American Journal of Epidemiology
Vol. 154, No. 4
Printed in U.S.A.

Analysis of Dynamic Cohort Data

John M. Williamson,[1] Glen A. Satten,[1,4] Jeffrey A. Hanson,[2] Hillard Weinstock,[1] and Somnath Datta[3]

Left-truncated and interval-censored data, termed *dynamic cohort* data, arise in longitudinal studies with rolling admissions and only occasional follow-up. The authors compared four approaches for analyzing such data: a constant hazard model; maximum likelihood estimation with flexible parametric models; the midpoint method, in which the midpoint of the last negative and first positive test result is used in a Cox proportional hazards model that accounts for left truncation; and a semiparametric method that uses imputed failure times in the Cox model. By using a simulation study, they assessed the performance of these approaches under conditions that can arise in observational studies: changes in disease incidence and changes in the underlying population. The simulation results indicated that the constant hazard model and midpoint method were inadequate and that the flexible parametric model was useful when enough parameters were used in modeling the baseline hazard. The semiparametric method ensured correct parameter (odds ratio) estimation when the baseline hazard was misspecified, but the trade-off increased computational complexity. In this paper, a study of the incidence of human immunodeficiency virus in patients repeatedly tested for the virus at a sexually transmitted disease clinic in New Orleans, Louisiana, illustrates the methods used. *Am J Epidemiol* 2001;154: 366–72.

Cox regression; incidence; interval censoring; survival analysis; truncation

Wide-scale monitoring of the human immunodeficiency virus (HIV) epidemic has involved two components: 1) reporting of persons with a recent diagnosis of acquired immunodeficiency syndrome (AIDS) and 2) large HIV prevalence surveys. However, at a time when the life expectancy of persons with HIV is increasing because of effective antiretroviral treatment (1), the monitoring of AIDS incidence and HIV prevalence tells less and less about current patterns of HIV transmission. Most US states have adopted some form of reporting of HIV cases (i.e., persons who have tested positive for HIV but do not have AIDS). Although HIV reporting data help to assess health care needs by reflecting the numbers of persons who know their infection status, the unknown duration between infection and a person's first positive HIV test result makes HIV reporting data a difficult tool for monitoring HIV incidence. Questions such as which age groups are at the highest risk of acquiring HIV infection can be answered only by conducting incidence studies. Thus, direct measurements of HIV incidence and determination of risk factors associated with recent HIV infection are now critically important for monitoring current patterns of HIV transmission and implementing successful HIV prevention programs in high-risk populations.

Direct measurement of HIV incidence (and risk factors associated with recent acquisition of HIV infection) has traditionally required a cohort study. However, start-up and follow-up costs, aging (gradually depleting the number of persons at high risk), and ethical requirements to provide HIV prevention counseling make cohort studies difficult to use as surveillance tools. Some recent cross-sectional approaches to the measurement of HIV incidence (2–4) show promise but require either blood specimens and special laboratory techniques (2) or a combination of high incidence and a large sample (3, 4).

A recent approach to making inferences about current HIV incidence is to use longitudinal samples of convenience, such as repeat attendees at a sexually transmitted disease (STD) clinic. By using historical data (5, 6) or stored specimens (7), some studies can be carried out without any additional follow-up. In this paper, we compare the ways in which the data on longitudinal convenience samples (8–12) can be analyzed. Typically, these data present several statistical challenges. Unlike a clinical trial or intervention study, the appropriate time scale for the risk of HIV acquisition is calendar time, not time in the study. Study enrollment dates (i.e., the date of the first record for a study participant) are not concurrent for all participants. In addition, the time of

Received for publication May 5, 2000, and accepted for publication January 31, 2001.

Abbreviations: AIDS, acquired immunodeficiency syndrome; HIV, human immunodeficiency virus; STD, sexually transmitted disease.

[1] Division of HIV/AIDS Prevention: Surveillance and Epidemiology, National Center for HIV, STD and TB Prevention, Centers for Disease Control and Prevention, Atlanta, GA.

[2] HIV/AIDS Program, Louisiana Office of Public Health, New Orleans, LA.

[3] Department of Statistics, University of Georgia, Athens, GA.

[4] Present affiliation: National Center for Environmental Health, Centers for Disease Control and Prevention, Atlanta, GA.

