Arch Public Health
1999, 57, 89-105

Adjusting for confounding when estimating a time trend in HIV prevalence based on pooled serum samples

by

Vansteelandt S.*, Goetghebeur E., Verstraeten T.

Abstract

Over the last decade, many epidemiological studies have demonstrated the successful use of pooled sera for screening purposes or HIV risk estimation (1-11). The method was originally designed as a cost-reductive tool, but also appears to lower the error rates associated with diagnosis and in low prevalence areas it produces, at most, a slight loss in the estimation accuracy. In this paper, we use test results on pooled sera to estimate a time evolution in HIV prevalence where we need to account for the presence of important confounders. An adjusted time trend estimate is proposed, assuming confounding variables are measured for each subject, but there is only one diagnostic test result per pool.

Experimental design is important if one is to achieve a precise and cost-efficient estimate of an evolution of risks over time. Specifically, the distribution of known prognostic factors over the pools is influential and

* Research Assistant of the Fund for Scientific Research – Flanders (Belgium) (F.W.O.).
Department of Applied Mathematics and Computer Science, University of Ghent, Krijgslaan 281 – S9, B-9000 Gent, Belgium. E-mail: Stijn.Vansteelandt@rug.ac.be.

can be influenced. Choosing pools to be covariate-homogeneous increases the amount of information at virtually no additional cost. The cost-precision balance is further optimized by calculating pool sizes in function of the distribution of covariates over the pools.

Our study was motivated by the planning of a growing database to monitor a time trend in HIV prevalence (14). The methods are used to adjust for age as a confounder in data obtained from Kenyan pregnant women. Analysis of the Kenyan data shows that age homogeneous pools of optimal size reduce cost to 44% of the original price, whilst precision remains close to the one obtained from non-pooled samples.

Key-words

Epidemiologic research design, HIV seroprevalence, sentinel surveillance, seroepidemiological methods, standardization, statistical models.

1. Introduction

Controlling the HIV epidemic is still one of the major challenges to public health, particularly in Africa. To monitor progress, the estimate of a time evolution in HIV prevalence is valuable. To this end, one may collect serum samples from one part of the population at regular time intervals. In developing countries, sera from pregnant women who present themselves at health centers are relatively easy to obtain. Hence, we test these and estimate HIV prevalence at each time point as the proportion of positive test results. The observed association between calendar time and estimated prevalence is then an estimate for the HIV risk evolution. However, such association need not be evidence of a trend in disease control. For instance, age is a likely confounder of the HIV risk evolution in a population of Kenyan pregnant women (see also section 3). Indeed, with a growing development women may tend to have children at an older age, whilst the probability of being HIV positive is likely to increase with age. Hence, the need to adjust and ultimately standardize for age when exploring a time trend (12-13).

Because budget constraints limit the number of available diagnostic tests, one has sought to increase the amount of information provided by a single diagnostic test. Subjecting pools of serum samples to the HIV test

is a commonly used technique in regions of low HIV prevalence (1-11). The observations then consist of one indicator per pool: "at least one HIV positive sample in this pool" or not. Estimation may then proceed by first estimating the probability of a positive test result for the pool as the proportion of positive test results and next backtransforming it to a point estimate of the mean HIV prevalence. As less tests are required for the same total number of samples, major cost reductions can be achieved. More surprisingly, in low prevalence settings precision remains high for the same total sample size – and is enhanced in the presence of substantial measurement error on the diagnostic test – as long as the *pool size*, i.e. the number of serum samples per pool, does not grow too large (11).

To adjust the HIV risk evolution for age, we calculate age-standardized risks. They are typically obtained by averaging age-specific risks over the age distribution in a "frozen" standard population (12). The age-specific risks can be estimated from *generalized linear models* (15-17), which assume that a monotonic transformation of the risk relates linearly to age. However, the new context of pooling implies that outcome values are only observed for pools of subjects. This prevents straightforward fitting of generalized linear models and calls for a feasible extension.

In this paper, we introduce methods for estimating age-specific risks on the basis of pooled testing data. We look at optimizing experimental design through careful choices for pool composition and pool size. The obtainable improvement in cost and precision is illustrated using data from a Kenyan HIV prevalence study in pregnant women.

2. Methods

2.1 Age-specific HIV prevalence estimates

Suppose we have N serum samples on pregnant women and enough diagnostic tests with sensitivity Se and specificity Sp. The definitions of Se and Sp are slightly different for tests performed on pools: Se (Sp) is now the probability of correctly classifying a *truly HIV positive (negative) pool*, i.e. a pool in which at least one of the samples is truly HIV positive (all samples are truly HIV negative). Our development assumes that both validity measures are pool size independent. The number of samples diluting an HIV positive sample is thus anticipated not to affect the probability of detecting the truly positive status of that sample. Numerous studies provide empirical evidence for this assumption to hold

up to pools of size 16 (3, 5, 8, 9). This is not surprising since single serum samples are diluted anyway before a diagnostic test is performed. In pools, the several samples serve as diluents for each other (11).

