The AIDS Epidemic: Estimating Survival After AIDS Diagnosis From Surveillance Data

XIN MING TU, XIAO-LI MENG, and MARCELLO PAGANO*

Survival analysis based on reported acquired immune deficiency syndrome (AIDS) cases from the surveillance data maintained by the Centers for Disease Control (CDC) leads to severe bias because of a sizable fraction of unreported deaths. One approach to this problem is to condition the analysis on reported deaths, but this approach presents several difficulties. First, only the individuals who die within a chronologic time interval defined by the analysis may be observed. Second, this right-truncated sampling process is complicated by the delay in reporting death to the surveillance system, and as a result, only a proportion of those who die within the defined time interval are reported. Third, the time each death is reported to the surveillance system, which is essential for the construction of a joint likelihood of survival and reporting delay, may not be available. Fourth, deaths that occurred before 1984 all had unknown times of death. As a consequence, the direct implementation of the joint likelihood approach appears to be complicated and difficult. On the other hand, this data set is worth analyzing; it is the only data base that comes close to covering the entire AIDS population and thus provides invaluable information to assess the current and predict the future AIDS epidemic. In this article we apply several missing-data techniques to deal with these difficulties. In particular, we discuss how to estimate the delay and survival distributions separately, and then combine the two sources of information to make valid inferences for the survival distributions using multiple imputation. The EM algorithm is used to facilitate computations. The methodology developed here can also be applied to other studies that share similar data structures, such as data from local and other national surveillance systems. Analysis of the CDC surveillance data as of March 1991, for some high-risk groups considered, indicates that there has been a steady and notable increase in survival after an AIDS diagnosis over chronologic time, especially for those who had *Pneumocystis carinii* pneumonia (PCP) as one of the AIDS-defining diagnoses.

KEY WORDS: Acquired immune deficiency syndrome; Censoring; Expectation-maximization algorithm; Incomplete data; Missing data; Multiple imputation; Proportional hazards model; Supplemented-expectation-maximization algorithm; Truncation.

Information about survival after a diagnosis of acquired immune deficiency syndrome (AIDS) is important for clinical research, health services, and policy planning. Research findings from some follow-up studies, such as those conducted by the AIDS Clinical Trials Group (Volberding et al. 1990), are extremely important but may be more useful for medical and clinical research than for health care planning, because patients enrolled in these studies receive state-of-the-art health care and tend to have longer survival times (Pagano et al. 1992). For health departments, what is of vital concern are the estimates of mortality for the general AIDS population residing in the United States as well as those residing in the areas serviced by the local health departments, which provide an indispensable source of information for assessing current health care needs as well as for future planning. The national surveillance system set up by the Centers for Disease Control (CDC) (CDC 1991) and systems maintained by the local health departments provide natural and better data sources for these estimates, because they are collected from these populations.

Analysis of survival time based on the reported AIDS cases from the CDC surveillance system (and local surveillance systems as well) is complicated by the fact that not all deaths are reported; of the diagnosed AIDS cases reported to the surveillance system, a proportion will never have their death reported. Survival estimates will be biased upward if cases without reported death certificates are simply censored at the time defined by the analysis (see Section 4). It is impossible to remove this bias unless we have knowledge from sources other than the CDC surveillance data to identify this confounding group of individuals. In this article, however, we focus on statistical methods for survival analysis without the help of such external sources.

One way to avoid the confounding problem is to estimate survival using only the *reported deaths* from the surveillance system. These reported deaths, which can be viewed as realizations of a sampling process that is right-truncated by a chronologic time x^* defined by the analysis, constitute only a portion of the sample: those who were diagnosed with AIDS, died, and were reported to the surveillance system by the time x^*. On the one hand, given the time of an AIDS diagnosis, times of death, and reporting of death, a joint likelihood may be constructed for simultaneously estimating the survival and the reporting delay distributions. However, the direct implementation of such an approach for analyzing the CDC data is very complicated because the time of reporting of death is not available for each reported death. On the other hand, a simplified analysis without adjustment for the reporting delay will be biased toward those whose death reporting is more likely to be complete, such as those who were diagnosed and died earlier in the epidemic and those

* Xin Ming Tu is Assistant Professor in the Department of Mathematics and Statistics, University of Pittsburgh, PA 15260. Xiao-Li Meng is Assistant Professor in the Department of Statistics, University of Chicago, IL 60637. Marcello Pagano is Professor in the Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115. Xin Ming Tu was Research Fellow in the Department of Biostatistics, Harvard School of Public Health at the time this article was written. The authors thank V. DeGruttola and R. A. Royce for constructive comments and helpful discussions, C. C. Clogg and R. J. A. Little for editorial guidance, and especially two anonymous reviewers for very helpful comments that led to the great improvement in the presentation. Tu and Pagano's research was supported in part by grants from the National Institutes of Health NIAID RO1-AI28076, T32-AI07358, R29-AI28905 and NO1-AI-95030, and Meng's research in part by NSF grant DMS-92-04504 and by the University of Chicago/AMOCO fund.

who were residents in the geographic locations that had better health surveillance systems.

In this article we apply several missing-data techniques to develop methods for survival analysis of surveillance data that may suffer from these problems. We develop and illustrate the methods by focusing on the CDC AIDS surveillance data, but the proposed methodology can be applied to data from some other surveillance systems, particularly those maintained by the local health departments. In Section 1, we first describe a discrete-time regression model for survival analysis of surveillance data and then apply the EM algorithm (Dempster, Laird, and Rubin 1977) to fit a general discrete-time regression model to data that can be censored as well as truncated in arbitrary intervals under no reporting delay. In Section 2 we discuss how to estimate the reporting delay distributions and to incorporate this information into estimation and inference of survival distributions using multiple imputation (Rubin 1987a). In Section 3 we present and discuss results from the analysis of the CDC surveillance data as of March 1991 for several high-risk groups. In Section 4 we provide some further discussions on the analysis and comparison of the proposed methodology with the standard alternative.

1. SURVIVAL DISTRIBUTION IN THE ABSENCE OF REPORTING DELAY

1.1 A Discrete Proportional Hazards Model

The observed deaths in the surveillance data consist of those who were diagnosed with AIDS, died, and were reported by some chronological time x^* defined by the analysis. Even if there were no reporting delay, these observations would still constitute only a portion of the sample observable in the time interval $[0, x^*]$, where 0 is chosen to be the time of the earliest reported death(s) in the surveillance data base. For the CDC data, a further complication occurs when the death times were also missing for deaths that occurred before 1984. (These relatively few cases, which may be excluded without affecting the main conclusions, were included to illustrate how to handle censoring as well as truncation, a problem that may arise in other studies.) Inclusion of these individuals in the analysis results in left-censored observations. Nonparametric estimation with no covariates of such truncated and censored data may be accomplished using the methods proposed by Turnbull (1976), Wang, Jewell, and Tsai (1986), and Wang (1992). But our main interest is to study the effect of covariates on survival. Of particular interest is to see how this survival distribution is changing with the time of diagnosis, an objective that cannot be achieved using these methods.

In this section we discuss a regression model and an associated fitting algorithm for data that are censored as well as truncated in intervals or unions of disjoint intervals. Methods for right-truncated data alone have been developed in the analysis of AIDS incidence and latency distributions (Brookmeyer and Liao 1990; Harris 1990a; Kalbfleisch and Lawless 1989, 1991; Finkelstein, Moore, and Schoenfeld, in press).

Let X be the calendar time of an AIDS diagnosis and let S be the time from diagnosis to death. Let $F(s|\mathbf{z})$ and $f(s|\mathbf{z})$ denote the cumulative distribution function (cdf) and probability density function (pdf) of the random variable S, conditional on a covariate vector $\mathbf{z}$. Also let $F(s)$ and $f(s)$ be the baseline cdf and pdf of S obtained when covariates are set to zero. The density for an individual observed in the absence of reporting delay is expressed as

$$f(s|\mathbf{z}, X = x, X + S \le x^*) = \frac{f(s|\mathbf{z})}{F(x^* - x|\mathbf{z})}.$$

In general we can only model and estimate the conditional distribution $F(s|\mathbf{z})/F(s^*|\mathbf{z})$, where s^* is the longest observed survival time. For $F(s|\mathbf{z})$ to be fully identifiable for $s \le s^*$, it is necessary that $F(s^*|\mathbf{z})$ be known. Estimating the unconditional survival distribution as well as other unconditional quantities, such as the survival median, depends on the assumption $F(s^*|\mathbf{z}) = 1$, which is a reasonable approximation for our application (Sec. 4).