Correspondence to Dr. John M. Williamson, Division of HIV/AIDS Prevention: Behavioral Intervention Research Branch, MS E-37, Centers for Disease Control and Prevention, 1600 Clifton Road, NE, Atlanta, GA 30333 (e-mail: jow5@cdc.gov).

HIV seroconversion (i.e., the time when detectable antibody to HIV develops, used as a proxy for time of HIV infection) is not measured exactly, but it is known to lie in the interval defined by the last negative and first positive test result (i.e., the data on seroconversion time are interval censored). Another feature of these data is that some persons are HIV positive the first time they are seen at a clinic. Data from these persons are not used, because modeling the incidence of HIV prior to the beginning of the study period is problematic. Excluding persons with prevalent HIV infection results in left-truncated data (13). Left-truncated and interval-censored survival data resulting from serial enrollment studies, in which the data are analyzed by calendar time rather than time in the study, are termed *dynamic cohort* data because the study population at risk for HIV seroconversion changes as some persons enter the study and others seroconvert (14). Finally, the demographics of the sample may change. For example, an STD clinic that initially serves a predominately White, gay clientele may see an increasing proportion of minority heterosexuals during the time the data are collected.

Several approaches are available for analyzing dynamic cohort studies. The simplest is to assume a constant incidence of HIV over time (i.e., the distribution of times to seroconversion is exponential). For a low incidence, maximum likelihood estimates assuming a constant hazard for seroconversion are equivalent to dividing the number of observed seroconverters by total person-time in the study. A second approach is to assume a more flexible parametric model for the hazard of becoming HIV infected (such as a piecewise linear model). In such a model, estimates of the parameters can be obtained by using the method of maximum likelihood. A third approach to fitting dynamic cohort data is to assign the date of seroconversion as the midpoint between the last negative and first positive test result and fit a Cox proportional hazards model that accounts for left truncation to the resulting data. The midpoint approach can result in severely biased parameter estimation and underestimation of the standard errors of the parameter estimates when used to analyze interval-censored data (15–17). A major question regarding these three approaches is what is the effect on regression parameter estimates (i.e., the effect of covariates on the hazard of seroconversion) if the hazard of becoming HIV infected changes during the study or if the demographics of the sample change? Datta et al. (14) attempted to minimize these biases by using a robust semiparametric procedure in which imputed times of seroconversion are used in a left-truncated Cox model. Although a parametric model for the hazard of becoming HIV infected is used for the imputation, robustness is achieved because the Cox model uses only the rank order of the imputed failure times.

MODELS FOR ANALYZING DYNAMIC COHORT DATA

Parametric models

Let t_{0i} denote the time of study enrollment for person i, $1 \leq i \leq N$, and take as time 0 the smallest of the t_{0i}'s (the time when the first person enrolls in the study). Persons were enrolled in the study only if they tested HIV negative at their first visit (t_{0i}). Those who were seropositive at their first visit were excluded from the analysis. We assumed that study participants were seen at random times independent of the time when they seroconverted. This assumption is not unreasonable, because the time between infection and seroconversion to HIV is fairly long (estimated median and mean of 45 and 65 days, respectively) (18). For persons who seroconverted, let $\delta_i = 1$ and let L_i and U_i denote the time of their last HIV-negative and first HIV-positive test results, respectively. For persons whose last HIV test result was negative, let L_i be the time of the last test and let $\delta_i = 0$. These conversion times are right censored (censoring time L_i). For each person, we also observed a vector $\boldsymbol{x_i}$ of covariates that modulated the hazard of becoming HIV infected.

A parametric model is easily specified by its hazard function. In the proportional hazards family, write $\lambda(t|\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{x_i}) = \lambda_0(t|\boldsymbol{\theta})e^{\boldsymbol{\beta}\boldsymbol{x_i}}$, where $\lambda_0(t|\boldsymbol{\theta})$ is the baseline hazard (a function of parameters $\boldsymbol{\theta}$) and $\boldsymbol{\beta}$ are the regression parameters of interest. Let $S(t) = Pr(T > t)$ denote the survival distribution for the failure time random variable T. For example, the random variable T denotes the time of seroconversion for a person, that is, the time from the beginning of the study ($t = 0$, the first enrollment) to seroconversion, not the time from entry into the study to seroconversion. According to Klein and Moeschberger (13), maximum likelihood estimates of the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ are obtained by maximizing the likelihood

$$L = \prod_{i=1}^{N} \left[\frac{S(L_i|\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{x_i}) - S(U_i|\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{x_i})}{S(t_{0i}|\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{x_i})} \right]^{\delta_i} \left[\frac{S(L_i|\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{x_i})}{S(t_{0i}|\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{x_i})} \right]^{1-\delta_i}. \quad (1)$$

