Group testing case identification with biomarker information

Dewei Wang[a], Christopher S. McMahan[b], Joshua M. Tebbs[a,*], Christopher R. Bilder[c]

[a]*Department of Statistics, University of South Carolina, Columbia, SC 29208, U.S.A*
[b]*Department of Mathematical Sciences, Clemson University, Clemson, SC 29634, U.S.A.*
[c]*Department of Statistics, University of Nebraska-Lincoln, Lincoln, NE 68583, U.S.A.*

Abstract

Screening procedures for infectious diseases, such as HIV, often involve pooling individual specimens together and testing the pools. For diseases with low prevalence, group testing (or pooled testing) can be used to classify individuals as diseased or not while providing considerable cost savings when compared to testing specimens individually. The pooling literature is replete with group testing case identification algorithms including Dorfman testing, higher-stage hierarchical procedures, and array testing. Although these algorithms are usually evaluated on the basis of the expected number of tests and classification accuracy, most evaluations in the literature do not account for the continuous nature of the testing responses and thus invoke potentially restrictive assumptions to characterize an algorithm's performance. This article revisits the evaluation of commonly used case identification algorithms in group testing but takes a different approach. Instead of treating testing responses as binary random variables (i.e., diseased/not), evaluations are made by exploiting an assay's underlying continuous biomarker distributions for positive and negative individuals. In doing so, a general framework to describe the operating characteristics of group testing case identification algorithms is provided when these distributions are known. The methodology is illustrated using two HIV testing examples taken

*Corresponding author
Email address: tebbs@stat.sc.edu (Joshua M. Tebbs)

from the pooling literature.

Keywords: Classification, Measurement error, Pooled testing, Screening, Sensitivity, Specificity

---

1. Introduction

Testing individual specimens in pools, which is known as group testing (or pooled testing), is widespread in disease screening applications. Individuals in pools that test negatively are declared to be negative, and positive pools are resolved (or "decoded") to determine which individuals are positive. The origins of group testing are usually traced back to Dorfman (1943), who proposed that it be used to screen World War II soldiers for syphilis. Since this seminal work, group testing has been applied to numerous infectious disease applications. A literature review reveals recent public health and surveillance applications for HIV (Krajden et al., 2014), HBV and HCV (Page-Shafer et al., 2008; Candotti and Allain, 2009), chlamydia and gonorrhea (Lewis et al., 2012), West Nile virus (Busch et al., 2005), and influenza (Edouard et al., 2015). Group testing is also routinely used by national organizations around the world to screen blood and plasma donations for HIV/HBV/HCV and other diseases (see, e.g., Schmidt et al., 2010; O'Brien et al., 2012; Stramer et al., 2013).

The original procedure proposed by Dorfman (1943) is a two-stage hierarchical algorithm; i.e., non-overlapping pools are tested in the first stage and individuals from positive pools are tested in the second. Hierarchical algorithms using a larger number of stages can reduce the number of tests needed when the disease prevalence is small. For example, Mehta et al. (2011) describe a three-stage algorithm for HIV testing in San Diego that uses master pools of size 10 in the first stage, subpools of size 5 in the second stage, and individual testing in the third. The most common non-hierarchical algorithm is two-dimensional array testing (Phatarfod and Sudbury, 1994; Hudgens and Kim, 2011; McMahan et al., 2012b), where individuals are tested in the rows and columns of an array. A recent HIV application in New Jersey (Martin et al., 2013) illustrates

how array testing can even be used in higher dimensions (Kim and Hudgens, 2009). Comprehensive summaries of group testing algorithms and their operating characteristics are found in Kim et al. (2007) and Westreich et al. (2008).

When faced with the task of choosing an appropriate case identification algorithm for screening purposes, public health officials and lab technicians are interested in cost and accuracy. Laboratories with large budgets may opt to test specimens individually as pooling can reduce an assay's sensitivity. In the group testing literature, this reduction is known as "the dilution effect" and can result in an increased number of false negative diagnoses. Group testing algorithms can be selected on the basis of minimizing the expected number of tests per individual to minimize costs (Kim et al., 2007; Westreich et al., 2008) or perhaps in a way that incorporates both the expected number of tests and classification accuracy (see, e.g., Malinovsky et al., 2016). Of course, additional practical considerations such as testing platform constraints, the time needed for testing, and the availability of individuals to pool should also be carefully considered.

When an individual or pooled specimen is tested, an assay typically elicits a binary diagnosis (positive/negative) that is derived from measuring a continuous biomarker; large values of this continuous measurement are usually evidence that the disease is present. Although it is widely known that dichotomizing a continuous outcome can lead to a loss in information, previous evaluations in group testing have largely ignored this underlying aspect and instead have relied explicitly on binary results. Doing so helps to facilitate the derivation of closed-form expressions for the expected number of tests and classification accuracy probabilities; however, this also usually requires one to make assumptions such as (a) the sensitivity and specificity are unaffected by pool size; i.e., there is no dilution effect, and (b) testing outcomes on pools containing common individuals are independent conditional on the true pool statuses. An important contribution of this article is to provide a general framework for case identification evaluation where these assumptions are not needed.

In offering this framework, our approach exploits the underlying continuous

biomarker distributions associated with positive and negative individuals. In other words, we do not dichotomize testing outcomes into "positive" or "negative" categories, but instead we make our evaluations in terms of the biomarker distributions themselves. Our work is related to the methodology in Wein and Zenios (1996), who proposed using biomarker concentrations to determine an optimized Dorfman algorithm for HIV testing. However, our article takes a somewhat different perspective. We are not focused on determining optimal designs for specific group testing procedures per se; instead, our goal is to enhance previous case identification algorithm evaluations, such as those in Kim et al. (2007) and Westreich et al. (2008), in group testing applications where biomarker distributions are known. Our evaluations can be performed for any group testing procedure, including Dorfman testing, higher-stage hierarchical algorithms, and array testing. We obtain closed-form expressions for operating characteristics for normally distributed biomarkers in specific algorithms; however, even these expressions may be of limited utility for practitioners. We therefore use simulation to overcome the computational challenges when incorporating biomarker information.

2. Notation and Preliminaries

We modify the notation from Wang et al. (2015), who used biomarker distributions to acknowledge the dilution effect in group testing regression. Let $T_i = 1$ if the ith individual is truly positive; $T_i = 0$ otherwise. We assume the T_i's are independent and identically distributed statuses with $\text{pr}(T_i = 1) = p$, the prevalence of the population. Generalizing our evaluation framework to allow for unequal individual disease probabilities (McMahan et al., 2012a; 2012b) or correlated individuals (Lendle et al., 2012) is straightforward; see Section 6. Let $\widetilde{\mathcal{C}}_i$ denote the true biomarker level of the ith individual (e.g., viral load, optical density reading, antibody concentration, etc.). We assume the $\widetilde{\mathcal{C}}_i$'s are mutually independent random variables and that the conditional

probability density function of $\widetilde{\mathcal{C}}_i$ given the true status $T_i = t$ is

$$f_{\widetilde{\mathcal{C}}_i|T_i=t}(u) = tf_{\widetilde{\mathcal{C}}^+}(u) + (1-t)f_{\widetilde{\mathcal{C}}^-}(u),$$

where $f_{\widetilde{\mathcal{C}}^+}$ and $f_{\widetilde{\mathcal{C}}^-}$ denote the true biomarker density functions for positive and negative individuals, respectively. In other words, positive individuals in the population have true biomarker levels described by the common density $f_{\widetilde{\mathcal{C}}^+}$; similarly, negative individuals' true biomarker levels are described by $f_{\widetilde{\mathcal{C}}^-}$.

We are interested in calculating quantities like the expected number of tests per individual and classification accuracy probabilities commonly seen in the group testing case identification literature (i.e., pooling sensitivity, pooling specificity, predictive values). To set our ideas, we assume a hierarchical group testing algorithm is used in $S \geq 2$ stages, although we later modify our notation to account for array testing in two dimensions (Phatarfod and Sudbury, 1994; Hudgens and Kim, 2011; McMahan et al., 2012b); see Section 3.3. An S-stage hierarchical algorithm begins by testing a master pool of individual specimens. If the master pool tests negatively, all individuals are declared to be disease-free and no further testing is performed. Otherwise, non-overlapping subpools are formed and are tested in the second stage. Any second-stage subpool that tests positively is split again while subpools that test negatively in the second stage are declared to be disease-free. This process continues until all subpools in a particular stage test negatively or until individual testing (in stage S) is performed.

