Pooled Testing for HIV Prevalence Estimation: Exploiting the Dilution Effect

Stefanos A. Zenios
Lawrence M. Wein

#3810-95-MSA March 1995

POOLED TESTING FOR HIV PREVALENCE ESTIMATION: EXPLOITING THE DILUTION EFFECT

Stefanos A. Zenios

Operations Research Center, M.I.T

and

Lawrence M. Wein

Sloan School of Management, M.I.T.

Abstract

We study pooled (or group) testing as a method for estimating the prevalence of HIV; rather than testing each sample individually, this method combines various samples into a pool and then tests the pool. Existing pooled testing procedures estimate the prevalence using dichotomous test outcomes. However, HIV test outcomes are inherently continuous, and their dichotomization may eliminate useful information. To overcome this problem, we develop a parametric procedure that utilizes the continuous outcomes. This procedure employs a hierarchical pooling model and estimates the prevalence using the likelihood equation. The likelihood equation is solved using an iterative algorithm, and a simulation study shows that our procedure yields very accurate estimates for a fraction of the cost of existing procedures.

March 1995

1 Introduction

Estimation of the prevalence of HIV is important for evaluating the current status of the AIDS epidemic and planning effective intervention programs. However, precise estimates are very difficult to obtain because HIV infection is not associated with any clinical symptoms. Because the interval between the infection time and the onset of the clinical symptoms of AIDS is very long, the total number of AIDS cases does not provide information on the recent incidence of HIV. Consequently, population surveys are the primary mode of HIV prevalence estimation.

This method screens an unbiased population sample using an antibody test, such as ELISA, and estimates the prevalence from the number of positive tests. However, even this method has its limitations. Because participants in a survey must give their prior consent, self-selection bias can occur that results in an underestimation of the prevalence (Gill, Adler and Day 1988). Also, the number of tests is frequently limited by financial constraints that adversely affect the accuracy of the estimates.

One possible way to improve the accuracy of the estimates and reduce the self-selection bias is *pooled testing*. The rationale behind pooled testing is simple and intuitive: suppose that a pool is formed from the blood of ten (for example) individuals and is tested with a single antibody test. If the prevalence is sufficiently small then there is a high probability that all ten samples will be HIV negative. Moreover, even if the pool is positive, it will rarely contain more than one infected individual. Under these conditions, a single pooled test provides nearly the same information as ten individual tests, and the pooling method achieves almost the same accuracy as individual testing but for significantly lower cost. Furthermore, the pooling procedure preserves the anonymity of the participating individuals, and therefore can reduce the self-selection bias (Gastwirth and Hammich 1989).

In his seminal paper, Dorfman (1943) showed how pooled testing, which is called *group*

testing in the statistical literature and *composite testing* in the environmental statistics literature, can be employed to efficiently detect the defective members of a population. Group testing was researched aggressively in the subsequent years and a large literature now exists on the topic (see Sobel and Groll 1958 and Johnson, Kotz and Wu 1991 and references therein).

Group testing has also been used to estimate the proportion of the defective members of a population (e.g., Sobel and Elassoff 1975, Chen and Swallow 1989 and Lovison, Gore and Patil 1994). However, these studies assume that tests have no misclassification errors. Motivated by the HIV prevalence estimation problem, Tu, Litvak and Pagano (1994) and Gastwirth and Hammich provide new insights into pooled testing by developing pooling procedures for imperfect tests. Nevertheless, existing studies suffer from two shortcomings. First, they assume that pooling does not affect the sensitivity and specificity of the test. However, if the group size is large, some positive sera may be excessively diluted by negative sera and become undetectable in the pool. Failure to explicitly capture this *dilution effect* may result in prevalence underestimation. Second, they assume that the test results are binary (i.e., infected or non-infected), even though the outcomes of ELISA, known as the optical density readings, are inherently continuous. To implement their pooling procedures, they adopt the dichotomous classification of the test outcomes dictated by the test kit manufacturers; this classification, in addition to eliminating useful information, is specifically designed for individual testing.

The purpose of this paper is to propose a parametric procedure that overcomes these two shortcomings. Our procedure employs the *hierarchical pooling model* of Wein and Zenios (1995), which is described in Section 2, and explicitly captures both the dilution effect and the continuous nature of the test outcomes. This model is used in Section 3 to develop an iterative estimation algorithm that utilizes the continuous optical density

readings. To implement the algorithm, it is necessary to calculate a quantity that cannot be expressed in closed form. We overcome this problem by developing an asymptotic expansion that is described in Section 4. An upper bound on the asymptotic variance is derived in Section 5, which is used in Section 6 to determine the sample size and the pool size that effectively balance the tradeoff between the testing cost and the accuracy of the estimate. A Monte Carlo simulation study is carried out in Section 7 to assess the performance of our estimation procedure. The results indicate that the parametric procedure can achieve more accurate estimates and attain significant cost savings relative to existing pooling procedures. Concluding remarks appear in Section 8.

2 The Hierarchical Pooling Model

ELISA detects HIV antibodies in the blood of infected individuals and then translates the antibody concentration into a continuous quantity, known as the optical density (OD) reading. To capture the continuity of the OD readings, we describe this mechanism as a two-level hierarchical statistical model. The probability density of the antibody concentration is specified at the first level, and the conditional density of the OD reading *given the antibody concentration* is determined at the second level. The dilution effect is captured by assuming that the antibody concentration in a pool equals the average of the individual concentrations.

Readers are referred to our companion paper (Wein and Zenios) for a detailed description of the hierarchical model. In that paper, the hierarchical model was validated on dilution series data and pooled testing data using a generalized linear model. In this section, we briefly summarize our hierarchical model and show how to use it to derive the probability density of pooled OD readings.

Let p denote the unknown HIV prevalence and consider a pool consisting of m individual blood samples. The pool's OD reading and antibody concentration are defined by $X^{(m)}$

and $Y^{(m)}$, respectively. These random variables have probability densities $f_X^{(m)}(x;p,\phi,\gamma)$ and $\pi_Y^{(m)}(y;p)$, where ϕ and γ are nuisance parameters that appear in the second level of the hierarchical model. Without loss of generality, we assume that the OD reading is normalized so as to fall between 0 and 1.

The model's first level specifies the density $\pi_Y^{(m)}(y;p)$. Let $\pi_+(y)$ and $\pi_-(y)$ be the probability density for the antibody concentration of infected and non-infected individuals, respectively, and define $\pi^{*(k,m)} = \pi_+^{*k} * \pi_-^{*(m-k)}$, where $*$ is the convolution operator. Let $\pi_Y^{(k,m)}(y)$ be the conditional density of $Y^{(m)}$ *given that the pool consists of* k *infected and* $m-k$ *non-infected individuals*. Then

$$\pi_Y^{(k,m)}(y) = m\pi^{*\{k,m\}}(my) \tag{1}$$

and the first level of the model states that

$$\pi_Y^{(m)}(y;p) = \sum_{k=0}^{m} \binom{m}{k} p^k (1-p)^{m-k} \pi_Y^{(k,m)}(y). \tag{2}$$

The model's second level specifies the conditional density of the OD reading given the antibody concentration:

$$f(x;\phi,\gamma|y) = \frac{1}{\sqrt{2\pi\phi\frac{y^\gamma}{(1+y^\gamma)^2}}} e^{-\frac{1}{2}\frac{(1+y^\gamma)^2}{\phi y^\gamma}\left(x-\frac{y^\gamma}{1+y^\gamma}\right)^2}, \tag{3}$$

where the nuisance parameters ϕ and γ depend upon the test kit employed. By equations (2) and (3), the density of $X^{(m)}$ is

$$f_X^{(m)}(x;p,\phi,\gamma) = \sum_{k=0}^{m} \binom{m}{k} p^k (1-p)^{m-k} f_X^{(k,m)}(x;\phi,\gamma), \tag{4}$$

where

$$f_X^{(k,m)}(x;\phi,\gamma) = \int_0^\infty \pi_Y^{(k,m)}(y) f(x;\phi,\gamma|y)dy. \tag{5}$$

To complete the description of the model, it remains to determine the parameters ϕ and γ, and the densities π_- and π_+. Because our goal is to estimate p, we assume that

these quantities can be obtained from an off-line analysis of historical data, and that the only unknown parameter is p. The off-line estimation of these quantities is discussed in Subsection 7.2.

It is worth noting that our model not only captures the dilution effect and the continuity of the OD readings, but is also consistent with empirical evidence that identify two sources of variability in ELISA: within sample (due to measurement errors) and between sample. The first level of the model captures the between sample variability. The second level of the model, in particular equation (3), allows repeated measurements from the same sample to be different, and hence captures the within sample variability.

To improve readability, in the remainder of this paper we suppress the notational dependence on the parameters ϕ and γ; for example, we use $f(x|y)$ instead of $f(x;\phi,\gamma|y)$.

3 The Iterative Algorithm

In this section, we develop an algorithm to estimate the prevalence of HIV. This is done by deriving the likelihood equation from the probability density function (4), and then iteratively solving this equation.

