University of North Carolina at Chapel Hill
The University of North Carolina at Chapel Hill Department of Biostatistics Technical Report Series

Year 2010 *Paper* 16

Group Testing for Case Identification with Correlated Responses

Samuel D. Lendle* Michael Hudgens†
Bahjat F. Qaqish‡

*University of North Carolina at Chapel Hill
†University of North Carolina at Chapel Hill, mhudgens@bios.unc.edu
‡University of North Carolina, Chapel Hill, qaqish@bios.unc.edu

Group Testing for Case Identification with Correlated Responses

Samuel D. Lendle, Michael Hudgens, and Bahjat F. Qaqish

Abstract

This paper examines group testing procedures where units within a group (or pool) may be correlated. The expected number of tests per unit (i.e., efficiency) of hierarchical and matrix based procedures is derived based on a class of models of exchangeable binary random variables. The effect of the arrangement of correlated units within pools on efficiency is then examined. In general, when correlated units are arranged in the same pool, the expected number of tests per unit decreases, sometimes substantially, relative to arrangements which ignore information about correlation.

Group testing for case identification with correlated responses

Samuel D. Lendle, Michael G. Hudgens,* and Bahjat F. Qaqish

Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, 27599, U.S.A.

*email: mhudgens@bios.unc.edu

October 9, 2010

Abstract

This paper examines group testing procedures where units within a group (or pool) may be correlated. The expected number of tests per unit (i.e., efficiency) of hierarchical and matrix based procedures is derived based on a class of models of exchangeable binary random variables. The effect of the arrangement of correlated units within pools on efficiency is then examined. In general, when correlated units are arranged in the same pool, the expected number of tests per unit decreases, sometimes substantially, relative to arrangements which ignore information about correlation.

Keywords: Composite sampling; Epitope mapping; Exchangeable binary random variables; Group testing; HIV; Matrix testing; Pooled testing

1 Introduction

Group testing is a method used to reduce the average number of tests needed to identify cases of a disease in a population. The first use of group testing was proposed by Dorfman (1943). Dorfman proposed pooling blood samples of groups of men inducted into the military, and testing the combined samples for antigens to identify the presence of syphilis. If the combined samples tested negative for the antigens, the men were declared syphilis free with only one test. Otherwise, samples from each man were tested individually. Specimen pooling or group testing has been applied to screening for various infectious diseases and has also found broader application in genetics, the pharmaceutical and blood bank industries, entomology, and many other areas (Lancaster and Keller-McNulty, 1998; Kim et al., 2007). Group testing can also be used to reduce the average number of tests needed to estimate the prevalence of a disease (Hughes-Oliver and Rosenberger, 2000; Tebbs and Swallow, 2003), but this paper focuses on case identification.

Dorfman's two stage procedure has been generalized to three or more stages. If the initial (or "master") pool tests positive, the specimens may be pooled into smaller non-overlapping subpools. If a subpool tests positive, individuals can be tested, or subpools can be divided further into smaller non-overlapping subpools nested within the previous subpool. This is known as a hierarchical procedure (Finucan, 1964; Johnson et al., 1991). Another common

group testing algorithm entails an array based procedure (Phatarfod and Sudbury, 1994; Berger et al., 2000; Kim and Hudgens, 2009). In the simplest scenario, a group of n^2 units is arranged into an $n \times n$ matrix, and pools of size n are constructed from units in each row or column. The $2n$ row and column pools are then tested, and positive units are identified by testing the units at the intersections of positive row and column pools.

Prior research regarding group testing procedures typically assumes that all individual units are independent. This assumption may not be reasonable in certain situations. For example, in the infectious disease setting, responses to a screening test may be positively correlated for individuals from the same geographical area or the same household. A second example arises in human immunodeficiency virus (HIV) vaccine development, where group testing methods are used to detect T-cell responses to specific epitopes induced by a candidate vaccine (Malhotra et al., 2007a,b; Yan et al., 2007). T-cell responses to one or more peptides are identified by using ELISpot, intracellular cytokine staining, or other assays. Li et al. (2006) developed a potential T-cell epitope peptide set designed to contain epitopes found in commonly circulating strains of HIV. The peptide set is made of 15-mer peptides, some of which overlap by 10 or more amino acids. It is reasonable to expect that T-cell responses from the same individual are likely to be correlated for overlapping peptides. Indeed, Malhotra et al. (2007a) observed that T-cells of HIV infected individuals can recognize multiple peptides containing variants of the same epitope. Roederer and Koup (2003) evaluated possible group testing procedures to be used in this setting using Monte Carlo simulation, but did not consider that T-cell responses may be correlated. Below we show that accounting for this correlation when using group testing for case identification can reduce the average number of tests needed to identify all peptides that elicit a T-cell response.

2 Preliminaries

Suppose that a unit has either a positive or negative response to some classification. For example, the unit could represent an individual classified by a disease screening test, or a peptide classified according to a particular assay. Assume the probability of an incorrect classification is zero. Similarly suppose if the units are pooled together, the pool will be classified as positive if and only if at least one unit in the pool is positive. The efficiency of a group testing procedure is defined as the expected number of tests per unit required to classify all units as either positive or negative. In order to evaluate the efficiency, the probabilities that pools of units do not have any positive responses need to be calculated. These calculations require knowledge about correlation among units within each pool. Suppose there are n units total which can be partitioned into l clusters of size m and the following assumption holds:

Assumption 1 *Units in different clusters are independent, and the joint distribution of units in the same cluster is the same for all clusters.*

Without loss of generality, let $\tilde{X} = (X_1, \ldots, X_m)$ be a vector of binary random variables representing the responses for a particular cluster, where $X_i = 1$ if the ith unit in that cluster is positive, and $X_i = 0$ otherwise for $i = 1, \ldots, m$. Let $\dot{X} = \sum_{i=1}^{m} X_i$, let $\tilde{x}$ be a possible

realization of $\tilde{X}$, and let $\dot{x}$ be the sum of the values of $\tilde{x}$. Let $\tilde{X}' = (X'_1, \ldots, X'_{m'})$ be a subset of any m' elements from $\tilde{X}$ where $m' \in \{1, \ldots, m\}$ and $\dot{X}' = \sum_{i=1}^{m'} X'_i$. Deriving the efficiency of a group testing procedure requires assumptions about the distribution of $\tilde{X}$. A class of models for $\tilde{X}$ is defined by Assumptions 2 and 3 below by factoring the probability mass function for $\tilde{X}$ as $\text{pr}(\tilde{X} = \tilde{x}) = \text{pr}(\dot{X} = \dot{x})\text{pr}(\tilde{X} = \tilde{x} \mid \dot{X} = \dot{x})$.

