The Optimal Group Size Controversy for Group Testing: Much Ado About Nothing?

Brianna D. Hitt*
Department of Statistics, University of Nebraska-Lincoln

Christopher R. Bilder
Department of Statistics, University of Nebraska-Lincoln

Joshua M. Tebbs
Department of Statistics, University of South Carolina

Christopher S. McMahan
Department of Mathematical Sciences, Clemson University

August 24, 2018

Abstract

Group testing, the process of testing items as an amalgamated group rather than individually, is used in a wide variety of applications, including human infectious disease screening, virus monitoring of insect carriers, food surveillance, discovery of new pharmaceutical drugs, and quality control of manufactured products. No matter the application, an important decision that needs to be made prior to implementation is determining what group sizes to use. In best practice, an objective function is chosen and then minimized to determine an optimal set of these group sizes, known as the

*Brianna D. Hitt is a PhD Student, Department of Statistics, University of Nebraska-Lincoln, Lincoln, NE 68583 (email: brianna.hitt@huskers.unl.edu); Christopher R. Bilder is a Professor, Department of Statistics, University of Nebraska-Lincoln, Lincoln, NE 68583 (email: chris@chrisbilder.com); Joshua M. Tebbs is a Professor, Department of Statistics, University of South Carolina, Columbia, SC 29208 (email: tebbs@stat.sc.edu); and Christopher S. McMahan is an Associate Professor, Department of Mathematical Sciences, Clemson University, Clemson, SC, 29634 (email: mcmaha2@clemson.edu). This research was supported by Grant R01AI121351 from the National Institutes of Health. The authors thank their colleagues at the Centers for Disease Control and Prevention, Innovative Blood Resources, the Nebraska Public Health Laboratory, the Nebraska Veterinary Diagnostic Center, and the State Hygienic Laboratory at the University of Iowa for their discussions and collaboration to make infectious disease testing more efficient.

optimal testing configuration (OTC). There are a few options for objective functions, and they differ based on how the expected number of tests, assay characteristics, and testing constraints are taken into account. These varied options have led to a recent controversy in the literature regarding which objective function is best. In our paper, we examine the most commonly proposed objective functions. We show that this controversy may be "much ado about nothing" because the OTCs and corresponding results (e.g., number of tests, accuracy) are largely the same for standard testing algorithms in a wide variety of situations. Supplemental materials for this article are available online.

Keywords: Binary response; Pooled testing; Screening; Sensitivity; Specificity

1. Introduction

Laboratories throughout the world test high volumes of clinical specimens for infectious diseases, including HIV, hepatitis C, and West Nile virus. In such situations, it has become standard practice to test amalgamations of specimens as a "group" or "pool" rather than to test individual specimens. The reason is simple: members of a negatively testing group can be declared negative all at once. Thus, for a group of size I, say, just one test is needed to declare all members negative, rather than the I separate tests that would be needed with individual testing. Fortunately, when disease prevalence is small, the majority of groups will test negatively when sensibly chosen group sizes are used. For members of a positive testing group, there are many algorithmic retesting procedures available to determine which specific individuals are positive. The first retesting procedure was proposed by Dorfman (1943) and simply involved individually retesting each member of a positive group. Since this seminal work, group testing has been used to efficiently test for infectious diseases in a vast number of human applications, including blood donation screening (American Red Cross 2018), antiretroviral treatment failure detection for HIV-positive individuals (Kim et al. 2014; Tilghman et al. 2015), chlamydia and gonorrhea testing for US government-sponsored STD assessment and prevention programs (Centers for Disease Control and Prevention 2012), and influenza outbreak surveillance (Hourfar et al. 2007). Outside of infectious disease testing in humans, group testing is used in an extensive number of applications, including cow milk surveillance (Græsbøll et al. 2017), disease detection in cattle and buffaloes (Abdellrazeq et al. 2014), West Nile virus monitoring in mosquitoes (Khan et al. 2017), food contamination detection (Pasquali et al. 2014), drug discovery (Kainkaryam and Woolf 2009), and diagnosis of faulty network sensors (Lo et al. 2013). Due to this wide variety of applications and somewhat diverse terminology within them, we will use the language associated with infectious disease testing throughout our paper for the ease of exposition.

