A review of statistical methods for estimating the risk of vertical human immunodeficiency virus transmission

David T Dunn

Background
Estimation of the risk of vertical transmission of human immunodeficiency virus (HIV) has been complicated by the lack of a reliable diagnostic test for paediatric HIV infection.

Methods
A literature search was conducted to identify all statistical methods that have been used to estimate HIV vertical transmission risk. Although the focus of this article is the analysis of birth cohort studies, ad hoc studies are also reviewed.

Conclusions
The standard method for estimating HIV vertical transmission risk is biased and inefficient. Various alternative analytical approaches have been proposed but all involve simplifying assumptions and some are difficult to implement. However, early diagnosis/exclusion of infection is now possible because of improvements in polymerase chain reaction technology and complex estimation methods should no longer be required. The best way to analyse studies conducted in breastfeeding populations is still unclear and deserves attention in view of the many intervention studies being planned or conducted in developing countries.

Keywords
Human immunodeficiency virus, statistical methods, vertical transmission, breastfeeding

Accepted 27 April 1998

Early studies of vertical (mother-to-child) transmission of human immunodeficiency virus (HIV) were often poorly designed, and tended to overestimate the risk of transmission because of selective inclusion of infected infants. It is now appreciated that vertical transmission risk can only be reliably estimated from birth cohort studies of children born to HIV-infected mothers and that multicentre recruitment is necessary to achieve adequate sample size. The first studies began in Europe and have been replicated in many areas where HIV infection is prevalent.

Diagnosis of HIV Infection
The standard diagnostic test for HIV infection is based on the detection of HIV immunoglobulin class G (IgG) antibody. However, all children born to infected mothers initially test HIV antibody positive since maternal IgG antibody crosses the placenta. The absence of infection can be deduced from any subsequent negative antibody test, but a positive diagnosis is only possible from a test performed after the age when maternal antibody is no longer detectable. In most studies a 15-months or 18-months threshold has been used, although maternal antibody has exceptionally been detected up to age 24 months.

In contrast, a positive virus test (p24 antigen, culture, polymerase chain reaction [PCR]) at any age is indicative of infection, although there have been rare reports of children with early positive virus tests who appear to have subsequently ‘cleared’ the virus. As a safeguard against laboratory error, the convention is to classify a child as infected only after two positive tests. Virus tests have not been regarded as sufficiently sensitive to allow infection to be excluded on the basis of a negative test result; sensitivity is particularly low in the first month of life.

Since death may preclude a definitive laboratory diagnosis it is important to also consider clinical criteria when diagnosing HIV infection. A child who died from HIV infection or who was diagnosed with AIDS would always be classified as infected, although less specific signs and symptoms have also been used.

Birth Cohort Studies
Standard method of estimation
In most birth cohort studies the analysis has proceeded as follows: (1) define infection status criteria based on immunological, virological, and clinical data available; (2) classify each child on the basis of these criteria as infected, uninfected, or of indeterminate infection status; (3) estimate the vertical transmission rate as the number of infected children relative to the number of infected plus uninfected children, ignoring indeterminate children.
Infection status may be unknown because of loss to follow-up, reporting delay, death which cannot be definitely ascribed to HIV infection, or simply because the child was born shortly before the date of analysis. Even if all of these processes were unrelated to true infection status, which is unlikely, this simple estimator would still generally be biased due to selective use of information in classifying infection status. For example, a child who developed AIDS would be defined as infected; but not developing AIDS, which suggests but does not prove the absence of infection, is not taken into account. Similar biases arise through ignoring negative virus test results and early positive antibody results.

A common technique in interim analyses is restricting the analysis to children who are old enough to have a definitive serological diagnosis of infection. Thus, if antibody persistence beyond 18 months was regarded as proof of infection then children born within 18 months of the date of analysis would be excluded. A related technique is to base the transmission risk estimate on children with an antibody test after the threshold age (taking into account children who died from HIV infection before this age) but this may bias estimates upwards as clinicians may discontinue testing seronegative children. When these rules were developed, analysis centred on the results of antibody tests since virus tests were not widely performed. This is no longer the case, and it is clearly inefficient to exclude from the analysis infants whose infection has been diagnosed, or excluded, by PCR or viral culture.

Alternative estimation methods

Tsai et al. described an estimator of the transmission risk under the idealized scenario where all children have only one HIV antibody test at the threshold age for antibody persistence (they assume 15 months). Their method is based on an independent competing risks model where the two events are development of clinical evidence of HIV infection and censorship (loss to follow-up or death unrelated to HIV infection). The maximum likelihood estimator of the transmission risk was shown to be

$$\hat{\pi} = 1 - \frac{n_u}{NS_c(15)}$$

where \(n_u\) is the number of children who test antibody negative, \(N\) the size of the initial cohort, and \(S_c(15)\) the survivor function for the censoring time \(C\) at age 15 months estimated by the Kaplan-Meier method. In effect, one minus the transmission risk is estimated as the proportion of children who test antibody negative at age 15 months, where the term \(S_c(15)\) is an adjustment to obtain the correct denominator.

