Title

Estimating the Distribution of Times from HIV Seroconversion to AIDS Using Multiple Imputation

Authors

Jeremy M. G. Taylor
Alvaro Munoz
Sue M. Bass
et al.

Publication Date

2011-10-24

STATISTICS IN MEDICINE, VOL. 9, 505–514 (1990)

ESTIMATING THE DISTRIBUTION OF TIMES FROM HIV SEROCONVERSION TO AIDS USING MULTIPLE IMPUTATION

JEREMY M. G. TAYLOR

Department of Biostatistics, UCLA School of Public Health and Jonsson Comprehensive Cancer Center, Los Angeles, CA 90024, U.S.A.

ALVARO MUÑOZ, SUE M. BASS AND ALFRED J. SAAH

Departments of Epidemiology and Biostatistics, Johns Hopkins School of Public Health, Baltimore, MD, 21205, U.S.A.

JOAN S. CHMIEL

Cancer Center Biometry Section, Northwestern University Medical School, Chicago, IL, 60611, U.S.A.

LAWRENCE A. KINGSLEY

Departments of Infectious Diseases and Microbiology/Epidemiology, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, PA, 15213, U.S.A.

AND

the
MULTICENTRE AIDS COHORT STUDY

Centres and investigators

Baltimore—Johns Hopkins School of Public Health: A. Saah, H. Farzadegan, R. Fox, J. Margolick and J. McArthur. *Chicago*—Howard Brown Memorial Clinic, Northwestern University Medical School: J. P. Phair, J. S. Chmiel, B. Cohen, K. Sheridan and S. Wolinsky. *Pittsburgh*—University of Pittsburgh School of Public Health: C. R. Rinaldo, P. Gupta, M. Ho, L. A. Kingsley, R. O. Valdiserri and A. Winkelstein. *Los Angeles*—University of California Schools of Public Health and Medicine: R. Detels, B. R. Visscher, J. Dudley, J. L. Fahey, J. V. Giorgi, D. Imagawa, J. Taylor and P. Nishanian. *Data Co-ordinating Center*—Johns Hopkins School of Public Health: A. Muñoz, S. Bass, V. Carey, L. P. Jacobson, K. Y. Liang, S. Su and S. Zeger. *National Institutes of Health*—National Institute of Allergy and Infectious Diseases: S. Vermund, P. Fleming, W. C. Blackwelder, R. A. Kaslow and M. J. van Raden; National Cancer Institute: I. Obrams.

SUMMARY

Multiple imputation is a model based technique for handling missing data problems. In this application we use the technique to estimate the distribution of times from HIV seroconversion to AIDS diagnosis with data from a cohort study of 4954 homosexual men with 4 years of follow-up. In this example the missing data are the dates of diagnosis with AIDS. The imputation procedure is performed in two stages. In the first stage, we estimate the residual AIDS-free time distribution as a function of covariates measured on the study participants with data provided by the participants who were seropositive at study entry. Specifically, we assume the residual AIDS-free times follow a log-normal regression model that depends on the covariates measured at enrolment on the seropositive participants. In the second stage we impute the date of AIDS diagnosis for the participants who seroconverted during the course of the study and are AIDS-free with use of the log-normal distribution estimated in the first stage and the covariates from each seroconverter's latest

0277–6715/90/050505–10$05.00

Received April 1989
Revised October 1989

visit. The estimated proportions developing AIDS within 4 and within 7 years of seroconversion are 15 and 36 per cent respectively, with associated 95 per cent confidence intervals of (10, 21) and (26, 47) per cent.

We discuss the Bayesian foundations of the multiple imputation technique and the statistical and scientific assumptions.

INTRODUCTION

Estimation of the distribution of times from HIV infection to AIDS has been the subject of much recent work,[1-7] and is of particular importance for modelling the growth of the AIDS epidemic.[8,9] For some data sets from longitudinal studies one can estimate this distribution almost directly[2,3] because the date of infection for each participant in the study is approximately known. For other data sets some statistical modelling is necessary to account for various features of the data, such as unknown date of infection, left truncation or biased sampling. Even for data sets in which the distribution can be measured directly there is still uncertainty in the estimates because of short follow-up times, small sample sizes, incomplete follow-up or the fact that the date of infection (or seroconversion) is known to be included in a relatively wide interval rather than known exactly.

