Published in final edited form as:

J R Stat Soc Ser C Appl Stat. 2015 August 1; 64(4): 693–710. doi:10.1111/rssc.12097.

Optimal retesting configurations for hierarchical group testing

Michael S. Black,

Department of Mathematics, University of Wisconsin-Platteville, Platteville, WI 53818, USA,
blackmi@uwplatt.edu

Christopher R. Bilder, and

Department of Statistics, University of Nebraska-Lincoln, Lincoln, NE 68583, USA

Joshua M. Tebbs

Department of Statistics, University of South Carolina, Columbia, SC 29208, USA,
tebbs@stat.sc.edu

SUMMARY

Hierarchical group testing is widely used to test individuals for diseases. This testing procedure works by first amalgamating individual specimens into groups for testing. Groups testing negatively have their members declared negative. Groups testing positively are subsequently divided into smaller subgroups and are then retested to search for positive individuals. In our paper, we propose a new class of informative retesting procedures for hierarchical group testing that acknowledges heterogeneity among individuals. These procedures identify the optimal number of groups and their sizes at each testing stage in order to minimize the expected number of tests. We apply our proposals in two settings: 1) HIV testing programs that currently use three-stage hierarchical testing and 2) chlamydia and gonorrhea screening practices that currently use individual testing. For both applications, we show that substantial savings can be realized by our new procedures.

Keywords

Classification; HIV; Infertility Prevention Project; Informative retesting; Pooled testing; Retesting

1. INTRODUCTION

Screening for diseases, such as HIV, West Nile Virus, and hepatitis C, has benefited greatly from the use of group testing (also known as “pooled testing”). For these applications, specimens (e.g., blood, urine) from separate individuals are amalgamated into a single specimen. Individuals within negative testing groups are declared negative. Individuals within positive testing groups are retested in some predetermined manner to decode the positive individuals from the negative individuals. As long as group sizes are judiciously chosen, group testing can significantly reduce the overall number of tests required which, as a consequence, reduces costs. Group testing has been successfully applied in numerous

settings, including chlamydia and gonorrhea testing as part of the Infertility Prevention Project (Lewis et al. 2012); blood-donor screening for HIV, hepatitis B, and hepatitis C (American Red Cross 2014); and testing known HIV-positive individuals to detect antiretroviral treatment failure (Smith et al. 2009).

Group testing methods are generally divided into hierarchical and non-hierarchical categories, where we focus on hierarchical in this paper. Hierarchical methods divide positive groups into two or more non-overlapping subgroups which are then tested. If any of these subgroups test positively, additional stages of dividing are used until individual testing takes place at the last stage. For any hierarchical application, selecting the number of subgroups, their sizes, and the number of stages are all important decisions to make in order to minimize the number of tests and corresponding costs.

This paper proposes new hierarchical group testing methods by taking advantage of recent advances in group testing collectively known as “informative retesting” (Bilder et al. 2010, McMahan et al. 2012a, McMahan et al. 2012b, Black et al. 2012). Succinctly put, informative retesting algorithms incorporate individual probabilities of positivity into the testing process. To obtain these probabilities, binary regression models are estimated using individual disease statuses and risk factors from a training data set. By exploiting the heterogeneity among the estimated probabilities, we propose a new class of three-stage and higher hierarchical methods that reduce the number of tests needed in comparison to those hierarchical methods which do not use informative retesting. Our procedures identify the optimal number of subgroups and their sizes at each stage, while allowing for multiple stages. No other informative retesting procedure is as flexible in its application.

Hierarchical group testing is already applied in a large number of disease testing settings. In particular, we examine publicly funded HIV testing programs at testing sites throughout the United States. Sherlock et al. (2007) summarizes the algorithms in use at some of these testing sites. For example, the site in Los Angeles initially pools specimens into groups of size 90. If a group tests positively, it is divided into 9 subgroups of size 10. If any of these subgroups test positively, individual testing is then used in the third and final stage. When discussing the different ways that group testing is implemented, Sherlock et al. (2007) remark that “the most efficient approach remains to be determined.” With this as motivation, we apply our methods to determine if they could reduce the number of tests needed to diagnose all individuals.

Our second example focuses on chlamydia and gonorrhea screening activities sponsored by the Centers for Disease Control and Prevention and the Office of Population Affairs in the United States. These activities, which were part of the Infertility Prevention Project (IPP) until 2013, involve testing urine and swab specimens at laboratories across the country. Due to the high volume of specimens, many states use group testing to save money. With each specimen tested, risk factor information, such as gender, sexual history, and clinician observations, are available. This has prompted at least one state, Idaho, to implement the “threshold optimal Dorfman” informative retesting procedure proposed in McMahan et al. (2012a) (Lewis et al. 2012). As we demonstrate herein, our new methods are typically better and are only minimally more complex to implement.

An outline of our paper is as follows. In Section 2, we derive the expected number of tests and measures of classification accuracy for hierarchical group testing when individuals have different probabilities of positivity. Using these derivations, we develop new hierarchical methods that attempt to minimize the expected number of tests. In Section 3, we investigate our methods in a controlled setting to examine the effects of heterogeneity on the expected number of tests and classification accuracy. In Sections 4 and 5, we apply our methods to the HIV and IPP examples, respectively. Finally, in Section 6, we summarize our work and discuss future research.

2. HIERARCHICAL GROUP TESTING

2.1 Expected number of tests

Consider a group of I individuals that are to be screened for a disease using group testing. Define G_{sj} as a binary random variable denoting the test status for group (or subgroup) j at the s^{th} stage, where 0 denotes a negative test result and 1 denotes a positive test result. For example, G_{11} denotes the outcome for the initial group test. The number of individuals within the group corresponding to G_{sj} is defined as I_{sj} , where $I_{11} = I$. If $G_{sj} = 0$, all individuals within the corresponding group are declared negative. If $G_{sj} = 1$, the corresponding group is divided into m_{sj} subgroups for the next stage of testing. Define c_s as the total number possible of subgroups tested at stage s , where $c_1 = 1$ and $c_s = \sum_{j=1}^{c_{s-1}} m_{s-1,j}$ for $s = 2, \dots, K$.

