3
Pooling Experiments for Blood Screening and Drug Discovery

JACQUELINE M. HUGHES-OLIVER

Pooling experiments date as far back as 1915 and were initially used in dilution studies for estimating the density of organisms in some medium. These early uses of pooling were necessitated by scientific and technical limitations. Today, pooling experiments are driven by the potential cost savings and precision gains that can result, and they are making a substantial impact on blood screening and drug discovery. A general review of pooling experiments is given here, with additional details and discussion of issues and methods for two important application areas, namely, blood testing and drug discovery. The blood testing application is very old, from 1943, yet is still used today, especially for HIV antibody screening. In contrast, the drug discovery application is relatively new, with early uses occurring in the period from the late 1980s to early 1990s. Statistical methods for this latter application are still actively being investigated and developed through both the pharmaceutical industries and academic research. The ability of pooling to investigate synergism offers exciting prospects for the discovery of combination therapies.

1   Introduction

The use of pooling experiments began as early as 1915 and was, initially, used in dilution studies for estimating the density of organisms in some medium. Examples quoted by Halvorson and Ziegler (1933) include investigations of densities of bacteria in milk and protozoa in soil. Prior to 1915, most dilution methods were inadequate because they failed to account for chance or error in observation. In 1915, McCrady presented a method of estimation based on probability which was then expanded by Halvorson and Ziegler (1933) to provide an estimator of density based on pooled data. Fisher (1921) also used a similar pooling-based estimator.

These early uses of pooling experiments were born of necessity, as explained below for bacterial density estimation. In order to determine the absence or presence of bacteria in a fluid, cultures are made of a number of samples (small amounts) of the fluid. Growth of a colony of bacteria within a fluid sample indicates the presence of bacteria and no growth indicates absence of bacteria. The act of culturing this fluid can be viewed as applying a test, the result of which is "good"

or "bad". This test is applied simultaneously to every molecule present in that sample of fluid and the results for all the molecules are pooled. The combined test results (from all samples) are then used to estimate the density of bacteria present in the source fluid. Although it is virtually impossible to perform this test on individual molecules, it is quite simple to ascertain whether a culture from the pooled molecules is free of colony growth.

Today, pooling studies are not typically used from necessity. Rather, they are used because of the economic gains, savings in time, or precision gains that can result. A useful review of pooling experiments from the point of view of composite sampling methods is offered by Lancaster and Keller-McNulty (1998). This chapter focuses on current usage for populations in which individuals are labeled with respect to one or more traits and where pooling experiments are optional. More specifically, the discussion addresses applications in blood testing and drug discovery. This is not meant to be an exhaustive review, but rather a vehicle for highlighting some important aspects of pooled screening in these two areas.

Applications in drug discovery require the identification of "hit compounds," which are those compounds having activity greater than some prespecified threshold in one or more biological assays. Good hit compounds need to be identified quickly to allow progression to other phases of drug discovery (see Chapter 4). One application in blood screening requires the identification of individuals with sero-prevalence (detectability in blood) of one or more diseases. Cost effectiveness is important here because a balance must be struck between the cost of testing, which can be high, and the large populations that must be screened. A second issue that arises in blood screening is the need to estimate prevalences, possibly as a function of covariates.

In order to address the two areas of application simultaneously, the term *individual* is used to mean either a person (in the context of blood testing) or a compound (drug discovery); the term *active* means either positive for one or more diseases (blood testing) or exceeding an activity threshold (drug discovery); and the term *population* means either a group of people being screened (blood testing) or a compound library being screened (drug discovery).

Two fundamentally different problems arise from pooling experiments, namely, *estimation* and *classification*. Estimation involves the use of pooled samples for decreasing the cost-per-unit information when estimating the prevalence of active individuals in a population. These estimation results may then be used as the end-product of analysis or they may be incorporated into a classification scheme. The estimation results serve as the end-product of analysis when the goal of the study is to estimate the prevalence of active individuals but there is no interest expressed in actually identifying these active individuals. In a classification scheme, the ultimate goal is screening for the purpose of identifying active individuals. The performance of a classification scheme is typically assessed by considering the expected number of tests required to identify active individuals with particular attributes. Drug discovery is considered to be a classification problem, but results from the estimation problem can also be used to inform classification decisions. For blood testing, the

application may be either a classification or an estimation problem, depending on the context.

The most common assumptions of pooling experiments are briefly critiqued in Section 2. In Section 3, general accomplishments and advances in pooling experiments are reviewed, irrespective of their particular applications. Sections 4 and 5 provide details specific to blood testing and drug discovery, respectively.

2   Types of Pooling and Assumptions

Pooling experiments are of two basic types, *simple* or *orthogonal*. In simple pooling, each individual appears in exactly one pool; see Figure 1(a), where each circle in the box represents an individual, and individuals in the same column are in the same pool. An active pool response must be followed by individual testing to determine which specific individuals in the pool are active. In orthogonal pooling
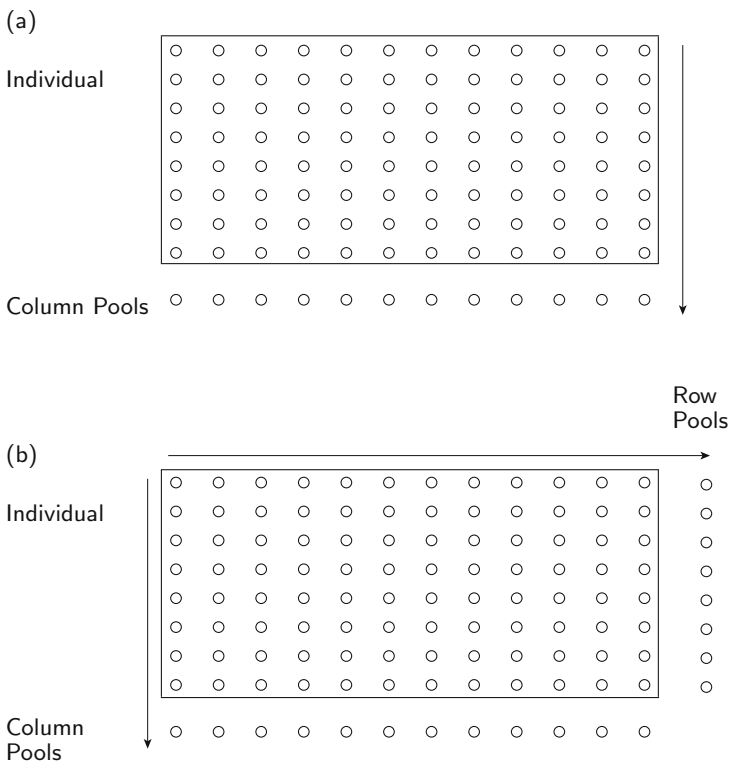

FIGURE 1. (a) Simple pooling: Individuals, represented by circles, are pooled according to their column location; each individual appears in exactly one pool. (b) Two-way orthogonal pooling: Individuals are pooled according to both their column and row locations; each individual appears in exactly two pools.

using d dimensions, individuals appear in exactly d pools; see Figure 1(b) for $d = 2$. The determination of active individuals from orthogonal pooling is easier than that from simple pooling. Consider orthogonal pooling over $d = 2$ dimensions corresponding to rows and columns. If a compound lies simultaneously in an active row and an active column, then it is reasonable to believe that this compound is active and that it should thus be assigned a favorable rank for individual testing. Despite its benefits, orthogonal pooling adds many complexities and is, consequently, not as popular as simple pooling. All further discussion is limited to simple pooling.