Let A_i denote the region in which the ith individual is censored and let B_i denote the truncation region associated with the ith individual. Note that A_i reduces to a single point if the ith individual is not censored, and in our context $B_i = [0, x^* - x_i]$, where x_i is the time the ith individual is diagnosed of AIDS. The log-likelihood based on a sample of size N is

$$L = \sum_{i=1}^{N} \left\{ \log\left[\int_0^{s^*} I_{A_i} f(s|\mathbf{z}_i)\, ds \right] - \log\left[\int_0^{s^*} I_{B_i} f(s|\mathbf{z}_i)\, ds \right] \right\}, \quad (1)$$

where $\mathbf{z}_i$ denotes the value of the covariate vector for the ith individual and I_C denotes the indicator having value 1 if $s \in C$ and 0 otherwise.

To estimate $F(s|\mathbf{z})$, consider the case where time is discrete and S has masses only at $s_0, s_1, \ldots, s_J$ ($s_J = s^*$). This discrete formulation is reasonable for the CDC surveillance data, as events are recorded in units such as a month or a quarter of a year, and would give better estimates when used to analyze such data where ties are numerous (Kalbfleisch and Lawless 1991). Because the time unit in our analysis is a quarter of a year, we let $s_j = j$ ($0 \le j \le J$) with s_0 denoting the mass point for death occurring during the same quarter as AIDS diagnosis, though this simplification is not necessary for the following development.

Let $\xi_{ij} = 1$ if $j \in A_i$ and 0 otherwise, and let $\eta_{ij} = 1$ if $j \in B_i$ and 0 otherwise. The log-likelihood (1) then has the form

$$L = \sum_{i=1}^{N} \left\{ \log\left[\sum_j \xi_{ij} f(j|\mathbf{z}_i) \right] - \log\left[\sum_j \eta_{ij} f(j|\mathbf{z}_i) \right] \right\}. \quad (2)$$

Note that if there is no censoring or truncation in the data, this expression reduces to

$$L = \sum_{i=1}^{N} \sum_{j=0}^{J} I_{ij} \log f(j|\mathbf{z}_i), \quad (3)$$

where $I_{ij} = 1$ if the ith individual fails at time $s = j$ and 0 otherwise.

We have chosen a discrete analog of the proportional hazards model (Cox 1972; Prentice and Gloeckler 1978) to model the dependence of survival on covariates. Specifically, we consider the model

$$f(j|\mathbf{z}) = (p_0 \cdots p_{j-1})^{\exp\{\mathbf{z}^T\beta\}}(1 - p_j^{\exp\{\mathbf{z}^T\beta\}}) \quad \text{if } 0 \le j \le J-1$$
$$= (p_0 \cdots p_{J-1})^{\exp\{\mathbf{z}^T\beta\}} \quad \text{if } j = J, \tag{4}$$

where $p_j = \Pr(S > j+1 \mid S > j)$ $(0 \le j \le J-1)$ are conditional baseline probabilities, which also correspond to $\mathbf{z} = 0$ in our model. Note that to facilitate computation, the range restriction on p_j is often removed by a reparameterization: $a_j = \log[-\log(p_j)]$ for $1 \le j \le J-1$. This model implies that the hazard function $h(s|\mathbf{z})$ is approximately

$$h(j|\mathbf{z}) \approx \exp\{a_j + \mathbf{z}^T\beta\}. \tag{5}$$

To facilitate the discussion, we let $\alpha = (a_0, \ldots, a_{J-1})$ and $\theta = (\alpha, \beta)$ and explicitly write $f(s|\mathbf{z}, \theta)$ for the dependence of f on θ.

With this model possible temporal trends in survival over the time of diagnosis, x, is investigated by allowing $\mathbf{z}$ to incorporate x. If this dependence is linear, then the distribution is stationary if the component of the regression coefficient vector β corresponding to x is not significantly different from 0, with an increase or decrease in survival suggested by whether that component carries a negative or a positive sign. This model is also applied to estimating the reporting delay distribution (Sec. 2), where $\mathbf{z}$ includes covariates such as regions of residence to explore the possible heterogeneity in reporting delays.

1.2 Parameter Estimation Via the EM Algorithm

Under the discrete proportional hazards model, estimation can proceed by maximizing the log-likelihood (2) with respect to the vector θ. Such an estimation can be simplified by using the EM algorithm (Dempster et al. 1977). The algorithm we describe, which is an application of their general approach, can also be viewed as an extension of Turnbull's algorithm (Turnbull 1976) for regression analysis of censored and truncated data. This algorithm in essence converts a difficult incomplete-data problem (resulting from censoring and truncation in our setting) into a sequence of pseudo-complete-data problems (involving no censoring and truncation). In the current context this approach is especially attractive, because it not only simplifies programming but also permits the use of existing software packages to estimate the model parameters.

Under the framework of incomplete-data analysis, censoring can be viewed as a type of incomplete data in the sense that the failure time for a censored subject is not observed at a mass point, but instead is only known to lie in an interval or a union of disjoint intervals of mass points. When truncation occurs, only a portion of the sample is observed, and the unobserved portion owing to truncation constitutes the missing observations. Because the complete-data log-likelihood (3) is linear in the data, the E step of the EM algorithm amounts to imputing the unobserved portion of the sample as well as the failure times for these and the censored individuals, using the observed data and estimates of model parameters from the previous M step. The model parameters are reestimated in the M step by maximizing the complete-data log-likelihood, treating the imputed data as though they were observed. We now describe the details.

Let $I_{ij} = 1$ if the ith individual fails at time $s = j$ and 0 otherwise, and let J_{ij} be the number of individuals who have covariate $\mathbf{z}_i$ and failure time at $s = j$. Given some initial estimate for θ, the algorithm at the $(k+1)$st cycle of iteration proceeds as follows:

1. E step: For each censored individual, the value of I_{ij} is not known. But its expectation, conditional on the previous estimate $\theta^{(k)}$ and all the observed data, denoted Y_{obs}, is given by

$$c_{ij}^{(k)} \equiv E(I_{ij}|Y_{\text{obs}}, \theta^{(k)}) = \xi_{ij}\frac{f(j|\mathbf{z}_i, \theta^{(k)})}{\sum_{r=0}^{J}\xi_{ir}f(r|\mathbf{z}_i, \theta^{(k)})}$$

for $1 \le i \le N$ and $0 \le j \le J$. Note that with an observed failure at time j, $c_{ij}^{(k)} \equiv 1$ and $c_{il}^{(k)} \equiv 0$ for all $l \ne j$. Similarly, because of truncation the value of J_{ij} may not be known. To calculate the conditional expectation of J_{ij}, suppose that $\sum_{r=0}^{J}(1 - \eta_{ir})J_{ir}$ follows a negative-binomial distribution

$$NB[m|n_i, P(\mathbf{z}_i)] = \binom{m + n_i - 1}{m}[1 - P(\mathbf{z}_i)]^m P(\mathbf{z}_i)^{n_i},$$

where $P(\mathbf{z}_i) = \sum_{r=0}^{J}\eta_{ir}f(r|\mathbf{z}_i, \theta^{(k)})$ and $n_i = \sum_{r=0}^{J}\eta_{ir}J_{ir}$. The conditional expectation of J_{ij} is then given by

$$g_{ij}^{(k)} \equiv E(J_{ij}|Y_{\text{obs}}, \theta^{(k)}) = \frac{(1 - \eta_{ij})f(j|\mathbf{z}_i, \theta^{(k)})}{\sum_{r=0}^{J}\eta_{ir}f(r|\mathbf{z}_i, \theta^{(k)})}$$

for $1 \le i \le N$ and $0 \le j \le J$.