To estimate HIV prevalence $\pi(x)$ in age group x, we assume that prevalence relates linearly to age on the scale of a monotonic function $g(.)$

$$\begin{aligned} & g(\pi(x)) = \alpha + \beta x \\ & \Rightarrow \pi(x) = g^{-1}(\alpha + \beta x); \end{aligned} \tag{1}$$

(1) is called a *generalized linear model* with *link function* $g(.)$. It encompasses the *logistic model* $\pi(x) = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$, the *complementary log-log model* $\pi(x) = 1 - e^{-e^{\alpha+\beta x}}$ and the *probit model* $\pi(x) = \Phi^{-1}(\alpha + \beta x)$ in which $\Phi^{-1}(.)$ denotes the inverse cumulative standard normal distribution function. Routines for generalized linear model fitting are available in most standard statistical software packages.

First, we construct pools combining c_i samples of the same age x_i, which are independent, random draws conditional on x_i. The observations consist of n independent test results Y_i for n pools and, hence, carry direct information about the probability $f_i(c_i)$ of a positive test result for the pool. More specifically, $f_i(c_i)$ is the probability that the pool is truly HIV positive and correctly classified, or truly HIV negative and misclassified:

$$\begin{aligned} f_i(c_i) &= \left[1 - (1 - \pi(x_i))^{c_i}\right] Se + (1 - \pi(x_i))^{c_i} (1 - Sp) \\ &= Se + (1 - Se - Sp)(1 - \pi(x_i))^{c_i}. \end{aligned} \tag{2}$$

It follows that α and β in (1) can only be estimated from the data via the pool risk $f_i(c_i)$, i.e. by combining (1) and (2) into

$$f_i(c_i) = Se + (1 - Se - Sp)(1 - g^{-1}(\alpha + \beta x_i))^{c_i}.$$

For the complementary log-log model, this equation translates into a new generalized linear model

$$\log\left(-\log\left(\frac{f_i(c_i) - Se}{1 - Se - Sp}\right)\right) = \alpha + \beta x_i + \log c_i.$$

The advantage is that little computational effort may suffice to estimate α and β, provided $\pi(x) = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$ is the "true" model for individual risk. By contrast, a probit or logistic model for individual risk does not yield the corresponding model for the test result for a pool of common age. Here, one has to invoke non-linear optimizing routines for maximizing the loglikelihood l, which measures the agreement between observed values Y_i and expected outcomes $f_i(c_i)$:

$$l = \sum_{i=1}^{n} Y_i \log f_i(c_i) + (1-Y_i)\log\left(1-f_i(c_i)\right) \quad . \tag{3}$$

Secondly, when arbitrary age components are allowed in each pool, the risk $f_i(c_i)$ has to account for the observed sources of heterogeneity in the pool. In particular,

$$f_i(c_i) = Se + (1-Se-Sp)\prod_{j=1}^{c_i}\left(1-g^{-1}\left(\alpha+\beta x_{ij}\right)\right), \tag{4}$$

with x_{ij} the age of the j-th sample in the i-th pool. The resulting model is no longer of the generalized linear form. Hence, direct maximization of the loglikelihood (3) is required; even for the complementary log-log model.

For computational ease (see discussion), further results will be given for the complementary log-log model.

2.2 Precision

It follows from appendix A that the asymptotic variance on the prevalence estimates

$$Var\left[\hat{\pi}(x)\right] \approx \frac{e^{-2e^{\alpha+\beta x}} e^{2(\alpha+\beta x)}}{\sum_{i=1}^{n} w_i(c_i)}\left[1+\frac{1}{S}(\tilde{x}-x)^2\right] \tag{5}$$

is a quadratic function of age x with $\tilde{x}$ and S expressing measures of location and spread for "pool-specific" ages; they are respectively called the informative age and pool-age variance. Specifically, when the pool-age is defined as a well-specified weighted average of age in the considered pool (appendix A), $\tilde{x}$ and S are the weighted sample mean and variance of the pool-ages with weights $w_i(c_i)$. The variance achieves its minimum $\frac{e^{-2e^{\alpha+\beta\tilde{x}}} e^{2(\alpha+\beta\tilde{x})}}{\sum_{i=1}^{n} w_i(c_i)}$ at $x = \tilde{x}$ and mainly depends on the study design through the sum of weights $\sum_{i=1}^{n} w_i(c_i)$. S relates to the rate of change of the variance with age x.

2.3 Cost

The expected cost of a prevalence study

$$C = NC_s + \sum_{i=1}^{n} C_{p,i},$$

can be summarized as a combination of the constant cost C_s for collecting each sample and the pool-specific cost $C_{p,i}$ for constructing and testing each pool. A detailed expression for $C_{p,i}$ in the Kenyan study is given in appendix B.

2.4 Optimal pool composition

A random pool composition, conditional on age, is required to yield consistent prevalence estimates. The age distribution in each pool can be chosen by the experimenter and influences the amount of information carried by each pool. Should we make the effort to distribute ages over pools in a prescribed way, given the fact that generalized linear models can be used to adjust for age?

