

ESTIMATING THE CUMULATIVE INCIDENCE OF HIV INFECTION AMONG PERSONS WITH HAEMOPHILIA IN THE UNITED STATES OF AMERICA

PHILIP S. ROSENBERG* AND JAMES J. GOEDERT

National Cancer Institute, 6130 Executive Blvd, EPN/403, Rockville, MD 20852, U.S.A.

SUMMARY

We used complementary approaches to estimate the cumulative incidence of infection with the human immunodeficiency virus (HIV) among persons with haemophilia in the United States of America. One approach, ratio estimation, divided the cumulative number of haemophiliacs diagnosed with acquired immunodeficiency syndrome (AIDS) in the United States by the corresponding cumulative proportion with AIDS among HIV-positive subjects in the Multicenter Hemophilia Cohort Study (MHCS). The other approach, back-calculation, reconstructed past HIV incidence from national surveillance of AIDS using essentially non-parametric estimates of the hazard functions for AIDS and for pre-AIDS death. We derived confidence intervals that fully incorporated uncertainty about the hazard functions. Results from the two approaches were consistent. Around 9200 haemophiliacs became infected during the course of the epidemic. Of them, around 7000 were living with HIV or AIDS as of 31 December 1992 and at least 5630 as of 31 December 1995. Credible calculations for this group must account for those who die before AIDS and for the significantly longer incubation times in those infected as children or adolescents. The consistency of back-calculation and cohort data in haemophiliacs supports the use of back-calculation to estimate prevalence in other populations. © 1998 John Wiley & Sons, Ltd.

Statist. Med., **17**, 155–168 (1998)

1. INTRODUCTION

Before the advent of antibody testing for the human immunodeficiency virus (HIV) in March 1985, thousands of persons with haemophilia in the United States became infected with HIV by exposure to contaminated blood or blood products. While much is known about the general pattern of infections in haemophiliacs, there remains uncertainty about the total number who became infected during that era and the number who are currently living with the infection. These estimates are of interest to the haemophilia community.¹ Furthermore, the haemophilia setting provides an opportunity to evaluate one method used to estimate HIV prevalence in the general population – back-calculation from national surveillance of acquired immunodeficiency syndrome (AIDS)² – in a setting where much is known.

* Correspondence to: Philip S. Rosenberg, National Cancer Institute, 6130 Executive Boulevard, EPN/403, Rockville, MD 20852, U.S.A. e-mail: philip_rosenberg@nih.gov.

In this report we estimate the cumulative incidence of infection using two approaches. One approach, ratio estimation, divides the cumulative number of haemophiliacs with AIDS in the United States by the corresponding proportion of HIV-positive (HIV+) haemophiliacs with AIDS in the Multicenter Haemophilia Cohort Study³ (MHCS). The other approach, back-calculation,⁴ uses a statistical model to reconstruct the historical incidence of infection that best accounts for the observed epidemic of AIDS cases.

Back-calculation requires knowledge of the distribution of the incubation period between infection and AIDS, called the incubation distribution. As haemophiliacs progress to AIDS more slowly than homosexual men of the same age at infection,⁵ it would not be appropriate to apply incubation distributions obtained from homosexual cohorts. Furthermore, many HIV-infected haemophiliacs acquired the infection as children or adolescents and are progressing to AIDS more slowly than haemophiliacs infected at older ages.³ For these reasons, we derived age-appropriate estimates of the incubation distribution from the extensive follow-up of the MHCS.

A substantial number of HIV+ haemophiliacs die of haemophilia-related conditions that are not AIDS-defining.⁶ Both approaches adjust for this competing mortality to avoid underestimating HIV prevalence.

In January 1993, the Centers for Disease Control and Prevention (CDC) expanded the AIDS case definition to include criteria based on severe immunosuppression (CD4+ T-lymphocyte count <200 cells/ μ l or CD4+ T-lymphocyte per cent <14).⁷ The expansion has distorted national trends in AIDS since 1993.⁸ To avoid bias we did not use AIDS incidence beyond 31 December 1992 to estimate prevalence. This exclusion does not unduly limit the analysis because few haemophiliacs have become infected since 1986.

2. DATA

2.1. The Multicenter Hemophilia Cohort Study

Since 1982, the MHCS has enrolled 2119 haemophiliacs in thirteen United States and four Western European treatment centres and followed them for HIV infection and its immunologic, virologic, and clinical sequelae, including AIDS. A total of 1257 subjects became HIV+, 999 in the United States and 258 in Europe. AIDS is defined in the MHCS according to the 1987 case definition augmented by opportunistic illnesses added in 1993. New 1993 criteria based on severe immunosuppression are not considered as AIDS.

For each HIV+ subject, an estimated date of seroconversion has been calculated using the methods of Kroner *et al.*⁹ The estimates derive from HIV antibody testing of all stored serum samples and additional information from the medical records documenting first and last exposures to non-heat-treated factor concentrates, the cause of almost all of the infections. We based ratio estimates of national prevalence on the experience of the 999 HIV+ subjects in the thirteen United States centres.

