HHS Public Access

Author manuscript

Am Stat. Author manuscript; available in PMC 2019 December 06.

Published in final edited form as:

Am Stat. 2019 ; 73(2): 117–125. doi:10.1080/00031305.2017.1366367.

Revisiting Nested Group Testing Procedures: New Results, Comparisons, and Robustness

Yaakov Malinovsky[a], **Paul S. Albert**[b]

[a]Department of Mathematics and Statistics University of Maryland, Baltimore County, Baltimore, MD

[b]Biostatistics Branch, Division of Cancer Epidemiology and Genetics National Cancer Institute, Rockville, MD

Abstract

Group testing has its origin in the identification of syphilis in the U.S. army during World War II. Much of the theoretical framework of group testing was developed starting in the late 1950s, with continued work into the 1990s. Recently, with the advent of new laboratory and genetic technologies, there has been an increasing interest in group testing designs for cost saving purposes. In this article, we compare different nested designs, including Dorfman, Sterrett and an optimal nested procedure obtained through dynamic programming. To elucidate these comparisons, we develop closed-form expressions for the optimal Sterrett procedure and provide a concise review of the prior literature for other commonly used procedures. We consider designs where the prevalence of disease is known as well as investigate the robustness of these procedures, when it is incorrectly assumed. This article provides a technical presentation that will be of interest to researchers as well as from a pedagogical perspective. Supplementary material for this article available online.

Keywords

Coding theory; Information theory; Optimal design

1. Introduction

Commonly, individual samples are assessed for the presence of a condition in order to identify disease status. Group testing is concerned with finding efficient algorithms to test groups of individual samples that provide these identifications with a minimum number of tests. Group testing (GT) procedures are cost- and time-saving identification procedures that have broad applications to blood screening for HIV, hepatitis, and other infectious diseases (Gastwirth and Johnson 1994; Bilder, Tebbs, and Chen 2010; Stramer et al. 2011; Tebbs, McMahan, and Bilder 2013; Bar-Lev et al. 2017), quality control in product testing (Sobel and Groll 1959; Bar-Lev, Boneh, and Perry 1990), veterinary medicine (Graaesbøll et al.

CONTACT Yaakov Malinovsky, yaakovm@umbc.edu, Department of Mathematics and Statistics, University of Maryland, Baltimore County, Baltimore, MD 21250, USA.

Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/TAS.

2016), drug discovery (Zhu, Hughes-Oliver, and Young 2001), DNA screening (Du and Hwang 2006; Cao and Sun 2016), communication and security networks (Wolf 1985; Laarhoven 2013), and experimental physics (Brady and Greighton 2000; Meinshausen, Bickel, and Rice 2009), among others.

Although group testing has its roots in the complete identification of a given population with respect to a particular disease, the estimation of parameters from probability models has also been considered. For example, there is an extensive literature on group testing for disease prevalence estimation (Thompson 1962; Tu, Litvak, and Pagano 1995; Delaigle and Hall 2012; Liu et al. 2012; Warasi et al. 2016; and Haber, Malinovsky, and Albert 2018). This article deals only with the identification problem.

There are two main classifications of group testing models for identification purposes: probabilistic group testing (PGT) and combinatorial group testing (CGT). In PGT, a probability model is assumed for the joint distribution of N binary population items. In CGT, it is assumed that there are a certain number of infected individuals among N individuals and combinatorial techniques are used to identify them (Du and Hwang 1999). For a more detailed discussion of this classification, see Bar-Lev, Stadje, and van der Duyn Schouten (2005). This article focuses on only PGT.

Throughout this article, we assume that the tests are not subject to misclassification (i.e., a gold-standard test is assumed). In many practical situations, a test is only considered for screening when the misclassification is very small. Thus, it is important to present a careful comparison of designs when tests are not subject to error. However, we do recognize the work of others who have addressed the issue of screening with a misclassified test (Kim et al. 2007; McMahan, Tebbs, and Bilder 2012; Malinovsky, Albert, and Roy 2016).

The purpose of this article is to provide a theoretical framework for comparing commonly used designs for group testing. In order to rigourously perform these comparisons, we develop new and discuss previously known results in the group testing (GT) literature. An early PGT formulation was a simple procedure proposed by Dorfman (1943), followed by an extension proposed by Sterrett (1957). Although these two procedures have been investigated, no one has presented a careful comparison of these two designs relative to an optimal nested procedure that we will define later.

We start the discussion with three commonly used GT procedures that were designed for the binomial group testing problem, where a set of N individuals have to be classified either as positive or not under assumptions that each individual has the same probability p of being positive and the outcomes of different individuals are independent. In the group testing literature such a set is called the *binomial set* (Sobel and Groll 1959).

Dorfman procedures: Procedures D and D'

We begin by introducing the Dorfman blood testing problem (see also Feller 1950, p. 189). The motivation was the need to administer blood tests for syphilis to millions of people drafted into the U.S. army during World War II. Dorfman (1943) suggested to group the blood samples of each of the k people and apply a single blood test to the entire group. If the

group test is negative, then only the single test is required for identification of the k individuals. If the group test is positive then each of the k individuals is tested separately, resulting in $k+1$ tests. This procedure is commonly referred to as the Dorfman two-stage group testing procedure (Procedure D). The intuition behind Procedure D is that for small p, a second stage will rarely be required. This method is used by the American Red Cross in the screening of blood donations for HIV and hepatitis (Dood et al. 2002).

Dorfman's two-stage GT procedure belongs to the nested class of GT procedures, which we will define later. The performance of any GT procedure will be evaluated with respect to the expected number of tests. We call a GT procedure *optimal* within a particular class of procedures if it has the minimal expected number of tests.

There is a logical inconsistency in Procedure D. It is clear that any "reasonable" group testing plan should satisfy the following property: "A test is not performed if its outcome can be inferred from previous test results" (Ungar 1960, p. 50). Procedure D does not satisfy this property since if the group is positive and all but the last person are negative, the last person is still tested. The modified Dorfman procedure (Sobel and Groll 1959) (defined as D') would not test the last individual in this case.

Even though the two procedures D and D' are very similar, we will show later in this article that particularly, when the prevalence is high, the efficiency gain of D' over D with respect to the expected number of tests may be substantial.