The pool-age variance S measures age variability between pools and, hence, increases with greater age homogeneity within the pools. The rate of change of $\mathrm{Var}\,\hat{\pi}(\tilde{x})$ with age is thus minimized for *age homogeneous pools*; the central variance, $\mathrm{Var}\,\hat{\pi}(\tilde{x})$, and cost depend only weakly on the age distribution of the pools.

2.5 Optimal pool size

The pool size influences both cost and precision. General pool size calculations will therefore try to achieve a determined cost-precision balance. In this section, an optimal pool size $c_{opt}(x)$ is obtained by evaluating how the pool size ideally depends on the age distribution over the pool and by choosing an optimal average pool size. We give specific results for pool size calculation in section 3; technical details and related sample size calculations can be found in (20).

When both the total number of samples N and pools n are kept fixed, central variance $\mathrm{Var}\,\hat{\pi}(\tilde{x})$ grows drastically (moderately) with increasing size of a high-risk (low-risk) pool. The constraints thus imply that enlarging low-risk pools and shrinking high-risk pools improves precision. Minimum variance is then achieved for pools of common risk $f_i(c_i)$. For perfectly homogeneous pools, a common probability f of a positive test result for the pool, i.e.

$$f_i(c_i) = Se + (1 - Se - Sp)\left(1 - \pi(x_i)\right) c_{opt}(x_i) = f, \qquad (6)$$

implies large pools of low-risk samples and small pools of high-risk samples. More precisely, it follows from (6) that the optimal pool size decreases with individual risk via

$$c_{opt}(x) = \frac{\kappa}{\log(1-\pi(x))} = \frac{\kappa}{\log(1-g(\alpha+\beta x))}, \qquad (7)$$

where we will identify $\kappa = \log\left(\frac{Se-f}{Se+Sp-1}\right)$ rather than f. For pools combining different ages, the average $c_{opt}(x)$-value in the pool is a good approximation to the "true" optimal pool size (20). Notice that in practice, one has to round the continuous values for $c_{opt}(x)$ so as to find integer optimal pool sizes greater than 0.

By construction, pools of optimal size $c_{opt}(x)$ minimize the central variance $\mathrm{Var}\,\hat{\pi}(\tilde{x})$ once N and n are given. As seen in (20), they also further attenuate the change of the variance with x!

A choice for κ in (7) follows from a chosen cost-precision tradeoff once expected cost and central variance are evaluated over a grid of κ-values. This is straightforward for the complementary log-log model: assuming a value for α and β based on prior information, central variance $\mathrm{Var}\,\hat{\pi}(\tilde{x})$ and expected cost C are calculated by substituting $f_i(c_i)$, $w_i(c_i)$ and n in the expressions (5) and (9) for $\mathrm{Var}\,\hat{\pi}(\tilde{x})$ and C by

$$Se + (1-Se-Sp)e^{\kappa}, \frac{(1-Sp-Se)^2\kappa^2 e^{2\kappa}}{f(1-f)} \quad \text{and} \quad n = \frac{1}{\kappa}\sum_{i=1}^{N}\log(1-\pi(x_i))$$

respectively (20) (see also section 3 and figure 3).

When precision is ignored, $\kappa = -\infty$ and minimum cost is achieved from one pool containing all serum samples. When cost is ignored and measurement error recognized, a single diagnostic test result for one pool of c sera may contain more information than the c corresponding individual test results (11). Surprisingly, the optimal pool size may thus be greater than 1 for some ages and κ can still reasonably be calculated from an equation (see (20)) involving *Se* and *Sp*.

2.6 Adjusting the estimated HIV prevalence evolution over time

Standardization (12-13) offers one way to adjust time-dependent prevalence estimation for age confounding. Age-standardized risks $\hat{\pi}_{st}$ are estimated by averaging age-specific risks over the age distribution in a standard population that is "frozen" over time; that is

$$\hat{\pi}_{st} = \hat{E}_x\left[\hat{\pi}(x)\right] = \hat{E}_x\left[g^{-1}\left(\hat{\alpha}+\hat{\beta}x\right)\right].$$

An approximate variance expression is given in (20). When age-standardized risks are estimated over regular time intervals, the observed time trend provides a corrected estimate for the time evolution in HIV risk. The original optimal design stays good for estimating a time evolution. However, because some extra averaging is involved, results are less sensitive now to age variations when confounding is weak. A practical illustration is given in the next section.

Alternatively, one can parameterize the time effect along with age, for instance through the extended generalized linear model

$$g(\pi(x,t)) = \alpha + \beta x + \gamma t\,, \tag{8}$$

where t represents time of collection. Details are given in (20). Using maximum likelihood estimation for model (8), the original optimal design stays good for estimating a precise and cost-efficient time effect, provided the pools are time homogeneous as well. As short-period effects of time on HIV risk are small, pools can be constructed homogeneously w.r.t. time of collection at virtually no expense of age homogeneity.