An analysis of the European Collaborative Study exploited the fact that infected children are persistently HIV antibody positive whereas uninfected children ultimately lose antibody. The survivor function of children who remain antibody positive should therefore asymptote at a value equal to the transmission risk. The advantages of this approach over the standard method of analysis are: (1) no need to assume a threshold age for antibody persistence; (2) all antibody test results are taken into consideration, including positive tests before the threshold age. Special survival analysis techniques were used to account for the fact that age at antibody disappearance is known only to lie in the interval between the last positive test and first negative test (interval censoring). Another difficulty in the analysis is death due to HIV infection, since this censoring mechanism is not independent of the outcome event i.e. antibody disappearance. This was addressed by setting the censoring age to the age the children would have been at the date of analysis had they not died.

A related analysis was performed in the ACTG-076 trial, although this was based on the results of all virus culture assays (at ages 12, 24, and 78 weeks) rather than serological findings. Kaplan-Meier analyses, ignoring interval censoring, were performed of age to first positive virus culture in the intervention and placebo groups. In a mirror image of the previously described analysis, the risk of vertical transmission was estimated as the asymptote of one minus the survivor function. No child tested positive for the first time at the 78-week sample and the asymptote was attained at age 24 weeks. In effect, a child who tested negative at the 24-week sample was deemed to be uninfected. This is strictly valid only if virus culture sensitivity at this age is 100%, but a sensitivity of around 90% between 1 and 6 months of age has been reported. This raises concerns about the appropriateness of a survival analysis model, which implicitly assumes that once virus can be cultured in a sample from an infected child then it can be cultured in all subsequent samples. However, reassuringly similar estimates were obtained using the standard method of estimation.

Methods based on modelling the disappearance of HIV antibody or appearance of virus are inefficient as they essentially rely on a single variable. An approach that takes into account clinical, serological, and virological data by formulating the problem as a two-group mixture model (comprising infected and uninfected children) was recently described. Polymerase chain reaction and virus culture were assumed to be perfectly specific, so that infection is inferred from a single positive result, and to have fixed, unknown sensitivities after age one month. The latter assumption could be criticized as sensitivity is likely to vary systematically between infected children and with age, related to variation in viral load. Estimation was performed by an EM algorithm which iterated between (1) estimating the conditional probability of infection for each child, and (2) estimating (non-parametrically) the vertical transmission risk, the sensitivities of the virus tests, the AIDS incubation period, and the distribution of disappearance of maternal antibody. In common with all the other estimators described above, censoring mechanisms were assumed to be independent of infection status.

Indirect method for studies in developing countries

In developed countries it is usually straightforward to determine whether a childhood death was or was not due to HIV infection. However, in developing countries with high underlying infant mortality rate and limited diagnostic facilities an element of subjectivity is unavoidable. This lead to the concept of including a control group of children born to HIV-uninfected mothers in the study. If the mortality experienced by this group reflects that of infected children born to infected mothers, the excess mortality due to HIV infection can be estimated indirectly. This can then be combined with serological or virological data on the surviving children to derive an estimate of transmission.
risk. When there is only a single antibody test at a specified age the transmission risk is estimated by

$$\hat{\pi} = \frac{(1 - M_j) S + M_1 - M_0}{1 - M_0}$$  \hspace{1cm} (2)

where $S$ is the HIV prevalence among surviving children, and $M_1$ and $M_0$ are the estimated risks of mortality before the age when the antibody test is conducted among children born to HIV-infected mothers and HIV-uninfected mothers. $M_1$ and $M_0$ are estimated by standard survival analysis methods. A derivation of this formula is given in the Appendix.

This estimator is biased if maternal infection per se has an adverse effect on child mortality. A more serious concern is that this estimator does not capture HIV transmission through breastfeeding after the antibody test, with an estimated incidence of approximately 7 per 100 child-years of breast milk exposure.

