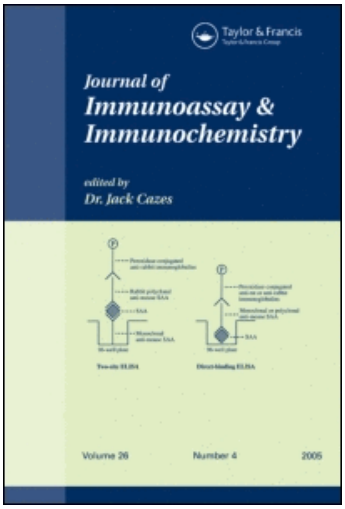

Journal of Immunoassay and Immunochemistry

Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713597271

Beyond Simple Pooling for HIV Screening

Leh-Hun Gwa[a]; Chung-Cheng Hsieh[a]; Yuan-Ching Ko[b]; Shou-Jen Lan[ac]
[a] The Department of Epidemiology, Harvard School of Public Health, Boston, U.S.A. [b] Institute of Public Health, College of Medicine, National Taiwan University, Taipei, Taiwan [c] School of Public Health, Kaohsiung Medical College, Kaohsiung, Taiwan

To cite this Article Gwa, Leh-Hun , Hsieh, Chung-Cheng , Ko, Yuan-Ching and Lan, Shou-Jen(1992) 'Beyond Simple Pooling for HIV Screening', Journal of Immunoassay and Immunochemistry, 13: 4, 545 — 557
To link to this Article: DOI: 10.1080/15321819208019835
URL: http://dx.doi.org/10.1080/15321819208019835

JOURNAL OF IMMUNOASSAY, 13(4), 545-557 (1992)

Beyond Simple Pooling for HIV Screening

Leh-Hun Gwa[1], Chung-Cheng Hsieh[1],
Yuan-Ching Ko[2], Shou-Jen Lan[1,3]

[1]The Department of Epidemiology, Harvard School of Public Health, Boston, U.S.A.; [2]Institute of Public Health, College of Medicine, National Taiwan University, Taipei, Taiwan; [3]the School of Public Health, Kaohsiung Medical College, Kaohsiung, Taiwan.

Abstract

We discuss some theoretical features underlying the successful uses of pooling in testing HIV seroprevalence. In particular it is shown that there is a scaling relation for the distribution of positive sera among the pools. A multi-stage pooling method consisting of repeatedly halving the positive pools is proposed. Concentrating on the number of tests required for screening all positive individuals the method is shown to be highly efficient in low prevalence situations.

Introduction

The AIDS epidemic has taken a terrible toll in Africa (1,2) and the outcome of this disease may have on the world is uncertain. It is therefore important that large scale testing of HIV virus be a viable option to public health officials, to facilitate early treatment of affected individuals and to prevent further spreading of the disease. Several investigation on the feasibility of using pooled sera to reduce the costs of large scale testing have been done, and some encouraging results have been found (4–7). Therefore, it is worthwhile to study pooling procedures further and to find efficient procedures to be used in practice.

The main obstacle to screening in a general population has been the high cost of such testing. This is especially true in less affluent countries (3). The cost efficiency of a pooling method depends on many variable factors, such as the cost of test material, the wage of a technician, the cost for constituting pools, and so on. Many of these considerations vary from place to place and even from time to time, as improvements in testing methods, procedures and training are implemented. Thus results for studies on the cost of a pooling procedure are not likely to be universally valid. One factor which can be studied independently of all the unstable factors is the expected number of tests for a pooling procedure. We shall focus exclusively on this factor in the following discussions. Our main purpose is to propose a multi-stage pooling procedure — the bisection procedure. We shall make an excursion and discuss a well-known formula in probability. The purpose is two fold: 1. To illustrate its relationship with an often-used formula for computing prevalence. 2. It is indispensable for analyzing the bisection procedure later.

