Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

ARTICLE TYPE

The Objective Function Controversy for Group Testing: Much Ado About Nothing?

Brianna D. Hitt*[1] | Christopher R. Bilder[1] | Joshua M. Tebbs[2] | Christopher S. McMahan[3]

[1]Department of Statistics, University of Nebraska-Lincoln, Lincoln, Nebraska
[2]Department of Statistics, University of South Carolina, Columbia, South Carolina
[3]School of Mathematical and Statistical Sciences, Clemson University, Clemson, South Carolina

Correspondence
*Brianna D. Hitt, Hardin Hall North 340, Lincoln, NE 68583.
Email: brianna.hitt@huskers.unl.edu

Funding information
National Institutes of Health, Grant/Award Number: R01 AI121351

Abstract

Group testing is an indispensable tool for laboratories when testing high volumes of clinical specimens for infectious diseases. 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 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 of two different objective functions is better. In our paper, we examine these objective functions over a number of realistic situations for infectious disease testing. 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.

KEYWORDS:
Binary response; Infectious disease; 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 negative 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[1] 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,[2] antiretroviral treatment failure detection for HIV-positive individuals,[3,4] chlamydia and gonorrhea testing,[5] and influenza outbreak surveillance.[6] Outside of infectious disease testing in humans, group testing is used in an extensive number of applications, including cow milk surveillance,[7] disease detection in cattle and buffaloes,[8] West Nile virus monitoring in mosquitoes,[9] food contamination detection,[10] drug discovery,[11] and diagnosis of faulty network sensors.[12]

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 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[13,14] 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 testing.[15,16] 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 testing algorithms, see Hughes-Oliver[17] and Bilder.[18]

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[19] 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[20] and McMahan et al[21] offered rejoinders to Malinovsky et al[19] 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 in realistic infectious disease testing situations.

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 testing algorithms. We show under what conditions these operating characteristics will be the same and when they will be different. Section 4 examines the objective function controversy in the context of actual assays used for infectious disease detection. To conclude, Section 5 summarizes our findings, discusses alternative objective functions, and provides recommendations for practice. We also discuss R functions that we provide with our paper to find the OTCs and to reproduce our work.

2. Objective Functions

Define T as a random variable representing 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 random variable (values of 1 and 0 indicate a positive and a negative test result, respectively) representing the outcome 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 (see Black et al[22]; an example diagram is given in the Supporting Information available online to further explain the notation). 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 being positive for each individual, and the sensitivity S_e and specificity S_p of the assay each time it is applied. We do not provide further detailed expressions for $E(T)$ here to avoid distraction from the main points of our paper and because expressions are already provided elsewhere. For example, Kim et al[16] provides expressions for the case of each individual having the same true probability of being positive, say p, and Black et al[22] provides expressions for the case of each individual potentially having a different probability of being truly positive, say p_i for $i = 1, \ldots, I$. The latter case is known as informative group testing,[23,24,25] 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 testing algorithms[16,26] 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. However, one will still examine separately the accuracy of the OTC to judge if it is satisfactory. As an alternative approach, Malinovsky et al[19] proposed an objective function that simultaneously takes into account accuracy and the expected number of tests. To examine the accuracy aspect, define Y_i as the final positive/negative (1/0) outcome based on the group testing algorithm, and define $\tilde{Y}_i$ as the true positive/negative (1/0) status of individual i, for $i = 1, \ldots, I$. Commonly used accuracy measures for a group testing algorithm as a whole are the pooling sensitivity $PS_{e,i} = P(Y_i = 1|\tilde{Y}_i = 1)$ and the pooling specificity $PS_{p,i} = P(Y_i = 0|\tilde{Y}_i = 0)$ for individual i. As an overall measure of accuracy, define C as the number of correct classifications for a group of size I. The expected number of correct classifications is

$$\begin{aligned} E(C) &= \sum_{i=1}^{I} \left\{ P(Y_i = 0, \tilde{Y}_i = 0) + P(Y_i = 1, \tilde{Y}_i = 1) \right\} \\ &= \sum_{i=1}^{I} \left\{ PS_{p,i}(1 - p_i) + PS_{e,i} p_i \right\}, \end{aligned} \tag{1}$$

where $P(\tilde{Y}_i = 1) = p_i$ is the probability that individual i is truly positive.