Assumption 2 *Units within a cluster are exchangeable in the sense that*

$$\text{pr}(\tilde{X} = \tilde{x} \mid \dot{X} = \dot{x}) = \binom{m}{\dot{x}}^{-1}.$$

Assumption 3 *The distribution of $\dot{X}$ is a mixture of binomial distributions such that*

$$\text{pr}(\dot{X} = \dot{x}) = E_\pi \left\{ \binom{m}{\dot{x}} \pi^{\dot{x}} (1-\pi)^{m-\dot{x}} \right\} \tag{1}$$

where $E_\pi\{g(\dot{x}, \pi)\} = \int_0^1 g(\dot{x}, \pi) dF(\pi)$ for any function g, and π is a random variable with support $[0, 1]$ and cumulative distribution function F.

Note there are connections between Assumption 3 and de Finetti's Theorem. In particular, if the cluster $\tilde{X}$ can be viewed as a subset of an infinite sequence of exchangeable binary random variables, then Assumption 3 and Lemma 2, below, follow immediately from de Finetti's Theorem (de Finetti, 1975). However, in settings motivating this work, such as epitope mapping studies, the focus is on clusters of finite size, in which case (1) does not hold in general (Diaconis, 1977).

Let $E(X_i) = p$ be the probability that any unit i tests positive and let $\text{cor}(X_i, X_j) = \sigma$ be the pairwise correlation between any two units i and j for $i \neq j$. The lemmas below establish certain properties about the class of models under Assumptions 2 and 3 that are needed for evaluating the efficiencies of group testing procedures in Sections 3 and 4. Lemma 1 shows that any distribution of exchangeable binary random variables approaches a known limiting distribution as σ approaches one. Lemma 2 shows that the distribution of a subset of units from a cluster with the properties defined in Assumptions 2 and 3 is of the same form as the distribution of the units in the cluster. By specifying a distribution for π where the first and second moments are p and $\sigma p(1-p) + p^2$, respectively, the distribution of a vector of exchangeable binary random variables with specified marginal means and pairwise correlations is defined by Lemma 3. Proofs of the lemmas are given in the appendix.

Lemma 1 *Under the conditions in Assumption 2, as σ approaches 1 the distribution of $\dot{X}$ converges to a two-point distribution where $\text{pr}(\dot{X} = 0) \to 1 - p$ and $\text{pr}(\dot{X} = m) \to p$.*

Lemma 2 *Under the conditions in Assumptions 2 and 3, the distribution of $\dot{X}'$ is a mixture of binomial distributions of the same form as $\dot{X}$, such that*

$$\text{pr}(\dot{X}' = \dot{x}') = E_\pi \left\{ \binom{m'}{\dot{x}'} \pi^{\dot{x}'} (1-\pi)^{m'-\dot{x}'} \right\} \tag{2}$$

for $\dot{x}' = 1, \ldots, m'$.

Lemma 3 *Under the conditions in Assumptions 2 and 3, if* $E(\pi) = p$ *and* $E(\pi^2) = \sigma p(1-p) + p^2$*, then* $E(X_i) = p$ *for all* i *and* $\text{cor}(X_i, X_j) = \sigma$ *for all* $i \neq j$*.*

In the sequel, three models are considered to examine how the efficiencies of group testing procedures are affected by correlated responses. The first is a beta-binomial model (Skellam, 1948), where π has a beta distribution with mean p and variance $\sigma p(1-p)$. Madsen (1993) described multiple distributions that can be used to model exchangeable binary data. One of those models can be constructed by letting $\pi = p$ with probability $1-\sigma$, $\pi = 0$ with probability $\sigma(1-p)$, and $\pi = 1$ with probability σp; this will be referred to as a Madsen model. A third model, described in Morel and Neerchal (1997) can be constructed by letting $\pi = p(1-\sqrt{\sigma}) + \sqrt{\sigma}$ with probability p and $p(1-\sqrt{\sigma})$ with probability $1-p$.

The efficiency derivations in Sections 3 and 4 below rely on the following additional notation. Let $q_0 = 1$ and $q_{m'} = \text{pr}(\dot{X}' = 0)$ denote the probability m' units from the same cluster are negative for $m' \in \{1, \ldots, m\}$. For the three models above $q_1 = 1-p$ and $q_{m'}$ is given by (2) for $\dot{x}' = 0$. Finally let T denote the number of tests required by a particular group testing procedure to classify n units as positive or negative.

3 Hierarchical procedures

3.1 Notation and general hierarchical procedures

Consider a hierarchical procedure where $n_1 = n$ units are combined to form a master pool. In the first stage, the master pool is tested, and if it is positive, w_2 non-overlapping pools of n_2 units are each tested in the second stage. In a two stage procedure, $n_2 = 1$ and each unit in a positive master pool is tested individually. In a general h stage procedure, for each pool that tests positive in stage $s-1$, n_{s-1}/n_s non-overlapping pools of n_s units are tested. There are a total of $w_s = n_1/n_s$ pools that could be tested at stage s if all of the pools in stage $s-1$ test positive. At the hth stage each pool is made up of individual units, so $n_h = 1$. The total number of tests $T = T_1 + \ldots + T_h$, where T_s is a random variable representing the number of tests at stage s. The efficiency of a hierarchical procedure is $E(T)/n_1 = \sum_{s=1}^{h} E(T_s)/n_1$. The master pool is always tested, so $E(T_1)$ is always one.

In this section, let Y_{sij} be 1 if the jth unit in the ith pool of the sth stage is positive, and 0 otherwise. Let $V_{si} = \max(Y_{si1}, \ldots, Y_{sin_s})$. If the ith pool in the sth stage is tested, then V_{si} is the observed response. For a particular arrangement of clusters, let m_{sik} be the number of units from cluster k in pool i of stage s where $k = 1, \ldots, l$. For $s > 1$, $E(T_s) = w_s/w_{s-1} \times \sum_{i=1}^{w_{s-1}} \text{pr}(V_{(s-1)i} = 1)$, where

$$\text{pr}(V_{(s-1)i} = 1) = 1 - \prod_{k=1}^{l} q_{m_{(s-1)ik}} \tag{3}$$

is the probability that pool i in stage $s-1$ tests positive. To determine the efficiency of a particular hierarchical procedure, (3) is evaluated based on how the clusters are arranged.