Pooling experiments are based, historically, on several assumptions that are often blatantly unjustified. The first assumption is that individuals have equal probabilities of being active. In blood testing, genetic characteristics, environmental exposures, and demographic identities are widely accepted as sources of variability for disease status, thus suggesting that probabilities of activity are not constant across the population (Dorfman, 1943). In drug discovery, it is well recognized that structure–activity relationships (SARs, see Chapter 4), where activity is related to chemical structural features of a compound, lead to nonconstant probabilities of activity; see McFarland and Gans (1986).

A second assumption generally used is that interactions do not occur within a pool; that is, activity is neither enhanced nor degenerated by testing multiple compounds using a single test on a pool. It is possible, however, that individually inactive compounds can give an active test result when pooled together (Borisy et al., 2003), thus providing a case of "activity enhancement" by pooling. This phenomenon is called *synergism* and its detection is crucial to the development of combination therapies in the pharmaceutical industry. The reverse situation can also occur in that pooled testing of individually diseased samples can result in disease-free pool results (Phatarfod and Sudbury, 1994), thus providing a case of "activity degeneration" by pooling. This phenomenon is called *antagonism* or *blocking* and is considered an undesirable potential effect of pooling in the blood testing application. Blocking relationships that occur in drug discovery applications can have a positive impact on screening outcomes in that they provide further implicit evidence of structure–activity relationships.

A third assumption concerns absence of errors in testing. Both blood testing and drug discovery have strong potential for false negatives and false positives. Errors in testing are inherently linked to assumptions regarding interactions within a pool. Both concepts are, in turn, related to the sometimes arbitrarily chosen threshold value used for categorizing a continuous assay response into only two classes of "active" or "inactive".

3   History of Pooling Experiments

Dorfman (1943) has been credited with the origin of pooling experiments in the statistical literature. His ideas were popularized through the books of Feller (1957, page 225) and Wilks (1962) and became known as "the blood testing problem".

Many efforts were then made to refine Dorfman's proposal by extending the number of stages using various retesting schemes and by relaxing assumptions. This section provides a brief summary of some key results from these efforts; see also Chapter 9 for related work in factorial experiments.

3.1   The Dorfman Model and Assumptions

Suppose that, in a large population of f individuals, each individual has, independently, the same probability p of being active. In this context, p represents a latent propensity for an individual to be active; some individuals ultimately express this latent feature and are thus labeled as actives, whereas others never express the latent feature and thus are labeled as inactive. If individuals are pooled into groups of size k and if pooling does not alter the behavior of individuals, then the resulting $g = f/k$ pools will, independently, have the same probability $\theta = 1 - (1 - p)^k$ of being active. Hence, the number of active pools, X, follows a binomial distribution with parameters g and θ. Of course, activity of pools or individuals must be revealed by some testing system and for now this system is assumed to be perfect. In other words, *sensitivity* (the probability that a test will identify, by testing outcome, an individual as active given that the individual is truly active) and *specificity* (the probability that a test will identify, by testing outcome, an individual as inactive given that the individual is truly inactive) are both assumed to be 1.0. Dorfman himself did not believe these assumptions strictly but was able to build from the strength of the overall approach to make worthwhile reductions in the required number of tests over one-at-a-time testing. Aspects of sensitivity and specificity are also discussed in Chapters 4 and 6.

3.1.1   Classification

Dorfman's application was the need to identify World War II Selective Service inductees whose blood contained syphilitic antigens. In other words, his was a classification problem and he wanted to minimize the number of tests required to classify all inductees. All individuals in inactive pools were declared to be inactive, without further testing. All individuals in active pools were subjected to one-at-a-time testing, thus leading to a random total number of tests $T = f/k + Xk$ (where f, k, and X are defined above). Dorfman then needed to determine a pool size to minimize the expected total number of tests. Pooling would only be advantageous if, on average, the total number of tests is less than f, which is the number of tests required by one-at-a-time testing. Dorfman minimized the expected relative cost, for given p,

$$\frac{E(T)}{f} = \frac{1}{k} + 1 - (1 - p)^k, \qquad \text{for } k > 1,$$

with respect to k, to determine the best possible improvements offered by pooling experiments over one-at-a-time testing. For example, by pooling, he obtained an 80% cost savings in tests over one-at-a-time testing when $p = .01$ and $k = 11$. The savings decrease as p increases but are still appreciable even for larger p with, for example, 28% savings when $p = .15$ and $k = 3$. In fact, pooling, based

on Dorfman retesting for classification, is better than one-at-a-time testing when $1/k < (1 - p)^k$. The approximation $(1 - p)^k \approx e^{-pk}$ is sometimes used to claim that pooling is better than one-at-a-time testing when $p < (\ln k)/k$. When p and k are both large, many active pools are observed and, consequently, more individual retests are required and this reduces the desirability of pooling.

Despite its simplicity, Dorfman's retesting strategy is still very widely used today, especially in blood testing and drug discovery applications. His rough guidelines for choosing k such that $p < (\ln k)/k$, coupled with recommendations by Thompson (1962), Kerr (1971), Loyer (1983), and Swallow (1987) to use an a priori upper bound on p, is also commonly used today. Indeed, the attraction of the Dorfman strategy is its simplicity. Improved methods for classification, some of which are discussed in this article, add various levels of complications that users may not yet be ready to accept.

3.1.2   Estimation

Dorfman (1943) did not really address the problem of estimating the prevalence p but, using his assumptions, others did. Gibbs and Gower (1960) and Thompson (1962) investigated the maximum likelihood estimator of p:

$$\hat{p} = 1 - \left(1 - \frac{X}{g}\right)^{1/k},$$

where $g = f/k$ is the number of pools and X the number of active pools.

This is a positively biased, but consistent, estimator for p. Based on the asymptotic variance of $\hat{p}$, Peto (1953) and Kerr (1971) determined that the optimum group size k satisfies $(1 - p)^k = .203$. Based on asymptotic considerations, Thompson (1962) suggested that the group size should be approximately $k = (1.5936/p) - 1$. He also argued, however, that the asymptotic results can be very misleading and offered small-sample exact bias and variance formulae. Gibbs and Gower (1960), Griffiths (1972), Loyer (1983), and Swallow (1985, 1987) also gave small-sample results.