For an S-stage hierarchical algorithm, let $\mathcal{P}_{sl}$ denote the index set of individuals in the lth pool formed at the sth stage of testing, for $l = 1, 2, ..., n_1/n_s$ and $s = 1, 2, ..., S$, where $n_s = |\mathcal{P}_{sl}|$ is the number of individuals in $\mathcal{P}_{sl}$. To illustrate this notation, Figure 1 displays the $S = 3$ stage hierarchical algorithm described in Mehta et al. (2011) from Section 1. In this example, the master pool is $\mathcal{P}_{11} = \{1, 2, ..., 10\}$, the two second-stage pools are $\mathcal{P}_{21} = \{1, 2, ..., 5\}$ and $\mathcal{P}_{22} = \{6, 7, ..., 10\}$, and the singleton pools $\mathcal{P}_{31} = \{1\}, \mathcal{P}_{32} = \{2\}, ..., \mathcal{P}_{3,10} = \{10\}$ are for individual testing in the third stage. These pools are of size $n_1 = 10$, $n_2 = 5$, and $n_3 = 1$. Additional examples of hierarchical algorithms used in HIV

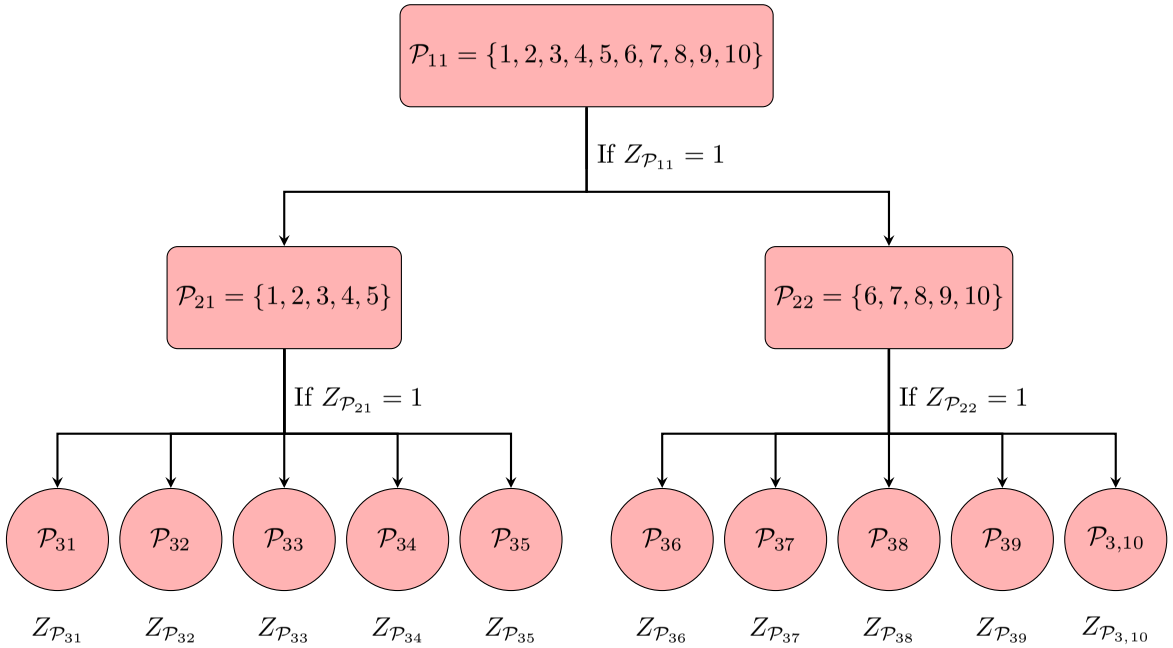

Figure 1: $S = 3$ stage hierarchical group testing algorithm $H(n_1 = 10 : n_2 = 5 : n_3 = 1)$.

testing are found in Sherlock et al. (2007). Henceforth, a general S-stage hierarchical algorithm is denoted by $H(n_1 : n_2 : \cdots : n_S)$, where $n_S = 1$. Note that Dorfman's seminal strategy uses $S = 2$ stages.

Let $T_{\mathcal{P}_{sl}} = 1$ if the lth pool in the sth stage is truly positive; i.e., $\mathcal{P}_{sl}$ contains at least one truly positive individual, $T_{\mathcal{P}_{sl}} = 0$ otherwise. Similarly, let $Z_{\mathcal{P}_{sl}} = 1$ if $\mathcal{P}_{sl}$ tests positively, $Z_{\mathcal{P}_{sl}} = 0$ otherwise. To acknowledge the continuous nature of the diagnostic assay, we assume that $Z_{\mathcal{P}_{sl}} = I(\mathcal{C}_{\mathcal{P}_{sl}} > \tau_{\mathcal{P}_{sl}})$; i.e., the pool $\mathcal{P}_{sl}$ tests positively if $\mathcal{C}_{\mathcal{P}_{sl}}$, the measured biomarker level of the pool, exceeds a threshold $\tau_{\mathcal{P}_{sl}}$ which potentially depends on the pool size n_s at stage s. To acknowledge the potential of error when measuring the true biomarker level $\widetilde{\mathcal{C}}_{\mathcal{P}_{sl}}$, we assume that $\mathcal{C}_{\mathcal{P}_{sl}}|\widetilde{\mathcal{C}}_{\mathcal{P}_{sl}} \sim f_\epsilon$, where $f_\epsilon = f_\epsilon(\cdot|\widetilde{\mathcal{C}}_{\mathcal{P}_{sl}})$ is a known probability density function. Therefore, our framework utilizes three distributions: the true biomarker distributions for positive and negative individuals, $f_{\widetilde{\mathcal{C}}^+}$ and $f_{\widetilde{\mathcal{C}}^-}$, respectively, and f_ϵ, which incorporates the effect of assay measurement error. Threshold selection for $\tau_{\mathcal{P}_{sl}}$ is discussed in Section 4.

As noted in Section 1, previous evaluations of case identification algorithms have largely assumed the sensitivity and specificity are constant and hence are unaffected by pool size. Although this assumption may be reasonable when testing negative pools (i.e., constant specificity), it is potentially more dubious when testing positive pools. Using the Law of Total Probability, note that the sensitivity associated with testing $\mathcal{P}_{sl}$ can be written as

$$\text{pr}(Z_{\mathcal{P}_{sl}} = 1|T_{\mathcal{P}_{sl}} = 1) = \frac{\sum_{m=1}^{n_s} \text{pr}(Z_{\mathcal{P}_{sl}} = 1|\sum_{i\in\mathcal{P}_{sl}} T_i = m)\ \text{pr}(\sum_{i\in\mathcal{P}_{sl}} T_i = m)}{\text{pr}(T_{\mathcal{P}_{sl}} = 1)},$$

where the random variable $\sum_{i\in\mathcal{P}_{sl}} T_i$ counts the number of positive individuals in $\mathcal{P}_{sl}$. Therefore, for the sensitivity to remain constant throughout the testing process, one would have to require that $\text{pr}(Z_{\mathcal{P}_{sl}} = 1|\sum_{i\in\mathcal{P}_{sl}} T_i = m)$ are equal for each $m = 1, 2, ..., n_s$, $l = 1, 2, ..., n_1/n_s$ and $s = 1, 2, ..., S$. Clearly, this requirement may be unsuitable−especially when testing results are heavily influenced by dilution.

On the other hand, when written in terms of the true biomarker distributions, $f_{\widetilde{\mathcal{C}}^+}$ and $f_{\widetilde{\mathcal{C}}^-}$, and the measurement error density f_ϵ, the sensitivity of $\mathcal{P}_{sl}$

is given by

$$S_e(n_s) = \text{pr}(Z_{\mathcal{P}_{sl}} = 1 | T_{\mathcal{P}_{sl}} = 1) = \text{pr}\left(\mathcal{C}_{\mathcal{P}_{sl}} > \tau_{\mathcal{P}_{sl}} \middle| \sum_{i \in \mathcal{P}_{sl}} T_i > 0\right) = \frac{\sum_{m=1}^{n_s} \binom{n_s}{m} p^m q^{n_s - m} S_e(n_s : m)}{1 - q^{n_s}},$$

where $q = 1 - p$ and

$$S_e(n_s : m) = \int_{\tau_{\mathcal{P}_{sl}}}^{\infty} \int_{-\infty}^{\infty} f_\epsilon(u|v) n_s f^{m(n_s-m)}_{\sum_{i \in \mathcal{P}_{sl}} \widetilde{\mathcal{C}}_i}(n_s v) dv du, \tag{1}$$

where

$$f^{m(n_s-m)}_{\sum_{i \in \mathcal{P}_{sl}} \widetilde{\mathcal{C}}_i}(v) = \int_{\sum_{i=1}^{n_s} v_i = v} \prod_{i=1}^{m} f_{\widetilde{\mathcal{C}}^+}(v_i) \prod_{i=m+1}^{n_s} f_{\widetilde{\mathcal{C}}^-}(v_i) dv_1 dv_2 ... dv_{n_s}. \tag{2}$$

The expression in Equation (2) is the density of $\sum_{i \in \mathcal{P}_{sl}} \widetilde{\mathcal{C}}_i$, the sum of the mutually independent biomarker levels in $\mathcal{P}_{sl}$ when $\mathcal{P}_{sl}$ contains exactly $m \geq 1$ positive and $n_s - m$ negative individuals; we obtain this density by convolving the true individual biomarker densities $f_{\widetilde{\mathcal{C}}^+}$ and $f_{\widetilde{\mathcal{C}}^-}$ m and $n_s - m$ times, respectively.