To help explain this notation, consider the three-stage algorithm described in Section 1 for HIV testing in Los Angeles. Specimens from individuals are placed into initial groups of size $I = 90$. If $G_{11} = 0$, all individuals are diagnosed as negatives. If $G_{11} = 1$, the initial group is divided into $m_{11} = 9$ non-overlapping subgroups of size $I_{21} = \dots = I_{29} = 10$ for stage 2 testing. For any subgroup at stage 2 that tests negatively, i.e., $G_{2j} = 0$, the corresponding individuals are declared negative. Any subgroup with $G_{2j} = 1$ is further divided into $m_{2j} = 10$ subgroups of size one. The maximum number of subgroups tested at stage 3 is $c_3 = 90$ because individual testing occurs. We provide additional examples of other hierarchical algorithms in the on-line supplementary materials.

Choosing the number of stages K , the number of subgroups m_{sj} , and the subgroup sizes I_{sj} are all important decisions. Our goal is to develop methods that determine these quantities by minimizing the expected number of tests for a potential application. The expected number of tests for an initial group of I individuals is

$$E(T) = 1 + \sum_{s=1}^{K-1} \sum_{j=1}^{c_s} m_{sj} P \left(\bigcap_{\{(s', j') : G_{s'j'} = 1\}} \{G_{s'j'} = 1\} \right) \quad (1)$$

over K stages, where T is number of tests. This expression arises by noting that m_{sj} more tests are performed whenever the j^{th} group of size $I_{sj} > 1$ tests positively at stage s . The probability in Equation (1) is a joint probability representing a succession of groups testing positively up to and including $G_{sj} = 1$. For example, $E(T)$ for the Los Angeles application is found as

$$\begin{aligned}
E(T) &= 1 + m_{11}P(\{G_{11}=1\}) \\
&\quad + m_{21}P(\{G_{11} \\
&\quad = 1 \cap \{G_{21} \\
&\quad = 1 + m_{22}P(\{G_{11} \\
&\quad = 1 \cap \{G_{22} \\
&\quad = 1\}) \\
&\quad + \cdots + m_{29}P(\{G_{11} \\
&\quad = 1\} \cap \{G_{29}=1\}).
\end{aligned}$$

To find the joint probability in Equation (1), we need to re-express it as a function of the true group statuses $\tilde{G}_{sj}$ to account for testing error. Suppose a single, properly calibrated assay is used to test groups so that the sensitivity $S_e = P(G_{sj}=1|\tilde{G}_{sj}=1)$ and the specificity $S_p = P(G_{sj}=0|\tilde{G}_{sj}=0)$ are constant for all group sizes to be considered. The joint probability in Equation (1) is

$$\begin{aligned}
P\left(\bigcap_{\{(s',j'):G_{sj}=1\}} \{G_{s'j'}=1\}\right) &= (1 - S_p)^s \left\{ \prod_{i=1}^I (1 - p_i) \right\} \\
&\quad + \sum_{a=1}^{s-1} S_e^a (1 - S_p)^{s-a} \left\{ \prod_{i \in B_{a+1,j'}} (1 - p_i) \right\} \left\{ 1 - \prod_{i \in \bar{B}_{a+1,j'}} (1 - p_i) \right\} \\
&\quad + S_e^s \left\{ 1 - \prod_{i \in B_{sj}} (1 - p_i) \right\},
\end{aligned} \quad (2)$$

where p_i is the probability that individual i is truly positive. The notation “ $i \in B_{sj}$ ” is understood to mean those individuals who belong to the j^{th} ordered group at the s^{th} stage, and “ $i \in \bar{B}_{sj}$ ” denotes the set of individuals within the parent group of B_{sj} excluding those in B_{sj} itself. For example, consider again the application in Los Angeles. The notation $i \in B_{29}$ denotes indices of those individuals who belong to the ninth subgroup formed at stage 2. If individuals were placed into groups by numerical order, individuals 81 to 90 would correspond to $i \in B_{29}$, and individuals 1 to 80 would correspond to $i \in \bar{B}_{29}$.

The right-hand side of Equation (2) is written as the sum of three distinguishable

components. The first component represents $P\left(\bigcap_{\{(s',j'):G_{sj}=1\}} \{G_{s'j'}=1\} \cap \{\tilde{G}_{s'j'}=0\}\right)$ i.e., the probability that all groups included in the intersection test positively when they are all truly negative. The third component is the same as the first except that all groups are truly positive. The middle component includes terms within the summand that correspond to a truly positive groups and $s - a$ truly negative groups, for $a = 1, \dots, s - 1$. One will notice that Equation (2) is written the same way as Equation (2) in Black et al. (2012), which examined the special case where positive groups are halved. This equivalence arises because

of our B_{sj} and $\bar{B}_{sj}$ notation. However, unlike Black et al. (2012), we attempt to find the specific hierarchical algorithm which provides the smallest expected number of tests, as described next.

2.2 Optimal retesting configurations

Before an application of group testing begins, we will not necessarily know the *retesting configuration* (i.e., number of subgroups, their sizes, and their members at each stage) that would result in the smallest number of tests. However, we can examine potential retesting configurations before testing begins and choose the one that minimizes $E(T)$. We refer to this configuration as “optimal.” Of course, $E(T)$ depends on the individual probabilities p_i that are likely unknown. In practice, these probabilities would be estimated, and a chosen configuration would then minimize estimates of $E(T)$. For now, we assume all individual probabilities are known. We address estimation in Section 5.

To find the retesting configuration that minimizes $E(T)$, we first order individuals by their probability of positivity within an initial group that tests positively. This helps to isolate those individuals with small and large probabilities while also decreasing the number of possible configurations that need to be examined. Let $p_{(1)} \ p_{(2)} \ \dots \ p_{(I)}$ denote these ordered individual probabilities. Whenever a group tests positively, we assume that individuals are assigned to subgroups successively by this ordering. For example, a group of size $I = 6$ could be divided into $m_{11} = 3$ subgroups of size $I_{21} = 3$, $I_{22} = 2$, and $I_{23} = 1$. The first subgroup in this retesting configuration contains those individuals having probabilities $p_{(1)}$, $p_{(2)}$, and $p_{(3)}$; the second subgroup contains those individuals having probabilities $p_{(4)}$ and $p_{(5)}$; and the third subgroup contains the individual with probability $p_{(6)}$. Ordering in this manner is intuitive because it allows larger (smaller) subgroups to be formed among the low-probability (high-probability) individuals. This, in turn, can help lead to a reduction in the number of tests. For a group of size $I > 1$, we define the *optimal retesting configuration* (ORC) as the configuration which minimizes $E(T)$ when ordered individuals are successively placed into subgroups.