Malinovsky et al[19] proposed to find the OTC by maximizing the expected number of correct classifications per individual divided by the expected number of tests per individual. Equivalently, this results in 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 assays 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

$$\log(O_{MAR}) = \log\{E(T)\} - \log\{E(C)\}. \tag{2}$$

For hierarchical testing, the pooling sensitivity is always the same for every individual tested in the same number of stages.[16,22] 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

$$E(C) = I \left\{ PS_p(1 - p) + PS_e p \right\}, \tag{3}$$

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 (1) to Equation (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 (3) into Equation (2), 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

$$\log\left\{PS_p(1-p)+PS_ep\right\}. \tag{4}$$

Unfortunately, further definitive statements cannot be made regarding Equation (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 $\log(O_{ET})$.

3. Comparisons

Because definitive statements are not possible for Equation (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.[16] We apply these group testing algorithms over thirty different values of p ranging from 0.005 to 0.150 by 0.005 and over five separate sets of accuracy levels (S_e and S_p values range from 0.90 to 0.99). These values of p, S_e, and S_p are chosen because they correspond to when group testing is used for infectious disease testing. The assay accuracies are assumed to not change based on group size, meaning that the assays have been properly tested and calibrated for group testing.

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 load 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 Supporting Information available on the publisher's website. These tables show only one case with differences between the OTCs.

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 are the same for the two objective functions. When differences exist, these differences occur more often for smaller values of 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,[27] three-stage hierarchical,[22] and array testing without a master pool implemented via the gradient method.[26] 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 $P_i \sim \text{beta}\{\alpha, \alpha(1-p)/p\}$ for $i = 1, \ldots, I$ in the same manner as in Black et al.[22] 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[22] for further discussion regarding the choice of α).

Table 3 displays the results for $E(P_i) = 0.01$, and the Supporting Information 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 testing algorithm. Expressions for these averages are provided in the Supporting Information on the publisher's website and are based on accuracy definitions given by Altman and Bland.[28] 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 vast majority of the differences likely would not be of practical importance due to similar operating characteristic values.

For three-stage hierarchical, the maximum difference in PS_e for some settings, such as $S_e = 0.90$ and $S_p = 0.99$, may be somewhat concerning at a first examination. Further investigation revealed that this occurred when the OTC for O_{MAR} had more sub-groups in the second stage of testing with a size of 1 than did the OTC for O_{ET}. This is important because 1) a third stage of testing is unnecessary for those individuals with a sub-group size of 1 in the second stage of testing; 2) pooling sensitivity for each individual is S_e^L, where L is the number of stages that the individual is tested within[22]; and 3) PS_e^W is a weighted average of each individual's pooling sensitivity. Especially when p is large for three-stage hierarchical testing, the initial group size can be quite small, so each individual's pooling sensitivity plays a larger role in the weighted average. Thus, while there are some differences in the weighted averages of the pooling sensitivities, it is due to those few individuals who are not tested in the third stage. The individuals tested in the same number of stages still have the same pooling sensitivity values.

4. Applications

We present two different applications comparing the OTCs obtained from using O_{ET} or O_{MAR} for infectious disease testing. To find the OTC for these and other applications, the value of p or p_i for $i = 1, \ldots, I$ is needed. Of course, these quantities would most likely be unknown. Instead, some type of past experience would be used by laboratories to estimate these quantities so that an "estimated" OTC could be chosen. Also to find the OTC, the values of S_e and S_p are needed because $E(T)$ and the pooling sensitivities/specificities depend upon them. Laboratories can obtain these values from a number of sources, including internal validations, research articles, product inserts for assays, and summaries provided by organizations such as the Centers for Disease Control and Prevention (CDC) and the Association of Public Health Laboratories. For each source, the sensitivity and specificity are actually observed through taking a large sample. For instance, a set of known positive specimens may be tested to evaluate the sensitivity of an assay. Alternatively, clinical-based evaluations may be performed by applying the assay in practice and using other means to validate true positive/negative statuses. The observed sensitivities and specificities usually are treated as constants and sometimes confidence intervals are stated along with them. Our purpose in this section is not to evaluate these procedures but rather use the accuracy measures as they are in practice to determine OTCs.

