Journal of Biopharmaceutical Statistics, 4(2), 129–146 (1994)

A MODEL-BASED APPROACH TO ESTIMATE THE AIDS-FREE TIME DISTRIBUTION IN HOMOSEXUAL MEN USING LONGITUDINAL DATA

Dorothy D. Dunlop,[1] *Ajit C. Tamhane,*[2] *Joan S. Chmiel,*[3] *and John P. Phair*[4]

[1]Center for Health Services and Policy Research
Northwestern University
Evanston, Illinois 60208

[2]Departments of Statistics and Industrial Engineering and Management Sciences
Northwestern University
Evanston, Illinois 60208

[3]Cancer Center Biometry Section
Northwestern University Medical School
Chicago, Illinois 60611

[4]Department of Medicine
Comprehensive AIDS Center
Northwestern University Medical School
Chicago, Illinois 60611

Key words. Survival analysis; AIDS incubation time; Generalized linear models; Classification trees; Multiple imputation; Jackknife

Abstract

A model-based approach is developed to estimate the distribution of time from seroconversion to diagnosis with acquired immunodeficiency syndrome (AIDS) as a function of selected time-dependent covariates. The approach is applied to longitudinal data collected over 4 years of follow-up from 450 men seropositive for the human immunodeficiency virus (90 AIDS cases) and 62 seroconverters (nine AIDS cases) participating in the Chicago part of the Multicenter AIDS Cohort Study. Because of the periodic nature of monitoring, the seroconversion time is interval-censored for seroconverters and left-censored for seroprevalent cohort members; the end-point is right-censored for 413 individuals. Since serological monitoring is not continuous but only at regularly scheduled visit times, a model for the *discrete* hazard rate (DHR) is proposed that is a generalized linear model that relates the DHR to the covariate history through the complementary log-log link. Classification trees are used for preliminary screening of covariates to identify predictors of AIDS that should be incorporated into the DHR model. The missing seroconversion times for all men are imputed 100 times to obtain 100 completed datasets from which the parameters of the DHR are then estimated using the maximum-likelihood method. The final DHR model includes the following infection progression (marker) variables: CD4%, hemoglobin, p24 antigen, and CD4% × p24 antigen interaction. Using this DHR model, the discrete survival distribution of AIDS-free time is estimated for the given population. The jackknife procedure is used to assess the precision of the estimated survival distribution.

Introduction

The problem of estimating the distribution of time from initial infection with human immunodeficiency virus (HIV) to the onset of acquired immunodeficiency syndrome (AIDS) (also referred to as the *AIDS-free time* or *incubation time*) has received much attention in recent years (1–5). In the present paper we address this problem in a novel way by exploiting the information on time-dependent covariates available from longitudinal data and incorporating it into a model for the distribution of AIDS-free time. We apply this model to data from the Chicago part of the Multicenter AIDS Cohort Study (MACS), a prospective longitudinal study of homosexual or bisexual men recruited between April 1984 and March 1985 and followed at semiannual intervals (6). The homosexual male population is of interest because it forms a large risk group for HIV infection and AIDS (7).

There are serious missing data problems associated with such a homosexual cohort making the estimation of the AIDS-free (survival) time distribution a formidable task. First, the exact dates of infection are unknown for these individuals. Because of the difficulty in determining the infection date, the seroconversion data (i.e., the date at which antibody to HIV is first detectable) is often used as the beginning point of the AIDS-free stage of the HIV infection, and we shall follow the same practice. But even the seroconversion date is unknown for someone who was seropositive (HIV+) on entry to the study, and it is only known to lie in a given interval for a seroconverter. Second, although many AIDS cases were observed among the prevalent cohort, very few cases were observed among the incident cohort. Muñoz *et al.* (1,2) provided a solution to the first problem by imputing the seroconversion times of the prevalent individuals based on data from selected infection progression (marker) variables measured at enrollment by using a model that relates these variables to the time since seroconversion; this model was estimated using the data from seroconverters whose seroconversion dates were known (within ±4 months). The uncertainty introduced due to imputing (rather than "knowing") the seroconversion times was assessed by using the multiple imputation methodology (8,9). Taylor *et al.* (3) employed the same methodology, but imputed the AIDS onset times for seroconverters who had not had an event and who would otherwise be right-censored. Note that while Muñoz *et al.* (1,2) imputed events in the past, Taylor *et al.* (3) imputed events in the future. Once a "complete" dataset was obtained by imputing the missing values, these investigators used Kaplan-Meier-type methods to estimate the survival distribution.

In the present paper we build on these two methods by incorporating the following enhancements:

1. We identify the most important longitudinally observed covariates predictive of AIDS onset and incorporate them into a model for the distribution of the AIDS-free time. They turn out to be certain key laboratory variables (selected by screening a large group of behavioral, laboratory, and other variables), which may be regarded as marker variables (10).
2. We employ the resulting model to estimate the AIDS-free time distribution by using the information on the longitudinal covariate trends after seroconversion. This is a model-based approach to the estimation of the survival time distribution, in contrast to the nonparametric Kaplan-Meier-type approaches employed by Muñoz *et al.* (1,2) and Taylor *et al.* (3), which do not utilize any covariate information.
3. We explicitly take into account the interval-censored nature of sampling in our modeling. The previous investigators (1–3) assumed continuous distributions.
4. We use the Muñoz *et al.* (1,2) method for imputing the seroconversion dates of seroprevalent individuals, but instead of using only

