Inference Based on Retrospective Ascertainment: An Analysis of the Data on Transfusion-Related AIDS

Author(s): J. D. Kalbfleisch and J. F. Lawless

Source: *Journal of the American Statistical Association*, Jun., 1989, Vol. 84, No. 406 (Jun., 1989), pp. 360-372

Published by: Taylor & Francis, Ltd. on behalf of the American Statistical Association

Stable URL: http://www.jstor.com/stable/2289919

Inference Based on Retrospective Ascertainment: An Analysis of the Data on Transfusion-Related AIDS

J. D. KALBFLEISCH and J. F. LAWLESS*

In some epidemiologic studies, identification of individuals for study is dependent on the occurrence of some event. Once an individual is identified, the time of a previous event, termed an *initiating event,* is determined retrospectively. This article considers problems of estimation when initiating events occur as a nonhomogeneous Poisson process, and the time s from the initiating event to the final event has pdf $f(s)$ independent of the time of the initiating event. A simple form for the likelihood function is obtained and methods of parametric and nonparametric estimation are developed and considered. In particular, the model is related to a Poisson process in the plane, and for the parametric case simple algorithms are developed for parameter estimation. Regression models are also considered as well as various generalizations of the basic problem. Parallel to the theoretical development, data on patients diagnosed with acquired immune deficiency syndrome (AIDS) are considered and a detailed analysis is given. The data report the dates of diagnosis with AIDS and infection with human immunodeficiency virus, for patients reported to the Centers of Disease Control in Atlanta, Georgia, and thought to be infected by blood or blood-product transfusion. The analysis of these data was considered by Medley, Anderson, Cox, and Billard (1987), Lui et al. (1986), and others. It is shown that nonparametric analysis leads to simple estimates of certain parameters and indicates clearly the nature of an identifiability problem that arises with data of this kind. Problems arise in the estimation of the total number of infectives or percentiles of the distribution of the induction period, s.

KEY WORDS: Acquired immune deficiency syndrome; Human immunodeficiency virus; Infection by transfusion; Nonparametric estimation; Poisson process in the plane; Retrospective studies.

1. INTRODUCTION

We consider situations in which an individual is ascertained for study from some population after some event occurs; for that individual, the time of an initiating event is determined retrospectively. It is assumed that in the population under investigation initiating events occur according to a nonhomogeneous Poisson process, although certain methods remain valid when initiating events follow other types of point process. Primary questions concern the rate of the nonhomogeneous process and the distribution of the time from the initiating event to the ascertaining event, called the *induction time.* Retrospective ascertainment is common in actuarial, demographic, epidemiologic, and other studies. The specific problems we consider concern studies of acquired immune deficiency syndrome (AIDS), in which the ascertaining event is the diagnosis of AIDS in an individual, and the initiating event is the infection of the individual with the human immunodeficiency virus (HIV), which is assumed to cause AIDS.

We consider nonparametric procedures as well as estimation based on parametric assumptions, and address questions of identifiability and estimability in both contexts.

The methods are illustrated in the analysis of data made available to us by T. A. Peterman of the Centers for Disease Control (CDC) in Atlanta, Georgia. For patients who were thought to be infected with HIV by blood or blood-product transfusion (not including hemophiliacs), the data record the sex and age of the patient, the date of reporting AIDS, the date of diagnosis, and (when it could be determined) the date of infection. Primary interest here centers on the process of infection, the distribution of the induction or "incubation" time (the time from HIV infection to the clinical manifestation of full-blown AIDS) and questions concerning the dependence of the induction period on covariates such as sex and age.

The data include 494 cases reported to the CDC prior to January 1, 1987, and diagnosed prior to July 1, 1986. Of the 494 cases, 295 had consistent data, and the infection could be attributed to a single transfusion or short series of transfusions. Our analyses are restricted to this subset, the data for which are reported in Table 1. Methods for handling uncertain time of infection are, however, presented in Section 6. Strictly speaking, in these data ascertainment is by reporting to the CDC; the times of diagnosis and infection are then recorded. For most analyses we suppose that the ascertainment is by diagnosis; the six-month lag from July 1, 1986, to January 1, 1987, assures that most diagnosed cases up to July 1, 1986, are in the data. In Section 6 models that adjust the analyses to take account of the reporting lag are discussed.

Analyses of these data, or earlier versions of them, have been considered by several authors, and estimates of the induction time distribution and the number of infected individuals based on them have been widely reported. Lui et al. (1986) postulated a Weibull distribution for the induction time and constructed a truncated likelihood for

* J. D. Kalbfleisch is Professor and Chairman and J. F. Lawless is Professor, Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada. The authors are grateful to David Cash for helpful discussion, and for his assistance with the computations. This research was supported by grants from the Natural Sciences and Engineering Research Council of Canada to both authors, and National Institute of Drug Abuse Grant 1-R01-DA04722, coordinated by the Societal Institute of the Mathematical Sciences.

Table 1. The CDC Transfusion Data on 295 Patients Used in the Analyses

INF	DIAG	AGE	INF	DIAG	AGE	INF	DIAG	AGE	INF	DIAG	AGE	INF	DIAG	AGE	INF	DIAG	AGE
23	27	4	36	38	56	41	42	62	46	42	67	65	29	58	60	38	68
38	14	2	12	62	58	35	48	26	68	20	60	61	33	66	41	57	61
38	15	56	50	24	52	67	16	59	76	12	29	84	10	61	53	46	37
36	18	65	45	29	62	35	48	51	64	24	62	41	53	58	48	51	54
27	28	57	33	41	67	71	12	71	50	38	69	83	11	4	36	63	67
45	10	1	55	19	68	67	16	72	59	29	67	58	36	21	72	27	70
23	34	20	37	37	34	36	47	22	80	8	1	45	49	77	60	39	55
48	10	1	46	28	69	69	14	61	46	43	66	83	12	55	36	63	57
25	34	46	62	13	2	45	38	78	78	11	61	80	15	2	48	51	27
42	17	46	57	18	61	76	8	1	26	63	62	84	11	55	55	44	51
33	29	53	34	41	32	66	18	67	62	27	80	59	36	65	58	41	70
45	17	39	56	19	70	22	62	70	19	70	17	85	10	1	60	39	35
33	29	54	57	18	23	74	10	42	22	67	59	59	36	53	89	10	1
39	23	2	64	11	1	49	35	62	26	63	48	91*	4	1	37	63	67
34	29	34	57	18	64	42	42	73	76	13	66	65	31	51	82	18	60
48	15	63	42	33	62	15	69	30	27	62	65	47	49	51	41	59	65
50	13	2	54	21	3	29	55	54	62	27	4	64	32	3	68	32	71
35	29	62	28	48	42	62	22	38	75	15	68	73	23	3	71	29	44
26	38	56	36	40	66	65	20	2	76	14	70	81	15	36	63	37	63
43	21	67	66	10	32	75	10	81	57	33	51	79	17	67	49	51	5
3	61	29	19	58	69	65	20	67	58	32	56	36	60	66	37	63	59
53	12	2	45	32	60	69	16	34	59	32	32	67	29	72	60	40	49
40	25	61	53	24	44	62	23	54	8	83	66	87	9	57	78	22	41
53	12	46	67	11	68	56	30	73	41	50	46	85	11	59	52	48	56
28	38	46	68	10	67	52	34	73	70	21	53	29	67	69	60	40	73
34	32	26	54	25	58	82	4	46	57	34	78	72	24	76	85	15	73
42	24	68	40	39	50	46	40	64	58	33	33	77	19	77	68	33	71
66	0	63	60	19	70	75	11	59	63	29	78	72	24	44	37	64	49
19	48	61	54	25	50	27	59	60	37	55	80	67	29	65	39	62	69
21	46	30	42	37	82	56	30	26	75	17	52	53	43	46	22	79	33
37	30	25	74	5	39	70	16	62	58	34	65	54	43	47	12	89	38
33	34	51	71	8	1	49	37	4	39	53	37	17	80	54	63	38	85
31	37	4	43	37	76	66	20	67	38	54	68	61	36	28	80	21	62
48	21	70	61	20	68	73	13	54	73	19	81	65	32	68	45	56	64
32	37	62	68	13	63	76	10	50	72	20	77	48	49	60	47	54	53
43	26	61	52	29	24	43	43	62	41	52	72	57	40	66	85	16	38
17	53	33	75	6	68	50	36	46	56	37	66	29	68	60	29	72	73
64	6	2	46	35	65	41	46	68	50	43	4	36	61	63	60	41	74
58	13	39	62	19	58	49	38	59	79	14	71	30	68	29	84	17	49
67	4	1	67	14	45	68	19	74	83	10	38	45	53	71	70	31	71
49	22	57	48	33	3	39	48	64	76	17	2	40	58	28	61	40	72
67	4	29	68	13	2	67	20	2	29	64	67	75	23	41	69	32	3
35	37	57	58	23	64	61	26	42	17	76	78	43	55	59	77	24	70
12	60	21	55	27	68	82	5	84	69	25	65	76	22	68	63	38	69
19	53	52	70	12	63	69	18	2	86	8	1	66	32	65	97*	4	36
60	13	69	33	49	71	65	23	66	29	65	53	86	12	36			
53	20	78	56	26	71	56	32	66	65	29	72	57	41	11			
56	17	66	47	35	23	59	29	52	75	19	61	75	23	65			
65	8	73	64	18	35	57	31	41	74	20	63	71	27	68			
53	20	56	15	68	6	57	31	69	42	52	64	51	47	65			

NOTE: INF is the month of infection with 1 = January 1978; DIAG is the duration of the induction period in months; and AGE is the age + 1 in years at the time of transfusion. For the continuous analyses, the time of infection was taken to be INF − .5, and the time of induction was as given in the table, except that 0 was replaced with .5.