Group testing is used widely for HIV testing in applications including blood donation screening[2] and health surveillance via public health clinics[14]. Branson et al.[29] provided the CDC's recommendations for HIV testing by laboratories. To make these recommendations, the authors examined over 30 research articles and product inserts, and they included the sensitivities and specificities associated with each assay examined. Observed sensitivities ranged from 96.3% to 100%, and observed specificities ranged from 99.03% to 100%. For our investigation here, we use the lowest values in these ranges to find the OTC. Our reason for using these particular values is because differences between OTCs would most likely occur with the lowest accuracies. Table 5 provides the OTCs from non-informative group testing algorithms. For these calculations, we use an overall HIV prevalence of $p = 0.004$ based on CDC estimates of HIV[30] and Census Bureau estimates of population[31] in the United States from 2016. Overall, the table shows that O_{ET} and O_{MAR} lead to the same OTCs for all group testing algorithms considered. While the OTCs for array testing with master pooling are the same for O_{ET} and O_{MAR}, a master pool with a 44×44 array may be too large

to use in practice (the largest group size that we have seen used for HIV testing is 128 [32]). A laboratory may need to choose a sub-optimal array size for such a situation.

Group testing is used widely for chlamydia testing as well. High volumes of clinical specimens are tested each year in this manner by public health laboratories across the United States as part of statewide surveillance projects (e.g., see Lewis et al. [24] and Bilder et al. [33]). Black et al. [34] examined the testing performed by the Nebraska Public Health Laboratory (NPHL) with the BD ProbeTec ET CT/GC Amplified DNA Assay. A main purpose of this paper was to evaluate how well an informative group testing algorithm could perform in comparison to their current implementation of individual testing. For our purpose here, we use the observed data from the urine specimen testing in 2009 to examine OTCs over a number of group testing algorithms. The overall observed chlamydia prevalence was 0.080 for females and 0.081 for males. We use these observed prevalences as our values for p when performing non-informative group testing by gender. To implement informative group testing, we used the beta distribution fits given by Black et al. [22] for the individual probabilities of being positive p_i and implement methods similar to those in Section 3.2. We limit our maximum group sizes to be 20 due to large group sizes not being used in chlamydia testing [35]. The NPHL provided assay sensitivities of $S_e = 0.805$ and $S_e = 0.93$ and specificities of $S_p = 0.96$ and $S_p = 0.95$ for females and males, respectively. This assay had an unusually low sensitivity for female urine specimens, and the laboratory eventually switched after that year to the Aptima Combo 2 Assay which has a much higher sensitivity ($S_e = 0.947$) [36]. However, to be consistent with how the actual tests were performed, we use the accuracies for the BD assay. Table 6 provides the OTCs for non-informative and informative group testing algorithms. Overall, the table shows that O_{ET} and O_{MAR} lead to the same OTCs for all non-informative group testing algorithms considered. While differences do exist for females when using informative hierarchical testing algorithms, these small differences likely would not be of practical importance.

5. Conclusion

We have shown that the choice between O_{ET} and O_{MAR} 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: both can be used in practice because they lead to very similar results. Some individuals may prefer to state that they used O_{MAR} because it directly takes into account accuracy at the beginning of the process. 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 unnecessary extra step, especially for laboratory directors and technicians who choose the OTC.

