Human or Cylon?

Group testing on Battlestar Galactica

Christopher R. Bilder
Department of Statistics
University of Nebraska-Lincoln
Lincoln, NE 68583
chris@chrisbilder.com
http://www.chrisbilder.com

1 Introduction

The statistical science has made significant contributions to medicine, engineering, social science, agriculture, and a multitude of other important areas. Does statistics have a place though in the world of science fiction? Because science fiction writers try to merge the sci-fi world with the real world in a believable way, one might think that statistics could make a significant contribution to solving sci-fi problems. After all, many science fiction works already rely on science to rescue characters from the brink of disaster. In the hit Sci Fi Network television show, the new Battlestar Galactica (a re-imagined version of the 1970's show), there is an attempt to use science to solve a very important problem. Due to the excessive amount of time the proposed solution would take to complete, it is deemed impractical and never implemented. This paper shows how the problem could have been solved instead using a statistical technique called "group testing." Scientists use this technique to solve many real-world problems, including the screening of blood donations for diseases. When applied to the problem on Battlestar Galactica, it will be shown that group testing could have a made dramatic difference to the course of the show.

2 Battlestar Galactica

The Emmy and Peabody award-winning Battlestar Galactica television show has been rated one of the best science fiction shows of all time – #2 since 1982 by *Entertainment Weekly* – and even one of the best among all television shows – #8 in 2008 and top 100 of all time by *Time*, top 10 in 2008 by the *Chicago Tribune*, and one of the American Film Institute's programs of the year in 2005 and 2006. The show is about the struggle between humans and Cylons in a distant part of our galaxy. Cylons are cybernetic life forms originally created by humans. These Cylons evolved and rebelled against the humans by destroying their home planets. Approximately 47,000 humans survived the Cylon attack, and they all banded together in a ragtag fleet of spaceships to escape from the Cylons. The fleet is led by a military Battlestar spaceship named Galactica, which is the source of the show's name.

There are two different types of Cylons. One type has a metallic form, such as the "Centurion" on the left of Figure 1. Earlier versions of Centurions were created by the humans to be workers and soldiers. Eventually, the Cylons themselves secretly created a new humanoid form of a Cylon, and one model is shown on the right of Figure 1. This new form played an important role in the almost complete destruction of humanity due to it being indistinguishable from humans.

Early on, while fleeing from the pursuing Cylons, the humans discover the existence of the new humanoid form of a Cylon, and their top priority becomes figuring out how to distinguish a human from a Cylon. The person charged with developing a "Cylon detector" is Dr. Gaius Baltar. Fortunately for him, the number of Cylons in the fleet is expected to be small, but all 47,905 individuals in the fleet must be tested. Baltar creates a blood test for Cylon indicators, and in the episode “Tigh Me Up, Tigh Me Down” he says a single test will take 11 hours to complete. Extrapolating to include everyone in the fleet, Baltar says it will take about 61 years to complete all of the testing! (From Baltar’s calculations, one can deduce that the fleet, like Earth, observes a 24-hour day and 365-day year.)

Baltar planned to use “individual testing.” This involves testing each individual blood specimen one at a time for Cylon indicators. The obvious problem with this testing strategy is that it will take a very long time. Another problem is these humans have very limited resources. When the surviving humans left their home planets, they were fleeing for their lives, so they did not have time to pack supplies. Overall, the Cylon testing needs to be done quickly, while using as little resources as possible.

3 Group testing

The “Tigh Me Up, Tigh Me Down” episode was the ninth episode of the series that lasted for over 70 episodes from 2004-2009. Also, this was the last episode where testing for Cylon indicators is mentioned, and the testing is never carried out. Perhaps the writers wanted to put Baltar into an impossible situation. Alternatively, perhaps the writers did not consult with scientists to merge the sci-fi world with the real world in this instance, because individual testing would not be used in the real world to solve this type of problem. Rather, scientists would use group testing.

Group testing (also known as pooled testing) is used in a wide variety of real-world applications already – see Sidebar 1 for a partial list. In this situation, group testing would begin by putting each individual into a group. Within a group, parts of each individual’s specimen are composited together, so that one test can be performed on it. If the composited specimen tests negative for Cylon indicators, all individuals within the group are declared to be negative. If the composited specimen tests positive for Cylon indicators, there is at least one Cylon in the group and there are a number of retesting procedures that can be implemented to find the positive(s). The potential advantage to using group testing is a smaller number of tests will be required with less resources expended. In general, these advantages occur when the overall prevalence of the trait of interest (e.g., being a Cylon here or having a particular disease in other applications) in a population is small. Otherwise, if the prevalence is large, one may have a great number of groups test positive resulting in potentially a large number of retests.

