AIDS

AAAS SYMPOSIA PAPERS 1988

Edited by Ruth Kulstad

The American Association for the Advancement of Science

6. The Incubation Period for HIV-1

Jeffrey E. Harris

A striking feature of the pathobiology of HIV-1 is the long and variable incubation period between initial viral infection and the emergence of AIDS. Data on the duration of the HIV-1 incubation period have now been reported separately from three sources: homosexual men (*1*), hemophiliacs (*2*), and transfusion recipients (*3*, *4*). Here, I combine the evidence from these three sources to produce an overall statistical description of the HIV-1 incubation period. My statistical method closely parallels that of DuMouchel and Harris (*5*), who combined the results of various dose-response experiments in the assessment of human cancer risks.

Data

Figure 1 shows the proportion of HIV-1–infected homosexual men who have contracted AIDS by a given time after HIV-1 seroconversion. Also shown are the 68% confidence limits around each step in the estimated cumulative distribution. The estimates in Fig. 1 were derived by the Kaplan-Meier technique on data from 155 men followed in the San Francisco Clinic Cohort (*1*).

Figure 2 shows the corresponding cumulative distribution, along with 68% confidence limits, for HIV-1–infected adult hemophiliacs. The estimates were derived by the Kaplan-Meier technique on data from 40 subjects followed at the Hemophilia Center of Central Pennsylvania (*2*).

Figure 3 shows the proportion of transfusion recipients who have AIDS by a given time after transfusion with HIV-1–infected blood. The figure has been estimated by a nonparametric maximum likelihood technique (*6*) from data on 297 transfusion-associated AIDS cases reported during 1978–1986, for which the date of transfusion was known (*3*). The cumulative proportion of AIDS cases by 8 years is exactly 100% because the investigators observed only the transfusion-infected cases that have been so far diagnosed with AIDS.

Statistical Methods

Figures 1 through 3 provide separate estimates of the cumulative distribution function for the incubation period of HIV-1. My task is to combine the three sources of data to produce an overall, synthetic estimate of the cumulative distribution.

The nonparametric estimates in Figs. 1 through 3 are step functions with different support points that depend on the observed incubation periods for the subjects under study. From each cumulative distribution, however, we can derive the estimated incidence of AIDS at evenly spaced intervals. In the present analysis, I work with the estimated annual incidence rates (and their sampling errors) derived from the three studies.

Jeffrey E. Harris is professor of economics at the Massachusetts Institute of Technology, Cambridge, MA 02139.

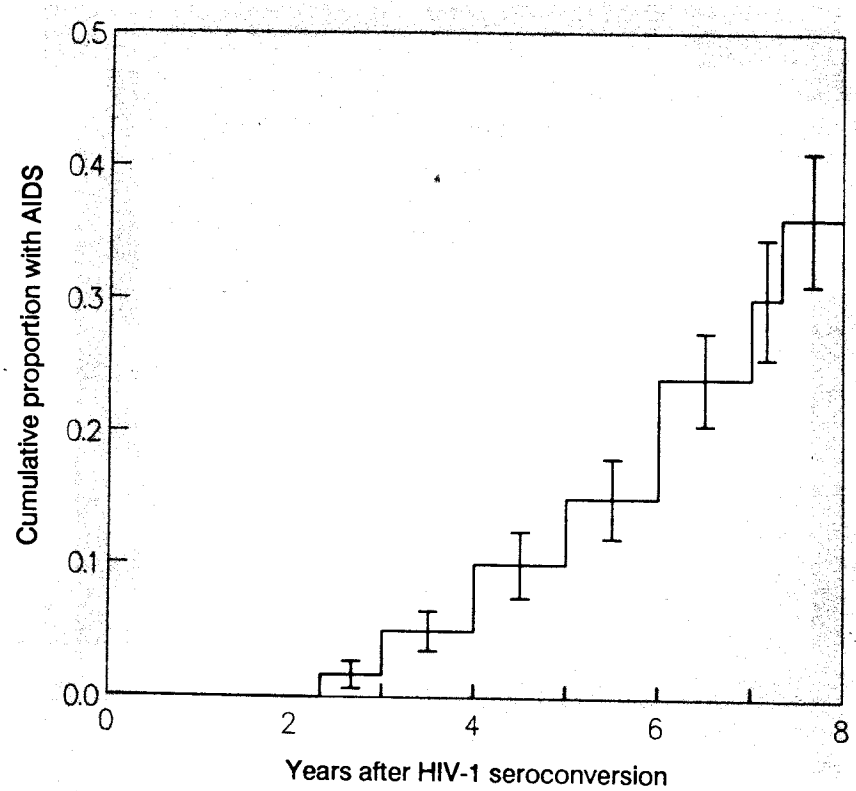

Fig. 1. Estimated cumulative proportion of homosexual men with AIDS in relation to the time since HIV-1 seroconversion. The error bars show the estimated 68% confidence intervals (1 standard error). The data are from (*1*).