Suppose that we collect blood samples from nm individuals, pool them together to form n independent pools of size m, and then test the pools. The raw data are $\mathbf{x} = (x_1, \ldots, x_n)$, where x_i denotes the OD reading of the ith pool. Equation (4) implies that the overall likelihood function is

$$L(\mathbf{x};p) = \prod_{i=1}^{n} \left(\sum_{k=0}^{m} \binom{m}{k} p^k (1-p)^{m-k} f_X^{(k,m)}(x_i) \right). \tag{6}$$

To obtain the maximum likelihood estimate (m.l.e.) $\hat{p}$ of p, we solve the likelihood equation

$$\frac{\partial \log L(\mathbf{x};p)}{\partial p} = 0. \tag{7}$$

It is convenient to define $\tau_k(x;p)$, the conditional probability that a pool contains k infected individuals *given that its OD is* x,

$$\tau_k(x;p) = \frac{\binom{m}{k}p^k(1-p)^{m-k}f_X^{(k,m)}(x)}{\sum_{j=0}^{m}\binom{m}{j}p^j(1-p)^{m-j}f_X^{(j,m)}(x)}. \tag{8}$$

After some tedious but straightforward algebra, equation (7) reduces to

$$\frac{1}{p}\sum_{i=1}^{n}\sum_{k=0}^{m}k\tau_k(x_i;p) - \frac{1}{1-p}\sum_{i=1}^{n}\sum_{k=0}^{m}(m-k)\tau_k(x_i;p) = 0. \tag{9}$$

Because $\sum_{k=0}^{m}\tau_k(x;p) = 1$, expression (9) simplifies to the self-consistency equation

$$nmp = \sum_{i=1}^{n}\sum_{k=0}^{m}k\tau_k(x_i;p). \tag{10}$$

Intuitively, the left side of (10) is the expected number of infected individuals and the right side is the conditional expected number of infected individuals, given the data.

Motivated by (10), we propose the following iterative algorithm, where ϵ denotes the desired tolerance level:

Step 1. Start the procedure with an initial estimate $p^{(o)}$.

Step 2. Obtain improved estimates by taking

$$p^{(s+1)} = \frac{1}{nm}\sum_{i=1}^{n}\sum_{k=0}^{m}k\tau_k(x_i;p^{(s)}). \tag{11}$$

Step 3. Terminate if $|p^{(s+1)} - p^{(s)}| \leq \epsilon$; otherwise, return to Step 2.

To prove that the algorithm converges, we show that it is an Expectation-Maximization (EM) algorithm (see Dempster, Laird and Rubin 1977). The EM algorithm is applicable whenever the raw data can be viewed as incomplete observations from a "complete" data set, and the maximum likelihood estimates for the complete data set are easily obtainable. Here, the complete data set consists of both the observable OD readings, and the unobservable number of infected individuals in the pooled samples. The details are described in Appendix A.

4 Analytical Approximation

The algorithm of Section 3 requires the computation of the conditional probabilities $\tau_k(x;\phi)$ for $k = 0,\ldots,m$. Since we have been unable to obtain these probabilities in closed form, Monte Carlo simulation is employed in Subsection 7.3 to estimate these probabilities. An analytical approximation to these probabilities is derived in this section; this approximation, which assumes large pool sizes, is much less computationally intensive than the simulation procedure, and is used in Section 6 to derive effective sampling designs. The large pool size approximation is derived in Subsection 4.1, and is employed in Subsections 4.2 and 4.3, respectively, to calculate asymptotic expansions for the probability densities defined in equations (4) and (5). A discussion of the appropriateness of the large group size approximation is deferred to Subsection 4.4.

4.1 A Large Pool Size Approximation

Let $Y^{(k,m)}$ denote the conditional antibody concentration *given that the pool consists of k infected and $m-k$ non-infected individuals*; in this subsection, we prove that $Y^{(k,m)}$ converges in distribution to a normal random variable as $k,m \to \infty$. Let μ_+ and μ_- denote, respectively, the mean antibody concentrations of infected and non-infected individuals, and let σ_+ and σ_- be the respective standard deviations. In addition, define $\mu(p) = p\mu_+ + (1-p)\mu_-$.

Proposition 1 *Let $k,m \to \infty$, such that $k/m \to s$. Then,*

$$\sqrt{m}(Y^{(k,m)} - \mu(s)) \xrightarrow{d} N\left(0, s\sigma_+^2 + (1-s)\sigma_-^2\right).$$

Proof: Let Y_i^+ and Y_i^-, $i = 1,2,\ldots$ denote infinite sequences of independent, identically distributed (iid) random variables with respective densities $\pi_+(y)$ and $\pi_-(y)$. The definition

of $Y^{(k,m)}$ implies

$$Y^{(k,m)} = \frac{Y_1^+ + \ldots + Y_k^+ + Y_1^- + \ldots + Y_{m-k}^-}{m}. \tag{12}$$

If we define

$$Y^{(m)+} = \frac{Y_1^+ + \ldots + Y_{[sm]}^+}{[sm]} \tag{13}$$

and

$$Y^{(m)-} = \frac{Y_1^- + \ldots + Y_{m-[sm]}^-}{m - [sm]}, \tag{14}$$

then the multivariate Central Limit Theorem (Billingsley 1987, p. 398) implies that

$$\begin{pmatrix} \sqrt{m}(Y^{(m)+} - \mu_+) \\ \sqrt{m}(Y^{(m)-} - \mu_-) \end{pmatrix} \xrightarrow{d} N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_+^2 & 0 \\ 0 & \sigma_-^2 \end{pmatrix}\right) \quad \text{as} \quad m \to \infty. \tag{15}$$

Furthermore, $Y^{(k,m)} \xrightarrow{d} sY^{(m)+} + (1-s)Y^{(m)-}$ as $m \to \infty$. Proposition 1 now follows from the Converging Together Theorem and the Continuous Mapping Theorem (Billingsley, Theorems 25.4 and 25.7). ■

Therefore, the density of $Y^{(k,m)}$ is

$$\pi_Y^{(k,m)}(y) \approx \sqrt{\frac{m}{2\pi(\frac{k}{m}\sigma_+^2 + (1-\frac{k}{m})\sigma_-^2)}} \exp\left(-\frac{1}{2}m\frac{\left(y - \mu(\frac{k}{m})\right)^2}{\left(\frac{k}{m}\sigma_+^2 + (1-\frac{k}{m})\sigma_-^2\right)}\right), \tag{16}$$

where the error is $O(m^{-2})$ (see Barndorff-Nielsen and Cox 1989, p. 91).

4.2 Approximating $f_X^{(k,m)}(x)$

In this subsection, we use the asymptotic distribution of $Y^{(k,m)}$ to develop an asymptotic expansion for $f_X^{(k,m)}(x)$, which is defined in (5). The density $f_X^{(k,m)}(x)$ can be expressed as

$$f_X^{(k,m)}(x) = \int_{-\infty}^{\infty} \pi_Y^{(k,m)}(y) f(x|y)\, dy. \tag{17}$$

Substituting (16) into (17) and defining $r = k/m$ yields

$$f_X^{(k,m)}(x) = \int_{-\infty}^{\infty} \sqrt{\frac{m}{2\pi(r\sigma_+^2 + (1-r)\sigma_-^2)}} \exp\left(-\frac{1}{2}m\frac{(y-\mu(r))^2}{(r\sigma_+^2 + (1-r)\sigma_-^2}\right) f(x|y)dy + O(m^{-2}). \tag{18}$$

As m increases, the value of the integrand in (18) depends more and more completely on the values of y in the neighborhood of $\mu(r)$. Hence, we replace $f(x|y)$ by its Taylor expansion around $\mu(r)$,

$$f(x|y) = f(x|\mu(r)) + (y-\mu(r))\frac{\partial f}{\partial y}(x|\mu(r)) + \frac{1}{2}(y-\mu(r))^2\frac{\partial^2 f}{\partial y^2}(x|\mu(r)) + O(|y-\mu(r)|^3). \quad (19)$$

If we substitute (19) into (18), then (16) implies that $E(Y^{(k,m)} - \mu(r)) = O(m^{-2})$, $E(Y^{(k,m)} - \mu(r))^2 = \frac{r\sigma_+^2 + (1-r)\sigma_-^2}{m} + O(m^{-2})$ and $E(Y^{(k,m)} - (\mu(r)))^3 = O(m^{-2})$. Thus, we have

$$f_X^{(k,m)}(x) = f(x|\mu(r)) + \frac{r\sigma_+^2 + (1-r)\sigma_-^2}{2m}\frac{\partial^2 f}{\partial y^2}(x|\mu(r)) + O(m^{-2}). \quad (20)$$

The calculation of $\frac{\partial^2 f}{\partial y^2}$ is required to complete the derivation. This calculation was performed using the symbolic differentiation of a computer algebra system (Maple V) and the expression is given in Appendix B.

