{\upsilon_2} \binom{k}{l} S_e^l (1-S_e)^{k-l} \binom{n-k}{c-l} (1-S_p)^{c-l} S_p^{n-c+l-k} \right\}, \end{aligned} \tag{20}$$

with $\lambda_1 = max(0, r-n+j)$, $\upsilon_1 = min(j, r)$, $\lambda_2 = max(0, c-n+k)$, and $\upsilon_2 = min(k, c)$.

Proof This proof follows directly from the logic in the proof of Lemma A.4, where we now extend to all n rows and n columns.

□

Appendix D

Lemma D.1. *Let A denote the collection of samples from the original testing matrix that are at the intersection of both a row and a column that test positive, B denote the collection of samples from the original testing matrix that are at the intersection of discordant row and column tests, and C denote the collection of samples from the original testing matrix that are at the intersection of both a row and a column that test negative. Using the definitions provided in Lemma B.1, if the overall prevalence of disease is p, then the prevalences in sections A, B, and C, denoted p_A, p_B and p_C respectively, are as follows:*

$$p_A = \frac{P(r_i^+, c_j^+|S_{i,j}^+)\, P(S_{i,j}^+)}{P(r_i^+, c_j^+|S_{i,j}^+)\, P(S_{i,j}^+) + P(r_i^+, c_j^+|S_{i,j}^-)\, P(S_{i,j}^-)}, \tag{21}$$

$$\begin{aligned} p_B &= \frac{P(r_i^+, c_j^-|S_{i,j}^+)\, P(S_{i,j}^+)}{P(r_i^+, c_j^-|S_{i,j}^+)\, P(S_{i,j}^+) + P(r_i^+, c_j^-|S_{i,j}^-)\, P(S_{i,j}^-)} \\ &= P(S_{i,j}^+|r_i^-, c_j^+), \end{aligned}$$

$$p_C = \frac{P(r_i^-, c_j^- | S_{i,j}^+)\, P(S_{i,j}^+)}{P(r_i^-, c_j^- | S_{i,j}^+)\, P(S_{i,j}^+) + P(r_i^-, c_j^- | S_{i,j}^-)\, P(S_{i,j}^-)},$$

with $P(r_i^+, c_j^+ | S_{i,j}^+)$, $P(r_i^+, c_j^+ | S_{i,j}^-)$, $P(r_i^+, c_j^- | S_{i,j}^+)$, $P(r_i^+, c_j^- | S_{i,j}^-)$,$P(r_i^-, c_j^- | S_{i,j}^+)$, *and* $P(r_i^-, c_j^- | S_{i,j}^-)$ *defined in the proof below.*

Proof First note, because of the independence of samples, $P(r_i^*, c_j^* | S_{i,j}^*) = P(r_i^* | S_{i,j}^*) P(c_j^* | S_{i,j}^*)$. Now, if sample (i, j) is infected, the i^{th} pooled row sample will test positive with probability S_e and negative with probability $1 - S_e$ and likewise the j^{th} pooled column sample will test positive and negative with the same probabilities, so that

$$\begin{aligned} P(r_i^+ | S_{i,j}^+) &= P(c_j^+ | S_{i,j}^+) = Se \\ P(r_i^- | S_{i,j}^+) &= P(c_j^- | S_{i,j}^+) = 1 - S_e. \end{aligned} \tag{22}$$

If the (i, j) sample is uninfected, then suppose all the other samples in the i^{th} row pool are uninfected which occurs with probability $(1-p)^{n-1}$, then the pool will test negative with probability S_p and positive with probability $1 - S_p$. However, if any of the other samples in the i^{th} row pool are infected, which occurs with probability $1 - (1-p)^{n-1}$, then the pool tests negative with probability $1 - S_e$ and positive with probability S_e. The same reasoning applies to the j^{th} column pool so that

$$\begin{aligned} P(r_i^+ | S_{i,j}^-) &= P(c_j^+ | S_{i,j}^-) = (1 - S_p)(1-p)^{n-1} + S_e\{1 - (1-p)^{n-1}\} \\ P(r_i^- | S_{i,j}^-) &= P(c_j^- | S_{i,j}^-) = S_p(1-p)^{n-1} + (1 - S_e)\{1 - (1-p)^{n-1}\}. \end{aligned} \tag{23}$$