Let q_{jt} denote the observed annual incidence of AIDS in cohort j during year t, where $j = 1,2,3$ and $t = 0,1,2,\ldots,7$. That is, q_{jt} is the estimated probability that an HIV-1–infected person in cohort j will develop symptomatic AIDS in the half-open time interval $[t, t+1)$, given that AIDS was not diagnosed during $[0, t)$.

The data q_{jt} are estimates of the actual incidences Θ_{jt}. I model the sampling errors for each Θ_{jt} as follows:

$$\log q_{jt} = \log \Theta_{jt} + \varepsilon_{jt} \quad (1),$$

where the ε_{jt} are joint normally distributed with zero means and known variance-covariance matrix. This assumption was motivated by the fact that the data in Figs. 1 through 3 are maximum likelihood estimates. Since Eq. 1 is in logarithmic form, the standard deviations of the error terms ε_{jt} can be interpreted as relative standard errors.

I further assume that the actual incidences Θ_{jt} are interrelated by the following linear model:

$$\log \Theta_{jt} = \mu + \alpha_j + \gamma_t + \delta_{jt} \quad (2)$$

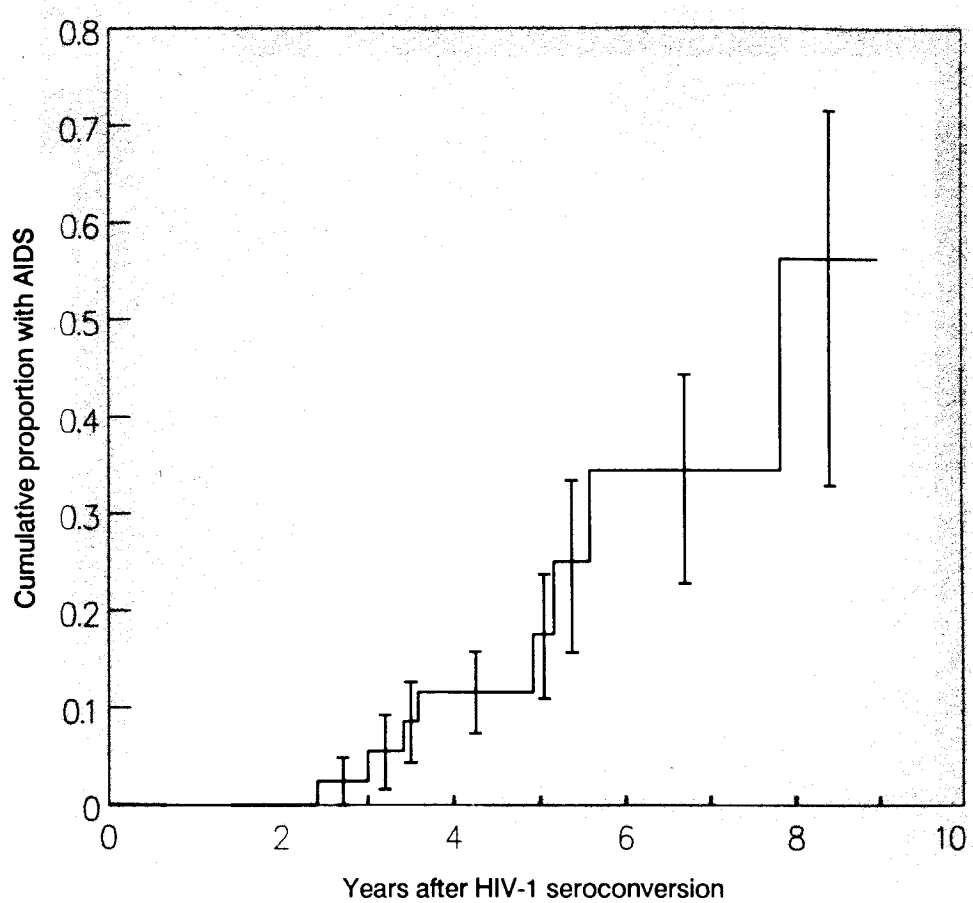

Fig. 2. Estimated cumulative proportion of adult hemophiliacs (aged 21 and over) with AIDS in relation to the time since HIV-1 seroconversion. The error bars show the estimated 68% confidence intervals (1 standard error). The data are from (*2*).