There are a number of retesting procedures that can be used to decode a positive group. The easiest and most used procedure is one originally proposed by Robert Dorfman in his 1943 *Annals of Mathematical Statistics* article, which suggests using group testing for syphilis screening of American soldiers during World War II. Simply, Dorfman's procedure retests each individual within a positive group to determine a diagnosis. Overall, an initial group of size I that tests positive would result in $I + 1$ tests in total. While this procedure typically results in significant savings when compared to individual testing, there are many other procedures which can do even better.

Andrew Sterrett's 1957 work in the *Annals of Mathematical Statistics* proposes a different retesting procedure, which leads to a smaller expected number of tests than Dorfman's procedure. For an initial positive group, the procedure begins by randomly retesting individuals until a first positive is found. Once found, the remaining individuals are pooled to form a new group. If this new group tests negative, the decoding process is complete, and the individuals in the new group are declared negative. Because group testing is used in low overall prevalence situations, having only one positive is a likely event for a reasonably chosen initial group size. If the new group tests positive though, the process begins again by randomly retesting individuals from this group until a second positive is found. Once the second positive is found, the remaining individuals again are pooled together to determine if more positives remain. The whole process of individually testing and repooling is repeated until no more retests are positive.

Another frequently used procedure involves creating subsets of a positive initial group. In one application, Eugene Litvak, Xin Tu, and Marcello Pagano in a 1994 *Journal of the American Statistical Association* article propose to subsequently halve positive groups until all individual positives have been found. For example, suppose an initial group of size eight tests positive. This group is divided into two halves of size four. Any new group that tests positive is subsequently divided into pools of size two. Finally, the remaining positive groups result in individual tests.

All of the above procedures assume an individual is initially assigned to one group only and retesting is performed in a hierarchical manner. There are other non-hierarchical testing procedures as well. In particular, Ravi Phatarfod and Aidan Sudbury propose in a 1994 *Statistics in Medicine* article to place specimens into a matrix-like grid for their matrix pooling procedure. Specimens are pooled within each row, and specimens are pooled within each column. Positive individuals occur at the intersection of positive rows and columns. When more than one row and column test positive within a single grid, individuals at the intersections are further individually tested to complete the decoding. Matrix pooling is especially useful in high throughput screening. In this situation, specimens are arranged into a matrix grid of wells on a plate so that robotic arms can do the pooling automatically.

With many of these procedures, modifications must be made to implement them in practice. For example, the Sterrett and matrix pooling procedures could lead to an individual being declared positive without ever having been tested individually. For example, this can happen if a positive group contains only one positive individual, and this individual is the last to be selected using Sterrett's procedure. For this reason, it is better to add an extra test for this individual rather than declaring it to be positive by default. Also, when using the halving procedure, a group not of a size evenly divisible by two can still be tested. The successive dividing of positive groups may lead to an individual being tested at an earlier stage. For example, a group of size seven can be divided into groups of size four and three. For the group of size three, a further subdividing leads to one group of size two and a separate individual. Finally, when performing matrix pooling, there may be a 10×10 grid available for testing, but 122 individuals need to be tested. This works well for the first 100, but not for the last 22. These last individuals can be tested in two rows of size 10 and Dorfman's procedure can be performed on the last two.

4 Cylon detection

When Baltar started his testing, the show's fans now know there were most likely seven Cylons out of the 47,905 individuals in the fleet, and all of these Cylons were unknown to Baltar. Taking this seven as the true count, the overall prevalence becomes approximately 0.0001461, which meets the criteria of a low prevalence for group testing to be useful. In comparison to a real-world example, the HIV prevalence for American Red Cross blood donations was 0.00009754 in 2001, when Dorfman's procedure was used for screening purposes.

Baltar makes a number of implicit assumptions with his testing. First, Baltar assumes all of the individual testing needs to be done back-to-back. Of course, if Baltar was able to run individual tests on multiple specimens simultaneously, this would greatly reduce the testing time. Perhaps due to the limited resources available to him (a nuclear warhead was needed to create his Cylon detector), only back-to-back testing was mentioned. Back-to-back testing is used here as well when implementing group testing. Second, Baltar never discusses the possibility of testing error (i.e., the test diagnosis is incorrect). In the real world, one usually needs to account for testing error through knowing the sensitivity and the specificity of a diagnostic test. Perhaps though Baltar did develop a Cylon detector absent of error; after all, he is considered to be a genius by people in the fleet. No testing error is assumed here when implementing group testing, but it is further discussed in Section 5.