For use in back-calculation, we estimated hazard rates for AIDS and for pre-AIDS deaths from follow-up through August 1996 of all 1257 HIV+ subjects in the entire cohort. Five centres in the United States enrolled all patients seen through 1986 regardless of HIV status. Other centres differentially enrolled HIV+ subjects, with the majority entering in 1985–1986. While considerable effort has been made to enrol all HIV+ subjects from each centre in the MHCS, some rapid progressors may have been inadvertently excluded. If so, their omission might lengthen the apparent incubation distribution. To avoid this potential bias, we included subjects outside the five centres

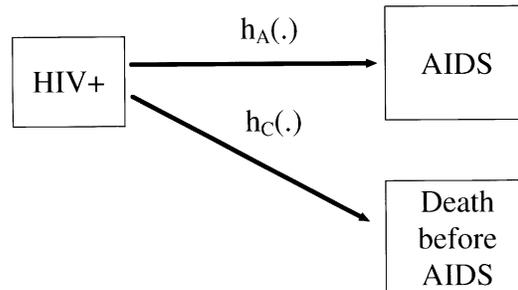


Figure 1. HIV+ haemophiliacs are at risk of AIDS and pre-AIDS death

in prospective analyses from the time of their first HIV+ antibody test rather than from their estimated date of seroconversion.

2.2. National AIDS Surveillance

We tabulated the monthly incidence of AIDS (1987 definition plus clinical conditions added in 1993) in haemophiliacs in the United States from the national registry of AIDS cases compiled by the CDC.¹⁰ We based our analysis on all cases reported to the CDC as of June 1996. We adjusted AIDS incidence for delays in reporting using standard CDC adjustment weights, and truncated the adjusted counts after 31 December 1992 to minimize the impact of immunologic criteria added in 1993. This restriction insured that national and MHCS data could be analysed on the basis of a comparable AIDS case definition.

We chose not to adjust AIDS incidence to reflect persons who develop AIDS but ultimately never get reported to the CDC. An estimated 90 per cent of AIDS cases in all groups are eventually reported.¹¹ Completeness is probably higher in groups such as haemophiliacs with relatively good access to health care.

We stratified AIDS surveillance and MHCS data according to the year of birth. We considered three birth cohorts: haemophiliacs born in 1942 or earlier; those born 1943–1957, and those born 1958 or later. Individuals in these respective groups were 35+, 20–34 and 0–19 years old at the onset of the HIV epidemic in January 1978.

3. METHODS

3.1. Ratio Estimation

Let N_{US} be the unobserved cumulative number of haemophiliacs in the United States who became infected as of calendar time $T = 31$ December 1992, let A_{US} be the cumulative number with AIDS obtained from AIDS surveillance, and let N and A be the corresponding observed values in the MHCS (United States centres). A simple ratio estimator of N_{US} is

$$\tilde{N}_{US} = \frac{A_{US}}{A/N}.$$

Some HIV+ haemophiliacs die before AIDS (Figure 1) and will not be counted as part of national AIDS surveillance. Let D_{US} be the cumulative number of deaths among haemophiliacs

with AIDS, C_{US} be the unobserved number of HIV+ haemophiliacs who died of competing causes before AIDS, and D and C be the corresponding observed values in the MHCS. A ratio estimator of cumulative mortality in HIV+ haemophiliacs from all causes, $M_{US} = D_{US} + C_{US}$, is

$$\tilde{M}_{US} = \frac{D_{US}}{D/(D + C)}.$$

A ratio estimator of HIV prevalence as of T is

$$\tilde{P}_{US} = \tilde{N}_{US} - \tilde{M}_{US}.$$

To estimate the variability, we regard the set of all HIV+ haemophiliacs as a finite population for which A_{US} and D_{US} are known. Therefore, we conditioned on A_{US} , D_{US} and N , and noted that the 2×2 table

		AIDS		
		No	Yes	
Death	No	$N - A - C$	$A - D$	$N - C - D$
	Yes	C	D	$C + D$
		$N - A$	A	N

has a multinomial distribution. We derived variance formulae for \tilde{N}_{US} , \tilde{M}_{US} , and \tilde{P}_{US} using the delta method.¹²

3.2. Disease Progression

The hazard function for AIDS, $h_A(u)$, specifies the annual rate of AIDS u years after seroconversion among those still AIDS-free. We obtained smooth and essentially non-parametric estimates of $h_A(u)$ by modelling the hazard as a spline and using an appropriate likelihood function.¹³ The likelihood accounted for left-truncation (delayed entry of subjects outside the five centres) and for right-censoring of subjects in the MHCS who were AIDS-free at the end of follow-up. The corresponding smooth estimate of AIDS-free survivorship derives from the maximum likelihood estimate $\hat{h}_A(u)$ as $\hat{S}_A(u) \equiv \exp(-\int_0^u \hat{h}_A(s) ds)$. We calculated the fully non-parametric estimator of survivorship that accounted for left-truncation, $\widehat{KM}_A(u)$, for graphic assessment of how closely the smooth estimator $\hat{S}_A(u)$ fit the data.

We estimated the variability of $\hat{h}_A(u)$ using the bootstrap. We constructed 1000 replicate data sets by resampling MHCS subjects with replacement. We saved the replicates $\hat{h}_A^*(u)$ for subsequent analysis, the ensemble representing 1000 draws from G_{h_A} , the bootstrap distribution

$$\hat{h}_A^*(u) \sim G_{h_A}(\hat{h}_A(u)).$$

In the analysis of AIDS progression, AIDS events were censored by the end of follow-up or death, whichever occurred first. We considered death before AIDS to be an independent competing risk (Figure 1). We estimated the hazard for this event, $h_C(u)$, using the same approach as for $h_A(u)$ with AIDS or the end of follow-up treated as the censoring events. As before, we saved 1000 draws from the bootstrap distribution

$$\hat{h}_C^*(u) \sim G_{h_C}(\hat{h}_C(u))$$

for subsequent analysis.