In writing Equation (1), we assume the true biomarker level $\widetilde{\mathcal{C}}_{\mathcal{P}_{sl}}$ is the arithmetic average of the individual biomarker levels in $\mathcal{P}_{sl}$; i.e., $\widetilde{\mathcal{C}}_{\mathcal{P}_{sl}} = n_s^{-1} \sum_{i \in \mathcal{P}_{sl}} \widetilde{\mathcal{C}}_i$. This assumption is often viewed as sacrosanct in the biomarker pooling literature (see, e.g., Zhang and Albert, 2011; Malinovsky et al., 2012; Mitchell et al., 2014; Delaigle and Hall, 2015) and is likely reasonable when pools are formed from aliquots of equal volume. Under this assumption, the specificity of $\mathcal{P}_{sl}$ is given by

$$S_p(n_s) = \text{pr}(Z_{\mathcal{P}_{sl}} = 0 | T_{\mathcal{P}_{sl}} = 0) = \text{pr}\left(\mathcal{C}_{\mathcal{P}_{sl}} < \tau_{\mathcal{P}_{sl}} \middle| \sum_{i \in \mathcal{P}_{sl}} T_i = 0\right) = \int_{-\infty}^{\tau_{\mathcal{P}_{sl}}} \int_{-\infty}^{\infty} f_\epsilon(u|v) n_s f^{0(n_s-0)}_{\sum_{i \in \mathcal{P}_{sl}} \widetilde{\mathcal{C}}_i}(n_s v) dv du, \tag{3}$$

where $f^{0(n_s-0)}_{\sum_{i \in \mathcal{P}_{sl}} \widetilde{\mathcal{C}}_i}(\cdot)$ is the density that convolves $f_{\widetilde{\mathcal{C}}^-}$ n_s times—once for each of the negative individuals in $\mathcal{P}_{sl}$. Note that Equations (1) and (3) are similar in form to the analogous expressions found in McMahan et al. (2013) and Delaigle

and Hall (2015), both of whom incorporate biomarker and measurement error distributions in group testing regression.

As an example, suppose the true individual biomarker distributions for negative and positive individuals are $\widetilde{\mathcal{C}}^- \sim \mathcal{N}(3, 0.25)$ and $\widetilde{\mathcal{C}}^+ \sim \mathcal{N}(6, 1)$, respectively, and that the measurement error density is $\mathcal{N}(\widetilde{\mathcal{C}}, 0.0025)$. For these distribution choices, the threshold that maximizes Youden's index (Youden, 1950) for individual testing is $\tau^* = 4.11$, which provides values of sensitivity and specificity (for individual testing) equal to 0.970 and 0.987, respectively. To illustrate the effect of pooling, Figure 2 displays the densities of the measured biomarker level on $\mathcal{P}_{sl}$; i.e., $f_{\mathcal{C}_{\mathcal{P}_{sl}}}(u) = \int_{-\infty}^{\infty} f_\epsilon(u|v) n_s f^{m(n_s-m)}_{\sum_{i\in\mathcal{P}_{sl}} \widetilde{\mathcal{C}}_i}(n_s v) dv$, for different values of m when the pool size is $n_s = 5$ and $n_s = 10$. This figure illustrates how relevant operating characteristics in group testing could ultimately depend on the individual biomarker distributions, the pool size, the threshold used for pools (see Section 4), and the number of positive individuals in each pool. In other words, once one moves beyond treating pool and individual diagnoses as binary, case identification evaluation becomes far more complicated. Note that we have created Figure 2 assuming normality for $\widetilde{\mathcal{C}}^-$, $\widetilde{\mathcal{C}}^+$, and the measurement error so that $f_{\mathcal{C}_{\mathcal{P}_{sl}}}(u)$ can be calculated exactly. However, biomarkers in real applications are rarely normally distributed and calculating $f_{\mathcal{C}_{\mathcal{P}_{sl}}}(u)$ for non-normal biomarkers, if it is even possible to do so, potentially involves high-dimensional integration (i.e., of dimension equal to the pool size).

3. Operating Characteristics

3.1. Efficiency

The most important characteristic of a group testing case identification algorithm is its expected number of tests per individual, or *efficiency*. Because the cost of screening is usually highly correlated with the number of tests expended, algorithms with lower values of this expectation are generally preferred. For example, an algorithm whose efficiency is 0.5 is twice as efficient as individual testing. An algorithm whose efficiency is larger than 1 uses more tests than in-

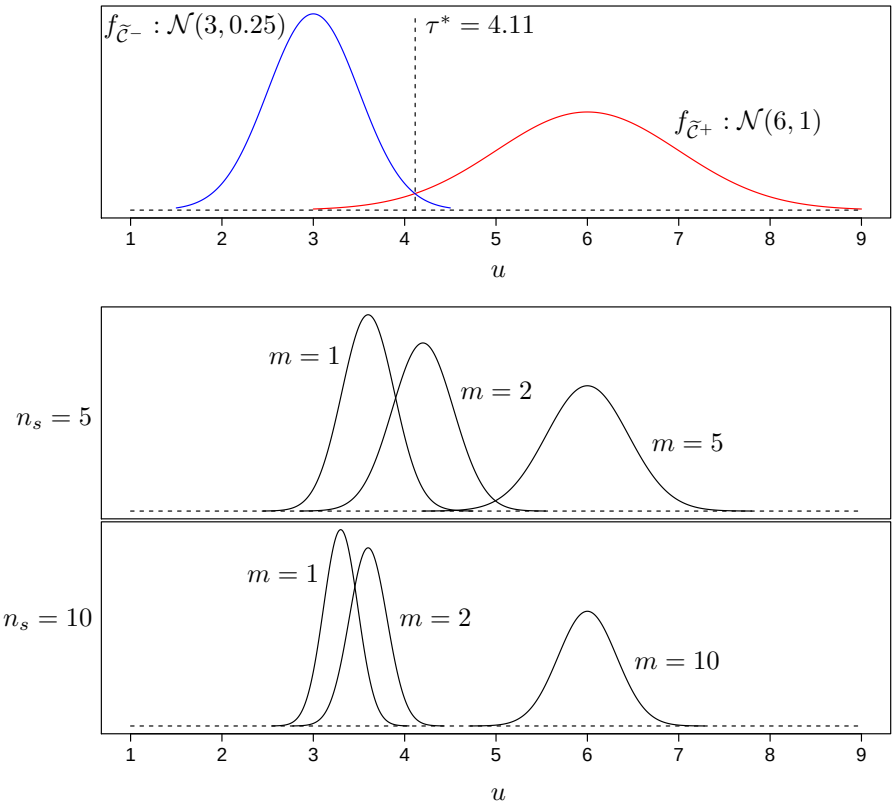

Figure 2: Top panel: Individual biomarker densities for $\widetilde{\mathcal{C}}^- \sim \mathcal{N}(3, 0.25)$ and $\widetilde{\mathcal{C}}^+ \sim \mathcal{N}(6, 1)$. Under our measurement error density assumption, $\mathcal{C}|\widetilde{\mathcal{C}} \sim \mathcal{N}(\widetilde{\mathcal{C}}, 0.0025)$, the threshold $\tau^* = 4.11$ maximizes Youden's index for individual testing. Middle and bottom panels: Densities of the pooled biomarker measurements, $f_{\mathcal{C}_{\mathcal{P}_{sl}}}(u) = \int_{-\infty}^{\infty} f_\epsilon(u|v) n_s f^{*m(n_s-m)}_{\sum_{i\in\mathcal{P}_{sl}} \widetilde{\mathcal{C}}_i}(n_s v) dv$, for different m (the number of positive individuals in $\mathcal{P}_{sl}$) when the pool size is $n_s = 5$ and $n_s = 10$.

dividual testing on average. In the group testing literature, optimal algorithms are usually identified as those that are the most efficient.