For all group testing applications, the choice of group sizes is extremely important for success. Choosing group sizes too large will lead to exceedingly many groups testing positively. This will subsequently lead to a large number of retests, perhaps even a larger

number of tests overall than what would be needed for individual testing. Similarly, choosing group sizes too small will lead to a larger number of tests than would otherwise be needed if the group sizes were chosen better. In best practice, laboratories choose group sizes by minimizing an objective function that takes into account the group testing algorithm to be implemented. There are a number of different algorithms in use, and they are best characterized as being either hierarchical or non-hierarchical in nature. Hierarchical algorithms begin by testing individuals in non-overlapping groups. For a group that tests positively, subsequent retesting stages occur in smaller, non-overlapping groups. The previously described Dorfman algorithm is a two-stage algorithm. Three- and four-stage algorithms are commonly used in practice (e.g., Quinn et al. 2000; Sherlock et al. 2007) because they are often more efficient (i.e., fewer tests). Non-hierarchical algorithms involve testing each individual in overlapping groups to reduce the number of retests. The most common type of non-hierarchical algorithm is known as array-based (matrix-based) testing (Phatarfod and Sudbury 1994; Kim et al. 2007). For this algorithm, individual specimens are arranged in a two-dimensional grid. These specimens are amalgamated by row and by column and then tested. Intersecting positive rows and columns indicate where retesting should be performed to determine which individuals are positive. For a thorough review of hierarchical and array-based algorithms, see Hughes-Oliver (2006).

While there are many different types of group testing algorithms, all laboratories are interested in minimizing the number of tests needed to assay their specimens. For this reason, objective functions are based on the expected number of tests, so that a set of group sizes for a testing algorithm, known as the optimal testing configuration (OTC), can be found by minimizing this function. Traditionally, group testing research has focused on objective functions expressed solely as the expected number of tests per individual. This is due to a close correspondence between the number of tests and testing costs. However, using an objective function that contains only the expected number of tests leaves out an important component of infectious disease testing: accuracy. Infectious disease testing is rarely perfect. Errors can occur for reasons such as improper laboratory implementation or a specimen being collected during the window period between disease contraction and the ability to detect it. Fortunately, known mathematical expressions are available for

the accuracy of most group testing algorithms. This enables laboratories to calculate the expected accuracy of a chosen testing configuration prior to implementation.

Malinovsky et al. (2016) recently proposed a new objective function that includes the expected number of tests and a measurement of accuracy. This allows laboratories to evaluate accuracy at the same time as the number of tests when choosing an OTC. As may be expected when breaking with tradition, the proposal generated controversy in the group testing research literature. Both Hudgens (2016) and McMahan et al. (2016) offered rejoinders to Malinovsky et al. (2016) that disagreed with this new objective function. All three of these works focused only on the Dorfman algorithm in their limited evaluations. The purpose of our paper is to examine a significant number of other group testing algorithms with respect to objective functions. This is important because other algorithms are widely used and known to result in a smaller number of tests and/or higher accuracy than the Dorfman algorithm. We present findings in our paper that interestingly show both the traditional and the new objective function are actually quite similar and very often lead to the same OTC.

The order of this paper follows. Section 2 explicitly defines the objective functions and provides a mathematical comparison between them. Section 3 calculates the OTC for each objective function along with their operating characteristics (expected number of tests and accuracy measures) in a wide variety of settings. These calculations are performed for both hierarchical and array-based group testing algorithms. We show under what conditions these operating characteristics will be the same and when they will be different. Section 4 summarizes our findings, discusses alternative objective functions, and provides recommendations for practice. We also provide R functions to find the OTCs and to reproduce our work.

2. Objective Functions

Define T as the total number of tests for an overall group of size I with a hierarchical algorithm. When using the traditional objective function, the OTC is found by minimizing

the expected number of tests per individual:

$$O_{ET} = E(T)/I.$$

For example, the expected number of tests for three-stage hierarchical testing is given by

$$E(T) = 1 + m_{11}P(G_{11} = 1) + \sum_{j=1}^{c_2} m_{2j}P(G_{11} = 1, G_{2j} = 1),$$