where μ, $\{\alpha_1, \alpha_2, \alpha_3\}$, and $\{\gamma_0,\ldots,\gamma_7\}$ are unknown parameters, and where the error terms δ_{jt} are independently and identically normally distributed with mean zero and variance σ^2. The parameters α_j are cohort effects, while the γ_t are time effects. By Eq. 2, the relative incidence rates for any two cohorts are approximately independent of duration t of HIV-1 infection. That is, for any two cohorts j and k and any year t, the quantity $\log \Theta_{jt} - \log \Theta_{kt}$ has time-independent mean $\alpha_j - \alpha_k$ and variance $2\sigma^2$. In essence, the parameter σ gauges the overall accuracy of approximation to this "constant relative risk" model. The critical statistical assumption here is that the approximation errors δ_{jt} and δ_{kt} have independent, identical distributions. This is what DuMouchel and Harris (*5*) called the "exchangeability assumption."

From Eqs. 1 and 2, the overall model is

$$\log q_{jt} = \mu + \alpha_j + \gamma_t + \varepsilon_{jt} + \delta_{jt} \quad (3).$$

If one has strong prior information, then all the parameters μ, α, γ, and σ can be estimated by Bayesian techniques (*5*). Here, I report only the classical maximum

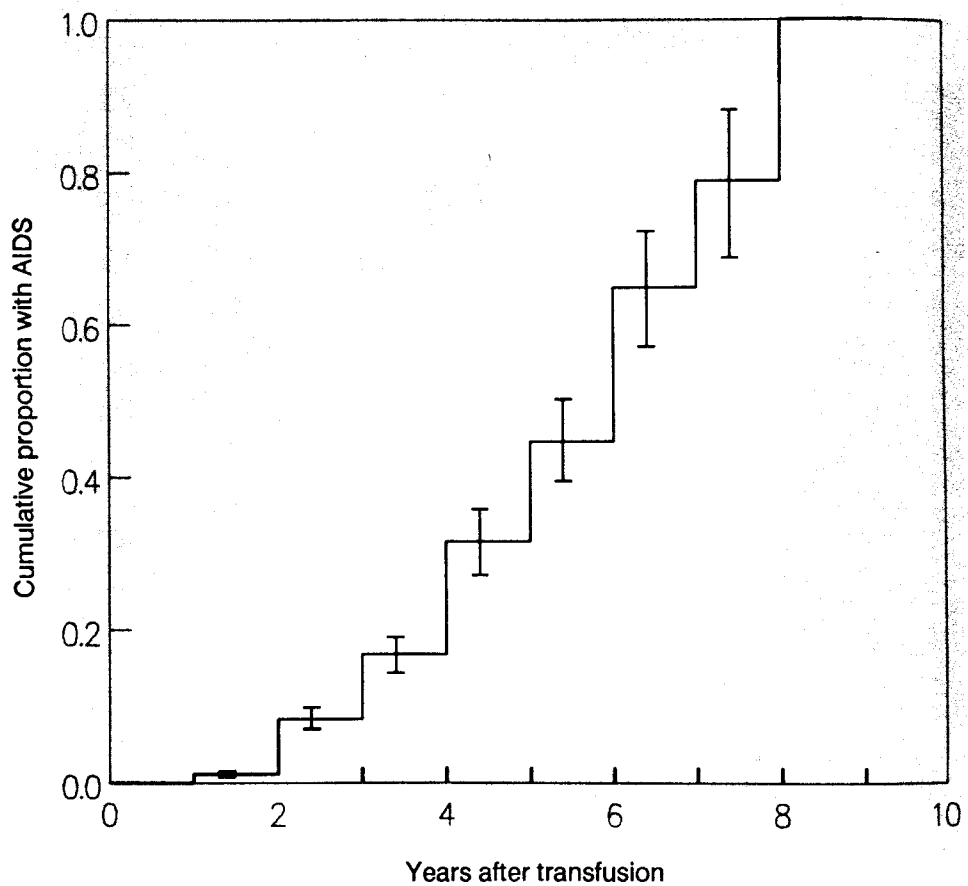

Fig. 3. Estimated cumulative proportion of transfusion recipients with AIDS in relation to the time since administration of HIV-1-infected blood. The error bars show the estimated 68% confidence intervals (1 standard error). A nonparametric maximum likelihood technique (*6*) was applied to the data reported in (*3*).