It is important to emphasize that laboratories would not use O_{ET} without still looking at accuracy. Rather than incorporating accuracy within the objective function, they would find the OTC and then examine the accuracy associated with it. If the accuracy resulting from O_{ET} (or O_{MAR}) was unsatisfactory, a new sub-optimal testing configuration would be chosen with accuracies that are acceptable. To help laboratories and those performing research in this area, we make available a set of R functions in the `binGroup` package that can be used to find the OTC or other suitable testing configurations by using O_{ET} and O_{MAR}. Examples of how to use these functions are available on our research website at www.chrisbilder.com/grouptesting and in the Supporting Information for this paper on the publisher's website.

Our evaluations of O_{ET} and O_{MAR} focus on realistic settings for infectious disease detection when group testing would be used. Thus, we focus on values of S_e and S_p close to 1 and small values of p. When smaller values of S_e and S_p and/or larger values of p are used, there can be differences in the OTCs and associated accuracy measures. For example, when $S_e = 0.75$, $S_p = 0.80$, and $p = 0.10$ for three-stage hierarchical testing, the OTC for O_{ET} has $I = 15$ and second-stage group sizes of 5 for each sub-group. For these same settings, the OTC for O_{MAR} has $I = 12$ and second-stage group sizes of 4 for each sub-group. However, the pooling sensitivity is only $PS_e = 0.42$ for both testing configurations, which makes the use of group testing unrealistic for this situation.

Laboratories may need to limit the particular values of I for which the OTC is searched over, similar to what we did in Section 4 for the chlamydia testing example. This may be due to physical constraints, such as a maximum group size that can be incorporated into an automated pooling platform. Also, this limit may be due to what is known as the "dilution effect" in group testing. Because specimens are pooled together, each individual specimen becomes a smaller part of the whole as the group size increases. This reduced portion can make it more difficult for an assay to identify its target, which in turn lowers its sensitivity. Laboratories may need to place an upper limit on I in this type of situation. Properly calibrated tests are needed whenever group

testing is used to make sure the dilution effect does not become a problem. Fortunately, the dilution effect is now much less likely to occur due to modern nucleic acid amplification testing methods.

There are other objective functions that could be used. For example, Malinovsky et al [19] 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. Objective functions can include penalties for making classification errors. For example, Graff and Roeloffs [37] proposed using an objective function that is a linear combination of the expected number of tests, the number of misclassified negatives, and the number of misclassified positives. Subjectively chosen weights are used with the misclassification measures to increase or decrease their importance. As would be expected, there will be weights then that result in an OTC which is quite different than what would be obtained from using O_{ET} and O_{MAR}. We provide examples in the Supporting Information illustrating these differences. However, the subjectiveness of these weights can depend on the infectious disease, the laboratory, or even particular individuals at a laboratory. Therefore, for general applications and research settings, it is difficult to use this or similar types of objective functions. We say this by no means to diminish the importance of taking into account the misclassification type. Because of its importance for specific applications, we provide tools in our `binGroup` package to find the OTC in those situations when this type of control is necessary. Examples are provided again on our research website and in the Supporting Information.

ACKNOWLEDGEMENTS

This research was supported by Grant R01 AI121351 from the National Institutes of Health. The authors thank their colleagues at the CDC, 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. The authors also thank the referees for helping to improve this paper.

References

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

2. American Red Cross . Infectious disease testing. 2019. https://www.redcrossblood.org/biomedical-services/blood-diagnostic-testing/blood-testing.html. Accessed May 26, 2019.

3. Kim SB, Kim HW, Kim HS, et al. Pooled nucleic acid testing to identify antiretroviral treatment failure during HIV infection in Seoul, South Korea. *Scandinavian Journal of Infectious Diseases* 2014; 46(2): 136–40.

4. Tilghman M, Tsai D, Buene TP, et al. Pooled nucleic acid testing to detect antiretroviral treatment failure in HIV-infected patients in Mozambique. *Journal of Acquired Immune Deficiency Syndromes* 2015; 70(3): 256–61.

5. Papp JR, Schachter J, Gaydos CA, Van Der Pol B. Recommendations for the laboratory-based detection of Chlamydia trachomatis and Neisseria gonorrhoeae-2014. Tech. Rep. RR-02, Centers for Disease Control and Prevention; Atlanta, GA: 2014.