The Fisher Formula

In Ref. (5) seroprevalence rate was estimated by setting the observed proportion of negative pools equal to the expected proportion of negative pools with the equation

$$\left(1 - \frac{S}{N}\right) = (1 - P)^A \qquad [1]$$

where P is the prevalence, S the number of positive pools, N the number of pools and A the number of sera in a pool. The computed prevalence is then compared to the rate of positive sera in the samples. Strictly speaking, Eq. [1] is an approximation, reflected in its usually being used in conjunction with a variance formula. The right-hand side of [1] assumes that the probability of being negative for each sera in a pool is independent; in reality, a finite part is being selected from a finite universe. Nevertheless, the approximation may be excellent if the part is negligibly small compared to the universe. However, occasions may arise when one will need to substitute the Fisher formula (8), which has no restriction on the sizes of the parts or the universe: Given a total of T objects containing T_+ positives and $T_- = T - T_+$ negatives, if one randomly picks A objects, then the probability that

one picks n positives and $k = A - n$ negatives is

$$P(n) = C_n^{T_+} C_k^{T_-} / C_A^T \qquad [2]$$

where $C_b^a = a!/b!(a-b)!$ is the number of ways to select b objects from a objects. In the context of pooled sera, $T = NA$ would be the total number of sera, T_+ (T_-) the total number of positive (negative) sera and n (k) the number of positive (negative) sera in a pool of size A. According to this formula the probability for finding negative pools is then

$$P(0) = \prod_{i=0}^{A-1} \frac{T_- - i}{T - i} \approx (1-P)^A \left(1 - \frac{PA}{2T}\frac{A-1}{1-P}\right) \qquad [3]$$

where only the order T^{-1} term has been retained in a large T expansion in the last part of the equation. Thus it is understandable why highly accurate results were obtained in Ref. (5) using Eq. [1] — Because with $T = 8000$, $A = 10$ and $P = 2.44\%$, the leading correction in Eq. [3] is less than 0.02%. Thus the prevalence computed from $P(0)$ would be indistinguishable from that computed using the simpler $(1-P)^A$.

Eq. [2] provides not only the expected number of negative pools, but also the expected number of pools containing one, two, etc. positive sera. Again, when the population is much greater than the pool size, a simpler formula generalizing Eq. [1] may be used instead, namely,

$$P(n) \approx C_n^A (1-P)^{A-n} P^n \equiv F(P; A, n) \; . \qquad [4]$$

For example, in the case studied in Ref. (5) , the number of pools containing 0,1,2,3,4,... positive sera were found to be 626, 155, 17, 2, 0, ... (The pools with one positive and one indeterminate has been counted as one positive for simplicity.) If one computes the expected number of pools with n positive sera, i.e., $800 \times F(2.44\%; 10, n)$ and Eq. [4], one finds 624.9, 156.3, 17.6, 1.2, 0.05, ... Thus the formula proves to yield an excellent estimate on the distribution of the pools.

Distribution of Initial Pools

Detailed information on the outcome of an initial pooling is valuable for planning a multi-stage pooling procedure. The basic idea behind any pooling procedure for screening

purposes is the expectation that a large number of tests spent on the individual negative sera will be avoided by a much smaller number of tests spent on negative pools. Whether this is the case or not depends on the pool size, the prevalence and the procedure. Here we have in mind a large scale screening program performed on the general population so that prevalence is very small and the universe is large, much larger than any technically feasible pool size. Under the circumstances the function $F(P; A, n)$ will be sufficiently accurate for computing the probabilities of finding 1-serum pools, 2-sera pools, etc. Therefore we need to have a general understanding of the behavior of $F(P; A, n)$. It turns out that there is a nice scaling behavior for this function, so that a single figure suffices to demonstrate the function for various pool sizes; this is shown in Fig. 1. The figure is drawn for $F(P; A, n)$ vs. $(A/16)P$ and it is clear that the curves for different A's nearly fall on the same curve. The horizontal axis is the actual prevalence for $A = 16$, but, for example, is $2P$ for $A = 32$ and hence the range of P covered in the figure is halved, i.e., from 0 to 5%. Within this range of scaled prevalence, the probability of finding pools with many positive sera fall off rapidly, which means that most of the positive pools contain very few positive sera. In such a case, many tests will be wasted on the negative sera if individual testing is done on all positive pools.