Another source of data for estimation of the incubation period distribution is transfusion associated AIDS cases.[10,11] Because there is no information on those who have not developed AIDS, this data set gives information only on the shape of the underlying density of the time to AIDS, that is on the conditional distribution of developing AIDS within t years of infection given that AIDS occurs prior to the length of follow-up in the sample. It gives no direct information about the cumulative proportion who develop AIDS within t years of infection, unless rather strong parametric assumptions are made.[12,13]

A cohort study of the natural history of AIDS can be thought of as consisting of two separate cohorts: the seropositive or prevalent cohort, and the seroconverter or incident cohort. The seropositive cohort consists of those subjects who already have HIV infection at enrolment. The seroconverter cohort consists of those subjects who became infected during the follow-up period. A problem with estimation of the time to AIDS distribution from seropositive cohorts is that infection is only known to have occurred prior to a given date. Some authors[1,6] have attempted to solve this problem by multiple imputation of the missing dates of infection and analysis of the resulting completed data sets. A problem with estimation of the time to AIDS distribution from seroconverter cohorts is that typically the follow-up times are short and the number of AIDS cases relatively small, so that accurate estimates of the distribution at long follow-up times are unattainable. In this article, to solve this problem, we take an imputation approach in which we impute the time of AIDS diagnosis for the seroconverter cohort. Compared with previous imputation schemes, we are imputing events in the future (AIDS diagnosis) rather than events in the past (HIV infection). Our approach is similar to but more formalized than Moss *et al.*'s analysis,[14] in which they showed that the immunologic profile at the latest visit of participants in a cohort study was so poor that they predicted at least three-quarters of the seropositive individuals in the cohort would eventually develop AIDS.

The technique of multiple imputation[15] is a model based scheme for analysing data with missing values, which has been used mainly in survey research. The basis of the method is to fill in the missing values to form multiple sets of complete data for further analysis. The missing values are imputed by drawing from the predictive distribution of the missing value given the observed data. In our application we do not evaluate explicitly the predictive distribution, but rather obtain the missing value from this predictive distribution in two stages: first we draw a parameter

value from the posterior distribution of the parameters, and then we draw the missing value from the conditional distribution given that parameter value. In this way, we propagate the uncertainty through the analysis in a Bayesian sense.

DATA DESCRIPTION

The data set used in the analysis is from the Multicenter AIDS Cohort Study (MACS).[16] The study consists of 4954 homosexual or bisexual men recruited in four cities between April 1984 and March 1985. Each participant is scheduled to return at six month intervals for laboratory tests, physical examination and completion of a questionnaire. The laboratory tests and questionnaire responses relevant to this paper are HIV antibody serology tests, both ELISA and Western blot, T-helper cell percentage, platelet count, haemoglobin and age. The data used in this article are from the first eight visits of the study.

The diagnosis of AIDS is not obtained at the semiannual visits but rather through contact with the participant, his family, friends or physician. We denote participants who are HIV antibody positve at the first visit as 'seropositives', and participants who changed from antibody negative to antibody positive during the follow-up period as 'seroconverters'. In defining the interval of seroconversion we consider an HIV Western blot antibody test positive if there is any detectable antibody, however weak, and clear evidence of antibody, both ELISA and Western blot, at later visits. Table I provides details of the number of participants in the four cities. Excluded are participants with missing covariates and seropositive individuals with no follow-up information, as well as seroconverters whose conversion interval is greater than 15 months. After these exclusions, the data available for analysis consist of 1631 seropositive subjects and 276 seroconverters. In this paper we estimate the distribution of times from HIV seroconversion to AIDS. The time between HIV infection and seroconversion is thought to average less than 3 months, but can be over a year. Thus the distribution we estimate will approximate closely the distribution of times from HIV infection to AIDS.