where G_{sj} is the binary outcome (values of 0 and 1 indicate a negative and a positive test result, respectively) for group j at stage s, m_{sj} is the number of subgroups that would be created if group j at stage s tests positively, and c_s is the number of groups at stage s. The probabilities $P(G_{11} = 1)$ and $P(G_{11} = 1, G_{2j} = 1)$ are both functions of the number of groups and their respective sizes, the probability of positivity for each individual, and the sensitivity S_e and specificity S_p of the assay. General expressions for $E(T)$ are available from Kim et al. (2007) for the case of each individual having the same probability of positivity p and from Black et al. (2015) for the case of each individual potentially having a different probability of positivity p_i for $i = 1, \ldots, I$. The latter case is known as informative group testing (Bilder et al. 2010; Lewis et al. 2012; Bilder and Tebbs 2012), because p_i can be estimated with the help of disease-risk information that may be available for each individual tested. We will refer to the former case then as non-informative group testing in our work here. Expressions for the expected number of tests are known for array-based algorithms (Kim et al. 2007; McMahan et al. 2012b) as well, where O_{ET} is still defined as the expected number of tests per individual.

While O_{ET} is the most commonly utilized objective function, it does not directly take into account the accuracy of the algorithm. When using O_{ET}, one usually examines the accuracy of an OTC separately through measures such as the overall sensitivity and specificity of an algorithm, most often known as the pooling sensitivity and pooling specificity, respectively. Each of these accuracy measures is a function of the group sizes used in the testing algorithm and the probabilities of positivity.

Rather than examining accuracy measures after obtaining the OTC, Malinovsky et al. (2016) proposed an alternative objective function that simultaneously takes into account

accuracy and the expected number of tests. Define C as the number of correct classifications for an overall group of size I. The OTC is found by minimizing

$$O_{MAR} = E(T)/E(C).$$

Because C is never larger than the number of individuals I, $E(C) \leq I$. By comparing O_{MAR} and O_{ET}, we see that

$$O_{ET} = \frac{E(T)}{I} \leq \frac{E(T)}{E(C)} = O_{MAR}$$

for the same initial group size I. In fact, O_{MAR} and O_{ET} will be quite close in value. This is because infectious disease testing algorithms will only be put into use if they have high accuracy. Thus, $E(C)$ will be quite close to I in practice.

To examine this closeness more precisely, consider minimizing the logarithm of each objective function:

$$\log(O_{ET}) = \log\{E(T)\} - \log(I)$$

and

$$\log(O_{MAR}) = \log\{E(T)\} - \log\{E(C)\}. \tag{2.1}$$

As shown in the Supplementary Material on the publisher's website, the expected number of correct classifications is

$$E(C) = \sum_{i=1}^{I} \{PS_{p,i}(1-p_i) + PS_{e,i}p_i\}, \tag{2.2}$$

where $PS_{p,i}$ and $PS_{e,i}$ are the pooling specificity and pooling sensitivity, respectively, for individual i. For hierarchical testing, the pooling sensitivity is always the same for every individual tested in the same number of stages (Kim et al. 2007; Black et al. 2015). The pooling specificity is the same for every individual as well, but only for non-informative group testing with equal group sizes within a stage. Under this scenario then, we can simplify the expression for the expected number of correct classifications to be

$$E(C) = I\{PS_p(1-p) + PS_e p\}, \tag{2.3}$$

where PS_p and PS_e are the pooling specificity and sensitivity, respectively, but now equal for each individual. For array testing, the same simplification for $E(C)$ from Equation (2.2) to Equation (2.3) occurs when the number of rows and the number of columns are the same (i.e., a square array), which is how array testing is usually applied.

By substituting Equation (2.3) into Equation (2.1), we obtain

$$\begin{aligned} \log(O_{MAR}) &= \log\{E(T)\} - \log\left[I\left\{PS_p(1-p) + PS_e p\right\}\right] \\ &= \log(O_{ET}) - \log\left\{PS_p(1-p) + PS_e p\right\}. \end{aligned}$$

Thus, any difference between the OTCs for the two objective functions is due to the "penalty" of

$$\log\left\{PS_p(1-p) + PS_e p\right\}. \tag{2.4}$$

Unfortunately, further definitive statements cannot be made regarding Equation (2.4), and we are left with making general statements regarding what will happen most often. In particular, we see that the penalty places a large weight on PS_p in comparison to PS_e because p is small for realistic group testing applications. Also, because PS_p and PS_e tend to be close to 1 for realistic applications, the penalty tends to be close to 0. Thus, $\log(O_{MAR})$ will most often be close to $\log\{E(T)\}$.