When c is the nontesting cost associated with obtaining an individual sample (for example, personnel time for drawing blood from an individual) divided by the cost of a test, Sobel and Elashoff (1975) showed that pooling is advantageous when

$$p < 1 - \frac{1 + 2c}{3 + 2c}.$$

For extremely costly tests, pooling can be beneficial for p as large as 2/3.

3.2   Some Alternative Models

Extensions of Dorfman's procedure follow four main branches:

  (i) Development of different retesting schemes;
 (ii) Strategies when p is unknown, as is usually the case;
(iii) Departures from binomial assumptions; and
(iv) Errors in testing.

The literature is quite extensive (see Hughes-Oliver, 1991), so only key papers are referenced here. For brevity, no attempt has been made to separate extensions for the goal of classification from extensions for the goal of estimation.

3.2.1   Retesting Schemes

Many different retesting schemes have been suggested in the literature, some of which require infinite testability of the units. For example, Sterrett (1957) proposed retesting individuals in an active pool only until an active individual is found. The remaining untested individuals are then retested as a single pool and the process is repeated. Sobel and Groll (1959) proposed a retesting scheme based on *nested halving* procedures. Active pools are subdivided into two pools of size approximately $k/2$, each of which is tested. Individuals in an inactive subpool are declared inactive but an active subpool is again halved. Halving terminates when pool size becomes one, that is, at individual testing. Sobel and Elashoff (1975) proposed a general nested retesting scheme for estimation, of which nested halving is a special case. They found that a certain class of nested halving procedures is highly efficient and the savings over one-at-a-time procedures is even greater for the estimation problem than for the classification problem. They also found that, when the cost of obtaining individuals relative to the cost of a test is negligible, the optimal testing scheme does not include retesting. Chen and Swallow (1990) confirmed the finding that retesting is not advantageous for estimation when testing costs far exceed costs of obtaining individuals, but they showed that data from retesting can provide useful information for testing model assumptions.

In contrast to the work of Sobel and Elashoff (1975) and Chen and Swallow (1990), where the stated goal was to reduce cost per unit information for estimation in the presence of perfect testing, retesting has been shown to be useful for classification, especially when test results may be inaccurate. Litvak et al. (1994) argued that, even when testing is correctly executed, it can lead to incorrect conclusions and, in these cases, retesting provides significant improvements over no-retesting for reducing error rates associated with labeling samples when screening low-risk HIV populations.

Based on nested halving, Litvak et al. (1994) also proposed a new retesting scheme where inactive pools are subjected to a repeat test; if they again test inactive then all individuals in those pools are declared inactive, otherwise the pool is halved and subjected to additional testing. Gastwirth and Johnson (1994), who were also concerned with error rates for labeling individuals assuming imperfect testing, proposed a "back-end" retesting stage where pooled testing is used to rescreen a subset of individuals who were declared inactive from "first-stage" pooled testing.

3.2.2   Choosing the Pool Size

The success of a pooling experiment depends heavily on the choice of a good value for the pool size k. Unfortunately, optimal pool size depends on the value of p. In the absence of a priori information on p, Le (1981) and a number of other authors

recommended that different pool sizes be used and the resulting data on the number of active pools for each pool size be combined to yield an estimator. Thompson (1962) argued that an a priori upper bound on p should be used to determine a single pool size, and Hughes-Oliver and Swallow (1994) and Hughes-Oliver and Rosenberger (2000) proposed two-stage adaptation to allow a single update of the pool size. These last authors also addressed the issue of pool size when there are multivariate responses from pools, motivated by the need to monitor prevalence rates for several diseases simultaneously.

3.2.3 Departures from Binomial Assumptions

On the issue of departures from binomial assumptions, Finucan (1964) considered a case where stratification occurs and results in different probabilities of activity for different individuals. A good early reference for various approaches to dealing with such situations is that of Hwang (1984). Chen and Swallow (1990) noted that model assumptions can be tested if data on unequal pool sizes are available. Many recent articles also consider the situation where probability of activity is dependent on covariates. For small numbers of covariates, Hung and Swallow (2000), Vansteelandt et al. (2000), Xie (2001), and Tebbs and Swallow (2003a,b) obtained estimates of prevalences in the different strata. For large numbers of covariates, Xie et al. (2001), Zhu et al. (2001), and Yi (2002) obtained estimates of prevalences in the different strata then ranked the estimated prevalences to define a testing order for the classification problem. Thus, the estimation problem was an intermediate step, not the ultimate goal, of the drug discovery applications of these authors. On a related note, Remlinger et al. (2005) considered the design problem of assigning individuals to pools based on their covariates; the goal was classification in the presence of covariate-dependent prevalences.

3.2.4 Errors in Testing

The problem of errors in testing has been examined by a host of investigators. References to investigators from a clinical/laboratory science viewpoint are given in Section 4.2. From a statistician's viewpoint, Gastwirth and Hammick (1989) and Hammick and Gastwirth (1994) used trinomial models in which either a confirmatory pool test or an independent pool test was used to reduce the number of false positives. They also incorporated sensitivities and specificities (Section 3.1) of the testing scheme into their estimator while maintaining individual anonymity. Tu et al. (1994, 1995) also incorporated sensitivities and specificities of the testing scheme and showed that this leads to improved estimation accuracy. Vansteelandt et al. (2000) took the same approach but with the added complication of covariate-adjusted estimation of prevalence. Hung and Swallow (1999) investigated robustness properties of the pooling estimator with respect to dilution effects and serial correlation models. Wein and Zenios (1996) also investigated dilution effects. In the area of drug discovery, Langfeldt et al. (1997), Xie et al. (2001), Zhu et al. (2001), Yi (2002), and Remlinger et al. (2005) all investigated

procedures that model possible interactions occurring within pools; such interactions may be mislabeled as errors in testing.

4   Pooling for Blood Screening

4.1   Background

Pooling is now considered to be a routine option in blood screening, especially for the human immunodeficiency virus (HIV). There are many reports espousing the benefits of pooled testing in countries across the world, using a variety of assay techniques.

There are actually three blood screening applications for which pooling has been beneficial. The two most common applications are the context of classification, where the goal of blood screening is to identify individuals with sero-prevalence of one or more diseases. One classification application arises from the need to screen donated blood and blood products and the other from the need to screen for individual diagnoses. Cost effectiveness, as measured by the reduction in the expected total number of tests, is the most commonly used assessment of pooling methods. The third application is the need to monitor changes in sero-prevalence over time for (possibly) different sets of individuals, where demarcation of individuals may occur along demographic lines or spatial/regional clusters.

4.1.1   Classification

Motivated by a more than 90% transmission rate of HIV by transfusion of blood and blood products, the World Health Organization (WHO) argued for 100% screening of donated blood. Recognizing that developing countries can ill-afford the cost of 100% one-at-a-time screening, WHO issued recommendations for testing for HIV antibody on serum pools (WHO, 1991) in areas where sero-prevalence is less than 2%. In fact, this figure of 2% sero-prevalence is much too restrictive. Many investigators have achieved success with much higher prevalences. For example, Soroka et al. (2003) described the successful use of pooling where prevalence was 9%. It is important to note that, for screening blood supplies, complete identification of sero-positive individuals is not necessary. All that is needed is a method for tracking the complete donated sample, without personal identifiers. This makes pooled screening very attractive for screening blood supplies because donors can be assured that their anonymity will be maintained.