For the four different group testing procedures discussed in Section 3 and a given group size, Figure 2 shows the expected number of tests for each procedure. While Dorfman's procedure is the simplest to implement, it generally results in a larger expected number of tests than the other procedures. Halving generally results in the smallest number of tests, reaching a minimum of 220.77 for a group size of 500, the maximum group size included in the plot. The actual minimum expected number of tests for halving is 172.39 when a group size of 4,080 is used. On the right-side y-axis of the plot in Figure 2, the number of tests has been translated into a year's length of time. Baltar says that individual testing will take about 61 years. Using a group size of 500, the expected time that halving takes is only 101 days. Even for Dorfman's procedure, this amount of time is still only 1.45 years (1,155.64 expected tests) using a group size of 80. Given Battlestar Galactica was on television for six years, there would have been plenty of time for testing to be completed!

In general practice, the group size corresponding to the minimum expected number of tests is considered to be the "optimal" size. Usually, this optimal group size is an estimate only because the overall prevalence, which is needed for its calculation, is unknown. In the Cylon example here, the overall prevalence is most likely known now, so Figure 2 allows one to see the optimal group sizes for each group testing procedure. These group sizes often can be found mathematically too. For example, one can show that the optimal group size for Dorfman's procedure is approximately the square root of the inverted overall prevalence, resulting in 82.73 here.

Of course, if Baltar implemented one of the group testing procedures, the actual number of tests most likely would not be the same as the expected number of tests. There will be variability from application to application. Figure 3 gives bands illustrating an expected range for the number of tests using the mean ± 3×(standard deviation). Applying Chebyshev's Theorem, one would expect to observe the number of tests to be within this range at least 89% of the time. For larger group sizes, both Dorfman's and Sterrett's procedures can produce a lot of variability leading to much uncertainty in the number of tests. Alternatively, both halving and matrix pooling lead to much less uncertainty, so an extremely large number of tests will not likely happen. When comparing the procedures based on their optimal group sizes though, Dorfman's and Sterrett's procedures result in a more reasonable amount of variability, but still more than halving and matrix pooling. Overall, if Baltar chose the halving procedure and a group size of 500, the expected number of tests is between 127.06 and 314.47 (58 to 144 days).

5 Additional considerations

While optimal group sizes are nice to know, they often can not be used in practice. Diagnostic testing procedures must be calibrated to ensure their accuracy is similar to that for individual testing. Using too large of a group size can dilute the composited specimen to a point which prevents detection. Also, optimal group sizes can be chosen based on other measures. Commonly, cost is included in its calculation. Labor and storage of specimens, which may be longer due to retesting, all can be factored into an optimal group size calculation. There are practicality issues that must be considered as well when choosing a group size. For example, if a 10×10 well plate is available, but a group size of 11 for matrix pooling is optimal, a group size of 10 would be likely used instead.

Testing error will likely occur at some time with any diagnostic procedure. Because of the additional uncertainty in test results, both individual and group testing will have more tests and more variability than when test results are perfectly accurate. For example, confirmatory tests may be needed to confirm a positive test result. Diagnostic accuracy is defined in terms of two quantities: sensitivity (S_e) and specificity (S_p). Sensitivity is the probability an individual or group is diagnosed to be positive from a single test given the individual or group is truly positive. Specificity is similarly defined for a negative diagnosis given a true negative. For example, if $S_e = 0.95$, $S_p = 0.95$, and groups of size 80 are used for Cylon detection, the expected number of tests for Dorfman's procedure increases to 3,495.22 with an expected range of 2,562.19 to 4,428.26 tests (without including confirmatory tests). Still, the number of tests using Dorfman's procedure is much less than for individual testing.

Due to the group testing protocols, the overall probability of diagnosing an individual to be positive given they are a true positive, called the "pooling sensitivity," for many procedures is lower than for individual testing. Exact formulas for the pooling sensitivity have been derived for some of the group testing procedures examined here to better gauge the overall effect of testing error. For example, the pooling sensitivity for Dorfman's procedure is S_e^2, while for individual testing it is just S_e because one test is performed for each individual. The pooling sensitivity can be increased by additional testing of individuals declared negative. In contrast, the pooling specificity, the probability an individual is declared negative through a group testing procedure given they are a true negative, is often higher than under individual testing.