The Bisection Procedure

In the following we describe the bisection procedure, a multi-stage pooling procedure for screening purpose. The bisection procedure starts by choosing a pool size of the form 2^m. The initial pooling consists of forming pools of this size. testing each pool and retaining the positive pools; this step is common to all pooling procedures, and its analysis is given in the previous section. After this initial step, the positive pools are then repooled by dividing each pool into two pools for the next stage testing. By "dividing" we mean making two half-size pools from the same set of sera in a positive pool. Now there must be at least one positive serum in one of the two pools, therefore, if a negative result is found in the first of the two pools, the second one must be positive and, hence, no testing is needed for the second pool. This step is then repeated until all individual positive sera have been

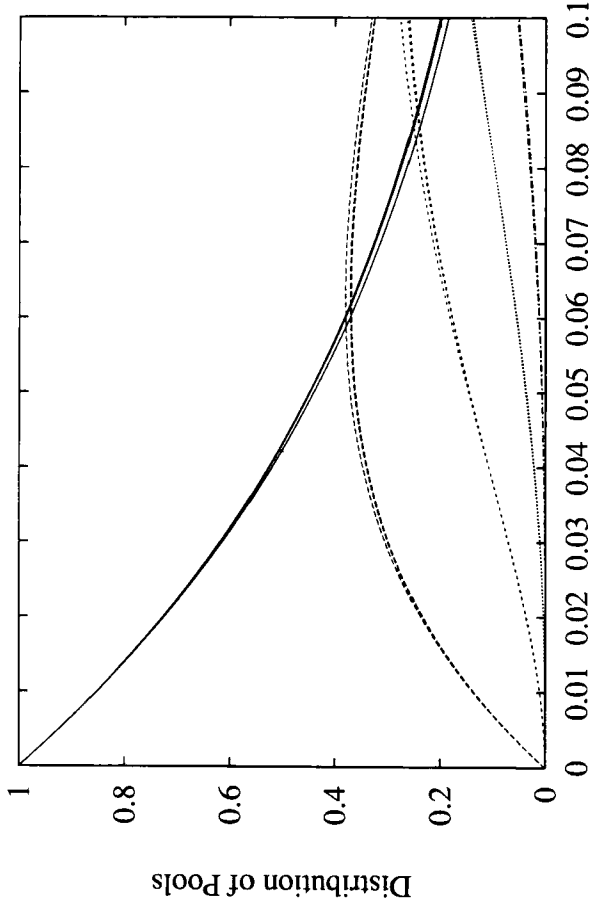

Fig. 1. Expected distribution of pools, $F(P; A, n)$, as a function of prevalence scaled by the pool size, $(A/16)P$. The different groups of curves from top to bottom (on the left part of the figure) are for pools with $n = 0, 1, 2, 3, 4$ positive sera, respectively. The three nearly overlapping sets of curves are for pool size $A = 16, 64, 256$, with the $A = 16$ curves slightly distinct.

identified. To be precise, we shall refer to the following steps as *stage m*: (1) Take a pool of size 2^m and divide into two pools of size 2^{m-1}. (2) Test one pool. If it is positive, retain it and test the other pool, which is retained if also positive; if it is negative, retain the other pool. Thus the bisection procedure requires a total of $m+1$ stages for an initial pool size 2^m; the initial stage is the ordinary pooling step and the particular bisections start from *stage m* and goes down to *stage 1*. At each stage, one starts with pools of size 2^m with at least one positive serum and ends with pools of size 2^{m-1} with at least one positive serum.