3.2 Nested hierarchical arrangement

Suppose clusters of size m are arranged such that for some $h' \in \{2, 3, \ldots, h\}$ all units from the same cluster are in the same pool for stages $1, 2, \ldots, h'-1$, and units in the same pool are from the same cluster at each stage for stages $h', \ldots, h$. That is, $m_{sik} = m$ or 0 for $s < h'$ and $m_{sik} = n_s$ or 0 for $s \geq h'$. Call this a nested hierarchical arrangement. By (3), if $1 < s \leq h'$ then $\text{pr}(V_{(s-1)i} = 1) = 1 - q_m^{n_{s-1}/m}$ for all i, and if $h' < s \leq h$ then $\text{pr}(V_{(s-1)i} = 1) = 1 - q_{n_{s-1}}$ for all i. Therefore

$$E(T_s) = \begin{cases} w_s(1 - q_m^{n_{s-1}/m}) & (1 < s \leq h'), \\ w_s(1 - q_{n_{s-1}}) & (h' < s \leq h). \end{cases} \tag{4}$$

If $\sigma = 0$ then $q_m^{n_{s-1}/m} = q_{n_{s-1}} = q_1^{n_{s-1}}$ so $E(T_s) = w_s(1 - q_1^{n_{s-1}})$. Johnson et al. (1991) derive the efficiency for a hierarchical procedure when incorrect classifications are possible. If $\sigma = 0$, $E(T)/n_1 = \sum_{s=1}^{h} E(T_s)/n_1$ is equivalent to their equation (6.19) when the probability of an incorrect classification equals zero.

3.3 Random hierarchical arrangement

Consider the case where units are arranged in a way that is unrelated to their cluster membership. Let $\tilde{M}_{si\cdot}$ be the random vector of length l of the number of units from each cluster $1, \ldots, l$ in pool i in stage s. Let each possible arrangement of the n_1 units have the same probability, so $\tilde{M}_{si\cdot}$ has a multivariate hypergeometric distribution such that

$$\text{pr}(\tilde{M}_{si\cdot} = \tilde{m}_{si\cdot t}) = \binom{n_1}{n_s}^{-1} \prod_{k=1}^{l} \binom{m}{m_{sikt}} \tag{5}$$

where $\tilde{m}_{si\cdot t} = (m_{si1t}, \ldots, m_{silt})$ is the tth possible value of $\tilde{M}_{si\cdot}$ for $t = 1, \ldots, \binom{n_1}{n_s}$. Then

$$\text{pr}(V_{(s-1)i} = 0 \mid \tilde{M}_{(s-1)i\cdot} = \tilde{m}_{(s-1)i\cdot t}) = \prod_{k=1}^{l} q_{m_{(s-1)ikt}}, \tag{6}$$

and therefore

$$\text{pr}(V_{(s-1)i} = 1) = 1 - \binom{n_1}{n_{s-1}}^{-1} \sum_{t=1}^{\binom{n_1}{n_{s-1}}} \left\{ \prod_{k=1}^{l} q_{m_{(s-1)ikt}} \binom{m}{m_{(s-1)ikt}} \right\}. \tag{7}$$

When n_1 is large, the number of possible arrangements $\binom{n_1}{n_s}$ becomes very large, and the exact calculation for (7) is computationally difficult. Monte Carlo simulation can be used to approximate (7). First values of $\tilde{M}_{(s-1)i\cdot}$ are repeatedly sampled from a multivariate hypergeometric distribution according to (5). Then the conditional probability (6) is evaluated for each sample. Finally, one minus the sample mean of the conditional probabilities will approximate (7).

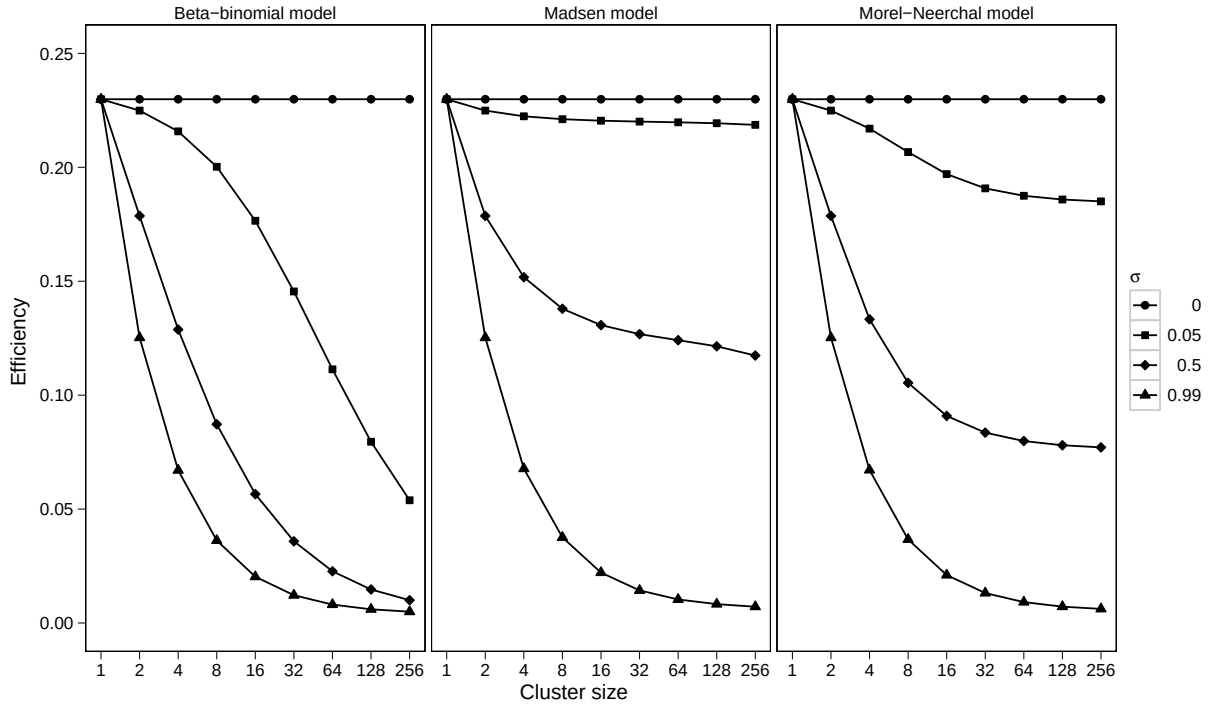

Figure 1: Efficiencies for a two stage hierarchical procedure where $n_1 = 256$ and $p = 0.001$ by cluster size m, pairwise correlation σ, and model

3.4 Comparison of hierarchical arrangements

For a two stage hierarchical procedure with a nested arrangement, the expected number of tests is $E(T) = 1 + E(T_2)$ where $E(T_2) = w_2(1 - q_m^{n_1/m})$ by (4). If $\sigma = 0$ then $q_m = q_1^m$ and the efficiency for all models equals $E(T)/n_1 = n_1^{-1} + 1 - q_1^{n_1}$ as in Dorfman (1943). Figure 1 illustrates the efficiency of a two stage hierarchical procedure for the three models and different values of σ as a function of m. For large clusters the expected tests per unit is reduced substantially as σ increases. When $\sigma = 0.99$, the efficiencies for all three models are almost identical, which is consistent with Lemma 1.