Another classification problem occurs when individual diagnosis is the required outcome of a screening campaign. In such a campaign, personal identifiers must be maintained for the purpose of reporting back to individuals about their sero-prevalence. Moreover, diagnostic testing requires that sero-positive pools be subjected to confirmatory gold-standard tests.

4.1.2 Estimation

Gastwirth and Hammick (1989) and Hammick and Gastwirth (1994) approached the blood testing problem with a keen eye towards preserving individual privacy rights. They proposed screening strategies designed for estimating prevalences. Rather than focus on the cost-saving advantages of pooling, these authors selected pooling because of the anonymity it provides to individuals being screened. They also reduced false predictive values by employing confirmatory tests to verify sero-prevalence.

4.2 HIV Testing

The standard practice in developed countries for determining HIV sero-prevalence is first to apply the cost-effective, but suboptimal, enzyme-linked immunosorbent assay (ELISA) test. For those individuals who are identified as sero-positive by the ELISA test, follow-up testing is then performed using the gold standard Western blot test. Unfortunately, the Western blot is very expensive, difficult to standardize and often results in no clear diagnosis for some individuals (Tamashiro et al., 1993). To relieve the cost burden, the WHO recommends a series of repeat testing that uses cheaper tests, namely ELISA or simple or rapid tests, to avoid the Western blot while still maintaining testing accuracy. In general, the Western blot best is up to six times as expensive as rapid or simple tests and 18 times as expensive as ELISA; see, for example, WHO (1992). Rapid and simple tests provide results in less than one hour (less than 30 minutes for rapid tests) and may be performed by personnel having little or no laboratory training. ELISA must be performed in a laboratory (so results are not immediately available) by extensively trained laboratory professionals.

The WHO (1992) recommendations are shown in Figure 2 and supporting text is given in Table 1. Strategy I is recommended for screening contributions to a blood supply. It says that a contribution should only be accepted if it is sero-negative according to either the ELISA test, or the rapid test, or the simple test. Sero-positive samples are not considered further. Strategy I is also recommended when prevalence is high and the goal is HIV surveillance.

WHO's Strategy III is recommended for diagnosing symptom-free individuals living in areas of low prevalence. It is the strategy that allows the greatest number of retests. If the first test is sero-positive, it is followed up with a second test that is not simply a repeat measurement of the first test. Specifically, the assay procedure should differ from the first assay procedure in some substantial way; for example, different antigen preparation, different test principle (such as indirect versus competitive) or both. The first test should be very sensitive but the other two tests should have higher specificity than the first. If this second test is again sero-positive, a third and last test is applied. Strategy II is similar but with only two stages.

Effective clinical pooling studies for HIV classification and surveillance have been reported by a large number of investigators. Emmanuel et al. (1988),
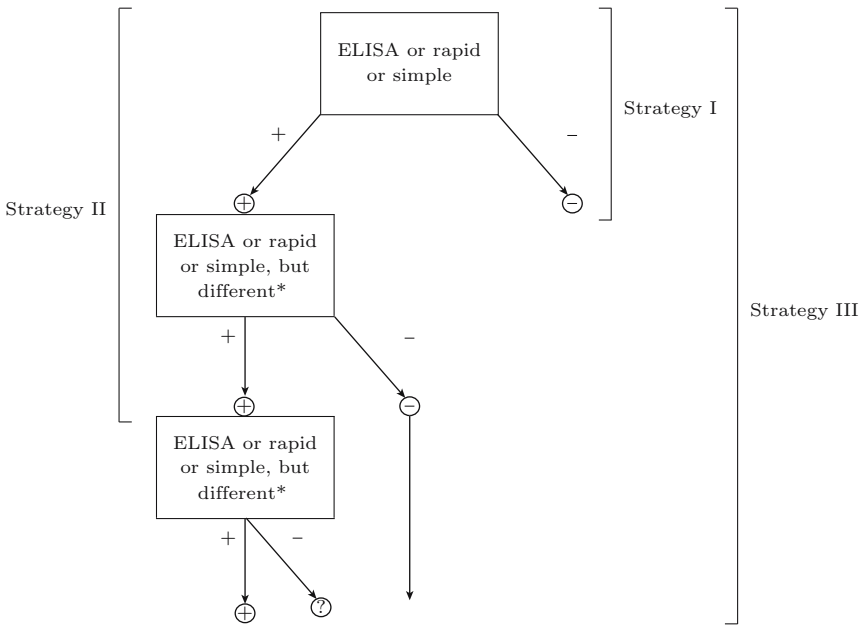
FIGURE 2. The World Health Organization 1992 recommendations on how to screen for HIV without using the Western blot test. * denotes that subsequent assays should differ from previous assays.

Cahoon-Young et al. (1989), Kline et al. (1989), Behets et al. (1990), Archbold et al. (1991), Ko et al. (1992), Babu et al. (1993), and Perriens et al. (1993) have all reported successes for several different countries, including countries in Africa and Asia. "Success" here is defined as the appropriate management of the logistics of pooling and the reduction of the amount of testing required. Moreover, successes have been achieved based on several different testing protocols, including ELISA, Western blot, and rapid testing techniques; see also Davey et al. (1991), Seymour et al. (1992), Raboud et al. (1993), McMahon et al. (1995), Verstraeten et al. (1998), and Soroka et al. (2003). These studies reported up to 80% reductions in cost for pooling experiments compared with one-at-a-time testing.

TABLE 1. WHO recommendations for the use of the strategies shown in Figure 2

Strategy	Limits on p	Objective
I	None	Blood supply
I	$p > 0.1$	HIV surveillance
II	$p \leq 0.1$	HIV surveillance
II	None	Diagnosis, HIV symptoms
II	$p > 0.1$	Diagnosis, no symptoms
III	$p \leq 0.1$	Diagnosis, no symptoms

Since the late 1980s, statistical contributions to pooling for blood testing have focused on the following aspects: assessing changes in sensitivity and specificity due to pooling, designing pooling strategies to accommodate both cheap initial screens and gold-standard confirmatory screens, and estimation of covariate-dependent prevalences.

