Informative Retesting Procedures to Assay High Volume Clinical Specimens

Bilder, Christopher R.
University of Nebraska-Lincoln, Department of Statistics
340 Hardin Hall North
Lincoln, NE 68583, USA
E-mail: chris@chrisbilder.com

Tebbs, Joshua M.
University of South Carolina, Department of Statistics
216 LeConte College
Columbia, SC 29208, USA
E-mail: tebbs@stat.sc.edu

1 Introduction

Group testing, where groups of individual specimens are composited to test for a binary outcome, is a procedure used to reduce the costs of screening large numbers of individuals. Applications of group testing include Bovine Viral Diarrhea virus detection in cattle herds (Kennedy et al., 2006), blood donation screening (Dodd et al., 2002), discovery of chemical compounds to use in new drugs (Remlinger et al., 2006), and opportunistic testing of individuals for chlamydia (Mund et al., 2008). In its simplest form, group testing works by pooling a number of individual specimens into a single group. If the group tests negative, all individuals within it are diagnosed as being negative. If the group tests positive, individual retesting is performed upon the specimens to decode the positive individuals from the negative individuals. This algorithmic approach is often referred to as Dorfman testing after the original proposal given by Dorfman (1943).

Traditionally, group testing research has assumed each individual to have the same probability of positivity when developing testing algorithms. However, this assumption is often unrealistic, especially when known risk factors can be used to estimate distinct probabilities of positivity for each individual. With this as motivation, a number of new procedures have been proposed recently. These *informative retesting* procedures try to take advantage of heterogeneity among individuals (as measured by estimated individual probabilities of positivity) to reduce the number of tests needed. They work in one or more of the following ways: 1) Retest individuals with higher probabilities first, 2) Select group sizes based on the individual probabilities, and 3) Organize the testing to reduce the number of retests that may be needed.

Our paper presents the first overall comparison of the informative retesting procedures that have been proposed. Section 2 provides brief reviews of procedures. Next, Section 3 compares these procedures through the use of data application and simulation. Finally, Section 4 gives overall recommendations on their use by practitioners.

2 Procedures

2.1 Informative Dorfman

Dorfman testing is the most widely used group testing procedure because of its ease in application. McMahan et al. (2011b) builds on this procedure by introducing informative retesting versions of it. Their goal is to maintain the ease of application by having only two steps (test initial group and retest individually if needed) and to reduce the number of tests expended in decoding. Next is a description of two of their proposals.

Threshold optimal Dorfman (TOD) uses a threshold probability to categorize individuals as high risk (individual probability above the threshold) or low risk (individual probability below the threshold). Individuals categorized as high risk are tested individually without using group testing. Individuals categorized as low risk are ordered by their probabilities of positivity and put into groups of equal size for testing (the group size is chosen to minimize the estimated expected number of tests). If a group tests positive, all individuals within the group are retested individually. The threshold probability used in TOD can be specified beforehand or can be found in an iterative manner using the individual probabilities (details are given in McMahan et al. (2011b)).

Pool-specific optimal Dorfman (PSOD) tries to find the most optimal grouping configuration through the use of the individual probabilities. Individuals are ordered by their probabilities and placed successively into groups. Those individuals with lower probabilities are placed into larger groups, while those individuals with smaller probabilities are placed into smaller groups (or tested individually). Through using a greedy optimization algorithm, the group sizes are automatically found for a prior specified set of individuals. For example, if a laboratory has 100 individuals to test during a day, the algorithm constructs the necessary groups for this "block" of individuals. Individual testing is performed on those groups that test positive.

2.2 Ordered halving

Successively splitting positive groups to decode positive from negative individuals is another commonly used group testing procedure. A specific form of it simply halves positively testing groups. For example, an initial group of size 8 that tests positive can be halved into two groups of size 4. If either sub-group tests positive, individual testing is performed on the positive sub-groups leading to at most three steps in total. Alternatively, a fourth step can be included where a positive sub-group of size 4 is halved again before individual testing. If an odd group size is encountered, as close as possible to equal-sized sub-groups can be created (e.g., split a group of size 9 into sub-groups of size 4 and 5).