For a three stage hierarchical procedure with a nested arrangement where all units from the same cluster fit into the same pool in stages 1 and 2 (i.e., $h' = 3$), the expected number of tests for stage 2 has the same form as in the two stage procedure above. Similarly, the expected number of tests for the third stage is $E(T_3) = w_3(1 - q_m^{n_2/m})$ so $E(T) = 1 + w_2(1 - q_m^{n_1/m}) + w_3(1 - q_m^{n_2/m})$. Figure 2 compares the efficiencies of three stage hierarchical nested and random arrangements by stage two pool size, n_2, as a function of σ. Efficiencies for the random arrangements were obtained by Monte Carlo simulation. In all cases in Fig. 2 the nested arrangements have better efficiency than random arrangements for $\sigma \in (0, 1)$. As σ approaches 1, the efficiencies for the three models converge for each arrangement as indicated by Lemma 1.

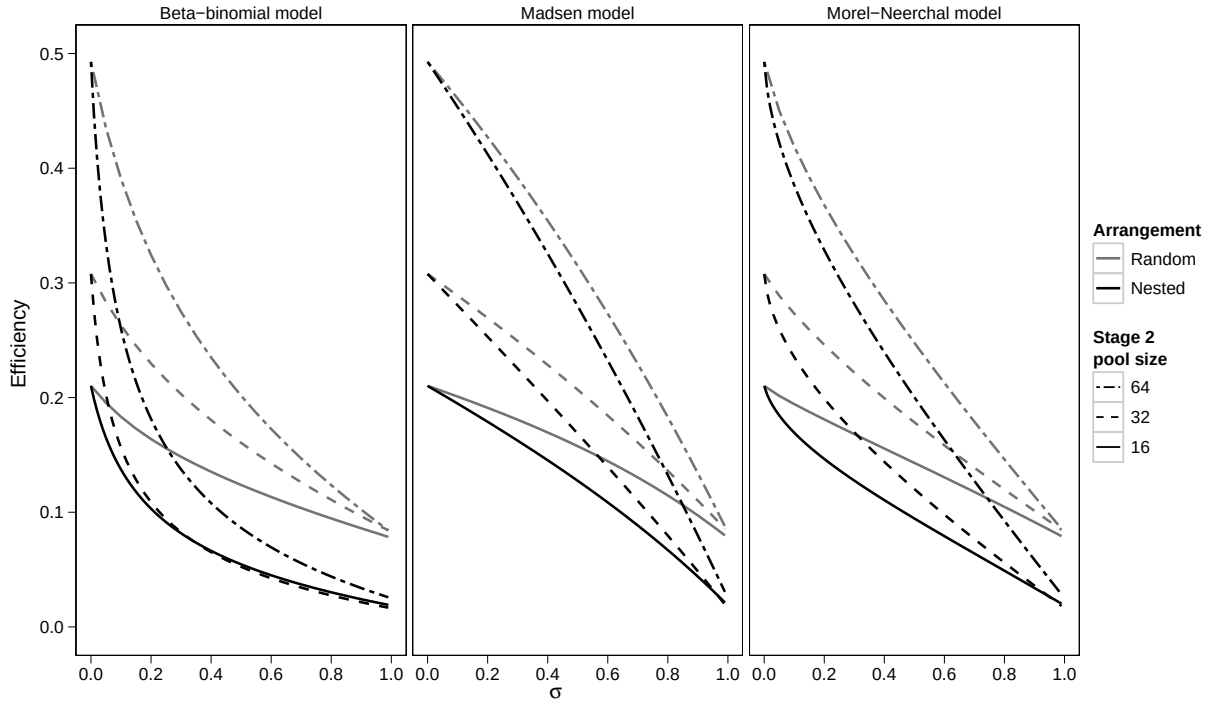

Figure 2: Efficiencies for three stage hierarchical procedures where $n_1 = 256$, $p = 0.01$, and $m = 32$ by pairwise correlation σ, stage two pool size n_2, arrangement, and model

4 Matrix procedures

4.1 Notation and general matrix procedures

Consider a matrix based procedure where $n = rc$ units are arranged in a matrix with r rows and c columns. First, r row pools and c column pools are tested. If any rows and columns test positive, then units at the intersections of positive rows and columns are tested. In this section, let Y_{ij} be 1 if the unit in the ith row and the jth column is positive and 0 otherwise, for $i = 1, \ldots, r$ and $j = 1, \ldots, c$. Let $R_i = \max(Y_{i1}, \ldots, Y_{ic})$ and $C_j = \max(Y_{1j}, \ldots, Y_{rj})$. The random variables R_i and C_j represent the observed responses for the tests corresponding to the ith row and the jth column, respectively. In general, for an $r \times c$ matrix, the expected number of tests equals

$$E(T) = r + c + \sum_{i=1}^{r} \sum_{j=1}^{c} \text{pr}(R_i = C_j = 1) \tag{8}$$

where

$$\text{pr}(R_i = C_j = 1) = 1 - \{\text{pr}(R_i = 0) + \text{pr}(C_j = 0) - \text{pr}(R_i = C_j = 0)\}. \tag{9}$$