Unfortunately, within the general framework we have outlined in this article, calculating the efficiency quickly becomes unmanageable−even for simple algorithms. For example, consider the $S = 3$ stage algorithm $H(10 : 5 : 1)$ depicted in Figure 1. It is easy to see that the efficiency of this algorithm is $\frac{1}{10} + \frac{1}{5}\text{pr}(Z_{\mathcal{P}_{11}} = 1) + \text{pr}(Z_{\mathcal{P}_{11}} = 1, Z_{\mathcal{P}_{21}} = 1)$, where recall $Z_{\mathcal{P}_{11}}$ and $Z_{\mathcal{P}_{21}}$ denote the (binary) testing responses of $\mathcal{P}_{11}$ and $\mathcal{P}_{21}$, respectively. When written in terms of the biomarker distributions, the first-stage probability is

$$\begin{aligned}\text{pr}(Z_{\mathcal{P}_{11}} = 1) &= \text{pr}(\mathcal{C}_{\mathcal{P}_{11}} > \tau_{\mathcal{P}_{11}} | T_{\mathcal{P}_{11}} = 1)\text{pr}(T_{\mathcal{P}_{11}} = 1) \\ &\qquad + \text{pr}(\mathcal{C}_{\mathcal{P}_{11}} > \tau_{\mathcal{P}_{11}} | T_{\mathcal{P}_{11}} = 0)\text{pr}(T_{\mathcal{P}_{11}} = 0) \\ &= \sum_{m=1}^{10} \binom{10}{m} p^m q^{10-m} S_e(10 : m) + \{1 - S_p(10)\}q^{10},\end{aligned}$$

where $S_e(10 : m)$ is calculated using Equations (1) and (2) and $S_p(10)$ is calculated using Equation (3) with $n_s = n_1 = 10$, $\tau_{\mathcal{P}_{sl}} = \tau_{\mathcal{P}_{11}}$, and $\mathcal{P}_{sl} = \mathcal{P}_{11}$. Even more daunting, a second-stage pool tests positively with probability $\text{pr}(Z_{\mathcal{P}_{11}} = 1, Z_{\mathcal{P}_{21}} = 1)$, which equals

$$\sum_{m_1=0}^{5}\sum_{m_2=0}^{5} \binom{5}{m_2} p^{m_2} q^{5-m_2} \binom{5}{m_1} p^{m_1} q^{5-m_1} \int_{\tau_{\mathcal{P}_{11}}}^{\infty}\int_{\tau_{\mathcal{P}_{21}}}^{\infty} \left\{ \int_{\mathbb{R}^{10}} \prod_{s=1}^{2} f_\epsilon \left(u_s \middle| \frac{1}{n_s}\sum_{i=1}^{n_s} v_i \right) \right.$$
$$\left. \times \prod_{i=1}^{m_2} f_{\tilde{\mathcal{C}}^+}(v_i) \prod_{i=m_2+1}^{5} f_{\tilde{\mathcal{C}}^-}(v_i) \prod_{i=6}^{m_1+5} f_{\tilde{\mathcal{C}}^+}(v_i) \prod_{i=m_1+6}^{10} f_{\tilde{\mathcal{C}}^-}(v_i) dv_1 dv_2 \cdots dv_{10} \right\} du_2 du_1,$$

where $n_1 = 10$ and $n_2 = 5$. In this expression, it is understood that products of the form $\prod_{i=a}^{b} f_{\tilde{\mathcal{C}}^+}(v_i)$ and $\prod_{i=a}^{b} f_{\tilde{\mathcal{C}}^-}(v_i)$, $a > b$, are vacuous.

As this simple example illustrates, offering a biomarker-based framework for group testing case identification presents nearly overwhelming computational challenges. Unfortunately, this is the price one must pay when relaxing assumptions used in previous evaluations. For example, under classical assumptions in Kim et al. (2007) and Westreich et al. (2008), the probabilities we have just presented reduce to $\text{pr}(Z_{\mathcal{P}_{11}} = 1) = S_e(1 - q^{10}) + (1 - S_p)q^{10}$ and

$$\text{pr}(Z_{\mathcal{P}_{11}} = 1, Z_{\mathcal{P}_{21}} = 1) = S_e^2(1 - q^5) + S_e(1 - S_p)(q^5 - q^{10}) + (1 - S_p)^2 q^{10},$$

respectively, where S_e and S_p are the assumed common sensitivity and specificity for pools of size $n_1 = 10$ and $n_2 = 5$. The simplified formula for $\text{pr}(Z_{\mathcal{P}_{11}} = 1, Z_{\mathcal{P}_{21}} = 1)$ above arises only when the testing responses $Z_{\mathcal{P}_{11}}$ and $Z_{\mathcal{P}_{21}}$ are conditionally independent given the true pool statuses $T_{\mathcal{P}_{11}}$ and $T_{\mathcal{P}_{21}}$. This assumption is required under classical evaluations because $\mathcal{P}_{11}$ and $\mathcal{P}_{21}$ contain common individuals.

In Appendix A in the Supplementary Material, we have derived a general expression for the efficiency of an S-stage hierarchical algorithm. This derivation has been described previously in the group testing literature; see, e.g., Kim et al. (2007) and the references therein. In our notation, the efficiency can be expressed as

$$\text{EFF}\{H(n_1 : n_2 : \cdots : n_S)\} = \frac{1}{n_1} + \sum_{s=1}^{S-1} \frac{1}{n_{s+1}} \text{pr}\left(\prod_{s'=1}^{s} Z_{\mathcal{P}_{s'1}} = 1 \right),$$

where the random variable $\prod_{s'=1}^{s} Z_{\mathcal{P}_{s'1}}$ equals 1 if and only if the first pool in each of the first s stages tests positively. Calculating $\text{pr}(\prod_{s'=1}^{s} Z_{\mathcal{P}_{s'1}} = 1)$ within our framework involves accounting for the joint uncertainty that arises in the correlated, error-laden biomarker measurements $\mathcal{C}_{\mathcal{P}_{11}}, \mathcal{C}_{\mathcal{P}_{21}}, ..., \mathcal{C}_{\mathcal{P}_{s1}}$, an extremely difficult problem analytically. Although this probability can be calculated exactly under normal biomarker assumptions, in general we recommend using Monte Carlo simulation and estimating $\text{EFF}\{H(n_1 : n_2 : \cdots : n_S)\}$ instead. Such a strategy is flexible and will accommodate any biomarker and measurement error distributions. In addition, one can quickly estimate the variance of the number of tests per individual (Kim et al., 2007), which would otherwise be an intractable calculation. A description of our simulation procedure is now given.

SIMULATION PROCEDURE

1. Generate $T_1, T_2, ..., T_{n_1} \sim$ iid Bernoulli(p). Generate $\widetilde{\mathcal{C}}_i \sim f_{\widetilde{\mathcal{C}}_i|T_i=t}(u) = tf_{\widetilde{\mathcal{C}}^+}(u) + (1-t)f_{\widetilde{\mathcal{C}}^-}(u)$, $i = 1, 2, ..., n_1$.
2. (Stage 1). Calculate $\widetilde{\mathcal{C}}_{\mathcal{P}_{11}} = n_1^{-1} \sum_{i \in \mathcal{P}_{11}} \widetilde{\mathcal{C}}_i$ and generate $\mathcal{C}_{\mathcal{P}_{11}}$ from $f_\epsilon(\cdot|\widetilde{\mathcal{C}}_{\mathcal{P}_{11}})$.

(a) If $Z_{\mathcal{P}_{11}} = I(\mathcal{C}_{\mathcal{P}_{11}} > \tau_{\mathcal{P}_{11}}) = 0$, stop and classify the n_1 individuals in $\mathcal{P}_{11}$ as negative.

(b) If $Z_{\mathcal{P}_{11}} = I(\mathcal{C}_{\mathcal{P}_{11}} > \tau_{\mathcal{P}_{11}}) = 1$, divide $\widetilde{\mathcal{C}}_i \in \mathcal{P}_{11}$ into subgroups of size n_2.

3. (Stage 2). Calculate $\widetilde{\mathcal{C}}_{\mathcal{P}_{2l}} = n_2^{-1} \sum_{i \in \mathcal{P}_{2l}} \widetilde{\mathcal{C}}_i$ for each subgroup in Step 2(b) and generate $\mathcal{C}_{\mathcal{P}_{2l}}$ from $f_\epsilon(\cdot|\widetilde{\mathcal{C}}_{\mathcal{P}_{2l}})$. Calculate $Z_{\mathcal{P}_{2l}} = I(\mathcal{C}_{\mathcal{P}_{2l}} > \tau_{\mathcal{P}_{2l}})$. For each l,

   (a) if $Z_{\mathcal{P}_{2l}} = I(\mathcal{C}_{\mathcal{P}_{2l}} > \tau_{\mathcal{P}_{2l}}) = 0$, classify the n_2 individuals in $\mathcal{P}_{2l}$ as negative (stop if all second-stage subgroups are negative).

   (b) if $Z_{\mathcal{P}_{2l}} = I(\mathcal{C}_{\mathcal{P}_{2l}} > \tau_{\mathcal{P}_{2l}}) = 1$, divide $\widetilde{\mathcal{C}}_i \in \mathcal{P}_{2l}$ into subgroups of size n_3.

4. (Stage 3). For each subgroup in Step 3(b), calculate $\widetilde{\mathcal{C}}_{\mathcal{P}_{3l}} = n_3^{-1} \sum_{i \in \mathcal{P}_{3l}} \widetilde{\mathcal{C}}_i$, generate $\mathcal{C}_{\mathcal{P}_{3l}}$ from $f_\epsilon(\cdot|\widetilde{\mathcal{C}}_{\mathcal{P}_{3l}})$, and calculate $Z_{\mathcal{P}_{3l}} = I(\mathcal{C}_{\mathcal{P}_{3l}} > \tau_{\mathcal{P}_{3l}})$. Continue this overall process until all subgroups in a particular stage test negatively or until individual testing (in stage S) is performed.