6. Hourfar MK, Themann A, Eickmann M, et al. Blood screening for influenza. *Emerging Infectious Diseases* 2007; 13(7): 1081–1083.

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

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

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

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

11. Kainkaryam RM, Woolf PJ. Pooling in high-throughput drug screening. *Current Opinion in Drug Discovery and Development* 2009; 12(3): 339–350.

12. Lo C, Liu M, Lynch JP, Gilbert AC. Efficient sensor fault detection using combinatorial group testing. *2013 IEEE International Conference on Distributed Computing in Sensor Systems (DCOSS)* 2013: 199–206. IEEE.

13. Quinn TC, Brookmeyer R, Kline R, et al. Feasibility of pooling sera for HIV-1 viral RNA to diagnose acute primary HIV-1 infection and estimate HIV incidence. *AIDS* 2000; 14(17): 2751–2757.

14. Sherlock M, Zetola N, Klausner J. Routine detection of acute HIV infection through RNA pooling: survey of current practice in the United States. *Sexually Transmitted Diseases* 2007; 34: 314–316.

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

16. Kim HY, Hudgens MG, Dreyfuss JM, Westreich DJ, Pilcher CD. Comparison of group testing algorithms for case identification in the presence of test error. *Biometrics* 2007; 63(4): 1152–1163.

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

18. Bilder CR. Group testing for identification. *Wiley StatsRef: Statistics Reference Online*: 1–11. https://doi.org/10.1002/9781118445112.stat08227.

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

20. Hudgens MG. Rejoinder to 'Reader reaction: a note on the evaluation of group testing algorithms in the presence of misclassification'. *Biometrics* 2016; 72(1): 304.

21. McMahan CS, Tebbs JM, Bilder CR. Rejoinder to 'Reader reaction: a note on the evaluation of group testing algorithms in the presence of misclassification'. *Biometrics* 2016; 72(1): 303–304.

22. Black MS, Bilder CR, Tebbs JM. Optimal retesting configurations for hierarchical group testing. *Journal of the Royal Statistical Society. Series C: Applied Statistics* 2015; 64(4): 693–710.

23. Bilder CR, Tebbs JM, Chen P. Informative retesting. *Journal of the American Statistical Association* 2010; 105(491): 942–955.

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

25. Bilder CR, Tebbs JM. Pooled-testing procedures for screening high volume clinical specimens in heterogeneous populations. *Statistics in Medicine* 2012; 31(27): 3261–3268.

26. McMahan CS, Tebbs JM, Bilder CR. Two-dimensional informative array testing. *Biometrics* 2012; 68(3): 793–804.

27. McMahan CS, Tebbs JM, Bilder CR. Informative Dorfman screening. *Biometrics* 2012; 68(1): 287–296.

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

29. Branson BM, Owen SM, Wesolowski LG, et al. Laboratory testing for the diagnosis of HIV infection: updated recommendations. 2014. http://stacks.cdc.gov/view/cdc/23447, Accessed May 26, 2019.

30. Centers for Disease Control and Prevention. Estimated HIV incidence and prevalence in the United States, 2010-2016. HIV Surveillance Supplemental Report 2019; 24 (No. 1). 2019. http://www.cdc.gov/hiv/library/reports/hiv-surveillance.html. Published February 2019. Accessed May 26, 2019.

31. U.S. Census Bureau, Population Division. Annual Estimates of the Resident Population: April 1, 2010 to July 1, 2018. 2018. https://www.census.gov/data/tables/time-series/demo/popest/2010s-national-total.html. Released December 2018. Accessed May 26, 2019.

32. Sullivan TJ, Patel P, Hutchinson A, Ethridge SF, Parker MM. Evaluation of pooling strategies for acute HIV-1 infection screening using nucleic acid amplification testing. *Journal of Clinical Microbiology* 2011; 49(10): 3667–3668.