For each row i, let $m_{i\cdot k}$ be the number of units from cluster k in row i, and for each column j and let $m_{\cdot jk}$ be the number of units from cluster k in column j. Let m_{ijk} be the

number of units from cluster k in either row i or column j. Then

$$\text{pr}(R_i = 0) = \prod_{k=1}^{l} q_{m_{i\cdot k}}, \tag{10}$$

$$\text{pr}(C_j = 0) = \prod_{k=1}^{l} q_{m_{\cdot jk}},$$

and

$$\text{pr}(R_i = C_j = 0) = \prod_{k=1}^{l} q_{m_{ijk}}.$$

Let $\tilde{m}_{i\cdot\cdot} = (m_{i\cdot 1}, \ldots, m_{i\cdot l})$, let $\tilde{m}_{\cdot j\cdot} = (m_{\cdot j1}, \ldots, m_{\cdot jl})$, and let $\tilde{m}_{ij\cdot} = (m_{ij1}, \ldots, m_{ijl})$. If the ordered values of $\tilde{m}_{i\cdot\cdot}$ and $\tilde{m}_{i'\cdot\cdot}$ are equal, the ordered values of $\tilde{m}_{\cdot j\cdot}$ and $\tilde{m}_{\cdot j'\cdot}$ are equal, and the ordered values of $\tilde{m}_{ij\cdot}$ and $\tilde{m}_{i'j'\cdot}$ are equal for all $i \neq i'$ and $j \neq j'$, then $\text{pr}(R_i = C_j = 1)$ is the same for all i and j, and (8) reduces to

$$E(T) = r + c + rc\text{pr}(R_i = C_j = 1). \tag{11}$$

For a matrix procedure where all units are independent, $\sigma = 0$ and the expected tests per unit is $E(T)/n = c^{-1} + r^{-1} + 1 - (q_1^c + q_1^r - q_1^{r+c-1})$ by (9) and (11), as in Phatarfod and Sudbury (1994).

4.2 Rectangular arrangement

In a rectangular arrangement, clusters of m units are arranged in sub-matrices of dimension $r' \times c'$ so $m = r'c'$. These sub-matrices are arranged in a matrix of dimensions $r \times c$. The number of rows r is assumed to be divisible by r' and the number of columns c is assumed to be be divisible by c'. In a rectangular arrangement, clusters are arranged in a way that (11) holds. Since $\text{pr}(R_i = 0) = q_{c'}^{c/c'}$, $\text{pr}(C_j = 0) = q_{r'}^{r/r'}$, and $\text{pr}(R_i = C_j = 0) = q_{(c'+r'-1)} q_{c'}^{c/c'-1} q_{r'}^{r/r'-1}$, by (9) and (11)

$$E(T) = r + c + rc\{1 - (q_{c'}^{c/c'} + q_{r'}^{r/r'} - q_{(c'+r'-1)} q_{c'}^{c/c'-1} q_{r'}^{r/r'-1})\}.$$

4.3 Diagonal arrangement

In a diagonal arrangement, assume $r = c = m$. Clusters of size m are arranged on diagonals of a matrix such that each row and each column has one and only one unit from each cluster. More precisely, for any $i \in \{1, \ldots, r-1\}$ and $j \in \{1, \ldots, c-1\}$, the responses Y_{ij} and $Y_{(i+1)(j+1)}$ will correspond to units from the same cluster in a diagonal arrangement. Clusters can wrap such that the last unit in a row of the matrix is a member of the same cluster as the first unit in the next row of the matrix. In this arrangement, clusters are arranged in a way that (11) holds. Since $\text{pr}(R_i = 0) = q_1^r$, $\text{pr}(C_j = 0) = q_1^r$, and $\text{pr}(R_i = C_j = 0) = q_1 q_2^{r-1}$, by (9) and (11)

$$E(T) = 2r + r^2 \left\{1 - \left(2q_1^r - q_1 q_2^{r-1}\right)\right\}.$$

4.4 Random arrangement

Now consider the case where units are arranged in a matrix randomly in a way that is unrelated to their clusters. Let $\tilde{M}_{i\cdot}$ be the random vector of the number of units from each cluster $1, \ldots, l$ in row i, let $\tilde{M}_{\cdot j}$ be the random vector of the number of units from each cluster $1, \ldots, l$ in column j, and let $\tilde{M}_{ij\cdot}$ be the random vector of the number of units from each cluster $1, \ldots, l$ in either row i or column j. Each possible arrangement of n units has the same probability, so $\tilde{M}_{i\cdot}$ has a multivariate hypergeometric distribution such that

$$\operatorname{pr}(\tilde{M}_{i\cdot} = \tilde{m}_{i\cdot t}) = \binom{n}{c}^{-1} \prod_{k=1}^{l} \binom{m}{m_{i\cdot kt}},$$

where $\tilde{m}_{i\cdot t} = (m_{i\cdot 1t}, \ldots, m_{i\cdot lt})$ is the tth possible vector of values of $\tilde{m}_{i\cdot t}$, $t = 1, \ldots, \binom{n}{c}$.

By (10),

$$\operatorname{pr}(R_i = 0 \mid \tilde{M}_{i\cdot t} = \tilde{m}_{i\cdot t}) = \prod_{k=1}^{l} q_{m_{i\cdot kt}},$$

so

$$\operatorname{pr}(R_i = 0) = \binom{n}{c}^{-1} \sum_{t=1}^{\binom{n}{c}} \left\{ \prod_{k=1}^{l} q_{m_{i\cdot kt}} \binom{m}{m_{i\cdot kt}} \right\}.$$

Additionally, $\operatorname{pr}(C_j = 0)$ and $\operatorname{pr}(R_i = C_j = 0)$ can be calculated in an analogous way. Similarly to $\operatorname{pr}(V_{(s-1)i} = 1)$ in a randomly arranged hierarchical procedure, calculating $\operatorname{pr}(R_i = 0)$, $\operatorname{pr}(C_j = 0)$, and $\operatorname{pr}(R_i = C_j = 0)$ becomes computationally infeasible as n increases, and Monte Carlo simulation can be used to approximate each of them. The efficiency, $E(T)/n$, can then be calculated by (9) and (11).

4.5 Comparison of matrix arrangements

Figure 3 shows the expected tests per unit for a square matrix of size 16×16 with clusters of size 16 for different rectangular arrangements, a diagonal arrangement, and a random arrangement. Efficiencies for the random arrangement were obtained by Monte Carlo simulation. For rectangular arrangements, the expected number of tests per unit decreases as σ increases. For the beta-binomial model, the expected tests per unit is lowest when clusters are arranged in a row, and the expected tests per unit increases as the arrangement of clusters moves from a single row to a 4×4 square. For the Madsen and Morel–Neerchal models, the rectangular arrangements perform about the same. Intuitively, a diagonal arrangement will perform worse than a rectangular arrangement, because positive responses in the same cluster will be in different rows and columns, and therefore more individual testing will be required. This intuition is supported by Fig. 3, where the diagonal arrangement performs much worse than the other arrangements as σ increases. In the diagonal arrangement, the most units from the same cluster that are tested together is two. The joint distribution for a cluster of size two is fully specified by the first and second moments, so the efficiency for the diagonal arrangement is the same for all three models. The efficiency for the randomly arrangement is worse than the rectangular arrangements, but better than the diagonal arrangement in this case.