Black et al. (2011) propose an informative retesting extension of halving that they call *ordered halving*. Sub-group halves are formed by minimizing one sub-group's probability of being negative and maximizing the other sub-group's probability of being positive. This is achieved by assigning to the first sub-group those individuals with a probability of positivity less than the median within the entire group; the remaining individuals are assigned to the second sub-group. By splitting a positive group in this way, the hope is that all positive individuals are in the same sub-group as long as possible through the retesting process, which avoids unnecessary group or individual retests. Note that if there are 0 or 1 positives in the initial group, ordered halving leads to the same number of tests as a normal application of halving.

2.3 Informative Sterrett

Sterrett (1957) proposes to start the same way as Dorfman (1943) by testing individuals in an initial set of groups. Rather than retesting all individuals within a positive group, Sterrett's procedure randomly selects individuals one-by-one for retesting until the first positive individual is found. The remaining individuals are pooled again to form a new group. If this new group tests negative, decoding is complete. If this new group tests positive, one-by-one testing is applied again until the second positive is found. This algorithmic process repeats itself until all individuals have been decoded as positive or negative. For most situations, the expected number of tests from Sterrett's procedure is smaller than for Dorfman testing; however, the possible number of steps for Sterrett ($2I - 1$, where I is the group size) can be much greater.

Rather than selecting an individual at random to test from a positive group, Bilder et al. (2010) select individuals in descending order by their individual probabilities of positivity. Thus, the first individual retested has the largest probability of being positive. Bilder et al. (2010) show that sizable

reductions in the expected number of retests can occur with this informative retesting procedure. They call their proposal *full informative Sterrett* (FIS).

Bilder et al. (2010) propose two simplifications of FIS. *One-stage informative Sterrett* (1SIS) proceeds in the same way as FIS until the first positive individual is found. If the new group of the remaining individuals tests positive, individual testing is used to decode these individuals. *Two-stage informative Sterrett* (2SIS) follows FIS until the second positive individual is found; if the group of the remaining individuals tests positive, individual testing is applied. Because the number of positive individuals within a group is usually small in group testing settings, 1SIS and 2SIS can perform almost as well as FIS.

2.4 Informative matrix testing

The group testing procedures discussed so far assign each individual to only one initial group for testing. In contrast, matrix (array) testing assigns individuals to two initial groups. This is done by placing individual specimens into a matrix-like grid and pooling individuals by row and by column. If rows and columns test positive, specimens lying at their intersection are decoded using individual testing. If a row (column) tests positive without a column (row) testing positive, all individuals within the row (column) are decoded using individual testing.

McMahan et al. (2011a) propose two grid layouts for specimens in order to minimize the number of rows and columns testing positive. This is desirable because it reduces the number of individual tests needed in the last testing step. In their *spiral array* layout, the individual with the largest probability is assigned to the (1,1) cell of the grid. Individuals with the 2nd, 3rd, and 4th largest probabilities are then put into the (1,2), (2,2), and (2,1) cells, respectively; thus, "spiraling" around the (1,1) cell. This process continues until the remaining cells of the grid all are filled. In the *gradient array* layout, individuals ordered by their probabilities are assigned to successive columns. Thus, column 1 has the specimens with largest individual probabilities (ordered with the (1,1) cell being the largest), column 2 has the next largest individual probabilities, and so on until all cells are filled.

3 Comparisons

We compare the informative retesting procedures through a combination of data application and Monte Carlo simulation. These comparisons are based on chlamydia and gonorrhea screening performed in the State of Nebraska. Across the state, individuals submit urine or swab specimens at health care clinics, and these specimens are sent to the Nebraska Public Health Laboratory (NPHL) for testing. Along with the specimen, each individual supplies information including their age, gender, and risk behavior. Clinical observations, such as the presence of urethritis, are available as well on each individual. Currently, the NPHL uses individual testing only. Our purpose here is to use their prior collected data to determine how well informative retesting procedures could work in this setting; we make this evaluation through the use of Monte Carlo simulation.