We implement this procedure B times and estimate the efficiency of $H(n_1 : n_2 : \cdots : n_S)$ using

$$\widehat{\text{EFF}}\{H(n_1 : n_2 : \cdots : n_S)\} = \frac{1}{n_1 B} \sum_{b=1}^{B} M_b,$$

where M_b is the number of tests observed in the bth replication. The variance of the number of tests per individual, denoted by $\text{var}\{H(n_1 : n_2 : \cdots : n_S)\}$, can be estimated using the sample variance of $M_1/n_1, M_2/n_1, ..., M_B/n_1$. Our simulation procedure is extremely fast and thus can be performed using very large values of B. Under normal biomarker assumptions, we show in Appendix B in the Supplementary Material that the difference between calculating $\text{EFF}\{H(n_1 : n_2 : \cdots : n_S)\}$ exactly and estimating it using a large number of replications is negligible.

3.2. Classification Accuracy

Although the efficiency of a group testing case identification algorithm is its most important characteristic, being able to quantify an algorithm's classification accuracy is also critical. Two commonly used measures of accuracy in the case identification literature are *pooling sensitivity* and *pooling specificity*. For an S-stage hierarchical algorithm, the pooling sensitivity

$$\text{PSE}\{H(n_1 : n_2 : \cdots : n_S)\} = \text{pr}\left(\prod_{s=1}^{S} Z_{\mathcal{P}_{s1}} = 1 \middle| T_1 = 1\right)$$

is the probability a truly positive individual is classified positively. Analogously, the pooling specificity

$$\text{PSP}\{H(n_1 : n_2 : \cdots : n_S)\} = 1 - \text{pr}\left(\prod_{s=1}^{S} Z_{\mathcal{P}_{s1}} = 1 \middle| T_1 = 0\right)$$

is the probability a truly negative individual is classified negatively. Values of PSE and PSP close to unity are preferred as this translates to a small percentage of false negative and false positive diagnoses. Simple formulae for PSE and PSP are available under classical assumptions (see, e.g., Kim et al., 2007). For example, $\text{PSE}\{H(n_1 : n_2 : \cdots : n_S)\} = S_e^S$ implies that a larger number of stages decreases pooling sensitivity. Of course, this formula no longer applies in our more general framework.

We derive expressions for $\text{PSE}\{H(n_1 : n_2 : \cdots : n_S)\}$ and $\text{PSP}\{H(n_1 : n_2 : \cdots : n_S)\}$ in terms of $f_{\widetilde{\mathcal{C}}^+}$, $f_{\widetilde{\mathcal{C}}^-}$, and f_ϵ in Appendix B in the Supplementary Material. However, as with the efficiency, these expressions may ultimately be too complicated for practical use. Therefore, simulation details to estimate PSE and PSP for an S-stage hierarchical algorithm are also provided. With these estimates in hand, one can also estimate the *pooling positive predictive value*

$$\text{PPV} = \frac{p\text{PSE}}{p\text{PSE} + (1-p)(1-\text{PSP})}$$

and the *pooling negative predictive value*

$$\text{NPV} = \frac{(1-p)\text{PSP}}{(1-p)\text{PSP} + p(1-\text{PSE})}.$$

These probabilities measure how likely an individual is truly positive (negative) given that the individual has been classified positively (negatively).

3.3. Array Testing

Our simulation methodology can be extended to estimate the operating characteristics of array testing algorithms. In two-dimensional array testing, individuals are first assigned to the cells of an array with R rows and C columns (Phatarfod and Sudbury, 1994; McMahan et al., 2012b). In the first stage, the rows and the columns of the array are tested. The second stage uses individual testing for individuals not classified as negative after the first stage. When the prevalence p is small, two-dimensional array testing can be more efficient than hierarchical algorithms (Kim et al., 2007; Westreich et al., 2008).

We modify our notation from Section 2 to accommodate array testing in two dimensions. Let $T_{r,c}$ denote the true binary status of the individual in the (r,c)th position, and let $\widetilde{\mathcal{C}}_{r,c}$ denote this individual's true biomarker level so that $f_{\widetilde{\mathcal{C}}_{r,c}|T_{r,c}=t}(u) = tf_{\widetilde{\mathcal{C}}^+}(u) + (1-t)f_{\widetilde{\mathcal{C}}^-}(u)$, for $r = 1,2,...,R$ and $c = 1,2,...,C$. The rth row and cth column pools are denoted by $\mathcal{P}_{r+} = \{(r,1),(r,2),...,(r,C)\}$ and $\mathcal{P}_{+c} = \{(1,c),(2,c),...,(R,c)\}$, respectively. Let $\widetilde{\mathcal{C}}_{\mathcal{P}_{r+}} = C^{-1}\sum_{c=1}^{C}\widetilde{\mathcal{C}}_{r,c}$ and $\mathcal{C}_{\mathcal{P}_{r+}}$ denote the true and measured biomarker level of $\mathcal{P}_{r+}$, respectively. Let $\widetilde{\mathcal{C}}_{\mathcal{P}_{+c}} = R^{-1}\sum_{r=1}^{R}\widetilde{\mathcal{C}}_{r,c}$ and $\mathcal{C}_{\mathcal{P}_{+c}}$ be defined analogously for $\mathcal{P}_{+c}$. In the first stage, row and column testing provide $Z_{\mathcal{P}_{r+}} = I(\mathcal{C}_{\mathcal{P}_{r+}} > \tau_{\mathcal{P}_{r+}})$ and $Z_{\mathcal{P}_{+c}} = I(\mathcal{C}_{\mathcal{P}_{+c}} > \tau_{\mathcal{P}_{+c}})$, where $\tau_{\mathcal{P}_{r+}}$ and $\tau_{\mathcal{P}_{+c}}$ are first-stage thresholds (see Section 4) and where $\mathcal{C}_{\mathcal{P}_{r+}}|\widetilde{\mathcal{C}}_{\mathcal{P}_{r+}} \sim f_\epsilon(\cdot|\widetilde{\mathcal{C}}_{\mathcal{P}_{r+}})$ and $\mathcal{C}_{\mathcal{P}_{+c}}|\widetilde{\mathcal{C}}_{\mathcal{P}_{+c}} \sim f_\epsilon(\cdot|\widetilde{\mathcal{C}}_{\mathcal{P}_{+c}})$. We follow the convention in Kim et al. (2007) when identifying which individuals to test in the second stage; i.e., those individuals in

$$\mathcal{M} = \Bigg\{(r,c) : Z_{\mathcal{P}_{r+}} = Z_{\mathcal{P}_{+c}} = 1 \ \text{ or } \ Z_{\mathcal{P}_{r+}} = 1,\ \sum_{c'=1}^{C} Z_{\mathcal{P}_{+c'}} = 0 \ \text{ or } \ \sum_{r'=1}^{R} Z_{\mathcal{P}_{r'+}} = 0,\ Z_{\mathcal{P}_{+c}} = 1\Bigg\}.$$

The event $\{Z_{\mathcal{P}_{r+}} = Z_{\mathcal{P}_{+c}} = 1\}$ occurs at the intersection of the rth row and cth column. The other two events in $\mathcal{M}$ represent ambiguous first-stage outcomes that could arise from testing error. Second-stage testing observes $Z_{r,c} = I(\mathcal{C}_{r,c} > \tau)$ for each individual in $\mathcal{M}$, where $\mathcal{C}_{r,c}|\widetilde{\mathcal{C}}_{r,c} \sim f_\epsilon(\cdot|\widetilde{\mathcal{C}}_{r,c})$ and τ is a threshold

for individual testing. Figure 3 illustrates this notation when $R = C = 5$ (i.e., for a square array). Complete simulation details to estimate the efficiency and accuracy probabilities are provided in Appendix C in the Supplementary Material.

4. Threshold Selection

There are different types of assays used for infectious disease detection, including antibody tests (e.g., ELISA, Western Blot, combination tests which also detect antigens, etc.) and more modern tests which utilize amplification methods. Before an assay is approved for commercial use, it is usually applied to known positive and known negative specimens to determine suitable thresholds for individual testing. Ideally, these thresholds provide high levels of sensitivity and specificity when testing individual specimens. A complete list of screening assays for HIV/HBV/HCV and other infectious agents in the United States is available at `www.fda.com`. An approved assay's product insert typically recommends which threshold should be used to identify positive individuals.

When an assay is applied to pooled specimens, choosing the appropriate threshold can be more subjective. Early work in group testing estimation (see, e.g., Chen and Swallow, 1990; Tu et al., 1994) suggested that individual testing assay thresholds might also be used for pools; see Stephens et al. (2000) and Currie et al. (2004) for specific applications. In the infectious disease pooling literature, a common strategy is to take the individual testing threshold, say τ, and divide it by the number of individuals in the pool; e.g., $\tau_{\mathcal{P}_{11}} = \tau/n_1$ for a master pool in an S-stage hierarchical algorithm $H(n_1 : n_2 : \cdots : n_S)$, $\tau_{\mathcal{P}_{2l}} = \tau/n_2$ for a second-stage pool, and so on. Note that selecting a pooled threshold inappropriately large will decrease the pooling sensitivity, thereby increasing the number of false negative diagnoses. On the other hand, a pooled threshold that is too small will provide far too many false positive pools, thereby weakening the efficiency of group testing.