2.2.1 All possible configurations—The most direct approach to find the ORC is to calculate $E(T)$ for all possible configurations and to choose the configuration that minimizes $E(T)$. For a three-stage procedure, the number of configurations equates to the number of ways to choose subgroups at stage 2. Splitting the individuals into b (say) subgroups, while respecting the ordering $p_{(1)} \ p_{(2)} \ \dots \ p_{(I)}$, is equivalent to placing $b - 1$ “partitions” in the $I - 1$ “spaces” between the ordered probabilities; there are ${}^{I-1}_{b-1}C_{b-1}$ ways to do this. Because we could choose b to be any number between 1 and I , the total number of possible configurations for three stages is $\sum_{i=0}^{I-1} {}^{I-1}_{i-1}C_{i-1} = 2^{I-1}$.

To illustrate for a simple case, consider an initial group of size $I = 4$. There are $2^{4-1} = 8$ possible configurations of subgroups at stage 2 with sizes: $[4]$, $[3,1]$, $[2,2]$, $[1,3]$, $[2,1,1]$, $[1,2,1]$, $[1,1,2]$ or $[1,1,1,1]$, where the notation “[$\cdot$]” denotes possible subgroup configurations. For instance, $[3,1]$ means there are two subgroups with individuals corresponding to $p_{(1)}$, $p_{(2)}$, $p_{(3)}$ in the first subgroup and the individual corresponding to $p_{(4)}$

in the second subgroup. Of course, configurations like [3,1] and [1,3] are different because subgroups are formed based on ordered individual probabilities. If needed, a third stage for positive testing subgroups of size two or more leads to individual testing. Note that the enumeration above contains configurations that would not typically be implemented, such as [4] (retest the entire group again), and those that would not allow for a third stage, such as [1,1,1,1]. Therefore, the number of configurations that actually would be considered in practice for three stages in general is less than 2^{I-1} .

Similarly, the number of all configurations involving four stages (the number of ways to choose subgroups at stages 2 and 3) is

$$\sum_{c_2=0}^{I-1} I-1 C_{c_2} \sum_{c_3=0}^{c_2-1} c_2-1 C_{c_3} = \sum_{c_2=0}^{I-1} I-1 C_{c_2} 2^{c_2-1} = 3^{I-1}.$$

For five stages, there are 4^{I-1} possible configurations. Through binomial expansion, this pattern can be shown to continue so that a K -stage procedure has $(K-1)^{I-1}$ possible configurations (see on-line supplementary materials). While the number of configurations that would actually be considered in practice is somewhat less than this value, it can still be very computationally time consuming to list those that are possible. For example, calculation times for $E(T)$ with $I = 12$ are less than one second for $K = 3$, approximately 2.4 minutes for $K = 4$, and approximately one hour for $K = 5$ using R 2.15.0 (R Development Core Team, 2012) and a 2.40 GHZ core of a processor. For very large I , such as $I = 90$ in the Los Angeles example, computational time will be impractical even for $K = 3$.

It is important to note that we have limited the possible configurations to those constructed sequentially with ordered individual probabilities. In addition to this being intuitive, past research has shown that ordering is the preferred choice. For halving algorithms, Black et al. (2012) proved that ordering always produces $E(T)$ as small or smaller than when not ordering. McMahan et al. (2012a) used ordering in applications of two-stage hierarchical group testing. Also for two-stages, Hwang (1975) showed that groups with a larger number of individuals should always have smaller probabilities than groups with fewer individuals.

2.2.2 Steepest descent search algorithm—For some applications, the maximum allowable initial group size will be small due to the size of the overall prevalence, laboratory constraints, or fear that large groups will dilute positive specimens below a detection threshold. In those cases, the all-possible-configuration approach of Section 2.2.1 will work fine. For other applications, these limitations may not be present or not as severe, allowing for large initial groups such as in the Los Angeles example. Examining all possible configurations then may not be possible. To consider these applications, we formulate the optimality problem as an integer program and use the method of steepest descent in an attempt to find the retesting configuration that minimizes $E(T)$. We call this retesting configuration the *candidate retesting configuration* (CRC) to distinguish it from the ORC.

The method of steepest descent begins by first choosing a starting configuration for a specified number of subgroups at each stage. For each possible subgroup pair, we alter the

starting configuration by adding one member to a subgroup and subtracting one member from a different subgroup. We then choose a “better” retesting configuration that has the lowest $E(T)$ among the new ones created. We continue this same process, keeping the number of subgroups the same at each stage, until no other configuration can be found with a smaller $E(T)$. We then repeat this process for all other possible numbers of subgroups. The configuration that minimizes $E(T)$ overall is the CRC. Details regarding the application of this algorithm are given in the on-line supplementary materials.

The CRC and ORC will coincide when $E(T)$ is convex upward as function of the subgroup sizes. Unfortunately, this is not always the case, which could lead to a local minimum being found rather than a global minimum. We give special cases in the on-line supplementary materials where convexity fails. However, despite the absence of convexity in general, we show in Sections 3 and 5 that the CRC results in an $E(T)$ which is the same as or very close to that resulting from the ORC for the cases examined. Therefore, we regard the CRC as a convenient computational alternative to the ORC when all possible configurations cannot be easily enumerated.

2.3 Accuracy measures

In addition to the expected number of tests, the accuracy of correctly classifying truly positive and negative individuals is also important. Define $Y_i = 1(0)$ as the positive (negative) diagnosed status of the i^{th} individual ($i = 1, \dots, I$), and define $\tilde{Y}_i = 1(0)$ similarly to denote the true status. The probability of a correct positive diagnosis, the pooling sensitivity, is $PS_e^{(i)} = P(Y_i = 1 | \tilde{Y}_i = 1)$ for individual i . Similarly, the pooling specificity is $PS_p^{(i)} = P(Y_i = 0 | \tilde{Y}_i = 0)$ for a correct negative diagnosis. We also define the pooling positive predictive value and the pooling negative predictive value as $PPPV^{(i)} = P(\tilde{Y}_i = 1 | Y_i = 1)$ and $PNPV^{(i)} = P(\tilde{Y}_i = 0 | Y_i = 0)$, respectively. These predictive values can be useful once a diagnosis has been made.