Another general purpose of group testing is to estimate the prevalence of a trait in a population. Sometimes, estimation is of interest only, so that identification is not a goal. In addition to identifying individuals who are Cylon, Baltar may be interested in using all of the initial group tests from a hierarchical procedure to estimate prevalence. This would give him an initial impression of the overall Cylon prevalence in the fleet. After first placing individuals randomly into groups, Table 1 gives the estimates, standard deviations, and 95% confidence intervals for the overall prevalence (supplementary materials at www.chrisbilder.com/grouptesting provide further details for the formulas used). For example, the estimated prevalence is 0.0001462 when only the group test outcomes from groups of size five are used. Overall, all of the estimates are all quite close to the true prevalence of 0.0001461. Larger group sizes produce estimates a little farther from the true prevalence, but this should be expected because less is observed from the population. Notice as well that all confidence intervals capture the true prevalence. While these calculations are made for individuals randomly assigned to groups, very similar results should be expected for other random groupings due to the small number of Cylons. In fact, it is quite unlikely to have groups with more than one Cylon in it, except in the case of the larger group sizes.

6 Conclusions

As in the real world, statistics could have played a significant role in solving this sci-fi problem. However, the consequences for implementing these procedures might have prematurely stifled fans' enthusiasm for Battlestar Galactica, because the humanoid Cylons would have been identified earlier in the show. While the television series ended in 2009, a prequel, Caprica, premieres in 2010. Caprica will investigate topics such as how Cylons were first developed by humans. Of course, this makes one wonder if the use of the statistical science could have played a role in preventing the Cylon attack on the humans in the first place. Fans can only hope the producers will ask a statistician to serve as a consultant in the writing of the new show!

Further Reading

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.

Peck, C. (2006). "Going after BVD." *Beef* 42:34-44.

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.

Tebbs, J. and Bilder, C. (2004). "Confidence interval procedures for the probability of disease transmission in multiple-vector-transfer designs." *Journal of Agricultural, Biological, and Environmental Statistics* 9:75-90.

Verstraeten, T., Farah, B., Duchateau, L., and Matu, R. (1998). "Pooling sera to reduce the cost of HIV surveillance: a feasibility study in a rural Kenyan district." *Tropical Medicine and International Health* 3:747-750.

Figure 1. Two types of Cylons; the Cylon on the left is a Centurion, and the Cylon on the right is the new humanoid type.