* These cases had infection after 90 months and are excluded from the analyses of Sections 4 and 5.

the estimation of the parameters of that distribution. Brookmeyer and Gail (1988) constructed a conditional likelihood that enables estimation of some parameters in the infection process as well as those of a parametric model postulated for the induction period. Our work is in large part stimulated by that of Medley, Anderson, Cox, and Billard (1987) and Medley, Billard, Cox, and Anderson (1988), who postulated a parameterized Poisson process for infections by transfusion and parametric models for the induction time. They gave maximum likelihood estimates (MLE's) of the infection rate, the parameters of the induction distribution under Weibull and gamma models, and various derived quantities of interest, but they did not discuss the precision of their estimates.

This article has two general aims. First, we give some statistical models and methodology for the analysis of data collected by retrospective ascertainment, and indicate what these data can and cannot estimate. Second, we provide a detailed analysis of the transfusion data to illustrate the methodology and indicate the nature of conclusions that are warranted on these data. Special problems associated with the transfusion data are discussed in some detail. Accordingly, in Sections 3.1 and 4.1 we describe methodologic development in the context of a general problem involving retrospective ascertainment. In Sections 3.2 and 4.2 we illustrate the methods on the transfusion data, and discuss the results.

Section 2 discusses likelihood construction and illus-

trates the strong connection between the various methods of analysis used. In Section 3, we explore nonparametric methods of estimating both the intensity of the process of initiating events and the distribution of the induction time, and relate this discussion to recent work by Lagakos, Barraj, and De Gruttola (1988). Section 4 deals with parametric estimation and develops efficient algorithms for fitting parametric models in the general context of a Poisson process in the plane. In both Sections 3 and 4, questions of estimability and identifiability play a major role. Section 5 considers methods for incorporating covariates in the distribution of induction time. Section 6 considers more general problems, such as when the times of more than one event are determined retrospectively, when the time of the initiating event is uncertain, and when mortality is incorporated as a competing risk. In Section 7, we identify and discuss several problems that require further study.

2. LIKELIHOOD CONSTRUCTION

We suppose the following:

Assumption 1. Initiating events occur according to a Poisson process of intensity $h^*(x)$, $x > -\infty$.

Assumption 2. Independent of the time x of the initiating event, the duration of the induction period s has subdistribution function $F^*(s)$ and subdensity function $f^*(s) = dF^*(s)/ds$.

Assumption 3. Observation is over the period $(-\infty, T]$, and an individual is observed only if the time of the second event, $t = x + s$, is less than or equal to T.

In this formulation, we allow $p = F^*(\infty) \leq 1$ so that not all individuals who experience the first event necessarily experience the second, as would be the case, for example, if individuals were subject to removal because of competing risks. Note that even if $T = \infty$, there is no information in the data on the proportion $1 - p$ of individuals who experience the initiating event but die without experiencing the second event. It is therefore convenient to restrict attention to the following:

1. Initiating events that eventually result in the occurrence of the second event. This is a (screened) Poisson process of intensity $h(x) = h^*(x)p$.
2. The cdf $F(s) = F^*(s)/p$ and pdf $f(s) = f^*(s)/p$ of the induction period s of an individual who experiences the second event.

With $T = \infty$, both $h(x)$ and $F(s)$ are directly estimable.

Let (x_i, t_i) $(i = 1, \ldots, n)$ be the observed data, where for the ith individual x_i is the time of the initiating event and t_i is the time of the second event. Let $s_i = t_i - x_i$ be the induction time. From Assumptions 1–3, it follows that the process of observed initiating events is a Poisson process of intensity $h(x)F(T - x)$, because an infection at time x leads to observation iff the second event follows after an induction time of at most $T - x$. Therefore, it follows that the number of observed events N has the Poisson probability function

$$\Pr\{N = n\} = A^n e^{-A}/n!, \qquad n = 0, 1, 2, \ldots, \quad (1)$$

where $A = \int_0^T h(x)F(T - x)\,dx$. Further, given $N = n$, it is well known that the ordered times of initiating events have the distribution of the order statistic in a sample of size n from the density $h(x)F(T - x)/A$, or

$$f(x_1, \ldots, x_n \mid N = n) = n! \prod_{i=1}^{n} \frac{h(x_i)F(T - x_i)}{A}. \quad (2)$$

Finally, given $x_1, \ldots, x_n$, the distribution of $t_1, \ldots, t_n$ has the truncated density

$$\prod_{i=1}^{n} \{f(t_i - x_i)/F(T - x_i)\}, \quad (3)$$

which arises because the individual is observed only if the induction time is less than $T - x_i$. The full likelihood is therefore the product of (1), (2), and (3):

$$L = \prod_{i=1}^{n} \{h(x_i)f(s_i)\}\exp(-A), \quad (4)$$

where, as before,

$$A = \int_0^T \int_0^{T-x} h(x)f(s)\,ds\,dx. \quad (5)$$

In subsequent sections, (4) is used for estimation. The model specifies a Poisson process in the (x, t) plane with intensity $h(x)f(t - x) = h^*(x)f^*(t - x)$ at the point (x, t) $(0 \leq x \leq t < \infty)$ and 0 elsewhere. The sampling scheme is such that all events in the triangular region $0 \leq x \leq t \leq T$ are observed; the likelihood (4) is appropriate in this instance.

This likelihood was obtained and used by Medley et al. (1987, 1988), and is related to other estimation procedures used in the literature in the analysis of the transfusion data. The conditional likelihood based on the distribution of the data, given $N = n$, is the ratio of (4) to (1),

$$L_c^{(1)} = n!\Pi h(x_i)f(s_i)/A^n, \quad (6)$$

which was used for estimation by Brookmeyer and Gail (1988). The conditional likelihood, given $N = n$ and $x_1, \ldots, x_n$, is proportional to (3); it is the likelihood arising from the truncated distribution of the induction times $s_1, \ldots, s_n$, and was used for inference by Lui et al. (1986) and Lagakos et al. (1988).

Parametric estimation based on (4) is discussed in detail in Section 4. We note, however, that (4) and (6) have a close relationship. For example, suppose that $h(x) = \alpha h(x; \lambda)$ and $f(s) = f(s; \theta)$ is a parametric model where $\alpha > 0$ and λ, θ are vectors of parameters. It is then easily seen that the likelihood (4) involves all parameters, whereas in (6) α is eliminated. Under conditions of differentiability and as established in Appendix A, (4) and (6) lead to both identical MLE's of θ and λ and identical observed and estimated Fisher information matrices. This is potentially useful in estimation based on a likelihood of the form (6), since (4) is simpler. It also demonstrates that the Poisson-based methodology we present for getting estimates and standard errors of θ and λ remains valid when initiating events follow more general point processes, particularly ones that have the order-statistic property (e.g., see Deff-

ner and Haeusler 1985). Brookmeyer and Gail (1988, app. A) assumed that initiating events follow such a process in their development of (6). Note that (3) uses only the independence assumption of Assumption 2 and is valid irrespective of the process generating the initiating events. When (4) or (6) is appropriate, however, (3) generally gives a different estimate of θ and is typically much less precise. Brookmeyer and Gail explored this point with regard to (3) and (6). The parametric form assumed for $h(x)$ generally contributes information that allows more precise estimation of θ.

The Poisson-based nonparametric methods in Section 3 also have a broader validity. Moreover, we show that the likelihoods (3), (4), and (6) in fact yield the same estimates of $F(s)$. This makes it clear that the parametric assumptions are responsible for the more precise estimation of θ from (4) or (6) than from (3).

3. NONPARAMETRIC ESTIMATION

3.1 General Comments

Although the results obtained in this section can also be derived as nonparametric MLE's based on observations from the continuous model of the last section, it is convenient to consider a discrete time version of the model on which (4) is based. Accordingly, suppose that x and s can take only integer values 0, 1, 2, . . . , and suppose that all individuals with $x + s \le T$ are observed. The discrete version of the likelihood (4) is

$$L = \prod_{i+j\le T} (h_i f_j)^{n_{ij}} \exp\left(-\sum_{i+j\le T} h_i f_j\right) = \prod_{i=0}^{T} h_i^{a_i} \prod_{j=0}^{T} f_j^{b_j} \exp\left(-\sum_{i+j\le T} h_i f_j\right), \quad (7)$$

where n_{ij}, the number of individuals with an initiating event at time i and terminal event at time $i + j$, is assumed to have a Poisson distribution with mean $h_i f_j$ $(i + j \le T)$, independently. Here, h_i is the mean number of infections at time i that ultimately result in the terminal event, f_j is the probability that the induction period is of duration j, $a_i = \sum_{j\le T-i} n_{ij}$, and $b_j = \sum_{i\le T-j} n_{ij}$. We consider maximization of (7) with respect to the $2T + 2$ parameters $\{h_i\}$ and $\{f_j\}$.

The likelihood (7) arises from Poisson sampling in the context of triangular incomplete contingency tables under a model of quasi independence (e.g., see Bishop, Fienberg, and Holland 1975, sec. 5.2). Goodman (1968, pp. 1112–1114) has shown that for such tables, estimates can be obtained in closed form; Equations (11) and (12) that follow are equivalent to ones given there.