Sterrett procedure: Procedure *S*

Sterrett (1957) suggested an improvement of Procedure D' in the following way. If in the first stage of Procedure D' the group is positive, then in the second-stage individuals are tested one-by-one until the first positive individual is identified. Then, the first stage of Procedure D' is applied to the remaining (nonidentified) individuals. The procedure is repeated until all individuals are identified.

When p is small, the probability of having two or more positive individuals in a group is very small. Therefore, when retesting a positive group, it is most probable that we will test sequentially until the first positive and then the remaining individuals grouped together will be negative. Thus, it is intuitively clear that Procedure S will be more efficient than Procedure D' in this situation.

Efficiency of a GT procedure

When comparing the different procedures throughout this article, we make the distinction between an infinite and a finite population of size N. This is done since for a large (infinite) population it is natural to obtain optimality results under the assumption of equal group sizes, and for a finite population we need to consider situations, where the population cannot be divided into equal group sizes. More formally, for an infinite population we are looking for a common group size k_A^* (for a given Procedure A ($A \in (D, D', S)$)) that minimizes the expected number of tests per person. For a finite N, we partition the population into subsets, where we apply a Procedure A ($A \in (D, D', S)$) within each subset. In this case, the

optimality is defined by finding a partition such that the expected total number of tests is minimal. For example, when $N = 10$, we may consider a design where we apply the procedure to the entire population, partition the 10 individuals into two groups of size 5, or partition them into three groups of size 3, 3, and 4, and apply the procedure separately in each group. In Section 3, we will show that the optimal partition has equal subgroup sizes if N is divisible by k_A^*. This motivates us to start the discussion with the infinite population case in order to obtain k_A^*, and then to apply the optimality results from the infinite to the finite population case.

For a general group testing problem, a procedure is optimal if, for a given N and p, it achieves the minimum expected total number of tests $E(N, p)$. A general optimal procedure is unknown, and its characterization was conjectured as an intractable problem (Du and Ko 1987). Du and Ko (1987) determined the computational complexity of a wide class of group testing models, where they proved that a general version of binomial group testing is NP-complete (no polynomial time solution is known) (Garey and Johnson 1979).

There are only a few fundamental results in binomial group testing that provide insights on the structure of $E(N, p)$. Ungar (1960) characterized the optimality of any group testing algorithm and proved that if $p \geq p_U = (3 - 5^{1/2})/2 \approx 0.38$, then there does not exist an algorithm that is better than individual one-by-one testing. That is,

$$E(N, p) = N, \quad \text{for} \quad p \geq p_U.$$

Web Appendix G provides additional details about Ungar's result. From now on, we will refer to p_U as Ungar's universal cut-off point (UCP). Another important result is due to Yao and Hwang (1988), who showed that $E(N, p)$ is increasing in p for $0 < p < p_U$ and $N \geq 2$.

A nested algorithm has the property that if the positive subset I is identified, the next subset I_1 that we will test is a proper subset of I, that is, $I_1 \subset I$. This natural class of GT procedures was defined by Sobel and Groll (1959) and Sobel (1960), and it is clear that Procedures D, D', and S all belong to this class. The optimal nested algorithm is not optimal among all possible GT algorithms, but they are simple to implement due to their sequential nature. Further, the optimal nested algorithm is nearly optimal over all algorithms (Sobel 1960, 1967).

This article is organized as follows. In order to compare procedures, we present in Section 2.1 the expected number of tests per person under Procedures D, D', and S. A new short proof for the expected number of tests under Procedure S is obtained. In Sections 2.2.1–2.2.3, we present the optimization problem under Procedures D, D', and S for the infinite population, and in Section 3, we present an optimal partition of the finite population under Procedures D, D', and S. In Section 4.1, we present the optimal nested procedure that can be found (even for a very large population size) with dynamic programming. Section 5 investigates the robustness of an optimum nested procedure versus D, D', and S in the case, where the only available information is an upper bound on parameter p. The proofs of key optimality results are presented in Appendices A–E. Web Appendices F–I provide

theoretical derivations to support other results stated in this article. Web Appendix J provides Matlab code for the optimum nested procedure.

2. Optimality Under an Infinite Population

In this section, we will compare two simple procedures that are useful for the case of (large) infinite populations. In order to make these comparisons, we need to present some known as well as new theoretical results.

2.1. Expected Number of Tests for the Procedures D, D′, and S

We denote the expected number of tests per person in a group of size k under Procedure A as $E_A(k, p)$. Under the binomial group testing model, we have the following characteristics under Procedures D, D', and S.

Procedure D—For $k \geq 2$, the total number of tests is 1 with probability q^k ($q = 1 - p$) and $k + 1$ with probability $1 - q^k$. Therefore,

$$E_D(k, p) = \begin{cases} 1 - q^k + \frac{1}{k} & for \quad k \geq 2 \\ 1 & for \quad k = 1. \end{cases} \tag{1}$$

Procedure *D′*—For $k \geq 2$, the total number of tests is 1 with probability q^k, k with probability $q^{k-1}(1 - q)$, and $k + 1$ with probability $1 - q^k - q^{k-1}(1 - q) = 1 - q^{k-1}$. Therefore,

$$E_{D'}(k, p) = 1 - q^k + 1/k - (1/k)(1 - q)q^{k-1}. \tag{2}$$

It is easy to check that $E_{D'}(1, p) = 1$, and that $E_{D'}(k, p) \leq E_D(k, p)$.

Procedure *S*—Sterrett (1957) provided an expression for $E_S(k, p)$ in terms of a finite sum of terms for which each element involves a binomial coefficient, and the resulting expression is very complex. In fact, this finite sum has a simple closed-form expression.

Result 1.

$$E_S(k, p) = \frac{1}{k}\left[2k - (k - 2)q - \frac{1 - q^{k+1}}{1 - q}\right]. \tag{3}$$

It is easy to check that $E_S(1, p) = 1$. Sobel and Groll (1959) provided this closed-form expression for $E_S(k, p)$ as a consequence of the general recursive equations. We prove this result with alternative short arguments in Appendix A.

In the remainder of this section, we present results needed for a careful comparison of Procedures D, D', and S.

2.2. Determining the Optimal Design for Procedures *D*, *D′*, and *S*

For an infinite population, our goal is to find the optimal group size $k_A^*(p)$ for a given Procedure A. It should be recognized that $k_A^*(p)$ is a function of p and this dependence on p is suppressed in the notation. The difficulty in developing a closed-form expression for k_A^* lies in the discreteness of the problem. Also, equations (1) and (2) are not unimodal as a function of k for a given p.