IMPUTATION METHODS

Let $F(V|X, \theta)$ denote the distribution of times to AIDS measured from enrolment time for the seropositive group, given covariates X and parameters θ. Let $\hat{\theta}$ and cov $(\hat{\theta})$ denote the maximum likelihood estimates and their covariance obtained from the observed information matrix. In our case, the sample size is large enough that we can approximate the posterior distribution of θ by $N(\hat{\theta}, \text{cov}(\hat{\theta}))$. For the seroconverter group let T denote the time from HIV seroconversion to AIDS, and let U denote the follow-up time (that is the time from seroconversion to last follow-up or AIDS, whichever comes first). Let δ denote the occurrence of AIDS at time U, that is $\delta = 1(0)$ denotes AIDS (no AIDS). Let V denote the residual AIDS-free time if $\delta = 0$ (that is the time from last follow-up to AIDS) and $V = 0$ if $\delta = 1$; thus $T = U + V$. The basis of the method is to estimate the distribution of T for the seroconverter group, with use of the (approximately) known values of U and information on the distribution of V obtained from the seropositive group. We use the imputation technique to 'complete' the data. We draw a value of V from the estimated F given the values of the covariates at the last visit on each seroconverter who had not developed AIDS. We then add this value to the known value of U. Finally, we use standard survival analysis techniques to estimate the distribution of T. We perform the whole procedure multiple times and combine the results. The above scheme is the standard multiple imputation approach[15] slightly tailored to the aims of the analysis. Table II provides complete details of the algorithm.

Table I. HIV seropositivity and AIDS occurrence among participants in the study during four years of follow-up

	Baltimore	Chicago	Los Angeles	Pittsburgh	Total
Number of participants	1153	1102	1637	1062	4954
Number of seropositives	340	453	815	215	1823
Number of seropositives known to develop AIDS	78	89	157	33	357
Number of seroconverters	87	69	89	68	313
Number of seroconverters known to develop AIDS	3	8	4	4	19

Table II. Imputation algorithm

1. DO $j = 1, J$
2. Draw θ_j from $N(\hat{\theta}, \text{cov}(\hat{\theta}))$
3. DO $i = 1, N$
4. Draw O_{ij} from $U(P_i, A_i)$
5. If $\delta_i = 1$
6. $T_{ij} = DL_i - O_{ij}$, $\eta_{ij} = 1$
7. Else if $\delta_i = 0$
8. Draw V_{ij} from $F(V|X_i, \theta_j)$
9. $VC_{ij} = \min(V_{ij}, 48)$
10. $T_{ij} = DL_i - O_{ij} + VC_{ij}$
11. If $V_{ij} \leq 48$, $\eta_{ij} = 1$
12. Else if $V_{ij} > 48$, $\eta_{ij} = 0$
13. Form Kaplan–Meier estimate (K_j) and
14. Greenwood variance (G_j) using (T_{ij}, η_{ij})
15. $$\bar{K} = \frac{1}{J}\sum K_j$$
16. $$SE = \left[\frac{1}{J}\sum G_j + \left(\frac{J+1}{J}\right)\frac{1}{J-1}\sum (K_j - \bar{K})^2\right]^{1/2}$$
17. $$\bar{f} = \frac{1}{J}\sum f(K_j)$$
18. $$SE(\bar{f}) = \left\{\frac{1}{J}\sum [f'(K_j)]^2 G_j + \left(\frac{J+1}{J}\right)\frac{1}{J-1}\sum [f(K_j) - \bar{f}]^2\right\}^{1/2}$$
19. 95% confidence interval $= (f^{-1}[\bar{f} - 1{\cdot}96SE(\bar{f})], f^{-1}[\bar{f} + 1{\cdot}96SE(\bar{f})])$

J is number of imputed data sets; N is sample size of seroconverters; P_i (A_i) is date of last visit before seroconversion (first visit after seroconversion); DL_i is min (date of AIDS, date of last follow-up); $\delta_i = 1$ (0) if AIDS (no AIDS) at time DL_i; X_i are covariate values at last visit; $f(K) = \text{arcsine}(K^{1/2})$.

We note the following points regarding Table II:

(i) Line 2 is a random draw of θ from its posterior distribution; thus we incorporate into the analysis the uncertainty in the estimated value of θ from the seropositive group.