Table 1. Variables Analyzed from Chicago MACS Database

Laboratory tests	Clinical data	Behavioral data
Complete blood count[a]	Communicable diseases[b]	Recreational drug use
HIV enzyme assay	Sexually transmitted diseases[c]	Cocaine
p24 antigen		MDA[d]
Immunoglobulin: G, A, M	Anergy	Sexual practices
Cytomegalovirus antibody	Herpes simplex	Age at first sex with male
Hepatitis B surface antigen	Herpes zoster	Age began regular male sex
Hepatitis B surface antibody	AIDS-related symptoms	
Lymphocyte phenotyping	Constitutional symptoms	Receptive anal intercourse
T lymphocytes (total)	Lymphadenopathy	Insertive anal intercourse
T-suppressor lymphocytes	Thrush	Number of sexual partners
T-helper lymphocytes	Rectal trauma	Lifetime male partners
	Parenteral exposure	Male partners last 6 mo
	Weight	Female partners last 6 mo
	Age	

[a] White blood cells, red blood cells, hemoglobin, hematocrit, platelets, neutrophils, bands, lymphocytes, monocytes, eosinophils, basophils.
[b] Ameba, *Giardia*, lice, other parasites, scabies, *Shigella*.
[c] Syphilis, gonorrhea, urethritis, condylomata acuminata.
[d] Methylenedioxyamphetamine.

the marker variables at entry, we use data from all follow-up visits. This may yield more reliable imputed values of their seroconversion dates.

The following statistical methodologies are used in the present paper: classification and regression trees (CART) (11) to do exploratory screening of potential predictor variables and interactions among them, discrete (interval-censored) survival models (12), generalized linear models (13), multiple imputations of missing data (8,9), and jackknifing (15) to estimate the survival probabilities at prescribed time points and their standard errors.

Background of the Study

The ensuing analyses use data gathered from 1102 homosexual or bisexual men enrolled into the Chicago MACS study between April 1984 and March 1985. Data were gathered at baseline and succeeding semiannual visits from an interview, physical examination, and laboratory evaluation. A summary of variables we analyzed is shown in Table 1. The data are from the first nine semiannual visits of the Chicago MACS study.

HIV antibody status was determined at each visit. An individual was considered to be HIV antibody positive based on a positive ELISA HIV antibody serology test confirmed by a positive Western blot. The cohort may be divided into three groups based on HIV antibody status over 4 years of follow-up: 589 men who were negative for the entire span of follow-up comprise the seronegative (HIV-) cohort and are not included in the present analyses; 450 men who were seropositive for HIV at entry comprise the prevalent cohort; and 63 men who were HIV negative at the entry time but tested positive during the follow-up comprise the incident or seroconverter cohort. During the first 4 years of the study, 99 cases of AIDS were diagnosed, 90 from the prevalent cohort and nine from the incident cohort. One seroconverter was HIV- at all visits but was later diagnosed with AIDS and is excluded from analyses. This follow-up period represents the experience of the cohort prior to the availability of Zidovudine and other treatments; we do not concern ourselves with effects of antiviral treatment that later become important.

Statistical Methods

Development of the Model

As noted before, the data are interval-censored because the visits are approximately 6 months apart. To develop a simple, discrete survival model, a 6-month time unit is used. (The actual data show that 47% of all visits took place within 5.5–6.5 months of the previous visits and 84% of the visits took place within 4.5–7.5 months of the previous visits. The mode of the intervisit time distribution was 6 months for earlier visits and 4–5 months for later visits.)

Let T_C denote the AIDS-free time measured on a continuous scale; T_C is the time between HIV serconversion and the diagnosis of AIDS. This time is not observable for various reasons. For prevalent cohort individuals, seroconversion occurred prior to study entry and is left censored. For both prevalent and incident cohort individuals, AIDS may not occur during the follow-up period; i.e., this event is right-censored. Even if both events occur during the study for an individual, they are interval-censored due to periodic monitoring. To account for these different types of censoring, we define a discrete approximation to T_C, denoted by T_D, which is the number of discrete time periods from HIV seroconversion to AIDS. In this example, a period is 6 months, thus $T_D = 2$ means 12 months. Discrete time is marked backward from the first HIV+ study visit to determine the period in which seroconversion occurred; this period defines the discrete time origin, as shown in Figure 1. Let V be the discrete time from seroconversion to the first HIV+ study visit. Note that $V = 0$ for the incident cohort and $V > 0$ for most prevalent cohort members. In fact, V is not known for prevalent cohort members and is estimated building on methods of Muñoz *et al.* (1,2) and Taylor

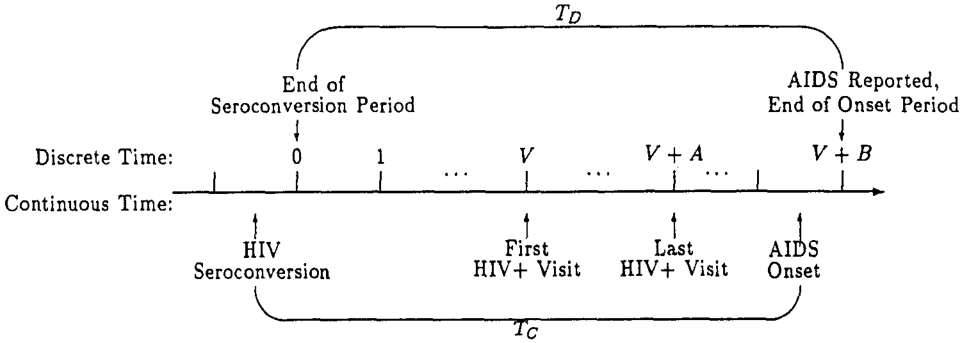

Figure 1. HIV incubation measured in discrete time.

et al. (3). Let A be the number of follow-up time periods from the first HIV+ visit to the last HIV+ visit and B be the number of time periods from the first HIV+ visit until the end of the time period in which AIDS onset occurs (i.e., the visit AIDS is reported); thus $T_D = V + B$. Note that for individuals who remain AIDS-free during follow-up, $B > A$; otherwise $B = A$.