2.2.1. Procedure D—In the original work of Dorfman (1943), there is no closed-form solution for $k_D^*(p)$, only numerical evaluations. For Procedure D, Samuels (1978) solved this optimization problem and showed that if $p < p_D = 1 - 1/3^{1/3} \approx 0.31$, then k_D^* is equal to $1 + [p^{-1/2}]$ or $2 + [p^{-1/2}]$ (where $[p]$ is denoted as the integer part of p); otherwise, $k_D^* = 1$. From his result, it follows that the applicability of Procedure D is limited by the value of p_D.

2.2.2. Procedure *D′*—Procedure D' was mentioned by Sobel and Groll (1959) and investigated in detail by Pfeifer and Enis (1978). They did not provide the closed-form solution for $k_{D'}^*$ but provided the following result, which immediately led to the solution.

Lemma 2 in Pfeifer and Enis (1978). Let $p \in (0, (3 - \sqrt{5})/2)$. Then (as a function of the continuous variable k) $E_{D'}(k, p)$ has an absolute minimum which is at the smallest zero of $E'_{D'}(k, p) = \frac{\partial E_{D'}(k, p)}{\partial k}$. This zero is unique in that portion of the domain of $E_{D'}$ for which $E_{D'}(k, p) < 1$.

From the above lemma, it follows that the optimal value $k_{D'}^*$ is the smallest k value which satisfies

$$E_{D'}(k, p) \le E_{D'}(k - 1, p) \quad \text{and} \quad E_{D'}(k, p) < E_{D'}(k + 1, p)\,. \tag{4}$$

Therefore, we have to sequentially check the above inequalities for $k = 2, 3, \ldots$ in order to find this smallest value of k.

It is clear from the above inequalities that there are nonunique solutions for some values of p (i.e., there are two solutions for the value of p where $E_{D'}(k, p) = E_{D'}(k - 1, p)$). From equations (1) and (2) and Lemma 2 of Pfeifer and Enis (1978), it follows that $k_{D'}^* \le k_D^*$ for $p < 1 - 1/3^{1/3}$. It was stated in Pfeifer and Enis (1978) that it does not seem possible to explicitly obtain a closed-form expression for the optimal group size $k_{D'}^*(p)$. Although we cannot prove it, we empirically verified the conjecture that the optimal group size $k_{D'}^*(p)$ is equal to $\lfloor p^{-1/2} \rfloor$ or $\lceil p^{-1/2} \rceil$ for $0 < p < p_U = \frac{3 - \sqrt{5}}{2}$ where $\lfloor x \rfloor (\lceil x \rceil)$ for $x > 0$ is defined as the largest (smallest) integer which is smaller (larger) than or equal to x. This conjecture was examined for values of p in the above range with incremental steps of 10^{-6} in the following way. For a given value of p, the optimal group size $k_{D'}^*(p)$ was found using (4), and it was then verified that it is equal to either $\lfloor p^{-1/2} \rfloor$ or $\lceil p^{-1/2} \rceil$. The curious reader can easily check

this conjecture empirically using the values of Table 1 and Table 1 in Pfeifer and Enis (1978).

2.2.3. Procedure S—Sterrett (1957) failed to provide the closed-form expression for $E_S(k, p)$ but instead provided a large-sample (infinite population) approximation. As a consequence, there are some inaccurate results in his Table I.

The following new result provides a way to find the optimal group size under Procedure S.

Result 2. Let $p \in (0, (3-\sqrt{5})/2)$. Then (as a function of continuous variable k, $k \geq 1$) $E_S(k, p)$ has an absolute minimum at the unique zero of $E'_S(k, p)$.

For the proof of Result 2, please see Appendix B. □

From Result 2, it follows that the optimal value $k^*_S(p)$ is equal to $\lceil l \rceil$ or $\lfloor l \rfloor$, where $E'_S(l, p) = 0$. Alternatively, to avoid solving the nonlinear equation $E'_S(l, p) = 0$, we can find the optimal k^*_S in the same manner as under Procedure D' using (4). We conjecture that the optimal group size $k^*_S(p)$ is equal to $\lfloor\sqrt{2/p}\rfloor$ or $\lfloor\sqrt{2/p}\rfloor + 1$ or $\lfloor\sqrt{2/p}\rfloor + 2$ for $0 < p < p_U$. Although we cannot prove it, we empirically verified this conjecture in the same manner as we did for $k^*_{D'}(p)$ in Section 2.2.2.

2.2.4. A Comparison of Procedures *D*, *D*′, and *S*—Table 1 shows a detailed comparison of the optimal group size and the corresponding total expected number of tests per 100 for Procedures D, D', and S as a function of p. In particular, for large p, Procedure D' has an impressive efficiency gain over Procedure D. Although Procedure S is uniformly better than D' for all p, for small values of p, Procedure S is substantially better than D'. Sterrett (1957) provided a similar comparison for Procedures D and S. However, he only used large-sample approximations for the optimal group size and expected number of tests, which were slightly inaccurate. Table 1 along with previous discussed theoretical results show that Procedures D' and S (but not D) achieve the same upper applicability bound (UCP) p_U (see also Web Appendix H).

3. Optimality Under a Finite Population

In Section 2, we have discussed the infinite population case, where for a given Procedure $A \in \{D, D', S\}$, we find the value k which minimizes the expected number of tests per person, $E_A(k, p) = \frac{h_A(k, p)}{k}$, where $h_A(k, p)$ is the expected total number of tests for a group of size k and for a prevalence of p. Define $h_A(k) = h_A(k, p)$.