3.3. Back-calculation

3.3.1. Accounting for Competing Mortality

The infection curve $v(s)$ describes the incidence of HIV infection as a function of time. The convolution equation

$$E(Y_t) = \int_{T_0}^t v(s) \{h_A(t-s)S_A(t-s)S_C(t-s)\} ds, \quad t = 1, \dots, T$$

relates expected AIDS incidence at time t , $E(Y_t)$, to the infection curve $v(s)$ in a manner that accounts for competing mortality. In this equation, the expression $h_A(u)S_A(u)S_C(u)$ specifies the crude probability¹⁴ of developing AIDS u years after infection. The crude probabilities need not integrate to 1 as subjects may die before AIDS. We did not model secular trends in the hazard rate for AIDS as these were not substantial in the MHCS, perhaps because most HIV+ subjects in the cohort became infected during a relatively narrow time period.

We discretized the model into monthly increments to replace the convolution integral with a numerically more tractable sum. We estimated parameters of the infection curve by maximizing a Poisson log-likelihood function.

3.3.2. Using Cubic Splines to Model the Infection Curve

Cubic splines provide a flexible family of models of the infection curve, and they can be constructed to allow one to test specific hypotheses about local smoothness. The cubic B -splines provide a convenient family of basis functions. Using the notation of de Boor,¹⁵ let

$$v(s) = \sum_k B_k(s)\beta_k, \quad \beta_k \geq 0$$

where k indexes a non-decreasing knot sequence $\{t_k\}$, $B_k(s)$ is the basis function of the spline associated with the knot t_k , and β_k are parameters to be estimated. We impose non-negativity constraints on the β_k to insure that $v(s) \geq 0$.

With three continuity constraints at t_k , the cubic spline and its first two derivatives are continuous at t_k . This represents the smoothest case. One obtains a rougher join-point by reducing the number of continuous derivatives at t_k . We present details in Appendix I.

We constrained the infection curve to have a long last step of six years, $v(s) = \text{constant}$ for $T - 6 \leq s \leq T$, as this assumption is parametrically appropriate for purposes of modelling the HIV epidemic in haemophiliacs. We also imposed the constraint that $v(T_0) = 0$. Both constraints are linear in the β_k .

3.3.3. Accounting for Uncertainty about the Natural History

Estimates derived by back-calculation are sensitive to changes in the assumed incubation distribution.¹⁶ We quantified this effect using a 'double bootstrap' that incorporated uncertainty about the natural history curves and intrinsic variation of AIDS incidence. We partitioned the variability explained by these two components using a one-way analysis of variance design.¹⁷ Details are given in Appendix II.

4. RESULTS

4.1. Ratio Estimation

We derived ratio estimates of cumulative HIV incidence in haemophiliacs from the data in Table I and present the results in Table II. In the MHCS, 97 haemophiliacs born in 1942 or earlier became infected during the course of the epidemic. As of 31 December 1992, 53 of them had developed AIDS, 45 of whom had died. Sixteen of the 44 subjects without an AIDS diagnosis had also died. In the entire United States, 437 haemophiliacs born in 1942 or earlier had been diagnosed with AIDS as of 31 December 1992 and 384 of these cases had died. Using the ratio approach, we estimated that $800 = 437/(53/97)$ haemophiliacs in this birth cohort became infected and $521 = 384/(45/61)$ had died, 384 of AIDS as observed and 137 imputed to have died before AIDS.

We estimated that 9260 haemophiliacs in the entire United States became infected (Table II). The estimated standard deviation for the total was 494, yielding a coefficient of variation of ± 5 per cent and a 90 per cent confidence interval (CI) of 8450–10,070. As of 31 December 1992, an estimated 2310 (90 per cent CI, 2170–2450) HIV+ haemophiliacs had died, 1718 with AIDS diagnosed and 592 imputed to have died before AIDS.

4.2. Natural History

In the entire MHCS 1257 HIV+ subjects contributed 9053 person-years of follow-up to AIDS and 448 AIDS events as of August 1996, including 1480 person-years and 130 events accrued 10 years after seroconversion. The median duration of follow-up was 7.5 years. In each birth cohort the hazard function for AIDS, $h_A(u)$, was modelled as a cubic spline with three knots set at quartiles of the observed times-to-AIDS.

The lower panel of Figure 2 shows the estimated annual hazard rates for AIDS and the upper panel shows corresponding estimates of AIDS-free survivorship. The spline survivorship functions (smooth curves) exhibit good agreement with corresponding non-parametric estimates (step functions). We show 95 per cent pointwise confidence envelopes for subjects born in 1942 or earlier and in 1958 or later. For visual clarity, we are not presenting the confidence envelopes for subjects born in 1943–1957; error bars show the limits for this group at 5 and 10 years following seroconversion.

Progression was fastest in subjects born in 1942 or earlier. For this older group, the median time-to-AIDS was 9 years and 80 per cent had progressed by 12 years. The hazard exceeded 10 per cent per year (%/yr) by 4.8 years after seroconversion.

Subjects born during 1943–1957 progressed at a significantly slower rate. For them the median time-to-AIDS was around 13 years. The hazard increased for about 7.5 years following seroconversion and then approached a plateau. At 10 years after seroconversion, the estimated hazard rate was 7.6%/yr with 95 per cent CI 5.7–9.8.