Define the discrete hazard rate (DHR) associated with T_D by

$$\theta_k = \Pr\{T_D = k \mid T_D > k - 1\} \; (k = 1, 2, \ldots). \tag{1}$$

The discrete survival function can be expressed as

$$S_D(k) = \Pr\{T_D > k\} = \prod_{r=1}^{k} (1 - \theta_r). \tag{2}$$

In this paper we model θ_k as a function of an individual's covariate history up to the beginning of the kth time period and then use this model to estimate $S_D(k)$ via (2). The covariate history for the ith individual is summarized in $x_{ik} = \{x_{ijk}(1 \leq j \leq J)\}$, $k = V_i, \ldots, V_i + A_i$, $i = 1, \ldots, N$, where N is the number of people, J is the number of covariates, V_i indexes the discrete time of the first postseroconversion visit, and A_i is the discrete follow-up time from the first postseroconversion visit until the individual either drops out of the study or is diagnosed with AIDS, whichever comes first. Three models for θ_k, complementary log-log (CLL), log logistic, and complementary log, were evaluated. These three models can be derived from a continuous hazard model, which assumes an underlying multiplicative or additive hazard rate. Breslow and Day (16) provide a method to compare models based on a log likelihood ratio criterion. Using this method, the CLL model was selected, which is associated with a multiplicative hazard rate. Details are given in Dunlop (17).

Let θ_{ik} denote the value of θ_k for the ith individual. We fitted the following generalized linear model with a complementary log-log (CLL) link to θ_{ik}

$$\ln[-\ln(1-\theta_{ik})] = \beta_{0k} + \sum_{j=1}^{J}\sum_{l=0}^{L_j} \beta_{jkl}x_{ijk-l-1}. \tag{3}$$

In (3), l indexes the number of lag periods, $x_{ijk-l-1}$ is the value of the jth covariate for the ith individual at the beginning of the $(k-l)$th period, β_{jkl} is the "effect" of the corresponding covariate value, and β_{0k} is the intercept term ($1 \le i \le N$, $1 \le j \le J$, $1 \le k \le K$, $1 \le l \le L_j$); here K is some maximum number of periods for which the model is to be estimated ($K \le \max_{1 \le i \le N}(V_i + A_i)$).

The β's in model (3) are unknown parameters to be estimated. The choice of the covariates x_j's and their lags L_j's is based on a preliminary exploration of the data, and data on the θ_{ik}'s are available through indicator variables Y_{ik}'s, where $Y_{ik} = 0$ or $Y_{ik} = 1$ depending on whether the ith individual is AIDS-free or is diagnosed with AIDS during the kth period. Thus

$$\theta_{ik} = \Pr(Y_{ik} = 1 \mid Y_{i0} = \cdots = Y_{ik-1} = 0).$$

Estimation of the Model for "Complete" Data

We now discuss how model (3) can be fitted if we have "complete" data of the following form:

$$\{(x_{ik}, y_{ik}),\ k = V_i, \ldots V_i + A_i\}_{i=1}^{N},$$

where y_{ik} is the observed value of Y_{ik} and the V_i are assumed to be known for all people. Note that when $k = V_i + A_i$ then $y_{ik} = 0$ if the ith individual is right-censored and $y_{ik} = 1$ otherwise. Assuming independence between individuals, the likelihood function can be written as

$$\mathfrak{L} = \prod_{i=1}^{N}\prod_{k=V_i+1}^{V_i+A_i} \theta_{ik}^{y_{ik}}(1-\theta_{ik})^{1-y_{ik}}. \tag{4}$$

Note that (4) is a product of ΣA_i independent Bernoulli probabilities. If there are any missing visits between two AIDS-free visits, the corresponding terms are omitted from this product; the likelihood is not affected in any other way (unless the model includes lagged covariate effects, in which case additional terms may be omitted because of lacking data). It is assumed that such visits are missing at random.

Prentice and Gloeckler (20) have derived the likelihood equations for the generalized linear model with a CLL link (3) whose solution yields the maximum likelihood estimate (MLE) $\hat{\beta}$ of the parameter vector β; they also have given a formula for the asymptotic variance-covariance matrix of $\hat{\beta}$. The iterative weighted least squares is an algorithmic method for obtaining $\hat{\beta}$. This is the method implemented in the GLIM (21) package that was used in the present work.

To have parsimony in the number of parameters, models were restricted to $\beta_{jkl} = \beta_{jl}$; i.e., the "effects" were assumed to be time-invariant. Also, instead of having a separate β_{0k} term for each $k = 1, 2, \ldots, K$, we approximated β_{0k} as a cubic in k,

$$\beta_{0k} \approx \alpha_0 + \alpha_1 k + \alpha_2 k^2 + \alpha_3 k^3. \tag{5}$$

Multiple Imputations of Seroconversion Dates

To fit the model (3) by maximizing the likelihood function (4), we need to know the seroconversion dates for all individuals, but these dates are unknown. Therefore, we impute them using the following modification of a method due to Muñoz *et al.* (1,2). The strategy is to draw an elapsed seroconversion time for each postseroconversion visit $a = 0, \ldots, A_i$ from a distribution estimated using the individual's laboratory data. The imputed seroconversion time is the median of the corresponding $(A_i + 1)$ seroconversion dates. This imputed time is used to estimate V, the discrete time from seroconversion to the first HIV+ visit. Multiple seroconversion times are imputed for each person to account for the uncertainty introduced by estimating rather than knowing this date.