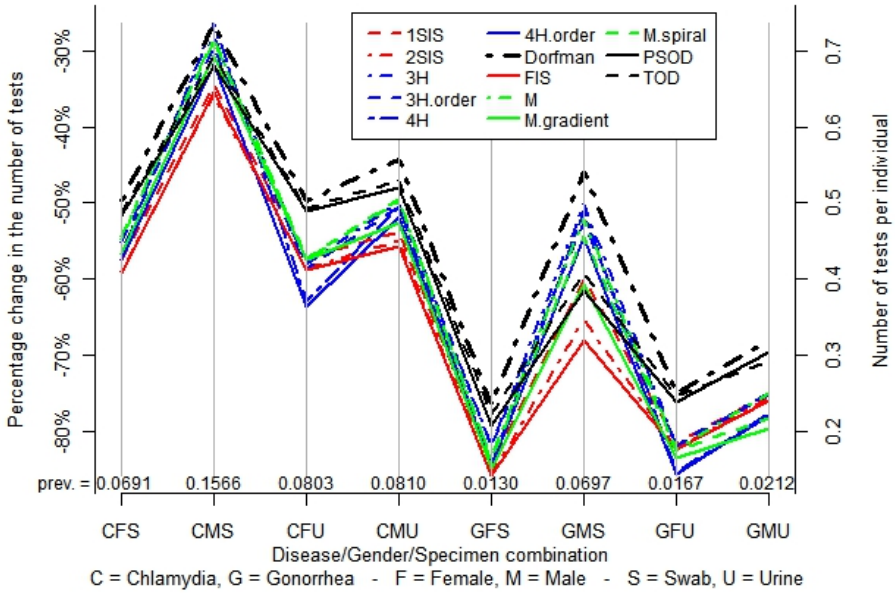
Using the available information on each individual and their diagnoses from testing (assumed absent of error), first-order logistic regression models are estimated to predict the probability an individual has a disease. This model uses training data from the year 2008 (23,146 individuals), and the model is applied to individuals screened in 2009 (27,521 individuals) in order to estimate their probability of having a disease. Separate models are estimated by disease (chlamydia or gonorrhea), gender (male or female), and specimen type (urine or swab) due to the separate assay sensitivity (S_e) and specificity (S_p) levels for each combination (please see Table 1).

Individuals are assigned to groups within their disease/gender/specimen combination, and a maximum group size of 20 is used. For group testing procedures that need a prior specified group size (all except TOD and PSOD), we use the 2008 data to make the best possible determination for what

Table 1: Assay sensitivity and specificity values provided by the NPHL. The table header gives the disease/gender/specimen combination as denoted by their first letters.

	CFS	CMS	CFU	CMU	GFS	GMS	GFU	GMU
S_e	0.928	0.925	0.805	0.930	0.966	0.985	0.849	0.970
S_p	0.960	0.950	0.960	0.950	0.980	0.960	0.980	0.960

Figure 1: Percentage change in the number of tests relative to individual testing. "H" denotes halving where legend names such as "3H.order" denote 3-step ordered halving. "M" denotes matrix pooling. Overall prevalence for a disease/gender/specimen combination is given in the row labeled by "prev".



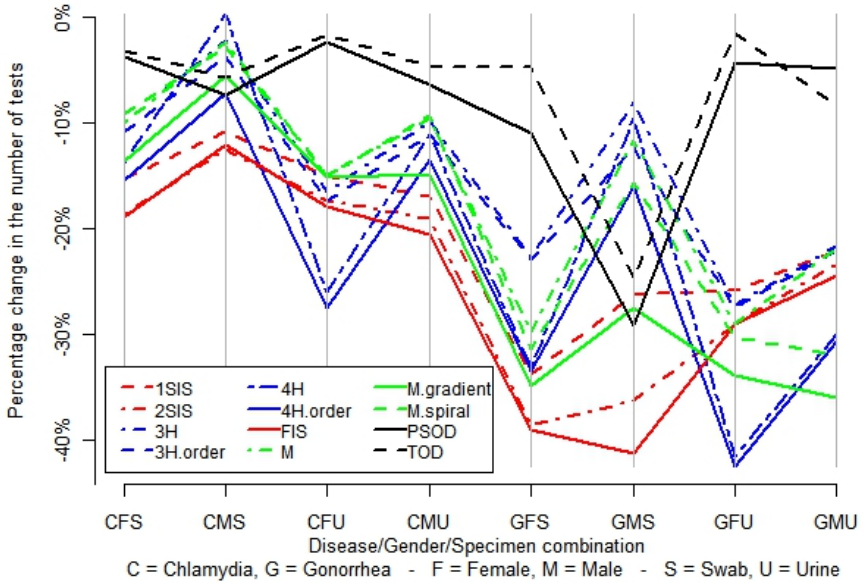
may be optimal. This is done by estimating the expected number of tests for each possible group size and choosing the group size with the lowest value. Also, for these group testing procedures, individuals are put into groups chronologically as would be done generally in practice. With respect to TOD and PSOD, they adaptively choose their group sizes within chronologically formed blocks of 64 individuals.