likelihood estimates. Accordingly, only the parameters μ and σ and the contrasts $\alpha_j - \alpha_1$ (for j = 2 and 3) and $\gamma_t - \gamma_0$ (for $t = 1,\ldots,7$) are separately identified.

The incidence data for homosexual men ($j = 1$) and hemophiliacs ($j = 2$) are derived from cohorts of HIV-1–seropositive subjects. By contrast, the estimates for transfusion recipients ($j = 3$) are derived solely from those recipients who have so far developed AIDS. That is, each q_{3t} is an estimate of Θ_{3t}/Φ, where Φ is the proportion of all persons receiving HIV-1–infected blood who will develop AIDS by 8 years. Accordingly, for the transfusion recipients ($j = 3$), Eq. 2 needs to be replaced by

$$\log \Theta_{3t} = \mu + \log \Phi + \alpha_3 + \gamma_t + \delta_{3t} \quad (4).$$

In a Bayesian framework, prior information on Φ might permit us to estimate this probability separately. In a classical framework, however, only the contrast $\log \Phi + \alpha_3 - \alpha_1$ can be identified (*7*).

Results

The incubation periods for the three data sources showed a close fit to the model of Eq. 3. The likelihood function for the parameter σ displayed an essentially flat top in the range $0 \leq \sigma \leq 0.05$. That is, with 68% probability, the assumption of constant relative incidence rates is accurate to within 5%.

Moreover, I could not reject the hypothesis that the annual incidence rates for HIV-1–infected homosexual men (cohort 1) and those for HIV-1–infected hemophiliacs (cohort 2) are equal. Accordingly, I shall present the estimated synthetic incidence rates under the restriction $\alpha_1 = \alpha_2$.

Table 1 shows the estimated annual incidence rates of AIDS during the years 0 through 7 after HIV-1 seroconversion. The left-hand column shows the original data q_{1t} from the San Francisco Clinic Cohort. The right-hand column shows the corresponding estimates of Θ_{1t}. The latter show the incidence of AIDS rising from 0.4% in the year immediately following HIV-1 infection to 12.9% during year 7.

In computing the incidence rates Θ_{1t} for homosexual men, the statistical procedure essentially borrows information from the other two cohorts. Hence, the relative standard deviations for the synthetic estimates Θ_{1t} in Table 1 are smaller than those of the original data q_{1t}. Still, the synthetic estimates are not too precise. The estimated incidence rate of 0.129 during year 7 has a 68% confidence interval of 0.092 to 0.180. Accordingly, while we can be reasonably sure that the incidence of AIDS rises during years 0 through 5 after HIV-1 infection, we cannot confidently conclude that it continues to rise during years 6 and 7.

Having obtained the estimates Θ_{1t}, we can compute a synthetic cumulative distribution function. This is done in Fig. 4, where the filled squares show the estimated cumulative proportion with AIDS at the start of each year. At the 8th year after HIV-1 infection, an estimated 41.2% have AIDS, with a 68% confidence interval of 36.2% to 45.8%.

Table 1. Estimated annual incidence of AIDS during years 0 through 7 after HIV-1 seroconversion.*

Year	San Francisco Clinic Cohort	Synthetic estimate
0	0.000	0.004 (0.364)
1	0.000	0.024 (0.273)
2	0.050 (0.300)	0.038 (0.211)
3	0.053 (0.411)	0.059 (0.246)
4	0.056 (0.370)	0.061 (0.256)
5	0.106 (0.250)	0.119 (0.201)
6	0.079 (0.523)	0.096 (0.403)
7	0.086 (0.474)	0.129 (0.397)

*The numbers in parentheses are the relative standard errors. The San Francisco Clinic Cohort data are from (*1*). The synthetic estimates are the maximum likelihood estimates from three data sources (homosexual men, hemophiliacs, transfusion recipients) based on the model of Eq. 3 with the restriction $\alpha_1 = \alpha_2$.