Let U denote the time since seroconversion at any given follow-up visit and let $F(u \mid z, \gamma) = \Pr\{U \leq u \mid z, \gamma\}$ denote the distribution function of U given the covariate vector z (which will in general be different from x used to model the distribution of T_D) at that visit; here $F(\cdot)$ is assumed to have a known functional form and γ is an unknown parameter vector. Note that U is the continuous time from seroconversion to a given follow-up visit, whereas T_D is the number of periods from seroconversion to AIDS onset. We estimate γ from the data on the incident cohort (denoted by I) that have known seroconversion intervals. Let z_{ia} be the vector of covariates and U_{ia} the time since seroconversion, both measured at the ath visit, and let w_i be the number of periods between the last HIV− visit and the first HIV+ visit for the ith individual, $i \in I$; for all except two incident cohort members we have $w_i = 1$; i.e., there are no missing visits straddling the change of serostatus. Conditioning on the fact that the ith individual is followed for A_i periods and hence $U_{ia} \leq A_i + w_i$, the MLE $\hat{\gamma}$ of γ can be found by maximizing the likelihood function

$$\mathfrak{L}\{ = \prod_{i \in I} \prod_{a=0}^{A_i} \Pr\{a < U_{ia} \leq a + w_i \mid U_{ia} \leq A_i + w_i\}$$

$$= \prod_{i \in I} \prod_{a=0}^{A_i} \left[\frac{F(a + w_i \mid z_{ia}, \gamma) - F(a \mid z_{ia}, \gamma)}{F(A_i + w_i \mid z_{ia}, \gamma)} \right], \tag{6}$$

which is obtained by treating the U_{ia} as independent r.v.'s. Liang and Zeger (22) have shown that this results in consistent estimates.

Once the MLE $\hat{\gamma}$ and its estimated covariance matrix, $\widehat{\text{Cov}}(\hat{\gamma})$, are obtained by using the standard maximum likelihood (ML) methods (23), the imputation of the seroconversion dates for individuals in the incident or prevalent cohort (denoted by I and P, respectively) proceeds according to the following algorithm.

Imputation Algorithm

1. To reflect the uncertainty due to estimating γ rather than "knowing" it, draw a random sample $\hat{\gamma}_1, \hat{\gamma}_2, \ldots, \hat{\gamma}_M$ from the posterior distribution of γ, which, for a large number of incident cohort observations, may be well approximated by a multivariate normal (MVN) distribution, $\gamma \sim$ MVN $(\hat{\gamma}, \widehat{\text{Cov}}(\hat{\gamma}))$.
2. For the ith individual ($i \in I, P$) and sampled vector $\hat{\gamma}_m(1 \leq m \leq M)$, draw u_{iam} from the distribution $F(\cdot \mid z_{ia}, \hat{\gamma}_m)$ for each nonmissing visit $a = 0, 1, \ldots, A_i$ for the ith individual, taking into account known bounds on u_{iam}. The elapsed seroconversion time at visit a is bounded, $a < U_{iam} \leq U^* + a$. We used $U^* = w_i$ for incident cohort members and $U^* = 20$ periods for prevalent cohort members; the latter value of U^* corresponds to January 1, 1975, which is a conservative estimate of the first HIV infection in the United States. (Values of U^* as high as 30 were used, but did not affect the results appreciably.)

   Thus draw u_{iam} from the conditional distribution function of U_{ia} given by

$$\frac{\{F(u \mid z_{ia}, \hat{\gamma}_m) - F(a \mid z_{ia}, \hat{\gamma}_m)\}}{\{F(U^* + a \mid z_{ia}, \hat{\gamma}_m) - F(a \mid z_{ia}, \hat{\gamma}_m)\}}.$$

3. Estimate the elapsed seropositive time prior to the first HIV+ visit for the ith individual by the median of $\{(u_{iam} - a), a = 0, 1, \ldots, A_i\}$. Denote the integer part of this median by $\hat{V}_{im}$, which is an estimate of V_i. Then we can assign discrete times $\hat{V}_{im}, \hat{V}_{im} + 1, \ldots, \hat{V}_{im} + A_i$ to this individual's $A_i + 1$ visits. Note that the first

$\hat{V}_{im} - 1$ "observations" are regarded as missing for this individual. However, this only causes the corresponding contributions to the likelihood function (4) to be omitted, as noted before.

4. Repeat steps 2 and 3 for all $i \in I, P$ and combine imputed times with the observed data, thus obtaining one "complete" data set.
5. Repeat step 4 for all $\hat{\gamma}_m(m = 1, 2, \ldots, M)$ thus obtaining M "complete" data sets.

Let ψ be a vector of parameters to be estimated. For example, ψ may be the vector of survival probabilities $\{S_D(k), k = 1, 2, \ldots, K\}$. Let $\hat{\psi}_m$ be an estimate of ψ obtained from the mth "complete" data set, and let $\widehat{\mathrm{Cov}}(\hat{\psi}_m)$ be the corresponding estimated "within" covariance matrix. An overall estimate of ψ may be taken to be $\hat{\psi} = (1/M) \Sigma_{m=1}^{M} \hat{\psi}_m$. Rubin and Schenker (9) propose the following as an estimate of the total covariance matrix of $\hat{\psi}$ (which accounts for both the "within" and "between" variability of the $\hat{\psi}_m$'s):

$$\widehat{\mathrm{Cov}}(\hat{\psi}) = \frac{1}{M} \sum_{m=1}^{M} \widehat{\mathrm{Cov}}\,(\hat{\psi}_m) + \left(\frac{M+1}{M}\right)\left(\frac{1}{M-1}\right) \sum_{m=1}^{M} (\hat{\psi}_m - \hat{\psi})(\hat{\psi}_m - \hat{\psi})'. \tag{7}$$

We used the grouped jackknife method to calculate the estimate $\hat{\psi}_m$ (where ψ is the vector of survival probabilities $S_D(k)$, $k = 1, \ldots, K$) and its estimated "within" covariance matrix $\widehat{\mathrm{Cov}}(\hat{\psi}_m)$, $m = 1, \ldots, M$. The method of calculation of the jackknife estimate of the survival probability vector for the population is described in the next section.