Finally, the results for Lemma D.1 follow directly from conditional probabilities, so that

$$p_A = P(S_{i,j}^+ | r_i^+, c_j^+) = \frac{P(S_{i,j}^+ \bigcap r_i^+, c_j^+)}{P(r_i^+, c_j^+)} = \frac{P(r_i^+, c_j^+ | S_{i,j}^+)\, P(S_{i,j}^+)}{P(r_i^+, c_j^+ | S_{i,j}^+)\, P(S_{i,j}^+) + P(r_i^+, c_j^+ | S_{i,j}^-)\, P(S_{i,j}^-)},$$

$$\begin{aligned} p_B &= P(S_{i,j}^+ | r_i^+, c_j^-) = \frac{P(S_{i,j}^+ \bigcap r_i^+, c_j^-)}{P(r_i^+, c_j^-)} = \frac{P(r_i^+, c_j^- | S_{i,j}^+)\, P(S_{i,j}^+)}{P(r_i^+, c_j^- | S_{i,j}^+)\, P(S_{i,j}^+) + P(r_i^+, c_j^- | S_{i,j}^-)\, P(S_{i,j}^-)} \\ &= P(S_{i,j}^+ | r_i^-, c_j^+), \end{aligned}$$

$$p_C = P(S_{i,j}^+ | r_i^-, c_j^-) = \frac{P(S_{i,j}^+ \bigcap r_i^-, c_j^-)}{P(r_i^-, c_j^-)} = \frac{P(r_i^-, c_j^- | S_{i,j}^+)\, P(S_{i,j}^+)}{P(r_i^-, c_j^- | S_{i,j}^+)\, P(S_{i,j}^+) + P(r_i^-, c_j^- | S_{i,j}^-)\, P(S_{i,j}^-)}.$$

□

Acknowledgements

This research is supported in part by the National Institutes of Health grants T32 AI007358 and R01 EB006195. The authors thank Victor DeGruttola, George Seage, and Laura Forsberg-White for their insightful comments.

References

[1] R. Dorfman. The detection of defective members of large populations. *Ann Math Stat*, 14:436–440, 1943.

[2] V.A. Lancaster and S. Keller-McNulty. A review of composite sampling methods. *J Am Stat Assoc*, 93:1216–1230, 1998.

[3] Robert C Venette, Roger D Moon, and William D Hutchison. Strategies and statistics of sampling for rare individuals. *Annu Rev Entomol*, 47:143–174, 2002.

[4] S. A. Zenios and L. M. Wein. Pooled testing for HIV prevalence estimation: exploiting the dilution effect. *Stat Med*, 17(13):1447–1467, Jul 1998.

[5] E. Litvak, X. Tu, and M. Pagano. Screening for the presence of a disease by pooling sera samples. *J Am Stat Assoc*, 89:424–434, 1994.

[6] E. Litvak, X. Tu, and M. Pagano. *Modeling the AIDS epidemic: planning, policy and prediction.*, chapter Screening for the presence of HIV by pooling sera samples: simplified procedures. Raven Press, 1994.

[7] R. M. Phatarfod and A. Sudbury. The use of a square array scheme in blood testing. *Stat Med*, 13(22):2337–2343, Nov 1994.

[8] M. Xie, K. Tatsouka, J. Sacks, and S.S. Young. Group testing with blockers and synergism. *J Am Stat Assoc*, 96:92–102, 2001.

[9] N.L. Kennedy. Multistage group testing procedure (group screening). *Commun Stat*, 33(3):621–637, 2004.

[10] B.L. Hedt and M. Pagano. Matrix pooling: An accurate and cost effective testing algorithm for detection of acute HIV infection.