For individual testing with threshold τ, it is easy to see that the sensitivity

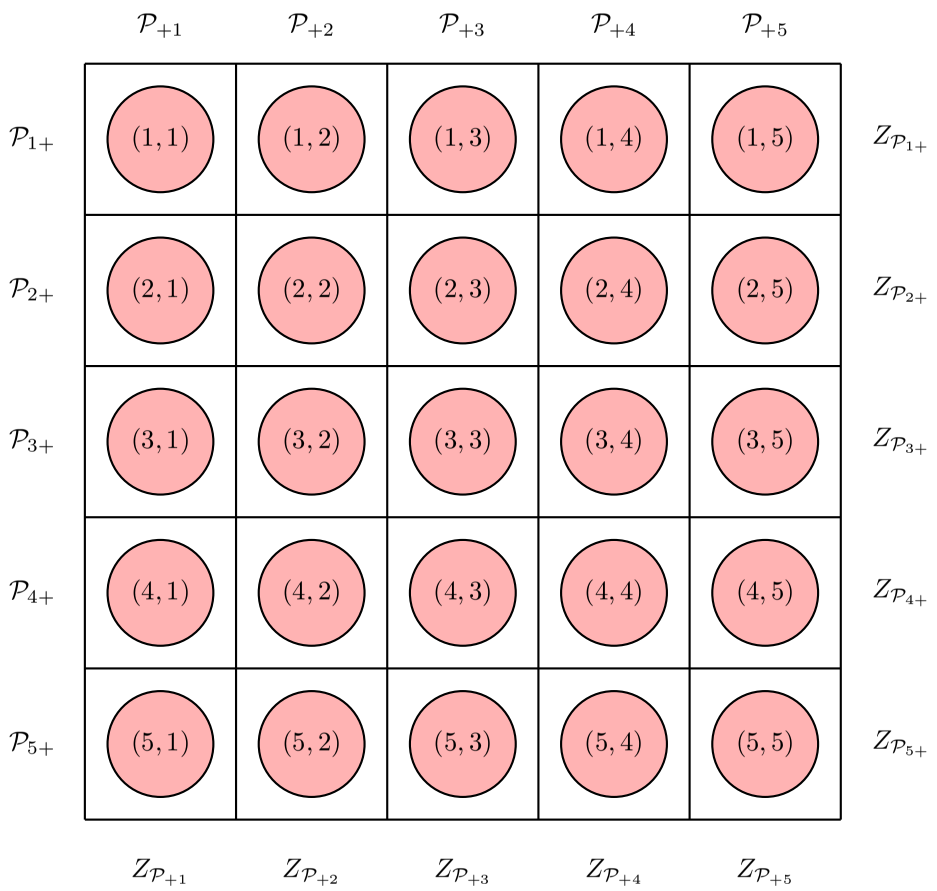

Figure 3: Two-dimensional array testing with $R = C = 5$. First-stage testing results on row pools are $Z_{\mathcal{P}_{1+}}, Z_{\mathcal{P}_{2+}}, ..., Z_{\mathcal{P}_{5+}}$, where $Z_{\mathcal{P}_{r+}} = I(\mathcal{C}_{\mathcal{P}_{r+}} > \tau_{\mathcal{P}_{r+}})$, for $r = 1, 2, ..., 5$. Testing results on column pools are $Z_{\mathcal{P}_{+1}}, Z_{\mathcal{P}_{+2}}, ..., Z_{\mathcal{P}_{+5}}$, where $Z_{\mathcal{P}_{+c}} = I(\mathcal{C}_{\mathcal{P}_{+c}} > \tau_{\mathcal{P}_{+c}})$, for $c = 1, 2, ..., 5$. Second-stage testing is performed on individuals as described in Section 3.3.

is a decreasing function of τ while the specificity is an increasing function of τ. Therefore, one way to choose an individual testing threshold is to maximize Youden's index (Youden, 1950); i.e., $\tau^* = \arg\max_{\tau\in\mathbb{R}}\{S_e(\tau) + S_p(\tau) - 1\}$, as this offers a balance between maximizing both sensitivity and specificity. For a pool generically denoted by $\mathcal{P}$ consisting of individuals whose true disease statuses are denoted by T_i, we propose a pooled threshold that is similar in spirit to Youden's index for individual testing; i.e.,

$$\tau^*_{\mathcal{P}} = \arg\max_{\tau\in\mathbb{R}} \left\{ \int_{\tau}^{\infty} f_{\mathcal{C}_{\mathcal{P}}|\sum_i T_i=1}(u)du + \int_{-\infty}^{\tau} f_{\mathcal{C}_{\mathcal{P}}|\sum_i T_i=0}(u)du - 1 \right\}.$$

The conditional density $f_{\mathcal{C}_{\mathcal{P}}|\sum_i T_i=1}(\cdot)$ describes the distribution of the measured biomarker level of pool $\mathcal{P}$ when there is exactly one positive individual in it. We have selected this density for two reasons. First, in low disease prevalence applications, it is almost always true that a truly positive pool is positive because there is only one positive individual in the pool. Therefore, $\tau^*_{\mathcal{P}}$ will be the appropriate threshold for a large majority of the positive pools. Second, as positive pools could conceivably contain more than one positive individual, $\tau^*_{\mathcal{P}}$ favors the adoption of a smaller-than-necessary threshold. Although this may inflate the efficiency slightly, it also promotes the detection of positive individuals.

5. Applications

We illustrate our simulation methodology using two examples taken from the HIV pooling literature. The first example is from Wein and Zenios (1996) and Zenios and Wein (1998), who consider HIV testing with an antibody assay. The second example from May et al. (2010) is not an classical HIV screening application, but instead describes a virological assay to detect treatment failure among HIV patients. The salient feature of each application is that biomarker distributions for $\mathcal{C}^+$ and $\mathcal{C}^-$ are presented as well as posited distributions for the assay measurement error. We illustrate our biomarker-based evaluations in each application using Dorfman testing, an $S = 3$ stage hierarchical algorithm using halving (Black et al., 2012), and two-dimensional array testing.

For Application 1 (Wein and Zenios, 1996; Zenios and Wein, 1998), the biomarker distributions provided by the authors are $\ln \mathcal{C}^+ \sim \mathcal{N}(0.958, 0.865^2)$, $\mathcal{C}^- \sim I(\mathcal{C}^- = 0.086)$, and the measurement error distribution is $\mathcal{C}_\mathcal{P}|\widetilde{\mathcal{C}}_\mathcal{P} \sim \mathcal{N}\{\widetilde{\mathcal{C}}_\mathcal{P}/(1+\widetilde{\mathcal{C}}_\mathcal{P}), 0.0088\times\widetilde{\mathcal{C}}_\mathcal{P}/(1+\widetilde{\mathcal{C}}_\mathcal{P})^2\}$. The use of a degenerate distribution for negative individuals is described in Zenios and Wein (1998). For this collection of distributions, the threshold that maximizes Youden's index for individual testing is $\tau^* = 0.0485$, which provides values of $S_e > 0.999$ and $S_p > 0.999$; i.e., individual testing is nearly perfect. For Application 2 (May et al., 2010), $\log_{10} \mathcal{C}^+$ is specified to have a two-component mixture $0.93G_1 + 0.07G_2$, where $G_1(G_2)$ is a three-parameter gamma random variable with shape parameter 1.6 (3.2), scale parameter 0.5 (0.5), and location parameter 2.7 (2.7). For negative individuals, $\mathcal{C}^- \sim 0.85U_1 + 0.05U_2 + 0.10U_3$, where $U_1 \sim \mathcal{U}(0, 50)$, $U_2 \sim \mathcal{U}(50, 100)$, and $U_3 \sim \mathcal{U}(100, 500)$, where $\mathcal{U}(a, b)$ denotes a uniform distribution from a to b. The measurement error distribution is specified as $\log_{10} \mathcal{C}_\mathcal{P}|\widetilde{\mathcal{C}}_\mathcal{P} \sim \mathcal{N}(\log_{10} \widetilde{\mathcal{C}}_\mathcal{P}, 0.12^2)$. The threshold that maximizes Youden's index for individual testing in Application 2 is $\tau^* = 436.11$, which provides values of $S_e = 0.989$ and $S_p = 0.980$.

For both applications, we illustrate the differences between our biomarker-based calculations of efficiency, variability, and classification accuracy and the same calculations which rely on classical assumptions (Kim et al., 2007; Westreich et al., 2008); i.e., constant $S_e(S_p)$ and conditional independence of testing responses given the true statuses. In doing so, we consider values of $p \in \{0.01, 0.05, 0.10\}$ while utilizing the three threshold options described in Section 4: τ^* (same for individual testing and pools), τ^* divided by pool size, and our proposed Youden index threshold for pools $\tau^*_\mathcal{P}$. For each combination of p and the threshold used, we calculate the efficiency, the standard deviation of the number of tests per individual, and the four accuracy probabilities in Section 3.2. All of our biomarker-based characteristics are estimated using $B =$ 1,000,000 Monte Carlo data sets. Operating characteristics under classical assumptions are calculated exactly using the expressions in Kim et al. (2007).