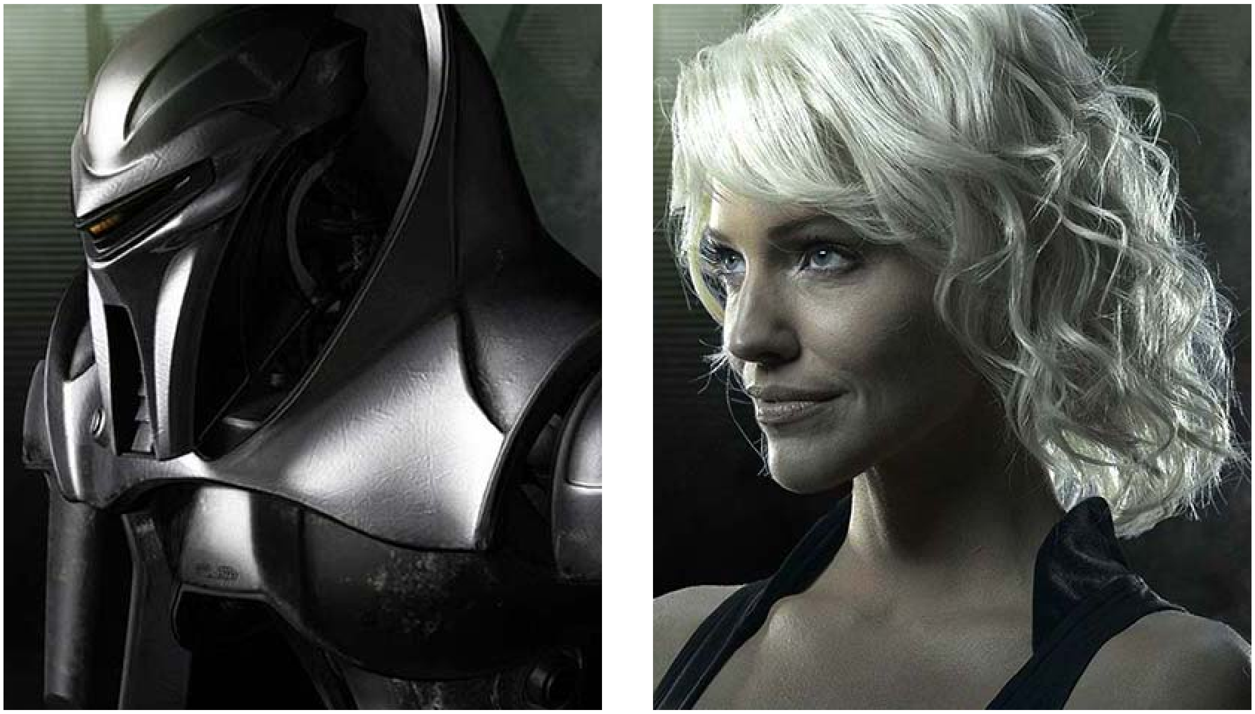

Figure 2. Expected number of tests versus group size using four different procedures; group size increments of size 10 are used for constructing the plot; conversion to years of testing is on the right axis.

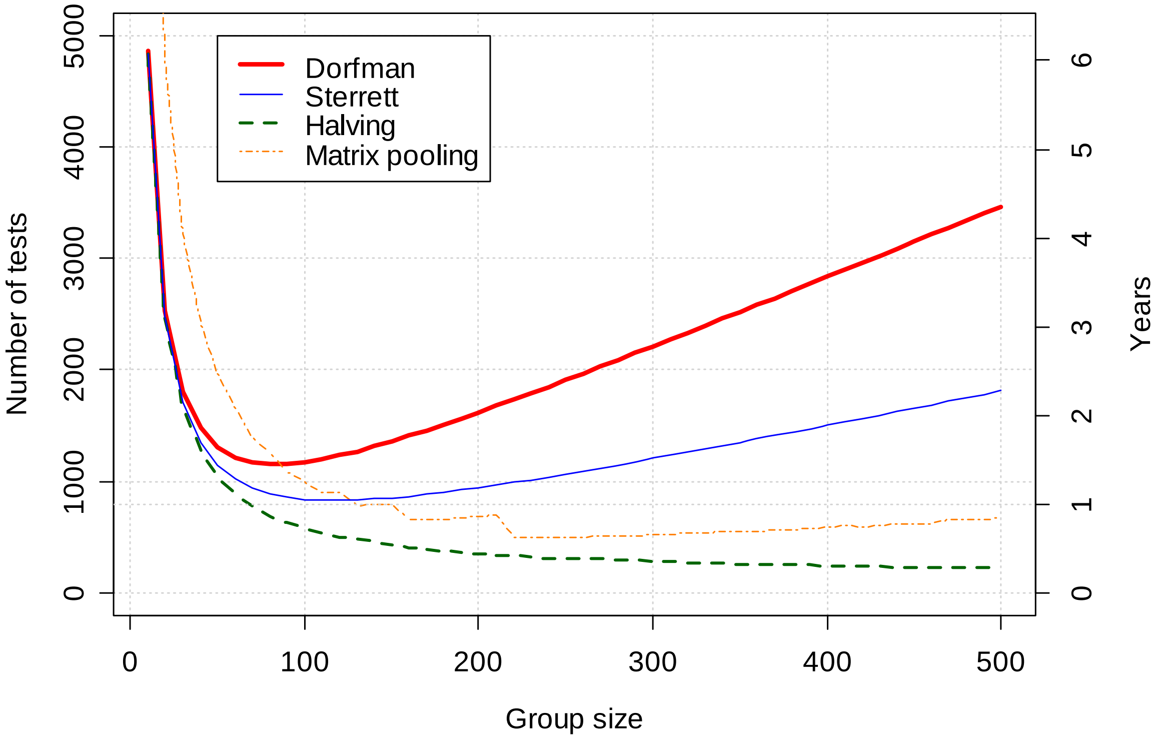

Figure 3. Expected tests ± 3×(standard deviation) bands plotted by group size for four procedures; group size increments of size 10 are used for constructing the plot; conversion to years of testing is on the right axis.

