

A MATHEMATICAL MODEL OF THE  
SPREAD OF HTLV-III

-Robert H. Evers, Jr., Ph.D.,  
W. Meade Morgan, Ph.D.,  
William W. Darrow, Ph.D., ✓  
Harold W. Jaffe, M.D.  
AIDS Branch, Statistics section

## INTRODUCTION

Since 1978 over 6,800 homosexual men from San Francisco have been enrolled for a study of the prevalence and incidence of hepatitis B. Blood samples from these men have been collected and stored at the Centers for Disease Control in Atlanta. It is now possible to test for antibodies to the virus suspected to cause AIDS. This makes the blood samples a source of valuable information concerning the spread of the virus in the San Francisco community since 1978.

Two random samples from the hepatitis study subjects have been selected for research on AIDS. The first sample consists of fifty percent of an early group enrolled to study prevalence. The second is a six percent sample of a later group enrolled to study the incidence of hepatitis. Six percent was chosen so that the two samples together would comprise a ten percent sample of the entire 6,800 plus subjects. We will refer to these samples as the 50% and 6% samples, respectively.

These data might be used to answer several questions. What percentage of the San Francisco gay subjects have been exposed to the virus? What has been the rate of spread of the virus? What is the probability that a contact between an infected and uninfected person will lead to a new infection? This paper deals with a statistical model which was designed to provide information relevant to these questions.

## METHODS

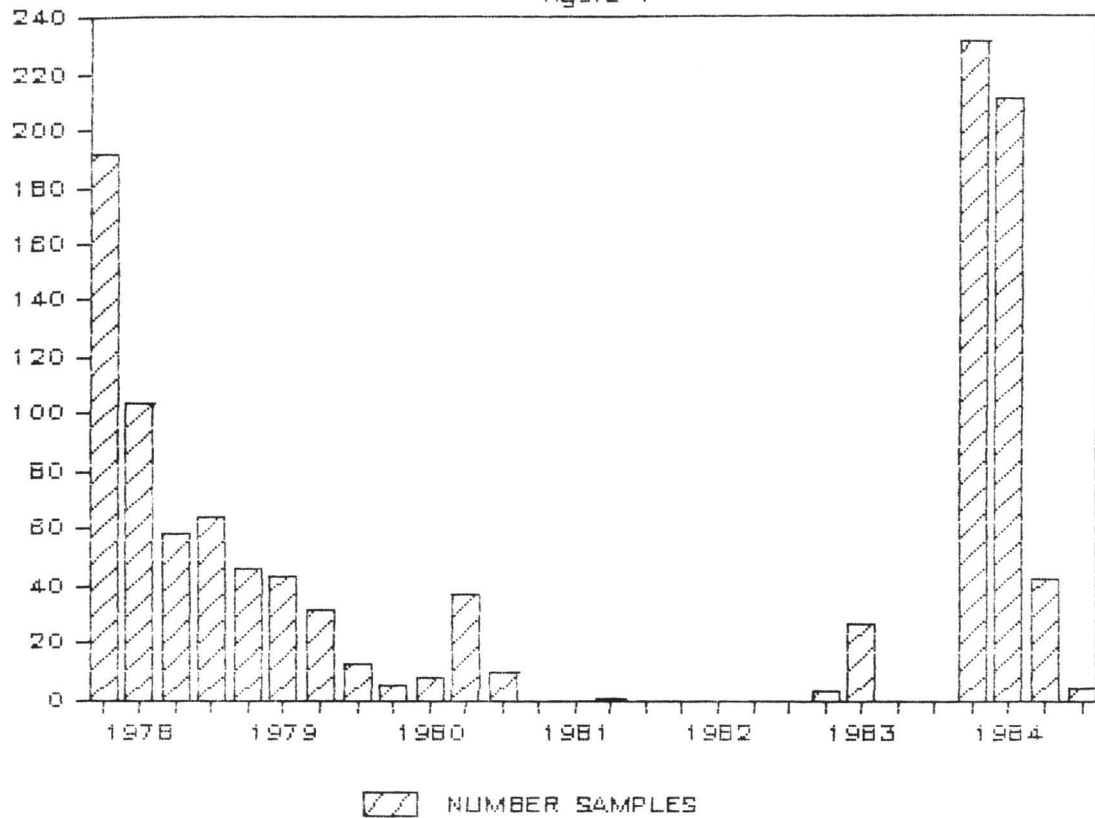
### THE DATA

Even though we are fortunate to have repeated observations over a long period of time, there are difficulties with analyzing these samples. Many subjects were found to have a negative antibody test in 1978 or 1979 and a positive test in 1984. A period in excess of five years remains in which the exposure could have taken place. We assume that the midpoint of the time interval is an adequate guess at exposure time. This gives the impression that there was a sudden increase in the rate of infection during the 1981-1982 period when few samples were taken.