For the i^{th} individual to be diagnosed as positive ($Y_i = 1$), the initial group and all subsequent subgroups containing the individual, including the last subgroup which contains only the i^{th} individual, need to test positively as well. Define subgroup j^* in stage L ($L = K$) as the stage where individual i could be tested individually with respect to the configuration. The pooling sensitivity for this individual is

$$PS_e^{(i)} = P(Y_i = 1 | \tilde{Y}_i = 1) = P\left(\bigcap_{\{(s', j') : G_{Lj^*} = 1\}} \{G_{s'j'} = 1\} \mid \bigcap_{\{(\tilde{s}', \tilde{j}') : \tilde{G}_{L\tilde{j}^*} = 1\}} \{\tilde{G}_{\tilde{s}'\tilde{j}'} = 1\}\right) = S_e^L,$$

where we use the standard assumption that test results are conditionally independent once the true status is known (see Litvak et al. 1994, p. 425, p. 430, and Kim and Hudgens 2009, p. 904, for more information). Thus, the pooling sensitivity is the same for each individual testing positively within L stages. Interestingly, this is the same as when all individual probabilities are homogeneous ($p_i = p$ for $i = 1, \dots, I$; see Kim et al. (2007).

The remaining accuracy measures are found similarly. The pooling specificity is

$$\begin{aligned} PS_p^{(i)} &= 1 - P(Y_i=1|\tilde{Y}_i=0) \\ &= 1 - P(\tilde{Y}_i=0)^{-1} \left[P(Y_i=1) - P(Y_i=1|\tilde{Y}_i=1) P(\tilde{Y}_i=1) \right] \\ &= 1 - (1 - p_i)^{-1} \left[P \left(\bigcap_{\{(s',j'): G_{Lj^*}=1\}} \{G_{s',j'}=1\} \right) - S_e^L p_i \right], \end{aligned}$$

where $P \left(\bigcap_{\{(s',j'): G_{Lj^*}=1\}} \{G_{s',j'}=1\} \right)$ is given by Equation (2). Predictive values follow from Bayes' rule:

$$PPPV^{(i)} = \frac{p_i PS_e^{(i)}}{p_i PS_e^{(i)} + (1 - p_i) (1 - PS_p^{(i)})}$$

and

$$PNPV^{(i)} = \frac{(1 - p_i) PS_p^{(i)}}{(1 - p_i) PS_p^{(i)} + p_i (1 - PS_e^{(i)})}.$$

Because $PPPV^{(i)}$ and $PNPV^{(i)}$ are individual specific, these measures could be useful if confirmatory testing was performed to check individual diagnoses (e.g., an individual with a small $PNPV^{(i)}$ could be tested again to ensure they are truly negative).

3. MEAN AND ACCURACY COMPARISONS

In this section, we simultaneously examine how $E(T)$ and measures of classification accuracy are affected by the initial group size, the number of stages, and heterogeneity among the individual probabilities. In addition, we examine how often the ORC and CRC coincide. We continue to assume that p_i is known for this investigation, postponing estimation until Section 5.

Suppose $p_i \sim \text{beta}(\alpha, \alpha(1 - p)/p)$ for $i = 1, \dots, I$, where $\alpha > 0$ and $0 < p < 1$, so that $E(p_i) = p$ and $\text{Var}(p_i) = p^2(1 - p)/(\alpha + p)$. In this family of distributions, note that p represents the overall prevalence for a population. We consider different combinations of p and α to examine how heterogeneity affects $E(T)$ and classification accuracy. Some of these combinations are motivated by the IPP data example in Section 5. Other combinations are chosen because they lead to extreme cases. For example, as $\alpha \rightarrow \infty$, the variance for p_i approaches 0 so that individual probabilities become homogeneous ($p_i = p$ for $i = 1, \dots, I$). Conversely, as α decreases, the variance for p_i increases, which induces more heterogeneity. In fact, one can show that the limiting distribution as $\alpha \rightarrow 0$ is Bernoulli with mean p (McMahan et al. 2012b); that is, $p_i = 1$ with probability p and $p_i = 0$ with probability $1 - p$.

While this specific distribution is unlikely in application, it is still useful to consider because it maximizes the amount of heterogeneity within a group.

Figure 1 shows values of $E(T)/I$, the expected number of tests per individual, for CRCs under different combinations of α , p , K , and I when $S_e = S_p = 0.95$. The on-line supplementary materials list the numerical values for $E(T)/I$ along with the group sizes at each stage. We replaced p_i with $E(p_{(i)})$, the expected value of the ordered individual probability, when calculating $E(T)$ in order to obtain an “average” assessment without simulation error. These values of $E(p_{(i)})$ are calculated in the same manner as described in Black et al. (2012) or through Monte Carlo simulation for the larger initial group sizes. When constructing Figure 1, we found that the ORC and the CRC always were the same for three stages. For four stages, the ORC and CRC always were the same when $I \leq 14$; we did not calculate the ORC for $I > 14$ when $K = 4$ due to the computational reasons outlined in Section 2.2.1.

Within Figure 1, we first note that four stages generally reduce $E(T)/I$ when compared to three stages for the same values of α and p . Exceptions can arise when individual probabilities are homogeneous with a large p or with an initial group size I that is small. Second, larger initial group sizes can greatly reduce $E(T)/I$ when the prevalence is very small (e.g., $p = 0.01$). This finding is consistent with previous research, but it is interesting to note that it persists for the CRC across different levels of heterogeneity and number of stages. Finally, and most importantly, $E(T)/I$ can be greatly reduced as the variability among the p_i values increases (α decreases). This figure clearly shows that our procedures can adeptly exploit differences among individuals as populations become more heterogeneous.

Figure 2 displays the classification accuracy measures described in Section 2.3 for $p = 0.05$ and $S_e = S_p = 0.95$. Plots for $p = 0.01, 0.10$, and 0.15 are in the on-line supplementary materials. When calculating the accuracy measures, we used the same $E(p_{(i)})$ substitution as mentioned above. The most important finding from these plots is that the accuracy measures tend to increase as the variability among the p_i values increases (α decreases). This is consistent across the different initial group sizes. At the same level of heterogeneity (same levels of α), $PS_p^{(i)}$ and $PPPV^{(i)}$ are larger for four stages than for three stages, while the opposite is generally true for $PS_e^{(i)}$ and $PNPV^{(i)}$. The reason why $PS_p^{(i)}$ and $PPPV^{(i)}$ are larger is that positive diagnoses occur only after multiple positive tests. Thus, a more stringent criterion is needed to be diagnosed as positive with four stages. Conversely, $PS_e^{(i)}$ and $PNPV^{(i)}$ tend to be larger for three stages because it takes only one negative test to produce a negative diagnosis.