(ii) In line 4 the date of seroconversion is randomly drawn from a uniform distribution that covers the possible interval during which we know seroconversion occurred. In this way, we incorporate into the analysis the uncertainty associated with the actual date of seroconversion. Strictly speaking, whether or not the seroconverter has developed AIDS does contain some information about the date of seroconversion within the interval. We restricted attention, however, to seroconverters whose interval of seroconversion was at most 15 months (for 91 per cent of these seroconverters the interval was less than 8 months). For these short intervals we assume that a uniform density approximates well the conditional density of the date of seroconversion. Choice of the date of seroconversion from a uniform distribution on the interval allows us to include seroconverters with somewhat wider intervals than if we had used the midpoint of the interval in the analysis.

(iii) The combination of line 2 and line 8 represents a draw from the predictive distribution of V given the covariate value. Thus we have integrated out the dependence on θ, that is $F(V|X) = \int F(V|X, \theta) P(\theta) \mathrm{d}\theta$.

(iv) In line 9 the distribution of V is censored at 48 months, which is approximately the largest follow-up time from the seropositives. We included this censoring to avoid extrapolation beyond the range of the observed times.

(v) In line 16 there are two components to the final estimate of uncertainty: the usual sampling variability (Greenwood's formulae) estimate, and the between-sample variability representing the uncertainty associated with the imputation. The $(J + 1)/J$ correction is an adjustment for small J,[17] which in our case is negligible since $J = 100$.

(vi) In lines 17 to 19 we use the variance stabilizing arcsine root transformation to improve the confidence intervals when the estimated proportion who develop AIDS is near zero.

(vii) For lines 13 to 19 an alternative and computationally simpler scheme is to estimate the cumulative hazard $H_j (= -\log K_j)$ and its variance for each imputed data set, calculate the final estimate and confidence interval for H, and then invert the transformation to obtain the estimates and confidence intervals for the required distribution. We would expect the arcsine root and log transformations to give similar results. We preferred the arcsine root transformation in this application because of its variance stabilizing properties.

The parametric model used for F is an accelerated failure time model with log-normal distribution, that is

$$\log V_i = \theta_0 + \sum_{k=1}^{p} \theta_k X_{ik} + \sigma e_i,$$

where V is the residual time to AIDS, X_{ik} is the kth baseline covariate measured on the ith person, (θ, σ) are the parameters and e has a Gaussian distribution. This model gave a better fit to the data then other models in which e was logistic, gamma or Weibull. The covariates used were (T-helper cell percentage)$^{1/2}$, platelet count, haemoglobin and age. We restricted attention to at most four covariates. We did consider other covariates (T-helper/T-suppressor cell ratio and T-helper cell number), but the combination (T-helper cell percentage)$^{1/2}$, platelet count, haemoglobin and age gave the largest maximized log-likelihood.

RESULTS AND DISCUSSION

Despite the fact that platelet count, haemoglobin and age were statistically significant predictors of AIDS in the seropositive group, they had negligible influence on the estimated distribution and standard errors from the imputation procedure and thus we omitted them in the final analysis. The parameter estimates, standard errors and correlation matrix for the model including only (T-helper cell percentage)$^{1/2}$ appear in Table III. Table IV shows the estimated residual AIDS-free distribution evaluated using the maximum likelihood estimates of the parameters for various values of T-helper cell percentage. The wide spread of the distribution illustrates how variable are the imputed dates of AIDS diagnosis for the seroconverters.

Table V gives the estimated distribution of times to AIDS and the associated standard errors. The estimates and standard errors are based on the Kaplan–Meier estimates (lines 13 and 14 in Table II) whereas the confidence intervals are based on the variance stabilizing transformation (line 19 in Table II). Also given are the estimates from separate analyses of both the seropositive and seroconverter data from each of the four centres, with use of only the (T-helper cell percentage)$^{1/2}$ as the covariate in the model. The estimated distribution is consistent with the information from other studies,[1–3,6] in particular that very few individuals, less than 3 per cent, develop AIDS within 2 years of infection but for longer times the hazard rate increases. Most studies suggest that about 11 and 32 per cent will develop AIDS within 4 and within 7 years of infection respectively, although the effect of other factors, particularly age and mode of transmission, may be important.[1,4] This analysis, by using the multiple imputation technique to extend the follow-up, estimates the distribution up to longer times after seroconversion than is available using the data from most other studies. The results in Table V apply to gay men; caution should be exercised in extrapolation to other risk groups.