Figure one shows the time of collection of blood samples by quarter. Figure two shows the effect of the assumption that conversion took place midway between the last negative and first positive tests. Gradual spread appears to become explosive toward the end of 1980. If samples were tested during the 1980 to 1984 period, we would expect a curve like the one labeled

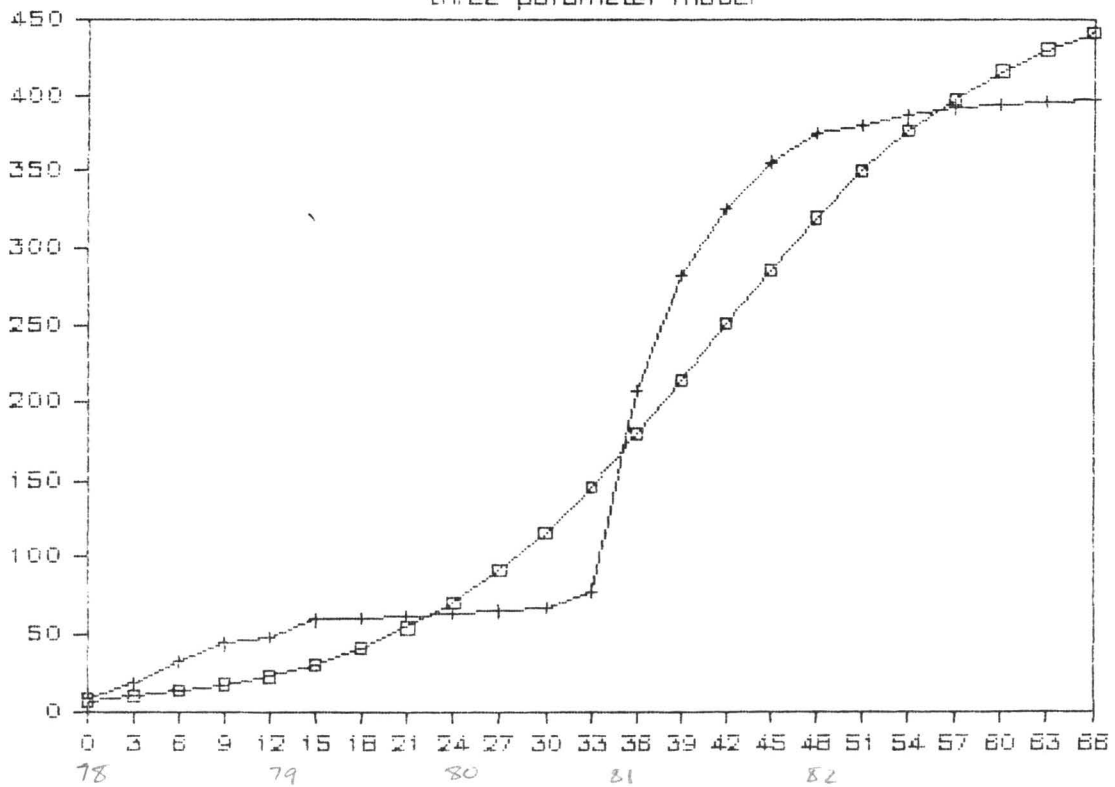
# QUARTER OF SAMPLE COLLECTION

figure 1



## all data

three parameter model



"expected" in figure two.