3. Results

3.1 The Kenyan study

In 1996, as part of the National AIDS Control Programme, 787 serum samples have been gathered from pregnant women who present themselves at health centers in Kenya. Base-line characteristics (age, region, parity, marrital status and education), as well as diagnostic test results were recorded, using the following test procedure. All serum samples were first subject to an Elisa test (Innotest HIV 1/HIV 2, Innogenetics, Belgium). Samples with a negative test result were HIV negatively declared; others were subject to a rapid assay test (Capillus HIV 1/HIV 2, Cambridge Biotech, Ireland). If this test did not contradict the former, the sample was declared HIV positive; other serum samples were subject to a second Elisa test (Viranostika, Organon, Holland), whose test result was then conclusive. For the Kenyan data, *Se* and *Sp* take the values 1 and 0.9997 respectively (18).

On the current data set an exploratory exercise was conducted to investigate whether the pooling of samples is worth considering in future studies, when ultimately the goal is to monitor any trend over time. For expository purposes and since age is expected to be the main confounder of a time trend, we adjust for this single covariate.

Descriptive statistics are given in table 1 and figure 1. Except for region, all measured variables suffer from some non-response. In particular, 36 test results, 27 ages, 55 parities, 31 marrital status and 30 educations are missing. We addressed missingness through the use of imputation methods (19), conditioning on all knowledge that was available in the base-line characteristics. Since the focus here is on the comparative value of several pooling strategies, we have restricted the procedure to just one imputation and treated the imputed data as if they were observed.

TABLE 1
Some descriptive statistics for the Kenyan study

Covariate	Type	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Missing
Age	Continuous	10	20	24	24.35	28	46	27
Parity	Discrete	0	0	1	1.38	3	10	55

Marr. Status	Single	Poly-gamous	Mono-gamous	Divorced	Widowed	Missing
Test res. –	81	140	449	5	2	22
Test res. +	9	6	34	1	2	0
Test res. missing	2	2	23	0	0	9
Number obs.	92	148	506	6	4	31

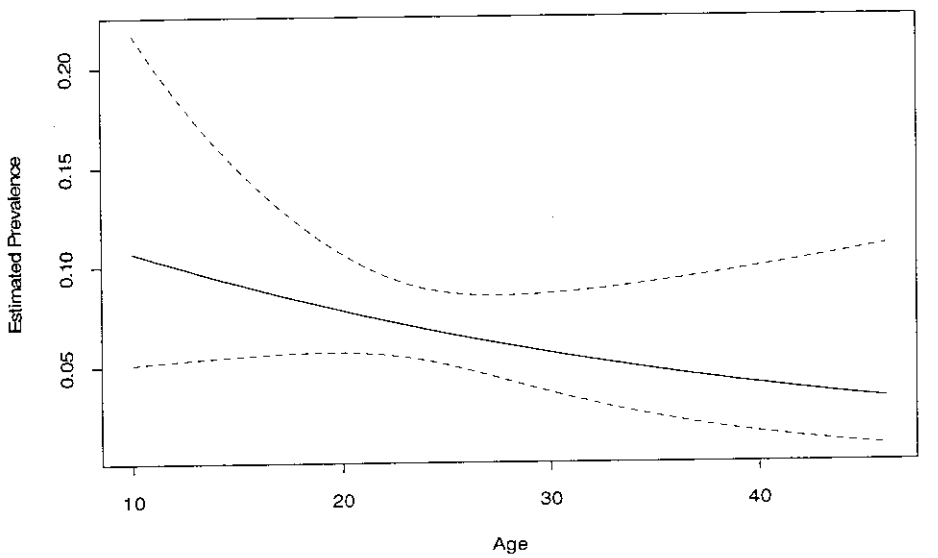

Fig. 1: Estimated age-specific HIV risk, along with 95% confidence bounds

3.2 Optimal pool composition in the Kenyan study

Consider the following recombinations of data where pools are constructed from the original samples in different compositions:

Chronological: 113 equally-sized ($c = 7$) pools constructed by time of collection. This reflects a practically convenient composition which comes close to a random composition.

Age ordered covariates: 113 equally-sized ($c = 7$), age homogeneous pools constructed by order of age.

Age common covariates: 113 variably-sized, age homogeneous pools combining 1 to 10 serum samples (mean and median size are 6.97 and 8 respectively) of the same age in years.

Diagnostic tests were not performed on the pools. For this comparative illustration, we defined the test results for a pool to be positive as soon as one test result among the contributing samples was positive. Our choice for 113 pools is explained in section 3.3. Finally, for reference purposes, results are also shown for the analysis of the non-pooled-samples and labeled *Individual*.