In the youngest group of subjects, the median incubation period exceeded 14 years and the hazard was the lowest of the three groups for 10 years. After 10 years, the hazard for the youngest and middle birth cohorts was similar, with 7 per cent to 8 per cent developing AIDS per year. This suggests that once the children had grown up, their hazard was similar to the hazard of surviving young adults.

We used the same methods to estimate the hazard of death before AIDS, $h_C(u)$ (Figure 3). We obtained a satisfactory fit with no interior knots, that is, we modelled the hazard as a cubic

Table I. Cumulative incidence of AIDS and death in United States' haemophiliacs, cases through 31 December 1992

Birth cohort	MHCS				National surveillance	
	HIV+	AIDS	Deaths after AIDS	Deaths before AIDS	AIDS	Deaths after AIDS
1942 or earlier	97	53	45	16	437	384
1943–1957	300	90	69	28	816	565
1958 or later	602	140	75	22	1335	769
Total	999	283	189	66	2588	1718

Table II. Cumulative incidence of HIV infection (90 per cent confidence interval) at selected dates among United States' haemophiliacs*

Birth cohort	Ratio estimation		Back-calculation [†]			
	31 Dec 92		31 Dec 86		31 Dec 92	
1942 or earlier	800	(680, 920)	820	(690, 950)	850	(780, 1130)
1943–1957	2720	(2330, 3110)	2380	(2040, 2790)	2570	(2200, 3700)
1958 or later	5740	(5040, 6440)	5960	(4500, 6540)	5960	(5450, 10890)
Total	9260	(8450, 10070)	9160	(8090, 10240) [‡]	9380	(6330, 12420) [‡]
Combined estimate [§]	9230	(8580, 9870)				

* Values rounded to the nearest 10

[†] Intervals based on percentile method and replicates generated by double bootstrap

[‡] Interval based on $1.645 \times$ bootstrap SD

[§] Point estimate is weighted mean of ratio estimate and back-calculated number infected to 31 December 86, weighted inversely to estimated variances. Confidence interval is for weighted mean

polynomial with four degrees of freedom. The risk of pre-AIDS death was lower than the risk of AIDS, increased over time, and was highest in the oldest cohort. At ten years the estimated hazard was 6.7%/yr in persons born in 1942 or earlier, 2.3%/yr in persons born in 1943–1957, and 1.3%/yr in persons born in 1958 or later.

Finally, we obtained similar hazard estimates for AIDS and for pre-AIDS death when the analysis was restricted to HIV+ subjects in the five U.S. centres that enrolled all patients into the MHCS regardless of HIV status (data not shown).

4.3. Back-calculation

We estimated the infection curve for each birth cohort using hazard rates for AIDS and pre-AIDS death derived from the MHCS. We fitted an array of cubic splines with one or two continuous derivatives at each interior knot. Table III presents the models and their Poisson deviances.

We selected model B as giving the best fit for haemophiliacs born in 1942 or earlier as model B did not fit significantly worse than the more complex model H ($\chi^2 = 2.5$ on 1 d.f.). We also selected model B for haemophiliacs born in 1943–1957. Model E gave the best fit for haemophiliacs born in 1958 or later. A double knot at 1 January 1986 was significant in each cohort, indicating that the slope of the infection curve changed abruptly at that time.

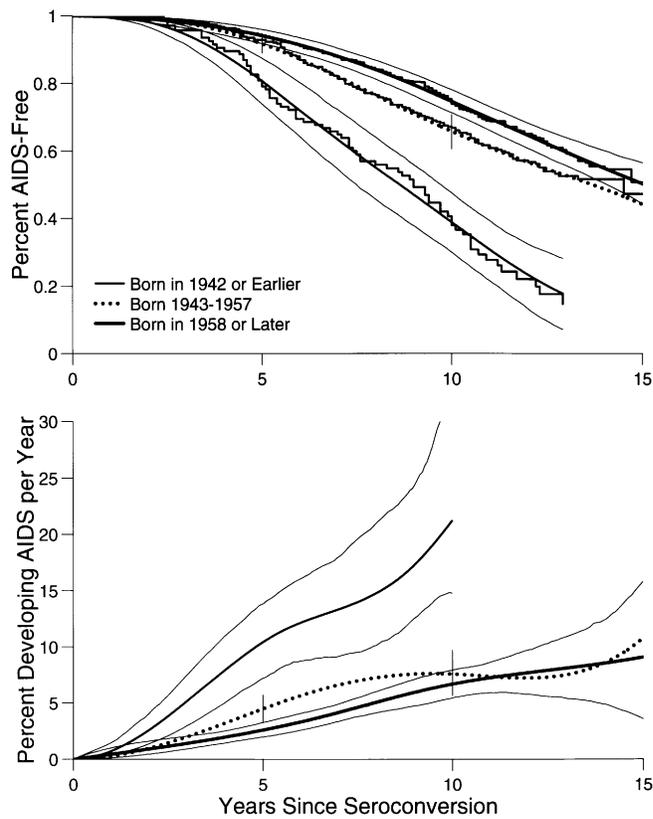


Figure 2. AIDS-free survivorship and hazard function in the MHCS



Figure 3. Hazard function of death before AIDS in the MHCS

Table III. Poisson deviance of selected cubic spline models of the infection curve in United States' haemophiliacs