Generally, for a finite population of size N and a given Procedure A, we have to solve the following optimization problem: find the optimal partition $\{n_1, \ldots, n_I\}$ with $n_1 + \cdots + n_I = N$ for some $I \in \{1, \ldots, N\}$ such that $h_A(N, p)$ is minimal (denote $H_A(N)$), that is, $\{n_1, \ldots, n_I\}$ is a solution of the following optimization problem:

$$H_A(N) = \min_{m_1, m_2, \ldots, m_J} \sum_{i=1}^{J} h_A(m_i), \text{ subject to, } \sum_{i=1}^{J} m_i = N, \quad J \in \{1, \ldots, N\}. \tag{5}$$

Recall from the Introduction that for $p \geq p_U$, the optimum is to test one-by-one for either finite or infinite population size. Therefore, in this case $I = N$ and the optimal partition is $\{n_i = 1, i = 1, \ldots, N\}$. A common method to solve (5) is dynamic programming (DP) (Bellman 1957); the first application of DP in group testing appeared in Sobel and Groll (1959), and that for Procedure D' was presented by Pfeifer and Enis (1978). The DP algorithm can be expressed as

$$h_A(1) = 1, \quad H_A(0) = 0, \quad H_A(1) = 1,$$
$$H_A(k) = \min_{0 \leq x \leq k-1} \{H_A(x) + h_A(k - x)\}, \quad k = 2, \ldots, N. \tag{6}$$

It is obvious that the computation effort of the above DP algorithm is $O(N^2)$, which makes it easy to implement, and it can be computationally fast enough for even large values of N.

We can compare D, D', and S for a finite population using this DP algorithm (6). For example, if $p = 0.05$, then the optimal group size for an infinite population under Procedure D is $k_D^* = 7$ (Table 1), and the optimal partition when $N = 13$ is $\{n_1, n_2, n_3\} = \{5, 4, 4\}$ with $H_D(13) = 5.615$; for Procedure D' is $k_{D'}^* = 5$ (Table 1), and the partition when $N = 13$ is $\{n_1, n_2, n_3\} = \{5, 4, 4\}$ with $H_{D'}(13) = 5.489$; for Procedure S with the same $p = 0\ 05$, $k_S^* = 7$ (Table 1) and the optimal partition when $N = 13$ is $\{n_1, n_2\} = \{6, 7\}$ with $H_S(13) = 4.685$.

The example above illustrates some interesting features of subgroup sizes of the optimal partition. Specifically, we see that the optimal partition subgroup sizes differ at most by one unit. This was conjectured by Lee and Sobel (1972) as a general result for Procedure D and they provided insight by using the convexity (with respect to k) property of an approximation to $E_D(k, p)$. Gilstein (1985) proved this result for Procedure D'. We will prove this result for Procedure S.

In the following result, we show that the optimal partition has equal subgroup sizes if N is divisible by k_S^*. It is clear from the proof that the same holds for the Procedures D and D'.

Result 3. Suppose we apply the group testing algorithm S for a finite population of size N for a given p. Also suppose that $N = sk_S^*(p)$, i.e., s subgroups of size k_S^*. Then, the optimal partition is $\{n_i = k_S^*, i = 1, \ldots, s\}$, that is, $I = s$ and the infinite population optimal solution is the subgroup size of the optimal partition for the finite population.

Proof. Please see Appendix C.

The following result establishes a relationship among the size of subgroups under the optimal partition for Procedure S.

Result 4. Suppose we apply group testing algorithm S for a finite population of size N for a given p ($p \in (0, 1)$) and we start with some partition $\{m_1, \ldots, m_J\}$. There exists a better (with respect to expected number of tests) partition $\{m'_1, \ldots, m'_J\}$ with $|m'_j - m'_i| \leq 1$ for all i, j.

Proof. The proof is based on the convexity property of $h_S(x)$ with respect to x (see also discussion on the previous page) and is presented in Appendix D. □

From Result 4, it follows that if we apply group testing algorithm S for a finite population of size N for a given p ($p \in (0, 1)$) and we start with an optimal number of subgroups I, then there exists an optimal partition containing groups whose sizes differ by at most 1.

The following result, which was conjectured for Procedures D and D' by Lee and Sobel (1972), provides a simple way to construct an optimal partition. Gilstein (1985) proved the result for D' and his method also applies to Procedure D. We prove it for Procedure S.

Result 5. Suppose we apply group testing algorithm S for a finite population of size N for a given p ($p \in (0, 1)$). Denote a to be an optimal group size under Procedure S for an infinite population, $s = \left\lfloor \frac{N}{a} \right\rfloor$ (i.e., s groups of size a) and $\theta = N - sa$ (i.e., remainder $0 < \theta < a$). Then, the optimal partition is one of the two following partitions:

- **i.** Distribute the remainder θ among s groups (with initial size a) in such a way that $|n_i - n_j| \leq 1$ for all $i, j \in \{1, \ldots, s\}$.
- **ii.** Build up an additional group (group $s + 1$) by taking the remainder θ and units from the above s groups (with initial size a) in such way that $|n_i - n_j| \leq 1$ for all $i, j \in \{1, \ldots, s, s + 1\}$.

Proof. Please see Appendix E. □

From Result 5, it follows that in order to find the optimal partition, we need to evaluate the total expected number of the tests in (i) and (ii) and choose the design ((i) or (ii)) that minimizes this quantity. Result 5 provides for finding the optimum partition without any computational cost. In Web Appendix F (Remark 1), we provide an alternative direct implementation of Result 5.

Table 2 provides the optimum partition (Procedures D, D', S) for the finite population case for different values of p in a similar way as in Table 1. The last two columns of Table 2 will be discussed in Section 4.

4. Optimal Nested Procedure and Connection with Coding Theory

4.1. Optimal Nested Procedure

A nested procedure, which was defined in the Introduction (Sobel and Groll 1959), requires that between any two successive tests:

i. future tests are concerned only with units not yet classified as good or defective,

ii. n units not yet classified have to be separated into only (at most) two sets. One set of size $m \geq 0$, called the "defective set," is known to contain at least one defective unit if $m \geq 1$ (it is not known which ones are defective or exactly how many there are). The other set of size $n - m \geq 0$ is called the "binomial set" because we have no knowledge about it other than the original binomial assumption. Either of these two sets can be empty in the course of experimentation; both are empty at termination.