2. M step: This step updates the parameter estimate by maximizing the expected complete-data log-likelihood

$$L^*(\theta|Y_{\text{obs}}, c_{ij}^{(k)}, g_{ij}^{(k)}) = \sum_{i=1}^{N}\sum_{j=0}^{J}[c_{ij}^{(k)} + g_{ij}^{(k)}]\log f(j|\mathbf{z}_i, \theta). \tag{6}$$

Denote the solution by $\theta^{(k+1)}$.

The estimate of the parameter vector $\theta = (\alpha, \beta)$ is obtained by cycling between the two steps until convergence. Note that the negative-binomial distribution in the E step, as described in Dempster et al. (1977, pp. 13–15), is a part of the complete-data formulation for applying the EM algorithm.

The expected complete-data log-likelihood (6) in the M step of the algorithm is in the same form as that of the log-likelihood (3). As for computation, this expected log-likelihood can be regarded as that of (3) under the pseudo-complete-data $\{c_{ij}^{(k)} + g_{ij}^{(k)}\}_{1 \le i \le N, 1 \le j \le J}$. The only difference is that the contribution to the expected log-likelihood for the pseudo-complete-data is $c_{ij} + g_{ij}$ rather than just 1 or 0 as in (3). Thus, the algorithm can simplify programming and use routines developed for complete-data analysis. For example, we obtained the estimates of our analysis (Sec. 3) using this

algorithm in conjunction with a computer routine for implementing the discrete proportional hazards model with observed failure times written in SAS IML (SAS Institute 1985). Because the discrete proportional hazards model is equivalent to the complementary log-log model, other packages developed for categorical data analysis such as GLIM (McCullach and Nelder 1989) may also be used.

Although the EM algorithm is not sensitive to starting values, the number of iterations required to convergence can be reduced by a sensible choice of initial values. For the proportional hazards model, the vector β can be set to 0, but for α we discuss two approaches. One is to use the starting values $1/J$ for each $f(j)$ $(0 \le j \le J)$, as in Turnbull (1976). Given $f(j)$, we can solve the set of equations

$$(p_0 \ldots p_{j-1})(1 - p_j) = f(j) \qquad (0 \le j \le J - 1)$$

and

$$p_0 \ldots p_{J-1} = f(J)$$

for p_j $(0 \le j \le J - 1)$ in closed form. The initial estimate for α is obtained by setting $a_j = \log[-\log(p_j)]$ for $0 \le j \le J - 1$. Alternatively, when the data are only right-truncated, a closed-form solution for $f(j)$ can be easily obtained by restricting attention to the noncensored portion of the data (Lagakos, Baraj, and DeGruttola 1988; Wang et al. 1986). If the amount of censoring is moderate, this approach would yield an initial estimate for α close to the MLE. The latter approach is used in our application, because there is a small proportion of deaths that occurred before 1984. The same approach is used for obtaining the initial values for estimating the reporting delay distributions (Sec. 2.1).

1.3 Calculation of the Asymptotic Variance Estimate

Hypotheses testing for the parameter vector, especially for regression coefficients, can be based either on the log-likelihood ratio test or on the Wald χ^2 test. The latter approach, as well as the method of multiple imputation (Sec. 2.2), requires an access to the observed information matrix, $\mathbf{I}_{\text{obs}}$. This matrix, of course, can be obtained by inverting the negative of the second-order derivative of the observed log-likelihood (2). A more efficient way, however, is to use the quantities calculated from the EM algorithm. An approach given in Louis (1982) is to correct the information matrix from the pseudo-complete-data log-likelihood, $\mathbf{I}_c$, for the additional variation due to missing data. It is readily shown that this approach leads to the following formula:

$$\begin{aligned}\mathbf{I}_{\text{obs}}(\theta^*) = \mathbf{I}_c(\theta^*) &- \sum_{i=1}^{N} \text{var}[\mathbf{S}(s|\mathbf{z}_i, \theta^*)|\mathbf{I}_{A_i} = 1] \\ &+ \sum_{i=1}^{N} \text{var}[\mathbf{S}(s|\mathbf{z}_i, \theta^*)|\mathbf{I}_{B_i} = 1] \\ &- \sum_{i=1}^{N} \frac{E[\mathbf{S}(s|\mathbf{z}_i, \theta^*)\mathbf{S}^{\mathrm{T}}(s|\mathbf{z}_i, \theta^*)]}{E[\mathbf{I}_{B_i}]}, \end{aligned} \tag{7}$$

where $\theta^* = (\alpha^*, \beta^*)$ denotes the MLE of $\theta = (\alpha, \beta)$ based on the observed data Y_{obs}, $\mathbf{S}(s|\mathbf{z}_i, \theta^*) = (\partial/\partial\theta) \times \log f(s|\mathbf{z}_i, \theta^*)$, and $\text{var}(\cdot\,|\mathbf{I}_A = 1)$ denotes the variance–covariance matrix conditional on $\mathbf{I}_A = 1$. Note that in case of no truncation, (7) reduces to the formula given by Louis (1982) for a multinomial distribution with missing cell counts. The use of the formula avoids the calculation of the second-order derivatives of the log-likelihood (2), because the quantity $\mathbf{S}(s|\mathbf{z}_i, \theta)$ is the score vector corresponding to an individual with covariate $\mathbf{z}_i$ and failure time s.

Alternatively, we can apply the SEM algorithm (Meng and Rubin 1991) to compute $\mathbf{I}_{\text{obs}}$ and completely avoid the analytical calculation of derivatives except for those required by the computation of $\mathbf{I}_c$. The key idea is to use the matrix of the rate of convergence of the EM algorithm, $\mathbf{DM}(\theta^*)$, which equals the matrix of fractions of missing information, to deflate the complete-data information matrix

$$\mathbf{I}_{\text{obs}}(\theta^*) = [\mathbf{I} - \mathbf{DM}(\theta^*)]\mathbf{I}_c(\theta^*).$$

The $\mathbf{DM}(\theta^*)$ matrix is obtained by numerical differentiation, which only requires the code for the EM algorithm because the mapping function $\theta^{(k+1)} = \mathbf{M}(\theta^{(k)})$, used in evaluating the numerical differentiation, is implicitly defined by the E and M steps. In the present context the SEM algorithm may be preferable if the EM algorithm is implemented using standard packages, because the formula in (7) requires the use of the score vector $\mathbf{S}(s|\mathbf{z}_i, \theta^*)$, which may not be readily available as standard output from these packages.

2. SURVIVAL DISTRIBUTION IN THE PRESENCE OF REPORTING DELAY

2.1 Estimating the Delay Distribution of Reporting Death

Given the lag time of reporting of death, r, and a covariate vector $\mathbf{z}$ that may include the time of death, t, estimation of the delay distribution can proceed using the methods described in Section 1. Following the notation in Section 1, the delay density is given by:

$$\begin{aligned}\delta(r|\mathbf{z}) &= (p_0 \ldots p_{r-1})^{\exp\{\mathbf{z}^{\mathrm{T}}\beta\}}(1 - p_r^{\exp\{\mathbf{z}^{\mathrm{T}}\beta\}}) \\ &\qquad\qquad \text{if } 0 \le r \le R - 1 \\ &= (p_0 \ldots p_{R-1})^{\exp\{\mathbf{z}^{\mathrm{T}}\beta\}} \qquad \text{if } r = R,\end{aligned}$$

where R is the maximal observed delay. To facilitate the following discussion, we denote this density by $\delta(r|\mathbf{z}, \psi)$, where $\psi = (a_0, \ldots, a_{R-1}, \beta)$ with $a_r = \log[-\log(p_r)]$.

The lag information needed for estimating the delay distribution was available for deaths that occurred after September 1987 when their reporting times (in calendar quarters) were recently added to the CDC Public Use data base as of July 1991. The data containing the lag information are also right-truncated, because only deaths with reporting delays less than $x^* - t$ are in the data base, where t is the calendar time of death. As a result of right truncation, the reporting delay distribution is identifiable only if

$$\Pr[\text{maximal observed delay} \le x^*] = 1. \tag{8}$$

If we restrict to the deaths with known times of reporting, this maximal observed delay is about 4 years. To increase the maximal delay, we also included deaths that occurred

before October 1987 but after January 1986, with their unknown times of reporting treated as left-censored. This allowed us to obtain a delay distribution conditional on deaths reported within about 5 years. Comparison of the number of reported deaths between the data set as of October 1988 and that as of July 1991 shows almost no change in the number of deaths that occurred before 1986. It is, therefore, quite likely that deaths not reported within 5 years will rarely or never be reported. Expression (8) thus approximately holds.