TABLE 2
Comparison of pool designs; standard errors are given between brackets; 1: Individual, 2: Chronological, 3: Age Ordered, 4: Age Common, 5: Age Ordered and Optimal-sized

	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\tilde{x}}$	$\hat{\pi}(\hat{\tilde{x}})$	$\hat{\pi}(46)$	$\hat{C}$	$\hat{S}$
1	−1.85 (0.66)	−0.034 (0.027)	23.48	0.069 (0.0091)	0.033 (0.021)	$2598	24.97
2	−2.89 (1.67)	0.011 (0.67)	24.65	0.070 (0.012)	0.088 (0.12)	$1150	4.35
3	−1.53 (0.71)	−0.050 (0.030)	23.36	0.066 (0.010)	0.022 (0.015)	$1139	23.71
4	−1.33 (0.77)	−0.057 (0.032)	23.28	0.067 (0.014)	0.019 (0.014)	$1121	24.34
5	−1.72 (0.67)	−0.037 (0.028)	23.27	0.073 (0.0097)	0.032 (0.014)	$1155	24.36

Table 2 and figure 2 confirm our previous statement that age-homogeneous pools carry the largest amount of information. In particular, the variance increases 6 times faster with age for chronological than for age ordered pools. The impact is most strikingly seen when considering ages towards the tail of the observed age distribution: for a 46-year old woman, a prevalence estimate of 0.088 (0.12) is found, whereas for age ordered samples the corresponding estimate 0.019 (0.014) is more precise. The striking difference between the risk estimates is suggestive of a possible finite sample size bias in the chronological case, and was

confirmed by a preliminary simulation study. We conclude that the extra effort for constructing age homogeneous pools is more than offset by the precision of the estimates and a fortiori the cost. As they require less effort than age common pools, age ordered pools enjoy our preference.

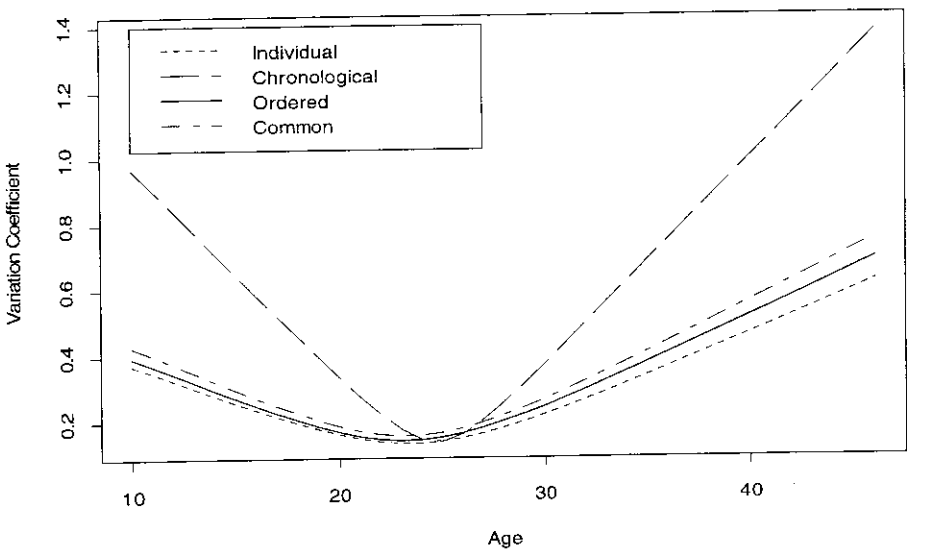

Fig. 2: Comparison of the variation coefficient $\left(\frac{\hat{\sigma}}{\hat{\pi}}\right)$ for four pool compositions.

3.3 Optimal pool size in the Kenyan study

We try to further increase the precision which we obtained in section 3.2 for age ordered pools, by optimizing the pool size. Given $(\alpha, \beta) = (-1.85, -0.034)$ and the same cost level $C = \$1139$ as for age ordered, equally-sized pools, figure 3 shows that maximum precision $\left(se\left[\hat{\pi}(\tilde{x})\right] \approx 0.00985\right)$ is achieved when κ equals –0.493; when cost is ignored, $c_{opt}(x)$ remains 1 within the observed age range as $\kappa = -0.022$.

Using the former result, we construct 116 age ordered pools of optimal sizes varying between 5 and 11 over the entire observed age range. The precision increase in figure 3 is twofold: a global precision increase, i.e. the variation coefficient of $\hat{\pi}(\tilde{x})$ decreases from 0.152 to 0.133, adds to a further attenuation of the rate of change of the variance with age, i.e. $\hat{S}$ increases from 23.71 to 24.36 (see table 2). The expected cost $\left(\hat{C} = \$1155\right)$ hardly differs from the expected cost for age ordered pools of size 7 $\left(\hat{C} = \$1146\right)$.