4. HIV TESTING

Sherlock et al. (2007) describe the use of three-stage hierarchical group testing for HIV testing at different locations in the United States. Our goal is to examine whether their implementations could be more efficient. Table 1 lists the locations with group sizes used at each stage and the overall prevalences observed during the study period. Note that testing in

San Francisco was performed using both two and three stages, but we consider this location to use only three stages for demonstration purposes.

There are two parts of our evaluation. First, we find the retesting configuration that minimizes $E(T)$ for each location by using the same initial group size and by treating the observed prevalence as the true probability of positivity for each individual. This would have been the ideal way to determine the retesting configuration in the absence of our research. We refer to this configuration as being “optimal” under the assumption of homogeneity. Second, we find the CRC and the corresponding $E(T)$ by taking into account potential heterogeneity among individuals at each location. Because the amount of heterogeneity is unknown, we use beta distributions to describe it as done in Section 3. Our choices for α are motivated by the IPP example in Section 5 and by referee suggestions. Because multiple assays are sometimes used at the same location and because different assays are used across locations, we assume $S_e = S_p = 0.99$ for simplicity.

Table 1 presents the results. First, we see that no locations used the optimal second stage under a homogeneity assumption (although some were very close). Second, among different levels of variability and across locations, we compute the percentage reduction in $E(T)$ from using the CRC rather than what would have been optimal assuming homogeneity. For example, Los Angeles has a reduction of 8.6% for $\alpha = 1$ when applying our CRC rather than using 10 groups of size 9 at stage 2. The corresponding CRCs for each location are included in the on-line supplementary materials. Overall, we generally see large reductions in $E(T)$, where the level of reduction increases as the heterogeneity increases. For example, if $\alpha = 0.1$, the reduction is 36.8% for Los Angeles.

While heterogeneity among the individuals being tested would certainly be expected, the actual amount is unknown. However, the levels of variability incorporated into this example are certainly not extreme. For example, for Los Angeles with $\alpha = 0.1$, the minimum and maximum values of $E(p_{(i)})$ are slightly larger than 0 and approximately 0.086, respectively.

5. INFERTILITY PREVENTION PROJECT

To further assess how well the ORC and CRC work in application, we examine a database of previously diagnosed individuals for chlamydia and gonorrhea in Nebraska for the IPP. Based on these data, we perform Monte Carlo simulations to estimate the number of tests that our own and other group testing methods would use in application. We also estimate measures of accuracy, which will help practitioners understand the classification characteristics of each method.

5.1 Data and models

We begin by estimating the probability of disease positivity for each individual using logistic regression. As would be done in practice, these estimated probabilities are used in place of the true probabilities when implementing informative retesting. The covariates used in the regression models describe clinical observations (symptoms, cervical friability, pelvic inflammatory disease, cervicitis, urethritis), demographic variables (age, race), and risk behavior (multiple partners, new partner in the last 90 days, contact with an individual who

has a sexually transmitted disease). Each covariate is binary except for age. The response variable is the observed disease diagnosis. We treat diagnoses as the true disease statuses when fitting regression models so that we can assess accuracy.

Separate models are estimated for each disease (chlamydia or gonorrhea), gender (male or female), and specimen type (urine or swab) combination. We estimate the models based on a training data set that includes 23,146 individuals who were tested in 2008 (parameter estimates are given in the on-line supplementary materials). These models are then applied to a test data set, which includes 27,521 individuals screened in 2009, in order to obtain individual estimates of positivity. Table 2 summarizes these individuals. Group testing is then applied to the individuals from 2009, where individuals are placed into initial groups chronologically by specimen date.

Because we would like to obtain measurements of accuracy, all group, subgroup, and individual responses that would be observed are simulated, assuming the actual observed individual responses as the true statuses. These binary responses are simulated from a Bernoulli distribution with the appropriate value of S_e and $1 - S_p$ as the probability parameter. For instance, if a group of size four has all negative responses in the test data set, we simulate a response for the group from a Bernoulli($1 - S_p$) distribution. If instead at least one of the four was positive, we simulate from a Bernoulli(S_e) distribution. The values for S_e and S_p are given in Table 2; these are for the assay used in Nebraska during this time period.

5.2 Group testing methods

We compare our proposed methods in this paper to other hierarchical group testing methods. First, we include Dorfman's (1943) original two-stage procedure (hierarchical group testing with $K = 2$ assuming equal individual probabilities) because it is the easiest method to apply. We also include the informative retesting procedure known as pool-specific optimal Dorfman (PSOD) proposed by McMahan et al. (2012a). PSOD is also a two-stage procedure that tries to minimize the estimated expected number of tests by strategically placing individuals within first-stage groups. We do not include the threshold optimal Dorfman method mentioned in Section 1 because McMahan et al. (2012a) showed it did not perform as well as PSOD.

We also include competing three- and four-stage hierarchical methods. First, we apply the informative retesting procedure known as "ordered halving" proposed by Black et al. (2012). It is limited by only being able to divide positive testing groups in half. Individuals are placed into subgroups that maximize one subgroup's probability of testing positively, while minimizing the other subgroup's probability of testing positively. We abbreviate ordered halving by OH3 and OH4 for three and four stages, respectively. Second, similar to Section 4, we find the retesting configuration that minimizes the estimated expected number of tests assuming homogeneity among the individuals. These hierarchical methods are abbreviated as H3 and H4 for three and four stages, respectively.

With respect to our proposed methods, there are two ways that the ORC or CRC can be found in practice. First, one can find a retesting configuration separately for *each* positive testing initial group. We refer to this as an "adaptive" procedure (A-ORC, A-CRC). Second,

a simpler approach is to estimate one *overall* retesting configuration from a training data set and then apply it to new specimens. One can implement this approach by first finding $\hat{p}_{(1)h}, \dots, \hat{p}_{(I)h}$, where the subscript “ $(i)h$ ” allows us to denote ordered probabilities within some group h , for all possible groups of the same size I in the training data (we cap I at 20 in this section). The ORC and CRC are found using the probabilities averaged over all groups, $\bar{\hat{p}}_{(1)}, \dots, \bar{\hat{p}}_{(I)}$, say, for the training data. Because only one configuration is applied to new specimens (those in our test data here), we refer to this as a “non-adaptive” procedure (N-ORC, N-CRC). In the results that follow, we denote the maximum number of stages by appending K to the end of each acronym; e.g., three-stage, non-adaptive ORC is denoted by N-ORC3. Note that we did not include N-ORC4 or any four-stage adaptive procedures in our study because of excessive computing time.