## THE MODEL

### Assumptions:

Assume that sexual contact between members of the cohort occurs randomly and with uniform frequency, and that the populations being studied are closed with no members entering or exiting. A new infection requires a contact between an infected and an uninfected person. Not every such contact results in infection. The probability of infection is then the product of the probability of contact and the probability of infection given that a contact has occurred. During a given time interval, the number of new infections will be a function of the total number of people infected and the probability that a contact will lead to an infection. Symbolically, let  $P_C$  be the probability of contact and  $P_{IIC}$  be the probability of infection given contact. The model states that the probability of an infection is

$$P_I = P_C P_{IIC}(t).$$

### Derivation:

If there are  $N$  people in the population at risk, and  $Y(t)$  of them are infected at some time  $t$ , then at that time there are  $Y(t)(N-Y(t))$  possible contacts between infected and uninfected people. There are  $N(N-1)/2$  possible contacts disregarding infection status. Then the probability of contact between infected and uninfected persons is:

$$P_C(t) = Y(t)(N-Y(t)) / [N(N-1)/2].$$

The total number of people infected at time  $t$  can then be written as

$$Y(t) = Y(t-1) + P_{IIC}(t)P_C(t)$$

or

$$Y(t) = Y(t-1) + 2P_{IIC} Y(t-1)(N-Y(t-1)) / (N-1)$$

Let  $K = 2P_{IIC} / (N-1)$ . The model can then be written as a differential equation in the usual way:

$$Y'(t) = KNY(t) - KY^2(t).$$

This is a Bernoulli equation with a known general solution. In this case the function  $Y$  is:

$$Y(t) = [1/R - (1/R - 1/Y(0))\exp(-Knt)]^{-1}$$

The parameter R is the number of people who will ultimately become infected. Y(0) is the number infected at the beginning.

The parameters R, Y(0) and P were estimated using the IIC method of least squares. The BMDPAR non-linear regression program performed the computations and gave estimates of the standard deviations of the parameters. The model was fit to the sample of thirty, the 6% sample, the 50% sample, the 6% and 50% samples combined, to all available data, and to the data not belonging to either random sample. The results are summarized in Tables 1 through 3.

The model can be rewritten to relax the assumption that the sample is closed with no subjects entering or leaving. By modeling the percent exposed rather than the number exposed, a model which does not explicitly depend on the sample size can be derived. Let Y(t) be the percent infected until time t. The probability of contact between uninfected and infected subjects is now written

$$P = 2Y(t)(1 - Y(t))$$

and the model is

$$Y'(t) = 2PY(t)(1 - Y(t)) \\ = 2PY(t) - 2PY^2(t)$$

which has as a solution

$$Y(t) = [1/R - (1/R - 1/Y(0))\exp(-2Pt)]^{-1}$$

R and Y(0) now have the dimension of percent exposed. When this model was fit to all available data the estimates were:

$$P = 0.51, \quad \text{s.e.} = 0.008445, \\ R = 77.4\%, \quad \text{s.e.} = 9.9, \\ Y(0) = 1.15\%, \quad \text{s.e.} = 0.56388.$$

When the percents are converted to sample subjects the estimates are the same as for the model in Tables 2 and 3.

## RESULTS

The model was fitted to several sets of samples. One set had thirty observations chosen because many samples had been taken on each subject. The 6% and 50% samples were analyzed separately and together, and the subjects not in any sample were fit. Finally, all data were merged and analyzed. Table I presents the sample sizes, percent infected at last test, and the probability of infection per month given contact between infected and uninfected individuals. The standard errors and 95% confidence limits for P are in the last two columns. The number exposed in the first quarter is fixed at the observed value. The number ultimately exposed and P are the two parameters estimated.

TABLE 1  
Two parameter model

	TOTAL	EXPOSED	%	P	S.E.	95% C. I.
test sample	30	12	40	0.017	0.0007	[.016, .018]
50% sample	257	171	67	0.031	0.0043	[.023, .039]
6% sample	176	106	60	0.039	0.0012	[.036, .041]
6% & 50%	433	277	64	0.042	0.0017	[.039, .046]
all data	614	400	65	0.049	0.0021	[.044, .054]
non-sample	181	123	68	0.067	0.0033	[.061, .074]

DISCUSSION

The column labeled P gives the probability of infection given contact. Since all groups were observed over the same time period, P is highly correlated with the final percent exposed.

Excluding the test sample, the extreme results are between the 50% sample with a P of 0.031 and the non-sample group with a P of 0.067. The final percentages exposed were 67% and 68% respectively. The big difference between these two groups is that the 50% sample had seven positives in the first three month period, while the other groups never had more than three. I ran the groups with a three parameter model, estimating the value of  $Y(0)$ .