2.2 Accommodating the Delay Distribution Via Multiple Imputation

To correct for the bias caused by the reporting delay, we can use the information contained in the reporting delay distribution to impute the unreported deaths (missing observations) and thus to incorporate them into the survival distributions. Given the delay distribution $\delta(r|\mathbf{z}, \psi)$, n_i observed deaths with failures at time t_i and covariate $\mathbf{z}_i$, we can impute the unreported deaths using the expected number of such cases under the assumption that the number of unreported deaths follows a negative-binomial distribution; that is

$$E[m|n_i, P(t_i, \mathbf{z}_i, \psi)] = \frac{1 - P(t_i, \mathbf{z}_i, \psi)}{P(t_i, \mathbf{z}_i, \psi)} n_i, \qquad (9)$$

where m denotes the number of unreported deaths and

$$P(t, \mathbf{z}, \psi) = \sum_{r=0}^{x^*-t} \delta(r|\mathbf{z}, \psi).$$

Although this mean-imputation process produces reasonable estimates for the survival distribution under the model assumption, the inference based on such imputation systematically underestimates the sampling variability, because the imputed observations are treated as if they were actually observed. In addition, because the model parameter vector ψ can be only estimated, the variation in the estimator itself should also be reflected with a proper imputation scheme.

Multiple imputation (Rubin 1987a) provides a natural tool for accommodating variability. (For a review of multiple imputation in health-care studies, see Rubin and Schenker 1991.) The idea underlying multiple imputation, in contrast to single imputation, is to let the variability in the missing data be manifested in a number of completed-data sets and then to combine the data analyses from these completed-data sets to reach valid inferences. In our analysis, the imputation task is accomplished by independent draws from the posterior predictive distribution of the missing data, under the assumption that the missing data are missing at random (Little and Rubin 1987; Rubin 1976).

To describe the details for implementing multiple imputation in our context, let $\hat{\psi}$ and $\hat{\Omega}$ be the MLE and the inverse of the observed information matrix for the reporting delay distribution obtained by applying the methods in Section 2.1. Under a noninformative prior for ψ, the standard large-sample asymptotic theory implies that the posterior distribution of ψ can be approximated by the multivariate normal distribution $\mathbf{N}(\hat{\psi}, \hat{\Omega})$. (The method of importance sampling [see, for example, Rubin 1987b] may be used to improve the approximation in case of a relatively small sample size.) The conditional distribution of the number of unreported deaths given n_i reported deaths with failures at time t_i and covariate $\mathbf{z}_i$ and the parameter vector ψ under our assumption is a negative-binomial:

$$NB(m|n_i, P(t_i, \mathbf{z}_i, \psi)) = \binom{m + n_i - 1}{m} [1 - P(t_i, \mathbf{z}_i, \psi)]^m P(t_i, \mathbf{z}_i, \psi)^{n_i},$$

where $P(t_i, \mathbf{z}_i, \psi) = \sum_{r=0}^{x^*-t_i} \delta(r|\mathbf{z}_i, \psi)$. Thus the independent draws from the posterior predictive distribution required by multiple imputation can be obtained as follows.

To obtain M completed-data sets by multiple imputation, we perform, for each l $(1 \le l \le M)$, the following steps with independent draws for all the random variables at each pass:

1. Draw a random sample ψ_l from $N(\hat{\psi}, \hat{\Omega})$.
2. Given ψ_l, for each observed n_i with death time t_i and covariate $\mathbf{z}_i$, draw a random sample $m_i^{(l)}$ from the negative-binomial distribution $NB[m|n_i, P(t_i, \mathbf{z}_i, \psi_l)]$.
3. For each observed n_i with death time t_i and covariate $\mathbf{z}_i$, impute the unreported deaths by $m_i^{(l)}$ to form the lth completed-data set Y_{*l}, which now contains $\{n_i + m_i^{(l)}\}$ number of deaths.

Note that because the maximum delay is about 5 years, only the unreported deaths with an AIDS diagnosis after January 1986 need to be imputed, in which case exact death times (in quarters) are available for the reported ones.

2.3 Analyzing the Multiply Imputed Data Sets

Applying the methods in Section 1 to each of the completed-data sets yields a set of ML estimates $\{\hat{\theta}_l = (\hat{\alpha}_l, \hat{\beta}_l); 1 \le l \le M\}$, as well as a set of the inverses of the observed information matrices $\{\hat{\mathbf{\Sigma}}_l; 1 \le l \le M\}$ for the survival distribution. Following Rubin (1987a), the multiple imputation estimate of θ is given by averaging over the estimates from the M imputations; that is,

$$\bar{\theta}_M = M^{-1} \sum_{l=1}^{M} \hat{\theta}_l.$$

The average of $\hat{\mathbf{\Sigma}}_l$ denoted by $\bar{\mathbf{\Sigma}}_M$, however, underestimates the variability associated with $\bar{\theta}_M$, because it does not take into account the variability due to the missing data. The appropriate estimate that measures the total variability of $\bar{\theta}_M$ is given by

$$\mathbf{T}_M = \bar{\mathbf{\Sigma}}_M + (1 + M^{-1})\mathbf{B}_M,$$

where

$$\mathbf{B}_M = (M - 1)^{-1} \sum_{l=1}^{M} (\hat{\theta}_l - \bar{\theta}_M)(\hat{\theta}_l - \bar{\theta}_M)^{\mathrm{T}}$$

measures the "between-imputation variability." Hypothesis testing for $\mathbf{C}\theta = 0$, where $\mathbf{C}$ is a $k \times d$ matrix and d $(\ge k)$ is the dimensionality of θ, can be accomplished by referring a modified Wald statistic,

$$\mathbf{D}_M = \frac{(\mathbf{C}\bar{\theta}_M)^{\mathrm{T}}(\mathbf{C}\bar{\mathbf{\Sigma}}_M\mathbf{C}^{\mathrm{T}})^{-1}\mathbf{C}\bar{\theta}_M}{k(1 + r_M)},$$

to an $F_{k,w}$ distribution, where

$$r_M = (1 + M^{-1})\text{trace}[\mathbf{C}\mathbf{B}_M\mathbf{C}^T(\mathbf{C}\bar{\boldsymbol{\Sigma}}_M\mathbf{C}^{\mathrm{T}})^{-1}]/k$$

and w is given by (Li, Raghunathan, and Rubin 1991):

$$w = w(r_M) = 4 + (v - 4)\left[1 + \left(1 - \frac{2}{v}\right)r_M^{-1}\right]^2, \quad \text{if } v = k(M-1) > 4$$
$$= \frac{v}{2}\left(1 + \frac{1}{k}\right)(1 + r_M^{-1})^2, \quad \text{otherwise.} \tag{10}$$

Alternatively, a recently proposed modified likelihood ratio test (Meng and Rubin 1992) can also be used to provide a p value for the null hypothesis $\mathbf{C}\theta = 0$. This approach is asymptotically equivalent to the previously described approach for any number of multiple imputations, but avoids the computation of the variance–covariance matrices $\mathbf{C}\hat{\boldsymbol{\Sigma}}_l\mathbf{C}^{\mathrm{T}}$ for $1 \le l \le M$, which could take up a lot of computing storage when k is large. Using this approach, the complete-data log-likelihood ratio

$$R(\hat{\theta}_0, \hat{\theta}) = -2\log\left[\frac{L(\hat{\theta}_0|\mathbf{Y})}{L(\hat{\theta}|\mathbf{Y})}\right]$$

is first computed as a function of the ML estimates $\hat{\theta}_0$ and $\hat{\theta}$ for each of the completed-data sets under the null model and full models to obtain $d_l = R(\hat{\theta}_{0l}, \hat{\theta}_l|\mathbf{Y}_{*l})$ for $1 \le l \le M$. The log-likelihood ratio is then computed again for the averaged ML estimates,

$$\bar{\theta}_{0M} = M^{-1}\sum_{l=1}^{M}\hat{\theta}_{0l}, \qquad \bar{\theta}_M = M^{-1}\sum_{l=1}^{M}\hat{\theta}_l,$$

under each of the completed-data sets to obtain $d_l^* = R(\bar{\theta}_{0M}, \bar{\theta}_M|\mathbf{Y}_{*l})$. Note that for our survival analysis, the appropriate complete-data log-likelihood function for computing these likelihood ratios is from expression (2), not from (3) or (6). Let $\bar{d}_M$ and $\bar{d}_M^*$ be the averages of d_l and d_l^*. Then the p value for $\mathbf{C}\theta = 0$ is $\Pr(F_{k,w(r_L)} \ge \mathbf{D}_L)$, where

$$r_L = \frac{M+1}{k(M-1)}(\bar{d}_M - \bar{d}_M^*), \qquad \mathbf{D}_L = \frac{\bar{d}_M^*}{k(1 + r_L)},$$

and the function $w(r)$ is given by (10).