4.3 Approximating $f_X^{(m)}(x;p)$

The methodology of Subsection 4.2 can also be used to derive an asymptotic expansion for $f_X^{(m)}(x;p)$, which is defined in (4). Let $\sigma(p) = \sqrt{p\sigma_+^2 + (1-p)\sigma_-^2 + p(1-p)(\mu_+ - \mu_-)^2}$. The Central Limit Theorem implies that

$$\sqrt{m}(Y^{(m)} - \mu(p)) \xrightarrow{d} N\left(0, \sigma^2(p)\right) \quad \text{as} \quad m \to \infty. \quad (21)$$

Repeating the intermediate steps of Subsection 4.2 yields

$$f_X^{(m)}(x;p) = f(x|\mu(p)) + \frac{\sigma^2(p)}{2m}\frac{\partial^2 f}{\partial y^2}(x|\mu(p)) + O(m^{-2}). \quad (22)$$

This expression is used in Section 5 to derive an upper bound for the asymptotic variance of the m.l.e. as the number of pools $n \to \infty$.

Objective
To Estimate the Prevalence p.
Known Parameters
$\mu_+, \mu_-, \sigma_+, \sigma_-, \phi, \gamma$
Raw Data
$\{x_1, \ldots, x_n\}$: OD readings from pools of size m
Approximations
$\tilde{f}_X^{(k,m)}(x) = f(x\|\mu(\frac{k}{m})) + \frac{\frac{k}{m}\sigma_+^2 + (1-\frac{k}{m})\sigma_-^2}{2m} \frac{\partial^2 f}{\partial y^2}(x\|\mu(\frac{k}{m}))$.
$\tilde{\tau}_k(x;p) = \frac{\binom{m}{k} p^k (1-p)^{m-k} \tilde{f}_X^{(k,m)}(x)}{\sum_{j=0}^{m} \binom{m}{j} p^k (1-p)^{m-j} \tilde{f}_X^{(j,m)}(x)}$.
The Algorithm
Step 1: Start the procedure with initial estimate $p^{(o)}$.
Step 2: $p^{(s+1)} = \frac{1}{nm} \sum_{i=1}^{n} \sum_{k=0}^{m} k \tilde{\tau}_k(x_i; p^{(s)})$.
Step 3: Terminate if $\|p^{(s+1)} - p^{(s)}\| \leq \epsilon$; otherwise, return to Step 2.

Figure 1: The estimation algorithm for the proposed estimate.

4.4 Discussion of the Approximation

The asymptotic representation for $f_X^{(k,m)}(x)$ can be used to approximate $\tau_k(x;p)$ and implement the estimation algorithm (see Figure 1). However, it is expected that the algorithm will converge to a fixed point that will be different from the m.l.e. To distinguish this fixed point from the m.l.e., we call it the *proposed* estimate and denote it by $\tilde{p}$. Using (20), we can show that the difference between the proposed estimate and the m.l.e. is of $O(m^{-2})$; however, the proof is omitted here and can be found in Zenios (1995).

Regarding the appropriateness of the large group size approximation, we note that the empirical results in Wein and Zenios display a large separation between OD readings for HIV positive and HIV negative individuals. Consequently, our parametric procedure can adopt very large pool sizes without significant loss of information, thereby partially justifying our

large group size approximation. However, the accuracy of the asymptotic expansion (20) depends on the rate that $Y^{(m)+}$ and $Y^{(m)-}$ converge to normal random variables in (15). Following the conventional rule-of-thumb (the sample mean of n iid random variables is approximately normal for $n \geq 15$), we conclude that if the pool size m and the number of infected individuals in the pool, k, satisfy $15 \leq k \leq m-15$, then the normal approximations should be sufficiently accurate for practical purposes.

The condition $k \leq m - 15$ is apt to hold for large pool sizes. Although $k \geq 15$ may be practical for other statistical applications, we would expect k to be much smaller than 15 for HIV prevalence studies. Approximation (15) would hold for small k if $\pi_+(y)$ was approximately normal; however, this does not appear to be the case (see Figure 7 of Wein and Zenios). In conclusion, we do not have a persuasive justification for (15); nevertheless, the large group size approximation developed in this section performs well in the computational study described in Section 7.

5 An Upper Bound on the Asymptotic Variance

This section contains a derivation, which is based on the information inequality, of an upper bound on the variance of the proposed estimate. Let $\mu^{(m)}(p) = E(X^{(m)}|p)$, and let $I_1(\mu^{(m)}(p))$ and $I_1(p)$ denote, respectively, the Fisher's information about $\mu^{(m)}(p)$ and p contained in a single observation from $X^{(m)}$, which is the OD reading of a pool of size m. Because $X^{(m)}$ is an unbiased estimate of $\mu^{(m)}(p)$, the Cramer-Rao lower bound implies that

$$I_1(p) \geq \frac{\left(\frac{d\mu^{(m)}}{dp}(p)\right)^2}{\mathrm{Var}(X^{(m)}|p)}. \tag{23}$$

Since $\mathrm{Var}(\hat{p})$, which is the asymptotic variance of $\hat{p}$ as $n \to \infty$, equals $1/(nI_1(p))$, it follows that

$$\mathrm{Var}(\hat{p}) \leq \frac{\mathrm{Var}(X^{(m)}|p)}{n\left(\frac{d\mu^{(m)}}{dp}(p)\right)^2}. \tag{24}$$

Let us first calculate $\mu^{(m)}(p)$. From (22), we have

$$\mu^{(m)}(p) = \int_0^\infty x f(x|\mu(p))dx + \frac{\sigma^2(p)}{2m}\int_0^\infty x\frac{\partial^2 f}{\partial y^2}(x|\mu(p))dx + O(m^{-2}). \tag{25}$$

Equation (3) implies

$$\int_0^\infty x f(x|y)dx = \frac{y^\gamma}{1+y^\gamma}, \tag{26}$$

and differentiating twice under the integral sign yields

$$\int_0^\infty x\frac{\partial^2 f}{\partial y^2}(x|y)dx = \frac{\gamma(\gamma-1)y^{\gamma-2} - \gamma(\gamma+1)y^{2\gamma-2}}{(1+y^\gamma)^3}. \tag{27}$$

Thus, substituting (26) and (27) into (25) gives

$$\mu^{(m)}(p) = \frac{(\mu(p))^\gamma}{1+(\mu(p))^\gamma} + \frac{\sigma^2(p)}{2m}\frac{\gamma(\gamma-1)\,(\mu(p))^{\gamma-2} - \gamma(\gamma+1)\,(\mu(p))^{2\gamma-2}}{(1+(\mu(p))^\gamma)^3} + O(m^{-2}). \tag{28}$$

Differentiating (28) yields

$$\begin{aligned}\frac{d\mu^{(m)}}{dp}(p) = {} & \frac{\gamma\,(\mu_+-\mu_-)\,(\mu(p))^{\gamma-1}}{(1+(\mu(p))^\gamma)^2} + \\ & \gamma\frac{\sigma^2(p)}{2m}\left(\frac{\gamma(\gamma+1)\,(\mu(p))^{3\gamma-1} - 4\gamma^2\,(\mu(p))^{2\gamma-1} + \gamma(\gamma-1)\,(\mu(p))^{\gamma-1}}{(1+(\mu(p))^\gamma)^4}A + \right. \\ & \frac{(\gamma+1)^2\,(\mu(p))^{3\gamma-2} - (4\gamma^2-2)\,(\mu(p))^{2\gamma-2} - (\gamma+1)^2(\mu(p))^{\gamma-2}}{(1+(\mu(p))^\gamma)^4}B + \\ & \left.\frac{(\gamma+1)(\gamma+2)\,(\mu(p))^{3\gamma-3} - (4\gamma^2-4)\,(\mu(p))^{2\gamma-3} - (\gamma+1)(\gamma+2)\,(\mu(p))^{\gamma-3}}{(1+(\mu(p))^\gamma)^4}C\right),\end{aligned} \tag{29}$$

where $A = \mu_- - \mu_+$, $B = \sigma_+^2 - \sigma_-^2 + \mu_+^2 - \mu_-^2$ and $C = \sigma_-^2\mu_+ - \sigma_+^2\mu_- + \mu_+\mu_-^2 - \mu_+^2\mu_-$.

Carrying out a similar analysis for $\mathrm{Var}(X^{(m)}|p)$ leads to

$$\begin{aligned}\mathrm{Var}(X^{(m)}|p) = {} & \phi\frac{(\mu(p))^\gamma}{(1+(\mu(p))^\gamma)^2} + \frac{\sigma^2(p)}{2m}\left(\frac{\phi(\gamma-1)(\mu(p))^{\gamma-2}}{(1+(\mu(p))^\gamma)^4} - \right. \\ & \left.\frac{(2+(4\phi-6)\gamma+2\gamma^2)\,(\mu(p))^{2\gamma-2} + (\gamma-2)(\phi-2+2\gamma)(\mu(p))^{3\gamma-2}}{(1+(\mu(p))^\gamma)^4}\right).\end{aligned} \tag{30}$$

For brevity of notation, let h be the first term in the right side of equation (29) and l be the coefficient of $\frac{1}{m}$ in (29). Similarly, let v be the first term in the right side of (30), and w

be the coefficient of $\frac{1}{m}$ in (30). Substituting (29) and (30) into (24) and omitting terms of $O(m^{-2})$, we have the upper bound

$$\mathrm{Var}(\hat{p}) \leq \frac{v + \frac{w}{m}}{n\left(h^2 + \frac{2hl}{m}\right)} + O(m^{-2}). \tag{31}$$

We conclude this section with a brief discussion of confidence intervals. The Wald's $1 - \alpha$ confidence interval for p is

$$\hat{p} \pm \frac{z_{\alpha/2}}{\sqrt{nI_1(\hat{p})}}, \tag{32}$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ quantile of the standard normal density. Because a closed form expression for $I_1(p)$ is not available, we could attempt to use the observed information $-\frac{\partial^2 \log L(\mathbf{x};\hat{p})}{\partial p^2}$ instead. Although the details are omitted, the resulting confidence intervals were tested in the computational study in Section 7, and were not sufficiently reliable for practical use (e.g., the actual covering probability for the 95% confidence interval was roughly 85%). Alternatively, we could construct confidence intervals that use the upper bound (31). However, this gives a covering probability that is roughly 99%.