Results

Imputation of Seroconversion Dates

The estimation of model (3) required imputing of the seroconversion dates for incident and prevalent cohort members. The covariates and the functional form of the distribution of elapsed time from seroconversion, $F(u \mid z_{ik}, \gamma)$, to be used in the imputation procedure were selected using the incident cohort data. CART regression trees were used to do preliminary exploratory screening of the data set of 313 observations on 65 variables contributed by 62 seroconverters to identify covariates related to the elapsed time from the first HIV+ visit. A parametric regression model was then fitted with the suggested CART candidates using the ML methods for interval-censored data (23). Basophils and CD4% had statistically significant nonzero coefficients and were incorporated into $F(\cdot)$. Four parametric forms for the distribution $F(\cdot)$ were evaluated using the likelihood function (6), the gamma, the log logistic, the log normal, and the Weibull. The log logistic distribution was selected based on its simplicity and its goodness of fit. The goodness of fit was assessed by the linearity of a plot of the generalized residuals versus the expected unit exponential order statistics (24). The plot from the logistic model was the most linear based on the correlation coefficient $r = .90$ and visual examination. The log logistic distribution is given by

$$F(u|z_{ik}, \gamma) = \frac{(\lambda u)^{\gamma_0}}{1 + (\lambda u)^{\gamma_0}}, \tag{8}$$

where $\lambda = \exp\,[-\{\gamma_1 + \gamma_2(\text{CD4\%}_{ik}) + \gamma_3\,(\text{basophil}_{ik})\}]$.

The estimates of the log logistic model parameters and their covariance matrix were as follows:

$$\hat{\gamma} = \begin{bmatrix} \hat{\gamma}_0 \\ \hat{\gamma}_1 \\ \hat{\gamma}_2 \\ \hat{\gamma}_3 \end{bmatrix} = \begin{bmatrix} 1.773 \\ 3.711 \\ -0.060 \\ 1.186 \end{bmatrix},$$

and

$$\widehat{\text{Cov}}(\hat{\gamma}) = \begin{bmatrix} .0745 & -.0719 & .0020 & -.0670 \\ -.0719 & .1485 & -.0034 & .1981 \\ .0020 & -.0034 & .0003 & -.0005 \\ -.0670 & .1981 & -.0005 & .0572 \end{bmatrix} \tag{9}$$

Values of γ were randomly drawn from the MVN($\hat{\gamma}$, $\widehat{\text{Cov}}(\hat{\gamma})$) distribution and used in (8) to impute 100 seroconversion dates for each individual from the incident and prevalent cohorts (as described in the imputation algorithm given in the preceding section) resulting in 100 complete datasets. Four years of information was available from the seroconverter cohort to estimate γ; the imputed dates for prevalent cohort members were sampled as far back as 10 years prior to enrollment. This extrapolation seems unavoidable given that the data are available only from the early part of the epidemic; however, data from the latter part of the epidemic can be used to check the reasonableness of this imputation as they become available over time.

Estimated DHR Model

The covariate terms to be incorporated into the DHR model with a CLL link function were identified by a two-stage screening procedure. In the first stage, classification trees were used to simultaneously screen a dataset of 2927 observations on 73 risk factor and marker variables (including 6- and 12-month lagged values of CD4, CD4%, CD8, and CD8%) contributed by 450 prevalent and 62 seroconverter cohort members to identify those variables and two-factor interactions most strongly related to AIDS onset. Classification trees using equal and unequal priors were grown from a learning set of 2178

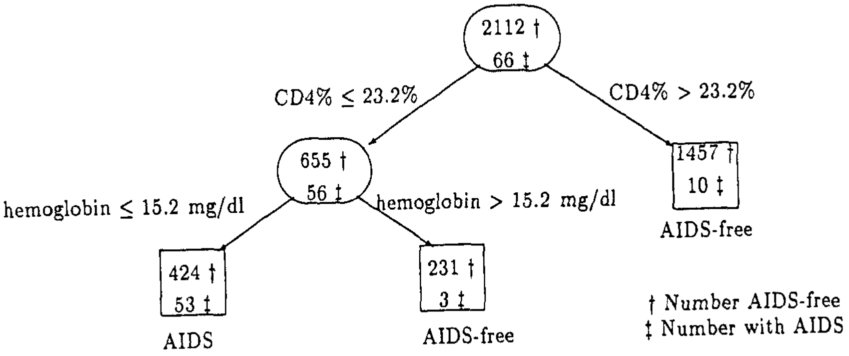

Figure 2. CART classification tree using equal priors (learning set tree: $\pi_1 = .5$ and $\pi_2 = .5$).

observations and pruned using a test set of the remaining 749 observations. The classification tree grown with equal priors, which tends to equalize the misclassification rates (11) for AIDS and AIDS-free cases, focused on variables that would identify AIDS cases. This tree, shown in Figure 2, indicates that AIDS cases are associated with low levels of CD4% and low hemoglobin. Among the AIDS learning set observations, 80% (53/66) had CD4% ≤ 23.2% and hemoglobin ≤ 15.2 mg/dl. The classification tree grown with unequal priors (reflecting the proportion of AIDS and AIDS-free seropositive individuals over the follow-up time) focused on variables that would identify AIDS-free cases. This tree, shown in Figure 3, indicates that AIDS-free cases are generally associated with a negative p24 antigen test; however, if p24 antigen is present, then the CD4% values become informative for identifying