3. RESULTS

3.1 Risk Groups Considered and Variations in Reporting Delays

As of July 1991, 142,605 AIDS cases were reported to the CDC. Of these, 120,553 were men (excluding pediatric and transfusion-related AIDS cases), who constitute about 85% of the reported AIDS population in the United States. Knowledge of variation in survival among these patients, which still forms a driving force for the current AIDS epidemic, would be of great help in assessing the current health care needs and for health policy planning.

The analysis was based on the 82,239 reported deaths with an AIDS diagnosis between the first quarter of 1983 and the first quarter of 1991. Note that deaths with AIDS diagnosed in the second quarter of 1991, the most recent quarter at the time of our analysis, were not used because of the severe underreporting. To study the difference in survival among the different risk groups as well as the difference resulting from manifestations of AIDS, we divided this subpopulation into four major risk groups on the basis of their sexual behavior and injecting drug (ID) use status: men who have sex with men (including bisexual contact) with ID use (6,329 deaths) and non-ID use (60,530), denoted IDMSM and MSM; and heterosexual men with ID use (14,533) and non-ID use (847), denoted IDMSW and MSW. Within each risk group we also defined three diagnosis strata in order: (1) *Pneumocystis carinii* pneumonia (PCP) (either definitively or presumptively diagnosed), (2) disease manifestations other than PCP and Kaposi's sarcoma (OTH), and (3) Kaposi's sarcoma (KS) (either definitively or presumptively diagnosed). (A similar classification scheme was used by Lemp et al. 1990.) Individuals with multiple diagnoses were classified according to the highest-ranked stratum. For example, individuals diagnosed with PCP and other opportunistic infection(s) were classified into the PCP category, those diagnosed with other opportunistic infection(s) and KS were classified into the OTH category, and so on. Thus under this classification scheme, classifying a patient PCP only indicates that the patient had PCP as a diagnosis, not necessarily the first diagnosis. Note that to minimize the influence of the change of definition in 1987 on survival, patients with a disease added to the case definition in 1987 (basically wasting, dementia, and disseminated TB; see CDC 1991) were excluded from the analysis.

Because of the underreporting of deaths resulting from delays in reporting, the total number of deaths that would have been reported within 5 years were estimated using the methods discussed in Section 2. Because time trends in delays of reporting AIDS incidence as well as variations among the five geographic regions consisting of metropolitan statistical areas with population at least 1 million (Northeast, Central, West, South, Mid-Atlantic) and a residual category consisting of areas with population less than 1 million, have been reported (Brookmeyer and Liao 1990; Harris 1990a; Zeger, See, and Diggle 1989), suspicion of the same phenomenon led us to model for the delay distributions the variation among the risk groups and a time trend for each of the six regions. The variation among the risk groups is modeled by coding an indicator for each group with IDMSM serving as a reference group; the time trend, by a covariate with values from 1 to 6 designating the year of death between 1986 and 1991. Shown in Table 1 are the estimates of the regression parameters for the discrete proportional hazards model when fitted to each region. All the regions show a decreasing delay in reporting except for the Northeast, which shows an increasing trend. Plotted in Figure 1 are the medians of the delay distributions for each of the regions, obtained by combining the risk groups. The Northeast region has the longest delay and is the only region in which the delay has increased between 1986 and 1991.

Plotted in Figure 2 are the CDF's of the delay distributions for the four risk groups, obtained under the stationarity assumption by combining the regions. It seems that the delays in reporting death are similar for all the risk groups except

Table 1. Regression Coefficients (Standard Errors) for the Discrete Proportional Hazards Model When Fitted to Each Region for Modeling the Reporting Delays

Covariates	Northeast	Central	West	South	Mid-Atlantic	Population < 1 million
Time of death	−.075 (.016)	.109 (.017)	.126 (.011)	.026 (.016)	.072 (.021)	.189 (.010)
IDMSW	.005 (.057)	−.035 (.098)	−.132 (.090)	−.379 (.076)	.027 (.101)	−.070 (.042)
MSM	.074 (.056)	−.005 (.075)	.039 (.044)	.059 (.055)	.129 (.082)	.022 (.037)
MSW	1.072 (.144)	.130 (.213)	−.174 (.219)	−.287 (.180)	.338 (.215)	−.018 (.082)

NOTE: Risk groups are coded as indicators, with IDMSM serving as a reference group and time of death a quantitative variable.

IDMSW, which shows a relatively longer delay. The plot shows that approximately 90% of the deaths that will ever be reported will be done so within 2 years.

3.2 Estimated Survival Distributions

Having the delay distributions, the survival distributions were estimated next, by adjusting the reported cases as described in Section 2. Because deaths with an AIDS diagnosis after 1982 were included in the analysis, we were able to obtain distributions conditional on death within about 8 years following diagnosis. Plotted in Figures 3–5 are the survival distributions (1 − CDF) for the four risk groups within each disease category obtained under the discrete proportional hazards model and the assumption that survival has not changed over the time of diagnosis. (The widest confidence bands in Figs. 3–5 are associated with the MSW due to its relatively small sample size.) These plots show that the MSM seems to have the longest survival and IDMSW the shortest survival after a diagnosis of AIDS. Figures 3–5 also indicate that the average medians of survival for the PCP, OTH, and KS cases are approximately 12, 9, and 15 months, all of which are comparable with the estimates reported by Harris (1990b), Lemp et al. (1990), and Friedland et al. (1991). The dip at the end of the third month for PCP, which was also observed in Harris (1990b) and in Friedland et al. (1991), seems to suggest that those diagnosed with PCP might be at a relatively high risk during the first 3-month period following diagnosis.

Note that the survival curves for OTH in Figure 5 show a close match when they are approximated by exponential distributions with means estimated from them. But, such exponential curves do not provide as good approximations for the survival curves in Figures 3 and 4. For KS, the exponential curves fall below the estimated survival curves during the first 2 years after diagnosis, with a maximum discrepancy of about 5%, then cross and stay above them after that, with a maximum difference of about 3%. The approximations for PCP are similar, except for the reduced ranges of discrepancy—3% and 2% for the respective periods—and some agreement between these curves during the initial 6 months following diagnosis.

The survival trend for each disease and risk group category was obtained by fitting the discrete proportional hazards model with a covariate coded from 1 to 9 designating the year of diagnosis between 1983 and 1991. Shown in Table 2 are the estimates of the regression parameters and the associated p values for the disease and risk group categories corresponding to both the single—using mean imputation (9)—and multiple imputations. It is clear from the table that

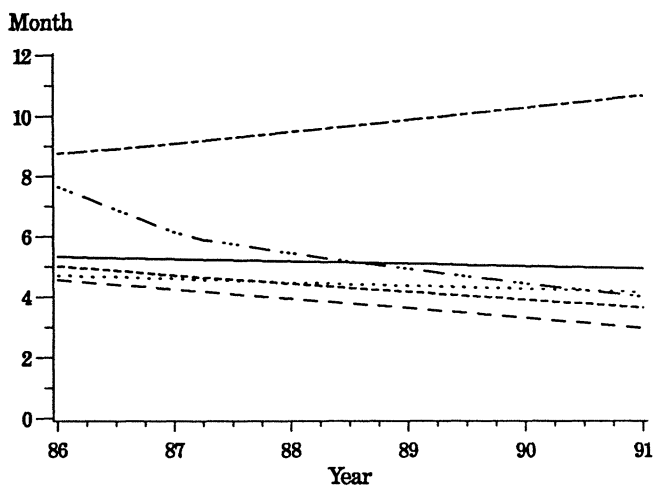

Figure 1. Medians of Reporting Delay Distributions for the Six Geographic Regions Classified by the CDC Surveillance System. The short- and long-dashed line (– – –) represents the Northeast region; the dashed line (– – – –), the West region; the dotted line (· · · ·), the Mid-Atlantic region; the long-dashed line (— — —), the Central region; the solid line (———), the South region; and the dotted and long-dashed line (· · · — · · · —), the population <1 million region.