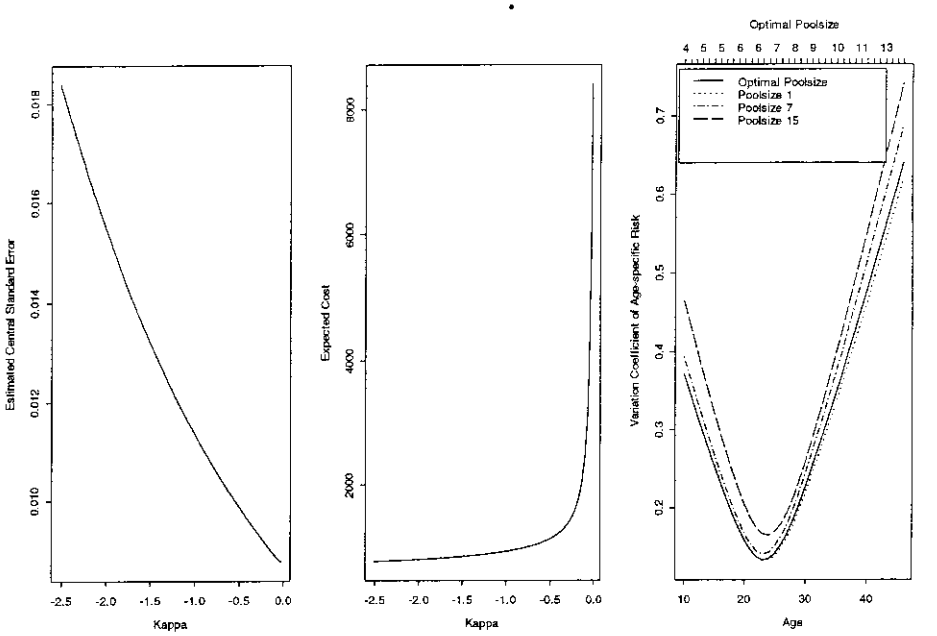

Fig. 3: Left and middle: estimated central standard error and expected cost in function of κ; right: variation coefficient $\left(\frac{\hat{\sigma}}{\hat{\pi}}\right)$ in function of age for age ordered pools of different sizes and with the top axis showing the optimal pool sizes.

3.4 Age-standardized risks in the Kenyan study

Currently, data are available for one year so that the estimate of a time evolution in HIV prevalence is not feasible yet. However, to evaluate how experimental design affects the precision of an estimated time trend, we generate an artificial standard population which is thought representative for a population of pregnant women in industrialized countries.

Columns 2 and 3 of table 3 show standardized risk estimates and variation coefficients when the standard and observed population equal each other. By contrast, the last two columns of table 3 show the results for the reference population. Here, the variation coefficient decreases by 26% from chronological to optimal pools, implying that an optimal design allows one to collect approximately 45%, i.e. $100(1-(1-0.26)^2)\%$, less serum samples and still reach the same precision as in the chronological case. This is not seen from the first two columns of table 3 where the standard and observed population equal each other, and, hence, the variance on the standardized prevalence estimates is mainly influenced by the stable central variance $\mathrm{Var}\left[\hat{\pi}(\tilde{x})\right]$. However, when temporal confounding is real, the age-standardized estimates will only be similar at one point in time and the impact of design tends to grow in importance.

Finally, notice how moving from the observed to the reference population causes a systematic drop in age-standardized prevalence estimates except for chronological pools. This once more illustrates how lack of design quality can yield a distorted view on the time evolution in HIV prevalence.

TABLE 3

Age-standardized prevalence estimates and variation coefficients $\left(\frac{\hat{\sigma}}{\hat{\pi}_{st}}\right)$ *when the standard population equals or differs from the observed population.*

Composition	Standard = Observed		Standard ≠ Observed	
	$\hat{\pi}_{st}$	$\frac{\hat{\sigma}}{\hat{\pi}_{st}}$	$\hat{\pi}_{st}$	$\frac{\hat{\sigma}}{\hat{\pi}_{st}}$
Individual	0.068	0.145	0.063	0.152
Chronological	0.070	0.136	0.072	0.211
Age Ordered	0.065	0.140	0.057	0.163
Age Common	0.066	0.155	0.057	0.176
Optimal	0.072	0.133	0.065	0.156

4. Discussion

We have proposed pooling methods for estimating a time evolution in HIV risk in the presence of individually measured confounders. For the results to hold, a random pool composition conditional on age is required, but the size and age distribution of each pool can be chosen by the investigator. Study design is important for the assessment of a precise and cost-efficient time evolution estimate. Covariate homogeneous pools carry the largest amount of information; pool sizes which attach the same risk to each pool further improve the cost-precision balance. In the Kenyan study, cost reduces to 44% of the orginal price by constructing age homogeneous pools of optimal size, whilst precision stays close to the precision obtainable from non-pooled samples. Calculations were carried out in S-Plus and programs are available from the first author on request.

The models for tests on pooled samples are extended in (20) to jointly account for several confounders. The proposed design construction extends and further leads to feasible estimators. The methods offer many new possibilities of which adjusting for confounding is an important one; others are testing for the significance of effects, regression imputation for studies dealing with missing values (14),... in the context of pooled testing data.

The complementary log-log link for the individual risk model generates a level of simplicity and thus arises as a "canonical" link function in this pooling context. However, as computations are still feasible for different link functions, one may wish to consider a link function for other than purely computational reasons. Often, a symmetric link around $\pi(x) = 0.5$, as in probit or logistic models, is the more common approach in contrast to the complementary log-log model, which is asymmetric (17). Logistic models further have a meaningful interpretation for their parameters in terms of odds ratios (17). Nonetheless, logistic and complementary log-log models yield approximately the same results in low-prevalence settings, so that the optimal pool size expression (7) remains approximately valid for the logistic model.