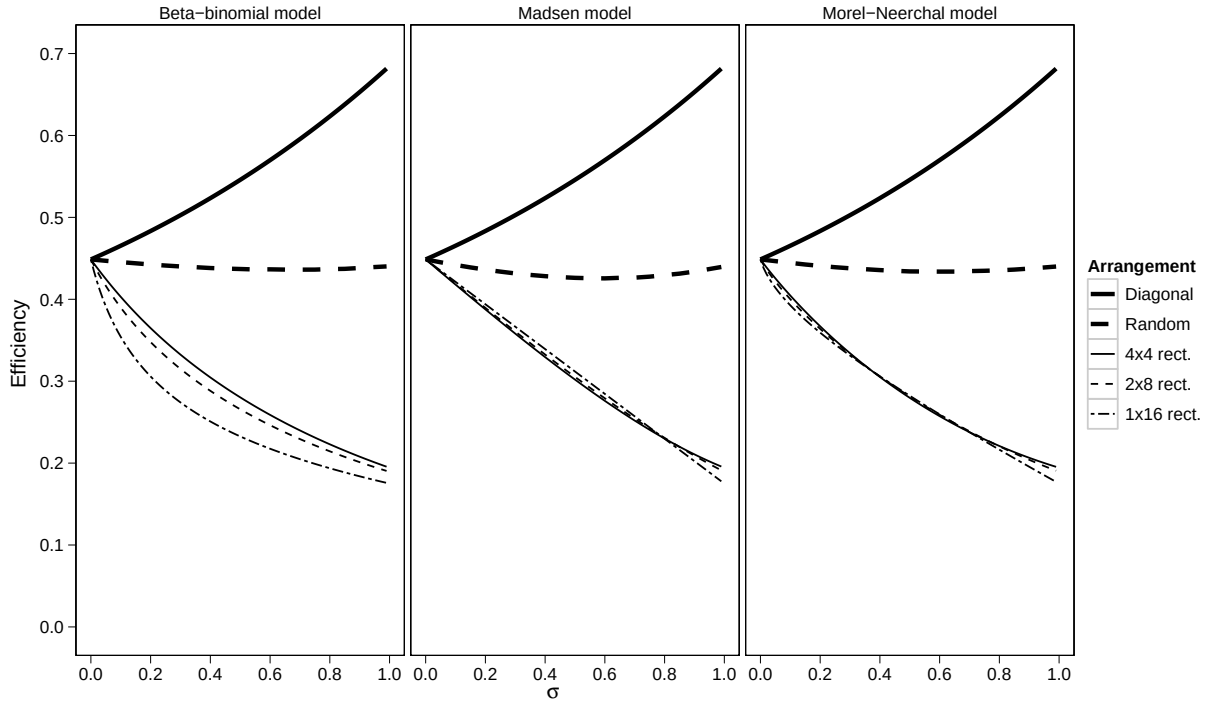

Figure 3: Efficiencies for a 16×16 matrix procedure where $p = 0.05$ and clusters are of size $m = 16$ by arrangement, pairwise correlation σ, and model

5 Application

Malhotra et al. (2007a) used a 9×10 matrix procedure to evaluate T-cell responses to 90 peptides. The matrix algorithm was used to test for peptide responses for each of 23 subjects in the study, so there were a total of 2030 T-cell responses to classify. The peptides were made up of 15 amino acids, with some pairs of peptides overlapping by 10 or more amino acids. To illustrate the potential gain in efficiency when clusters are arranged strategically for group testing, we consider the efficiency of the 9 x 10 matrix procedure for different possible peptide arrangements. Assume the 90 peptides can be partitioned into groups of size 5 or 10 such that T-cell responses to each group of peptides form an exchangeable cluster with positive pairwise correlations. From Fig. 2A of Malhotra et al. (2007a), there were a total of 151 positive responses to the set of 90 peptides for all subjects. Therefore suppose for this illustration the probability of a positive T-cell responses is 0.07 (i.e., approximately 151/2030).

Figure 4 shows the efficiency of the 9×10 matrix procedure if the clusters are in a rectangular arrangement compared to a random arrangement. For the rectangular arrangements, the clusters of size 5 are arranged in sub-matrices of size 1×5 and the clusters of size 10 are arranged in sub-matrices of size 1×10. The efficiencies for the random arrangements are obtained by Monte Carlo simulation. For both of these cluster sizes, the rectangular arrangements have a substantial gain in efficiency over the random arrangements for all three model choices. At $\sigma = 0.4$ for $m = 5$, the efficiency for the rectangular arrangement is 0.39