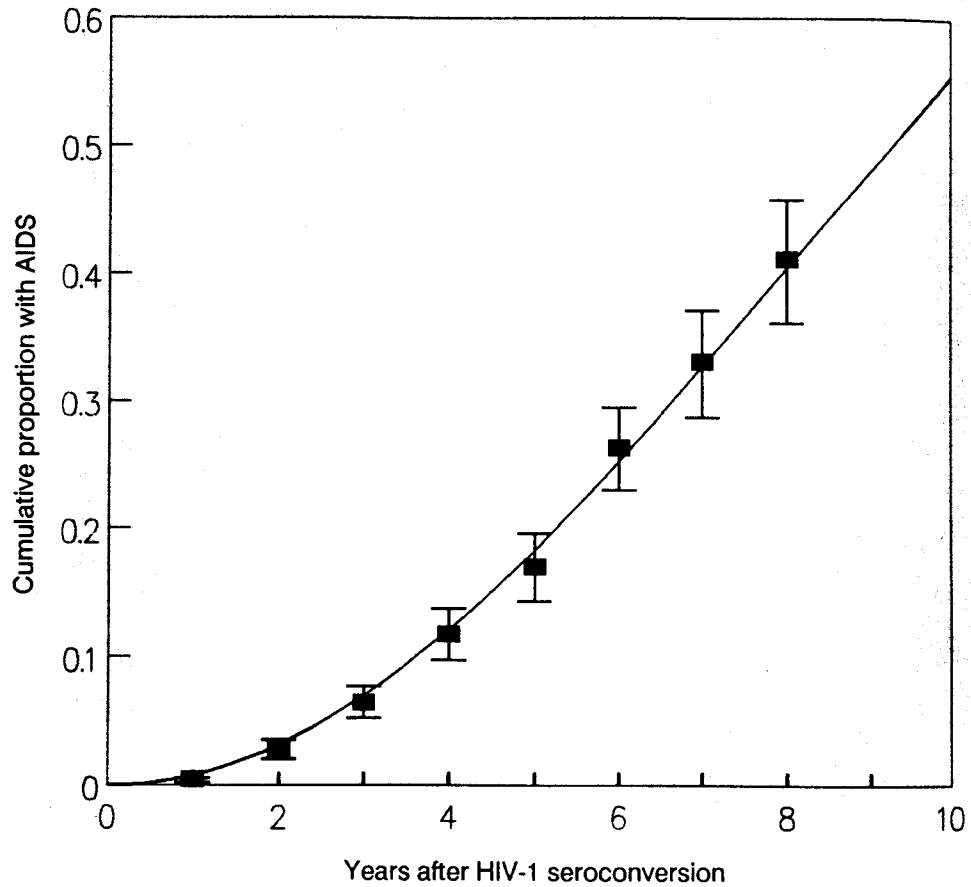

Fig. 4. Estimated cumulative proportion of HIV-infected persons with AIDS in relation to the time since HIV-1 seroconversion. The closed squares (along with 68% confidence intervals) are the combined estimates derived from Table 1. The continuous curve is the Weibull distribution $F(t) = 1 - \exp(-(0.09t)^2)$.

In this study, we have no direct observations on the incidence of AIDS past the 8th year of HIV-1 infection. Still, we can project the incidence beyond that point if we are willing to make parametric assumptions. As Fig. 4 shows, the synthetic estimates are consistent with a two-parameter Weibull cumulative distribution function $F(t) = 1 - \exp(-(0.09t)^2)$, where t is now considered a continuous variable. This two-parameter Weibull model predicts that 50% of HIV-1–infected subjects would have AIDS by 9.25 years. By the 15th year after infection, 84% would have AIDS. The data in Fig. 4 are insufficient to distinguish between this two-parameter Weibull model, in which everyone who is infected would eventually develop AIDS, and a three-parameter model in which some unknown proportion are forever AIDS-free.

Discussion

The duration of the incubation period for HIV-1 may be determined by many factors, including the initial dose of the virus; the initial port of entry; the site of latent infection; the initial or subsequent state of the immune response; the infected person's age; coinfection with other viruses; and reinfection by other strains of HIV. The current study has detected no striking differences in the distribution of HIV-1 incubation among homosexual men and two other groups who presumably received the virus through the blood-borne route. Some men in the San Francisco Clinic Cohort may have been infected by sharing of HIV-1 contaminated needles. Still, the evidence so far does not reveal a difference in the duration of the HIV-1 incubation period between those infected by the blood-borne route and those infected by sexual transmission. If there are as yet unidentified "cofactors" that influence the natural history of HIV-1 infection, the present results suggest that such cofactors are common to diverse risk groups.