Setting derivatives of log L from (7) equal to 0 gives

$$\frac{\partial \log L}{\partial h_i} = \frac{a_i}{h_i} - F_{T-i} = 0, \qquad i = 0, 1, \ldots, T, \quad (8)$$

and

$$\frac{\partial \log L}{\partial f_j} = \frac{b_j}{f_j} - H_{T-j} = 0, \qquad j = 0, 1, \ldots, T, \quad (9)$$

where $F_r = f_0 + \cdots + f_r$ and $H_r = h_0 + \cdots + h_r$. It is apparent from (7) or (8) and (9) that $\{h_i\}$ and $\{f_j\}$ are identifiable only up to a multiplicative constant: If sets of values $\{\hat{h}_i\}$ and $\{\hat{f}_j\}$ are a solution to (8) and (9) that satisfies $\hat{F}_T = \sum \hat{f}_j \le 1$, then $\{c^{-1}\hat{h}_i\}$ and $\{c\hat{f}_j\}$, where c is any constant such that $0 < c\hat{F}_T \le 1$, is also an admissible solution. The data are therefore sufficient to estimate the shape of the continuous cumulative intensity function $H(x) = \int_0^x h(u)\,du$ or the cdf $F(s)$ only; neither $H(x)$ nor $F(s)$ can be individually estimated. The truncated cdf $F(s)/F(T)$ $(0 \le s \le T)$ is estimable, as are $H(x)/H(T)$ $(0 \le x \le T)$ and $F(s)H(x)$. Nevertheless, even small percentage points of $F(\cdot)$ cannot be estimated with data of this type. Lagakos et al. (1988) made the same point.

Since only conditional probabilities can be estimated, it is convenient to define

$$g_j = f_j/F_T, \quad G_j = g_0 + \cdots + g_j = F_j/F_T, \qquad j = 0, 1, \ldots, T,$$

and to obtain the MLE's of these. In addition, we let

$$g_j^* = f_j/F_j = \Pr\{s = j \mid s \le j\}, \qquad j = 0, \ldots, T. \quad (10)$$

It is easily seen, by substitution of h_i from (8) into (9), that

$$\hat{g}_j^* = b_j \Big/ \left(\sum_{i=0}^{T-j} a_i - \sum_{l=j+1}^{T} b_l\right), \quad (11)$$

and hence we find that

$$\hat{G}_T = 1, \quad \hat{G}_{j-1} = (1 - \hat{g}_j^*)\hat{G}_j, \qquad j = T, T-1, \ldots, 1. \quad (12)$$

Finally, $\hat{h}_i = a_i/\hat{G}_{T-i}$ $(i = 0, \ldots, T)$ is obtained from (9), where it is understood that $\hat{h}_i = 0$ if $a_i = \hat{G}_{T-i} = 0$, and $\hat{h}_i = \infty$ if $a_i > 0$ and $\hat{G}_{T-i} = 0$. The estimate (11) of the conditional probability (10) arises as the ratio of the number of observed induction times equal to j to the number of induction times that are less than or equal to j but with the time of the initiating event at $T - j$ or earlier. Figure 1 gives an illustration.

The estimates (11) and (12) are identical to those obtained by Lagakos et al. (1988) by another approach. Their analysis is based on the truncated likelihood (3), and relates the estimation problem to one of left truncation in a reversed time process. In reverse time, (10) can be interpreted as a discrete hazard function and the estimate (12) is of the product-limit type. The relationship between nonparametric estimation based on the full likelihood (4) and the truncated likelihood (3) is further discussed in Appendix B.

Note that if $\hat{g}_j^* = 1$ for some j, then $\hat{G}_{j-1} = \cdots = \hat{G}_0 = 0$ even though some of the corresponding b_l's might be positive. If $b_l > 0$, then $\hat{g}_l = 0$ and $\hat{h}_l = \infty$; the maximum likelihood estimate in this case arises as a supremum of the likelihood corresponding to a limit as $h_l \to \infty$ and $g_l \to 0$, but with the expected frequency $h_l g_l$ approaching a positive constant. It is important to note that the estimate

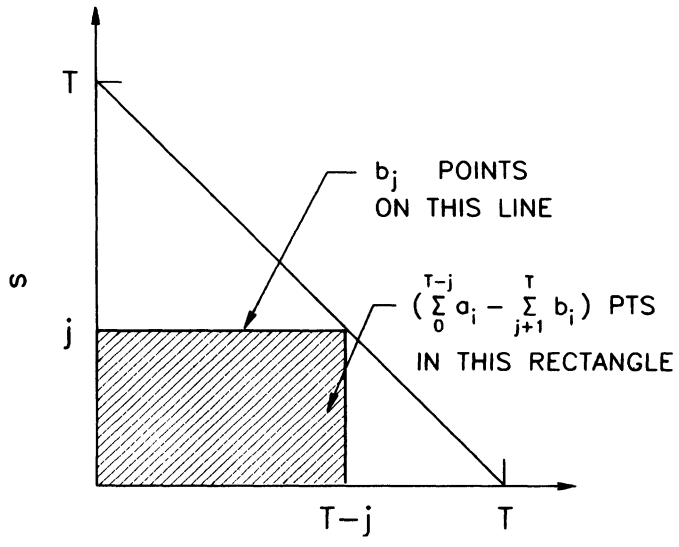

Figure 1. Display Showing the Genesis of the Estimate $\hat{g}_j^*$ of (11). Note that there are b_j points on the line $s = j$, which forms the top boundary of the shaded rectangle, and $(\sum_{l=0}^{T-j} a_l - \sum_{l=j+1}^{T} b_l)$ points in the entire rectangle.

of G_j can be very imprecise, especially for large values of j. As Figure 1 makes clear, the risk set in the denominator of (11) is small when j is large, and as a consequence the corresponding $\hat{g}_j^*$ can be quite variable. Of course, it is possible to estimate F_j/F_r for $j \le r < T$ instead of G_j to avoid the small risk sets associated with later times.

3.2 Nonparametric Estimation in the Transfusion Data

The transfusion data are summarized by year of transfusion and year of diagnosis in Table 2. The year 1986 extends only from January 1 to June 30. The original data, given in Table 1, provide the months in which the events occurred, and the analyses were based on these. Thus, in the notation of Section 3.1, i ranges over the months of possible transfusion, with 0 representing January 1978 and $T = 101$ representing June 1986. Note that i corresponds to INF − 1 in Table 1. Similarly, the possible induction times are indexed by $j = 0, 1, 2, \ldots, 101$.

Estimation was done separately for each of the age groups considered by Medley et al. (1987), specifically, "children" aged 0–4 years, "adults" aged 5–59 years, and "elderly patients" aged 60 and over at the time of transfusion. As Medley et al. noted, this division is sensible based on current theories suggesting that immuno competence is low for very young patients and for elderly patients. Of particular interest are the functions $H(x) = \int_0^x h(u)\,du$, which represent the expected number of infective transfusions administered prior to time x that will ultimately result in a diagnosis of AIDS, and $F(s)$, the cdf of the induction time distribution.

Figures 2 and 3 give the nonparametric MLE's $\hat{H}(t) = \sum_{i \le 12t} \hat{h}_i$ and $\hat{F}(t) = \sum_{j \le 12t} \hat{f}_j$, respectively, for each of the three age groups, where time t is measured in years. The estimates are step functions similar to the empirical cdf, but here they are smoothed by connecting them at six-month intervals with straight lines. In view of the identifiability problem discussed previously and the fact that there are no induction times over 7.5 years, we scale the estimates of $F(t)$ and $H(t)$ so that for each age group the probability that the induction period is less than or equal to 7.5 years is an unspecified constant $c = F(7.5)$.

As discussed in Kalbfleisch and Lawless (1988), estimates of $F(t)$ for young children (displayed in Fig. 3) suggest that the whole induction time distribution may have been seen; in what follows, we choose $c = 1$. Although there are no long induction times observed in the youngest age group, times that exceed 7.5 years may be seen with additional follow-up. For the other two age groups, nothing in the graph suggests a value of c, and further

Table 2. Observed and Expected Frequencies Based on the Nonparametric Estimates From (9) and (10)

Year of diagnosis	Year of transfusion									
	1978	1979	1980	1981	1982	1983	1984	1985	1986	Total
1978	0									0
	.00									.00
1979	0	0								0
	.05	.06								.11
1980	0	0	0							0
	.09	.88	.13							1.10
1981	0	0	0	0						0
	.12	1.24	1.56	.72						3.64
1982	0	2	2	5	1					10
	.19	2.06	2.68	4.86	1.84					11.64
1983	1	3	10	7	7	4				32
	.39	2.46	4.38	8.47	8.92	2.12				26.74
1984	0	5	7	15	18	20	2			67
	.29	4.15	5.68	14.08	14.54	17.81	2.47			59.02
1985	1	5	7	15	24	23	25	5		105
	.88	4.57	9.63	16.32	23.75	28.36	22.61	3.79		109.91
1986	0	3	4	16	13	23	14	7	1	81
	.00	2.59	5.93	13.55	13.95	21.70	15.92	8.21	1.00	82.85
Total	2	18	30	58	63	70	41	12	1	295
	2.0	18.0	30.0	58.0	63.0	70.0	41.0	12.0	1.0	

NOTE: The data are grouped by year, although the analyses were done using monthly data. 1986 denotes January 1–June 30.