Let us first consider approaches to assessing changes in sensitivity and specificity due to pooling. As defined in Section 3.1, sensitivity is the probability that a test correctly detects antibodies in a serum sample, and specificity is the probability that a test correctly identifies an antibody-free serum sample. These probabilities have been a major area of concern in pooling studies for blood testing (WHO, 1991). The over-arching issue when screening a blood supply is whether dilution effects will cause a single sero-positive individual to be missed when combined in a pool with several (perhaps as many as 14) sero-negative individuals. This issue relates to the false negative predictive value as follows. A predictive value is the probability of truth given an individual's testing outcome; a false negative predictive value is the probability that the individual is truly active but is labeled as inactive from testing; a false-positive predictive value is the probability that the individual is truly inactive but is labeled as active from testing. When screening for diagnostic purposes, the major concern is that sero-negative individuals will be labeled sero-positive; this relates to the false positive predictive value. Repeatedly, however, studies have indicated that, under their worst performance, these possible pooling effects are negligible. In fact, Cahoon-Young et al. (1989), Behets et al. (1990), Archbold et al. (1991), Sanchez et al. (1991) all reported reductions in the number of misclassified sero-negative individuals; for example, Cahoon-Young et al. (1989) found that there were seven misclassified sero-negative individuals out of 5000 tested, but no misclassified sero-negative pools out of 500 tested.

For understanding sensitivity, specificity, false negative predictive value, and false positive predictive value, consider the four cells and two column margins of Table 2, where individuals are cross-classified with respect to their true sero-status versus observed sero-status. Sensitivity is represented by $S_e = P$(testing outcome $+ \mid$ truth is $+$) and specificity is $S_p = P$(testing outcome $- \mid$ truth is $-$). With these definitions and with p denoting the probability of an individual having

TABLE 2. Cross-classification of individuals for "true" versus "observed" sero-status $(+, -)$ in terms of sensitivity S_e, specificity S_p, and probability p of positive sero-status

		True	
		$+$	$-$
Observed	$+$	pS_e	$(1-p)(1-S_p)$
	$-$	$p(1-S_e)$	$(1-p)S_p$

positive sero-status, the false negative predictive value is

$$FNPV = P(\text{ truth is } + \mid \text{ testing outcome } -) = \frac{p(1 - S_e)}{p(1 - S_e) + (1 - p)S_p}$$

and the false positive predictive value is

$$FPPV = P(\text{ truth is } - \mid \text{ testing outcome } +) = \frac{(1 - p)(1 - S_p)}{(1 - p)(1 - S_p) + pS_e}.$$

Large false negative predictive values are particularly troubling when screening a blood supply because they allow sero-positive samples to enter the blood supply system, thus leading to possible transmission of deadly diseases. Minimizing the false negative predictive value is probably more important than increasing cost efficiency of pooling for this application. Of course, large false negative predictive values can arise even when screening is accomplished using one-at-a-time testing. False positive predictive values are of greater concern in diagnostic testing because they can cause undue stress for the falsely identified individuals and increase testing costs. Notice that if $S_e = S_p = 1$, then $FNPV = FPPV = 0$ and no misclassifications will occur.

Litvak et al. (1994) compared three pooling strategies and one-at-a-time testing with respect to their abilities to reduce $FNPV$, $FPPV$, the expected numbers of tests required, and the expected numbers of tests performed for each individual. The first pooling study considered was Dorfman retesting with pool size $k = 15$; that is, all individuals in sero-positive pools were tested one-at-a-time but no retesting was applied to individuals in sero-negative pools. The pool size of 15 was selected because, at the time, it was the largest acceptable size from a laboratory perspective for maintaining high sensitivity and specificity after pooling. Litvak et al. (1994) called this screening protocol T_0. Their second pooling protocol, T_2, was essentially the retesting method proposed by Sobel and Groll (1959) whereby sero-positive pools are recursively halved and testing of the subpools continues until no further splits are possible. In this strategy with $k = 15$, a serum sample must be positive four or five times before being declared sero-positive. Their third pooling protocol, T_2^+, is similar to T_2 except that each sero-negative pool is subjected to one confirmatory pool test before all its individuals are labeled as sero-negative. It was found that T_2 and T_2^+ were comparable and that both provided huge reductions in $FPPV$ compared with one-at-a-time testing but smaller reductions compared with T_0. For $FNPV$, T_2^+ was the best protocol. In short, pooling reduced both false negative and false positive predictive values.

The result from estimating sero-prevalence of HIV in the presence of errors in testing is really quite startling. Tu et al. (1994, 1995) found that pooling actually increases estimator efficiency by reducing the effect of measurement errors. Vansteelandt et al. (2000) extended the procedure to account for covariate adjustments. These results, along with the large number of empirical findings from investigators such as Emmanuel et al. (1988), clear the way for heavy reliance on pooling strategies to eliminate the backlog and reduce the cost of screening large populations. This is of particular importance to developing countries that are often

cash-strapped but might benefit the most from 100% screening. Even developed countries might want to rethink their screening strategies to take advantage of fewer but more informative pooled test results.

5   Pooling for Screening in Drug Discovery

5.1   Background

Twenty percent of sales from the pharmaceutical industry for the year 2000 were reinvested into research and development activities. This percentage is higher than in most other industries, including the electronics industry. At the same time, it is getting increasingly difficult to introduce (that is, discover, demonstrate efficacy and safety, and receive approval for marketing of) new drugs in order to recoup investment costs. On average, one new drug requires investment of $880 million and 15 years development (Giersiefen et al., 2003, pages 1–2). The days of profitability of "runner-up" or "me-too" drugs have long passed and the simple current reality is that survival and financial security of a pharmaceutical company demands that they find the best drugs as fast as possible. This means that the five major phases of drug discovery, as illustrated in Figure 3, need to be traversed aggressively. Details on the phases of drug discovery can be found in Chapter 4. Here, attention is directed to the third phase, Lead Identification, which is where pooling experiments for screening in drug discovery usually occur.
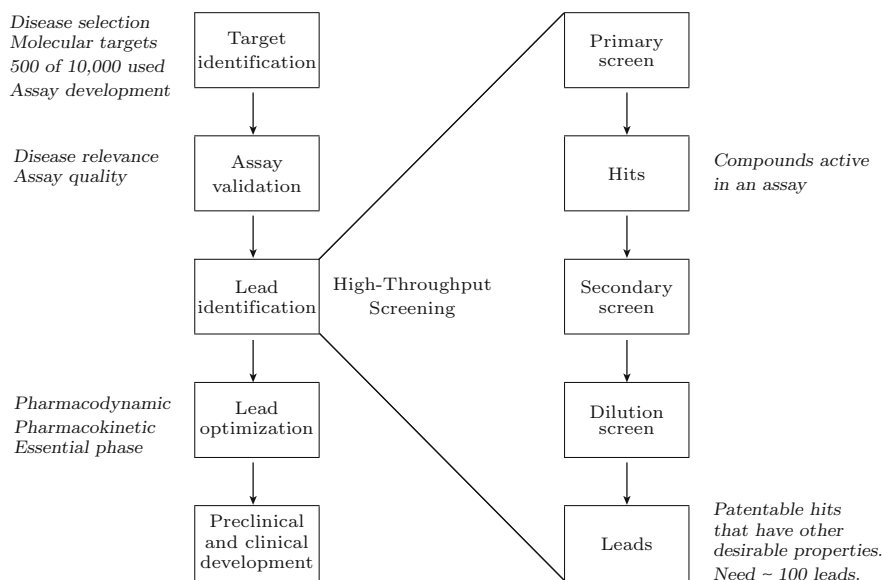

FIGURE 3. Phases of drug discovery.