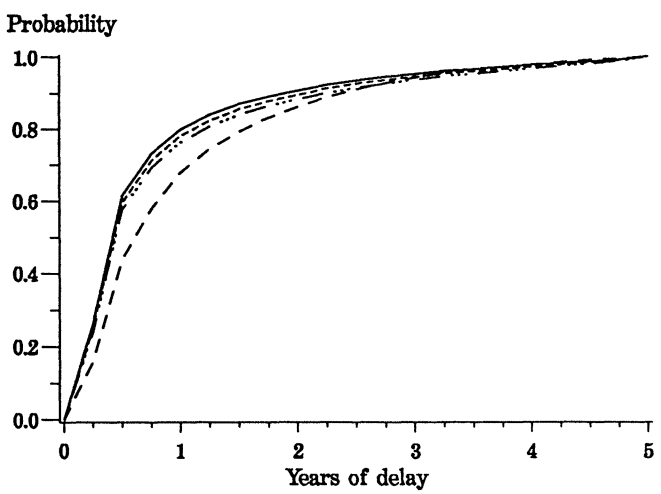

Figure 2. Cumulative Reporting Delay Distributions for the Four Risk Groups. The dotted and long-dashed line (· · · — · · · —) represents the IDMSM group; the long-dashed line (— — —), the IDMSW group; the short-dashed line (– – –), the MSM group; and the solid line (———), the MSW group.

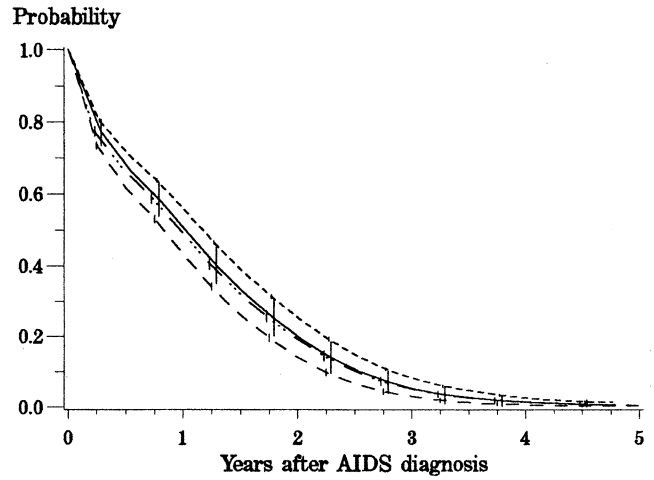

Figure 3. Survival Distributions and 95% Confidence Intervals (Vertical Bars) for Pneumocystis carinii Pneumonia for the Four Risk Groups. The dotted and long-dashed line (· · · — · · · —) represents the IDMSM group; the long-dashed line (— — —), the IDMSW group; the short-dashed line (– – –), the MSM group; and the solid line (———), the MSW group.

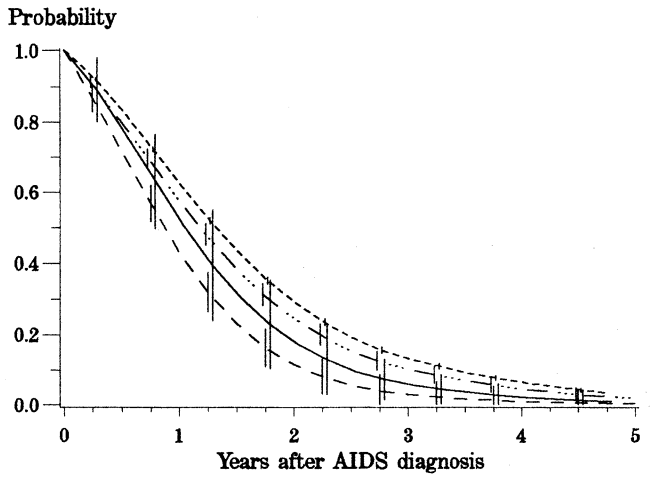

Figure 5. Survival Distributions and 95% Confidence Intervals (Vertical Bars) for Kaposi's Sarcoma for the Four Risk Groups. The dotted and long-dashed line (· · · — · · · —) represents the IDMSM group; the long-dashed line (— — —), the IDMSW group; the short-dashed line (– – –), the MSM group; and the solid line (———), the MSW group.

even though the single imputation yields very similar estimates for the regression coefficients, it systematically underestimates the associated variabilities (often by an order of magnitude), with the degree of underestimation determined by the fraction of missing information (Rubin 1987a). Thus the p values that correspond to single imputation are far too small for some disease and risk group categories and will be very misleading when used for statistical inferences. In contrast, the p values based on multiple imputation, even though they are slightly different under different number of imputations, essentially imply the same practical conclusions and provide valid inferences. It seems that in our context, ten imputations will suffice to yield reliable results.

The results in Table 2 indicate that significant increase in survival between 1983 and 1991 is mostly confined to the homosexual population (MSM and IDMSM) with PCP as one of the diagnoses, even though the patients who were diagnosed with opportunistic infections other than PCP and KS within the same population also show some improvement, especially for non-ID users. Under the model assumption the estimated coefficients shown in Table 2 imply an average 13% annual reduction rate in mortality for the PCP patients in this risk population and about a 7% reduction rate for those in the OTH category within the same population.

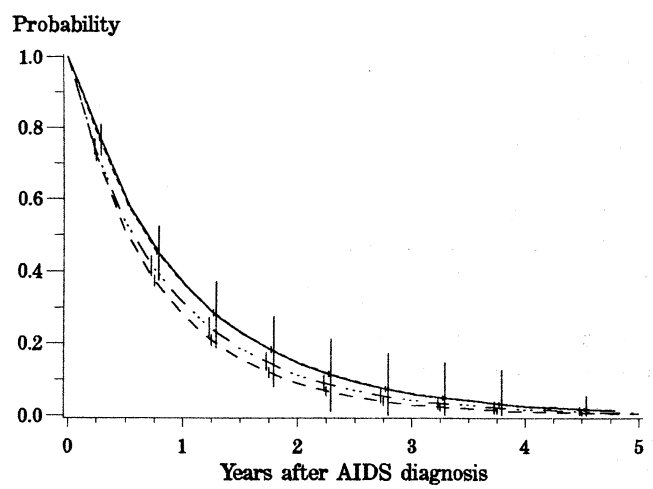

Figure 4. Survival Distributions and 95% Confidence Intervals (Vertical Bars) for Diagnosis Other Than Pneumocystis carinii Pneumonia and Kaposi's Sarcoma for the Four Risk Groups. The dotted and long-dashed line (· · · — · · · —) represents the IDMSM group; the long-dashed line (— — —), the IDMSW group; the short-dashed line (– – –), the MSM group; and the solid line (———), the MSW group.