The term $S(t_{0i}|\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{x_i})$ in the denominator of equation 1 accounts for left truncation because we are including only those persons initially seronegative (i.e., seroconvert after their study enrollment time t_{0i}). Therefore, the first term in the likelihood corresponds to the probability that a person with covariates $\boldsymbol{x_i}$ will seroconvert between L_i and U_i, and the second term corresponds to the probability that he or she will be right censored at L_i. Choosing $\lambda_0(t|\theta) = \theta$ results in an exponential distribution in which the hazard of seroconversion is independent of calendar time; choosing $\lambda_0(t|\boldsymbol{\theta}) = \theta_1 t^{\theta_2}$ results in a Weibull model for the distribution of times to seroconversion. Parametric models such as these often are too restrictive for modeling the baseline distribution. For instance, a Weibull distribution can accommodate only increasing or decreasing hazard functions.

Therefore, to allow greater flexibility in modeling $\lambda_0(t|\boldsymbol{\theta})$, we also considered a piecewise linear model. The hazard function was assumed to be linear and continuous between a series of change points. Here, the parameters $\boldsymbol{\theta}$ are combined to estimate the slopes of the piecewise linear hazard functions. In these models, the values of the change points are fixed and are not considered parameters. (Choosing the values of the change points is considered in the Simulations

section of this paper.) For details on the use of these models, refer to Ramsay (19), Rosenberg (20), or Herndon and Harrell (21). The parameters $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ are estimated simultaneously by maximizing equation 1, as with the constant hazard model. An example of a piecewise linear hazard model with five parameters (five change points) is shown in figure 1, which was used in an HIV incidence study conducted in New Orleans, Louisiana.

Midpoint method

When the exact failure times are known, the regression parameters $\boldsymbol{\beta}$ can be estimated by using the Cox proportional hazards model without specifying a model for $\lambda_0(t|\boldsymbol{\theta})$ (22). With dynamic cohort data, the failure times are known only to fall in the interval (L_i, U_i). Data analysts often substitute the midpoint of the failure interval, $(L_i + U_i)/2$, for the unknown failure time and use the left-truncated Cox model to analyze the resulting data. This approach may result in biased parameter estimation and underestimation of the parameter estimate standard errors when used to analyze nontruncated interval-censored data (16).

Semiparametric model

A recent semiparametric approach for analyzing dynamic cohort data involves imputing a seroconversion time between L_i and U_i, for those participants with $\delta_i = 1$, from a specified parametric baseline distribution (14). As with the midpoint method, the data are then analyzed by using the left-truncated Cox model. The estimation procedure is iterative. Multiple imputations of the missing seroconversion times are made on the basis of a current estimate of the baseline hazard, leading to a new estimate of $\boldsymbol{\beta}$. Then, this new estimate of $\boldsymbol{\beta}$ is used to obtain a new estimate of the baseline hazard function. These two steps are repeated until the variability in $\hat{\boldsymbol{\beta}}$, the estimator of $\boldsymbol{\beta}$, is small.

The semiparametric approach is similar to the Bayesian technique of multiple imputation, except that it is not necessary to specify a prior distribution, and an iterative approach must be taken because the imputation depends on the parameters we are estimating. Refer to Rubin (23) for more information regarding multiple imputation. For details on choosing the number of iterations to repeat this process and the number of imputations per iteration, refer to Satten et al. (24) and Datta et al. (14). Initial estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ can be obtained from the exponential model. Datta et al. also give an expression for the standard error of $\hat{\boldsymbol{\beta}}$ obtained with this procedure. When the observed censoring intervals are small enough that the intervals do not overlap, the midpoint method and semiparametric approach (unlike the parametric approaches described) reduce to the Cox proportional hazards model with left truncation (13).