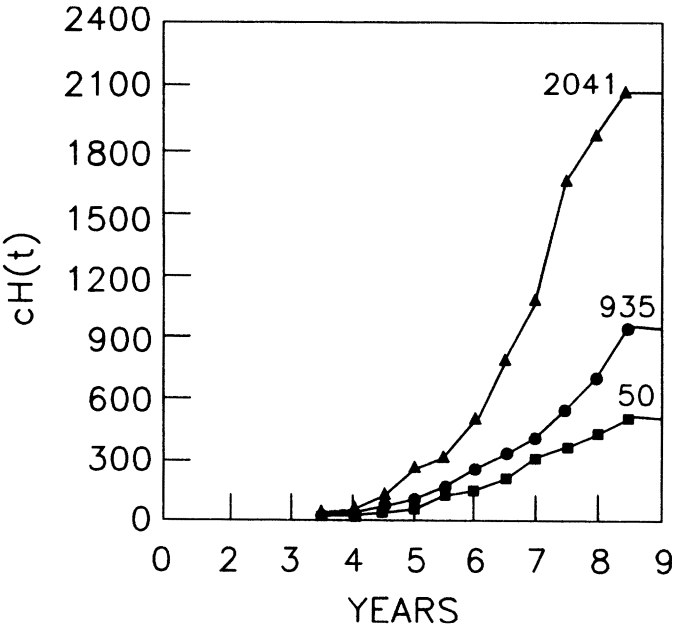

Figure 2. Nonparametric Estimates of H(t): the (unadjusted) Cumulative Number of Infections by Blood Transfusion for ■, Children; ●, Adults; and ▲, Elderly Patients. The curve for children has been scaled up by a factor of 10. Time is measured from January 1, 1978, so that 7.5 represents the initiation of screening of the blood supply in July 1985. The constant c is the probability that the induction time is less than 7.5 years. Note that c is not estimable on the transfusion data alone, and that different values of c may apply to the separate age groups.

progress in estimating the early percentiles of the induction distribution or the total number of infections requires external information to estimate c.

One possible approach to supplementing the data is to use cohort studies to obtain an estimate of, say, $c = F(7.5)$ or F at some earlier time. For example (as in Kalbfleisch and Lawless, in press), data from the San Francisco City Clinic cohort suggest that about 36% of HIV-positive individuals in that cohort are diagnosed with AIDS within approximately 7.5 years of infection. We actually seek an

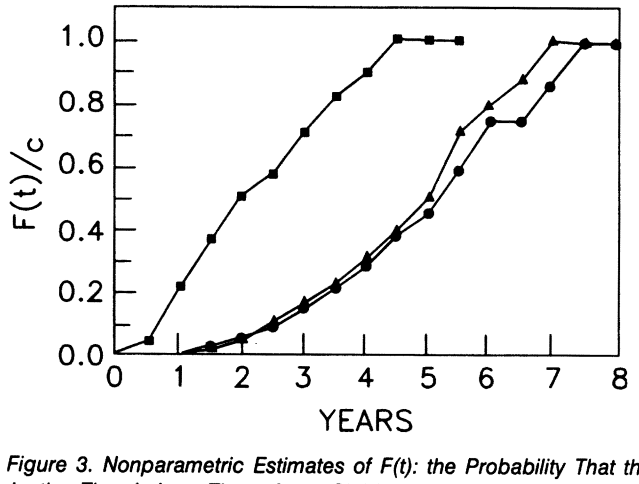

Figure 3. Nonparametric Estimates of F(t): the Probability That the Induction Time Is Less Than t for ■, Children; ●, Adults; and ▲, Elderly Patients. The unspecified constant c is the probability that the induction time is less than 7.5. Note that c is not estimable on the transfusion data alone, and that different values of c may apply to the separate age groups.

estimate of the proportion of seropositives who eventually develop AIDS and are diagnosed with an induction time of 7.5 years or less. Clearly, 36% is most likely an underestimate of this proportion within the San Francisco City Clinic cohort, allowing that every seropositive individual in the cohort will not develop AIDS. With high rates of mortality among transfusion recipients, 36% may be a considerable underestimate of c, especially in the elderly group, where competing mortality rates are high (e.g., see Peterman, Lui, Lawrence, and Allen 1987). It is also questionable whether the natural history of the disease is the same among those infected by transfusion as among those infected primarily by sexual transmission (as in the San Francisco cohort). If c is taken to be 1 for each age group, we obtain what might be considered a lower point estimate of $(2{,}041 + 935 + 50) \times 494 \div 295 = 5{,}900$ for the eventual number of diagnoses of AIDS where the cause is infection by blood transfusion prior to July 1, 1985. On the other hand, if $c = .36$ is assumed for the adult and elderly groups, the corresponding estimate of 13,930 is probably somewhat high. The suitability of $c = .36$ as an estimate in this context is discussed further in Section 6.3.

The expected frequencies (i.e., numbers of cases observed) that are generated by the nonparametric MLE's in Figures 2 and 3 are given in Table 2. This involves an amalgamation across the three age groups, and the overall fit appears good. The fitted model, however, leads to an expected frequency of 4.95 for the number of diagnoses in 1978–1981, whereas none is observed. It is possible that this is due to difficulties with diagnosis prior to 1982. Medley et al. (1987) suggested that there is some evidence of an improving probability of diagnosis in these data, and the fit to this part of the data is consistent with that claim.

Another way to portray the data and fitted model is with rows denoting the induction duration; note that in this case if we used the actual monthly figures, observed and estimated row and column totals would be equal.

It is important to stress that these same expected frequencies arise whatever values of $c = F(7.5)$ are chosen in the three groups. The fact that the expected frequencies are unchanged by varying c illustrates clearly that the transfusion data themselves are unable to discriminate between very high infection rates and long incubation times on the one hand, and low infection rates and short incubation times on the other. As $T \to \infty$, $F(T) \to 1$, so the identifiability problem will eventually disappear with longer follow-up.

3.3 Some Additional Comments on Nonparametric Estimation

Before discussing full parametric estimation in the next section, we remark that it is possible to introduce parametric assumptions about either $h(x)$ or $f(s)$. For example, an inspection of the nonparametric estimates for $\int_0^t h(x)\,dx$ suggests that an exponential form for the intensity function of initiating events is descriptive of the data. If it were supposed that the mean number of initiating events during

the ith month is ab^i ($i = 0, 1, \ldots, T$), the likelihood (7) becomes

$$a^n b^{n\bar{x}} \prod_{j=0}^{T} f_j^{b^j} \exp\left\{-\sum_{i=0}^{T}\sum_{j=0}^{T-i} ab^i f_j\right\}. \qquad (13)$$

It is a simple matter to construct algorithms that estimate a, b, and $\{f_j\}$. The nonidentifiability problems carry through to this case as well: The likelihood (13) is sufficient only to estimate b and $\{af_j\}$.

The estimates of $F(t)$ arising from (13) are given in Figure 4, again adjusted so that $F(7.5) = c$. The estimates of a and b^{12} for the children, adults, and elderly, respectively, are (.047/c, 1.69), (.389/c, 1.89), and (.195/c, 2.21). The estimates of b do not vary much across age groups, which is consistent with the reasonable hypothesis that they should be the same; if transfusion rates are constant in time for each of the age groups, the intensity associated with infectious transfusions should, for each group, be proportional to the prevalence of infected blood among the donor population. Here, properties of the estimators require some evaluation. Evidently, however, the estimate of b in each group is consistent with a one-year doubling time in the number of infections, for which $b^{12} \cong 2$. This is consistent with estimates of May and Anderson (1987), Medley et al. (1987), Lui et al. (1986), and others, concerning the early part of the HIV epidemic. The shapes of the estimates $\hat{F}(t)/c$ in Figures 3 and 4 agree closely for the two younger age groups, but differ to some degree for the elderly group. This is because $c\hat{H}(t)$ in Figure 2 is less well approximated by a geometric function for the elderly group than for the other two.

Further analyses are possible that postulate a common b in each of the age groups but allow the distribution of induction time to vary arbitrarily between groups. Other parametric models for h could also be fitted. Medley et al. (1987, 1988) fit both exponential and linear forms for h and showed that the exponential form better describes the data. Some authors (De Gruttola and Lagakos 1987; Hyman and Stanley 1988), however, have argued that quadratic or other subexponential growth in $h(x)$ is more likely, except early in the epidemic. Since the transfusion-related infections studied here occurred prior to mid-1985, and because the data do not allow one to discriminate effectively between, say, quadratic and exponential shapes, we have not carried this aspect of the analysis further. Finally, we note that if we were to leave h_i nonparametric and parameterize f_j, then it is clear from the discussion in Appendix B that the likelihoods (13) would lead to the same estimate of f as (6).

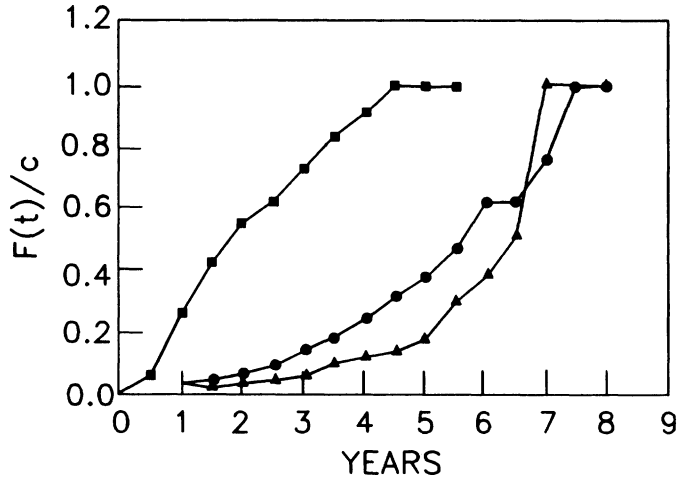

Figure 4. Estimates of F(t): the Probability That the Induction Time Is Less Than t for ■, Children; ●, Adults; and ▲, Elderly Patients. It is assumed that the mean number of infections in the ith month is ab^i. Note that c is the probability that the induction time is less than 7.5 years and is not estimable.