33. Bilder CR, Tebbs JM, McMahan CS. Informative group testing for multiplex assays. *Biometrics* 2019; 75(1): 278–288.

34. Black MS, Bilder CR, Tebbs JM. Group testing in heterogeneous populations by using halving algorithms. *Journal of the Royal Statistical Society. Series C: Applied Statistics* 2012; 61(2): 277–290.

35. Mund M, Sander G, Potthoff P, Schicht H, Matthias K. Introduction of Chlamydia trachomatis screening for young women in Germany. *Journal der Deutschen Dermatologischen Gesellschaft* 2008; 6(12): 1032–1037.

36. Food and Drug Administration . Gen-Probe® Aptima Combo 2® Assay Package Insert. 2018. https://www.fda.gov/media/74033/download. Accessed May 26, 2019.

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

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

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

Algorithm	S_e	S_p	Objective function	OTC	$E(T)/I$	PS_e	PS_p
Two-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
	0.99	0.90	O_{ET}	11-1	0.2841	0.9801	0.9815
			O_{MAR}	11-1	0.2841	0.9801	0.9815
	0.90	0.99	O_{ET}	11-1	0.1941	0.8100	0.9990
			O_{MAR}	11-1	0.1941	0.8100	0.9990
Three-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
	0.99	0.90	O_{ET}	24-6-1	0.1708	0.9703	0.9928
			O_{MAR}	24-6-1	0.1708	0.9703	0.9928
	0.90	0.99	O_{ET}	25-5-1	0.1229	0.7290	0.9997
			O_{MAR}	25-5-1	0.1229	0.7290	0.9997
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
	0.99	0.90	O_{ET}	23-1	0.1726	0.9703	0.9923
			O_{MAR}	23-1	0.1726	0.9703	0.9923
	0.90	0.99	O_{ET}	27-1	0.1279	0.7292	0.9995
			O_{MAR}	27-1	0.1279	0.7292	0.9995
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
	0.99	0.90	O_{ET}	529-23-1	0.1708	0.9606	0.9924
			O_{MAR}	529-23-1	0.1708	0.9606	0.9924
	0.90	0.99	O_{ET}	729-27-1	0.1151	0.6563	0.9996
			O_{MAR}	729-27-1	0.1151	0.6563	0.9996

TABLE 2 Largest differences between operating characteristics for OTCs under non-informative group testing. 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 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}.

				Largest difference		
Algorithm	S_e	S_p	Frequency	$E(T)/I$	PS_e	PS_p
Two-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
	0.99	0.90	7	0.0056	0.0000	0.0096
	0.90	0.99	0	-	-	-
Three-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
	0.99	0.90	7	0.0041	(0.0098)	0.0136
	0.90	0.99	1	0.0000	0.0000	0.0002
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
	0.99	0.90	5	0.0043	0.0005	0.0076
	0.90	0.99	1	0.0000	0.0006	0.0001
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
	0.99	0.90	5	0.0048	0.0005	0.0077
	0.90	0.99	2	0.0003	0.0026	0.0005

TABLE 3 OTC summary for $E(P_i) = 0.01$ under informative group testing. Multiple initial group sizes for two-stage hierarchical algorithms are found within a block size of 50, so they are not displayed here. The full OTCs are provided in the Supporting Information available on the publisher's website. Differences between O_{ET} and O_{MAR} are highlighted.