There is some evidence of difference among the four cities, with Chicago having the highest rate and Pittsburgh the lowest rate of occurrence of AIDS; but due to the large standard errors these differences should not be overinterpreted.

Also shown in Table V is the distribution of times from seroconversion to AIDS, for the combined data with no use of imputation to extend the follow-up time. The estimated distribution is similar to that from the full analysis of the combined data, but the longest time is only 42 months, compared with 84 months with use of the imputation scheme.

The estimated hazard is 2·5 per cent at 2 years, 8 per cent at 3 years, 8 per cent at 4 years, 9 per cent at 5 years and 10 per cent at 6 years. The hazard is calculated with the use of the formula $12[\hat{F}(t+\Delta) - \hat{F}(t-\Delta)]/2\Delta[1 - \hat{F}(t)]$, where Δ is 3 months. The apparent levelling of the hazard after 3 years is similar to the results of a different analysis of the MACS data,[6] and is not consistent with the Weibull distribution which has been used commonly in the estimation of the incubation period of AIDS. This hazard is more consistent with a log-logistic or a log-normal or a gamma distribution or a distribution with a hazard given by $a[\exp(bt) - 1]/[\exp(bt) - bt]$. Extrapolation of the incubation distribution beyond 7 years, with the assumption that the hazard remains constant at 10 per cent per year, suggests that the median time to AIDS is about 9·5 years. This is similar to the 10·7 years estimated by Muñoz *et al.*[6] and the 9·8 years estimated by Bacchetti and Moss.[7]

The results in Table V are based on 100 imputations ($J = 100$). The ratio of the between-imputation variance to the total variance varies depending on the time point. The ratio is 0.27 at 18 months and 0·48 at 6 years; this difference reflects the fact that the 18 month estimate is based mainly on measured data whereas at 6 years the imputation scheme plays an important role.

We fitted the accelerated failure time model to the seropositive group with the SAS procedure PROC LIFEREG. We wrote a FORTRAN program to perform the imputation step; the

Table III. Parameter estimates, standard errors and correlation matrix from analysis of seropositive group: residual AIDS-free times in months modelled as log-normal

Variable	Estimate	Standard error	Correlation θ_0	θ_1	σ
θ_0 (intercept)	1·1402	0·2239	1·00	− 0·96	− 0·18
θ_1 (T-helper %)$^{1/2}$	0·6624	0·0452	− 0·96	1·00	0·38
σ (scale)	1·1438	0·0498	− 0·18	0·38	1·00

Table IV. Percentiles* of estimated distribution of residual AIDS-Free Time

	T-helper 10	15	20	25	30	35	40
5th percentile	4	6	9	13	18	24	31
25th percentile	12	19	28	40	54	73	95
Median	25	41	60	86	†	—	—
75th percentile	55	88	—	—	—	—	—

* $\exp[\hat{\theta}_0 + \hat{\theta}_1 (\text{T-helper \%})^{1/2} + \hat{\sigma} Z_\alpha]$, where Z_α is the αth percentile of N(0, 1).
† Values larger than 100 months are not shown.