4. PARAMETRIC ESTIMATION

4.1 General Comments

The statistical model specified in Assumptions 1 and 2 of Section 2 is a Poisson process in the (x, t) plane, with intensity

$$\lambda(x, t) = h(x)f(t - x), \qquad t > x > 0$$
$$= 0, \qquad \text{otherwise.}$$

Under the assumptions specified, the data consist of complete observation of this process over the triangular region $R = \{(x, t): 0 < x < t < T\}$, as pictured in Figure 5. Viewed in this way, it is evident that although the data provide direct information on the mean number of outcomes $\int\int_D \lambda(x, t)\, dx\, dt$ in any specified subset D of R, they are insufficient to estimate either $h(x)$ or $f(s)$ without further assumption.

Consider a Poisson process in the plane with intensity

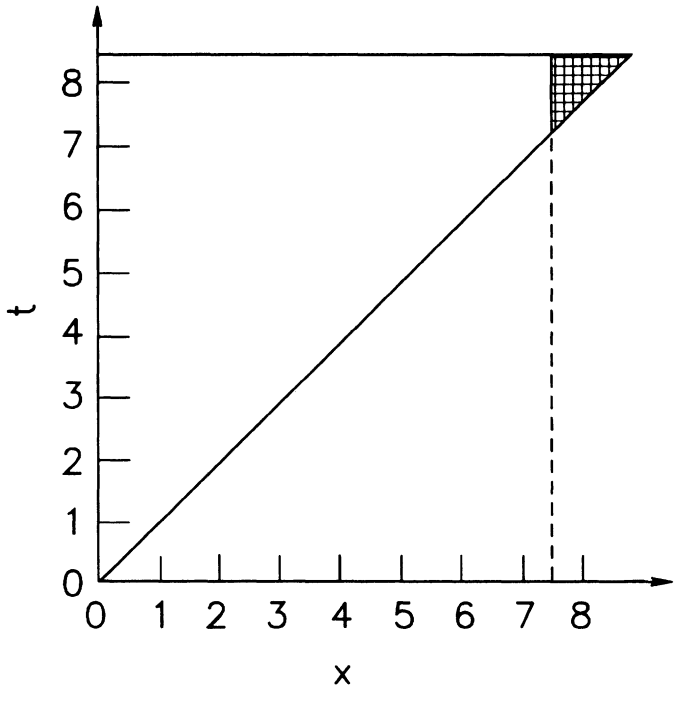

Figure 5. The large triangular region represents the region of observation of the Poisson process with intensity $\lambda(x, t) = h(x)f(x - t)$, where x represents the time of infection and t the time of diagnosis. Each observed individual is represented by a point in this region. Parametric analyses in Sections 4 and 5 are based on the region obtained by eliminating the small thatched triangle, which recognizes the introduction of blood screening in July 1985 at 7.5 years.

$\lambda(x, t; \theta)$, where θ is a vector of parameters, and suppose that the process is observed over a region $R \subset \mathbf{R}^2$. Let (x_i, t_i) $(i = 1, \ldots, n)$ represent the observed points in R. It follows that the likelihood of θ is

$$\prod_{i=1}^{n} \lambda(x_i, t_i; \theta) \exp\left\{-\int_R \lambda(x, t; \theta)\, dx\, dt\right\}. \quad (14)$$

We suppose that $\lambda(x, t; \theta)$ is twice differentiable with respect to θ for almost all x and t, and let $\lambda^{(1)}(x, t) = \partial \log \lambda/\partial\theta$ and $\lambda^{(2)}(x, t) = \partial^2 \log \lambda/\partial\theta\partial\theta'$. The score function arising from (14) is

$$U(\theta) = \sum_{i=1}^{n} \lambda^{(1)}(x_i, t_i) - \int_R \lambda^{(1)}(x, t)\lambda(x, t)\, dx\, dt,$$

and the observed information is

$$-\partial^2 \log L(\theta)/\partial\theta\partial\theta' = -\sum_{i=1}^{n} \lambda^{(2)}(x_i, t_i) + \int_R \lambda^{(2)}(x, t)\lambda(x, t)\, dx\, dt + \int_R \lambda^{(1)}(x, t)\lambda^{(1)}(x, t)'\lambda(x, t)\, dx\, dt. \quad (15)$$

It can be verified that $U(\theta)$ has expectation 0 and that the Fisher information, the expectation of (15), is

$$I(\theta) = \int_R \lambda^{(1)}(x, t)\lambda^{(1)}(x, t)'\lambda(x, t)\, dx\, dt.$$

Therefore, Fisher's scoring algorithm presents one approach to model fitting in this context. The algorithm is particularly simple to implement because it requires calculation of only first derivatives of $\lambda(x, t; \theta)$. At convergence $I(\hat{\theta})^{-1}$ provides an estimate of the covariance matrix of the MLE $\hat{\theta}$.

Some care is important in parametric analysis of data of this type. Interest may center on particular functions of $\lambda(x, t; \theta)$, which are poorly estimated. In such instances, point estimates can be extremely imprecise, and even supplemented with standard errors they may not indicate the true uncertainty in estimation. Nonparametric methods are useful in this regard because they suggest particular aspects of the model about which the data tend to be uninformative. A striking example is the transfusion data, where interest centers on the estimation of percentiles or moments of the induction time distribution, and on the expected number of infections that ultimately lead to diagnosis of AIDS. These quantities are essentially inestimable based on the transfusion data alone, as the nonparametric analysis clearly demonstrates. Although point estimates are available when parametric assumptions are made, these estimates are imprecise.

4.2 Parametric Estimation for the Transfusion Data

Medley et al. (1987, 1988) considered full parametric modeling of the transfusion data. In this section we consider a model closely related to one that they fit.

Suppose that the rate of infections is increasing exponentially beginning in January 1978 until June 1985, when the screening of blood for seropositivity was introduced, and that subsequent infection rates are 0. The data contain two individuals who were thought to be infected after June 1985, and in the following these two cases were omitted. Thus we assume that

$$h(x) = \exp(\alpha + \beta x), \quad 0 < x \le 7.5$$
$$= 0, \quad x > 7.5. \quad (16)$$

The choice of January 1978 as the time origin is arbitrary, but this is a reasonable estimate of the earliest possible time of HIV infection. To complete the model, we suppose that the duration of the induction period has a Weibull distribution with parameters $\lambda = \exp(-\mu)$ and $p = \sigma^{-1}$, having density function

$$f(s) = (\sigma s)^{-1} \exp\left\{\frac{\log s - \mu}{\sigma} - \exp\left\{\frac{\log s - \mu}{\sigma}\right\}\right\}, \quad s > 0. \quad (17)$$

The exponential form (16) is reasonably well supported by the nonparametric estimates obtained in Figure 2, and by mathematical models of the early stages of the epidemic (e.g., see Isham 1988). There is much less prima facie evidence for the Weibull model, but it is evident that similar conclusions hold for a wide class of models for the induction time. The choice of the particular parameterization in (17) was found to be advantageous in implementing the scoring algorithm.

The scoring algorithm outlined in the previous section allows models such as (16) and (17) to be fit to the data. The model is a Poisson process over the region indicated in Figure 5, with $\lambda(x, t) = h(x)f(s)$, where $s = t - x$. The components of $\lambda^{(1)}(x, t)$ are $\partial \log \lambda(x, t)/\partial\alpha = 1$, $\partial \log \lambda(x, t)/\partial\beta = x$, $\partial \log \lambda(x, t)/\partial\mu = -(1 - e^y)/\sigma$, and $\partial \log \lambda(x, t)/\partial\sigma = -(1 + y - ye^y)/\sigma$, where $y = (\log s - \mu)/\sigma$. Table 3 gives the MLE's for each of the three age groups, along with the inverse of the Fisher information matrix evaluated at the MLE. Convergence tends to be slow, which provides a warning that the likelihood may be somewhat ill-behaved and that caution should be

Table 3. Maximum Likelihood Estimates, Estimated Standard Errors (SE's), and Estimated Covariance Matrices Based on the Parametric Model (16) and (17)

Age group	MLE	SE	Covariance matrix			
0–4 years	$\hat{\alpha} = -.688$	.589	.347	−.0648	−.0190	.0054
	$\hat{\beta} = .560$	.118		.0140	.0088	−.0006
	$\hat{\mu} = .915$	.184			.0338	.0048
	$\hat{\sigma} = .542$	.084				.0070
5–59 years	$\hat{\alpha} = 2.186$	.778	.605	−.0059	.3365	.0187
	$\hat{\beta} = .637$	.073		.0053	.0116	−.0010
	$\hat{\mu} = 2.145$	.453			.2055	.0102
	$\hat{\sigma} = .425$	.040				.0016
≥60 years	$\hat{\alpha} = 1.183$	.461	.212	−.0070	.0941	.0086
	$\hat{\beta} = .800$	.079		.0060	.0076	−.0010
	$\hat{\mu} = 1.896$	.256			.0655	.0037
	$\hat{\sigma} = .384$	.032				.0010

NOTE: All times are measured in years, although analyses were done using monthly data.

exercised in applying asymptotic results in estimation.

A referee has strongly suggested that we make an explicit comparison of the point estimates in Table 3 with those of Medley et al. (1987, 1988). As noted before, estimates are very imprecise, and we include this comparison with some reluctance because it places undue emphasis on them. There are some differences between the estimates given by Medley et al. (1987) and Medley et al. (1988) and some apparent misprints in the former paper. From the latter paper, the estimates of α, β, μ, and σ, respectively, are $-.33$, .53, .98, and .52 (0–4 age group); 2.61, .64, 2.21, and .42 (5–59 age group); 1.23, .85, 1.83, and .34 (50 and over age group). These estimates are based on slightly different data than we have used, a different assumed origin for the infection process, and perhaps slightly different conventions in incorporating zeros or interpreting integer times. These facts and the very flat likelihood in the neighborhood of the maximum presumably account for the differences between these estimates and those in Table 3. The σ and β are essentially shape parameters and are more reliably estimated than α and μ.