Model*	d.f.	Birth cohort		
		1942 or earlier	1943–1957	1958 or later
A	{78 ⁴ , 80, 83, 86, 93 ⁴ }	134.5	178.8	201.1
B	{78 ⁴ , 80, 83, 86 ² , 93 ⁴ }	107.3	147.6	184.4
C	{78 ⁴ , 80, 83 ² , 86, 93 ⁴ }	114.4	153.2	191.8
D	{78 ⁴ , 80 ² , 83, 86, 93 ⁴ }	126.0	166.8	198.0
E	{78 ⁴ , 80, 83 ² , 86 ² , 93 ⁴ }	105.6	146.1	157.0
F	{78 ⁴ , 80 ² , 83, 86 ² , 93 ⁴ }	104.9	144.5	171.2
G	{78 ⁴ , 80 ² , 83 ² , 86, 93 ⁴ }	107.9	143.1	182.1
H	{78 ⁴ , 80 ² , 83 ² , 86 ² , 93 ⁴ }	104.8	142.6	157.0

*'78' refers to 1 January 78; exponents give the multiplicity of the knots

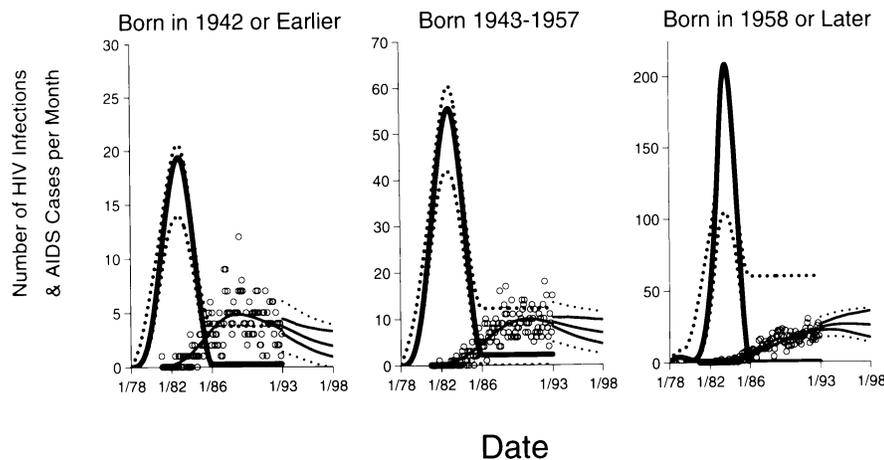


Figure 4. Back-calculation of HIV incidence among haemophiliacs in the United States

For each birth cohort, Figure 4 presents the estimated infection curves (thick solid curve) and the observed (open circles) and expected (thin solid curve) incidence of AIDS. In each cohort, the back-calculated rate of infection peaked in 1982–1983, declined sharply during 1984–1985, and fell to near-zero or zero levels after January 1986.

We projected the number of AIDS opportunistic illnesses (OIs) expected through December 1997. Even without therapeutic advances, AIDS-OIs are expected to decline in the oldest cohort as the population is depleted, and to plateau in the much larger group of HIV+ haemophiliacs born in 1958 or later.

We constructed bootstrap confidence intervals to account for random variation of the observed AIDS incidence counts and for uncertainty about the hazard functions. We generated 20 replicates of the AIDS incidence series ($R_i = 20$ in Appendix II) for each of 50 bootstrapped hazards ($R_g = 50$). Figure 4 shows 80 per cent confidence limits (dotted curves for the infection curve and projected observed AIDS-OIs, solid curves for projected expected AIDS-OIs). We calculated that

80 per cent of the total variation in the cumulative number infected to January 1986 was due to uncertainty about the hazard functions, as was one-half to three-quarters of the total uncertainty about the projections, depending on year and cohort.

Bootstrap replicates showed considerable upside uncertainty, especially for the youngest cohort. Thus, back-calculation could not determine with certainty that the rate of infection was low compared to the rate of AIDS.

Table II presents estimates of cumulative incidence through 31 December 1986 and 31 December 1992. Confidence intervals for the latter reflect the uncertainty about the rate of infection since 1986 as estimated by back-calculation.

As we know that few haemophiliacs have become infected since 1986, the estimated cumulative incidence through 1985 provides a more reliable measure. Hence, an estimated 9160 haemophiliacs became infected during the course of the epidemic. The standard error for this total was 655, yielding a coefficient of variation of ± 7 per cent.

The ranges derived by back-calculation and ratio estimation overlap. We derived a combined estimate equal to the weighted mean of the point estimates, weighted inversely by the variances. The combined estimate equalled 9230 (90 per cent CI 8580–9870). As of 31 December 1992, an estimated 2310 HIV+ haemophiliacs had died from all causes, yielding an estimate of HIV prevalence for that date of 6920 (90 per cent CI 6270–7560).

5. DISCUSSION

Back-calculation yielded credible reconstructions of HIV incidence among haemophiliacs in the United States. The estimated rate of infection peaked during 1982–1983, declined sharply during 1984–1985, and fell to near-zero levels after 1986. These trends are qualitatively consistent with direct estimates of HIV incidence in the five centres of the MHCS, and they are plausible because they are temporally correlated with public health interventions to protect the blood supply.⁹

More than half of those infected were born in 1958 or later, the generation that entered the AIDS epidemic as children or adolescents. In the MHCS, 63 per cent were aged 25 or younger at the time of infection.