To simulate the testing results obtained during group testing, observations are generated from a Bernoulli distribution using a success probability parameter of S_e or S_p; the specific parameter used depends on the 2009 diagnosed statuses, which are assumed correct. This process allows us to estimate measures of accuracy for the group testing procedures. Each group testing procedure is then applied using 500 separate simulation runs in order to account for the stochastic nature of the simulation process. We average summary measures, such as the number of tests used and the pooling sensitivity, over these simulation runs.

Figure 1 presents a parallel coordinates plot of the percentage reduction (relative to individual testing) in the number of tests needed to diagnosis each individual. As would be expected, smaller prevalences lead to larger reductions in tests. The group testing procedures that tend to have the largest reduction in tests are FIS and 4-step ordered halving. Dorfman, TOD, and PSOD usually have the smallest reduction in tests. Figure 2 gives a similar account where we have re-scaled each vertical axis to represent the percentage reduction in tests compared to Dorfman testing. This plot helps to display differences among procedures more dramatically.

Figure 3 displays accuracy measures for diagnosing individuals as positive or negative. The pooling sensitivity (specificity) is the probability that an individual is diagnosed as positive (negative) by going through the entire group testing process given that they are truly positive (negative). The

Figure 2: Percentage change in the number of tests relative to Dorfman testing.



pooling positive (negative) predictive value is the probability that an individual is truly positive given they were diagnosed as positive (negative) by going through the entire group testing process. We also provide the actual accuracy measures for individual testing on these plots using dark gray lines.

Overall, group testing has larger pooling specificity and pooling positive predictive values than individual testing, and group testing has smaller pooling sensitivity and pooling negative predictive values than individual testing. This is not necessarily a new finding, and it is due to only one negative test (group or individual) leading to a negative diagnosis and more than one positive test (group or individual) leading to a positive diagnosis. While the 4-step halving procedures tend to produce the most reduction in tests, they are the least accurate with respect to pooling sensitivity and pooling negative predictive values. The FIS, 1SIS, and 2SIS also tend to produce the most reduction in tests, but their accuracy levels are generally in the middle when compared to other procedures. PSOD tends to have the largest accuracy values when compared to Dorfman and TOD. The matrix pooling procedures largely have very similar accuracy measures as do the halving procedures for a particular number of steps. These results help to show that the informative retesting procedures maintain a similar level of accuracy when compared to their non-informative counterparts.