6 Designing Efficient Procedures

To conduct a population survey that adopts the algorithm of Figure 1, we must first specify the number of pools n and the group size m. An efficient design must balance the testing cost with the accuracy of the resulting estimate. The design problem that arises in practice can be posed as one of the following two constrained optimization problems: either choose n and m to achieve a prespecified variance at minimum cost, or choose n and m to minimize the variance of the estimate subject to a prespecified budget constraint. These two problems are closely related (more about that later), and for concreteness we will address the former problem.

Let Δ denote the prespecified variance threshold. Since the true variance of the estimate is not available in closed form, we employ the upper bound of the asymptotic variance derived in (31) as a surrogate. Our resulting design will be conservative, in that the actual variance will be less than or equal to Δ. Of course, the upper bound on the variance is a function of the prevalence, which must be estimated at the design stage.

We assume that the cost of testing a pool of size m is (see Behets et al. 1988 for a detailed discussion)

$$C(m) = \begin{cases} 1 & \text{if } m = 1, \\ em + f & \text{if } m > 1, \end{cases} \tag{33}$$

where e and f are positive constants. The design problem is to choose positive integers n and m to

$$\text{minimize} \quad nC(m) \tag{34}$$

$$\text{subject to} \quad n \geq \frac{v + \frac{w}{m}}{\Delta\left(h^2 + \frac{2hl}{m}\right)}. \tag{35}$$

It is possible to solve this integer programming problem using exhaustive search. However, this is cumbersome and time consuming. Instead, we obtain a closed form solution to the non-integer relaxation of this problem.

Proposition 2 *If* $(wh^2 - 2hlv)(fh^2 - 2ehl) \geq 0$, *then the solution to the non-integer relaxation of (34)-(35) is*

$$m^* = min\left(0, \frac{-2hl + \sqrt{\frac{(wh^2 - 2hlv)(fh^2 - 2ehl)}{ev}}}{h^2}\right), \tag{36}$$

$$n^* = \frac{v + \frac{w}{m^*}}{\Delta\left(h^2 + \frac{2hl}{m^*}\right)}. \tag{37}$$

The proof involves a standard Lagrangian argument and elementary calculus, and is omitted.

Similarly, we can solve the second optimization problem, which is to choose m and n to minimize the upper bound of the asymptotic variance subject to $nC(m) \leq B$, where B is

the prespecified available budget. The solution to the integer relaxation of this problem is given by (36) and

$$n^* = \frac{em^* + f}{B}. \tag{38}$$

Notice that the pool size in (36) is independent of the specified variance threshold and the imposed budget constraint. It can be shown that this pool size also minimizes the cost per unit information (i.e., cost times variance). Although the design in Proposition 2 employs an upper bound on the asymptotic variance and is clearly suboptimal, this result should provide a useful back-of-the-envelope calculation.

7 Simulation Study

The results from a Monte Carlo simulation study are reported in this section. The purpose of this study is to address the following questions:

1. How accurate are the approximations of Section 4?
2. What is the value of employing the continuous OD readings rather than the traditional dichotomous test results?
3. Is it necessary to explicitly incorporate the dilution effect into our analysis?
4. What are the expected benefits from pooled testing?

The first question is addressed by comparing the proposed parametric estimator (p.p.e) to the exact m.l.e. To answer question 2, we develop a class of binary estimators, and compare the best binary estimator to the exact m.l.e. To address question 3, we compare the p.p.e. to two other binary estimators, one that explicitly captures the dilution effect and one that ignores it. We also consider individual testing in order to provide an answer to question 4.

To obtain meaningful results from the simulation study, we employ real data. The data are described in Subsection 7.1 and are used in Subsection 7.2 to estimate the parameters of the simulation model. The various estimators are defined in Subsection 7.3, the results of the simulation study are reported in Subsection 7.4 and a hypothetical application is described in Subsection 7.5.

7.1 Data

We employ two data sets that were collected and tested by the National Reference Laboratory of Australia (NRL). The NRL1 data set consists of the OD readings for 4000 HIV negative and 3000 HIV positive sera. The sera were tested using a commercially available test kit, and the cutoff classifying the test outcomes as positive or negative was 0.05. The NRL2 data set consists of OD readings from 10 infected sera, diluted sequentially to a fixed negative serum to produce a series of 10 two-fold dilutions, with ratios $1:16, 1:32, \ldots, 1:8192$. The samples were tested in duplicate using 10 different test kits. No individuals are common to the two data sets, and the test kit in NRL1 is not one of the 10 kits used in NRL2. Finally, we note that all individual samples in NRL1 were diluted according to the test kit manufacturer's instructions before being tested.

7.2 The Simulation Model

The simulation model randomly generates antibody levels and OD readings for both individual and pooled samples via the hierarchical model defined in Section 2. In this subsection, we specify the model parameters ϕ and γ, and the densities $\pi_+(y)$ and $\pi_-(y)$.

The parameters ϕ and γ are solely dependent on the test kit employed, and can be estimated from dilution series data by fitting the data to a generalized linear model; Wein and Zenios carry out this procedure for the 10 different test kits in NRL2. The ten estimated

values for γ (one for each test kit) ranged from 0.09 to 1.11, and averaged 0.54. Evidently, the value for γ is highly dependent on the test kit employed. Unfortunately, we do not have dilution series data for the test kit employed in the NRL1 data set, which is the data set used to estimate $\pi_+(y)$ and $\pi_-(y)$. For simplicity, we use the value $\gamma = 1.0$ in the simulation model. Sensitivity analysis in Wein and Zenios shows that $\gamma = 0.54$ and $\gamma = 1.0$ yield qualitatively similar results.

We employ the parameter ϕ and the densities $\pi_+(y)$ and $\pi_-(y)$ that are used in Wein and Zenios, and readers are referred to that paper for details; here we provide only a brief summary. Rather than estimate ϕ in a manner similar to the estimation of γ, Wein and Zenios obtain the estimate $\phi = 0.0088$ as a by-product of estimating $\pi_-(y)$ from the NRL1 data. The densities $\pi_+(y)$ and $\pi_-(y)$ are difficult to obtain because the antibody concentration is not an observable quantity. Moreover, these densities are population dependent and may vary if the epidemic is nonstationary. However, if the epidemic is progressing slowly, then these densities can safely be assumed to be constant, and hence do not have to be re-estimated every time a population survey is conducted.

Using a combination of exploratory data analysis and the EM algorithm to estimate $\pi_-(y)$ from the 4000 OD readings for HIV negative individuals in NRL1, Wein and Zenios conclude that the variability of OD readings for HIV negative individuals is due nearly entirely to within sample variability and contains almost no between sample variability. Consequently, for $\pi_-(y)$ we use the degenerate density with mean $\mu_- = 0.0086$ and $\sigma_- = 0.0$. Exploratory data analysis in Wein and Zenios suggests that the within sample variability is substantially larger than the between sample variability for HIV positive individuals, and they make the simplifying assumption that the within sample variability is zero for infected individuals. Thus, $X|Y = E(X|Y)$ for infected individuals, and combining this