We conclude that for relatively low prevalence, a well-designed pooling strategy for serum samples is a feasible and useful instrument to estimate prevalence conditional on individually measured covariates from imperfect diagnostic tests. We hope that the methods in this paper will find their way to the practical settings where cost-efficiency is a real concern.

Appendix A

We derive the asymptotic variance expression (5) for the complementary log-log model. General results for vectors of covariate values and arbitrary link functions are found in (20).

Standard calculations yield the following components of the Fisher information matrix $I = (I_{kl})$ for (α, β):

$$I_{kl} = \sum_{i=1}^{n} \frac{(1-Sp-Se)^2}{f_i(c_i)(1-f_i(c_i))} \left(\sum_{j=1}^{c_i} x_{ij}^{k-1} e^{-e^{\alpha+\beta x_{ij}}} e^{\alpha+\beta x_{ij}} \prod_{m \neq j} e^{-e^{\alpha+\beta x_{im}}} \right)$$

$$* \left(\sum_{j=1}^{c_i} x_{ij}^{l-1} e^{-e^{\alpha+\beta x_{ij}}} e^{\alpha+\beta x_{ij}} \prod_{m \neq j} e^{-e^{\alpha+\beta x_{im}}} \right) .$$

By introducing the positive weights

$$w_{ij}(c_i) = \frac{(1-Sp-Se) e^{-e^{\alpha+\beta x_{ij}}} e^{\alpha-\beta x_{ij}} \prod_{k \neq j} e^{-e^{\alpha+\beta x_{ik}}}}{\sqrt{f_i(c_i)(1-f_i(c_i))}} \quad \text{and} \quad w_i(c_i) = \left[\sum_{j=1}^{c_i} w_{ij}(c_i) \right]$$

and the pool-ages $\bar{x}_i^w = \frac{\sum_{j=1}^{c_i} w_{ij}(c_i) x_{ij}}{\sum_{j=1}^{c_i} w_{ij}(c_i)}$, we reduce the Fisher information matrix to

$$I = \left(\sum_{i=1}^{n} w_i(c_i)\right)\begin{pmatrix} 1 & \sum_{i=1}^{n} w_i(c_i)\bar{x}_i^w \left(\sum_{i=1}^{n} w_i(c_i)\right)^{-1} \\ \sum_{i=1}^{n} w_i(c_i)\bar{x}_i^w \left(\sum_{i=1}^{n} w_i(c_i)\right)^{-1} & \sum_{i=1}^{n} w_i(c_i)\left(\bar{x}_i^w\right)^2 \left(\sum_{i=1}^{n} w_i(c_i)\right)^{-1} \end{pmatrix}.$$

Defining informative age $\tilde{x}$ and pool-age variance S as

$$\tilde{x} = \frac{\sum_{i=1}^{n} w_i(c_i)\bar{x}_i^w}{\sum_{i=1}^{n} w_i(c_i)}$$

$$S = \frac{\sum_{i=1}^{n} w_i(c_i)\left(\bar{x}_i^w - \tilde{x}\right)^2}{\sum_{i=1}^{n} w_i(c_i)}$$

$$= \tilde{x}^2 - \tilde{x} \quad ,$$

the asymptotic covariance matrix for $(\hat{\alpha}, \hat{\beta})$ takes the following simplified form:

$$I^{-1} = \frac{1}{\sum_{i=1}^{n} w_i(c_i)}\begin{pmatrix} 1 + \tilde{x}^2 S^{-1} - \tilde{x}S^{-1} \\ -\tilde{x}S^{-1} \quad S^{-1} \end{pmatrix}.$$

The variance on the prevalence estimate is found via the delta method:

$$Var\left[\hat{\pi}(x)\right] \approx -2e^{\alpha+\beta x} e^{2(\alpha+\beta x)}\left[Var[\hat{\alpha}] + 2xCov\left[\hat{\alpha}, \hat{\beta}\right] + x^2 Var\left[\hat{\beta}\right]\right]$$

$$= \frac{e^{-2^{\alpha+\beta x}} e^{2(\alpha+\beta x)}}{\sum_{i=1}^{n} w_i(c_i)}\left[1 + \frac{1}{S}\left(\tilde{x} - x\right)^2\right] \quad .$$

Appendix B

The three-step sequential testing procedure in the Kenyan study (14) yields

$$C_{p,i} = I(c_i > 1)C_p + C_{el1} + \left(f(c_i) + f(c_i)^{+-}\right)C_{cap} + \left(f_i(c_i)^{+--} + f_i(c_i)^{+-+}\right)C_{el2}, \quad (9)$$

where $I(c_i > 1)$ takes the value 1 when the pool size c_i is larger than 1, and 0 otherwise. C_p is the cost for constructing one pool. The costs of the first and second Elisa, and the cost of the Capillus test are respectively written as C_{el1}, C_{el2} and C_{cap}. For our problem $C_s = C_p$ = \$0.8, $C_{el1} = C_{el2}$ = \$2.1 *and* C_{cap} = \$4.1 (including material and wages) (19).