Our results when $p = 0.05$ are provided in Table 1; the same tables for $p = 0.01$ and $p = 0.10$ are given in Appendix D in the Supplementary Ma-

terial. In each table, we first determine the most efficient Dorfman algorithm $H(n_1 : 1)$, three-stage halving algorithm $H(n_1 : n_1/2 : 1)$, and square array algorithm $A(n_1 \times n_1)$ under the classical assumptions in Kim et al. (2007) and then compare our biomarker-based evaluations to this optimal setting. Our goal is not to try to outperform the operating characteristics under classical assumptions per se, but instead to illustrate the differences between these calculations and those which exploit underlying biomarker distributions and measurement error, and, more pointedly, how these differences depend on the threshold used. This comparison simultaneously allows one to assess how robust group testing characteristics are under classical assumptions. To the best of knowledge, this is the first assessment of this type in the case identification literature.

From Table 1 and the additional tables in Appendix D, it is clear that the efficiency (EFF), the standard deviation of the number of tests per individual (SD), and the pooling sensitivity (PSE) of group testing are the most heavily influenced by the choice of threshold. One should not be deceived by the ostensibly efficient results that arise when the threshold for individual testing τ^* is also used with pools, as this is also accompanied by a decrease in PSE$-$sharply so in Application 2 where S_e and S_p are lower. On the other hand, dividing τ^* by the pool size leads to a threshold that is too small. This results in too many negative pools testing positively which inflates the efficiency. Our proposed threshold for pools $\tau^*_{\mathcal{P}}$ offers a nice compromise between these two extremes by providing approximately the same efficiency as under classical assumptions. Both applications show that accuracy probabilities under classical assumptions may be slightly optimistic, an important finding for practitioners who are concerned about classification accuracy. This is seen more noticeably in Application 2 where the error rates for individual testing are comparatively larger and also for lower values of p in both applications (e.g., $p = 0.01$, shown in Appendix D).

Table 1: Operating characteristics for Application 1 (Zenios and Wein, 1998) and Application 2 (May et al., 2010) when $p = 0.05$. Efficiency (EFF), standard deviation of the number of tests per individual (SD), and accuracy probabilities (PSE, PSP, PPV, and NPV) are provided. The threshold τ^* maximizes Youden's index for individual testing. The threshold $\tau^*_{\mathcal{P}}$ is calculated as in Section 4. Classical operating characteristics are calculated exactly from Kim et al. (2007). Biomarker-based characteristics are estimated using B = 1,000,000 Monte Carlo data sets. For each application, individual testing values of S_e and S_p are provided. The same tables for $p = 0.01$ and $p = 0.10$ are in Appendix D in the Supplementary Material.

			Biomarker-based evaluations			
			τ^*	τ^*/pool size	$\tau^*_{\mathcal{P}}$	Classical
Application 1 $S_e > 0.999; S_p > 0.999$	$H(5:1)$	EFF (SD)	0.426 (0.418)	0.788 (0.492)	0.426 (0.418)	0.426 (0.418)
		PSE	0.996	>0.999	0.997	>0.999
		PSP	>0.999	>0.999	>0.999	>0.999
		PPV	>0.999	>0.999	>0.999	>0.999
		NPV	>0.999	>0.999	>0.999	>0.999
	$H(8:4:1)$	EFF (SD)	0.391 (0.388)	0.704 (0.450)	0.393 (0.391)	0.395 (0.389)
		PSE	0.986	0.999	0.993	>0.999
		PSP	>0.999	>0.999	>0.999	>0.999
		PPV	>0.999	>0.999	>0.999	>0.999
		NPV	0.999	>0.999	>0.999	>0.999
	$A(9 \times 9)$	EFF (SD)	0.374 (0.115)	0.835 (0.161)	0.378 (0.116)	0.380 (0.117)
		PSE	0.973	0.999	0.986	>0.999
		PSP	>0.999	>0.999	>0.999	>0.999
		PPV	>0.999	>0.999	>0.999	>0.999
		NPV	0.999	>0.999	>0.999	>0.999
Application 2 $S_e = 0.989; S_p = 0.980$	$H(5:1)$	EFF (SD)	0.337 (0.344)	0.588 (0.487)	0.458 (0.437)	0.439 (0.426)
		PSE	0.633	0.987	0.952	0.978
		PSP	0.998	0.983	0.991	0.996
		PPV	0.931	0.756	0.853	0.930
		NPV	0.981	>0.999	0.997	0.999
	$H(8:4:1)$	EFF (SD)	0.258 (0.302)	0.575 (0.424)	0.406 (0.397)	0.396 (0.387)
		PSE	0.513	0.986	0.920	0.967
		PSP	0.999	0.985	0.994	0.997
		PPV	0.947	0.776	0.894	0.948
		NPV	0.975	0.999	0.996	0.998
	$A(10 \times 10)$	EFF (SD)	0.253 (0.053)	0.751 (0.169)	0.394 (0.121)	0.385 (0.120)
		PSE	0.412	0.989	0.892	0.967
		PSP	0.999	0.981	0.994	0.997
		PPV	0.964	0.733	0.889	0.948
		NPV	0.970	0.999	0.994	0.998

6. Discussion

We have proposed a simulation-based methodology to evaluate the operating characteristics of group testing case identification algorithms when individual biomarker distributions are known. Our approach allows the investigator to incorporate the effect of assay measurement error and proposes a new strategy for selecting thresholds when testing pools. Our research web site `www.chrisbilder.com/grouptesting` contains R programs that implement our simulation methods for hierarchical algorithms and two-dimensional array testing with normally distributed biomarkers. These programs can be changed to include other biomarker distributions; e.g., gamma, lognormal, or nonstandard choices like those found in Section 5. In addition, these programs can be modified to include other group testing strategies, such as array testing designs that include master pools (Kim et al., 2007), higher dimensional arrays (Kim and Hudgens, 2009), and other algorithms outside the $H(n_1 : n_2 : \cdots : n_S)$ family described in Section 2.

Our evaluations of case identification algorithms do not require one to assume anything about the sensitivity and specificity of testing pools, because operating characteristics are estimated directly from the biomarker distributions themselves. Our approach also does not force one to assume that testing results are conditionally independent given the true statuses of the individuals being tested. This assumption is required under classical evaluations because pools formed throughout the testing process can contain common individuals. Litvak et al. (1994) have described scenarios where the conditional independence assumption is reasonable empirically; however, there is a large body of evidence in the diagnostic testing literature suggesting that this assumption may be too restrictive. Finally, because the framework described in this article incorporates Monte Carlo simulation, it would be straightforward to generalize our evaluations to allow for unequal disease probabilities p_i, say, which may arise when covariate information is available on individuals (McMahan et al., 2012a; 2012b). For this same reason, our approach could also be extended to ac-

commodate individual disease statuses that are correlated (Lendle et al., 2012) or to applications where biomarkers are measured for multiple diseases (Tebbs et al., 2013).

Throughout this article, we have assumed that the biomarker distributions for positive and negative individuals, $f_{\widetilde{C}^+}$ and $f_{\widetilde{C}^-}$, respectively, and the measurement error density f_ϵ are known exactly. This assumption may be prohibitive in applications where biomarker and measurement error information is not available (e.g., in surveillance studies, etc.). It should be possible to estimate these distributions with continuous group testing responses on pools and individuals, although this would require the development of new deconvolution methods and hence we leave this to future work. In lieu of perfect knowledge about these distributions, an anonymous referee has suggested that one could perform a sensitivity analysis to assess the impact of misspecifying $f_{\widetilde{C}^+}$, $f_{\widetilde{C}^-}$, or f_ϵ. This is straightforward to accomplish within the framework outlined in this article because our methods make use of Monte Carlo simulation. In Appendix E in the Supplementary Material, we provide an example showing how such an analysis could be implemented.

Acknowledgements

We are grateful to two anonymous referees who provided insightful comments and suggestions. We thank Dr. Elizabeth Torrone at the Centers for Disease Control and Prevention for her consultation on infectious disease screening practices in the United States. This research was funded by Grant R01 AI121351 from the National Institutes of Health.

Supplementary Material

Supplementary material related to this article can be found online at [insert address here].