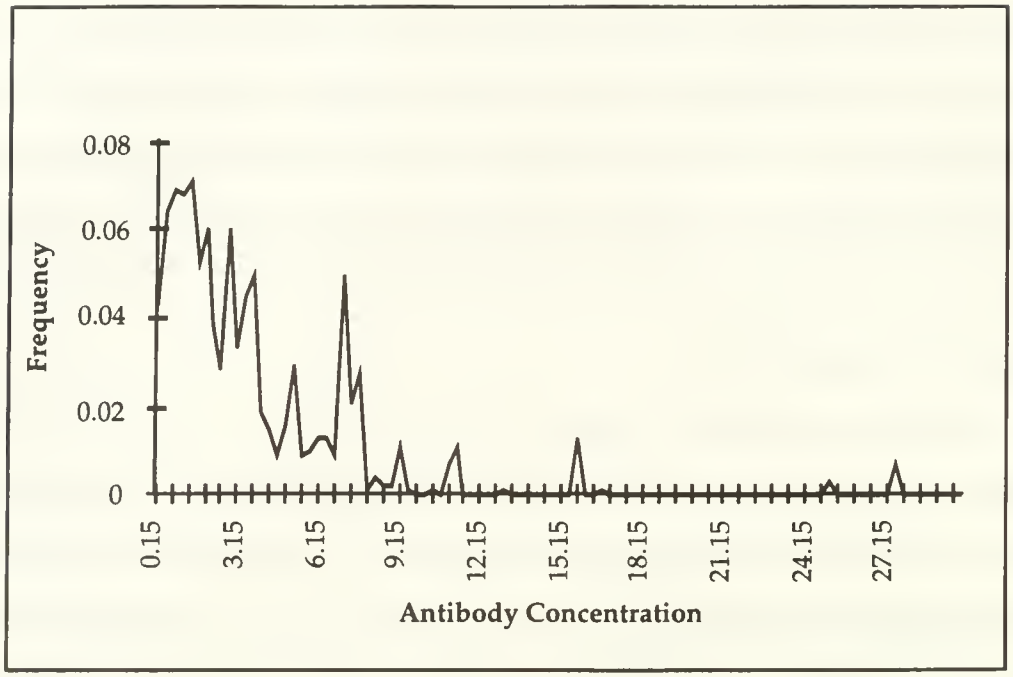

Figure 2: The density $\pi_+(y)$ used in the simulation model.

with equation (3) we obtain

$$Y = \left(\frac{X|Y}{1 - X|Y} \right)^{\frac{1}{\gamma}}. \tag{39}$$

Since our estimate for γ is one, if we let $x_1, \ldots, x_{3000}$ denote the observed OD readings for the 3000 infected individuals in the NRL1 data set, then the density $\pi_+(y)$ is specified by the empirical density of $\left(\frac{x_1}{1-x_1}\right), \ldots, \left(\frac{x_{3000}}{1-x_{3000}}\right)$. This density is displayed in Figure 2. For ease of reference, the parameters of the simulation model are provided in Table 1.

7.3 Estimation Procedures

In this subsection, we describe the procedures that are considered in this simulation study: the proposed parametric estimator, a numerical procedure for the exact m.l.e., three distinct binary estimators that employ binary test results rather than continuous OD readings, and individual testing.

$\mu_- = 0.0086$
$\mu_+ = 3.79$
$\sigma_- = 0.00$
$\sigma_+ = 4.00$
$\phi = 0.0088$
$\gamma = 1.00$

Table 1: The parameters of the simulation model.

Parametric Estimators. The proposed parametric estimator employs the algorithm in Figure 1 with the estimates in Table 1. The p.p.e. employs the pool size given by Proposition 2. The cost parameters are obtained from the field study of Behets et al., and are $e = 0.04$ and $f = 1.35$; details may be found in Wein and Zenios. Figure 2 displays the derived pool size m^* for $0 \leq p \leq 1$. Notice that $m^* \geq 80$ for $p \leq 0.22$. The hierarchical pooling model was validated on pools of size at most 80 in Wein and Zenios and, in fact, 80 appears to be the largest pool size reported in the literature (Cahoon-Young 1992). Consequently, we do not consider pool sizes larger than 80.

To derive the exact m.l.e., we use the algorithm in Figure 1 with $\tilde{f}_X^{(k,m)}(x;\phi)$ replaced by a lookup table representation of the density. The table entries are obtained from a simulated data set as follows. For each value of k and m, a sample of 3000 pools that each consist of k infected and $m - k$ non-infected individuals is generated from the antibody densities $\pi_+(y)$ and $\pi_-(y)$. The OD readings are generated using the density in (3), and the table representation is obtained from the empirical density of the simulated OD readings.

When embedded in the estimation algorithm in Figure 1, these representations produce an estimate that is approximately equal to the m.l.e. We call this estimator the *theoretical parametric estimator* (t.p.e.), and it provides the benchmark to which the performance of the p.p.e. is to be compared. The t.p.e. also uses the pool sizes shown in Figure 2.

Binary Estimators: To assess the value of employing continuous OD readings rather

than traditional binary test results, we also consider the following class of procedures: test n pools of size m, classify each pool outcome as HIV negative or positive using the OD cutoff u, and estimate the prevalence from the total number of positive pools. Let $\text{Se}(m,u)$ and $\text{Sp}(m,u)$ denote the sensitivity and specificity for this class of pooling procedures, and k be the total number of positive pools. The prevalence estimate is

$$\overline{p} = 1 - \left(\frac{\text{Se}(m,u) - k/n}{\text{Se}(m,u) + \text{Sp}(m,u) - 1} \right)^{1/m}, \tag{40}$$

and the asymptotic variance of $\overline{p}$ is

$$\text{Var}(\overline{p}) = \frac{1}{nm^2}(1-p)^{2-2m} \frac{f(m,u,p)\,(1-f(m,u,p))}{(\text{Se}(m,u) + \text{Sp}(m,u) - 1)^2}, \tag{41}$$

where $f(m,u,p) = [1-(1-p)^m]\text{Se}(m,u) + (1-p)^m(1-\text{Sp}(m,u))$. Also, the cost per unit information is

$$F(m,u) = (em+f)\frac{1}{m^2}(1-p)^{2-2m} \frac{f(m,u,p)\,(1-f(m,u,p))}{(\text{Se}(m,u) + \text{Sp}(m,u) - 1)^2}. \tag{42}$$

We consider three different bimary estimators that are based on (40). The first estimator, called the *simple binary estimator* (s.b.e), assumes that that $\text{Se}(m,u)$ and $\text{Sp}(m,u)$ are not a function of m (i.e., there is no dilution effect), and that the cutoff u is determined by the test manufacturer. Consequently, for all m, $\text{Se}(m,u)$ and $\text{Sp}(m,u)$ represent the sensitivity and specificity under individual testing, as reported by the test manufacturer. The pool size m is chosen to minimize the cost per unit information (42); recall that the pool size derived in Section 5 also minimized this quantity. The simple binary estimator was proposed by Tu et al. who were the first to derive equations (40)-(41). To implement this procedure, we set u equal to 0.05, which is the cutoff for the NRL1 data, and use Monte Carlo simulation (with $u = 0.05$ and $m = 1$) to derive very precise estimates of the sensitivity and specificity under individual testing; these two estimates are substituted into equations (42) and (40) to perform the computations described above.

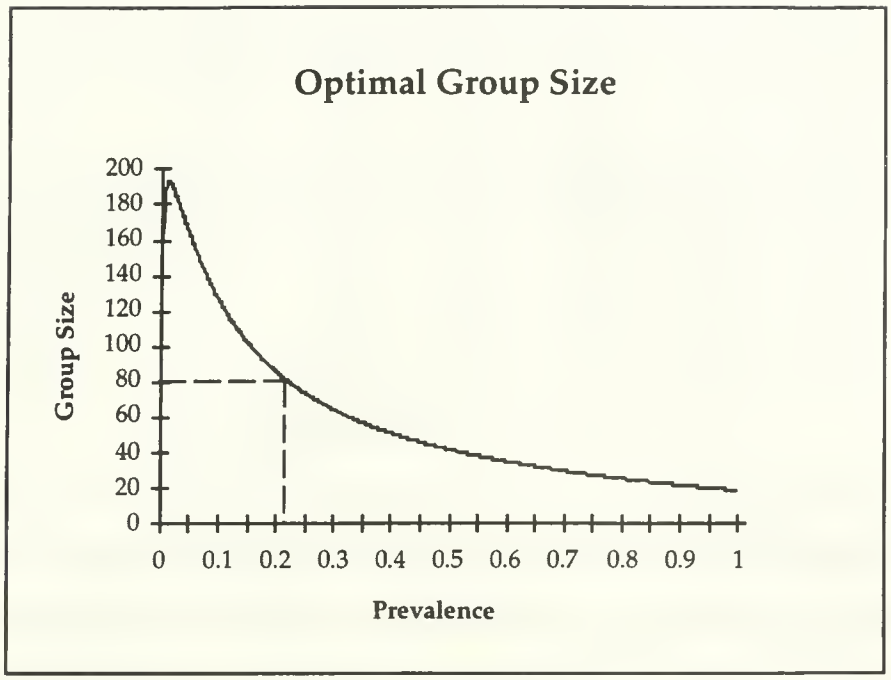

Figure 3: The derived pool size m^*.