In Medley et al. (1987), Brookmeyer and Gail (1988), and Lui et al. (1986), interest centers on the estimation of derived parameters, such as the total number of infections that leads to AIDS diagnosis, $H(7.5) = \int_0^{7.5} h(x)\,dx$, and properties of the induction distribution, such as its median. Point estimates of these quantities are easily obtained from Table 3, and standard errors can be estimated using the δ method. As an alternative and better approach, the likelihood ratio statistic can be used to provide confidence intervals. Table 4 summarizes the results for estimation of $F(7.5)$, $H(7.5)$, and the median induction time $m = e^{\mu}(-\log .5)^{\sigma}$. Except for the children, estimation of each of these quantities is imprecise.

Further insights can be obtained by examining the profile likelihood for $F(7.5)$ and $H(7.5)$,

$$R_M(\theta_1, \theta_2) = L(\tilde{\alpha}, \tilde{\beta}, \tilde{\mu}, \tilde{\sigma})/L(\hat{\alpha}, \hat{\beta}, \hat{\mu}, \hat{\sigma}), \quad (18)$$

where $(\tilde{\alpha}, \tilde{\beta}, \tilde{\mu}, \tilde{\sigma})$ are the MLE's subject to the constraint $F(7.5) = \theta_1$ and $H(7.5) = \theta_2$. For the elderly group, the 5% contour is plotted in Figure 6. This contour corresponds to an approximate 95% confidence region for (θ_1, θ_2). A similar plot based on the nonparametric analysis of Section 3 would give contours that are hyperbolas, since

Table 4. Maximum Likelihood Estimates of $F(7.5) = Pr\{S \leq 7.5\}$, $H(7.5) = \int_0^{7.5} h(x)\,dx$, and the Median Induction Time m by Age Group

Age group	F(7.5)	H(7.5)	Median (years)
0–4 years	1.00 (.82, 1.00)	53 (37, 126)	2.0 (1.5, 4.0)
5–59 years	.52 (ε, .90)	1,644 (594, ∞)	7.3 (4.6, ∞)
≥60 years	.74 (ε, .95)	1,643 (734, ∞)	5.8 (4.3, ∞)

NOTE: The interval specifies those values of the parameter for which the generalized likelihood ratio exceeds 14%. If large sample theory were appropriate, these would represent approximate 95% confidence intervals. ε denotes a small (<.0001) positive quantity and ∞ denotes a large positive quantity.

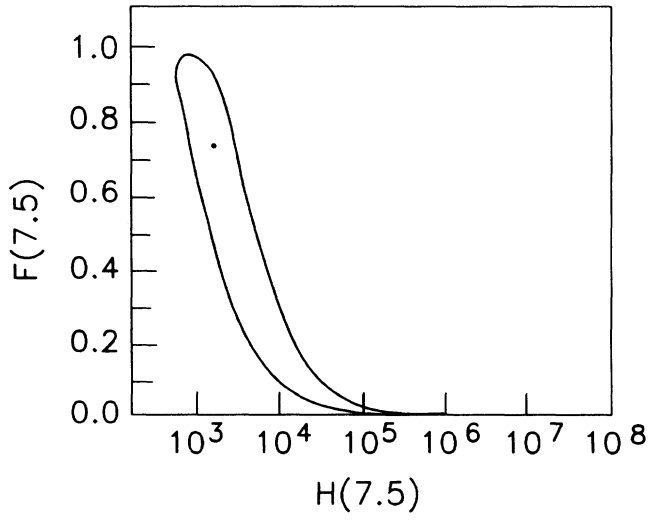

Figure 6. Contour of the Profile Likelihood (18) of $\theta_1 = F(7.5)$ and $\theta_2 = H(7.5)$, Based on the Elderly Patients and the Model (16) and (17). These parameters are nearly nonidentifiable, and this approximate 95% confidence region includes values of H(7.5) (the expected number of infections) that exceed 10^7.

only the product $H(7.5)F(7.5)$ is estimable; this same characteristic is reflected in the parametric plot. Figure 7 gives the 5% contour for the children. Here the parameters are better estimated because of the indication that the whole distribution of induction times has been seen (see Fig. 3).

It is apparent from Table 4 and Figures 6 and 7 that parametric assumptions do not circumvent the identifiability problems discussed in Section 3, and that point estimates from such models can be misleading. In addition, approximate confidence limits based on MLE's and

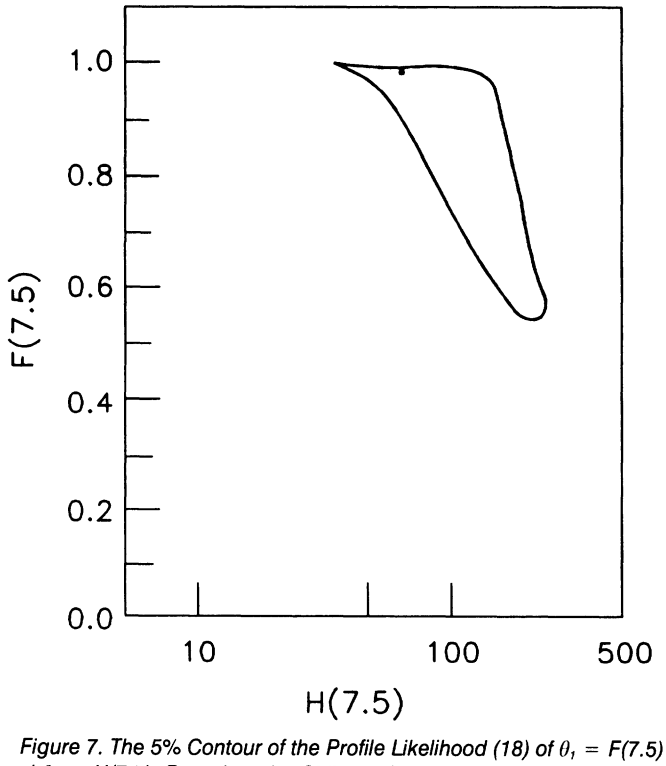

Figure 7. The 5% Contour of the Profile Likelihood (18) of $\theta_1 = F(7.5)$ and $\theta_2 = H(7.5)$, Based on the Children (age ≤4 years). The data are more informative about these parameters than for the elderly or adult groups.

asymptotic standard errors can be very different from those based on likelihood ratios. In this case, the latter should be considered more reliable.

5. COVARIANCE ANALYSIS

Suppose that z is a vector of discrete covariates that apply to individuals at the time they experience the initiating event. Let $g(z) = \Pr\{Z = z\}$ $(z \in \mathcal{Z})$ be the probability function of z, which is assumed to be the same at each time x of the initiating event, and suppose that the density function of the induction time, given z, is $f(s; z)$. The data consist of the time t_i of ascertainment, as well as a retrospective determination of the corresponding time of the initiating event x_i and the covariate z_i $(i = 1, \ldots, n)$.

As before, let $h(x)$ be the overall intensity of initiating events so that the intensity for covariate value z is $g(z)h(x)$. The likelihood is constructed by considering each $z \in Z$ separately and combining the independent likelihood components over z to obtain

$$\prod_{i=1}^{n} \{g(z_i)h(x_i)f(s_i; z_i)\} \times \exp\left\{-\sum_{z\in Z} g(z) \int\int h(x)f(s; z)\, dx\, ds\right\}, \quad (19)$$

where integration is over the region $0 < x + s < T$. Parametric models for g, h, and f can be specified, and a slight variation of the scoring algorithm outlined in Section 4.1, which applies to each level of z separately, provides a simple numerical method for fitting (19).

The intensity $h(x)$ should be proportional to the prevalence of HIV infection in blood available for transfusion. In the transfusion data, it seems reasonable to suppose, as before, that

$$\begin{aligned} h(x) &= \exp(\alpha + \beta x), & 0 < x \le 7.5 \\ &= 0, & x > 7.5. \end{aligned} \quad (20)$$

One possible approach to fitting (19) is to leave $g(z)$ arbitrary and postulate a parametric regression model for s; the Weibull regression model, for example, would have conditional density function

$$f(s; z) = (1/\sigma s)\exp\{y - e^y\}, \quad (21)$$

where $y = (\log s - z'\gamma)/\sigma$ and γ is a vector of regression parameters. Alternatively, it may be possible to specify a parametric model for $g(\cdot)$.

When $g(\cdot)$ is left arbitrary, it is easily verified that the likelihood (19) is maximized with respect to g at

$$\tilde{g}(z) = c_z \Big/ \int\int h(x)f(s; z)\, dx\, ds, \qquad z \in Z,$$

where $c_z = \#\{z_i = z\}$. Thus the profile likelihood, after maximization with respect to g, is proportional to

$$\prod_{i=1}^{n} \frac{h(x_i)f(s_i; z_i)}{\int\int h(x)f(s; z_i)\, dx\, ds}, \quad (22)$$

which also arises as a conditional likelihood obtained by conditioning on the observed values of c_z $(z \in Z)$.

The likelihood (22), with the assumptions (20) and (21), is most easily fitted by working with the related Poisson likelihood (see the comments in Sec. 2 and App. A)

$$\prod_{z\in Z}\left[\prod_{i|z_i=z} \{\alpha_z e^{\beta x_i} f(s_i; z)\} \times \exp\left\{-\alpha_z \int\int e^{\beta x} f(s; z)\, dx\, ds\right\}\right], \quad (23)$$

which gives rise to the same estimates of β, γ, and σ as (22), along with appropriate covariance estimates. The estimate of α_z can be considered as an estimate of $g(z)e^{\alpha}$. Calculations based on (23) are identical to those discussed earlier for the homogeneous case, except that a separate computation is required for each observed z in the data set.