Table V. Distribution of times to AIDS after HIV seroconversion

Number of months	Estimate * (standard error) Baltimore	Chicago	Los Angeles	Pittsburgh	Combined	95% confidence interval,† combined	No imputation,‡ combined
12	0·000 (0·003)	0·016 (0·016)	0·010 (0·015)	0·000 (0·004)	0·006 (0·006)	(0·000, 0·021)	0·006 (0·005)
18	0·002 (0·008)	0·033 (0·023)	0·023 (0·020)	0·016 (0·018)	0·019 (0·010)	(0·004, 0·041)	0·017 (0·009)
24	0·009 (0·015)	0·038 (0·026)	0·031 (0·025)	0·022 (0·021)	0·026 (0·012)	(0·008, 0·053)	0·019 (0·010)
30	0·029 (0·073)	0·078 (0·039)	0·045 (0·030)	0·027 (0·026)	0·048 (0·016)	(0·022, 0·083)	0·042 (0·017)
36	0·062 (0·032)	0·115 (0·046)	0·086 (0·040)	0·061 (0·040)	0·083 (0·020)	(0·048, 0·126)	0·098 (0·030)
42	0·082 (0·038)	0·151 (0·053)	0·129 (0·046)	0·103 (0·048)	0·121 (0·024)	(0·078, 0·171)	0·174 (0·050)
48	0·112 (0·047)	0·195 (0·061)	0·175 (0·056)	0·121 (0·053)	0·155 (0·027)	(0·105, 0·212)	
54	0·147 (0·056)	0·249 (0·068)	0·216 (0·062)	0·140 (0·059)	0·191 (0·031)	(0·134, 0·255)	
60	0·185 (0·063)	0·306 (0·076)	0·257 (0·067)	0·164 (0·067)	0·229 (0·034)	(0·166, 0·299)	
66	0·227 (0·069)	0·353 (0·081)	0·294 (0·073)	0·190 (0·078)	0·263 (0·037)	(0·195, 0·338)	
72	0·267 (0·071)	0·391 (0·087)	0·328 (0·081)	0·214 (0·090)	0·298 (0·041)	(0·221, 0·381)	
78	0·306 (0·079)	0·423 (0·090)	0·353 (0·089)	0·240 (0·108)	0·332 (0·045)	(0·247, 0·421)	
84					0·362 (0·052)	(0·264, 0·465)	

* Based on 100 imputed data sets.
† Using arcsine root transformation.
‡ Estimate when residual time is censored at time zero, that is no imputation used to extend the follow-up.

program included calls to IMSL subroutines (GGNSM and GGNPM) to generate multivariate and univariate Gaussian quantities.

Although the analysis is easy to perform and describe, it does involve certain scientific and statistical assumptions. A major assumption implicit in the title of the paper is that the distribution of times from infection to AIDS is constant over chronological time. There is some evidence of a lengthening of incubation times because of the greater availability of partially

effective treatments and because of the reduced incidence of Kaposi's sarcoma (a generally less severe clinical manifestation) relative to the total incidence of AIDS among homosexual men with AIDS. There are known biases[18] in the analysis of data from seropositive cohorts and with use of the enrolment date as time zero when the natural zero time is the unknown date of infection. In this situation, however, one can justify the analysis if all the information about the future course of the disease is contained in the current value of the covariates; such covariates are sometimes called surrogate response variables.[19] This assumption would seem a reasonable approximation to the truth given that the T-helper cell percentage is a good predictor of AIDS.

Providing notation for the above heuristics, let V denote the residual AIDS-free time at time U for individuals who seroconverted at time zero and who have not developed AIDS. Let $\{X(u), 0 \leqslant u \leqslant U\}$ be the history of the time varying covariates and let $\delta(u)$ be an indicator of AIDS, that is $\delta(u) = 1$ if AIDS has developed by time u and $\delta(u) = 0$ otherwise. Let T be the time from infection to AIDS. Let $F(V, U, X(U))$ be the distribution of V given $X(U)$ as a function of U, that is $F(v, U, X(U)) = P(V < v | U, X(U), \delta(U) = 0)$. To justify the above analysis we need to assert that $F(V, U, X(U))$ depends on U only through $X(U)$. This follows if we assume that $P(V < v | U, X(U), \delta(U) = 0) = P(V < v | X(U), \delta(U) = 0)$. We attempted to check this assumption in two ways. First, we estimated at each centre $P(V < v | X(U), \delta(U) = 0)$ for the seropositive groups. We found the four estimates with use of T-helper cell percentage as the covariate were similar, which, given that the infections in Los Angeles probably occurred slightly earlier and those in Pittsburgh slightly later than at the other two centres, provides evidence that favours the assumption. Secondly, for the seroconverter group, we estimated $P(V < v | X(U), U = U_i, \delta(U) = 0)$, where $U_i = 3$, 9, 15 and 21 months. Although the sample size, particularly the number of AIDS cases, is small, the four estimated models did appear similar. Thus, although we are unable to validate the assumptions, the available data did suggest they are not unreasonable.