Improvement of survival for PCP after 1987 has been reported by short-term survival analysis based on the surveillance data (Harris 1990b; Lemp et al. 1991). To accentuate the difference in survival between the two periods, we fitted the model to the PCP cases for the homosexual population using a binary covariate with 1 for people diagnosed after 1987 and 0 otherwise. The estimates for the regression coefficients are −.34 for ID users and −.52 for non-ID users, both of which are highly significant as a consequence of stationarity tests presented in Table 2. These estimates imply reduced mortality of 28% for ID users and 40% for non-ID users after 1987, which seem to be comparable to the reduced mortality of about 30% found in the San Francisco study for patients taking zidovudine (AZT) (CDC 1990; Lemp et al. 1990) and some other clinical trial studies (Volberding et al. 1990). Based on these other studies, it has been hypothesized that the improvement for PCP after 1987 was partly due to the intervention of the antiviral drug AZT (Harris 1990b), although improvement of survival is evident even before 1987, a fact that must also be part of any explanation that attributes the improvement to some key event that occurred in 1987 (Bennett et al. 1989; Cotton 1989). Note that the reduction in mortality for non-ID users seems to be slightly higher than that reported by the San Francisco study, which might suggest that the improved survival for the PCP cases could also be attributed to other treatments, such as

Table 2. Estimates (p Values) of Regression Coefficient of Time of Diagnosis (Coded as a Quantitative Variable) for the Discrete Proportional Hazards Model When Fitted to Each Risk Group and Disease Category for Modeling Survival under Mean Imputation and Multiple Imputation (with M Imputations)

Number of imputations	MSM			MSW		
	PCP	OTH	KS	PCP	OTH	KS
			ID			
Mean imputation	−.113 (.000)	−.071 (.007)	.014 (.082)	−.009 (.315)	.008 (.035)	.030 (.005)
$M = 10$	−.099 (.000)	−.066 (.046)	.016 (.546)	−.011 (.316)	.008 (.656)	.025 (.063)
$M = 50$	−.100 (.000)	−.064 (.069)	.012 (.629)	−.010 (.399)	.009 (.601)	.029 (.056)
			Non-ID			
Mean imputation	−.162 (.000)	−.080 (.000)	−.077 (.025)	.019 (.514)	.038 (.091)	.050 (.472)
$M = 10$	−.163 (.000)	−.079 (.000)	−.077 (.079)	.019 (.745)	.031 (.621)	.073 (.623)
$M = 50$	−.163 (.000)	−.079 (.000)	−.076 (.061)	.007 (.924)	.029 (.745)	.051 (.776)

prophylactic trimethoprim-sulfamethoxazole or aerosolized pentamidine, that have been found to be successful in preventing recurrence of PCP (Hirschel et al. 1991; Leoung et al. 1990). Of course, without controlled clinical trials, all these hypotheses can remain only speculations.

Plotted in Figure 6 are the medians of survival for the PCP cases from the risk groups IDMSM and MSM and for the OTH cases from the MSM risk group. It is seen that the medians for PCP in the MSM risk group averaged over the periods 1983–1985, 1986–1987, and 1988–1990 are approximately 10, 15, and 21 months. These estimates seem to be comparable to the medians 10.3, 17.9, and 21 months for the PCP cases diagnosed in the periods 1981–1985, 1986–1987, and 1988–1990, estimated from the San Francisco study (Lemp et al. 1991). Note that only the estimates for PCP in the MSM group are comparable to the estimates reported from the San Francisco study, which may not be surprising, as about 85% of the population in the San Francisco study were MSM. Significantly increased survival was also reported by Lemp et al. (1991) for cases diagnosed with infections other than PCP and KS. The medians of survival for OTH in the MSM risk group averaged over 1983–1985, 1986–1987, and 1988–1990 are approximately 7, 8, and 10 months, which are slightly lower than the average medians 9, 10, and 13 months over 1981–1985, 1986–1987, and 1988–1990 for the cases with infections other than PCP and KS reported in Lemp et al. (1991).

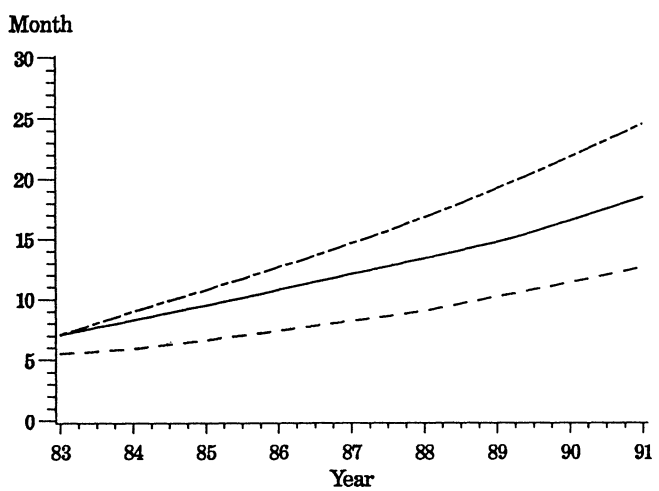

Figure 6. Median Survival Times for Pneumocystis carinii Pneumonia Cases in the MSM and IDMSM Risk Groups and for Diagnoses Other than Pneumocystis carinii Pneumonia and Kaposi's Sarcoma in the MSM Risk Group. The solid line (———) represents the IDMSM (PCP) group; the short- and long-dashed line (– – –), the MSM (PCP) group; and the long-dashed line (— — —), the MSM (OTH) group.

4. DISCUSSION

In this article we adopt an approach that combines EM with multiple imputation for properly analyzing survival data when the failure time is truncated and possibly censored and the reporting of such failures is constrained within a chronologic time interval. Another example involving similar data structures in the AIDS research, and thus our approach can be applied to, is the problem of estimating HIV incidence when the time of onset of AIDS is truncated and reporting delay is present. This approach not only accommodates the deficient features of most surveillance data sets of this nature in general and the CDC AIDS surveillance system in particular, but also leads to greatly simplified estimation. Because the simplification is achieved by splitting the problem of joint estimation of survival and delay distributions into separate estimations of each, the resulting estimates may not be as efficient as those based on the joint likelihood. However, this possible disadvantage is offset by the gain in simplicity of computing and the large sample sizes of the AIDS surveillance data set in particular and many other survey and surveillance data of similar magnitude in general.

In the analysis we have treated deaths occurring in the same quarter as diagnosis as having survival times less than 3 months. But a proportion of these individuals may represent a much delayed diagnosis of AIDS at a time that is very close to death rather than rapid disease progression (CDC 1990). It is difficult or even impossible to identify these individuals from the surveillance data alone. In an attempt to investigate this possible bias, we refitted the models after deleting the deaths occurring in the same quarter as diagnosis, there was very little change in the estimates.

An assumption used in modeling the survival distributions is that $F(s^*|\mathbf{z}) = 1$. We examined the deaths from AIDS

diagnoses before 1984 and found that less than 1% of them survived for 8 years. In light of this and results from other AIDS surveillance data and AIDS clinical trials studies, we think that it would be rare for an individual to survive for more than 8 years after an AIDS diagnosis. The estimates presented, therefore, seem to provide good approximations to the unconditional distributions.

It is quite possible that the change in definition of AIDS by the CDC in 1987, which broadened the case definition to include presumptive diagnosis of several conditions, including PCP (CDC 1991), early diagnosis, overall improvement of medical care and treatment, and so on, could all effect the estimates. The results presented suggest that any explanations must accommodate a differential effect on the risk groups, because such a significant increasing trend is not observed for all the risk groups considered.

The survival estimates presented in this article are based on reported deaths rather than on reported AIDS cases using standard survival methodology as in several similar analyses (Harris 1990b; Lemp et al. 1990, 1991). Unlike these other studies, where patients had been closely followed-up after an AIDS diagnosis, a sizable fraction of deaths in the CDC data will never be reported. Thus the standard approach is not appropriate for analyzing the CDC data. Simply censoring reported AIDS cases with no death certificates at the time of analysis will cause severe bias. Figure 7 shows the comparison between the two approaches for the PCP cases in the MSM risk group. The upper two survival curves are based on reported AIDS cases with death censored at two different times, one at the time of analysis and the other 1 year before that time. The bottom curve, which is the same curve in Figure 3 for the same risk group, is based on reported deaths. The survival curves based on the reported AIDS cases show that between 15% and 18% of PCP cases would still be alive at the end of 5 years, a survival rate for PCP much higher than those reported by other studies (Friedland et al. 1991; Lemp et al. 1990). This example explains the upward bias in the estimates reported by Rothenberg et al. (1987) that was first recognized by Lemp et al. (1990). Note that censoring death 1 year before the time of analysis produced a less-biased estimate. Nevertheless, such estimates from reported AIDS cases are always biased, unless we have knowledge to separate those whose deaths will never be reported from those who were still alive at the time of analysis.