The number of potential nested group testing algorithms is astronomical. For example, if $N = 5$, then there are 235,200 possible algorithms (Moon and Sobel 1977). Therefore, it is impossible to directly evaluate the expected number of tests for each algorithm, making a direct computation infeasible. Sobel and Groll (1959) overcame this problem by proposing a DP algorithm that finds the optimal nested algorithm, which Sobel and Groll termed "the Procedure R_1." There was a large research effort to reduce the computational complexity $O(N^3)$ of the original proposed algorithm (Sobel 1960; Kumar and Sobel 1971; Hwang 1976). With new theoretical results, Sobel (1960) reduced the complexity to $O(N^2)$. Further, Kumar and Sobel (1971) reduced the computation complexity by half as compared with Sobel (1960). Finally, Hwang (1976), using the results for optimal binary trees (Huffman trees, Huffman 1952) and optimal alphabetic binary trees (Hu and Tucker 1971), reduced the computational complexity to $O(N)$ (not including the sorting effort). In addition, Yao and Hwang (1990) proved that the pairwise testing algorithm is the unique (up to the substitution of equivalent items) optimal nested algorithm for all N if and only if $1 - \frac{1}{\sqrt{2}} < p < p_U$ (at the boundary values the pairwise testing algorithm is an optimal nested algorithm). Recently, Zaman and Pippenger (2016) provided an asymptotic analysis of the optimal nested procedure.

The development of the optimal nested algorithm (due to Sobel) of complexity $O(N^2)$ is presented in Web Appendix I. This result allows for computing the optimal total expected number of tests $H_1(N)$ (under the optimal nested Procedure R_1).

For example, if $p = 0.05$ and $N = 13$, then the expected number of tests under the optimum nested Procedure R_1 is $H_1(13) = 3.878$. For comparison, with the same values of p and N, we obtain $H_D(13) = 5.615$, $H_{D'}(13) = 5.489$ and $H_S(13) = 4.685$. Web Appendix I provides a thorough explanation of the construction of an optimal nested procedure in this case.

In Table 2, we present the expected number of tests (optimum nested procedure) per $N = 100$ individuals ($E_1(100)$) for different values of p.

4.2. Coding Theory and Information Lower Bound

In the previous subsection, we showed that DP can be used to obtain the optimal nested procedure. However, it does not speak more generally to optimality among all possible procedures. Deriving an information lower bound for the expected total number of tests of an optimal procedure provides insight into the efficiency of the optimal nested procedure. The information lower bound (ILB) was provided in Sobel and Groll (1959). Sobel (1960,

1967) used noiseless-coding theory to derive ILB. In Web Appendix H, we carefully demonstrate the development of the ILB using coding theory attributed to Sobel (1960, 1967). Web Appendix H provides a good pedagogical tool for this important development in GT. The key result is the information lower bound $H(p)$ for the expected number of tests under an optimal procedure that is the Shannon formula of entropy:

$$H(p) = N\left[p \log_2 \frac{1}{p} + q \log_2 \frac{1}{q}\right]. \tag{7}$$

The information lower bound $H(p)$ is not attainable but provides a benchmark for what is a close-to-attainable level for an optimal group testing procedure (for a detailed discussion, see Web Appendix H). In the last column of Table 2, we present the information lower bound $H(p)$ for different values of p when $N = 100$.

5. Robustness Investigation

In this section, we investigate the robustness of the procedures to the incorrect specification of the parameter p. In order to simplify this investigation, we will assume the large population setting, which will allow for a common group size for a given procedure. Optimal group sizes under Procedures D, D', S, and an optimal nested procedure are all functions of parameter p. However, p may not be known, and interest is on the comparison of different design strategies when p is not correctly specified. In some situations, there is only knowledge of an upper bound U of the design parameter p. Under a constant group size setting, such as in Procedures D, D', and S, we can follow the methodology developed by Malinovsky and Albert (2015) to calculate the minimax group size k_A^{**} for Procedures $A \in \{D, D', S\}$ as

$$k_A^{**} = \arg \min_{k \in \mathbb{N}^+} \sup_{p \in (0, U]} L_A(k, p), \tag{8}$$

where $L_A(k, p) = E_A(k, p) - E_A(k^*(p), p)$, $A \in \{D, D', S\}$.

Table 3 shows the expected number of tests per 100 individuals for minimax designs of D, D', and S along with nested Procedure R_1 using U instead of p ($H_1(100)$) and $U/2$ instead of $p\left(H_1^*(100)\right)$ for different p. We evaluated the nested Procedure R_1 at a value of U. Further, since values of p are often substantially lower than a specified upper bound U, we also evaluated the procedure at a value of $U/2$.

Table 3 shows that Procedure R_1 is generally more efficient than D, D' and S. However, in rare situations where the assumed upper bound (U) is substantially higher than the true unknown p, Procedure S may indeed be more efficient than R_1.

6. Summary

This article provides a unique perspective on group testing, where we tie together literature on infinite and finite populations GT, dynamic programming, and coding theory. This is done in order to compare important nested group testing procedures, including Dorfman (D

and D'), Sterrett (S), and an optimal nested Procedure R_1, with the theoretical information lower bound of efficiency serving as a reference. All theoretical developments were essential for making these comparisons.

Some of the results were provided previously in the literature, while others, particulary for Procedure S, are new. We demonstrated that, particularly when p is small, Procedure S has a large efficiency gain relative to Procedures D and D'. Further, there can be a sizable efficiency gain by using the optimal nested procedure that is based on DP. However, this efficiency gain needs to be weighed against the practical complications in implementing the different procedures. For example, although there is a sizable efficiency gain in using the optimal nested procedure, the complex nature of the design may make it less practical (see Web Appendix I). The less efficient Dorfman procedure is simple to implement in that it is a two-stage procedure, where testing within the second stage can be conducted in parallel (simultaneously). This is in contrast with the Sterrett procedure, where stages subsequent to the first stage are sequential and cannot be performed in parallel.

These results are based on the correct specification of p. Using our newly derived results on the Sterrett procedure, we were able to show that even when p is misspecified, the optimal nested procedure is generally more efficient than S. However, it is important to recognize that this may not be the case when our knowledge of p is far from the truth (i.e., when U is substantially larger than p). Importantly, for any p, Procedure S is more efficient than D'.

The results in this article highlight the importance of studying efficient procedures in group testing. Simplicity aside, the Sterrett and optimal nested procedures are more efficient than Dorfman's procedure. This is also true when p is misspecified. Modern applications of nested procedures have generally focused on using Dorfman's procedure (Hill et al. 2016; France et al. 2015), although Sterrett's procedure is also used (Bilder, Tebbs, and Chen 2010), rather than an alternative nested procedure. The results of this article clearly demonstrate the advantage of using the Sterrett procedure or an optimal nested procedure whenever it practicably feasible. Based on our results, we encourage the use of the Sterrett procedure and optimal nested procedures in practice.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors thank the editor, the associate editor, and two referees for their thoughtful and constructive comments. Also, the comments and suggestions of the associate editor resulted in significant improvements of the article. The authors thank Mattson Publishing Services for editing the article. This article is dedicated to the memory of David Assaf, a teacher and a friend.