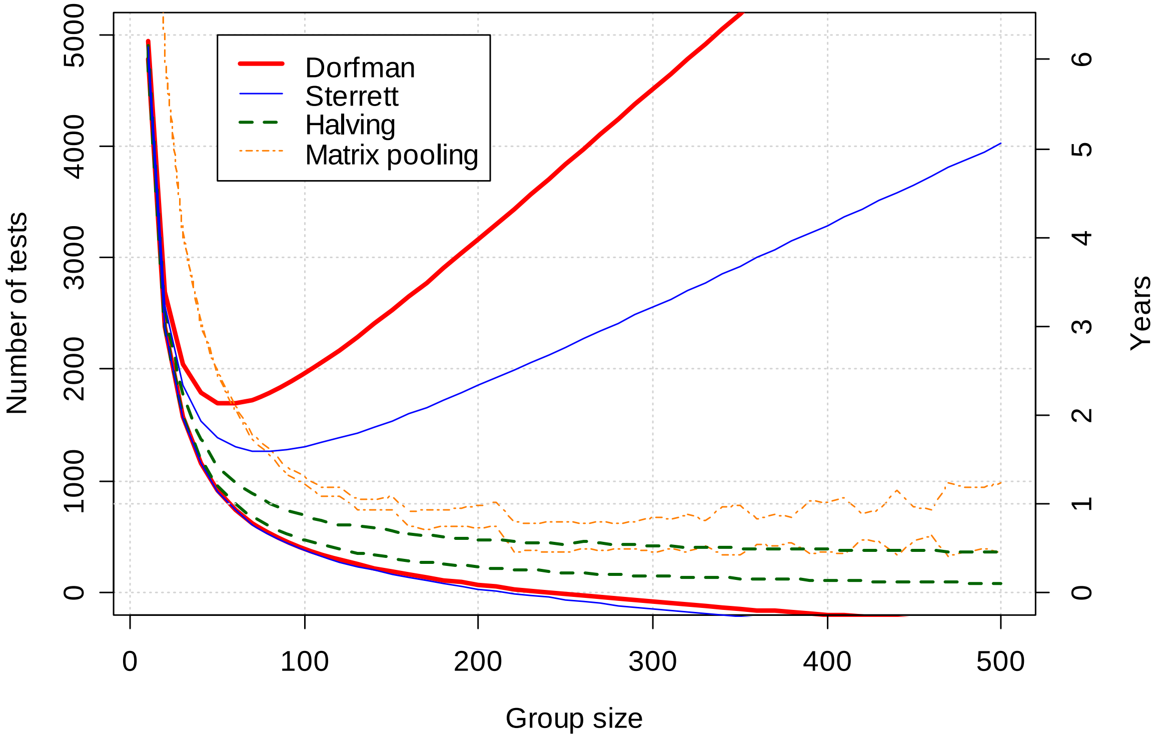

Table 1. Estimates of the 7/47,905 = 0.0001461 overall prevalence through using only the initial group test outcomes from a hierarchical procedure; individuals are randomly put into groups of the size indicated in the table.

		Standard	95% confidence interval		# of positive
Group size	Estimate	deviation	lower limit	upper limit	groups
5	0.0001462	0.00005524	0.00003789	0.0002544	7
10	0.0001462	0.00005526	0.00003791	0.0002545	7
50	0.0001466	0.00005542	0.00003802	0.0002553	7
100	0.0001472	0.00005629	0.00003816	0.0002562	7
500	0.0001293	0.00005281	0.00002584	0.0002328	6
1000	0.0001338	0.00005466	0.00002667	0.0002409	6

Sidebar 1 – Areas where group testing is used.

Areas	**Description**
Screening blood donations	All blood donations need to be screened for diseases such as HIV, hepatitis, and West Nile virus. The American Red Cross uses groups of size 16 and Dorfman's procedure for their screening. See Dodd et al. (2002).
Drug discovery experiments	Early on in the drug discovery process, hundreds of thousands of chemical compounds are screened to look for those that are active. The matrix pooling procedure is used often for this purpose. See Remlinger et al. (2006).
Plant pathology	In multiple vector transfer design experiments, groups of insect vectors are transferred to individual plants. After a sufficient amount of time, the plants are examined to determine if they have become infected by the insects. In this case, the plants provide the group responses. See Tebbs and Bilder (2004).
Veterinary	Among the many applications, cattle are screened for the bovine viral diarrhea virus. Groups of up to size 100 are formed from ear notches. See Peck (2006).
Public health studies	Group testing provides a cost efficient mechanism for poorer countries to obtain information on disease prevalence. See Verstraeten et al. (2000).