Efficiency Analysis

We now analyze the efficiency of the bisection procedure. As stated in the introduction, we shall consider exclusively the number of tests required to identify all positive individuals. In practice, there will be other factors to consider as well, but our purpose is to study this particular factor thoroughly so that its role is clearly understood. Since we do not assume a fixed universe, the proper indicator should be the expected number of tests *per pool*, $\mathcal{N}_m$, for an initial pool size of 2^m. Before going into the bisection procedure, let us first discuss the procedure in which an initial pooling is followed by individual testing for all positive pools. Denoting the corresponding indicator as $\bar{\mathcal{N}}_m$, we have the simple equation

$$\bar{\mathcal{N}}_m = 1 + [1 - F(P; 2^m, 0)]2^m \qquad [5]$$

Since one test is required to test the pool, then there is a probability of $1 - F(P; 2^m, 0)$ that the pool is positive and requires 2^m more tests. (Of course, the pool size need not be 2^m for this procedure. We write the equation in this form for comparison with the bisection procedure later.) The expected test *per serum* is then $\bar{\mathcal{N}}_m/2^m = 1 - F(P; 2^m, 0) + 2^{-m}$ Since this equation only involves $F(P; A, 0)$, a quick estimate can be obtained just by looking at the solid curves in Fig. 1. Let us demonstrate it using the example of Ref. (5): There $A = 10$ and $P = 2.44\%$, hence the scaled prevalence is $(10/16)2.44\% \approx 0.015$. The solid curve gives around 0.78, hence $\bar{\mathcal{N}}/A \approx 1 - 0.78 + 0.1 = 0.32$. There are 8000 samples, so one expects to make $8000 \times 0.32 = 2560$ tests, in good agreement with the actual number without retests, 2540.

Table 1. The expected number of tests in the bisection procedure for a pool of size 2^m with $1, 2, 3, 4$ positive sera, respectively.

m =	1	2	3	4	5	6	7	8	9	10
$A = 2^m$	2	4	8	16	32	64	128	256	512	1024
$M_1(m)$	1.5	3	4.5	6	7.5	9	10.5	12	13.5	15
$M_2(m)$	2	4.5	7.1	9.9	12.7	15.6	18.6	21.6	24.5	27.5
$M_3(m)$		5.5	9.1	13.0	17.1	21.4	25.7	30.1	34.6	39.0
$M_4(m)$		6	10.7	15.5	20.7	26.1	31.6	37.2	43.0	48.7

Now we analyze the bisection procedure. There is one test for the initial pool, as in Eq. [5], but the number of subsequent tests will be different for pools with different number of positive sera. Let $M_n(m)$ be the expected number of subsequent tests needed to screen a pool which contains n positive sera. Then we have

$$\mathcal{N}_m = 1 + F(P; 2^m, 1)M_1(m) + F(P; 2^m, 2)M_2(m) + F(P; 2^m, 3)M_3(m) + \cdots \quad [6]$$

$M_1(m)$ is simple: such a pool requires $\frac{3}{2}(= \frac{1}{2} \times 1 + \frac{1}{2} \times 2)$ tests at each stage, on the average, and the number of positive pools is unchanged after each bisection. Since there are m stages, we have

$$M_1(m) = \frac{3}{2}m \quad [7]$$

The efficiency of the procedure can be seen in this equation, which trades an exponential function 2^m with a linear function. Since we are particularly concerned with a small prevalence situation, in which most positive pools contain only one positive serum (see Fig. 1), the bulk of the repooled tests will be spent here. The other $M_n(m)$ can be calculated by repeated use of the Fisher formula [2]. The derivation for $n = 2, 3, 4$ is given in the Appendix and the results are shown in Table 1 for m up to 10. The favorable comparison to individual testing (the first row) is quite evident. As expected, pools with more positive sera require more tests to screen. but the Table suggests that one still realizes significant savings by using a larger pool.