The results highlight the weakness of back-calculation as a method for estimating infection rates in the recent past. It is known that few haemophiliacs have become infected since 1986, but this low incidence of infection could not be identified with certainty by back-calculation from AIDS cases.

The ratio estimator is attractive because it is simple and based on direct data. The method is valid if the infection curves in the MHCS and the United States are proportional and if the hazard rates for AIDS and pre-AIDS deaths are each homogeneous in the two groups. We tacitly make these assumptions. Unfortunately, in many settings the required direct data are not available.

Ratio estimation and back-calculation yielded consistent estimates of HIV prevalence in haemophiliacs. Both estimates are consistent with a previous one derived on the basis of infection rates in the MHCS and external data on the number of persons with haemophilia in the United States.¹⁸ The consistency between these three approaches in the haemophilia setting supports the use of back-calculation to make national estimates for other populations. Given the uncertainties of back-calculation, national estimates should ideally as here and elsewhere be derived by combining information from multiple independent sources.^{19, 20}

From the extensive follow-up of the MHCS we were able to identify a levelling of the hazard function for AIDS in three birth cohorts using essentially non-parametric methods. This plateau may

reflect a mixture of subjects with varying intrinsic frailty, existence of an endstage in the disease process, or lengthening of the incubation period due to therapy. We also noted a convergence of the hazard for subjects born in 1943–1957 and those born in 1958 or later. This phenomenon suggests that the protective effect of younger age at seroconversion may attenuate over the course of infection. Both of these aspects of the natural history were directly incorporated into the back-calculations.

Our calculations also accounted for HIV+ subjects who died before AIDS. The methods might usefully be applied to other populations with substantial competing mortality, such as injection drug users.²¹ However, the method presented here assumes that the hazards operate independently. To validate this assumption, one could check whether the rate of non-AIDS death was similar in groups with varying degrees of immunosuppression as measured by the CD4+ lymphocyte count or the HIV RNA level in serum. It would be worthwhile to do this in subsequent analyses.

We used cubic splines to model the infection curve because smooth reconstructions have intrinsic appeal and added plausibility. Our approach complements methods based on penalized likelihood of step function models with relatively narrow time increments.²² In these alternative procedures, the penalty function (often the sum of squared second differences) is a composite score of deviations over the entire span of the epidemic. In contrast, our approach is more parametric and allows us to assess the required degree of local smoothness.

The ‘double bootstrap’ effectively captured the substantial variation arising from uncertainty about the natural history. The method will be most appropriate when the natural history cohort is representative, as was likely the case here.

Our spline models of the infection curve used knots positioned in light of external information about likely change-points in HIV incidence among haemophiliacs. For other groups such external information may not be available. We obtained similar estimates for haemophiliacs by selecting the best-fit model from the set of models with three, four, or five equal length segments between January 1981 and January 1990. In each of these models, we constrained the rate of infection between January 1990 to December 1992 to be constant as there is essentially no information in AIDS incidence about trends in the rate of infection during the recent past. We also recommend that one examine the standardized residuals to diagnose possible lack of fit.

Recent studies have performed back-calculation in two dimensions to estimate age-specific infection rates from age-specific AIDS incidence data.² In contrast, the present analysis made a series of one-dimensional reconstructions for different birth cohorts. The former approach is more complex but can assess whether the infection rate in a given age group has changed over time. The latter is relatively simple and can assess generational differences in HIV incidence. With respect to the epidemic in haemophiliacs, the approach based on birth cohort analysis appears to be satisfactory.

The double bootstrap provides a clear incentive for doing back-calculation in one dimension (with stratification on birth cohort to account for age) as it provides a realistic assessment of stochastic uncertainty. Unfortunately, the computational resources needed to implement this approach in two dimensions are prohibitive.

We estimated that around 7000 haemophiliacs in the United States were living with HIV or AIDS as of 31 December 1992. It is problematic to estimate the number alive more recently. Since 1993 many haemophiliacs have been diagnosed on the basis of severe immunosuppression, an outcome of HIV infection that usually occurs years before the first opportunistic illness. These individuals are now followed to death as part of national AIDS surveillance, and the death is counted regardless of cause. As the competing mortality is relatively high in this population, some of these cases would not have lived long enough to be counted with AIDS prior to January 1993.

Using ratio estimation to account for non-AIDS mortality, we calculated *at most* 3600 cumulative deaths in HIV + haemophiliacs to 31 December 1995. This approach yields an upper bound because the AIDS case definition in the MHCS is more restrictive than in national surveillance, for example, the ratio $D/(C + D)$ is lower than the ratio needed to make an unbiased adjustment for non-AIDS deaths. Thus, at least 5630 HIV + haemophiliacs were living with HIV or AIDS as of 31 December 1995. Refining this estimate will require methods to adjust for the expanded case definition.¹⁰ The haemophilia experience may provide a useful test case to validate such methods.

Finally, we projected that AIDS incidence will remain stable, or more likely, decline by 1997, even without therapeutic advances. New highly active combination therapies that include a protease inhibitor may lead to further declines. Comparing expected incidences with what is observed may help to quantify the public health impact of these new therapies in persons with haemophilia.