Funding

The work was supported by the National Cancer Institute Intramural Program.

Appendix

A.: Proof of Result 1

Proof. Let X be the number of tests in order to identify k persons and let 1_j be an indicator function that is equal to 1 if the first positive identified person is the person j ($j = 1, \ldots, k$) tested. Also, denote 1_0 as an indicator function that is equal to 1 if no positive person is in the group of size k. We have

$$X = X1_0 + X1_1 + X1_2 + \cdots + X1_{k-1} + X1_k.$$

Define $E_k(X) \equiv E(k) = kE_S(k, p)$. It is clear that $E(1) = 1$. Therefore,

$$\begin{aligned} E(k) &= q^k + kq^{k-1}(1-q) + (1-q)(2 + E(k-1)) + q(1-q)(3 + E(k-2)) + q^2(1-q)(4 + E(k-3)) + \cdots \\ &+ q^{k-2}(1-q)(k + E(k-(k-1))) \\ &= 1 - (k-1)q^k + \frac{1-q^{k-1}}{1-q} + (1-q)[E(k-1) + qE(k-2) + q^2E(k-3) + \cdots + q^{k-2}E(1)]. \end{aligned}$$

Taking the difference $E(k+1) - E(k)$, we get

$$E(k+1) = E(k) + 2 - q - q^{k+1}.$$

Substituting the appropriate expression for $E(k)$, $E(k-1), \ldots, E(1) = 1$, we get

$E(k) = (2k-1) - (k-2)q - \frac{q - q^{k+1}}{1-q} = 2k - (k-2)q - \frac{1-q^{k+1}}{1-q}$. □

B.: Proof of Result 2

Proof. Denote $f(k) = E_S(k, p)$, $\dot{f}(k) = \frac{\partial f(k)}{\partial k}$, $\ddot{f}(k) = \frac{\partial^2 f(k)}{\partial k^2}$. Recall (see (3)) that

$f(k) = 2 - q - \frac{1 - q^{k+1} - 2q + 2q^2}{k(1-q)}$. We have $f(1) = 1$, $\lim_{k \uparrow \infty} f(k) = 2 - q$,

$\dot{f}(k) = \frac{1}{1-q}\left[\frac{1}{k^2}\left(1 - q^{k+1} - 2q + 2q^2\right) + \frac{1}{k}q^{k+1}ln(q)\right]$. So, $\dot{f}(1) < 0$ for $q \in \left(\frac{\sqrt{5}-1}{2}, 1\right)$ and, therefore, the function $f(k)$ (as a function of continuous variable $k \geq 1$) has a minimum in support $k \geq 1$. Further, $\ddot{f}(k) = \frac{1}{1-q}\left[-\frac{2(1-q)}{k}\dot{f}(k) + \frac{1}{k}q^{k+1}(\ln(q))^2\right]$ and if $\dot{f}(l) = 0$, then $\ddot{f}(l) > 0$, shows that l is the unique minimum. □

C.: Proof of Result 3

Proof. $N = sk_S^*(p)$. $\frac{h_S\left(k_S^*(p)\right)}{k_S^*(p)} = E_S\left(k_S^*(p), p\right) \le E_S(k, p) = \frac{h_s(k)}{k}$ for any $k = 1, 2,\ldots$ implies $sh_S\left(k_S^*(p)\right) = \frac{\sum_{i=1}^{J} m_i}{k_S^*(p)} h_S\left(k_S^*(p)\right) \le \sum_{i=1}^{J} h_S\left(m_i, p\right)$ for any partition $\{m_1, \ldots, m_J\}$ with $\sum_{i=1}^{J} m_i = N$, $j \in \{1, \ldots, N\}$, which completes the proof. □

D.: Proof of Result 4

Proof. It is easy to verify that for all $p \in (0, 1)$ the second derivative of $h_S(x) = xE_S(x, p)$ with respect to x is positive and, therefore, the function $h_S(x)$ is convex with respect to x. We start with some partition $\{m_1, \ldots, m_J\}$. Convexity of $h_S(x)$ implies that for any $m_j - m_i \ge 2$, $h_S(m_j - 1) + h_S(m_i + 1) \le h_S(m_j) + h_S(m_i)$. Applying this +1, −1 improvement for any i, j with $m_j - m_i \ge 2$, we obtain a better (with respect to expected number of tests) partition $\left\{m'_1, \ldots, m'_J\right\}$ with $\left|m'_j - m'_i\right| \le 1$ for all i, j. □

E.: Proof of Result 5

The proof for Procedure S is exactly the same as a proof for Procedures D' and M in Gilstein (1985) (p. 389) and is based on the fact that the function $f(x) = E_S(x, p)$ has a unique minimum for $x \ge 1$ as was shown in the proof of Result 2 in Appendix B.

References

Bar-Lev SK, Boneh A, and Perry D (1990), "Incomplete Identification Models for Group-testable Items," Naval Research Logistics, 37, 647–659.

Bar-Lev SK, Stadje A, and van der Duyn Schouten FA (2005), "Multinomial Group Testing Models with Incomplete Identification," Journal of Statistical Planning and Inference, 135, 384–401.

Bar-Lev SK, Boxma O, Kleiner I, and Perry D (2017), "Recycled Incomplete Identification Procedures for Blood Screening," European Journal Operational Research, 259, 330–343.

Bellman R (1957), Dynamic Programming, Princeton, NJ: Princeton University Press.

Bilder CR, Tebbs JM, and Chen P (2010), "Informative Retesting," Journal of the American Statistical Association, 105, 942–955. [PubMed: 21113353]

Brady P, and Greighton T (2000), "Searching for Periodic Sources with LIGO. II: Hierarchical Searches," Physical Review D, 61, 082001.

Cao C, and Sun X (2016), "Combinatorial Pooled Sequencing: Experiment Design and Decoding," Quantitative Biology, 4, 36–46.

Delaigle A, and Hall P (2012), "Nonparametric Regression with Homogeneous Group Testing Data," Annals of Statistics, 40, 131–158.