3. Comparisons

Because definitive statements are not possible for Equation (2.4) or for the more general cases of unequal group sizes and informative group testing, we provide in this section a thorough investigation of the OTCs when using the objective functions over a very large number of situations. For each of these situations, we calculate the OTCs along with corresponding operating characteristics. Our results for both non-informative and informative group testing algorithms are described next.

3.1. Non-informative group testing

We include in this investigation the following group testing algorithms: two-stage hierarchical, three-stage hierarchical, array testing without a master pool (row and column groups are tested first, as described in Section 1), and array testing with a master pool (all specimens in the array are tested together in one group before any row or column groups are formed). For the first three algorithms, we allow the initial group sizes to range from $I = 3, ..., 40$, but allow higher initial group sizes when the overall prevalence is very small (e.g., $p = 0.005$) so that the OTC does not include our arbitrary upper bound for I. For array testing with a master pool, we use the same range of group sizes for the row and column groups, leading to a maximum master pool size of I^2. All array testing algorithms use square arrays, and we account for potential testing ambiguities that can occur in arrays (e.g., a row tests positively without any columns testing positively) by the methods described in Kim et al. (2007). We apply these group testing algorithms over thirty different values of p ranging from 0.005 to 0.150 by 0.005 and over three separate sets of accuracy levels (low: $S_e = S_p = 0.90$, medium: $S_e = S_p = 0.95$, and high: $S_e = S_p = 0.99$).

Table 1 displays the results for $p = 0.01$. The OTCs are the same for both objective functions when using the hierarchical algorithms. Some small differences between OTCs exist for the array testing algorithms, but the differences are not of practical importance. For example, examine the results for array testing without master pooling and $S_e = S_p = 0.90$. The expected number of tests and the pooling sensitivities are the same to four decimal places. The pooling specificities are also quite close. In practical terms, for a testing volume of 100,000 individuals, there would be 98,267 correct negatives found when using the OTC for O_{ET} and 98,307 correct negatives found when using the OTC for O_{MAR}. While 40 additional false positives would result from the OTC for O_{ET}, these false positives would most likely be discovered from follow-up confirmatory testing that normally would occur. We also provide similar tables for $p = 0.05$ and $p = 0.10$ in the Supplementary Material available on the publisher's website. These tables show no differences among the OTCs when using either O_{ET} or O_{MAR}.

Table 2 summarizes the largest differences among the operating characteristics across all thirty different values of p included in our investigation. Most often, the OTCs found

Table 1: OTC summary for $p = 0.01$ under non-informative group testing.

Algorithm	S_e	S_p	Objective function	OTC	$E(T)/I$	PS_e	PS_p
2-stage hierarchical	0.99	0.99	O_{ET}	11-1	0.2035	0.9801	0.9990
			O_{MAR}	11-1	0.2035	0.9801	0.9990
	0.95	0.95	O_{ET}	11-1	0.2351	0.9025	0.9932
			O_{MAR}	11-1	0.2351	0.9025	0.9932
	0.90	0.90	O_{ET}	12-1	0.2742	0.8100	0.9816
			O_{MAR}	12-1	0.2742	0.8100	0.9816
3-stage hierarchical	0.99	0.99	O_{ET}	25-5-1	0.1354	0.9703	0.9996
			O_{MAR}	25-5-1	0.1354	0.9703	0.9996
	0.95	0.95	O_{ET}	24-6-1	0.1443	0.8574	0.9973
			O_{MAR}	24-6-1	0.1443	0.8574	0.9973
	0.90	0.90	O_{ET}	24-6-1	0.1562	0.7290	0.9938
			O_{MAR}	24-6-1	0.1562	0.7290	0.9938
Array w/o master pooling	0.99	0.99	O_{ET}	25-1	0.1378	0.9703	0.9995
			O_{MAR}	25-1	0.1378	0.9703	0.9995
	0.95	0.95	O_{ET}	25-1	0.1475	0.8575	0.9970
			O_{MAR}	24-1	0.1475	0.8575	0.9972
	0.90	0.90	O_{ET}	25-1	0.1611	0.7291	0.9926
			O_{MAR}	24-1	0.1611	0.7291	0.9930
Array w/ master pooling	0.99	0.99	O_{ET}	625-25-1	0.1364	0.9606	0.9995
			O_{MAR}	625-25-1	0.1364	0.9606	0.9995
	0.95	0.95	O_{ET}	625-25-1	0.1402	0.8146	0.9972
			O_{MAR}	576-24-1	0.1402	0.8146	0.9974
	0.90	0.90	O_{ET}	625-25-1	0.1450	0.6562	0.9934
			O_{MAR}	576-24-1	0.1450	0.6562	0.9937

NOTE: Equally sized groups are optimal at each stage; thus, an OTC of "24-6-1" means that stage 1 has a group of size 24, stage 2 has four groups of size 6, and stage 3 has twenty-four groups of size 1. Differences between O_{ET} and O_{MAR} are highlighted.