APPENDIX I: USING CUBIC SPLINES TO MODEL THE INFECTION CURVE

In de Boor's notation,¹⁵ the multiplicity of the knots in $\{t_k\}$ determines the smoothness of the spline at the knots. Consider splines on the interval $[T_0, T]$ with one interior knot at t_1 . Denote knot multiplicities by exponentiation, for example, t_1^2 refers to the subsequence $\{t_1, t_1\}$. The knot sequence $\{T_0^4, t_1, T^4\}$ allows arbitrary values of the spline at the endpoints, joins separate cubic pieces at t_1 , and requires that $v(s)$ and its two derivatives be continuous at t_1 . The sequence $\{T_0^4, t_1^2, T^4\}$ drops the restriction that $v''(s)$ be continuous at t_1 , $\{T_0^4, t_1^3, T^4\}$ allows $v'(s)$ to be discontinuous at t_1 , and $\{T_0^4, t_1^4, T^4\}$ allows for a *discontinuity* of $v(s)$ at t_1 .

In practice, several interior knots may be required to model $v(s)$. As each basis function has support over the interval $[t_k, t_{k+4}]$, the basis becomes more localized by adding a new knot or by increasing the multiplicity of existing knots.

With goodness-of-fit determined by the Poisson deviance, the significance of adding a new knot can be assessed by the corresponding decrease in the deviance. Conversely, the significance of breaking continuity constraints at an existing knot is found by increasing the multiplicity of the knot and testing the corresponding decrease in the deviance.

APPENDIX II: INCORPORATING UNCERTAINTY ABOUT THE NATURAL HISTORY

We estimated the variability using the following:

Algorithm:

$$\text{For } g = 1, \dots, R_g$$

$$\text{For } i = 1, \dots, R_i$$

Generate:

$$Y_{tgi}^* \sim \text{Poisson}(E(\hat{Y}_t)), t = 1, \dots, T \text{ and}$$

$$\hat{h}_{Ag}^* \sim G_{h_A}(\hat{h}_A(u))$$

$$\hat{h}_{Cg}^* \sim G_{h_C}(\hat{h}_C(u))$$

from the bootstrap distributions G_{h_A} and G_{h_C} .

Estimate $\hat{\beta}_{gi}^{**}$ = maximum likelihood estimate obtained from Y_{tgi}^* on the basis of \hat{h}_{Ag}^* and \hat{h}_{Cg}^* .

In practice, \hat{h}_{Ag}^* and \hat{h}_{Cg}^* are taken from the ensemble of replicates obtained by bootstrapping the natural history models.

Let $N(\hat{\beta}_{gi}^{**})$ be a real-valued functional of interest, such as the cumulative number infected. From the classical one-way ANOVA, and with a dot subscript denoting averaging over that subscript, $\text{var}(N(\hat{\beta}_{gi}^{**}))$ is proportional to

$$\sum_{g=1}^{R_g} \sum_{i=1}^{R_i} (N_{gi} - N_{..})^2 = \sum_{g=1}^{R_g} \frac{1}{R_i} \sum_{i=1}^{R_i} (N_{gi} - N_{g.})^2 + \frac{1}{R_g} \sum_{g=1}^{R_g} (N_{g.} - N_{..})^2$$

$$\text{TSS} = \text{WGSS} + \text{BGSS}.$$

The percentage of variation due to uncertainty about $h_A(u)$ and $h_C(u)$ is $100 \times \text{BGSS}/\text{TSS}\%$.

ACKNOWLEDGEMENTS

We gratefully acknowledge Robert Frey from the CDC for providing AIDS surveillance data, and Julie Smith from Information Management Services Inc. and Francis Yellin from Computer Sciences Corporation for their expert help with data management.

Institutions and investigators in the Multicenter Hemophilia Cohort Study include the following: Research Triangle Institute, Rockville, MD – Kevin Kiger, Barbara L. Kroner, PhD, Scott Royal and Susan E. Wilson; National Cancer Institute, Rockville – Mitchell H. Gail, MD, PhD, James J. Goedert, MD, Thomas R. O'Brien, MD, MPH, Charles Rabkin, MD, MSc and Philip S. Rosenberg, PhD; Mount Sinai Medical Center, New York, NY – Louis M. Aledort, MD and Stephanie Seremetes, MD; Pennsylvania State University School of Medicine, Hershey – W. Christopher Ehmann, MD and M. Elaine Eyster, MD; Cornell University Medical Center, New York – Donna Di Michele, MD, and Margaret W. Hilgartner, MD; Cardeza Foundation Hemophilia Center, Philadelphia, PA – Phillip Blatt, MD and Barbara Konkle, MD; University of North Carolina, Chapel Hill – Gilbert C. White II, MD; Children's Hospital of Philadelphia – Alan R. Cohen, MD; Children's Hospital National Medical Center, Washington, DC – Naomi Luban, MD and Anne Angiolillo, MD; George Washington University Medical Center, Washington DC – Craig M. Kessler, MD; Case Western Reserve University, Cleveland Ohio – Michael M. Lederman, MD; Tulane University Medical School, New Orleans, LA – Cindy Leissing, MD; University of Texas Health Science Center, Houston – W. Keith Hoots, MD; Hospital Cantonal Universitaire, Geneva, Switzerland – Philippe de Moerloose, MD; Athens University Medical School and Laikon General Hospital, Athens, Greece – Angelos Hatzakis, MD, Titika Mandalaki, MD and Giota Touloumi, PhD; Universität München, Munich, Germany – Wolfgang Schramm, MD; University of Vienna (Austria) – Sabine Eichinger, MD and Georg Stingl, MD.