References

Black, M., Bilder, C., Tebbs, J., 2012. Group testing in heterogeneous populations by using halving algorithms. Journal of the Royal Statistical Society: Series C 61, 277-290. DOI: 10.1111/j.1467-9876.2011.01008.x

Busch, M., Caglioti, S., Robertson, E., McAuley, J., Tobler, L., Kamel, H., Linnen, J., Shyamala, V., Tomasulo, P., Kleinman S., 2005. Screening the blood supply for West Nile virus RNA by nucleic acid amplification testing. New England Journal of Medicine 353, 460-467. DOI: 10.1056/NEJMoa044029

Candotti, D. Allain, J., 2009. Transfusion-transmitted hepatitis B virus infection. Journal of Hepatology 51, 798-809. DOI: 10.1016/j.jhep.2009.05.020

Chen, C., Swallow, W., 1990. Using group testing to estimate a proportion, and to test the binomial model. Biometrics 46, 1035-1046. DOI: 10.2307/2532446

Currie, M., McNiven, M., Yee, T., Schiemer, U., Bowden, F., 2004. Pooling of clinical specimens prior to testing for *Chlamydia trachomatis* by PCR is accurate and cost saving. Journal of Clinical Microbiology 42, 4866-4867. DOI: 10.1128/JCM.42.10.4866-4867.2004

Delaigle, A., Hall, P., 2015. Nonparametric methods for group testing data, taking dilution into account. Biometrika 102, 871-887. DOI: 10.1093/biomet/asv049

Dorfman, R., 1943. The detection of defective members of large populations. Annals of Mathematical Statistics 14, 436-440. DOI: 10.1214/aoms/1177731363

Edouard, S., Prudent, E., Gautret, P., Memish, Z., Raoult, D., 2015. Cost-effective pooling of DNA from nasopharyngeal swab samples for large-scale detection of bacteria by real-time PCR. Journal of Clinical Microbiology 52, 1002-1004. DOI: 10.1128/JCM.03609-14

Hudgens, M., Kim, H., 2011. Optimal configuration of a square array group testing algorithm. Communications in Statistics−Theory and Methods 40, 436-448. DOI: 10.1080/03610920903391303

Kim, H., Hudgens, M., 2009. Three-dimensional array-based group testing algorithms. Biometrics 65, 903-910. DOI: 10.1111/j.1541-0420.2008.01158.x

Kim, H., Hudgens, M., Dreyfuss, J., Westreich, D., Pilcher, C., 2007. Comparison of group testing algorithms for case identification in the presence of testing error. Biometrics 63, 1152-1163. DOI: 10.1111/j.1541-0420.2007.00817.x

Krajden, M., Cook, D., Mak, A., Chu, K., Chahil, N., Steinberg, M., Rekart, M., Gilbert, M., 2014. Pooled nucleic acid testing increases the diagnostic yield of acute HIV infections in a high-risk population compared to 3rd and 4th generation HIV enzyme immunoassays. Journal of Clinical Virology 61, 132-137. DOI: 10.1016/j.jcv.2014.06.024

Lendle, S., Hudgens, M., Qaqish, B., 2012. Group testing for case identification with correlated responses. Biometrics 68, 532-540. DOI: 10.1111/j.1541-0420.2011.01674.x

Lewis, J., Lockary, V., Kobic, S., 2012. Cost savings and increased efficiency using a stratified specimen pooling strategy for *Chlamydia trachomatis* and *Neisseria gonorrhoeae*. Sexually Transmitted Diseases 39, 46-48. DOI: 10.1097/OLQ.0b013e318231cd4a

Litvak, E., Tu, X., Pagano, M., 1994. Screening for the presence of a disease by pooling sera samples. Journal of the American Statistical Association 89, 424-434. DOI: 10.1080/01621459.1994.10476764

Malinovsky, Y., Albert, P., Roy, A., 2016. A note on the evaluation of group testing algorithms in the presence of misclassification. Biometrics 72, 299-302. DOI: 10.1111/biom.12385

Malinovsky, Y., Albert, P., Schisterman, E., 2012. Pooling designs for outcomes under a Gaussian random effects model. Biometrics 68, 45-52. DOI: 10.1111/j.1541-0420.2011.01673.x

Martin, E., Salaru, G., Mohammed, D., Coombs, R., Paul, S., Cadoff, E., 2013. Finding those at risk: Acute HIV infection in Newark, NJ. Journal of Clinical Virology 58, 24-28. DOI: 10.1016/j.jcv.2013.07.016

May, S., Gamst, A., Haubrich, R., Benson, C., Smith, D., 2010. Pooled nucleic acid testing to identify antiretroviral treatment failure during HIV infection. Journal of Acquired Immune Deficiency Syndromes 53, 194-201. DOI: 10.1097/QAI.0b013e3181ba37a7

McMahan, C., Tebbs, J., Bilder, C., 2012a. Informative Dorfman screening. Biometrics 68, 287-296. DOI: 10.1111/j.1541-0420.2011.01644.x

McMahan, C., Tebbs, J., Bilder, C., 2012b. Two-dimensional informative array testing. Biometrics 68, 793-804. DOI: 10.1111/j.1541-0420.2011.01726.x

McMahan, C., Tebbs, J., Bilder, C., 2013. Regression models for group testing data with pool dilution effects. Biostatistics 14, 284-298. DOI: 10.1093/biostatistics/kxs045

Mehta, S., Nguyen, V., Osorio, G., Little, S., Smith, D., 2011. Evaluation of pooled rapid HIV antibody screening of patients admitted to a San Diego hospital. Journal of Virological Methods 174, 94-98. DOI: 10.1016/j.jviromet.2011.04.002

Mitchell, E., Lyles, R., Manatunga, A., Danaher, M., Perkins, N., Schisterman, E., 2014. Regression for skewed biomarker outcomes subject to pooling. Biometrics 70, 202-211. DOI: 10.1111/biom.12134

O'Brien, S., Yi, Q., Fan, W., Scalia, V., Fearon, M., and Allain, J. (2012). Current incidence and residual risk of HIV, HBV and HCV at Canadian Blood Services. Vox Sanguinis 103, 83–86.

Page-Shafer, K., Pappalardo, B., Tobler, L., Phelps, B., Edlin, B., Moss, A., Wright, T., Wright, D., O'Brien, T., Caglioti, S., Busch, M., 2008. Testing strategy to identify cases of acute hepatitis C virus (HCV) infection and to project HCV incidence rates. Journal of Clinical Microbiology 46, 499-506. DOI: 10.1128/JCM.01229-07

Phatarfod, R., Sudbury, A., 1994. The use of a square array scheme in blood testing. Statistics in Medicine 13, 2337-2343. DOI: 10.1002/sim.4780132205

Schmidt, M., Pichl, L., Jork, C., Hourfar, M., Schottstedt, V., Wagner, F., Seifried, E., Muller, T., Bux, J., Saldanha J., 2010. Blood donor screening with cobas s 201/cobas TaqScreen MPX under routine conditions at German Red Cross institutes. Vox Sanguinis 98, 37-46. DOI: 10.1111/j.1423-0410.2009.01219.x

Sherlock, M., Zelota, N., Klausner, J., 2007. Routine detection of acute HIV infection through RNA pooling: Survey of current practice in the United States. Sexually Transmitted Diseases 34, 314-316. DOI: 10.1097/01.olq.0000263262.00273.9c

Stephens, G., Raboud, J., Karakas, L., Sherlock, H., 2000. Can pooling be used for seroprevalence studies of hepatitis C? Journal of Clinical Microbiology 38, 4264-4265.

Stramer, S., Krysztof, D., Brodsky, J., Fickett, T., Reynolds, B., Dodd, R., Kleinman S., 2013. Comparative analysis of triplex nucleic acid test assays in United States blood donors. Transfusion 53, 2525-2537. DOI: 10.1111/trf.12178

Tebbs, J., McMahan, C., Bilder, C., 2013. Two-stage hierarchical group testing for multiple infections with application to the Infertility Prevention Project. Biometrics 69, 1064-1073. DOI: 10.1111/biom.12080

Tu, X., Litvak, E., Pagano, M., 1994. Screening tests: Can we get more by doing less? Statistics in Medicine 13, 1905-1919. DOI: 10.1002/sim.4780131904

Wang, D., McMahan, C., Gallagher, C., 2015. A general regression framework for group testing data, which incorporates pool dilution effects. Statistics in Medicine 34, 3606-3621. DOI: 10.1002/sim.6578

Wein, L., Zenios, S., 1996. Pooled testing for HIV screening: Capturing the dilution effect. Operations Research 44, 543-569. DOI: 10.1287/opre.44.4.543

Westreich, D., Hudgens, M., Fiscus, S., Pilcher, C., 2008. Optimizing screening for acute human immunodeficiency virus infection with pooled nucleic acid amplification tests. Journal of Clinical Microbiology 46, 1785-1792. DOI: 10.1128/JCM.00787-07

Youden, W., 1950. Index for rating diagnostic tests. Cancer 3, 32-35. DOI: 10.1002/1097-0142(1950)3:1<32::AID-CNCR2820030106>3.0.CO;2-3

Zenios, S., Wein, L., 1998. Pooled testing for HIV prevalence estimation: Exploiting the dilution effect. Statistics in Medicine 17, 1447-1467. DOI: 10.1002/(SICI)1097-0258(19980715)17:13<1447::AID-SIM862>3.0.CO;2-K

Zhang, Z., Albert, P., 2011. Binary regression analysis with pooled exposure measurements: A regression calibration approach. Biometrics 67, 636-645. DOI: 10.1111/j.1541-0420.2010.01464.x