This study focused on HIV-1–infected adults. Children with hemophilia may have a different incubation curve (*2*), and have been excluded from the present analysis. In future research, we also need to analyze separately infants who have received infected blood transfusions (*3*) and newborns who have acquired HIV-1 by vertical transmission.

The data in Fig. 4 fit quite closely to a parametric Weibull distribution of the form $F(t) = 1 - \exp(-(\rho t)^2)$. A reasonably good Weibull fit has also been reported for the transfusion data alone, though some authors have suggested the gamma, log-logistic, normal and log-normal distributions (*3, 4, 8*). In the absence of direct observations on AIDS incidence rates for persons infected for over 8 years, such parametric models need to be viewed mainly as within-sample smoothing devices. Their use in predicting the incidence of AIDS in the second decade after HIV-1 infection needs to be cautious and qualified.

It is interesting that the cumulative hazard of AIDS appears to rise with the square of time since initial HIV-1 infection. This mathematical form arises in reliability theory when a device consists of a large number of paired elements. Any single pair "fails" when both of its elements have failed, and the device fails as soon as the first pair fails (*9*). If there are M pairs and if each element of each pair has a constant failure rate ω, then the cumulative distribution of time to failure is $F(t) = 1 - \exp(-(\rho t)^2)$, where $\rho = \omega\sqrt{M}$.

The biological analogy would be that HIV-1 initially infects a large number of host cells (for example, lymphocytes, macrophages, Langerhans cells). AIDS would become manifest after any single infected cell undergoes two separate changes (for example, coinfection with another virus, activation by another foreign antigen, production of cytotoxic proteins). We have estimated $\rho = 0.09$, so if $M = 100$ cells were initially infected, then each of the two subsequent steps would have a transition rate of $\omega = 0.009$ per year.

References and Notes

1. N. Hessol *et al.*, paper presented at the *Third Int. Conf. on AIDS, Washington, DC, Abstr.* (U.S. Dept Health and Human Services and WHO, Washington, DC, 1987); J. W. Curran *et al., Science* **239**, 610 (1988) (also this volume, p. 19).

2. M. Eyster *et al., Ann. Int. Med.* **107**, 1 (1987).

3. G. F. Medley *et al., Nature* **328**, 719 (1987).

4. K. Lui *et al., Proc. Natl. Acad. Sci. U.S.A.* **83**, 3051 (1986).

5. W. H. DuMouchel and J. E Harris, *J. Am. Stat. Assn.* **78**, 293 (1983).

6. J. Harris, *Delay in Reporting Acquired Immune Deficiency Syndrome*, Working Paper Number 452 (M.I.T. Department of Economics, Cambridge, MA, May 1987). Results nearly identical to those in Fig. 3 were obtained by the reverse hazard method of S. W. Lagakos *et al., Nonparametric Analysis of Truncated Survival Data, with Application to AIDS*, Technical Report Number 544Z (Dana-Farber Cancer Institute, Boston, August 1987).

7. One problem in applying Eqs. 3 and 4 is that some of the observed annual incidence rates q_{jt} are zero, so that $\log q_{jt} = -\infty$. In principle, each q_{jt} equals a ratio n_{jt}/N_{jt}, where N_{jt} is the number of subjects at risk in cohort j at date t, and n_{jt} is the number of new AIDS cases in $[t, t+1)$. As a binomial random variable, q_{jt} thus has relative standard error $(1-q_{jt})/(q_{jt}\sqrt{N_{jt}})$, which becomes arbitrarily large as $n_{jt} \rightarrow 0$. In essence, the data points for which $q_{jt} = 0$ contribute no information to the model and can be dropped. In fact, I obtained virtually the same results whether I omitted the data points for which $q_{jt} = 0$ or included them by setting the corresponding n_{jt} to an arbitrarily small positive number.

8. M. Rees, *Nature* **326**, 343 (1987); M. Rees, *ibid.* **330**, 427 (1987).

9. N. L. Johnson and S. Kotz, *Distributions in Statistics: Continuous Univariate Distributions 1* (Wiley, New York, 1970), ch. 20; B. V. Gnedenko *et al., Mathematical Methods in Reliability Theory* (Academic Press, New York, 1968).