				$\alpha = 2$				$\alpha = 0.5$			
Algorithm	S_e	S_p	Objective function	Initial group size for OTC	$E(T)/I$	PS_e^W	PS_p^W	Initial group size for OTC	$E(T)/I$	PS_e^W	PS_p^W
Two-stage hierarchical	0.99	0.99	O_{ET}	-	0.1947	0.9801	0.9991	-	0.1683	0.9801	0.9992
			O_{MAR}	-	0.1947	0.9801	0.9991	-	0.1683	0.9801	0.9992
	0.95	0.95	O_{ET}	-	0.2264	0.9025	0.9931	-	0.2019	0.9025	0.9943
			O_{MAR}	-	0.2264	0.9025	0.9931	-	0.2019	0.9025	0.9943
	0.90	0.90	O_{ET}	-	0.2657	0.8100	0.9822	-	0.2439	0.8100	0.9843
			O_{MAR}	-	0.2657	0.8100	0.9822	-	0.2439	0.8100	0.9843
	0.99	0.90	O_{ET}	-	0.2754	0.9801	0.9813	-	0.2511	0.9801	0.9837
			O_{MAR}	-	0.2754	0.9801	0.9813	-	0.2511	0.9801	0.9837
	0.90	0.99	O_{ET}	-	0.1854	0.8100	0.9990	-	0.1611	0.8100	0.9993
			O_{MAR}	-	0.1854	0.8100	0.9990	-	0.1611	0.8100	0.9993
Three-stage hierarchical	0.99	0.99	O_{ET}	26	0.1285	0.9703	0.9996	33	0.1197	0.9703	0.9996
			O_{MAR}	26	0.1285	0.9703	0.9996	33	0.1197	0.9703	0.9996
	0.95	0.95	O_{ET}	26	0.1375	0.8574	0.9974	28	0.1291	0.8574	0.9977
			O_{MAR}	26	0.1375	0.8574	0.9974	28	0.1291	0.8574	0.9977
	0.90	0.90	O_{ET}	26	0.1497	0.7290	0.9939	29	0.1422	0.7290	0.9942
			O_{MAR}	26	0.1497	0.7290	0.9939	29	0.1422	0.7290	0.9942
	0.99	0.90	O_{ET}	26	0.1638	0.9703	0.9930	28	0.1554	0.9703	0.9935
			O_{MAR}	26	0.1638	0.9703	0.9930	28	0.1554	0.9703	0.9935
	0.90	0.99	O_{ET}	26	0.1168	0.7290	0.9997	37	0.1078	0.7290	0.9997
			O_{MAR}	26	0.1168	0.7290	0.9997	37	0.1078	0.7290	0.9997
Array w/o master pooling	0.99	0.99	O_{ET}	25	0.1349	0.9703	0.9995	28	0.1277	0.9703	0.9995
			O_{MAR}	25	0.1349	0.9703	0.9995	28	0.1277	0.9703	0.9995
	0.95	0.95	O_{ET}	25	0.1448	0.8575	0.9972	28	0.1379	0.8574	0.9971
			O_{MAR}	25	0.1448	0.8575	0.9972	27	0.1379	0.8574	0.9972
	0.90	0.90	O_{ET}	25	0.1585	0.7291	0.9929	28	0.1519	0.7290	0.9927
			O_{MAR}	25	0.1585	0.7291	0.9929	27	0.1519	0.7290	0.9930
	0.99	0.90	O_{ET}	23	0.1699	0.9703	0.9926	25	0.1631	0.9703	0.9926
			O_{MAR}	23	0.1699	0.9703	0.9926	25	0.1631	0.9703	0.9926
	0.90	0.99	O_{ET}	27	0.1251	0.7293	0.9996	30	0.1181	0.7291	0.9996
			O_{MAR}	27	0.1251	0.7293	0.9996	30	0.1181	0.7291	0.9996

TABLE 4 Largest differences between operating characteristics for OTCs under informative group testing. 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}.

					Largest difference		
Algorithm	α	S_e	S_p	Frequency	$E(T)/I$	PS_e^W	PS_p^W
Two-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
		0.99	0.90	12	0.0011	(0.0008)	0.0022
		0.90	0.99	2	0.0003	0.0052	0.0000
	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
		0.99	0.90	16	0.0012	(0.0011)	0.0022
		0.90	0.99	11	0.0006	0.0078	(0.0002)
Three-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
		0.99	0.90	4	0.0035	0.0006	0.0066
		0.90	0.99	14	0.0180	0.0500	0.0003
	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
		0.99	0.90	5	0.0022	0.0034	0.0070
		0.90	0.99	9	0.0057	0.0355	0.0003
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
		0.99	0.90	4	0.0028	0.0003	0.0053
		0.90	0.99	0	-	-	-
	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
		0.99	0.90	8	0.0024	0.0001	0.0041
		0.90	0.99	1	0.0003	0.0005	0.0003

TABLE 5 OTC summary for HIV testing using $p = 0.004$, $S_e = 0.963$, and $S_p = 0.9903$, with non-informative group testing. 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. There are no differences between the OTCs.