4 Conclusions

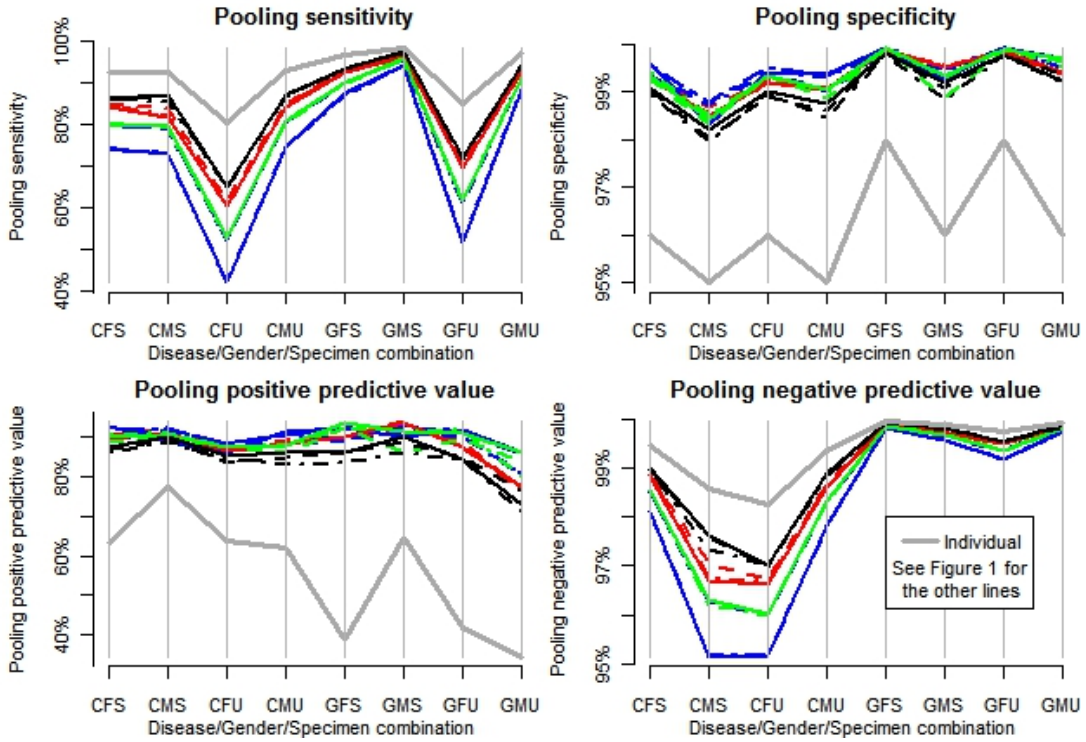
There is not one informative retesting procedure that strictly does better than others. Practitioners can use the results here as general gauge for how a procedure will work for a given disease prevalence, S_e, and S_p. Along with the prevalence, the distribution of the individual probabilities is important too. Due to space restrictions, we were not able to provide histograms here, but they are available at `http://www.chrisbilder.com/grouptesting`. This website also contains additional material on our group testing research.

REFERENCES

Bilder, C., Tebbs, J., and Chen, P. (2010). Informative Retesting. *Journal of the American Statistical Association*, 105:942-955.

Black, M., Bilder, C., and Tebbs, J. (2011). Group Testing in Heterogeneous Populations Using Halving Algorithms. In revision for *Journal of the Royal Statistical Society: Series C (Applied Statistics)*.

Figure 3: Accuracy measures for testing. Note that each plot has a different y-axis scale.



Dodd, R., Notari, E., and Stramer, S. (2002). Current prevalence and incidence of infectious disease markers and estimated window-period risk in the American Red Cross donor population. *Transfusion*, 42:975-979.

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

Kennedy, J., Mortimer, R., and Powers, B. (2006). Reverse transcription-polymerase chain reaction on pooled samples to detect bovine viral diarrhea virus by using fresh ear-notch-sample supernatants. *Journal of Veterinary Diagnostic Investigation*, 18:89-93.

McMahan, C., Tebbs, J., and Bilder, C. (2011a). Informative Array Screening. Under preparation.

McMahan, C., Tebbs, J., and Bilder, C. (2011b). Informative Dorfman Screening. In revision for *Biometrics*.

Mund, M., Sander, G., Pottho, P., Schicht, H., and Matthias, K. (2008). Introduction of Chlamydia Trachomatis screening for young women in Germany. *Journal der Deutschen Dermatologischen Gesellschaft*, 6:1032-1037.

Remlinger, K., Hughes-Oliver, J., Young, S., and Lam, R. (2006). Statistical design of pools using optimal coverage and minimal collision. *Technometrics*, 48:133-143.

Sterrett, A. (1957). On the detection of defective members of large populations. *Annals of Mathematical Statistics*, 28:1033-1036.

ACKNOWLEDGMENT

This research is supported by Grant R01 AI067373 from the National Institutes of Health.