Dodd RY, Notari IV EP, and Stramer SL (2002), "Current Prevalence and Incidence of Infectious Disease Markers and Estimated Window-period Risk in the American Red Cross Blood Donor Population," Transfusion, 42, 975–979. [PubMed: 12385406]

Dorfman R (1943), "The Detection of Defective Members of Large Populations," The Annals of Mathematical Statistics, 14, 436–440.

Du D, and Hwang FK (1999), Combinatorial Group Testing and its Applications, Singapore: World Scientific.

Du D, and Hwang FK (2006), "Pooling Design and Nonadaptive Group Testing: Important Tools for DNA Sequencing," Singapore: World Scientific.

Du DZ, and Ko KI (1987), "Some Completeness Results on Decision Trees and Group Testing," SIAM Journal on Algebraic and Discrete Methods, 8, 762–777.

Feller W (1950), An Introduction to Probability Theory and Its Application, New York: Wiley.

France B, Bell W, Chang E, and Scholten T (2015), "Composite Sampling Approaches for Bacillus Anthracis Surrogate Extracted from Soil," PLoS One, 10, 1–18.

Garey MR, and Johnson DS (1979), Computers and Intractability. A Guide to the Theory of NP-Completeness, San Francisco, CA: W. H. Freeman and Co.

Gastwirth J, and Johnson W (1994), "Screening with Cost Effective Quality Control: Potential Applications to HIV and Drug Testing," Journal American Statistical Association, 89, 972–981.

Gilstein CZ (1985), "Optimal Partitions of Finite Populations for Dorfman-type Group Testing," Journal Statistical Planning Inference, 12, 385–394.

Graesbøll K, Andresen L-O, Halasa T, and Toft N (2016), "Opportunities and Challenges when Pooling Milk Samples Using ELISA," Preventive Veterinary Medicine, 139 Part B, 93–98.

Haber G, Malinovsky Y, and Albert PS (2018), "Sequential Estimation in the Group Testing Problem," Sequential Analysis, 37, 1–17.

Hill JA, HallSedlak R, Magaret A, Huang ML, Zerr DM, Jeromeb KR, and Boeckh M (2016), "Efficient Identification of Inherited Chromosomally Integrated Humanherpesvirus 6 Using Specimen Pooling," Journal of Clinical Virology, 77, 71–76. [PubMed: 26921738]

Hu TC, and Tucker AC (1971), "Optimum Computer Search Tree," SIAM Journal on Applied Mathematics, 21, 514–532.

Huffman DA (1952), "A Method for the Construction of Minimum-Redundancy Codes," Proceedings of the I.R.E, 40, 1098–1101.

Hwang FK (1976), "An Optimal Nested Procedure in Binomial Group Testing," Biometrics, 32, 939–943.

Kim HY, Hudgens MG, Dreyfuss JM, Westreich DJ, and Pilcher CD (2007), "Comparison of Group Testing Algorithms for Case Indentification in the Presence of Test Error." Biometrics, 63, 1152–1162. [PubMed: 17501946]

Kumar S, and Sobel M (1971), "Finding a Single Defective in Binomial Group-testing." Journal of the American Statistical Association, 66, 824–828.

Laarhoven T (2013), "Efficient Probabilistic Group Testing Based on Traitor Tracing," 51st Annual Allerton Conference on Communication, Control and Computing, At Monticello IL, USA.

Lee JK, and Sobel M (1972), "Dorfman and R_1-type Procedures for a Generalized Group Testing Problem," Mathematical Biosciences, 15, 317–340.

Liu A, Liu CL, Zhang Z, and Albert PS (2012), "Optimality of Group Testing in the Presence of Misclassification," Biometrika, 99, 245–251. [PubMed: 23049137]

Malinovsky Y, and Albert PS (2015), "A Note on the Minimax Solution for the Two-stage Group Testing Problem," The American Statistician, 69, 45–52. [PubMed: 28042146]

Malinovsky Y, Albert PS, and Roy A (2016), "Reader Reaction: A Note on the Evaluation of Group Testing Algorithms in the Presence of Misclassification," Biometrics, 72, 299–302. [PubMed: 26393800]

McMahan CS, Tebbs JM, and Bilder CR (2012), "Informative Dorfman Screening," Biometrics 68, 287–296. [PubMed: 21762119]

Meinshausen N, Bickel P, and Rice J (2009) "Efficient Blind Search: Optimal Power of Detection Under Computational Cost Constraints," The Annals of Applied Statistics, 3, 38–60.

Moon JW, and Sobel M (1977), "Enumerating a Class of Nested Group Testing Procedures," Journal of Combinatorial Theory, Series B, 23, 184–188.

Pfeifer CG, and Enis P (1978), "Dorfman-type Group Testing for a Modified Binomial Model," Journal of the American Statistical Association, 73, 588–592.

Samuels SM (1978), "The Exact Solution to the Two-stage Group-testing Problem," Technometrics, 20, 497–500.

Sobel M (1960), "Group Testing to Classify Efficiently all Defectives in a Binomial Sample," in Information and Decision Processes, ed. Machol RE, New York: McGraw Hill, pp. 127–161.

Sobel M (1967), "Optimal Group Testing," Proc. Colloq. on Information Theory, Bolyai Math. Society, Debrecen, Hungary.

Sobel M, and Groll PA (1959), "Group Testing to Eliminate efficiently All Defectives in a Binomial Sample." Bell System Technical Journal, 38, 1179–1252.

Sterrett A (1957), "On the Detection of Defective Members of Large Populations," The Annals of Mathematical Statistics, 28, 1033–1036.

Stramer SL, Wend U, Candotti D, Foster GA, Hollinger FB, Dodd RY, Allain JP, and Gerlich W (2011), "Nucleic Acid Testing to Detect HBV Infection in Blood Donors," The New England Journal of Medicine, 364, 236–247. [PubMed: 21247314]

Tebbs J, McMahan C, and Bilder C (2013), "Two-Stage Hierarchical Group Testing for Multiple Infections with Application to the Infertility Prevention Project," Biometrics, 69, 1064–1073. [PubMed: 24117173]

Thompson KH (1962), "Estimation of the Proportion of Vectors in a Natural Population of Insects." Biometrics, 18, 568–578.

Tu XM, Litvak E, and Pagano M (1995), "On the Information and Accuracy of Pooled Testing in Estimating Prevalence of a Rare Disease: Application to HIV Screening," Biometrika, 82, 287–297.