Table 2: Largest differences between operating characteristics for OTCs under non-informative group testing.

Algorithm	S_e	S_p	Frequency	Largest difference $E(T)/I$	Largest difference PS_e	Largest difference PS_p
2-stage hierarchical	0.99	0.99	0	-	-	-
	0.95	0.95	3	0.0018	0.0000	0.0049
	0.90	0.90	4	0.0023	0.0000	0.0054
3-stage hierarchical	0.99	0.99	0	-	-	-
	0.95	0.95	1	0.0014	0.0000	0.0051
	0.90	0.90	3	0.0015	0.0000	0.0049
Array w/o master pooling	0.99	0.99	0	-	-	-
	0.95	0.95	5	0.0010	0.0018	0.0026
	0.90	0.90	8	0.0028	0.0022	0.0054
Array w/ master pooling	0.99	0.99	2	0.0005	0.0006	0.0008
	0.95	0.95	4	0.0012	0.0017	0.0026
	0.90	0.90	8	0.0015	0.0018	0.0051

NOTE: Values of p range from 0.005 to 0.150 by 0.005. The frequency column denotes the number of times a different OTC was found for O_{ET} and O_{MAR} among these values of p. Differences between operating characteristics are rounded to four decimal places. Note that operating characteristics are always smaller for O_{ET} than for O_{MAR} when differences exist.

are the same for the two objective functions. When differences exist, these differences occur more often for smaller values of S_e and S_p, but again are not of practical importance. Overall, these findings help confirm what was strongly suspected in Section 2 through our mathematical analysis. Namely, the objective functions lead to the same OTCs or OTCs with similar operating characteristics when differences exist.

3.2. Informative group testing

We include in this investigation the following group testing algorithms: two-stage hierarchical implemented via the pool-specific optimal Dorfman (PSOD) method (McMahan et al. 2012a), three-stage hierarchical (Black et al. 2015), and array testing without a master pool implemented via the gradient method (McMahan et al. 2012b). For the PSOD method, we use a block size of 50 and replace its greedy optimization algorithm with examination of all possible testing configurations. Array testing with a master pool is not included in our investigations because there have been no informative group testing algorithms proposed

for it. We continue to allow the initial group sizes to range from $I = 3, ..., 40$ and allow for higher initial group sizes when the overall prevalence is very small.

To provide different levels of heterogeneity among the p_i for $i = 1, \ldots, I$, we use the expected value of order statistics from $P_i \sim$ beta $\{\alpha, \alpha(1-p)/p\}$ for $i = 1, \ldots, I$ in the same manner as in Black et al. (2015). This beta distribution has $E(P_i) = p$, and we once again consider values of p ranging from 0.005 to 0.150 by 0.005. The amount of heterogeneity is controlled by α, where lower levels indicate a larger amount of heterogeneity (see Black et al. 2015 for further discussion regarding the choice of α).

Table 3 displays the results for $E(P_i) = 0.01$, and the Supplementary Material available on the publisher's website provides the results for $E(P_i) = 0.05$ and $E(P_i) = 0.10$. The displayed pooling sensitivity, PS_e^W, and pooling specificity, PS_p^W, are weighted averages of individual pooling sensitivities and pooling specificities, respectively, for all individuals within the initial group for a hierarchical algorithm or within the entire array for an array-based algorithm. Expressions for these averages are provided in the Supplementary Material on the publisher's website and are based on accuracy definitions given by Altman and Bland (1994). The largest differences for each operating characteristic across all values of p are given in Table 4. Overall, while differences exist more often for some algorithms than in the non-informative group testing setting, O_{ET} and O_{MAR} still result in the same or very similar OTCs the majority of the time, and, when differences exist, the differences likely would not be of practical importance due to similar operating characteristic values.