Algorithm	Objective function	OTC	$E(T)/I$	PS_e	PS_p
Two-stage hierarchical	O_{ET}	17-1	0.1313	0.9274	0.9993
	O_{MAR}	17-1	0.1313	0.9274	0.9993
Three-stage hierarchical	O_{ET}	49-7-1	0.0732	0.8931	0.9998
	O_{MAR}	49-7-1	0.0732	0.8931	0.9998
Array w/o master pooling	O_{ET}	44-1	0.0749	0.8931	0.9997
	O_{MAR}	44-1	0.0749	0.8931	0.9997
Array w/ master pooling	O_{ET}	1936-44-1	0.0721	0.8600	0.9998
	O_{MAR}	1936-44-1	0.0721	0.8600	0.9998

TABLE 6 OTC summary for chlamydia testing. The test accuracies are $S_e = 0.805$ and $S_p = 0.96$ for females and $S_e = 0.93$ and $S_p = 0.95$ for males. For non-informative group testing, $p = 0.08$ for females, $p = 0.081$ for males, and equally sized groups are optimal at each stage (see Table 1's caption for how group sizes are denoted). For informative group testing, $P_i \sim \text{beta}(1.1, 11.54)$ for females, $P_i \sim \text{beta}(1.8, 18.20)$ for males, and full OTCs are provided in the Supporting Information available on the publisher's website. Results are not displayed for informative array testing with master pooling because no group testing algorithms have been proposed for it. Differences between O_{ET} and O_{MAR} are highlighted.

	Algorithm	Objective function	Non-informative				Informative			
			OTC	$E(T)/I$	PS_e	PS_p	Stage 1 size	$E(T)/I$	PS_e^W	PS_p^W
Female	Two-stage hierarchical	O_{ET}	5-1	0.5008	0.6480	0.9897	-	0.4757	0.6480	0.9901
		O_{MAR}	5-1	0.5008	0.6480	0.9897	-	0.4761	0.6480	0.9910
	Three-stage hierarchical	O_{ET}	12-4-1	0.4099	0.5217	0.9937	19	0.4102	0.5217	0.9930
		O_{MAR}	12-4-1	0.4099	0.5217	0.9937	14	0.4113	0.5479	0.9933
	Array w/o master pooling	O_{ET}	9-1	0.4327	0.5240	0.9931	10	0.4187	0.5226	0.9929
		O_{MAR}	9-1	0.4327	0.5240	0.9931	10	0.4187	0.5226	0.9929
	Array w/ master pooling	O_{ET}	81-9-1	0.3485	0.4218	0.9945	-	-	-	-
		O_{MAR}	81-9-1	0.3485	0.4218	0.9945	-	-	-	-
Male	Two-stage hierarchical	O_{ET}	4-1	0.5523	0.8649	0.9877	-	0.5462	0.8649	0.9867
		O_{MAR}	4-1	0.5523	0.8649	0.9877	-	0.5462	0.8649	0.9867
	Three-stage hierarchical	O_{ET}	9-3-1	0.4931	0.8044	0.9924	8	0.5081	0.8206	0.9905
		O_{MAR}	9-3-1	0.4931	0.8044	0.9924	8	0.5081	0.8206	0.9905
	Array w/o master pooling	O_{ET}	8-1	0.5015	0.8056	0.9901	8	0.5105	0.8053	0.9900
		O_{MAR}	8-1	0.5015	0.8056	0.9901	8	0.5105	0.8053	0.9900
	Array w/ master pooling	O_{ET}	64-8-1	0.4667	0.7492	0.9908	-	-	-	-
		O_{MAR}	64-8-1	0.4667	0.7492	0.9908	-	-	-	-