It is a common practice to retest a positive result. We do not recommand retesting in the bisection procedure, because the chance that a positive serum has not been tested

Table 2. Same as Table 1, but with modified *stage 1*.

m =	1	2	3	4	5	6	7	8	9	10
$M_1(m)$	2	3.5	5	6.5	8	9.5	11	12.5	14	15.5
$M_2(m)$	2	5.2	8	10.8	13.7	16.6	19.6	22.5	25.5	28.5
$M_3(m)$		6	10.2	14.3	18.6	22.8	27.2	31.6	36.1	40.5
$M_4(m)$		6	11.8	17.1	22.5	27.9	33.5	39.2	44.9	50.6

twice by the end of the procedure is small. However, to add more assurance, we can modify the final stage to ordinary individual testing, i.e., testing both sera regardless of the result of the first serum. If this is the case, Eq. [7] is then changed to $M_1(m) = \frac{3}{2}(m-1)+2$. The other $M_n(m)$ are also changed. But as the results in Table 2 show, the added cost is not significant. (See the Appendix for detail.)

In principal, to calculate $\mathcal{N}_m$ requires knowing all the $M_n(m)$, which can be computed similarly as is done in the Appendix. However, since in the regime of interest $F(P;2^m,n)$, $n > 4$ are typically small (Fig. 1), we use instead two simple bounds for an estimate. The function $3m/2$ for $n = 1$ is a lower bound for all $M_n(m)$, since the more positive sera in a pool the more tests are required. A simple upper bound is $3mn/2$, because a pool with n positive sera can result in no more than n positive pools at any stage. Thus with the actual $M_n(m)$ for $n \leq 4$ and the bounds for $n \geq 5$, we compute $\mathcal{N}_m$ from Eq. [6] and compare it to $\bar{\mathcal{N}}_m$. Fig. 2 gives the results of $\mathcal{N}_m/\bar{\mathcal{N}}_m$ vs. the scaled prevalence, $(A/16)P$, for pool sizes 16, 32, 64, 128, 256, 512 and 1024. Thus the figure shows the advantage of the bisection procedure over individual testing of positive pools. Since the "efficiency" in Fig. 2 is defined as the ratio of test numbers *per serum*, the smaller the ratio the more efficient the bisection procedure. It should be remembered that the ranges of prevalence in the figure are different for different pool sizes. To view it in the actual scale, one should imagine squeezing each curve to the left to half of the width of the curve above it.

Conclusions and Discussions

Apparently the smaller the prevalence the more efficient a larger pool size is. In practice, one has to take into account other cost factors and weigh it against the test-