4. Conclusion

We have shown that the choice of objective function most often does not change the OTC, and even when the OTC is different, there are not practical differences in the operating characteristics. Therefore, our work helps to close the case on the recent controversy regarding objective functions: they both can be used in practice. However, we tend to favor the traditionally used O_{ET} for one main reason. Simply, laboratories need to know the number of tests to be expected and the corresponding costs involved. In many instances, the expected costs are directly proportional to the expected number of tests. While the expected number of tests could also be stated when using O_{MAR}, this seems to be an

Table 3: OTC summary for $E(P_i) = 0.01$ under informative group testing.

Algorithm	α	S_e	S_p	Objective function	Initial group size for OTC	$E(T)/I$	PS_e^W	PS_p^W
2-stage hierarchical	2	0.99	0.99	O_{ET}	-	0.1947	0.9801	0.9991
				O_{MAR}	-	0.1947	0.9801	0.9991
		0.95	0.95	O_{ET}	-	0.2264	0.9025	0.9931
				O_{MAR}	-	0.2264	0.9025	0.9931
		0.90	0.90	O_{ET}	-	0.2657	0.8100	0.9822
				O_{MAR}	-	0.2657	0.8100	0.9822
	0.5	0.99	0.99	O_{ET}	-	0.1683	0.9801	0.9992
				O_{MAR}	-	0.1683	0.9801	0.9992
		0.95	0.95	O_{ET}	-	0.2019	0.9025	0.9943
				O_{MAR}	-	0.2019	0.9025	0.9943
		0.90	0.90	O_{ET}	-	0.2439	0.8100	0.9843
				O_{MAR}	-	0.2439	0.8100	0.9843
3-stage hierarchical	2	0.99	0.99	O_{ET}	26	0.1285	0.9703	0.9996
				O_{MAR}	26	0.1285	0.9703	0.9996
		0.95	0.95	O_{ET}	26	0.1375	0.8574	0.9974
				O_{MAR}	26	0.1375	0.8574	0.9974
		0.90	0.90	O_{ET}	26	0.1497	0.7290	0.9939
				O_{MAR}	26	0.1497	0.7290	0.9939
	0.5	0.99	0.99	O_{ET}	33	0.1197	0.9703	0.9996
				O_{MAR}	33	0.1197	0.9703	0.9996
		0.95	0.95	O_{ET}	28	0.1291	0.8574	0.9977
				O_{MAR}	28	0.1291	0.8574	0.9977
		0.90	0.90	O_{ET}	29	0.1422	0.7290	0.9942
				O_{MAR}	29	0.1422	0.7290	0.9942
Array w/o master pooling	2	0.99	0.99	O_{ET}	25	0.1349	0.9703	0.9995
				O_{MAR}	25	0.1349	0.9703	0.9995
		0.95	0.95	O_{ET}	25	0.1448	0.8575	0.9972
				O_{MAR}	25	0.1448	0.8575	0.9972
		0.90	0.90	O_{ET}	25	0.1585	0.7291	0.9929
				O_{MAR}	25	0.1585	0.7291	0.9929
	0.5	0.99	0.99	O_{ET}	28	0.1277	0.9703	0.9995
				O_{MAR}	28	0.1277	0.9703	0.9995
		0.95	0.95	O_{ET}	28	0.1379	0.8574	0.9971
				O_{MAR}	27	0.1379	0.8574	0.9972
		0.90	0.90	O_{ET}	28	0.1519	0.7290	0.9927
				O_{MAR}	27	0.1519	0.7290	0.9930

NOTE: Multiple initial group sizes for 2-stage hierarchical algorithms are found within a block size of 50, so they are not displayed here. The full OTCs are provided in the Supplementary Material available on the publisher's website. Differences between O_{ET} and O_{MAR} are highlighted.

Table 4: Largest differences between operating characteristics for OTCs under informative group testing.

Algorithm	α	S_e	S_p	Frequency	Largest difference $E(T)/I$	Largest difference PS_e^W	Largest difference PS_p^W
2-stage hierarchical	2	0.99	0.99	0	-	-	-
		0.95	0.95	7	0.0006	(0.0023)	0.0011
		0.90	0.90	12	0.0010	(0.0052)	0.0023
	0.5	0.99	0.99	0	-	-	-
		0.95	0.95	3	0.0003	(0.0035)	0.0011
		0.90	0.90	15	0.0008	(0.0103)	0.0022
3-stage hierarchical	2	0.99	0.99	1	0.0000	(0.0019)	0.0002
		0.95	0.95	2	0.0035	0.0219	0.0033
		0.90	0.90	6	0.0044	0.0152	0.0062
	0.5	0.99	0.99	1	0.0000	0.0001	0.0001
		0.95	0.95	0	-	-	-
		0.90	0.90	3	0.0010	0.0250	0.0033
Array w/o master pooling	2	0.99	0.99	1	0.0003	0.0004	0.0005
		0.95	0.95	2	0.0011	0.0012	0.0027
		0.90	0.90	5	0.0016	0.0012	0.0040
	0.5	0.99	0.99	0	-	-	-
		0.95	0.95	4	0.0003	0.0004	0.0015
		0.90	0.90	14	0.0015	0.0004	0.0032