SIMULATIONS

We conducted a simulation trial to compare the performance of these approaches for analyzing dynamic cohort data in a situation in which the correct answer was known. The data were generated to represent a fairly extreme case in which the hazard of seroconversion and the study demographics changed over time. The goals of the simulation were to compare the bias among the four methods and to assess the efficiency of the semiparametric method relative to the parametric approaches. For data that are interval censored but not left truncated, Satten et al. (24) showed that the semiparametric estimate of $\boldsymbol{\beta}$ is robust to the misspecification of $\lambda_0(t|\boldsymbol{\theta})$ and very efficient when $\lambda_0(t|\boldsymbol{\theta})$ is specified correctly. However, because the left-truncated version of the full data proportional hazards model uses a restricted risk set at each failure time (13), it is not clear that the efficiency performance of the semiparametric model will be retained with dynamic cohort data.

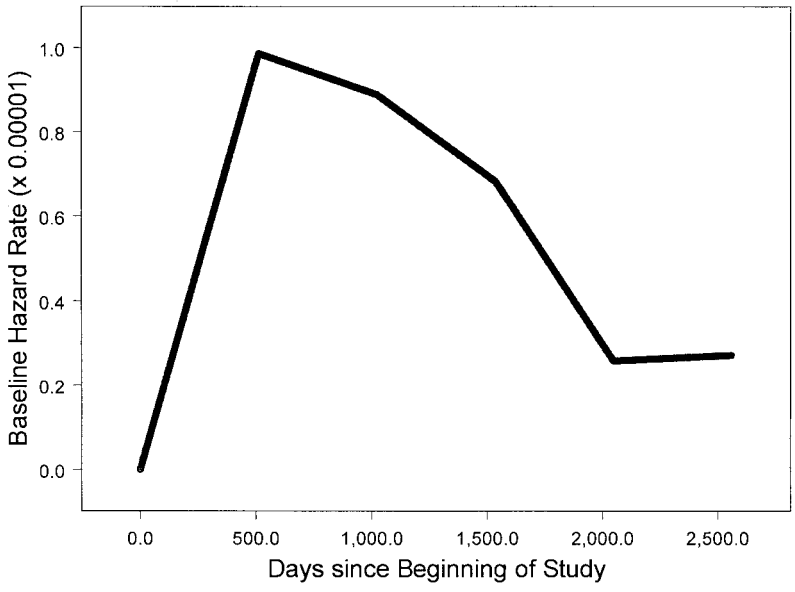

FIGURE 1. A piecewise linear hazard model with five parameters (five change points).

We generated 500 data sets, each containing 2,500 failure times, a binary covariate, and a left-truncation time. In each data set, 1,250 observations were generated with $x = 0$ and the other 1,250 observations with $x = 1$; we used $\beta = \ln(2) \approx 0.693$ corresponding to a hazard ratio for covariate values $x = 1$ to $x = 0$ of $\exp(\beta) = 2.0$. We chose $\lambda_0(t|\boldsymbol{\theta}) = \theta_1\theta_2 t^{\theta_1 - 1}/(1 + \theta_2 t^{\theta_1})$ corresponding to a log-logistic distribution for the times of HIV seroconversion, with a shape parameter of $\theta_1 = 4$ and a scale parameter of $\theta_2 = 0.01$. We chose this distribution because $\lambda_0(t|\boldsymbol{\theta})$ is nonmonotone, first rising from 0 and then falling asymptotically to 0.

To generate interval-censored data, for each observation we also generated potential visit times by assuming that times between visits for each person were independent and were identically exponentially distributed with mean 10. To allow for changes in the demographics of the sample, we assumed that each person was observed at his or her potential visit times only with probability $\phi(t|x_i) = 0.2 + 0.6\,x_i + (0.6\,t/200.0)(1.0 - 2.0\,x_i)$. We used the first observed visit time as the left-truncation time t_{0i}. It was assumed that the study closed at time 65, so all visit times after time 65 were discarded. We chose the last observed visit time before seroconversion as L_i. If a person had observed visit times after seroconversion, we chose U_i as the earliest of these times; otherwise, the observation was considered right censored at time L_i. Accordingly, $\phi(t)$ decreased from 0.8 to 0.605 for persons with covariate value $x = 1$ and increased from 0.2 to 0.395 for persons with covariate value $x = 0$ as t ranged from 0 to 65. Observations were rejected for which the left-truncation time was later than the failure time, and new pairs of failure and left-truncation times were generated until 1,250 observations were recorded for each covariate value. Approximately 90 percent of the observations were right censored, similar to the percentage for a low incidence study. On average, 943.1 observations with a covariate value of $x = 0$ and 82.4 observations with a covariate value of $x = 1$ were discarded because of left truncation (i.e., because seroconversion occurred before a person's first observed visit) for each data set generated.