The other two binary estimators explicitly incorporate the dilution effect, and hence are expected to outperform the simple binary estimator. Both estimators use the same OD cutoff u and pool size m; the estimators' differentiating characteristics will be defined after we show how these two parameters are calculated. As with the s.b.e., u and m are chosen to minimize (42). However, the quantities $\mathrm{Se}(m,u)$ and $\mathrm{Sp}(m,u)$ in (42) are difficult to estimate for a testing procedure that deviates from the manufacturer's instructions; therefore, we develop approximate closed form expressions for these quantities in order to find the cost-minimizing values of u and m. These expressions are obtained from the simplified pooling model of Wein and Zenios. This model assumes that the OD reading X and the antibody concentration Y are related via $\ln\left(\frac{X}{1-X}\right) = \gamma \ln(Y) + e$, where e is a Gaussian random variable with mean zero and variance σ^2. Proposition 1 and the simplified pooling model lead to an asymptotic approximation for the probability density of $\ln\left(\frac{X^{(k,m)}}{1-X^{(k,m)}}\right)$, where $X^{(k,m)}$ is the

conditional OD reading for a pool of m *given that it contains* k *infected individuals.* This approximation can be used to derive the expressions:

$$\mathrm{Se}(m,u) = \frac{1}{1-(1-p)^m}\left(\sum_{k=1}^{m}\binom{m}{k}p^k(1-p)^{m-k}\left(1-\Phi\left(\frac{u-\gamma\ln(\frac{k}{m}\mu_+ + (1-\frac{k}{m})\mu_-)}{\sqrt{\gamma^2\frac{k\sigma_+^2+(m-k)\sigma_-^2}{(k\mu_++(m-k)\mu_-)^2}+\sigma^2}}\right)\right)\right), \tag{43}$$

$$\mathrm{Sp}(m,u) = \Phi\left(\frac{u-\gamma\ln(\mu_-)}{\sqrt{\sigma^2}}\right). \tag{44}$$

In (43)-(44), we set γ equal to 1, and use the approach in Subsection 7.1 of Wein and Zenios to estimate σ^2 and modify μ_+ and σ_+ so that a conservative estimate for the left tail of $\pi_+(y)$ is obtained. The resulting estimates are $\sigma = 0.42$, $\mu_+ = 2.732$ and $\sigma_+ = 1.3032$. The approximations (43) and (44) are substituted into equations (41) and (42) in steps 2 and 3, respectively, of the following algorithm:

Step 1. Set $m = 1$.

Step 2. Use steepest descent to obtain a cutoff $u(m)$ that achieves a local minimum of $\mathrm{Var}(\overline{p})$.

Step 3. If $F(m,u(m)) \geq F(m-1,u(m-1))$ or $m \geq 80$, stop; otherwise, set $m \leftarrow m+1$ and return to Step 2.

Upon termination, this algorithm gives a locally optimal procedure that satisfies the constraint $m \leq 80$. Table 2 displays the resulting pool sizes and cutoffs for the seven prevalence values that are considered in the simulation study of the next subsection.

Notice that the binary estimate $\overline{p}$ in (40) is a function of $\mathrm{Se}(m,u)$ and $\mathrm{Sp}(m,u)$, where m and u are derived from the three-step algorithm described above. Our two binary estimators differ by how they approximate $\mathrm{Se}(m,u)$ and $\mathrm{Sp}(m,u)$ in (40). The *proposed binary*

p ($\times 10^2$)	Pool Size	Cutoff	Se	Sp	Se	Sp
0.1	66	0.026	0.8704	0.996	0.7385	0.9812
0.5	62	0.022	0.9407	0.988	0.7804	0.9425
1	54	0.021	0.9671	0.9858	0.8562	0.9362
5	20	0.027	0.9978	0.9974	0.9361	0.9857
10	12	0.033	0.9996	0.9992	0.9680	0.9975
15	8	0.038	0.9998	0.9999	0.9780	0.9993
20	7	0.042	0.9999	0.9999	0.9804	1.0000

Table 2: Optimal binary pooling procedures.

estimator (p.b.e.) uses equations (43)-(44) to approximate these two functions; these approximate estimates for the sensitivity and specificity are reported in the fourth and fifth columns of Table 2. We also used Monte Carlo simulation to derive precise estimates of $\text{Se}(m, u)$ and $\text{Sp}(m, u)$. In this case, equation (40) gives the best possible binary estimator, which we call the *theoretical binary estimator*. The simulated estimates for $\text{Se}(m, u)$ and $\text{Sp}(m, u)$ are reported in the last two columns of Table 2.

The three binary estimators serve different purposes. The theoretical estimator is very tedious to implement in practice and is only included to answer question 2. In contrast, both the simple and the proposed binary estimators are rather easy to implement and can be used to address questions 3 and 4.

Individual Testing: Finally, to assess the value of employing pooled testing rather than individual testing, we consider the *individual estimator*, which is identical to the simple binary estimator, except that we choose $m = 1$.

7.4 Simulation Results

In this subsection, we report the simulation results from three experiments. First we compare the p.p.e. to the theoretical parametric estimator, then we compare the theoretical

parametric estimator to the theoretical binary estimator, and finally we compare the p.p.e., the p.b.e., the s.b.e. and the individual estimator.

Experiment 1: A sample of 80,000 antibody concentrations is generated from the probability mixture $p\pi_+(y) + (1-p)\pi_-(y)$. The simulated blood samples are divided into $n = 1000$ pools of size $m = 80$, and the OD readings are generated using the conditional density (3). The p.p.e. and the theoretical parametric estimator are obtained under seven different scenarios of varying prevalence, and the experiment is repeated 400 times for every scenario. The simulation model was implemented in C, on a Sun Sparc Station 20.

For the p.p.e. $\hat{p}$, the estimator mean $E(\hat{p}|p)$ and the estimator variance $\mathrm{Var}(\hat{p}|p)$ are obtained from the simulation experiment. We also compute the relative bias

$$\mathrm{rbias}(\hat{p}) = \frac{p - E(\hat{p}|p)}{p}, \tag{45}$$

the relative variance error

$$\mathrm{revar}(\hat{p}) = \frac{\dfrac{\mathrm{Var}(X^{(m)}|p)}{n\left(\frac{d\mu^{(m)}}{dp}(p)\right)^2} - \mathrm{Var}(\hat{p}|p)}{p} \tag{46}$$

and the mean squared error (mse), $E[(p-\hat{p})^2|p]$. The analogous quantities are also determined for the theoretical parametric estimate $\tilde{p}$.

The results are given in Table 3. Not surprisingly, the theoretical parametric estimator is unbiased. In contrast, the p.p.e. is biased, and tends to understate the true prevalence. The relative bias decreases as the prevalence increases. Also, the relative variance error is positive, which confirms that (31) provides an upper bound on $\mathrm{Var}(\hat{p}|p)$.

Experiment 2: Once again, 80,000 (different) simulated blood samples are generated for the seven scenarios in Table 2. The blood samples are divided into pools of size 80 to obtain the theoretical parametric estimator. In addition, the same blood samples are divided into pool sizes given in Table 2 to derive the theoretical binary estimator.

Prevalence	Performance Measures	Theoretical	Proposed
0.001	mean	0.00096	0.00085
	rbias (%)	0	15
	mse $\times 10^6$	0.02	0.04
	variance $\times 10^6$	0.02	0.02
	revar (%)	34	39
0.005	mean	0.0050	0.0046
	rbias (%)	0	8
	mse $\times 10^6$	0.09	0.27
	variance $\times 10^6$	0.09	0.10
	revar (%)	39	31.8
0.01	mean	0.01	0.0094
	rbias (%)	0	5.7
	mse $\times 10^6$	0.20	0.60
	variance $\times 10^6$	0.20	0.23
	revar (%)	31	19
0.05	mean	0.05	0.048
	rbias (%)	0	3.6
	mse $\times 10^6$	1.25	4.20
	variance $\times 10^6$	1.25	1.23
	revar (%)	14	15
0.1	mean	0.10	0.097
	rbias (%)	0	3.2
	mse $\times 10^6$	2.8	12.9
	variance $\times 10^6$	2.76	2.74
	revar (%)	6	7
0.15	mean	0.149	0.146
	rbias (%)	0	3.0
	mse $\times 10^6$	4.50	24.2
	variance $\times 10^6$	4.22	4.05
	revar (%)	7	10
0.2	mean	0.20	0.194
	rbias (%)	0.1	2.8
	mse $\times 10^6$	6.10	38.30
	variance $\times 10^6$	6.05	5.74
	revar (%)	2	7.4

Table 3: Comparison of the theoretical and proposed parametric estimates.

Confidence Intervals				
p ($\times 10^2$)	Parametric ($\times 10^2$)	Binary ($\times 10^2$)	Relative Efficiency	Relative Cost Efficiency
0.1	0.096 ± 0.0028	0.104 ± 0.0032	1.590	1.395
0.5	0.500 ± 0.0059	0.530 ± 0.0069	1.757	1.479
1	1.000 ± 0.0088	0.993 ± 0.0093	1.673	1.289
5	5.000 ± 0.0219	5.114 ± 0.0255	5.401	2.551
10	10.000 ± 0.0328	10.041 ± 0.0298	5.495	2.217
15	14.900 ± 0.0416	15.000 ± 0.0364	7.633	2.808
20	20.000 ± 0.0484	20.080 ± 0.0468	10.639	3.831

Table 4: Comparison of the theoretical parametric and binary estimators.

Table 4 contains a comparison of the theoretical parametric and binary estimators. The first column of Table 4 gives the prevalence and the next two columns contain 95% confidence intervals for the prevalence estimates under the parametric and binary procedures. The third column shows the *relative efficiency* of the parametric with respect to the binary estimator, and the fourth column gives the corresponding *relative cost efficiency*; these two quantities are defined as follows: Suppose that the parametric estimator uses a pool of size m' and requires n' observations to attain a predetermined variance τ^2. Furthermore, suppose that the binary estimator uses a pool of size m and requires n observations to attain τ^2. The relative efficiency is the ratio n'/n, and the relative cost efficiency is $n'C(m')/(nC(m))$.