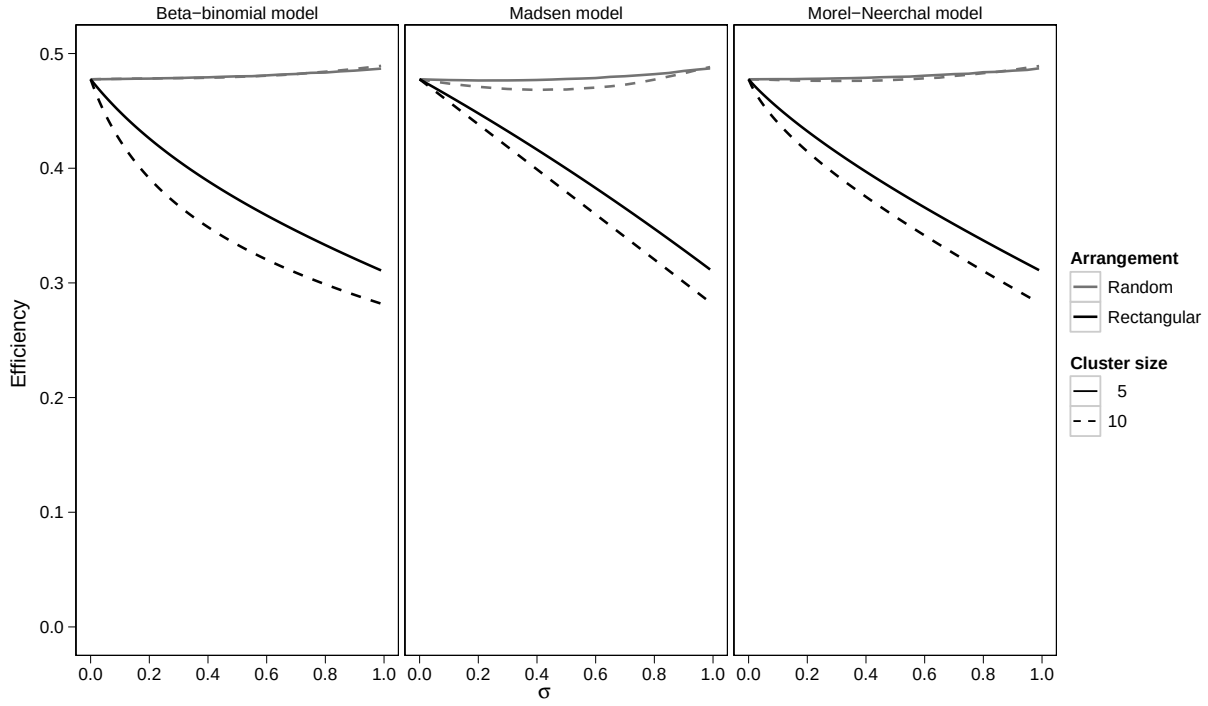

Figure 4: Efficiencies for a 9×10 matrix procedure where $p = 0.07$ by pairwise correlation σ, cluster size m, arrangement, and model.

versus 0.48 for the random arrangement from the beta-binomial model, resulting in 0.09 fewer tests per peptide on average. For each of the 23 subjects, 90 peptides are evaluated, so there is a potential savings of about 186 tests by strategically arranging peptides within a matrix. Malhotra et al. (2007a) also tested an additional 49 peptides, for which similar savings may be possible. This study only examines peptides associated with the Nef gene, but other studies evaluate a much larger number of peptides across the HIV genome (Russell et al., 2003; Koup et al., 2010). For such large scale studies, the savings from a strategic arrangement of peptides can be substantial.

6 Discussion

This paper considers group testing where units may be correlated. For the models considered, the efficiencies for hierarchical and matrix based procedures are expressed in closed form. These results allow investigation into the effect of the arrangement of clusters on a procedure's efficiency. In general, if units from the same cluster can be tested together, then the efficiency of a particular procedure can be improved, sometimes substantially, relative to random arrangements which ignore information about cluster membership. The results in this paper can be easily generalized to handle different cluster sizes, different arrangements of clusters for both hierarchical and matrix based procedures, and different prevalences between clusters, but not within a cluster. In future research, other correlation structures

could be considered, such as an autoregressive structure. Kim et al. (2007) examined the operating characteristics of both hierarchical and matrix based procedures in the presence of test error, and this could be explored further by relaxing the assumption of independence between units. Also, methods to identify optimal group testing procedures when units are correlated could be developed.

Acknowledgements

Samuel D. Lendle and Michael G. Hudgens were supported by National Institutes of Health grant R01 AI029168.

References

Berger, T., Mandell, J., and Subrahmanya, P. (2000). Maximally efficient two-stage screening. *Biometrics* **56,** 833–840.

de Finetti, B. (1975). *Theory of probability. A critical introductory treatment.*, Volume 2. John Wiley & Sons Ltd.

Diaconis, P. (1977). Finite forms of de Finetti's theorem on exchangeability. *Synthese* **36,** 271–281.

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

Finucan, H. M. (1964). The blood testing problem. *Applied Statistics* **13,** 43–50.

Hughes-Oliver, J. and Rosenberger, W. (2000). Efficient estimation of the prevalence of multiple rare traits. *Biometrika* **87,** 315.

Johnson, N. L., Kotz, S., and Wu, X. (1991). *Inspection errors for attributes in quality control*. Chapman and Hall/CRC.

Kim, H. and Hudgens, M. G. (2009). Three-dimensional array-based group testing algorithms. *Biometrics* **65,** 903–910.

Kim, H., Hudgens, M. G., Dreyfuss, J. M., Westreich, D. J., and Pilcher, C. D. (2007). Comparison of group testing algorithms for case identification in the presence of test error. *Biometrics* **63,** 1152–1163.

Koup, R., Roederer, M., Lamoreaux, L., Fischer, J., Novik, L., Nason, M., Larkin, B., Enama, M., Ledgerwood, J., Bailer, R., et al. (2010). Priming immunization with DNA augments immunogenicity of recombinant adenoviral vectors for both HIV-1 specific antibody and T-cell responses. *PloS ONE* **5,** e9015.

Lancaster, V. A. and Keller-McNulty, S. (1998). A review of composite sampling methods. *Journal of the American Statistical Association* **93,** 1216–1230.

Li, F., Malhotra, U., Gilbert, P. B., Hawkins, N. R., Duerr, A. C., McElrath, J. M., Corey, L., and Self, S. G. (2006). Peptide selection for human immunodeficiency virus type 1 CTL-based vaccine evaluation. *Vaccine* **24,** 6893–6904.

Madsen, R. (1993). Generalized binomial distributions. *Communications in Statistics-Theory and Methods* **22,** 3065–3086.

Malhotra, U., Li, F., Nolin, J., Allison, M., Zhao, H., Mullins, J., Self, S., and McElrath, M. (2007a). Enhanced detection of human immunodeficiency virus type 1 (HIV-1) Nef-specific T cells recognizing multiple variants in early HIV-1 infection. *Journal of Virology* **81,** 5225.

Malhotra, U., Nolin, J., Mullins, J., and McElrath, M. (2007b). Comprehensive epitope analysis of cross-clade Gag-specific T-cell responses in individuals with early HIV-1 infection in the US epidemic. *Vaccine* **25,** 381–390.

Morel, J. and Neerchal, N. (1997). Clustered binary logistic regression in teratology data using a finite mixture distribution. *Statistics in Medicine* **16,** 2843.

Phatarfod, R. M. and Sudbury, A. (1994). The use of a square array scheme in blood testing. *Statistics in Medicine* **13,** 2337–2343.

Roederer, M. and Koup, R. A. (2003). Optimized determination of T cell epitope responses. *Journal of Immunological Methods* **274,** 221–228.

Russell, N., Hudgens, M., Ha, R., Havenar-Daughton, C., and McElrath, M. (2003). Moving to human immunodeficiency virus type 1 vaccine efficacy trials: defining T cell responses as potential correlates of immunity. *The Journal of Infectious Diseases* **187,** 226–242.