Given a large collection of compounds, say $f = 500{,}000$, the goal of lead identification is to find about 100 compounds that are

(i) Active for the assay—this allows them to be called "hits";
(ii) Patentable; that is, their structures are novel and not already under patent;
(iii) Have good chemical properties such as stability, can be synthesized, are not toxic, and so on;
(iv) Something is understood about what makes them active; that is, their structure–activity relationships have been, at least partially, identified;
(v) Each compound is fairly different from the other ninety-nine.

Compounds that satisfy all these requirements are called *leads* or *lead compounds*. The need for properties (i)–(iii) is clear, but additional comments are warranted for the other properties.

Knowledge of structure–activity relationships allows chemists to focus on the essential substructures of the compound without wasting time with the portions that do not affect activity. The drug discovery phase that follows lead identification is *lead optimization*. In this phase, chemists expend enormous energies "tweaking" the leads to increase the chances of compounds making it through the grueling stages of preclinical and clinical development. It is imperative that the lead optimization phase produces very strong lead compounds to be investigated during preclinical and clinical development. Once a compound reaches the preclinical and clinical development phase, extensive additional financial and time investments are made, so that heavy losses would be incurred if the compound had to be abandoned further down the drug discovery channel because it possesses undesirable features (see also Chapter 4).

5.2 Differences and Similarities Between Blood Screening and Screening for Drug Discovery

The goals of drug discovery, as stated above, seem to be very similar to those of the blood screening for classification problem, but this is not at all the case. As mentioned, in earlier sections of this chapter, approaches to solving the blood testing for classification problem do not routinely incorporate covariate information. For the HIV blood testing problem, relevant covariate information for an individual may include the following: number of blood transfusions received, number of sexual partners, number of sexual partners who are HIV-infected, syringe use, drug use, sexual preference, and HIV status of parents. Recent investigations have allowed the estimation of prevalence in different covariate-defined strata, but the number of strata is never large and is quite typically less than 10. In screening for drug discovery, on the other hand, the number of covariates is quite often at least twice the number of pooled responses available. Indeed, the significant challenges that arise from the high-dimensional-with-low-sample-size data sets that usually result from "high-throughput screening" in drug discovery present major obstacles to analysis, even for one-at-a-time testing results. These difficulties are magnified

in the presence of pooled responses. More information is given by Langfeldt et al. (1997), Xie et al. (2001), Zhu et al. (2001), Yi (2002), and Remlinger et al. (2005).

Arguably, the biggest difference between the two application areas discussed in this chapter is the potential for synergistic relationships between compounds in pools for drug discovery, whereas no such concept has arisen for blood testing. Synergism has recently become the major supporting argument for pursuing pooling experiments in drug discovery (Xie et al., 2001; Yi, 2002; Remlinger et al., 2005). Synergistic relationships can only be discovered through pooling studies where compounds are forced together, and it is these synergistic relationships that form the basis of *combination therapies*. These therapies involve deliberate mixing of drugs and they are now the standard of care for life-threatening diseases such as cancer and HIV. Current combination therapies were discovered by combining individually active compounds after they had been approved by the Food and Drug Administration. By investigating synergistic relationships in vitro, it is expected that one could find a combination where, individually, the compounds are inactive but, when pooled, their activities exceed all other combinations. Borisy et al. (2003) demonstrated this quite nicely using several real experiments. For example, chlorpromazine and pentamidin were more effective than paclitaxel (a clinically used anticancer drug), even though individually neither drug was effective at tolerable doses. Similar ideas were discussed by Tan et al. (2003).

So, are cost considerations no longer important for drug discovery? The answer is "not really," or at least not as much as they used to be. Before the advent of high-throughput screening (HTS, see Chapter 4) and ultrahigh-throughput screening (uHTS), pooling was necessary for processing the large compound libraries typically encountered. In those days, a large screening campaign might screen a total of 50,000 compounds, and it would take months to complete. Today, uHTS can screen 100,000 compounds in a single day; see Banks (2000) and Niles and Coassin (2002). HTS and uHTS systems are centralized, highly automated, and are under robotic control so they can work almost around the clock with very small percentages of down-time.

The two applications of drug discovery and blood testing are similar in how they process screening outcomes. Comparing Strategy III of Figure 2 with the extended view of Lead Identification in Figure 3, it can be seen that both methods use three tests in labeling the final selected individuals. The selected individuals are the gems for drug discovery applications but, for the blood testing problem, they actually cause concern because they are blood samples that have been confirmed to be diseased.

5.3   Design and Analysis Techniques

A commonly used technique for analyzing drug discovery screening data from individuals is recursive partitioning (RP), more commonly known as "trees" (see, for example, Blower et al., 2002). In very recent times, efforts based on multiple trees (Svetnik et al., 2003) have become the method of choice, despite the additional difficulties associated with them, because of their good predictive abilities.

The number of researchers working to develop methodology appropriate for pooled drug screening data and who are allowed to discuss these issues outside the big pharmaceutical companies is very small. Papers from these researchers have been reviewed earlier in this chapter, but a few additional comments are warranted. The bulk of the work has been divided into two major paths. One path concerns the search for the efficient placement of individuals within pools; that is, the design of pooling studies. Because of the very large number of covariates, this is a difficult problem that requires computer-intensive techniques. Remlinger et al. (2005) obtained structure-based pooling designs to assign pool placement in response to covariate-adjusted prevalences. Zhu (2000) developed model-based designs for the same problem.

The second major path concerns analysis methods, including nonparametric, semi-parametric, fully parametric, and Bayesian approaches. Nonparameteric results are based on recursive partitioning on pooled data and require the formation of pooled summaries and decisions of whether and how to include the retested data in the analysis without violating independence assumptions. For the semi-parametric work, Yi (2002) modeled data from pooling experiments as missing data scenarios where missingness occurs at random. This was a novel use of the semi-parametric methodology to an area that had never before been considered. Another interesting finding is that random retesting of both active and inactive pools can lead to improved estimators. Litvak et al. (1994) and Gastwirth and Johnson (1994) were able to improve their estimators in the blood testing problem by retesting inactive pools.

Zhu et al. (2001) described a trinomial modeling approach that incorporates the phenomenon of blocking and used this model to develop criteria for creating pooling designs. These fully parametric models were also extended by Yi (2002) who considered pairwise blocking probabilities. Xie et al. (2001) used a Bayesian approach for modeling blockers and synergism. Finally, Remlinger et al. (2005) also considered design pooling strategies, but from a completely structure-based approach.

When it comes to designing and analyzing pooling studies for drug discovery, many open questions remain. Single-dose pooling studies, which is an area still in its infancy, have been the focus of this chapter. Multiple-dose pooling studies, which constitute a more mature area of research and application, can bring yet another level of interesting questions and evidence of the utility of pooling; see, for example, Berenbaum (1989).