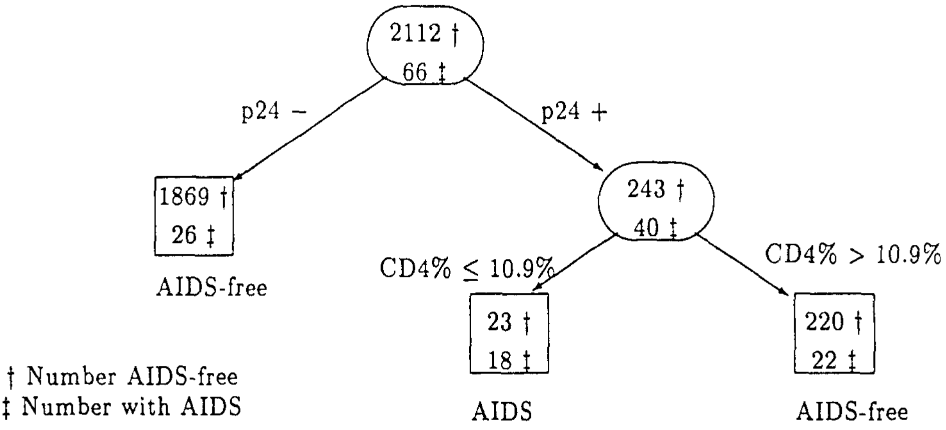

Figure 3. CART classification tree using unequal priors (learning set tree: $\pi_1 = .8$ and $\pi_2 = .2$).

AIDS-free cases. Among the AIDS-free learning set observations, 88% (1869/2112) had negative p24 antigen results. An additional 10% of the AIDS-free observations (220/2112) with positive p24 antigen results had CD4% > 10.9%. In the second stage, these covariate terms together with some additional terms (strong surrogate or competitive variables suggested by the CART analysis) were further screened by fitting the DHR model to a dataset where the estimated seroconversion date for each individual was the mean of 100 imputed dates. All candidate covariates were entered into the model; the final set of covariates were those associated with a statistically significant change in deviance (13), namely, CD4%, hemoglobin, p24 antigen, and a CD4% × p24 antigen interaction.

CART was the primary tool used to screen and select covariates. The secondary screening, which used imputed data, only confirmed the covariates selected by CART. Thus multiple imputation did not have any effect on the final selection of covariates.

The MLEs of the parameter vector from the DHR model with a cubic approximation (5) of the β_{0k} term and time-invariant covariate coefficients were obtained from each of the 100 complete datasets and combined into an overall estimated parameter vector and its estimated covariance matrix, as described in the preceding section. The signs of the estimated α_1, α_2, and α_3 coefficients in that approximation were found to flip back and forth across the 100 datasets, which resulted in highly unstable and nonsignificant coefficient estimates. In contrast, the α_0 term was consistently stable in sign and significant; the coefficients of the covariate terms were also consistently stable in magnitude and sign and were always significant. This suggested that the β_{0k} term be modeled as a constant, $\beta_{0k} \equiv \beta_k$. The modeling of β_{0k} as a constant does not imply a constant hazard rate, however; the DHR is a function of time-dependent covariates and hence is not constant. But now all the parameters are independent of time. This estimated DHR model is:

$$\begin{aligned}\ln[-\ln(1-\hat{\theta}_k)] = {} & -5.283 - 0.154[\text{CD4\%}_k - 27] \\ & - 0.294[\text{hemoglobin}_k - 15] + 2.020[\text{p24}_k] \\ & + 0.055[\text{p24}_k][\text{CD4\%}_k - 27]. \end{aligned} \qquad (10)$$

Standard errors of the estimated parameters are shown in Table 2.

Distribution of AIDS-Free Time After Seroconversion

The discrete survival function of AIDS-free time, $S_D(k)$, given by (2), was estimated using

$$\hat{S}_D(k) = \prod_{r=1}^{k} (1 - \hat{\theta}_r),$$

where the $\hat{\theta}_r$ were evaluated by substituting in (3) the representative popu-

Table 2. Parameters Estimates for Complementary Log-Log Model

Term	$\hat{\beta}$	SE($\hat{\beta}$)
Constant	−5.283	0.3517
CD4%	−0.154	0.0231
Hemoglobin	−0.294	0.0834
p24 antigen	2.020	0.4783
CD4% × p24	0.055	0.0304

lation values at time period r of covariates CD4%, hemoglobin, and p24 antigen.

We now explain in more detail the above estimation procedure. The seropositive individuals were divided into an early AIDS group with average incubation times under 4 years and a late AIDS group with average incubation times greater than 4 years (which includes people with right-censored events) based on imputed seroconversion dates. This was done because different immunological characteristics are associated with early and late AIDS individuals (14) and the covariate values for the two groups are quite different. The discrete survival function was estimated separately for each group using median covariate values (or the mean in the case of p24 antigen, a binary variable) obtained from each complete dataset at each time period or by extrapolating from a fitted trend when data were sparse (less than five observations). Seropositive individuals in the study group represent the prevalent and incident cohorts, but exclude the "unseen cohort" of those exposed who developed AIDS prior to enrollment (18,19). The contribution of this unseen cohort was estimated by using incident cases who developed AIDS during the 6 years following enrollment (obtained from additional records through 1989) to represent "lost" cases from 1978 to the start of the study. With this adjustment, there were 42 early AIDS individuals and 499 late AIDS individuals (57 AIDS and 442 right-censored cases). The early AIDS group exerts influence on the early part of the survival curve, while the changes in the survivorship of the late AIDS group influence the latter part of the survival curve (after 4 years).