The large sample size for the accelerated failure time analysis of the seropositive group suggests that $N(\hat{\theta}, \text{cov}(\hat{\theta}))$ is a reasonable approximation to the posterior distribution of θ, so this should not be a concern with the validity of the analysis. In addition, when we set θ equal to $\hat{\theta}$ in the imputation scheme rather than draw it from $N(\hat{\theta}, \text{cov}(\hat{\theta}))$, it made very little difference to the estimated distribution and reduced the standard errors by less than 2 per cent.

In estimation of the distribution of times to AIDS from a cohort study, one must consider the statistical issues of length-biased sampling and left-truncation.[18] Because the final estimated distribution comes from only the seroconverter group, left-truncation does not occur. The problem of length-biased sampling for the seropositive group is handled by estimation of the residual AIDS-free time distribution conditional on the T-helper cell percentage value. We assume that given this covariate the residual AIDS-free time is independent of the time interval from seroconversion to the time of measurement of the covariate.

T-helper cell percentage, the covariate used as a predictor of the residual time in this analysis, is subject to measurement error. This suggests that we could obtain a better predictor of AIDS-free residual time with use of multiple measurements on each individual at different time points. Incorporation of repeated measurements in the analysis would probably reduce the uncertainty in the final estimated distribution, although we suspect not by a large amount, because most of the information about the future course of the disease is contained in the current covariate measurement and the addition of prior measurements may not dramatically improve our knowledge for most individuals. In addition, this would require a more complicated model than $F(V|X, \theta)$, which is a drawback from the appeal of our intuitively simple analysis.

Other markers, in particular Neopterin, P24 antigen and beta-2-microglobulin,[14,20] are possible covariates to include in the model. These will likely be more useful additions to the T-helper cell percentage than platelet count, haemoglobin and age. These variables, however, were measured only at the necessary visits on a small fraction of the study participants.

An analogous, more non-parametric multiple imputation approach would be a 'hot deck' procedure. In such an analysis, we would find multiple matches for each seroconverter from the seropositive group. For each seroconverter we would extend the follow-up by using the exact follow-up and AIDS incidence information from the matching seropositive participant. The matching would be based on the covariates; we would use the baseline values of the seropositive group to match the covariate values from the last visit of the seroconverter group. This scheme avoids the need for the parametric modelling in the first stage of the analysis, but it does require a method to perform the matching.

Another point worth mentioning is that although we have incorporated many sources of uncertainty in the final estimate, we did not incorporate the uncertainty in the model selection stage. That is, we essentially mined the data to discover that the T-helper cell percentage, platelet count, haemoglobin and age were the best covariates to use with a log-normal distribution. We then performed the analysis conditional on this selected model and covariates. Strictly speaking, one should incorporate the uncertainty associated with the model selection stage, possibly with use of the methods suggested by Hodges.[21] This, however, is not standard practice and the computations involved make it unattractive.

In summary, we used data from a study with 4 years follow-up in conjunction with multiple imputation techniques to estimate the distribution of times from HIV seroconversion to AIDS up to 7 years after seroconversion.

ACKNOWLEDGEMENTS

This work was partially supported by National Institute of Health grants and contracts CA-16042, AI-72631, AI-72632, AI-72634, AI-72676, AI-32535.