For each data set, we estimated the parameter β and its standard error by using the constant hazard model, parametric models with $K = 1, \ldots, 5$ piecewise linear hazard functions (denoted parametric(K)), the midpoint method, and the semiparametric method using parametric(K) ($K = 1, \ldots, 5$) as the baseline hazard models. The SAS procedure LIFEREG was used to fit the constant hazard models (25). The left-truncation (study entry) time was subtracted from each observation time (L_i and U_i), and the resulting data were analyzed by assuming an underlying exponential distribution. The SAS procedure PHREG was used to fit the midpoint method models (26). The value $(L_i + U_i)/2$ was used for the failure time for persons with $\delta_i = 1$, and no changes were made for right-censored subjects. The resulting data were analyzed by using the left-truncated version of the Cox model (13). The parametric piecewise linear models and the semiparametric models were fit with Fortran programs (available on request from the first author). For the piecewise linear models, we chose equally spaced change points. For example, the change points for the parametric(5) model were 0, 13, 26, 39, 52, and 65. Note that the model parametric(1) does not have change points but varies linearly between times 0 and 65.

The results of our simulations are summarized in table 1. For each method, table 1 gives a statistic (z) for testing the hypothesis that the mean value of $\hat{\beta}$ is equal to the true value $\log(2) \approx 0.693$. Although the value of z depends on the number of simulations conducted, it is useful in ordering the methods with regard to bias, because the standard errors of $\hat{\beta}$ for the various methods were all similar. For the constant hazard model, the average estimate of β for all 500 data sets was 0.509, rather different from the true value of 0.693 used

TABLE 1. Simulation results

Method	Model for $\lambda_0(t\,\vert\,\theta)$	β*	Var† $(\beta)^{0.5}$	Emp SE‡ (β)	z §
Maximum likelihood	Constant	0.509	0.134	0.143	–28.81
Maximum likelihood	Parametric(1)	0.611	0.133	0.143	–12.76
Maximum likelihood	Parametric(2)	0.698	0.135	0.147	0.68
Maximum likelihood	Parametric(3)	0.712	0.136	0.148	2.79
Maximum likelihood	Parametric(4)	0.703	0.135	0.147	1.51
Maximum likelihood	Parametric(5)	0.701	0.135	0.146	1.22
Midpoint¶	None	0.671	0.109	0.146	–3.35
Semiparametric	Parametric(1)	0.680	0.134	0.147	–2.00
Semiparametric	Parametric(2)	0.700	0.135	0.146	1.02
Semiparametric	Parametric(3)	0.701	0.135	0.146	1.14
Semiparametric	Parametric(4)	0.701	0.135	0.146	1.18
Semiparametric	Parametric(5)	0.701	0.135	0.146	1.18

* Average of 500 simulated data sets. True value of β is log(2) ≈ 0.693.

† Square root of the average estimated variance of $\hat{\beta}$ of the 500 simulated data sets.

‡ Empirical standard error of $\hat{\beta}$ $((\sum_{l=1}^{500}(\hat{\beta}_l - \hat{\beta}_\cdot)^2/499)^{1/2}$, where $\hat{\beta}_\cdot = \sum_{l=1}^{500}\hat{\beta}_l/500)$.

§ z statistic resulting from testing $\beta = 0$, by using the empirical standard of $\hat{\beta}$.

¶ Method of assigning the date of seroconversion as the midpoint between the last negative and first positive test result and using the Cox proportional hazards model with left truncation.

to generate the data, as indicated by the z-statistic value of –28.81. For the parametric piecewise linear models, the average $\hat{\beta}$ varied from 0.611 (one parameter) to 0.701 (five parameters). Adding parameters to the parametric piecewise linear model improved the estimation of β as measured by a decrease in the z statistic (12.76 to 1.22). Although parameter estimation was unbiased with the two-parameter model (z-statistic value of 0.68), $\hat{\beta}$ and z did not change steadily with increasing parameters, and the value of $\hat{\beta}$ did not plateau until five parameters were used to specify the baseline (average $\hat{\beta}$ = 0.703 using parametric(4) and average $\hat{\beta}$= 0.701 using parametric(5)).