To ensure a fair comparison, we use the optimal initial group size for each method by minimizing the estimated expected number of tests in the training data for group sizes 5, 6, ..., 20 and then use that as the initial group size for the test data. While groups only up to size 10 have been used for chlamydia and gonorrhea screening (see the references within the review paper of Mund et al. 2008), we decided to increase our maximum size to 20 in order to determine if benefits may occur from a larger group size. We would like to emphasize that the estimated expected number of tests are found *only* when working with the training data, as would likely be done in practice. Note that PSOD does not have one overall initial group size due to the nature of the method itself.

5.3 Monte Carlo simulations

To account for the variation introduced by simulating responses, we repeat the application of group testing over 500 data sets. Tables 3 and 4 provide summaries of our results for chlamydia and gonorrhea testing, respectively. Within the tables, the mean column provides the average number tests used across the simulation runs, while the SD column provides the standard deviation for the number of tests. The PS_e , PS_p , $PPPV$, and $PNPV$ columns provide the corresponding *observed* accuracy measures (Altman and Bland 1994a, 1994b) as averaged over the simulated data sets. For example, the observed PS_e for one simulated data set is the observed proportion of true positives that are diagnosed to be positive by a group testing method.

Within the tables, we first note that all three- and four-stage methods provide dramatic reductions in the average number of tests in comparison to Dorfman-based methods. The one exception is for PSOD when testing male swabs, where the average number of tests is similar to our proposed methods. Second, with respect to the accuracy measures, Dorfman-based methods have higher PS_e and slightly higher $PNPV$ than other methods but lower values of PS_p and $PPPV$. These findings are not new in the comparison of hierarchical procedures (e.g., see Kim et al. 2007 and McMahan et al. 2012a), and they are simply due to the number of stages as discussed in Section 3. Third, the standard deviations given in the tables can be used to obtain a rough range for the number of tests observed with each simulated data set (e.g., Mean $\pm$ 3SD). Additional summary measures (minimum, maximum, and percentiles) are given in the on-line supplementary materials. Although there is

variability in the number of tests, we did not find a pattern that would suggest certain methods are consistently more precise.

Specifically when compared to H3, ORC3 and CRC3 generally reduce the average number of tests, where reductions can be as much as 22.0% (testing males for gonorrhea with swabs). In the two cases where H3 has a smaller average, it is by a relatively small amount. In comparison to H4, CRC4 always reduces the average number of tests, where the largest reduction is 27.9% (testing males for gonorrhea with swabs). With respect to accuracy, the ORC and CRC almost always provide a larger PS_e and $PNPV$ than those for H3 and H4. In contrast, the ORC and CRC almost always result in a smaller PS_p and $PPPV$. When compared to ordered halving, the ORC and CRC reduce the average number of tests with similar or larger accuracy in almost all cases.

Interestingly, the differences in the average number of tests among corresponding adaptive and non-adaptive, three-stage procedures are small and are not consistently positive or negative (largest relative difference is 2.7% when testing males for gonorrhea using urine). The accuracy levels are similar too. These results suggest that the non-adaptive approach may be preferred due to its ease in application. Furthermore, there are only very small differences between corresponding applications of the ORC and CRC. This suggests that the CRC may be preferred to the ORC when the ORC takes an excessive amount of time to compute.

6. DISCUSSION

We have demonstrated that the ORC and CRC can significantly reduce the number of tests for hierarchical group testing while maintaining classification accuracy. We also showed two ways that our proposed methods can be implemented, either adaptively or non-adaptively, and we obtained similar results for each. This is especially important because implementing the non-adaptive approach allows practitioners to use only one retesting configuration throughout the screening process. Of course, our evaluations in Sections 3, 4, and 5 are limited to the specific cases we considered. However, we have purposely used real-life scenarios in our evaluations. We expect our conclusions to be similar at least in other disease-testing applications.

Finding the ORC over three stages is closely related to the PSOD procedure described in Section 5. Essentially, the last two stages of a three-stage ORC resemble PSOD. For a positive testing group at stage two of the ORC procedure (or stage one of PSOD), one allocates individuals to subgroups to minimize the expected number of remaining tests. Stage three for ORC (or stage two of PSOD) concludes with individual testing. The advantage that ORC has over PSOD is that ORC's first stage can immediately diagnose all I individuals as negative. If the initial group tests positively for ORC, PSOD will always result in an expected number of tests greater than or equal to the remaining tests of ORC (this is not necessarily true for CRC). This is because finding the ORC requires an examination of all possible configurations, whereas the greedy search algorithm used by PSOD is not guaranteed to find the optimal configuration.

While the ORC and CRC often result in the same expected number of tests, future research could explore other ways to find the optimal configuration, especially for large initial group sizes. In particular, the enumeration of all possible configurations falls into a setting commonly known as “embarrassingly parallel,” so gains in computational time potentially could be achieved through high performance computing. Also, a genetic search algorithm could potentially be used to search among “good” configurations that lead to new generations of “better” configurations. Finally, as suggested by a referee, Hwang (1981) and Hwang and Rothblum (2012) present examples of how three-stage hierarchical testing, in the absence of testing error, can be envisioned as an ordered optimal partition problem. This could potentially lead to determining an optimal configuration in our context.

R functions to compute the ORC and CRC are provided at www.chrisbilder.com/grouptesting/BBT. We include a program at this web address showing how the computations in Sections 3 and 4 can be performed. Our methodology can be applied in practice by simply inputting the estimated individual probabilities within a group into our functions along with S_e , S_p , and K .

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGEMENTS

This research was supported by Grant R01 AI067373 from the National Institutes of Health. We thank the Editor, Associate Editor, and an anonymous referee for their comments which helped us to improve this paper considerably.