6   Discussion

Many modern developments and applications point to a bright future for pooling experiments. First, blood testing is ready to support heavy-duty use of pooling studies all across the world. The evidence of success is overwhelming whereas the costs are minimal. Secondly, the drug discovery application still has a long way to go before it is fully developed, but researchers are making great strides. The

ability to uncover synergistic relationships for discovering combination therapies is very exciting and offers many new challenges and possibilities.

References

Archbold, E., Mitchel, S., Hanson, D., Hull, B., Galo, L. M., and Monzon, O. (1991). Serum-pooling strategies for HIV screening: Experiences in Trinidad, Ecuador, and the Phillipines. *VII International Conference on AIDS Abstract Book*, **2**, 331.

Babu, P. G., Saraswathi, N. K., Vaidyanathan, H., and John, T. J. (1993). Reduction of the cost of testing for antibody to human immunodeficiency virus, without losing sensitivity, by pooling sera. *Indian Journal of Medical Research Section A–Infectious Diseases*, **97**, 1–3.

Banks, M. (2000). Automation and technology for HTS in drug development. In *Approaches to High Throughput Toxicity Screening*. Editors: C. K. Atterwill, P. Goldfarb, and W. Purcell, pages 9–29. Taylor and Francis, London.

Behets, F., Bertozzi, S., Kasali, M., Kashamuka, M., Atikala, L., Brown, C., Ryder, R. W., and Quinn, T. C. (1990). Successful use of pooled sera to determine HIV-1 seroprevalence in Zaire with development of cost efficiency models. AIDS, **4**, 737–741.

Berenbaum, M. C. (1989). What is synergy? *Pharmacological Reviews*, **41**, 93–141.

Blower, P., Fligner, M., Verducci J., and Bjoraker, J. (2002). On combining recursive partitioning and simulated annealing to detect groups of biologically active compounds. *Journal of Chemical Information and Computer Sciences*, **42**, 393–404.

Borisy, A. A., Elliott, P. J., Hurst, N. W., Lee, M. S., Lehar, J., Price, E. R., Serbedzija, G., Zimmerman, G. R., Foley, M. A., Stockwell, B. R., and Keith, C. T. (2003). Systematic discovery of multicomponent therapies. *Proceedings of the National Academy of Sciences*, **100**, 7977–7982.

Cahoon-Young, B., Chandler, A., Livermore, T., Gaudino, J., and Benjamin, R. (1989). Sensitivity and specificity of pooled versus individual sera in a human immunodeficiency virus antibody prevalence study. *Journal of Clinical Microbiology*, **27**, 1893–1895.

Chen, C. L. and Swallow, W. H. (1990). Using group testing to estimate a proportion, and to test the binomial model. *Biometrics*, **46**, 1035–1046.

Davey, R. J., Jett, B. W., and Alter, H. J. (1991). Pooling of blood donor sera prior to testing with rapid/simple HIV test kits. *Transfusion*, **31**, 7.

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

Emmanuel J. C., Bassett M. T., Smith H. J., and Jacobs, J. A. (1988). Pooling of sera for human immunodeficiency virus (HIV) testing: An economical method for use in developing countries. *Journal of Clinical Pathology*, **41**, 582–585.

Feller, W. (1957). *An Introduction to Probability Theory and its Applications*. John Wiley and Sons, New York.

Finucan, H. M. (1964). The blood testing problem. *Applied Statistics*, **13**, 43–50.

Fisher, R. A. (1921). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London*, A, **222**, 309–368.

Gastwirth, J. L. and Hammick, P. A. (1989). Estimation of the prevalence of a rare disease, preserving the anonymity of the subjects by group testing: Application to estimating the prevalence of AIDS antibodies in blood donors. *Journal of Statistical Planning and Inference*, **22**, 15–27.

Gastwirth, J. L. and Johnson, W. O. (1994). A two-stage adaptive group-tesing procedure for estimating small proportions. *Journal of the American Statistical Association*, **89**, 972–981.

Gibbs, A. J. and Gower, J. C. (1960). The use of a multiple-transfer method in plant virus transmission studies: Some statistical points arising in the analysis of results. *Annals of Applied Biology*, **48**, 75–83.

Giersiefen, H., Hilgenfeld, R., and Hillisch, A. (2003). Modern methods of drug discovery: An introduction. In *Modern Methods of Drug Discovery*. Editors: A. Hillisch and R. Hilgenfeld. Birkhauser Verlag, Basel and Boston, pages 1–30.

Griffiths, D. A. (1972). A further note on the probability of disease transmission. *Biometrics*, **28**, 1133–1139.

Halvorson, H. O. and Ziegler, N. R. (1933). Application of statistics to problems in bacteriology. *Journal of Bacteriology*, **25**, 101–121.

Hammick, P. A. and Gastwirth, J. L. (1994). Group testing for sensitive characteristics: Extension to higher prevalence levels. *International Statistical Review*, **62**, 319–331.

Hann, M., Hudson, B., Lewell, X., Lifely, R, Miller, L., and Rarnsden, N. (1999). Strategic pooling of compounds for high-throughput screening. *Journal of Chemical Information and Computer Sciences*, **39**, 897–902.

Hughes-Oliver, J. M. (1991). Estimation using group-testing procedures: Adaptive iteration. PhD dissertation. North Carolina State University, Department of Statistics.

Hughes-Oliver, J. M. and Rosenberger, W. F. (2000). Efficient estimation of the prevalence of multiple rare traits. *Biometrika*, **87**, 315–327.

Hughes-Oliver, J. M. and Swallow, W. H. (1994). A two-stage adaptive group-testing procedure for estimating small proportions. *Journal of the American Statistical Association*, **89**, 982–993.

Hung, M. and Swallow, W. (1999). Robustness of group testing in the estimation of proportions. *Biometrics*, **55**, 231–237.

Hung, M. and Swallow, W. (2000). Use of binomial group testing in tests of hypotheses for classification or quantitative covariables. *Biometrics*, **56**, 204–212.

Hwang, F. K. (1984). Robust group testing. *Journal of Quality Technology*, **16**, 189–195.

Kerr, J. D. (1971). The probability of disease transmission. *Biometrics*, **27**, 219–222.

Kline, R. L., Brothers, T. A., Brookmeyer, R., Zeger, S., and Quinn, T. C. (1989). Evaluation of human immunodeficiency virus seroprevalence in population surveys using pooled sera. *Journal of Clinical Microbiology*, **27**, 1449–1452.

Ko, Y. C., Lan, S. J., Chiang, T. A., Yen, Y. Y., and Hsieh, C. C. (1992). Successful use of pooled sera to estimate HIV antibody seroprevalence and eliminate all positive cases. *Asia Pacific Journal of Public Health*, **6**, 146–149.

Lancaster, V. A. and Keller-McNulty, S. (1998). A review of composite sampling methods. *Journal of the American Statistical Association*, **93**, 1216–1230.

Langfeldt, S. A., Hughes-Oliver, J. M., Ghosh, S., and Young, S. S. (1997). Optimal group testing in the presence of blockers. *Institute of Statistics Mimeograph Series 2297*. North Carolina State University, Department of Statistics.