The simulation results also indicated that the midpoint method results in biased parameter estimation (z-statistic value of –3.35) and that the standard error of the parameter estimate is grossly underestimated in comparison to the empirical standard error of the parameter estimate. For the semiparametric approach, we imputed 50 data sets for each estimate of the baseline hazard function, and we updated the baseline hazard function 400 times (for a total of 20,000 imputed data sets). The average values of $\hat{\beta}$ for the semiparametric method, in which the model for $\lambda_0(t|\boldsymbol{\theta})$ was the same as in parametric(l) and parametric(5), were 0.680 and 0.701, respectively. The semiparametric method was a noticeable improvement over the parametric method for estimating β when the baseline distribution was misspecified, as is the case with parametric(l) (z-statistic values of –2.00 for the semiparametric method vs. –12.76 for the parametric method). However, when the baseline hazard was closely modeled, as with parametric(5), the semiparametric method offered no improvement over the parametric approach. However, in comparison with any of the parametric approaches, there was no noticeable loss of efficiency when the semiparametric method was used.

DYNAMIC COHORT OF PATIENTS ATTENDING AN STD CLINIC

A study was conducted to assess the incidence of HIV infection and the risk factors associated with HIV seroconversion among patients repeatedly tested for HIV at an STD clinic in New Orleans (5). Although this sample of repeat attendees at an STD clinic is not representative of the US population, it is an important group to monitor because of its high HIV incidence. Medical records in databases and paper charts were reviewed. The clinic serves a predominately inner-city minority population, and persons attending the clinic are routinely offered HIV testing if they have not been tested within the preceding 3 months. We examined the records of all patients who initially tested HIV negative in 1991–1998 and who received at least one additional HIV test during the study period. Demographic data, risk behaviors, and clinical information, in addition to the patient's HIV test results, were obtained from the collected records. A total of 9,183 persons for whom we had complete data (no missing covariate information) were included in this analysis: 2,311 women, 6,506 heterosexual men, and 366 men who have sex with men. The seroconversion rate was the highest for the latter group (14 seroconverters/975.43 years of follow-up = 1.44 percent/year), followed by the rates for women (26 seroconverters/4,883.40 years of follow-up = 0.53 percent/year) and heterosexual men (72 seroconverters/15,924.19 years of follow-up = 0.45 percent/year).

We first analyzed these data by using the semiparametric model with parametric(5) as the baseline distribution; demographic and risk factor variables were examined until a final model that included all significant covariates was formulated. This final model was also fit with the semiparametric approach by using parametric(l) as the baseline distribution, the parametric approach by using parametric(l) and parametric(5), the constant hazard model, and the midpoint method. The various regression results are presented in table 2. Parameter estimates from the constant hazard model did not differ from those from the semiparametric model with five parameters by more than 3 percent, with the exception of the covariates *reactive nontreponemal test result* (4.9 percent) and *gonorrhea* (5.6 percent). However, for any of the seven covariates shown in this table, the parametric model with five parameters and the midpoint method produced parameter estimates that did not differ by more than 1 percent from the five-parameter semiparametric approach.

TABLE 2. Parameter estimates from a human immunodeficiency virus incidence study

Covariate	Constant hazard	Par(1)*	Par(5)*	Midpoint†	SP(1)‡	SP(5)‡
Reactive nontreponemal test result§,¶	1.632 (0.254)#	1.826 (0.260)	1.551 (0.253)	1.571 (0.253)	1.576 (0.264)	1.552 (0.262)
Exchange drugs/money for sex§	0.588 (0.214)	0.623 (0.215)	0.590 (0.214)	0.593 (0.214)	0.585 (0.222)	0.586 (0.222)
Genital ulcer disease§	1.023 (0.320)	1.114 (0.325)	0.998 (0.318)	1.010 (0.317)	1.003 (0.327)	0.999 (0.326)
Gonorrhea§	0.630 (0.206)	0.713 (0.205)	0.595 (0.207)	0.607 (0.207)	0.605 (0.209)	0.595 (0.208)
MSM§,**	3.426 (1.034)	3.241 (1.012)	3.432 (1.030)	3.398 (1.023)	3.435 (1.004)	3.437 (1.005)
Age (years) for non-MSMs††	0.001 (0.011)	–0.003 (0.011)	0.002 (0.011)	0.001 (0.012)	0.002 (0.012)	0.002 (0.012)
Age (years) for MSMs††	–0.083 (0.038)	–0.076 (0.037)	–0.083 (0.038)	–0.082 (0.038)	–0.083 (0.036)	–0.084 (0.036)