Table 4 shows that both estimators are unbiased and their variances are comparable. However, because the parametric estimator adopts a larger pool size, both the efficiency and cost efficiency ratios are always greater than one. Therefore, the parametric estimator can achieve the same variance as the binary estimator for a smaller sample size and testing cost. Furthermore, the benefits from the theoretical parametric estimator appear to be more substantial when the prevalence is high. Intuitively, when the prevalence is high, the binary estimator is forced to adopt moderate pool sizes in order to minimize the probability that a

pool contains more than one infected individual; pools that contain more than one infected individual *hide* information on the true prevalence. In contrast, the theoretical parametric estimator has the flexibility to employ substantially larger pool sizes because it can infer the total number of infected individuals in each pool from the OD reading. Therefore, the parametric estimator can adopt a substantially larger pool size and still achieve the same variance as the binary procedure. This situation changes when the prevalence is low because then the optimal group size for the binary estimator is close to 80, and the group size for the parametric procedure is exactly 80; recall that we have introduced the restriction $m \leq 80$. Hence, the upper bound restricts the flexibility of the parametric procedure and forces it to behave like the binary procedure.

This experiment confirms our intuition that the continuous OD readings can be used to produce very precise estimates. Unfortunately, the estimators in this experiment are hard to employ in practice. In the next experiment we use more practical estimators that contain several approximations. Although these estimators are expected to be biased, we will see that the bias is offset by their high efficiency, and the procedures that utilize the hierarchical pooling model produce very precise estimates.

Experiment 3: Once again, 80,000 (different) simulated blood samples are generated for the seven scenarios in Table 2. The blood samples are divided into pools of size 80 to obtain the proposed parametric estimator. The same blood samples are also divided into the pool sizes given in Table 2 to derive the proposed binary estimator. In addition, these blood samples are divided into the pool sizes dictated by the simple binary procedure in order to derive the s.b.e. Finally, the samples are also tested individually to calculate the individual estimator.

Table 5 gives the mean squared error (mse) for the four estimators. We observe that the mse for the simple and individual estimators are from 4 to 40 times larger than the mse

p ($\times 10^2$)	Individual	s.b.e.	p.b.e.	p.p.e.
0.1	1.000	0.850	0.028	0.046
0.5	4.000	10.464	0.073	0.216
1	9.000	35.705	0.465	0.471
5	59.983	91.113	14.278	4.082
10	123.936	79.731	30.211	10.902
15	190.891	74.199	38.991	20.606
20	220.886	60.135	59.618	38.354

Table 5: Comparison of the mean squared errors ($\times 10^6$) of the individual, simple binary, proposed binary and proposed parametric estimators.

of the two new estimators. By failing to incorporate the dilution effect, the simple binary estimator introduces a substantial bias that dramatically inflates the mse. Although the individual estimator is the only unbiased estimator of the four, its variance is much larger than the estimators that capture the dilution effect. Hence, pooled testing, if used properly, can provide very precise estimates for the prevalence of HIV.

Now we compare the two practical estimators: the p.p.e and the p.b.e. Table 6 gives 95% confidence intervals for the prevalence estimates together with the relative efficiency of the p.p.e. with respect to the p.b.e., and the relative cost efficiency. Surprisingly, the binary procedure outperforms the parametric procedure when the prevalence is less than or equal to 0.01. Table 4 shows that, at least in theory, the parametric procedure should be more efficient than the binary procedure. However, its relative efficiency in this table is rather small when $p \leq 0.01$. It appears that the proposed approximations are more accurate in the binary setting than the parametric setting (at least when the prevalence is small), and this improved accuracy more than offsets the difference in Table 4 when the prevalence is sufficiently small.

We also carried out all three experiments with sample sizes of 800 and 8000. The

Confidence Intervals				
p ($\times 10^2$)	p.p.e. ($\times 10^2$)	p.b.e. ($\times 10^2$)	Relative Efficiency	Relative Cost Efficiency
0.1	0.085 ± 0.0030	0.110 ± 0.0026	0.911	0.798
0.5	0.465 ± 0.0060	0.500 ± 0.0053	1.007	0.847
1	0.949 ± 0.0090	0.944 ± 0.0075	1.028	0.794
5	4.832 ± 0.0220	4.638 ± 0.0214	3.774	1.783
10	9.713 ± 0.0320	9.468 ± 0.0267	4.651	1.866
15	14.592 ± 0.0390	14.400 ± 0.0339	7.575	2.770
20	19.429 ± 0.0470	19.260 ± 0.0432	9.615	3.460

Table 6: Comparison of the p.p.e. and p.b.e.

resulting point estimates and efficiency ratios under the smaller sample sizes were nearly identical to the corresponding quantities in Tables 3-6; however, as expected, the sample size of 80,000 gave more precise estimates for the various quantities. It is worth noting that the bias of the proposed estimator is less than 0.006 for all seven scenarios and all three sample sizes. Because such a small error is unlikely to significantly affect any public health policy, the proposed estimator appears to be sufficiently accurate for policy making purposes.

In summary, our results demonstrate that in order to obtain the most precise prevalence estimates it is necessary to adopt a pooled testing strategy that exploits the dilution effect. As a rough guideline, the proposed binary estimator should be adopted when the prevalence is low, and the p.p.e. should be adopted when the prevalence is higher, where the cutoff is somewhere between 1% and 5%.

7.5 Application

To better estimate the benefits that are achievable from our proposed procedure, we consider the following hypothetical application: it is known that the prevalence is close to 5%, and three options are under consideration, the individual estimator, the proposed binary esti-

mator and the proposed parametric estimator. A budget constraint is imposed such that, if individual testing is adopted, at most 1000 samples can be tested. The objective is to employ the procedure that will give the smallest mse.

Individual estimator: Here $n = 1000$ and $m = 1$. From the results of experiment 3 (not shown), the variance is expected to be 6.2×10^{-4}. Because the individual estimator is unbiased, this coincides with the mse.

Proposed binary estimator: From Table 2, $m = 20$, and to satisfy the budget constraint $n = \frac{1000}{1.35+0.04(20)} \approx 465$. From Table 6, the variance is $\left(\sqrt{100}\frac{0.000214}{1.96}\right)^2 \frac{80000/20}{465} = 1.00 \times 10^{-5}$, and the mse is $(0.04636 - 0.05)^2 + 1.00 \times 10^{-5} = 2.30 \times 10^{-5}$.

Proposed parametric estimator: Here $m = 80$, and to satisfy the budget constraint $n = \frac{1000}{1.35+0.04(80)} \approx 220$. From Table 6, the variance is expected to be $\left(\sqrt{100}\frac{0.00022}{1.96}\right)^2 \frac{80000/80}{220} = 5.72 \times 10^{-6}$, and the mse is $(0.04832 - 0.05)^2 + 5.72 \times 10^{-6} = 8.54 \times 10^{-6}$.

Hence, the proposed parametric estimator has the smallest mse and is the preferred procedure.

8 Concluding Remarks

We have developed a parametric procedure to estimate the prevalence of HIV from pooled samples. Our approach is novel in that it captures the dilution effect and estimates the prevalence directly from the continuous OD readings. This procedure was developed specifically for the HIV estimation problem but is applicable whenever liquids or gases are pooled together and tested for estimation purposes; many applications can be found in industrial or environmental quality control.

The procedure was tested on simulated data that were generated from actual blood samples, and the results indicate that it is more accurate and roughly an order-of-magnitude more efficient than existing binary pooling procedures. As expected, the benefits from this

procedure are larger when the prevalence is high. We also derived a new binary procedure that explicitly accounts for the dilution effect. When the prevalence is above 5% our proposed parametric estimator was several times more efficient than the proposed binary estimator. However, the approximations embedded in our procedures appear to more accurate in the binary setting than in the continuous OD setting, and when the prevalence is below 1% the proposed binary procedure performs better than proposed parametric procedure.

In conclusion, the numerical results provide strong evidence supporting the adoption of our parametric pooled testing procedures for population surveys aimed at estimating HIV prevalence. The next step is to determine whether the conclusions of this paper are confirmed when the procedure is tested on real data.

Acknowledgement

This research is supported by National Science Foundation grant DDM-9057297 and American Foundations for AIDS Research (AmFAR) grant 02100-15-RG. The authors are grateful to Elizabeth Dax for providing the data, and John Karon and Glenn Satten for helpful discussions.