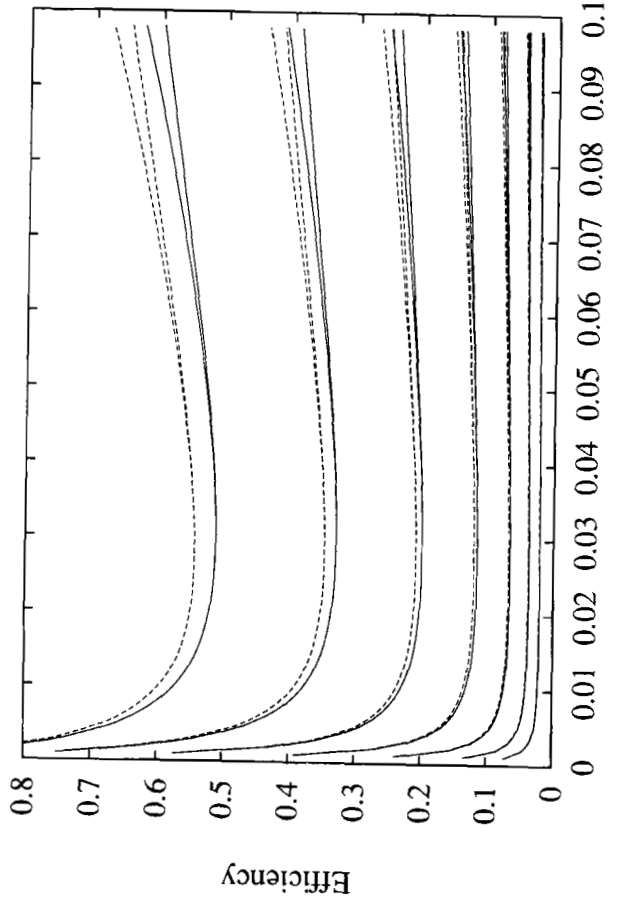

Fig. 2. Efficiency of the bisection procedure (as compared to individual testing of positive pools), $\mathcal{N}_m/\bar{N}_m$, as a function of the scaled prevalence, $(A/16)P$. The group of curves from top to bottom are for initial pool size 2^m, $m = 4, 5, \ldots 10$, respectively. The solid curves are for bisection without modified *stage 1* (Table 1) and the dashed curves are with modification (Table 2). The two curves on the right part which appear to split from a single curve on the left part are the lower and upper bounds.

number factor in deciding an appropriate pool size. We have shown the results for the first 10% scaled prevalence because the figures here are less cluttered. Similar computation can of course be extended to a higher range. In particular, the flatness of the bottom few curves in Fig. 2 suggest that the large savings may extend to a much higher prevalence than is covered in Fig. 2.

For high prevalence, the bisection procedure may be modified so that a large pool size can still be efficient. This may be particularly useful if a technique like PCR (9) becomes more accessible. One may consider using PCR for the first few stages when the pool sizes are large and switch to less sensitive, but less expensive, testing methods at later stages. A simple modification suggests itself from the numbers in Table 1. Notice that $M_n(m)$ is typically smaller than the pool size, except for high n and low m. This means that there is waste at the final few stages on pools with many positive sera. Therefore if the prevalence is high, and hence there are many pools with many positive sera, it would be better to terminate the bisection at an earlier stage. For example, if individual testing is performed at pool size 4, then the entries in the second column would be uniformly 4. This should result in the numbers for $M_n(m)$ with large n to improve.

The bisection procedure has a further practical advantage. Once the initial pooling size has been chosen, the procedure is completely straightforward; it can easily be made into a standard routine for laboratory technicians to follow. This is important, since our motivation is not so much to give the mathematically most efficient procedure, but rather to provide a sufficiently efficient yet practical procedure to be used in real life, in order that a terrible disease may be prevented from taking as much toll on human lives as it has and may.

Appendix

For pools containing two positive sera, there are two possible outcomes after one bisection: either both pools contain one positive serum, with probability $p_{1+1}(m)$, or one pool contains both of the sera, with probability $p_{0+2}(m) = 1 - p_{1+1}(m)$. In the latter case, if the first test is negative, then there is no need to test the second pool; this occurs with

a probability $p_{0+2}(m)/2$. Thus the expected number of tests for *stage* m is $\frac{1}{2}p_{0+2}(m)+2[p_{1+1}(m)+\frac{1}{2}p_{0+2}(m)]$. The resulting pools partly contain one positive serum, which then require $M_1(m-1)$ further tests, and partly contain two positive sera, which requires $M_2(m-1)$ further tests. Putting these together, we have a recursion formula,

$$M_2(m)=\frac{1}{2}[3+p_{1+1}(m)]+2p_{1+1}(m)M_1(m-1)+[1-p_{1+1}(m)]M_2(m-1) \qquad [A.1]$$

From Eq. [2], we have $p_{1+1}(m)=2^{m-1}/(2^m-1)$ Then $M_2(m)$ can be calculated recursively using $M_2(1)=2$ and $M_1(m)=3m/2$. Note that one must use Eq. [2] and not Eq. [4] in analyzing the bisection procedure, since the parts are half of the universe.