* Parametric piecewise linear models with one parameter and five parameters, respectively.
† Method of assigning the date of seroconversion as the midpoint between the last negative and first positive test result and using the Cox proportional hazards model with left truncation.
‡ Semiparametric models with a piecewise linear hazard function using one parameter and five parameters, respectively.
§ 1 = yes; 0 = no.
¶ Nontreponemal serologic test result for syphilis.
# Numbers in parentheses, standard error of the parameter estimate.
** MSM, men who have sex with men.
†† Age was calculated at the midpoint of the interval between the last negative and first positive test results for seroconverters and at the midpoint of the interval between the second-to-last and last visits for nonseroconverters.

These results indicate that an assumption of constant incidence of HIV over time may not be fully adequate for these data, although a parametric model with five parameters seemed to model the underlying survival distribution with enough accuracy to estimate parameters adequately. Parameter estimate standard errors were only slightly smaller with the parametric approaches than with the semiparametric approach. According to the regression results, persons with genital ulcer disease, gonorrhea, or a positive nontreponemal test result for syphilis are at higher risk for HIV infection than patients without these characteristics. Persons who have exchanged drugs or money for sex are also more likely to seroconvert than are those who have not. Men who have sex with men are less likely to seroconvert as they get older. A plot of the baseline hazard function from the five-parameter spline model is shown in figure 1 and indicates that the hazard rate is nonconstant (a p value of <0.001 for testing the equality of the five parameters). The baseline hazard is constrained at 0 when this model is used. We initially fit the hazard with an extra parameter representing the hazard at $t = 0$, but we dropped it from the model because the resulting estimate was 0.

DISCUSSION

In dynamic cohort studies, serial admission of persons who have not yet experienced the event of interest produces left-truncated data; subsequent periodic observation produces interval-censored data. We considered four approaches to analyzing dynamic cohort data: a constant hazard model (at low incidence, equivalent to estimating incidence by dividing the number of seroconverters by person-time of observation), a flexible parametric model, the midpoint method, and a semiparametric approach. In a simulation study, we considered how these approaches perform given the types of sampling biases that can occur in an observational study. These included changes in incidence over calendar time and changes in the underlying population. The constant hazard model and midpoint method gave inadequate results in the presence of these changes. The simulation results also indicated that valid parameter estimation can be achieved by using a piecewise linear model conditioning on the study enrollment date.

We recommend the use of the piecewise linear model for analyzing data of these types if the baseline distribution is not thought to vary greatly with time and if the baseline model is repeatedly fit with an increasing number of parameters (change points) until the estimate of $\hat{\beta}$ stabilizes. If $\hat{\beta}$ does not stabilize with an increasing number of parameters, then the semiparametric approach should be used to protect against misspecification of the baseline (protection that warrants the extra computational complexity). Akaike's Information Criterion can also be used to select the piecewise linear hazard model with the optimal number of parameters (27). For the New Orleans HIV incidence study, the four-parameter piecewise linear hazard model was chosen as optimal according to this criterion, and the resulting parameter estimates differed from those of the five-parameter piecewise linear model only in the fourth decimal place. The semiparametric approach can always be used to verify the parameter estimates resulting from the parametric piecewise linear models.

We compared these four approaches to analyzing dynamic cohort data by analyzing a study conducted to assess the incidence of HIV infection and the risk factors associated with HIV seroconversion (5). Although the absolute value of HIV incidence may be difficult to interpret, it is hoped that trends in HIV incidence and cofactors that affect the relative risk of acquiring HIV infection may be generalizable. Because HIV incidence information is so difficult to obtain and is critically important, it is worth exploring the usefulness of such information.