Ungar P (1960), "Cutoff Points in Group Testing," Communications Pure Applied Mathematics, 13, 49–54.

Warasi M, Tebbs J, McMahan C, and Bilder C (2016), "Estimating the Prevalence of Multiple Diseases from Two-Stage Hierarchical Pooling," Statistics Medicine, 31, 185–191.

Wolf JK (1985), "Born Again Group Testing: Multiaccess Comunications," IEEE Transactions on Information Theory, 31, 185–191.

Yao YC, and Hwang FK (1988), "A Fundamental Monotonicity in Group Testing," SIAM Journal Discrete Mathematics, 1, 256–259.

Yao YC, and Hwang FK (1990), "On Optimal Nested Group Testing Algorithms," Journal Statistical Planning Inference, 24, 167–175.

Zaman N, and Pippenger N (2016), "Asymptotic Analysis of Optimal Nested Group-testing Procedures," Probability Engineering Informational Sciences 30, 547–552.

Zhu L, Hughes-Oliver JM, and Young SS (2001), "Statistical Decoding of Potent Pools Based on Chemical Structure," Biometrics, 57, 922–930. [PubMed: 11550946]

Table 1.

The minimal (optimal) expected number of tests per 100 individuals $\left(100E_A\left(k_A^*, p\right)\right)$ for Procedure A ($A \in \{D, D', S\}$) using an optimal group size k_A^* for different p.

	D		D'		S	
p	k_D^*	$100E_D(k_D^*, p)$	$k_{D'}^*$	$100\,E_{D'}(k_{D'}^*, p)$	k_S^*	$100E_S(k_S^*, p)$
0.001	32	6.2759	32	6.2729	45	4.5844
0.005	15	13.91	15	13.879	21	10.535
0.01	11	19.557	10	19.47	15	15.172
0.03	6	33.369	6	32.94	9	27.305
0.05	5	42.622	5	41.807	7	35.977
0.07	4	50.195	4	48.787	6	43.167
0.10	4	59.39	4	57.567	5	52.288
0.13	3	67.483	3	64.203	4	60.042
0.15	3	71.921	3	68.308	4	64.784
0.20	3	82.133	3	77.867	3	74.933
0.25	3	91.146	2	84.375	3	83.854
0.27	3	94.432	2	86.855	2	86.855
0.30	3	99.033	2	90.5	2	90.5
0.32	1	100	2	92.88	2	92.88
0.35	1	100	2	96.375	2	96.375
0.38	1	100	2	99.78	2	99.78

Table 2.

The minimal (optimal) expected number of tests per 100 individuals (H_A (100)) for Procedure A ($A \in \{D, D', S\}$), and Procedure R_1 (E_1(100)) and comparison with the information lower bound $H(p)$. OP means an optimal partition, and $s \times a$ means s groups of size a.

	D		D'		S			
p	OP	H_D(100)	OP	$H_{D'}$(100)	OP	H_S(100)	E_1(100)	$H(p)$
0.001	2 × 33, 1 × 34	6.281	2 × 33, 1 × 34	6.278	2 × 50	4.605	1.766	1.141
0.005	5 × 14, 2 × 15	13.917	5 × 14, 2 × 15	13.884	5 × 20	10.537	4.749	4.541
0.01	10 × 10	19.562	10 × 10	19.470	5 × 14, 2 × 15	15.181	8.320	8.079
0.03	12 × 6, 4 × 7	33.402	12 × 6, 4 × 7	32.993	10 × 9, 1 × 10	27.325	19.693	19.439
0.05	20 × 5	42.622	20 × 5	41.807	5 × 6, 10 × 7	36.018	28.958	28.640
0.07	25 × 4	50.195	25 × 4	48.787	2 × 5, 15 × 6	43.184	36.916	36.592
0.10	25 × 4	59.390	25 × 4	57.567	20 × 5	52.288	47.375	46.900
0.13	32 × 3,1 × 4	67.492	32 × 3, 1 × 4	64.258	25 × 4	60.042	56.183	55.744
0.15	32 × 3,1 × 4	71.956	32 × 3, 1 × 4	68.396	25 × 4	64.784	61.485	60.984
0.20	32 × 3,1 × 4	82.210	2 × 2, 32 × 3	77.872	32 × 3, 1 × 4	74.974	72.875	72.192
0.25	32 × 3,1 × 4	91.234	50 × 2	84.375	2 × 2, 32 × 3	83.875	82.191	81.128
0.27	32 × 3,1 × 4	94.518	50 × 2	86.855	50 × 2	86.855	84.864	84.146
0.30	32 × 3,1 × 4	99.117	50 × 2	90.5	50 × 2	90.5	88.889	88.129
0.32	100 × 1	100	50 × 2	92.88	50 × 2	92.88	91.574	90.438
0.35	100 × i	100	50 × 2	96.375	50 × 2	96.375	95.633	93.407
0.38	100 × 1	100	50 × 2	99.78	50 × 2	99.78	99.730	95.804

Table 3.

Robustness of the nested Procedure R_1 vs. Procedures D, D', and S.

	$U = 0.05$				$U = 0.10$				$U = 0.20$			
p	0.001	0.005	0.01	0.05	0.001	0.01	0.05	0.10	0.001	0.01	0.1	0.2
$100E_D(k_D^{**}, p)$	10.185	14.455	19.557	52.211	13297	20.226	46.158	69.453	13.297	20.226	69.453	95.723
$100E_D(k_{D'}^{**}, p)$	10.985	14.841	19.470	49.811	13.285	20.109	45.721	68.855	14.969	20.945	65.697	92.565
$100E_s(k_S^{**}, p)$	7.975	11.241	15.185	41.899	10.628	16.138	37.760	59.381	13.024	17.647	55.928	85.889
$H_1(100)$	7.468	9.311	11.567	28.958	15.287	18.007	30.979	47.375	33.233	35.221	53.271	72.875
$H_1^*(100)$	4.511	6.578	9.194	30.242	7.468	11.567	28.958	50.282	15.287	18.007	47.375	79.988
k_D^{**}	11	11	11	11	8	8	8	8	8	8	8	8
$k_{D'}^{**}$	10	10	10	10	8	8	8	8	7	7	7	7
k_S^{**}	14	14	14	14	10	10	10	10	8	8	8	8