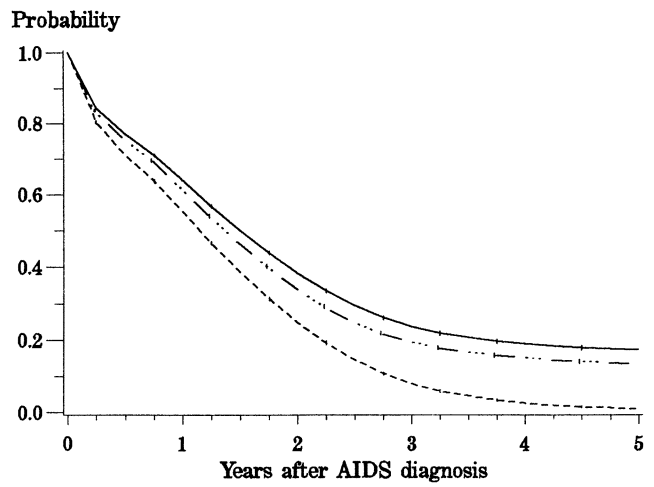

Figure 7. Survival Distributions and 95% Confidence Intervals (Vertical Bars) for Pneumocystis carinii Pneumonia in the MSM Risk Group. Based on reported AIDS cases with death censored at the time of analysis March 1991 (upper curve), reported AIDS cases with death censored after March 1990 (middle curve), and reported deaths (bottom curve).

[*Received May 1991. Revised August 1992.*]

REFERENCES

Bennett, C. L., et al. (1989), "The Relation Between Hospital Experience and In-Hospital Mortality for Patients With AIDS-Related PCP," *Journal of the American Medical Association,* 261, 2975–2979.

Brookmeyer, R., and Liao, J. (1990), "The Analysis of Delays in Disease Reporting: Methods and Results for Acquired Immunodeficiency Syndrome," *American Journal of Epidemiology,* 132, 355–365.

Centers for Disease Control (1990), "Survival After Diagnosis of AIDS," *Morbidity and Mortality Weekly Report,* 39, 25–31.

——— (1991), "AIDS Public Information Data Set."

Cotton, D. J. (1989), "Improving Survival in Acquired Immunodeficiency Syndrome: Is Experience Everything?" *Journal of the American Medical Association,* 261, 3016–3017.

Cox, D. R. (1972), "Regression Models and Life Tables" (with discussion), *Journal of the Royal Statistical Society,* Ser. B, 34, 187–220.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum Likelihood Estimation from Incomplete Data via the EM Algorithm" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 39, 1–38.

Finkelstein, D. M., Moore, D. F., and Schoenfeld, D. A. (in press), "A Proportional Hazards Model for Truncated AIDS Data," *Biometrics.*

Friedland, G. H., et al. (1991), "Survival Differences in Patients With AIDS," *Journal of Acquired Immune Deficiency Syndromes,* 4, 144–153.

Harris, J. E. (1990a), "Reporting Delays and the Incidence of AIDS," *Journal of the American Statistical Association,* 85, 915–924.

——— (1990b), "Improved Short-Term Survival Among AIDS Patients Initially Diagnosed with *Pneumocystis carinii* Pneumonia, 1984 Through 1987," *Journal of the American Medical Association,* 263, 397–401.

Hirschel, B., et al. (1991), "A Controlled Study of Inhaled Pentamidine for Primary Prevention of *Pneumocystis carinii* Pneumonia," *The New England Journal of Medicine,* 324, 1079–1083.

Kalbfleisch, J. D., and Lawless, J. F. (1989), "Inference-Based Retrospective Ascertainment: An Analysis of the Data on Transfusion-Related AIDS," *Journal of the American Statistical Association,* 84, 360–372.

——— (1991), "Regression Models for Right-Truncated Data With Application to AIDS Incubation Times and Reporting Lags," *Statistica Sinica,* 1, 19–32.

Kaplan, E. L., and Meier, P. (1958), "Nonparametric Estimation for Incomplete Observations," *Journal of the American Statistical Association,* 53, 457–481.

Lagakos, S. W., Baraj, L. M., and DeGruttola, V. (1988), "Nonparametric Analysis of Truncated Survival Data, With Application to AIDS," *Biometrika,* 75, 515–523.

Lemp, G. F., Hirozawa, A. M., Araneta, M. R., Young, K., and Nieri, G. (1991), "Improved Survival for Persons With AIDS in San Francisco," *VII International Conference on AIDS, Florence 1991, Abstract Book: Vol. 1;* TU.C. 41.

Lemp, G. F., Payne, S. F., Neal, D., Temelso, T., and Rutherford, G. W. (1990), "Survival Trends for Patients With AIDS," *Journal of the American Medical Association,* 263, 402–406.

Leoung, G. S., et al. (1990), "Aerosolized Pentamidine for Prophylaxis Against *Pneumocystis carinii* Pneumonia," *The New England Journal of Medicine,* 323, 769–775.

Li, K. H., Raghunathan, T. E., and Rubin, D. B. (1991), "Large Sample Significance Levels From Multiply Imputed Data Using Moment-Based Statistics and an *F* Reference Distribution," *Journal of the American Statistical Association,* 86, 1056–1073.

Little, R. J. A., and Rubin, D. B. (1987), *Statistical Analysis With Missing Data,* New York: John Wiley.

Louis, T. A. (1982), "Finding the Observed Information Matrix When Using the EM Algorithm," *Journal of the Royal Statistical Society,* Ser. B, 44, 226–233.

McCullach, P., and Nelder, J. A. (1989), *Generalized Linear Models* (2nd ed.), London: Chapman and Hall.

Meng, X. L., and Rubin, D. B. (1991), "Using EM to Obtain Asymptotic Variance-Covariance Matrices: The SEM Algorithm," *Journal of the American Statistical Association,* 86, 899–909.

——— (1992), "Performing Likelihood Ratio Tests With Multiply Imputed Data Sets," *Biometrika,* 79, 103–111.

Pagano, M., DeGruttola, V., MaWhinney, S., and Tu, X. M. (1992), "The HIV Epidemic in New York City: Statistical Methods for Projecting AIDS Incidence and Prevalence," in *Statistical Methodology for the Study of the AIDS Epidemic,* eds. K. Dietz, V. Farewell, and N. P. Jewell, Boston: Birkhäuser-Boston, pp. 123–140.

Prentice, R., and Gloeckler, L. (1978), "Regression Analysis of Grouped Survival Data With Application to Breast Cancer Data," *Biometrics,* 34, 57–67.

Rothenberg, R., et al. (1987), "Survival With the Acquired Immunodeficiency Syndrome," *The New England Journal of Medicine,* 317, 1297–1302.

Rubin, D. B. (1976), "Inference and Missing Data," *Biometrika,* 63, 581–592.

——— (1987a), *Multiple Imputation for Nonresponse in Surveys,* New York: John Wiley.

——— (1987b), Comment on "The Calculation of Posterior Distributions by Data Augmentation," by M. A. Tanner and W. H. Wong, *Journal of the American Statistical Association,* 82, 543–546.

Rubin, D. B., and Schenker N. (1991), "Multiple Imputation in the Health Care Databases: An Overview and Some Applications," *Statistics in Medicine,* 10, 585–598.

SAS Institute (1985), *SAS/IMSL Guide, Version 6,* Cary, NC.

Turnbull, B. W. (1976), "The Empirical Distribution Function With Arbitrarily Grouped, Censored, and Truncated Data," *Journal of the Royal Statistical Society,* Ser. B, 38, 290–295.

Volberding, P. A., et al. (1990), "Zidovudine in Asymptomatic Human Immunodeficiency Virus Infection: A Controlled Trial in Persons With Fewer Than 500 CD4-Positive Cells per Cubic Millimeter," *The New England Journal of Medicine,* 322, 941–949.

Wang, M.-C. (1992), "The Analysis of Retrospectively Ascertained Data in the Presence of Reporting Delays," *Journal of the American Statistical Association,* 87, 397–406.

Wang, M.-C., Jewell, N. P., and Tsai, W. Y. (1986), "Asymptotic Properties of the Product Limit Estimate Under Random Truncation," *The Annals of Statistics,* 14, 1597–1605.

Zeger, S. L., See, L., and Diggle, P. J. (1989), "Statistical Methods for Monitoring the AIDS Epidemic," *Statistics in Medicine,* 8, 3–22.