REFERENCES

1. Palella FJ Jr, Delaney KM, Moorman AC, et al. Declining morbidity and mortality among patients with advanced human immunodeficiency virus infection. HIV Outpatient Study Investigators. N Engl J Med 1998;338:853–60.
2. Janssen RS, Satten GA, Strainer SL, et al. New testing strategy to detect early HIV-1 infection for use in incidence estimates and for clinical and prevention purposes. JAMA 1998;280:42–8.
3. Brookmeyer R, Quinn TC. Estimation of current human immunodeficiency virus incidence rates from a cross-sectional survey using early diagnostic tests. Am J Epidemiol 1995;141:166–72.
4. Brookmeyer R, Quinn T, Shepherd M, et al. The AIDS epidemic in India: a new method for estimating current human immunodeficiency virus (HIV) incidence rates. Am J Epidemiol 1995; 142:709–13.
5. Weinstock H, Sweeney S, Satten GA, et al. HIV seroincidence and risk factors among patients repeatedly tested for HIV attending sexually transmitted disease clinics in the United States, 1991–1996. STD Clinic HIV Seroincidence Study Group. J Acquir Immune Defic Syndr Hum Retrovirol 1998; 19:506–12.
6. Murrill CS, Prevots DR, Miller MS, et al. Incidence of HIV among injection drug users entering drug treatment programs in four US cities. J Urban Health 2001;78:152–61.
7. Torian LV, Murrill CS, Makki HA, et al. High HIV incidence in nonwhite bisexual men making repeat visits to a New York City sexually transmitted disease clinic, 1994–1995: results of a blinded longitudinal survey. Presented at the 4th Conference on Retroviruses and Opportunistic Infections, Washington, DC, January 1977.
8. Kerndt PR, Weber M, Ford W, et al. Incidence among injection drug users enrolled in a Los Angeles methadone program. (Letter). JAMA 1995;273:1831–2.
9. Ford WL, Melia N, Weber MD, et al. HIV-1 seroprevalence rates and trends in select populations in Los Angeles County, 1988–1991. Los Angeles, CA: Los Angeles County Department of Health Services, 1992.
10. Longshore D, Anglin MD. HIV prevalence and incidence among injection drug users in Los Angeles. J Acquir Immune Defic Syndr 1994;7:738–9.
11. Siddiqui NS, Brown LS Jr, Meyer TJ, et al. Decline in HIV-1 seroprevalence and low seroconversion rates among injecting drug users at a methadone maintenance program in New York City. J Psychoactive Drugs 1991;25:245–50.
12. New York State Department of Health. AIDS in New York State. Albany, NY: New York State Department of Health, 1993:34.
13. Klein JP, Moeschberger ML. Survival analysis: techniques for censored and truncated data. New York, NY: Springer-Verlag, 1997.
14. Datta S, Satten GA, Williamson JM. Consistency and asymp-

totic normality of estimators in a proportional hazards model with interval censoring and left truncation. Ann Inst Stat Math 2000;52:160–72.
15. Odell PM, Anderson KM, D'Agostino RB. Maximum likelihood estimation for interval-censored data using a Weibull-based accelerated failure time model. Biometrics 1992;48:951–9.
16. Satten GA. Rank-based inference in the proportional hazards model for interval censored data. Biometrika 1996;83:355–70.
17. Law CG, Brookmeyer R. Effects of mid-point imputation on the analysis of doubly censored data. Stat Med 1992;11:1569–78.
18. Busch MP, Satten GA. Time course of viremia and antibody seroconversion following HIV exposure. Am J Med 1997;102 (suppl 5B):117–24.
19. Ramsay JO. Monotone regression splines in action. Stat Sci 1988;3:425–61.
20. Rosenberg PS. Hazard function estimation using B-splines. Biometrics 1995;51:874–87.
21. Herndon JE, Harrell FE. The restricted cubic spline as baseline hazard in the proportional hazards model with step function time-dependent covariables. Stat Med 1995;14:2119–29.
22. Cox DR. Regression models and life tables (with discussion). J R Stat Soc (B) 1972;34:187–220.
23. Rubin DB. Multiple imputation for nonresponse in surveys. New York, NY: John Wiley & Sons, 1987.
24. Satten GA, Datta S, Williamson JM. Inference based on imputed failure times for the proportional hazards model with interval-censored data. J Am Stat Assoc 1998;93:318–27.
25. SAS Institute, Inc. SAS/STAT user's guide, version 6, 4th ed. Cary, NC: SAS Institute, Inc, 1994.
26. SAS Institute, Inc. SAS/STAT software: changes and enhancements through release 6.12. Cary, NC: SAS Institute, Inc,1997.
27. Akaike H. Prediction and entropy. In: Atkinson AC, Feinberg SE, eds. A celebration of statistics: the ISI centenary volume. New York, NY: Springer-Verlag, 1985:1–24.