**Ad Hoc Studies**

**Neonatal blood specimens**

Anonymous testing of neonatal blood samples has been used to produce estimates of the prevalence of HIV infection among child-bearing women. Extending this approach, Comeau et al. tested by PCR stored blood spots of children who developed AIDS. In all, 35 (52%) of the 67 specimens were PCR positive. Based on this finding, it was proposed that a 'real time' estimate of the vertical transmission risk could be obtained by testing all antibody-positive blood spots by PCR, calculating the proportion of these specimens which were PCR positive, and multiplying this by a factor of 2. The method was applied to 48 585 blood spot specimens from the New England Regional Newborn Screening Programme. Overall, 161 samples were Western Blot confirmed antibody positive, with 156 yielding sufficient material for PCR analysis, of which 10 were PCR positive. The transmission risk was therefore estimated as $2 \times 10/156 = 12.8\%$ (95% CI: 6.2–23.0%, ignoring uncertainty in the multiplicative factor).

This method is sensitive to the assumption that 50% of vertically-infected children test PCR positive in the neonatal period. Recent studies indicate that the proportion of infected children with detectable levels of HIV DNA at birth is closer to 30–40%. Moreover, PCR sensitivity changes rapidly in the first 2 weeks of life, and the exact age at which the sample is obtained is therefore important. However, this information would normally be lost in the anonymization of the specimens.

**Medicaid claims data**

Hauck et al. utilized Medicaid claims data to estimate vertical transmission risk. They first identified HIV-infected Medicaid-enrolled women with live-born deliveries, whose records were then linked to their children's birth hospitalization and subsequent health care claims. This analysis was justified on the grounds of a large sample size which is more 'representative' than the study populations enrolled in prospective studies. However, laboratory results were not available for analysis and the children's infection status could only be ascertained from diagnoses coded according to the International Classification of Diseases on Medicaid claims. This meant that in no case could HIV infection be explicitly excluded. To solve this problem, the authors formulated a two-group mixture model which incorporated a survival analysis of 'time to HIV diagnosis' according to pre-defined clinical and serological criteria. Kaplan-Meier analyses were difficult to interpret and the main findings were based on parametric distributions. One potentially strong bias is that only 69% of mother and child records were linked, with failure to link likely to have been more common for uninfected children.

**Discussion**

Recognizing that the standard method for estimating the risk of vertical transmission is biased and inefficient, various alternative analytical approaches have been proposed. These all involve simplifying assumptions and some are also difficult to implement. It has recently been shown, however, that an accurate early diagnosis of infection is possible by PCR, and the results on a single blood sample obtained at 6 weeks of age may be sufficient for epidemiological analysis provided quality control procedures are implemented. The routine collection of an early blood sample for PCR analysis in study protocols should obviate the need for the complex analytical methods that have been developed, although these may find application in the context of other vertically-transmitted infections.

Difficult problems remain, however, in the analysis of studies conducted in breastfed populations, including randomized controlled trials of interventions to prevent intrapartum transmission. An analysis of PCR results on early samples can demonstrate the ineffectiveness of the intervention, as in the case of antiseptic cleansing of the birth canal. Interpretation is difficult, however, where a difference is found as an effect of the intervention on intrapartum transmission could be offset by transmission through breastfeeding. This makes it important to continue testing children until cessation of breastfeeding, requiring synthesis of clinical, immunological, and virological data in the analysis. The 'indirect' method for estimating the transmission risk is not immediately relevant since intervention studies have not included a control group of children born to HIV-uninfected mothers. Quantification of the risk and timing of transmission through breastfeeding is also important for appropriate feeding counselling for HIV-infected women. It is essential to take account of duration of breastfeeding and to calculate, at a minimum, the rate of infection per unit time of breast milk exposure or risk conditional on breastfeeding to a specified age.

**References**

METHODS FOR ESTIMATING VERTICAL HIV TRANSMISSION


**Appendix**

**Derivation of equation (2)**

Define $S$, $M_1$, and $M_0$ as previously. Let $\pi$ be the probability of vertical transmission, $I$ an indicator variable denoting whether a child is infected ($I = 1$) or uninfected ($I = 0$), and $X$ an indicator variable denoting whether a child dies ($X = 1$) or survives ($X = 0$). The proportion of children who survive to the threshold age and are infected is estimated by $(1 - M_1)S$. The proportion of children who die and are infected is estimated by $M_1P(I = 1|X = 1)$. Thus the transmission risk, $\hat{\pi}$, is estimated by

$$\hat{\pi} = (1 - M_1)S + M_1P(I = 1|X = 1) \quad (A1)$$

From Bayes’ Theorem,

$$P(I = 1|X = 1) = \frac{P(X = 1) - P(X = 1|I = 0)P(I = 0)}{P(X = 1)}$$

which is estimated by

$$\frac{M_1 - M_0}{M_1} \quad (A2)$$

Substituting in equation (A1),

$$\hat{\pi} = (1 - M_1)S + [M_1 - M_0(1 - \hat{\pi})]$$

which, on re-arranging, yields equation (2).

Note. Daniel Commenges (INSERM U 330, University of Bordeaux II) provided invaluable advice in the derivation of this formula.