For pools with three positive sera, there are again two possible outcomes after a bisection: either one pool contains one and the other contains two positive sera or one pool contains all three positive sera. The probability of the former case is denoted as $p_{1+2}(m)$ and that of the latter case is $p_{0+3}(m)=1-p_{1+2}(m)$. We have the recursion formula

$$M_3(m)=\frac{1}{2}[3+p_{1+2}(m)]+p_{1+2}(m)[M_1(m-1)+M_2(m-1)]+[1-p_{1+2}(m)]M_3(m-1) \quad [A.2]$$

Then $M_3(m)$ is computed using $M_3(2)=5\frac{1}{2}$ and $p_{1+2}(m)=(3/4)2^m/(2^m-1)$, which comes from Eq. [2].

For $M_4(m)$ we perform the same analysis, but now the possible outcomes are represented by three probabilities, $p_{1+3}(m)$, $p_{0+4}(m)$ and $p_{2+2}(m)$. The probability of testing negative on the first pool is $p_{0+4}(m)/2$; one test is required for this case and two tests otherwise for *stage* m. The subsequent test numbers are then described by appropriate *stage* $m-1$ functions. A recursion equation similar to $[A.1]$ and $[A.2]$ can be formulated and solved, with the additional condition $M_4(2)=6$. Table 1 gives the results from these computation.

If the procedure at *stage 1* is modified to testing every serum, without skipping the companion of a negative serum, the recursion relations remain valid. The change is in the

initial conditions. Thus one simply repeats the calculation using the following instead: $M_1(m) = 2 + \frac{3}{2}(m-1)$, $M_2(1) = 2$ and $M_3(2) = M_4(2) = 4$. The results are tabulated in Table 2.

Acknowledgements

This work is supported in part by the Fogarty International Center, National Institute of Health, under Grant No. D43TW00004. We wish to thank Dr. Tun-Hou Lee for stimulating discussions.

*Corresponding author: Dr. Chung-Cheng Hsieh, Department of Epidemiology, Harvard School of Public Health, 677 Huntington Avenue.

Telephone: (617)432-4564, Fax: (617)566-7805.

References

1. Brown, R. C. 1990. Seroprevalence and clinical manifestations of HIV-1 infection in Kananga, Zaire. AIDS 4:1267–1269.
2. Killewo, J., K. Nyamuryekunge, A. Sandstriom, U. Bredberg-Riadien, S. Wall, F. Mhalu, G. Biberfeld. 1990. Prevalence of HIV-1 infection in the Kagera region of Tanzania: a population-based study. AIDS 4:1081–1085.
3. Rudin, C., R. Berger, R. Tobler, P. W. Nars, M. Just, N. Pavic. 1990. HIV-1, hepatitis (A, B and C), and measles in Romanian children. Lancet **336**:1592–1593.
4. Behets, F., S. Bertozzi, M. Kasali, M. Kashamuka, L. Atikala, C. Brown, R. W. Ryder, and T. C. Quinn. 1990. Successful use of pooled sera to determine HIV-1 seroprevalence in Zaire with development of cost-efficiency models. AIDS 4:737–741.
5. Kline, R. L., T. A. Brothers, R. Brookmeyer, S. Zeger, T. C. Quinn. 1989. Evaluation of Human Immunodeficiency Virus seroprevalence in population surveys using pooled sera. J. Clin. Microbiol. **27**:1449–1452.
6. Cahoon-Young, B., A. Chandler, T. Livermore, J. Gaudino, R. Benjamin. 1989. Sensitivity and specificity of pooled versus individual sera in a human immunodeficiency virus antibody prevalence study. J. Clin. Microbiol. **27**:1893–1895.

7. Mariotti, M., J-J. Lefrére, B. Noel, F. Ferrer-Le-Coeur, D. Vittecoq, R. Girot, C. Bosser, A-M. Couroucé, C. Salmon, P. Rouger. 1990. DNA amplification of HIV-1 in seropositive individuals and in seronegative at-risk individuals. AIDS 4:633–637.

8. Fisher, R. A. 1921. On the mathematical foundation of theoretical statistics. Phil. Trans. Roy. Soc. London **A 222**:309–368.

9. Pershing D. H. 1991. Polymerase Chain Reaction: Trenches to Benches. J. Clin. Microbiol. **29**:1281–1285.