REFERENCES

- Altman D, Bland J. Diagnostic tests 1: Sensitivity and specificity. *BMJ*. 1994a; 308:1552. [PubMed: 8019315]
- Altman D, Bland J. Diagnostic tests 2: Predictive values. *BMJ*. 1994b; 309:102. [PubMed: 8038641]
- American Red Cross. Blood testing. 2014. <http://www.redcrossblood.org/learn-about-blood/what-happens-donated-blood/blood-testing>, retrieved Sept. 27, 2014
- Bilder C, Tebbs J, Chen P. Informative retesting. *Journal of the American Statistical Association*. 2010; 105:942–955. [PubMed: 21113353]
- Black M, Bilder C, Tebbs J. Group testing in heterogeneous populations by using halving algorithms. *Journal of the Royal Statistical Society: Series C*. 2012; 61:277–290.
- Dorfman R. The detection of defective members of large populations. *Annals of Mathematical Statistics*. 1943; 14:436–440.
- Hwang F. A generalized binomial group testing problem. *Journal of the American Statistical Association*. 1975; 70:923–926.
- Hwang F. Optimal partitions. *Journal of Optimization Theory and Applications*. 1981; 34:1–10.
- Hwang, F.; Rothblum, U. *Partitions: Optimality and Clustering: Single-Parameter*. World Scientific; Singapore: 2012.
- Kim H, Hudgens M, Dreyfuss J, Westreich D, Pilcher C. Comparison of group testing algorithms for case identification in the presence of test error. *Biometrics*. 2007; 63:1152–1163. [PubMed: 17501946]
- Kim H, Hudgens M. Three-dimensional array-based group testing algorithms. *Biometrics*. 2009; 65:903–910. [PubMed: 19053996]

- Lewis J, Lockary V, Kobic S. Cost savings and increased efficiency using a stratified specimen pooling strategy for Chlamydia trachomatis and Neisseria gonorrhoeae. Sexually Transmitted Diseases. 2012; 39:46–48. [PubMed: 22183846]
- Litvak E, Tu X, Pagano M. Screening for the presence of a disease by pooling sera samples. Journal of the American Statistical Association. 1994; 89:424–434.
- McMahan C, Tebbs J, Bilder C. Informative Dorfman screening. Biometrics. 2012a; 68:287–296. [PubMed: 21762119]
- McMahan C, Tebbs J, Bilder C. Two-dimensional informative array testing. Biometrics. 2012b; 68:793–804. [PubMed: 22212007]
- 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; 6:1032–1037. [PubMed: 18479502]
- R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing; Vienna, Austria: 2012. <http://www.R-project.org>. ISBN 3-900051-07-0
- 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. [PubMed: 17483725]
- Smith D, May S, Perez-Santiago J, Strain M, Ignacio C, Haubrich R, Richman D, Benson C, Little S. The use of pooled viral load testing to identify antiretroviral treatment failure. AIDS. 2009; 23:2151–2158. [PubMed: 19730348]

Figure 1.

Expected number of tests per individual for the CRC using $E(p_{(i)})$ from beta distributions when $S_e = S_p = 0.95$. Lines for $K = 3$ have twice the thickness as lines for $K = 4$. Note that the vertical axis scales are not the same for each plot.

Figure 2.

Accuracy measures for the CRC calculated using $E(p_{(i)})$ from beta distributions when $p = 0.05$ and $S_e = S_p = 0.95$. Lines for $K = 3$ have twice the thickness as lines for $K = 4$. Note that the vertical axis scales are not the same for each plot.

Table 1

Three-stage hierarchical group testing for HIV testing. The prevalences, initial group sizes, and actual second stage group sizes are obtained from Table 1 of Sherlock et al. (2007). The optimal second stage group sizes are calculated assuming every individual has a probability of positivity equal to the prevalence (i.e., homogeneity) at a particular location. The last four columns present the percentage reduction in $E(T)$ obtained by using the CRC rather than the optimal second stage under homogeneity, where α is a beta distribution parameter controlling the amount of heterogeneity. All third stages involved individual testing.

	Prevalence (p)	Initial group size	Under homogeneity		Reduction in $E(T)$ from CRC			
			Actual second stage	Optimal second stage	$\alpha = 2$	$\alpha = 1$	$\alpha = 0.5$	$\alpha = 0.1$
North Carolina	0.0021	90	9 groups of size 10	10 groups of size 9	4.2%	7.7%	13.6%	33.1%
Los Angeles	0.0045	90	9 groups of size 10	10 groups of size 9	4.4%	8.6%	15.2%	36.8%
San Francisco	0.0175	50	5 groups of size 10	2 groups of size 7 6 groups of size 6	4.7%	8.4%	15.1%	37.4%
Seattle-King County	0.0164	30	3 groups of size 10	6 groups of size 5	3.5%	7.2%	12.2%	32.0%
Atlanta	0.0030	48	6 groups of size 8	6 groups of size 7 1 group of size 6	3.2%	6.5%	11.1%	27.3%

Table 2

Summary statistics for chlamydia and gonorrhea testing in 2009. The overall observed prevalence is denoted by $\bar{p}$. The maximum likelihood estimates (MLE) for α and p are found by fitting beta distributions to all estimated individual probabilities.

Disease	Gender	Specimen	Count	$\bar{p}$	S_e	S_p	MLE of α	MLE of p
Chlamydia	Female	Swab	14,503	0.069	92.8%	96.0%	2.5	0.067
		Urine	4,970	0.080	80.5%	96.0%	1.1	0.087
	Male	Swab	1,909	0.157	92.5%	95.0%	1.0	0.149
		Urine	6,139	0.081	93.0%	95.0%	1.8	0.090
Gonorrhea	Female	Swab	14,503	0.013	96.6%	98.0%	0.5	0.011
		Urine	4,970	0.017	84.9%	98.0%	0.5	0.018
	Male	Swab	1,909	0.070	98.5%	96.0%	0.4	0.077
		Urine	6,139	0.021	97.0%	96.0%	0.2	0.014

Table 3

Mean number of tests, standard deviation (SD) for the number of tests, and mean accuracy measures for Nebraska IPP chlamydia screening based on 500 simulated data sets. The total number of individuals screened is given in Table 2.