Skellam, J. (1948). A probability distribution derived from the binomial distribution by regarding the probability of success as variable between the sets of trials. *Journal of the Royal Statistical Society. Series B (Methodological)* **10,** 257–261.

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

Yan, J., Yoon, H., Kumar, S., Ramanathan, M., Corbitt, N., Kutzler, M., Dai, A., Boyer, J., and Weiner, D. (2007). Enhanced cellular immune responses elicited by an engineered HIV-1 subtype B consensus-based envelope DNA vaccine. *Molecular Therapy* **15,** 411–421.

A

A.1 Technical Details

Proof of Lemma 1. For $n = 1, 2, \ldots$, let $\{\sigma_n\}$ be a sequence of real numbers with $\{0 \leq \sigma < 1\}$ for all n such that σ_n converges to 1, i.e. $\lim_{n\to\infty} \sigma_n = 1$. For $n = 1, 2, \ldots$, let $\dot{X}_n$ be the sum of m exchangeable binary random variables, each with mean p and with pairwise correlation

σ_n. Let $Z_n = \dot{X}_n/m$ so $E(Z_n) = p$, $\text{var}(Z_n) = m^{-2}\text{var}(\dot{X}_n) = m^{-1}p(1-p)\{1+(m-1)\sigma_n\}$, and $\lim_{n\to\infty}\text{var}(Z_n) = p - p^2$, implying $\lim_{n\to\infty} E(Z_n^2) = p$.

Let $A_n = \{0 < Z_n < 1\}$, $\text{pr}(A_n) = \alpha_n$, $\text{pr}(Z_n = 1) = \beta_n$, $E(Z_n \mid A_n) = \mu_n$ and $E(Z_n^2 \mid A_n) = \nu_n$. For all n, $E(Z_n) = E(Z_n \mid A_n)\text{pr}(A_n) + \text{pr}(Z_n = 1) = \mu_n\alpha_n + \beta_n = p$. For all n,

$$\mu_n - \nu_n = \sum_{i=1}^{m-1}\left[\left\{\frac{i(m-i)}{m^2}\right\}\text{pr}(Z_n = i/m \mid A_n)\right] \geq \frac{m-1}{m^2},$$

so $\nu_n \leq \mu_n - (m-1)/m^2$. Let $c = (m-1)/m^2$. This implies $E(Z_n^2) = E(Z_n^2 \mid A_n)\text{pr}(A_n) + \text{pr}(Z_n = 1) = \nu_n\alpha_n + \beta_n \leq (\mu_n - c)\alpha_n + \beta_n = \mu_n\alpha_n + \beta_n - c\alpha_n = p - c\alpha_n$. Because c is a positive constant and $\lim_{n\to\infty} E(Z_n^2) = p$, $\lim_{n\to\infty}\alpha_n = 0$. Therefore, $\lim_{n\to\infty}\text{pr}(0 < \dot{X}_n < m) = 0$.

For all n, $\mu_n \leq (m-1)/m < 1$ so $\lim_{n\to\infty}\mu_n\alpha_n = 0$. Since $\mu_n\alpha_n + \beta_n = p$, it follows $\lim_{n\to\infty}\beta_n = p$. This implies $\lim_{n\to\infty}\text{pr}(Z_n = 0) = 1-p$, so $\lim_{n\to\infty}\text{pr}(\dot{X}_n = 0) = 1-p$ and $\lim_{n\to\infty}\text{pr}(\dot{X}_n = m) = p$. □

Proof of Lemma 2.

$$\begin{aligned}
\text{pr}(\dot{X}' = \dot{x}') &= \sum_{\dot{x}=\dot{x}'}^{m-(m'-\dot{x}')} \text{pr}(\dot{X}' = \dot{x}', \dot{X} = \dot{x}) \\
&= \sum_{\dot{x}=\dot{x}'}^{m-(m'-\dot{x}')} \text{pr}(\dot{X} = \dot{x})\text{pr}(\dot{X}' = \dot{x}' \mid \dot{X} = \dot{x}) \\
&= \sum_{\dot{x}=\dot{x}'}^{m-(m'-\dot{x}')} E_\pi\left\{\binom{m}{\dot{x}}\pi^{\dot{x}}(1-\pi)^{m-\dot{x}}\right\}\frac{\binom{m'}{\dot{x}'}\binom{m-m'}{\dot{x}-\dot{x}'}}{\binom{m}{\dot{x}}} \\
&= E_\pi\left\{\binom{m'}{\dot{x}'}\pi^{\dot{x}'}(1-\pi)^{m'-\dot{x}'} \times \sum_{\dot{x}=\dot{x}'}^{m-(m'-\dot{x}')}\left(\binom{m-m'}{\dot{x}-\dot{x}'}\pi^{\dot{x}-\dot{x}'}(1-\pi)^{m-m'-(\dot{x}-\dot{x}')}\right)\right\} \\
&= E_\pi\left\{\binom{m'}{\dot{x}'}\pi^{\dot{x}'}(1-\pi)^{m'-\dot{x}'} \times \sum_{\dot{x}=0}^{m-m'}\left(\binom{m-m'}{\dot{x}}\pi^{\dot{x}}(1-\pi)^{m-m'-\dot{x}}\right)\right\} \\
&= E_\pi\left\{\binom{m'}{\dot{x}'}\pi^{\dot{x}'}(1-\pi)^{m'-\dot{x}'}\right\} \qquad \square
\end{aligned}$$

Proof of Lemma 3. Suppose $m' = 1$, so $E(\pi) = \text{pr}(\dot{X}' = 1) = p$. When $m' = 1$, $\text{pr}(\dot{X}' = 1) = \text{pr}(X_i = 1) = E(X_i)$ for all i, so $E(X_i) = p$ by Lemma 2. Suppose $m' = 2$, so $E(\pi^2) = \text{pr}(\dot{X}' = 2) = \sigma p(1-p) + p^2$. For all $i \neq j$, $E(X_iX_j) = \text{pr}(X_i = 1, X_j = 1) = \text{pr}(\dot{X}' = 2)$ and

$$\begin{aligned}
\text{cor}(X_i, X_j) &= \frac{E(X_iX_j) - E(X_i)E(X_j)}{\sqrt{\text{var}(X_i)\text{var}(X_j)}} \\
&= \frac{\sigma p(1-p) + p^2 - p^2}{p(1-p)} \\
&= \sigma \qquad \square
\end{aligned}$$