REFERENCES

1. Taylor, J. M. G., Schwartz, K. and Detels, R. 'The time from infection with human immunodeficiency virus (HIV) to the onset of AIDS', *Journal of Infectious Disease*, **154**, 694–697 (1986).
2. Eyster, M. E., Gail, M. H., Ballard, J. O., Al-Mandhing, H. and Goedert, J. J. 'Natural history of human immunodeficiency virus in hemophiliacs: effect of T-cell subsets, platelet counts and age', *Annals of Internal Medicine*, **107**, 1–6 (1987).
3. Curran, J. W., Jaffe, H. W., Hardy, A. M., Morgan, W. M., Selik, R. M. and Dondero, T. J. 'Epidemiology of HIV infection and AIDS in the United States', *Science*, **239**, 610–616 (1988).
4. Giesecke, J., Scalia-Tamber, G., Berglund, O., Berntorp, E., Schulman, S. and Stigendal, L. 'Incidence of symptoms and AIDS in 146 Swedish haemophiliacs and blood transfusion recipients infected with human immunodeficiency virus', *British Medical Journal*, **297**, 99–102 (1988).
5. Brookmeyer, R. and Goedert, J. J. 'Censoring in an epidemic with application to hemophilia associated AIDS', *Biometrics*, **45**, 325–335 (1989).
6. Muñoz, A., Wang, M.-C., Bass, S., Taylor, J. M. G., Kingsley, L. A., Chmiel, J. S. and Polk, B. F. for the Multicenter AIDS Cohort Study. 'AIDS-free time after HIV-1 seroconversion homosexual men', *American Journal of Epidemiology*, **130**, 530–539 (1989).
7. Bacchetti, P. and Moss, A. R. 'Incubation period of AIDS in San Francisco', *Nature*, **338**, 251–253 (1989).
8. Brookmeyer, R. and Gail, M. H. 'A method for projecting the AIDS epidemic', *Lancet*, **ii**, 99 (1988).
9. Taylor, J. M. G. 'Models for the HIV infection and AIDS epidemic in the United States', *Statistics in Medicine*, **8**, 45–58 (1989).
10. Lui, K. J., Lawrence, D. N., Morgan, W. M., Peterman, T. A., Haverkos, H. W. and Bregman, D. J. 'A model-based approach for estimating the mean incubation period of transfusion associated acquired immunodeficiency syndrome', *Proceedings of the National Academy of Sciences*, **83**, 3051–3055 (1986).
11. Medley, G. F., Anderson, R. M., Cox, D. R. and Billard, L. 'Incubation period of AIDS in patients infected via blood transfusion', *Nature*, **328**, 719–721 (1987).
12. Kalbfleisch, J. D. and Lawless, J. F. 'Estimating the incubation period for AIDS patients', *Nature*, **333**, 504–505 (1988).
13. Lagakos, S. W., Barraj, L. M. and De Gruttola, V. 'Nonparametric analysis of truncated survival data with applications to AIDS', *Biometrika*, **75**, 515–523 (1988).

14. Moss, A. R., Bacchetti, P., Ormand, D., Krampf, W., Chaisson, R. E., Stites, D., Wilber, J., Allain, J.-P. and Carlson, J. 'Seropositivity for HIV and the development of AIDS or AIDS related condition: three years follow up of the San Francisco General Hospital Cohort', *British Medical Journal*, **296**, 745–750 (1988).
15. Rubin, D. B. *Multiple Imputation for Survey Nonresponse*, Wiley, New York, 1986.
16. Chmiel, J. S., Detels, R., Kaslow, R. A., Van Raden, M., Kingsley, L. A., Brookmeyer, R. and the Multicenter AIDS Cohort Study Group. 'Factors associated with prevalent human immunodeficiency virus (HIV) infection in the Multicenter AIDS Cohort Study', *American Journal of Epidemiology*, **126**, 568–577 (1987).
17. Rubin, D. B. and Schenker, N. 'Multiple imputation for interval estimation from simple random samples with ignorable nonresponse', *Journal of the American Statistical Association*, **81**, 366–374 (1986).
18. Brookmeyer, R. and Gail, M. 'Biases in prevalent cohorts', *Biometrics*, **43**, 739–749 (1987).
19. Cox, D. R. 'A remark on censoring and surrogate response variables' *Journal of the Royal Statistical Society, Series B*, **45**, 391–393 (1983).
20. Melmed, R. N., Taylor, J. M. G., Detels, R., Bozorgmehri, M. and Fahey, J. L. 'Serum Neopterin changes in HIV infected subjects: indicator of significant pathology, CD4 T cell changes and development of AIDS', *Journal of Acquired Immune Deficiency Syndromes*, **2**, 70–76 (1989).
21. Hodges, J. S. 'Uncertainty, policy analysis and statistics', *Statistical Science*, **2**, 259–275 (1987).