Appendix

A Convergence Proof

In this appendix, we show that the estimation algorithm in Section 3 is an Expectation-Maximization (EM) algorithm. Let us define the unobserved vector $\mathbf{z} = (z_1, \ldots, z_n)$, where z_i is the number of infected individuals in pool i, and view the raw data $\mathbf{x}$ as the incomplete observations from the complete data set $(\mathbf{x}, \mathbf{z})$. Hence, the complete log-likelihood is given

by

$$l_c(\mathbf{x}, \mathbf{z}; p) = \sum_{i=1}^{n} \sum_{k=0}^{m} I(z_i = k) \left[\log \left(\binom{m}{k} p^k (1-p)^{m-k} \right) + \log \left(f_X^{(k,m)}(x_i) \right) \right]. \quad (47)$$

The expectation step of the EM algorithm calculates $E\left[l_c(\mathbf{x}, \mathbf{z}; p)|\mathbf{x}, p^{(s)}\right]$. Because

$$E\left[I(z_i = k)|x; p^{(s)}\right] = \tau_k(x_i; p^{(s)}), \quad (48)$$

it follows from equation (47) that

$$E\left[l_c(\mathbf{x}, \mathbf{z}; p)|\mathbf{x}, p^{(s)}\right] = \sum_{i=1}^{n} \sum_{k=0}^{m} \tau_k(x_i; p^{(s)}) \left[\log \left(\binom{m}{k} p^k (1-p)^{m-k} \right) + \log \left(f_X^{(k,m)}(x_i) \right) \right]. \quad (49)$$

The maximization step updates $p^{(s)}$ using

$$p^{(s+1)} = \arg\max_p E\left[l_c(\mathbf{x}, \mathbf{z}; p)|\mathbf{x}, p^{(s)}\right]. \quad (50)$$

The function $E\left[l_c(\mathbf{x}, \mathbf{z}; p)|x, p^{(s)}\right]$ is concave in p, and hence $p^{(s+1)}$ is the unique solution to the first-order optimality condition

$$\frac{\partial E\left[l_c(\mathbf{x}, \mathbf{z}; p)|\mathbf{x}, p^{(s)}\right]}{\partial p} = 0, \quad (51)$$

which, after some algebra, gives

$$p^{(s+1)} = \frac{1}{nm} \sum_{i=1}^{n} \sum_{k=0}^{m} k \tau_k(x_i; p^{(s)}). \quad (52)$$

The equivalence of (52) and (11) establishes that the iterative procedure is the EM algorithm. To prove that the algorithm converges to a fixed point, we need the following fact.

Theorem 2 (Dempster et al.): *If the incomplete log-likelihood is bounded from above and*

$$E\left[l_c(\mathbf{x}, \mathbf{z}; p^{(s+1)})|x, p^{(s)}\right] - E\left[l_c(\mathbf{x}, \mathbf{z}; p^{(s)})|x, p^{(s)}\right] \geq \lambda(p^{(s+1)} - p^{(s)})^2 \quad (53)$$

for some scalar λ and all $p^{(s)}$, then the sequence $p^{(s)}$ converges to some p^* in $[0, 1]$.

To show that $L(\mathbf{x};p)$ is bounded from above in our problem, we define

$$M = \sup_y \{\max(\pi_+(y), \pi_-(y))\}.$$

Then $\pi^{*\{k,m\}}(y) \leq M^m$ and from (1), $\pi_Y^{(k,m)}(y) \leq mM^m$. Thus, it follows from (6) that

$$L(\mathbf{x};p) \leq [mM^m]^n. \tag{54}$$

To establish (53), we use (49) to obtain

$$\begin{aligned} &E\left[l_c(\mathbf{x},\mathbf{z};p^{(s+1)})|x,p^{(s)}\right] - E\left[l_c(\mathbf{x},\mathbf{z};p^{(s)})|x,p^{(s)}\right] = \\ &\sum_{i=1}^{n}\sum_{k=0}^{m} \tau_k(x_i;p^{(s)}) \left(k\log p^{(s+1)} + (m-k)\log(1-p^{(s+1)})\right) - \\ &\sum_{i=1}^{n}\sum_{k=0}^{m} \tau_k(x_i;p^{(s)}) \left(k\log p^{s} + (m-k)\log(1-p^{s})\right). \end{aligned} \tag{55}$$

Since $\sum_{i=1}^{n}\sum_{k=0}^{m} k\tau_k(x_i;p^{(s)}) = nmp^{(s+1)}$, the right side of (55) reduces to

$$nm\left(p^{(s+1)}\log\left(\frac{p^{(s+1)}}{p^{(s)}}\right) + (1-p^{(s+1)})\log\left(\frac{1-p^{(s+1)}}{1-p^{(s)}}\right)\right). \tag{56}$$

To complete the convergence proof, we need the following lemma.

Lemma 1 *For all* $(x,y) \in [0,1]^2$,

$$x\log\left(\frac{x}{y}\right) + (1-x)\log\left(\frac{1-x}{1-y}\right) \geq (x-y)^2. \tag{57}$$

Proof: Fix $y \in (0,1)$ and define $h(x) = x\log\left(\frac{x}{y}\right) + (1-x)\log\left(\frac{1-x}{1-y}\right) - (x-y)^2$. Setting $h'(x) = 0$ gives

$$\log\left(\frac{x(1-y)}{y(1-x)}\right) = 2(x-y), \tag{58}$$

which is satisfied by $x = y$. Also, $h''(x) = \frac{1}{x(1-x)} - 2 \geq 0$. Therefore, $h(x)$ is convex and is minimized when $x = y$. Because $h(y) = 0$, it follows that

$$x\log\left(\frac{x}{y}\right) + (1-x)\log\left(\frac{1-x}{1-y}\right) \geq (x-y)^2 \quad \text{for all} \quad x \in (0,1). \tag{59}$$

Because y is arbitrary, relation (59) holds for every x and y in $(0,1)$. The result extends to the closed interval $[0,1]$ by continuity. ■

Lemma 1 and (56) imply that

$$E\left[l_c(\mathbf{x},\mathbf{z};p^{(s+1)})|x,p^{(s)}\right] - E\left[l_c(\mathbf{x},\mathbf{z};p^{(s)})|x,p^{(s)}\right] \geq nm(p^{(s)}-p^{(s+1)})^2 \text{ for all } p^*. \quad (60)$$

Therefore, the conditions of Theorem 2 hold and the algorithm converges to a fixed point p^*.

B Partial Derivative $\frac{\partial^2 f}{\partial y^2}(x|y)$

Define the constants

$$\begin{aligned}
A &= 4\gamma(x-1)^2, \\
B &= [4\gamma(x-1-\phi)+(1-\gamma)\phi](x-1), \\
C &= (4\phi-8\gamma x)(x-1)+\phi^2(\gamma-2), \\
D &= x(4\gamma x-4\phi(1+2\gamma)), \\
E &= 4\gamma x^2, \\
F &= -8\gamma x^2+(8\gamma-4\phi)x+\phi^2(\gamma+2).
\end{aligned}$$

Then

$$\frac{\partial^2 f}{\partial y^2}(x|y) = \frac{\gamma}{4\phi^2 y^2(1+y^\gamma)}\left(Ay^{3\gamma}+By^{2\gamma}+Cy^\gamma+D\frac{1}{y^\gamma}+E\frac{1}{y^{2\gamma}}+F\right)f(x|y). \quad (61)$$

References

Barndorff-Nielsen, O.E., and Cox, D.R. (1989), *Asymptotic Techniques for use in statistics*, Chapman and Hall: London.

Becker, N.G., Watson, L.F. and Carlin, J.B. (1991), "A method of non-parametric back-projection and its application to AIDS data", *Statistics in Medicine* **10**, 1527-1542.

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

Billingsley, P. (1987), *Probability and Measure*. Wiley: New York.

Cahoon-Young, B. (1992), "Optimal pool size for determination of HIV prevalence in low risk populations", presented at the HIV/AIDS Surveillance Workshop, South San Fransisco, CA.

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.

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977), "Maximum likelihood from incomplete data via the EM algorithm", (with Discussion), *Journal of the Royal Statistical Society, Series B* **339**, 1-22.

Dorfman, R. (1943), "The detection of defective members of large populations", *Annals of Mathematical Statistics* **44**, 436-441.

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.

Gills, O.N, Adler, M.W. and Day, N.E. (1989), "Monitoring the prevalence of HIV", *British Medical Journal* **299**, 1295-1299.

Hastie, T.J. and Pregibon, D. (1992), "Generalized Linear Models", in *Statistical Models in S,* eds. J.M. Chambers and T.J. Hastie, 1 Wadsworth & Brooks/Cole: California, pp. 195-247.

Johnson, N.L., Kotz, S., and Wu, X. (1991), *Inspection Errors for Attributes in Quality Control*, Chapman and Hall: London.

Lovison, G., Gore, S.D. and Patil, G.P. (1994), "Design and analysis of composite sampling procedures: A review", in *Handbook of Statistics, Vol. 12,* eds. G. P. Patil and C. R. Rao, Elsevier Science: Amsterdam, pp. 103-166.

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

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

Wein, L.M. and Zenios, S.A. (1995), "Pooled testing for HIV screening: Capturing the dilution effect", Working Paper, Sloan School of Management, MIT, Cambridge, MA.

Zenios, S. A. "Health Care Applications of Stochastic Optimal Control", unpublished Ph.D. dissertation, Operations Research Center, MIT, Cambridge, MA, in preparation.