Female/Urine								Male/Urine						
Method	I	Mean	SD	PS _e	PS _p	PPPV	PNPV	I	Mean	SD	PS _e	PS _p	PPPV	PNPV
Dorfman	5	2,495	44	65.0%	99.0%	84.6%	97.0%	5	3,418	42	86.6%	98.5%	83.2%	98.8%
PSOD	NA	2,436	45	65.1%	99.0%	85.2%	97.0%	NA	3,197	39	87.4%	98.8%	86.1%	98.9%
H3	12	1,996	59	52.1%	99.4%	88.3%	96.0%	9	3,039	40	80.5%	99.2%	90.3%	98.3%
OH3	9	2,059	55	52.3%	99.3%	87.2%	96.0%	7	3,029	40	80.3%	99.2%	89.5%	98.3%
N-CRC3	16	2,031	65	52.4%	99.4%	88.2%	96.0%	8	2,886	41	82.2%	99.2%	89.6%	98.4%
N-ORC3	16	2,026	66	52.2%	99.4%	88.2%	96.0%	8	2,890	41	82.3%	99.2%	89.5%	98.5%
A-CRC3	16	2,049	64	52.8%	99.3%	86.7%	96.0%	9	2,907	42	82.2%	99.1%	89.4%	98.5%
A-ORC3	16	2,046	66	52.7%	99.3%	86.9%	96.0%	9	2,913	42	82.3%	99.2%	89.5%	98.5%
H4	20	1,801	74	41.9%	99.6%	90.0%	95.1%	18	2,994	55	75.0%	99.3%	90.6%	97.8%
OH4	19	1,820	69	42.1%	99.4%	86.8%	95.2%	13	2,939	54	75.0%	99.3%	90.5%	97.8%
N-CRC4	20	1,747	69	42.0%	99.6%	90.7%	95.2%	16	2,862	52	76.9%	99.3%	90.7%	98.0%

Female/Swab								Male/Swab						
Method	I	Mean	SD	PS _e	PS _p	PPPV	PNPV	I	Mean	SD	PS _e	PS _p	PPPV	PNPV
Dorfman	5	7,292	60	86.1%	99.0%	86.2%	99.0%	4	1,403	22	85.5%	98.0%	88.7%	97.3%
PSOD	NA	7,014	57	86.4%	99.1%	87.6%	99.0%	NA	1,299	22	87.2%	98.2%	90.1%	97.6%
H3	9	6,366	62	79.9%	99.5%	92.2%	98.5%	9	1,359	31	79.0%	98.7%	91.7%	96.2%
OH3	7	6,388	58	79.8%	99.5%	91.5%	98.5%	7	1,371	27	79.1%	98.5%	90.9%	96.2%
N-CRC3	8	6,237	58	81.6%	99.4%	91.0%	98.7%	9	1,295	30	81.9%	98.4%	90.6%	96.7%
N-ORC3	8	6,235	57	81.6%	99.4%	91.1%	98.7%	9	1,292	28	81.8%	98.4%	90.7%	96.7%
A-CRC3	9	6,110	58	81.1%	99.5%	92.0%	98.6%	8	1,286	28	82.1%	98.5%	90.8%	96.8%
A-ORC3	9	6,111	61	81.1%	99.5%	92.0%	98.6%	8	1,286	30	82.1%	98.5%	90.9%	96.8%
H4	18	6,262	88	74.2%	99.6%	92.4%	98.1%	20	1,363	49	73.1%	98.8%	92.1%	95.2%
OH4	13	6,171	71	74.1%	99.5%	92.3%	98.1%	17	1,258	44	73.1%	98.6%	90.4%	95.2%
N-CRC4	16	5,970	80	76.3%	99.5%	92.4%	98.3%	20	1,252	45	76.4%	98.8%	92.1%	95.8%

Table 4

Mean number of tests, standard deviation (SD) for the number of tests, and mean accuracy measures for Nebraska IPP gonorrhea screening based on 500 simulated data sets. The total number of individuals screened is given in Table 2.

Method	I	Female/Urine					Male/Urine				
		Mean	SD	PS_e	PS_n	PPPV	I	Mean	SD	PS_e	PS_n
Dorfman	8	1,238	36	71.8%	99.8%	84.4%	10	1,958	50	94.1%	99.2%
PSOD	NA	1,185	42	72.3%	99.8%	84.5%	NA	1,862	55	94.5%	99.2%
H3	20	810	35	61.4%	99.9%	91.1%	20	1,417	32	91.3%	99.6%
OH3	13	861	32	61.0%	99.9%	89.7%	15	1,523	30	91.2%	99.5%
N-CRC3	20	769	32	61.2%	99.9%	92.6%	20	1,407	35	92.1%	99.6%
N-ORC3	20	766	34	61.5%	99.9%	92.6%	20	1,410	34	92.0%	99.6%
A-CRC3	20	783	33	62.5%	99.9%	91.6%	20	1,445	37	92.2%	99.5%
A-ORC3	20	783	34	62.0%	99.9%	91.6%	20	1,447	37	92.2%	99.5%
H4	20	721	33	52.5%	100.0%	94.2%	20	1,307	27	88.4%	99.8%
OH4	19	722	33	52.3%	99.9%	92.6%	19	1,293	25	88.5%	99.7%
N-CRC4	20	699	31	52.3%	100.0%	95.8%	20	1,265	28	90.2%	99.7%

Method	I	Female/Swab					Male/Swab				
		Mean	SD	PS_e	PS_n	PPPV	I	Mean	SD	PS_e	PS_n
Dorfman	10	3,385	52	93.4%	99.8%	83.6%	5	1,035	17	97.1%	98.8%
PSOD	NA	3,012	62	93.5%	99.8%	86.2%	NA	733	24	97.8%	99.2%
H3	20	2,316	34	90.1%	99.9%	92.4%	9	935	14	95.6%	99.4%
OH3	15	2,543	33	90.3%	99.9%	89.1%	7	861	13	95.5%	99.5%
N-CRC3	20	2,119	35	91.0%	99.9%	92.5%	13	734	19	96.5%	99.4%
N-ORC3	20	2,116	35	90.9%	99.9%	92.5%	13	734	19	96.4%	99.4%
A-CRC3	20	2,074	36	91.2%	99.9%	93.5%	13	730	19	96.3%	99.4%
A-ORC3	20	2,072	37	91.0%	99.9%	93.5%	13	729	18	96.4%	99.4%
H4	20	2,165	29	87.3%	99.9%	95.2%	9	945	15	95.2%	99.5%
OH4	19	2,170	32	87.2%	99.9%	93.8%	15	818	15	94.1%	99.5%
N-CRC4	20	1,963	27	90.1%	99.9%	93.9%	20	681	15	95.9%	99.5%