As a simple example, let z be a three-dimensional vector with $z_0 = 1$ and $z_1 = 1$ if age ≤ 4 and 0 otherwise, and $z_2 = 1$ if age ≥ 60 and 0 otherwise. The estimates arising from (23) with estimated standard errors are $\hat{\beta} = .686(.048)$, $\hat{\gamma}_0 = 2.45(.725)$, $\hat{\gamma}_1 = -1.53(.720)$, $\hat{\gamma}_2 = -.57(.660)$, and $\hat{\sigma} = .428(.025)$. A test for equality of induction period distribution across the three age groups can be based on a Wald-type statistic on 2 df. The observed value of 16.54 provides strong evidence of a dependence of induction period on age group. This effect is entirely due to the short induction periods of the younger age group; there is no evidence in the data of any difference between the adults and elderly patients.

6. OTHER COMMENTS

6.1 Accounting for the Reporting Lag

In the transfusion data, there is a lag between diagnosis and the time at which the case is reported. The models and methods discussed previously are easily generalized to allow for this reporting lag.

In the simplest case, let $k(y)$ be the pdf of the reporting lag y, and suppose that this is independent of the time t of diagnosis. The likelihood analogous to (4) is

$$L = \prod_{i=1}^{n} \{h(x_i)f(s_i)k(y_i)\}\exp\{-A^*\}, \quad (24)$$

where $A^* = \int\int\int h(x)f(s)k(y)\, dy\, ds\, dx$ and integration is over the region $0 < x + s + y < T$. Both nonparametric and parametric analyses can be applied as before, and joint estimation of h, f, and k considered. Hypotheses regarding a decrease in reporting lag with time can be addressed by allowing $k(y)$ to depend on the time t of diagnosis.

An alternative approach to the incorporation of reporting lag would be to specify a screened process of intensity $h_k(x)$ corresponding to the occurrence of new infections that ultimately lead to an AIDS diagnosis and a reporting lag of k months duration. We retain the assumption that the induction time is distributed independently of the time x of infection, with cdf and pdf $F(s)$

and $f(s)$ independent of k. Within the kth stratum, the observation on infection time x and diagnosis $x + s$ is bounded by $0 \le x \le x + s \le T - k$, and within the stratum g_j^* in (10) is estimated by

$$b_{jk} \Big/ \left(\sum_{i=0}^{T-k-j} a_{ik} - \sum_{l=j+1}^{T-k} b_{lk} \right), \qquad j = 0, \ldots, T-k,$$

in an obvious notation. Combining across strata gives the estimate

$$\hat{g}_j^* = \sum_{k=0}^{T-j} b_{jk} \Big/ \sum_{k=0}^{T-j} \left(\sum_{i=0}^{T-k-j} a_{ik} - \sum_{l=j+1}^{T-k} b_{lk} \right), \qquad j = 0, \ldots, T,$$

which can be used in (12) to obtain a nonparametric estimate of the conditional cdf G_j. Parametric analyses could also be done, allowing separate parameters for each intensity $h_k(x)$ $(k = 0, 1, 2, \ldots)$. This approach has the advantage of not using assumptions of stationarity in time of the distribution of the reporting lag.

It is evident that the methods considered here can be similarly extended to allow chains of events of arbitrary length leading to ascertainment.

6.2 Incorporation of Mortality

In the previous discussion, mortality has been ignored by restricting attention to a purged multistate process in which only individuals who are eventually diagnosed with AIDS are considered. It is also of interest to consider the process in which individuals may leave because of death by other causes. Thus we have a process of the form illustrated in Figure 8, where once again we consider separate models for the three age groups examined earlier. As in Section 2, $h^*(x)$ represents the rate of infective transfusions, whether or not these transfusions subsequently result in a diagnosis of AIDS. Death rates $\mu(t; x, a)$ would typically depend on the age a of the individual within each age group, as well as the time x of transfusion. As before, however, we might assume that the instantaneous rate of transition from infection to AIDS is $r(t; x) = r(s)$ and depends only on the time $s = t - x$ since infection.

Within a small specified age group in which the mortality can be seen as age-independent, we might assume that the death rate $\mu(t; x, a) = \mu(s)$ depends solely on the times

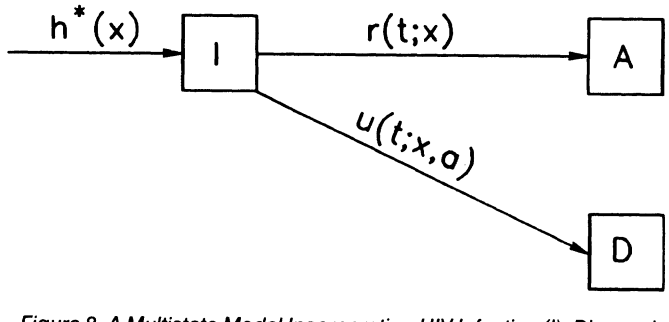

Figure 8. A Multistate Model Incorporating HIV Infection (I), Diagnosis With Full-Blown AIDS (A), and Death by Causes Other Than AIDS (D).

since transfusion. In this case, the likelihood is proportional to

$$\prod_{i=1}^{n} [h^*(x_i) f^*(s_i)] \exp\left\{ -\int\int h^*(x) f^*(s)\, dx\, ds \right\},$$

where $f^*(s) = r(s)\exp\{-\int_0^s r(u) + \mu(u)\, du\}$. The estimation problem is therefore unchanged. Specification of $\mu(s)$ allows estimation of the proportion of individuals who eventually develop AIDS. Of course, this estimation is subject to the same identifiability problems and arbitrary constant of proportionality c in using a nonparametric analysis. In view of the estimability problems with the transfusion data, more elaborate analyses do not seem warranted.

6.3 Uncertain Times of Infection

In some circumstances the precise time x_i of the initiating event may be unknown. In particular, for transfusion-related AIDS cases the date of HIV infection is frequently known only to lie in some interval. This is readily handled for either the nonparametric or parametric methods of analysis; we outline the nonparametric methods.

Suppose that for the ith individual observed we know the exact time t_i of the second event, but only that the first event lies in the interval $u_i \le x_i \le v_i$ $(i = 1, \ldots, n)$. The likelihood function for the parameters $\{h_i\}$ and $\{f_j\}$ of Section 3 is then proportional to

$$L = \prod_{l=1}^{n} \left(\sum_{i=u_l}^{v_l} h_i f_{t_l - i} \right) \exp\left\{ \sum_{i=0}^{T} \sum_{j=0}^{T-i} h_i f_j \right\}, \qquad (25)$$

under conditions of independent censoring. Define for convenience

$$\delta_{ir}^h = I(u_i \le r \le v_i), \quad \delta_{is}^f = I(t_i - v_i \le s \le t_i - u_i),$$

and

$$D_r = \sum_{i=u_r}^{v_r} h_i f_{t_r - i}.$$

Differentiation of $\log L$ then gives

$$\frac{\partial \log L}{\partial h_i} = \sum_{j=1}^{n} \frac{\delta_{ji}^h f_{t_j - i}}{D_j} - F_{T-i}, \qquad i = 0, 1, \ldots, T,$$

and

$$\frac{\partial \log L}{\partial f_j} = \sum_{i=1}^{n} \frac{\delta_{ij}^f h_{t_i - j}}{D_i} - H_{T-j}, \qquad j = 0, 1, \ldots, T.$$

An EM algorithm (Dempster, Laird, and Rubin 1977) provides a convenient method of maximizing (25). If as in Section 3 a_i and b_j represent the (unknown) total numbers of initiating event times equal to i and induction times equal to j, respectively, then

$$E\{a_r \mid (u_i, v_i, t_i), i = 1, \ldots, n\} = \sum_{i=1}^{n} \frac{\delta_{ir}^h h_r f_{t_i - r}}{D_i}, \qquad r = 0, 1, \ldots, T, \quad (26)$$

and

$$E\{b_s \mid (u_i, v_i, t_i), i = 1, \ldots, n\} = \sum_{i=1}^{n} \frac{\delta_{is}^{f} h_{t_i - s} f_s}{D_i}, \quad s = 0, 1, \ldots, T. \quad (27)$$

The EM algorithm consists of selecting initial estimates $\{\tilde{h}_i\}\{\tilde{f}_j\}$ and then alternating the E and M steps, as follows.

E step: Using the current estimates $\{\tilde{h}_i\}$ and $\{\tilde{f}_j\}$, impute values $\{a_r\}$ and $\{h_i\}$ equal to the right sides of (26) and (27), respectively.

M step: Compute new estimates $\{\tilde{h}_i\}$ and $\{\tilde{f}_j\}$ by using the current imputed a_r and b_s values ($r, s = 0, 1, \ldots, T$) in the full-data MLE's given in Section 3.1.

On convergence, this algorithm yields estimates maximizing (25). A more detailed discussion of (25) and the EM algorithm will be given elsewhere; however, we note that it is possible for some observed data sets to have nonunique MLE's for $\{f_j\}$ and $\{h_i\}$, even when we make the restriction $F_T = 1$.

7. DISCUSSION

Nonparametric methods, where they can be applied, provide a vehicle by which one can gain insight into those aspects of a statistical model that are essentially nonidentifiable. It is often the case that introduction of parametric assumptions makes these aspects technically identifiable but leads to imprecise estimation. Retrospective ascertainment (as in the transfusion data) cannot, either with or without parametric assumptions, lead to valid estimation of the mean or median incubation times, or to the size of the epidemic. The data must be supplemented to estimate these important parameters. The data do, however, give direct evidence on the shape of the intensity function for the epidemic and are consistent with an exponential increase early in the epidemic. They also provide evidence on the shape of the induction time distribution over the first seven years after infection by transfusion.