REFERENCES

1. Institute of Medicine. *HIV and the Blood Supply: An Analysis of Crisis Decisionmaking*, National Academy of Sciences, Washington, D.C., 1995.
2. Karon, J. M., Rosenberg, P. S., McQuillan, G., Khare, M., Gwinn, M. and Petersen, L. R. 'Prevalence of HIV Infection in the United States, 1984 to 1992', *Journal of the American Medical Association*, **276**, 126–131 (1996).

3. Goedert, J. J., Kessler, C. M., Aledort, L. M., Biggar, R. J., Andes, W. A., White, II G. C., Drummond, J. E., Vaidya, K., Mann, D. L., Eyster, M. E., Ragni, M. V., Lederman, M. M., Cohen, A. R., Bray, G. L., Rosenberg, P. S., Friedman, R. M., Hilgartner, M. W., Blattner, W. A., Kroner, B. and Gail, M. H. 'A prospective study of human immunodeficiency virus type 1 infection and the development of AIDS in subjects with hemophilia', *New England Journal of Medicine*, **321**, 1141–1148 (1989).
4. Brookmeyer, R. 'Reconstruction and future trends of the AIDS epidemic in the United States', *Science*, **253**, 37–42 (1991).
5. Rosenberg, P. S., Goedert, J. J. and Biggar, R. J., for the Multicenter Hemophilia Cohort Study and the International Registry of Seroconverters. 'Effect of age-at-seroconversion on the natural AIDS incubation distribution', *AIDS*, **8**, 803–810 (1994).
6. Darby, S. C., Ewart, D. W., Giangrande, P. L., Dolin, P. J., Spooner, R. J. and Rizza, C. R. 'Mortality before and after HIV infection in the complete UK population of haemophiliacs. UK Haemophilia Centre Directors' Organisation [see comments]', *Nature*, **377**, 79–82 (1995).
7. Centers for Disease Control and Prevention. '1993 revised classification system for HIV infection and expanded surveillance case definition for AIDS among adults and adolescents', *Morbidity and Mortality Weekly Report*, **41**, (RR-17), 1–19 (1992).
8. Centers for Disease Control and Prevention. 'Update: impact of the expanded AIDS surveillance case definition for adolescents and adults on case reporting – United States, 1993', *Morbidity and Mortality Weekly Report*, **43**, 160–161, 170 (1994).
9. Kroner, B. L., Rosenberg, P. S., Aledort, L. M., Alvord, W. G. and Goedert, J. J. 'HIV-1 infection incidence among persons with hemophilia in the United States and Western Europe, 1978–1990', *Journal of Acquired Immune Deficiency Syndromes*, **7**, 279–286 (1994).
10. Centers for Disease Control and Prevention. *HIV/AIDS Surveillance Report*, **8**, (no.1), 1–33 (1996).
11. Rosenblum, L., Buehler, J. W., Morgan, M. W., Costa, S., Hidalgo, J., Holmes, R., Lieb, L., Shields, A. and Whyte, B. M. 'The completeness of AIDS case reporting, 1988: a multisite collaborative surveillance project', *American Journal of Public Health*, **82**, 1495–1499 (1992).
12. Bishop, Y. M. M., Fienberg, S. E. and Holland, P. W. *Discrete Multivariate Analysis: Theory and Practice*, MIT Press, 1975.
13. Rosenberg, P. S. 'Hazard function estimation using B-splines', *Biometrics*, **51**, 874–887 (1995).
14. Gail, M. H. 'Competing risks', in Kotz, S., Johnson, N. L. and Read, C. B. (eds), *Encyclopedia of Statistical Sciences*, vol. 2, Wiley, 1982, pp. 75–81.
15. de Boor, C. *A Practical Guide to Splines*, Springer-Verlag, 1978.
16. Bacchetti, P., Segel, M. R. and Jewell, N. P. 'Uncertainty about the incubation period of AIDS and its impact on backcalculation' in Jewell, N. P., Dietz, K. and Farewell, V. T. (eds), *AIDS Epidemiology: Methodological Issues*, Birkhäuser, 1992, pp. 61–80.
17. Hay, J. W. and Wolak, F. A. 'A procedure for estimating the unconditional cumulative incidence curve and its variability for the human immunodeficiency virus', *Applied Statistics*, **43**, 599–624 (1994).
18. Brookmeyer, R. 'More on the relation between AIDS cases and HIV prevalence', *New England Journal of Medicine*, **321**, 1547–1548 (1989).
19. Day, N. E., Gore, S. M. and De Angelis, D. 'Acquired immune deficiency syndrome predictions for England and Wales (1992–97): sensitivity analysis, information, decision', *Journal of the Royal Statistical Society, Series A*, **158**, 505–524 (1995).
20. Karon, J. M., Khare, M. and Rosenberg, P. S. 'The current status of methods for estimating the prevalence of human immunodeficiency virus in the United States', *Statistics in Medicine*, **17**, 127–142 (1998).
21. van Haastrecht, H. J. A., van den Hoek, J. A. R. and Coutinho, R. A. 'High mortality among HIV-infected injecting drug users without AIDS diagnosis: implications for HIV infection epidemic modellers?', *AIDS*, **8**, 363–366 (1994).
22. Bacchetti, P., Segel, M. R. and Jewell, N. P. 'Backcalculation of HIV infection rates (with discussion)', *Statistical Science*, **8**, 82–119 (1993).