NOTE: Values of $E(P_i) = p$ range from 0.005 to 0.150 by 0.005. The frequency column denotes the number of times a different OTC was found among these values of p. Differences between operating characteristics are rounded to four decimal places. Note that the operating characteristic value for O_{ET} is always subtracted from the operating characteristic value for O_{MAR}. Thus, a negative value (indicated with parentheses) means that the value for O_{ET} was larger than the value for O_{MAR}.

unnecessary extra step, especially for laboratory directors and technicians who choose the OTC. For these users and also for those performing research in the area, we make available a set of R functions in the `binGroup` package that can be used to find the OTC with O_{ET} or O_{MAR}. Examples of how to use these functions are available on our research website at www.chrisbilder.com/grouptesting and in the Supplementary Material for this paper on the publisher's website.

Throughout this paper, we had to make the assumption that p or p_i for $i = 1, \ldots, I$ is known. Of course, this would not be known in actual practice. Instead, some type of past experience would be used by laboratories to estimate these quantities so that an "estimated" OTC could be chosen. These estimated OTCs still would be the same or very similar for the two objective functions because the same estimates for probabilities of positivity would be used with each function. Furthermore, even when there would be small differences, these differences would have less meaning in practice due to the true probabilities being unknown.

There are other objective functions that could be used. For example, Malinovsky et al. (2016) considered maximizing $E(C/T)$, but concluded this to be inferior to O_{MAR}. Therefore, we focused only on their O_{MAR} proposal in our paper. One could also use the objective function proposed by Graff and Roeloffs (1972). This function involves a linear combination of the expected number of tests, the number of misclassified negatives, and the number of misclassified positives. Subjectively chosen weights (or penalties) can be used with the misclassification measures to increase or decrease their importance. Of course, there will be weights then that result in an OTC different from using O_{ET} and O_{MAR}. The subjectiveness of these weights can depend on the infectious disease, the laboratory, or even particular individuals at a laboratory. For this reason, we do not examine this particular objective function in our paper.

Supplementary Material

The supplementary material contains the derivation for $E(C)$ (Equation 2.2), additional results for Sections 3.1 and 3.2, and an explanation of R functions available to reproduce our results.

References

Abdellrazeq, G., El-Naggar, M., Khallel, S., and Gamal-Eldin, A. (2014). Detection of Mycobacterium avium subsp. paratuberculosis from cattle and buffaloes in Egypt using traditional culture, serological and molecular based methods. *Veterinary World*, 7(8):586–593.

Altman, D. and Bland, J. (1994). Diagnostic tests 1: sensitivity and specificity. *BMJ*, 308:1552.

American Red Cross (2018). Infectious disease testing. https://www.redcrossblood.org/biomedical-services/blood-diagnostic-testing/blood-testing.html. Retrieved August 14, 2018.

Bilder, C. R. and Tebbs, J. M. (2012). Pooled-testing procedures for screening high volume clinical specimens in heterogeneous populations. *Statistics in Medicine*, 31(27):3261–3268.

Bilder, C. R., Tebbs, J. M., and Chen, P. (2010). Informative retesting. *Journal of the American Statistical Association*, 105(491):942–955.

Black, M. S., Bilder, C. R., and Tebbs, J. M. (2015). Optimal retesting configurations for hierarchical group testing. *Journal of the Royal Statistical Society. Series C: Applied Statistics*, 64(4):693–710.

Centers for Disease Control and Prevention (2012). STD-related reproductive health, prevention, training, and technical assistance centers. https://www.cdc.gov/std/stdrhpttac/default.htm. Retrieved August 14, 2018.

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

Græsbøll, K., Andresen, L. O., Halasa, T., and Toft, N. (2017). Opportunities and challenges when pooling milk samples using ELISA. *Preventive Veterinary Medicine*, 139:93–98.