A grouped jackknife estimate of the weighted average of $S_D(k)$ from the early and late AIDS groups and its estimated standard error were obtained from each of the 100 complete datasets and then combined into an overall estimate, $\hat{S}_D(k)$, and its standard error, SE[$\hat{S}_D(k)$], respectively. The resulting estimates with bars representing pointwise approximate 95% confidence intervals ($\hat{S}_D(k) \pm 1.96$SE [$\hat{S}_D(k)$] with lower limits truncated to zero) are shown in Figure 4. The estimated percentages of cases remaining AIDS-free at 4, 8, 12, and 15 years following seroconversion can be read from this distri-

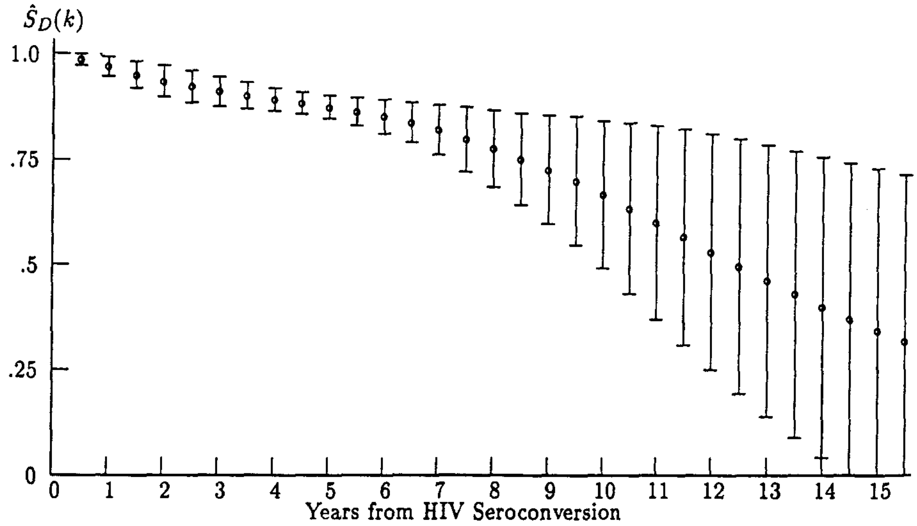

Figure 4. AIDS-free time distribution estimate and confidence intervals ($\hat{S}_D(k)$ ± 1.96 SE[$\hat{S}_D(k)$]).

bution to be 89.4%, 77.9%, 53.2%, and 34.6%, respectively. The median AIDS-free time obtained from this distribution is 12.5 years following seroconversion.

Discussion

In the fitted DHR model of (3) we have identified the key laboratory variables that are useful for tracking the progression to AIDS of HIV-infected individuals. The changes in the DHR over time are modeled through the time-dependent covariate values of CD4%, hemoglobin, and the presence or absence of the p24 antigen. The positive p24 antigen coefficient indicates that its presence is also indicative of AIDS. The interaction of CD4% with p24 antigen modifies the effect of CD4%; the presence of p24 antigen effectively makes the CD4% coefficient less negative owing to its interaction with a binary variable. Thus the presence of p24 antigen diminishes the importance of CD4% values for predicting AIDS. Other studies (2,25–29) have shown the separate predictive relationships of CD4% (or CD4) and p24 to AIDS. Our present study confirms the role of those marker variables when accompanied by hemoglobin; in addition, it shows that they have a joint predictive relationship to AIDS. What makes the present analysis striking is that these three marker variables were selected from a group of 73 clinical, behavioral, and laboratory variables. The strength of this model lies in the size of the dataset analyzed and the wide variety of variables that were evaluated as possible predictors of AIDS. It is also notable that the model is memoryless in the sense that θ_k depends only on the x_k values, but not on earlier covariate values (i.e., the current value of CD4% rather than the rate of change in CD4% affects the probability of AIDS onset). To a large extent, this Markovian nature is due to the elimination of the lagged covariates in the screening process.

Information from both the incident and prevalent cohorts was used to estimate the AIDS-free time distribution. It is well recognized that among prevalent cohorts the follow-up times associated with unknown infection times and undersampling of short incubation periods can lead to biased estimates of the incubation distribution (4,10). The imputation of 100 seroconversion times for each prevalent cohort member addresses the first issue. The goal of multiple imputation is to provide more precise estimates of the AIDS-free time distribution and assess the uncertainty in the parameter estimates contributed by the imputed times. Undersampling of short incubation times can be addressed by estimating the contribution of the unseen cohort (18). However, since the covariate information from the associated unseen visits was not known, this contribution was estimated in an *ad hoc* manner by using incident cohort cases who developed AIDS from 1984 to 1989 to represent the unseen cohort from 1978 to the start of the study.

It is evident from Figure 4 that the estimated confidence intervals become wider with time; this reflects the growing variance of the estimated $S_D(k)$ with increase in k as a result of an increasing number of terms in the product $\Pi_{r=1}^{k}(1 - \hat{\theta}_r)$. The increase in variance over time also reflects the uncertainty in estimates due to the smaller number of observations available at longer AIDS-free times. The median from this estimated distribution of AIDS-free time for homosexual men without treatment intervention should be interpreted cautiously owing to statistical uncertainty reflected by the wide confidence intervals of the survival function and uncertainty contributed by projected covariate values associated with time periods with sparse data. The estimated median AIDS-free time of 12.5 years obtained from this model-based approach is slightly longer than the 10.7 years estimated by Muñoz *et al.* (1) and the 9.5 years estimated by Taylor *et al.* (3), both of whom employed Kaplan-Meier-type analyses using natural history data from the entire MACS.