Finally, $f(c_i)^{+--}\left(f(c_i)^{+-+}\right)$ is the probability that a given pool reacts positively to the first Elisa, negatively to the Cappillus, and negatively (positively) to the second Elisa. Assuming independent test results for the same pool,

$$f_i(c_i)^{+--} = Se_{el1}\left(1-Se_{cap}\right)\left(1-Se_{el2}\right)\left(1-\pi(x_i)\right)^{c_i} + \left(1-Sp_{el1}\right)Sp_{cap}Sp_{el2}\left(1-\left(1-\pi(x_i)\right)^{c_i}\right) \quad (10)$$

$$f_i(c_i)^{+-+} = Se_{el1}\left(1-Se_{cap}\right)Se_{el2}\left(1-\pi(x_i)\right)^{c_i} + \left(1-Sp_{el1}\right)Sp_{cap}\left(1-Sp_{el2}\right)\left(1-\left(1-\pi(x_i)\right)^{c_i}\right). \quad (11)$$

Acknowledgements

We are greatful to Dr. F. Nang'ole, Medical Officer of Health Kajiado district, for his role in collecting the data. We also thank anonymous referees for valuable comments that improved an earlier version of the manuscript.

References

1. BABU P G, SARASWATHI N K, VAIDYANATHAN H, JOHN T J. Reduction of the cost of testing for antibody to human immunodeficiency virus, without losing sensitivity, by pooling sera. Indian J Med Res (A) 1993; 1-3.
2. BEHETS F et al. Successful use of pooled sera to determine HIV 1 seroprevalence in Zaïre with development of cost-efficiency models. AIDS 1990; 4: 737-741.
3. CAHOON-YOUNG B, CHANDLER A, LIVERMORE T, GAUDINO J, BENJAMIN R. Sensitivity and Specificity of Pooled versus Individual Sera in a Human Immunodeficiency Virus Antibody Prevalence Study. J Clin Microbiology 1989; 27: 1893-1895.
4. LEH-HUN GWA, CHUNG-CHENG HSIEH, YUAN-CHING KO, SHOU-JEN LAN. Beyond simple pooling for HIV screening. J Immunoassay 1992; 13: 545-557.
5. KLINE R L, BROTHERS T A, BROOKMEYER R, ZEGER S, QUINN T C. Evaluation of Human Immunodeficiency Virus Seroprevalence in Population Surveys Using Pooled Sera. J Clin Microbiology 1989; 27: 1449-1452.
6. YING-CHIN KO, SHOU-JEN LAN, TAI-AN CHIANG, YEA-YIN YEN, CHUNG-CHENG HSIEH. Successful Use of Pooled Sera to Estimate HIV Antibody Seroprevalence and Eliminate All Positive Cases. Asia-Pacific J Public Health 1993; 6: 146-149.
7. LITVAK E, TU X M, PAGANO M. Screening for the presence of a disease by pooling serum samples. J Am Stat Assoc 1994; 89: 424-434.
8. PERRIËNS J H et al. Use of rapid test and an ELISA for HIV antibody screening of pooled serum samples in Lubumbashi, Zaïre. J Vir Methods 1993; 41: 213-222.

9. SHERLOCK C H, STRATHDEE S A, LE T, SUTHERLAND D, O'SHAUGHNESSY M V, SCHECHTER M T. Use of pooling and outpatient laboratory specimens in an anonymous seroprevalence survey of HIV infection in British Columbia, Canada. AIDS 1995; 9: 945-950.
10. TU X M, LITVAK E, PAGANO M. Issues in Human Immunodeficiency Virus (HIV) Screening Programs. Am J Epidemiology 1992; 136 (2): 244-255.
11. TU X M, LITVAK E, PAGANO M. Studies of AIDS and HIV surveillance; Screening tests: can we get more by doing less? Stat Med 1994; 13: 1905-1919.
12. FIENBERG S E, BISHOP Y M M, HOLLAND W, LIGHT R J. Discrete Multivariate Analysis: theory and practice. MIT Press, Cambridge, 1980: 131-136.
13. FLEISS J L. Statistical Methods for Rates and Proportions. Wiley, New York, 1981: 7, 155-172.
14. NGUTI-MATU R. Group Screening in HIV Prevalence Estimation. Dissertation Biostatistics, Limburgs Universitair Centrum, 1997.
15. DOBSON A J. An introduction to Generalized Linear Models. Chapman & Hall, London, 1990.
16. COLLETT D. Modelling Binary Data. Chapman & Hall, London, 1996.
17. MCCULLAGH P, NELDER J A. Generalized Linear Models, second edition. Chapman & Hall, London, 1997.
18. LITTLE R J A, RUBIN D B. Statistical Analysis with missing data. Wiley, 1987: 255-259.
19. W.H.O. Global Programme on AIDS – Operational Characteristics of Commercially Available Assays to Detect Antibodies to HIV 1 and/or HIV 2 in Human Sera. Report 8, Geneva 1994.
20. VANSTEELANDT S, GOETGHEBEUR E, VERSTRAETEN T. Regression Models for Disease Prevalence from Diagnostic Tests on Pooled Serum Samples. Technical Report, 1999.