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

Hourfar, M. K., Themann, A., Eickmann, M., Puthavathana, P., Laue, T., Seifried, E., and Schmidt, M. (2007). Blood screening for influenza. *Emerging Infectious Diseases*, 13(7):1081–1083.

Hudgens, M. G. (2016). Rejoinder to 'Reader reaction: A note on the evaluation of group testing algorithms in the presence of misclassification'. *Biometrics*, 72(1):304.

Hughes-Oliver, J. M. (2006). Pooling experiments for blood screening and drug discovery. In Dean, A. and Lewis, S., editors, *Screening: Methods for Experimentation in Industry, Drug Discovery, and Genetics*, pages 48–68. Springer, New York, NY.

Kainkaryam, R. M. and Woolf, P. J. (2009). Pooling in high-throughput drug screening. *Current Opinion in Drug Discovery and Development*, 12(3):339–350.

Khan, S. A., Chowdhury, P., Choudhury, P., and Dutta, P. (2017). Detection of West Nile virus in six mosquito species in synchrony with seroconversion among sentinel chickens in India. *Parasites & Vectors*, 10(1):13.

Kim, H. Y., 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(4):1152–1163.

Kim, S. B., Kim, H. W., Kim, H.-S., Ann, H. W., Kim, J. K., Choi, H., Kim, M. H., Song, J. E., Ahn, J. Y., Ku, N. S., Oh, D. H., Kim, Y. C., Jeong, S. J., Han, S. H., Kim, J. M., Smith, D. M., and Choi, J. Y. (2014). Pooled nucleic acid testing to identify antiretroviral treatment failure during HIV infection in Seoul, South Korea. *Scandinavian Journal of Infectious Diseases*, 46(2):136–40.

Lewis, J. L., Lockary, V. M., and Kobic, S. (2012). Cost savings and increased efficiency using a stratified specimen pooling strategy for Chlamydia trachomatis and Neisseria gonorrhoeae. *Sexually Transmitted Diseases*, 39(1):46–48.

Lo, C., Bai, Y., Liu, M., and Lynch, J. P. (2013). Efficient sensor fault detection using combinatorial group testing. In *2013 IEEE International Conference on Distributed Computing in Sensor Systems*, pages 199–206, Cambridge, MA.

Malinovsky, Y., Albert, P. S., and Roy, A. (2016). Reader reaction: A note on the evaluation of group testing algorithms in the presence of misclassification. *Biometrics*, 72(1):299–302.

McMahan, C. S., Tebbs, J. M., and Bilder, C. R. (2012a). Informative Dorfman screening. *Biometrics*, 68(1):287–296.

McMahan, C. S., Tebbs, J. M., and Bilder, C. R. (2012b). Two-dimensional informative array testing. *Biometrics*, 68(3):793–804.

McMahan, C. S., Tebbs, J. M., and Bilder, C. R. (2016). Rejoinder to 'Reader reaction: A note on the evaluation of group testing algorithms in the presence of misclassification'. *Biometrics*, 72(1):303–304.

Pasquali, F., De Cesare, A., Valero, A., Olsen, J. E., and Manfreda, G. (2014). Improvement of sampling plans for Salmonella detection in pooled table eggs by use of real-time PCR. *International Journal of Food Microbiology*, 184:31–34.

Phatarfod, R. M. and Sudbury, A. (1994). The use of a square array scheme in blood testing. *Statistics in Medicine*, 13(22):2337–2343.

Quinn, T. C., Brookmeyer, R., Kline, R., Shepherd, M., Paranjape, R., Mehendale, S., Gadkari, D. A., and Bollinger, R. (2000). Feasibility of pooling sera for HIV-1 viral RNA to diagnose acute primary HIV-1 infection and estimate HIV incidence. *AIDS*, 14(17):2751–2757.

Sherlock, M., Zetola, N., and 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.

Tilghman, M., Tsai, D., Buene, T. P., Tomas, M., Amade, S., Gehlbach, D., Chang, S., Ignacio, C., Caballero, G., Espitia, S., May, S., Noormahomed, E. V., and Smith, D. M.

(2015). Pooled nucleic acid testing to detect antiretroviral treatment failure in HIV-infected patients in Mozambique. *Journal of Acquired Immune Deficiency Syndromes*, 70(3):256–61.