References

1. Muñoz A, Wang MC, Bass S, Taylor JMG, Kingsley LA, Chmiel JS, Polk BF, for the Multicenter AIDS Cohort Study: AIDS-free time after HIV-1 seroconversion in homosexual men. *Am J Epidemiol,* 130:530–549, 1989.

2. Muñoz A, Carey V, Taylor JMG, Chmiel JS, Kingsley, L, Van Raden M, Hoover DR, Polk BF: Estimation of time since exposure for a prevalent cohort. *Statist Med* 11:939–962, 1992.
3. Taylor JMG, Muñoz A, Bass SM, Saah AJ, Chmiel JS, Kingsley LA, and the Multicenter AIDS Cohort Study: Estimating the distribution of times from HIV seroconversion to AIDS using multiple imputation. *Statist Med* 9:505–514, 1990.
4. Jewell NP: Some statistical issues in studies of the epidemiology of AIDS. *Statist Med* 9:1387–1416, 1990.
5. Bacchetti P: Estimating the incubation period of AIDS by comparing population infection and diagnosis patterns. *J Am Statist Assoc* 85:1002–1008, 1990.
6. Kaslow, RA, Ostrow DG, Detels R, Phair JP, Polk BF, Rinaldo DR: The Multicenter AIDS Cohort Study: Rationale, organization, and selection of participants. *Am J Epidemiol* 154:310–318, 1987.
7. Kaslow RA, Francis DP: *The Epidemiology of AIDS*. Oxford University Press, New York, 1989.
8. Little RJA, Rubin DB: *Statistical Analysis with Missing Data*. Wiley, New York, 1987.
9. Rubin DB, Schenker N: Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *J Am Statist Assoc* 81:366–374, 1986.
10. Brookmeyer R, Gail MH, Polk BF: The prevalent cohort study and the acquired immunodeficiency syndrome. *Am J Epidemiol* 126:14–24, 1987.
11. Breiman JH, Friedman RA, Olshen CJ, Stone CJ: *Classification and Regression Trees*. Wadsworth International Group, Belmont, CA, 1984.
12. Kalbfleisch JR, Prentice RL: *The Statistical Analysis of Failure Time Data*. Wiley, New York, 1980.
13. McCullagh P, Nelder FA: *Generalized Linear Models*. Chapman and Hall, New York, 1983.
14. Phair J, Jacobson L, Detels R, Rinaldo C, Saah A, Schrager L, Muñoz A: Acquired immune deficiency syndrome occurring within 5 years of infection with human immunodeficiency virus type-1: The Multicenter AIDS Cohort Study. *J AIDS* 5:490–496, 1992.
15. Efron B: The jackknife, the bootstrap, and other resampling plans. Technical Report 163, Department of Statistics, Stanford University, Stanford, CA, 1980.
16. Breslow NE, Day NE: *Statistical Methods in Cancer Research. Vol. 2: The Design and Analysis of Cohort Studies*. Oxford University Press, New York, 1987.
17. Dunlop DD: Developing risk factor and disease progression marker models for longitudinal cohort data. Ph.D. dissertation, Northwestern University, Evanston, IL, 1990.
18. Hoover RD, Muñoz A, Taylor JMG, Chmiel JS, Odaka N, Armstrong J, for the Multicenter AIDS Cohort Study: The unseen sample in cohort studies; estimation of its size and effect. *Statist Med* 10:1993–2003, 1991.
19. Hoover RD, Muñoz A, Carey V, Chmiel JS, Taylor JMG: Estimating the 1978–1990 and future spread of human immunodeficiency virus type 1 in subgroups of homosexual men. *Am J Epidemiol* 134:1190–1205, 1991.
20. Prentice RL, Gloeckler LA: Regression analysis of group survival data with application to breast cancer data. *Biometrics* 34:57–67, 1978.
21. GLIM Rel 3.77, Numerical Algorithm Group Inc, Downers Grove, IL, 1987.
22. Liang K-Y, Zeger SL: Longitudinal data analysis using generalized linear models. *Biometrics* 73:13–22, 1986.
23. Finkelstein DM: A proportional hazards model for interval censored failure time data. *Biometrics* 42:845–854, 1986.

24. Kay R: Proportional hazard regression models and the analysis of censored data. *Appl Statist* 26:227–237, 1977.

25. Polk BF, Fox F, Brookmeyer R, Kanchanaraska S, Kaslow R, Visscher B, Rinaldo C, Phair J: Predictors of acquired immunodeficiency syndrome developing in a cohort of seropositive homosexual men. *N Engl J Med* 316:61–66, 1987.

26. Moss AR, Bacchetti P, Osmond D, Kramph W, Chaisson RE, Stites D, Wilber J, Allain JP, Carlson J: Seropositivity for HIV and the development of AIDS or AIDS related condition: Three year followup of the San Francisco General Hospital cohort. *Br Med J* 296:745–750, 1988.

27. MacDonell KB, Chmiel JS, Poggensee L, Wu S, Phair JP: Predicting progression to AIDS: Combined usefulness of CD4 lymphocyte counts and p24 antigen. *Am J Med* 89:706–712, 1990.

28. DeGruttola V, Lange N, Dafni U: Modeling the progression of HIV infection. *J Am Statist Assoc* 86:569–577, 1991.

29. Bacchetti P, Jewell NP: Nonparametric estimation of the incubation period of AIDS based on a prevalent cohort with unknown infection times. *Biometrics* 47:947–960, 1991.