Le, C. T. (1981). A new estimator for infection rates using pools of variable size. *American Journal of Epidemiology*, **14**, 132–136.

Litvak, E., Tu, X. M., and Pagano, M. (1994). Screening for presence of a disease by pooling sera samples. *Journal of the American Statistical Association*, **89**, 424–434.

Loyer, M. (1983). Bad probability, good statistics, and group testing for binomial estimation. *The American Statistician*, **37**, 57–59.

McCrady, M. H. (1915). The numerical interpretation of fermentation-tube results. *Journal of Infectious Diseases*, **17**, 183–212.

McFarland, J. W. and Gans, D. J. (1986). On the significance of clusters in the graphical display of structure-activity data. *Journal of Medicinal Chemistry*, **29**, 505–514.

McMahon, E. J., Fang, C., Layug, L., and Sandler, S. G. (1995). Pooling blood donor samples to reduce the cost of HIV-1 antibody testing. *Vox Sanguinis*, **68**, 215–219.

Niles, W. D. and Coassin, P. J. (2002). Miniaturization technologies for high-throughput biology. *In Integrated Drug Discovery Technologies.* Editors: H.-Y. Mei and A. W. Czarnik, pages 341–364. Marcel Dekker, New York.

Perriens, J. H., Magazani, K., Kapila, N., Konde, M., Selemani, U. Piot, P., and van der Groen, G. (1993). Use of a rapid test and an ELISA for HIV antibody screening of pooled serum samples in Lubumbashi, Zaire. *Journal of Virological Methods*, **41**, 213–221.

Peto, S. (1953), A dose response equation for the invasion of micro-organisms. *Biometrics*, **9**, 320–335.

Phatarfod, R. M. and Sudbury, A. (1994). The use of a square array scheme in blood testing. *Statistics in Medicine*, **13**, 2337–2343.

Raboud, J. M., Sherlock, C., Schecter, M. T., Lepine D. G., and O'Shaughnessy, M. V. (1993). Combining pooling and alternative algorithms in seroprevalence studies. *Journal of Clinical Microbiology*, **31**, 2298–2302.

Remlinger, K. S., Hughes-Oliver, J. M., Young, S. S., and Lam, R. L. H. (2005). Statistical design of pools using optimal coverage and minimal collision. *Technometrics*, in press.

Sanchez, M., Leoro, G., and Archbold, E. (1991). Workload and cost-effectiveness analysis of a pooling method for HIV screening. *VII International Conference on AIDS Abstract Book*, volume 2, page 330.

Seymour, E., Barriga, G., and Stramer, S. L. (1992). Use of both pooled saliva and pooled serum samples for the determination of HIV-1/HIV-2 antibodies by both conventional and rapid EIA techniques. *International Conference on AIDS*, **8**, 194.

Sobel, M. and Elashoff, R. M. (1975). Group testing with a new goal, estimation. *Biometrika*, **62**, 181–193.

Sobel, M. and Groll, P. A. (1959). Group testing to eliminate efficiently all defectives in a binomial sample. *The Bell System Technical Journal*, **38**, 1179–1252.

Soroka, S. D., Granade, T. C., Phillips, S., and Parekh, B. (2003). The use of simple, rapid tests to detect antibodies to human immunodeficiency virus types 1 and 2 in pooled serum specimens. *Journal of Clinical Virology*, **27**, 90–96.

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

Svetnik, V., Liaw, A., Tong, C., Culberson, J.C., Sheridan, R.P., and Feuston, B.P. (2003). Random forest: A classification and regression tool for compound classification and QSAR modeling. *Journal of Chemical Information and Computer Sciences*, **43**, 1947–1958.

Swallow, W. H. (1985). Group testing for estimating infection rates and probability of disease transmission. *Phytopathology*, **75**, 882–889.

Swallow, W. H. (1987). Relative mean squared error and cost considerations in choosing group size for group testing to estimate infection rates and probabilities of disease transmission. *Phytopathology*, **77**, 1376–1381.

Tamashiro, H., Maskill, W., Emmanuel, J., Fauquex, A., Sato, P., and Heymann, D. (1993). Reducing the cost of HIV antibody testing. *Lancet*, **342**(8863), 87–90.

Tan, M., Fang, H. B., Tian, G. L., and Houghton, P. J. (2003). Experimenal design and sample size determination for testing synergism in drug combination studies based on uniform measures. *Statistics in Medicine*, **22**, 2091–2100.

Tebbs, J. M. and Swallow, W. H. (2003a). Estimating ordered binomial proportions with the use of group testing. *Biometrika*, **90**, 471–477.

Tebbs, J. M. and Swallow, W. H. (2003b). More powerful likelihood ratio tests for isotonic binomial proportions. *Biometrical Journal*, **45**, 618–630.

Thompson, K. H. (1962). Estimation of the proportion of vectors in a natural population of insects. *Biometrics*, **18**, 568–578.

Tu, X. M., Litvak, E., and Pagano, M. (1994). Screening tests: Can we get more by doing less? *Statistics in Medicine*, **13**, 1905–1919.

Tu, X. M., Litvak, E., and Pagano, M. (1995). On the information and accuracy of pooled testing in estimating prevalence of a rare disease: Application to HIV screening. *Biometrika*, **82**, 287–297.

Vansteelandt, S., Goetghebeur, E., and Verstraeten, T. (2000). Regression models for disease prevalence with diagnostic tests on pools of serum samples. *Biometrics*, **56**, 1126–1133.

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.

Wein, L. and Zenios, S. (1996). Pooled testing for HIV screening: Capturing the dilution effect. *Operations Research*, **44**, 543–569.

World Health Organization (WHO). (1991). Recommendations for testing for HIV antibody on serum pool. *World Health Organization Weekly Epidemiological Record*, **66**(44), 326–327.

World Health Organization (WHO) (1992). Recommendations for the selection and use of HIV antibody tests. *World Health Organization Weekly Epidemiological Record*, **67**(20), 145–149.

Wilks, S. S. (1962). *Mathematical Statistics*. John Wiley and Sons, New York.

Xie, M. (2001). Regression analysis of group testing samples. *Statistics in Medicine*, **20**, 1957–1969.

Xie, M., Tatsuoka, K., Sacks, J., and Young, S. S. (2001). Group testing with blockers and synergism. *Journal of the American Statistical Association*, **96**, 92–102.

Yi, B. (2002). Nonparametric, parametric and semiparametric models for screening and decoding pools of chemical compounds. Unpublished PhD dissertation. North Carolina State University, Department of Statistics.

Zhu, L. (2000). Statistical decoding and designing of pooling experiments based on chemical structure. Unpublished PhD dissertation. North Carolina State University, Department of Statistics.

Zhu, L., Hughes-Oliver, J. M., and Young, S. S. (2001). Statistical decoding of potent pools based on chemical structure. *Biometrics*, **57**, 922–930.