Several problems of statistical interest can be identified. Additional investigation of the nonparametric methods in Sections 3 and 6 would be useful; although statistical properties under the discrete model of Section 3.1 are fairly straightforward, interesting questions remain concerning estimation and tests for the continuous time functions $H(x)$ and $F(s)$, for more complicated likelihoods such as (25), and for the multivariate analogs of the bivariate Poisson model. Lagakos et al. (1988) considered asymptotic properties of the nonparametric estimator of $F(s)$ based on results in the theory of counting processes.

It is of some interest to note that a reverse time proportional hazards model could be considered for the "hazard" $g^*(s) = \Pr\{S \in (s, s - \Delta s) \mid S \leq s\}$, whereby it is assumed that $g^*(s; z) = g_0^*(s)\exp(\mathbf{z}'\boldsymbol{\beta})$, for example, where z is a vector of regression parameters. For this model, it is easy to see that $F(s; z) = F(s; 0)^{\exp(\mathbf{z}'\boldsymbol{\beta})}$, and parameters have a simple interpretation. A partial likelihood can be developed to estimate β, with $g_0^*(s)$ or $F(s; 0)$ left completely unspecified. These methods, however, would not extend to allow interval censoring on the time of the initiating event.

Finally, interesting questions arise as to how best the transfusion data can be supplemented to gain full identifiability of the quantities of interest. One possible approach would be (a) to use historical records on post-transfusion mortality due to competing risks, to obtain an estimate of the death rate $\mu(s)$ (see Fig. 8) and (b) to use data gathered from stored serum at blood banks to obtain estimates of the total number of infective transfusions administered over some period of time. These two pieces of information would enable full identifiability of the cause-specific hazard $r(s)$.

Since this article was written there have been several updates of the CDC's data on transfusion-associated AIDS. As of June 30, 1988, the number of cases reported had risen from 494 to 1,788. Kalbfleisch and Lawless (in press) present a detailed analysis of these more recent data. The broad conclusions of Sections 3 and 4 remain valid; in particular, confidence intervals for quantities of interest remain very wide, even with the longer follow-up period and much larger sample size. For example, with the parametric model used in Section 4.2 the 95% confidence interval on the median induction time for persons aged 13–69 is (7.9, ∞). Some interesting new features emerge, however. In the more recent data, there is no longer evidence that very young patients have an induction time distribution substantially different from that for the adults. Unlike the earlier data reported in Table 1 and analyzed in Sections 3 and 4, the more recent data have very young patients that have induction times as long as 107 months. In addition, the more recent data indicate that, to estimate the total number of infections leading to AIDS, the reporting lag must be explicitly considered. The approach we and other authors have used, adjusting for reporting lag by including only cases diagnosed at least six months prior to the last reporting date, is inadequate and leads to substantial underestimation. Consequently, the analyses in Kalbfleisch and Lawless (in press) are based on the approach outlined in Section 6.1.

APPENDIX A: EQUIVALENCE OF ESTIMATION BASED ON THE LIKELIHOODS (4) AND (6)

Consider a Poisson process of intensity $\lambda(x, t)$ observed over a region R of the (x, t) plane. Suppose that $\lambda(x, t) = e^{\alpha}\lambda_0(x, t; \theta)$, where $\alpha \in \mathbf{R}$ and θ is a vector of parameters. Let $A_0 = \int\int \lambda_0(x, t; \theta)\, dx\, dt$, where integration is over R, and let $\psi = e^{\alpha}A_0$. The conditional likelihood (6) and the Poisson likelihood (4) are related by $L_p = L_c\psi^n e^{-\psi}/n!$ It is then easy to see that the MLE of ψ is n, and maximizing L_p with respect to θ is equivalent to maximizing L_c. Thus the two likelihoods give rise to the same estimates of θ, and the block-diagonal structure of the matrix of second derivatives of L_p in the parameterization ψ, θ assures that the observed and expected information for θ are also the same for the two likelihoods.

This very simple argument is attributable to D. R. Cox and was communicated to us by Valerie Isham.

APPENDIX B: NONPARAMETRIC ESTIMATION OF $F(s)/F(T)$

It is possible to base nonparametric estimation of the induction time distribution $F(s)$ on the likelihoods (3) or (6) of Section 2. The discrete time analog of (6) is

$$L_1 = \prod_{i+j\le T} (h_i f_j)^{n_{ij}} \Big/ \left(\sum_{i+j\le T} h_i f_j \right)^n, \tag{B.1}$$

where (as in Sec. 3) n_{ij} is the observed number of cases with the initiating event at i and the terminal event at $i + j$. The discrete likelihood corresponding to (3) is

$$L_2 = \prod_{i+j\le T} [f_j / F_{T-i}]^{n_{ij}}. \tag{B.2}$$

It is a simple matter to check that the profile likelihood from (7) or (B.1) obtained by maximizing with respect to $\{h_i\}$ is given by (B.2). Thus all three likelihoods, (3), (4), and (6), lead to the same nonparametric estimates of $F(s)/F(T)$.

This is, in essence, the well-known result that for triangular incomplete contingency tables, Poisson, multinomial, and product-multinomial sampling all yield the same estimates (e.g., see Bishop et al. 1975, sec. 5.2). Our result is a slight extension of the contingency table result, which assumes that the incomplete contingency table is inseparable.

The aforementioned results make it clear that the Poisson-based nonparametric estimates for the f_j's are valid, regardless of what the initiating event process happens to be, provided the induction time is independent of the time of the initiating event. (One can also test the independence assumption, if desired.) The Poisson-based estimates of the h_j's are generally valid as an estimate of the (unobserved) number of initiating events realized at time j. Of course, developing sampling properties of the estimates requires the specification of the initiating events process. We stress the use of the nonparametric estimates in assessing the shape of $H(x)$ and $F(s)$, and in view of the basic nonidentifiability of H and F we do not develop any distribution theory that could be used to obtain interval estimates. This is easily done, however, under the Poisson model. We remark also that by letting the number of discrete values for x or s become large, the nonparametric estimates are seen to give valid estimates of the analogous continuous-time functions such as $H(x)/H(T)$ and $F(s)/F(T)$. Lagakos et al. (1988) and Keiding and Gill (1987) provided some continuous-time asymptotics.

[*Received January 1988. Revised December 1988.*]

REFERENCES

Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W. (1975), *Discrete Multivariate Analysis*, Cambridge, MA: MIT Press.

Brookmeyer, R., and Gail, M. H. (1988), "A Method for Obtaining Short-Term Projections and Lower Bounds on the Size of the AIDS Epidemic," *Journal of the American Statistical Association*, 83, 301–308.

Deffner, A., and Haeusler, E. (1985), "A Characterization of Order Statistic Point Processes That Are Mixed Poisson Processes and Mixed Sample Processes Simultaneously," *Journal of Applied Probability*, 22, 314–323.

DeGruttola, V., and Lagakos, S. W. (1987), "The Value of Doubling Time in Assessing the Course of the AIDS Epidemic," Technical Report 564Z, Dana-Farber Cancer Institute, Boston.

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

Goodman, L. A. (1968), "The Analysis of Cross-Classified Data: Independence, Quasi-Independence, and Interactions in Contingency Tables With or Without Missing Values," *Journal of the American Statistical Association*, 63, 1091–1131.

Hyman, J. M., and Stanley, E. A. (1988), "Using Mathematical Models to Understand the AIDS Epidemic," *Mathematical Biosciences*, 90, 415–473.

Isham, V. (1988), "Mathematical Modelling of the Transition Dynamics of HIV Infection and AIDS: A Review" (with discussion), *Journal of the Royal Statistical Society*, Ser. A, 151, 5–49.

Kalbfleisch, J. D., and Lawless, J. F. (1988), "Estimating the Incubation Period for AIDS Patients," *Nature*, 333, 504–505.

——— (in press), "Estimating the Incubation Time Distribution and Expected Number of Cases for Transfusion-Associated Acquired Immune Deficiency Syndrome," *Transfusion*, 29.

Keiding, N., and Gill, R. (1987), "Random Truncation Models and Markov Processes," Research Report 87/3, University of Copenhagen, Statistical Research Unit.

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

Lui, K. J., Lawrence, D. N., Morgan, W. M., Peterman, T. A., Haverkos, H. W., and Bragman, D. J. (1986), "A Model-Based Approach for Estimating the Mean Incubation Period of Transfusion-Associated Acquired Immunodeficiency Syndrome," *Proceedings of the National Academy of Science, USA*, 83, 3051–3055.

May, R. M., and Anderson, R. M. (1987), "Transmission Dynamics of HIV Infection," *Nature*, 326, 137–142.

Medley, G. F., Anderson, R. M., Cox, D. R., and Billard, L. (1987), "Incubation Period of AIDS in Patients Infected Via Blood Transfusion," *Nature*, 328, 719–721.

Medley, G. F., Billard, L., Cox, D. R., and Anderson, R. A. (1988), "The Distribution of the Incubation Period for the Acquired Immunodeficiency Syndrome (AIDS)," *Proceedings of the Royal Society of London*, Ser. B, 233, 367–377.

Peterman, T. A., Drotman, D. P., and Curran, J. W. (1985), "Epidemiology of the Acquired Immunodeficiency Syndrome (AIDS)," *Epidemiologic Reviews*, 7, 1–21.

Peterman, T. A., Lui, K. J., Lawrence, D. N., and Allen, J. R. (1987), "Estimating the Risks of Transfusion-Associated Acquired Immune Deficiency Syndrome and Human Immunodeficiency Virus Infection," *Transfusion*, 27, 371–374.