TABLE 2  
Three parameter model

	TOTAL	EXPOSED	%	P	S.E.	Y(0)	
						obs.	est.
test sample	30	12	40	0.021	0.0013	1	2
50% sample	257	171	67	0.063 ✓	0.0032	7	1
6% sample	176	106	60	0.039	0.0041	0	4
6% & 50%	433	277	64	0.071	0.0031	7	1
all data	614	400	65	0.051	0.0084	8	7
non-sample	181	123	68	0.047	0.0059	1	4

In the 50% sample the model estimates only one person exposed in the first quarter, and the P value increases from 0.031 to 0.063. In the previous case the number exposed in the first quarter was fixed at seven, the observed value. The P value then accounts for an increase from 7 to 171 exposed over a 45 month period. Estimating only one exposure in the first quarter means a larger P value is required to fit the larger increase from 1 to 171.

One of the parameters estimated was the maximum number ever to become exposed. Table 3 summarizes the results.

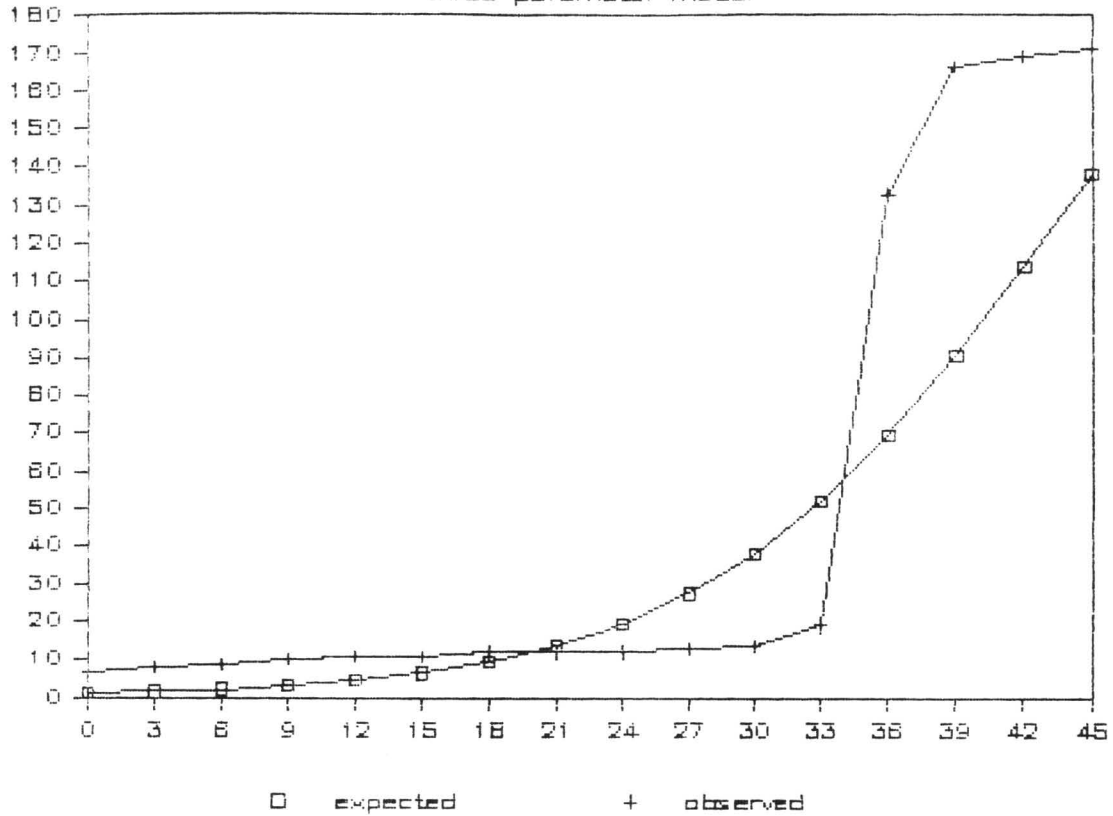
TABLE 3  
 PREDICTED MAXIMUM NUMBER EXPOSED

<u>sample</u>	<u>max. pred.</u>	<u>s.e.</u>	<u>time to achieve</u> (months since 1/1/1978)
50%	257 (100%)	0	88
6%	175 (100%)	0	118
6 & 50%	326 (75%)	38	91
nonsample	135 (75%)	14	179
all data	475 (77%)	60	101 (May, 1986)

The parameters estimated from all the data indicate that the probability of infection given contact is about 5%. In 1984 the exposure incidence rate was 2% per month; in June 1985 it is estimated to become 1.8% per month. The maximum estimated exposure of 77% will be achieved around the middle of 1986.

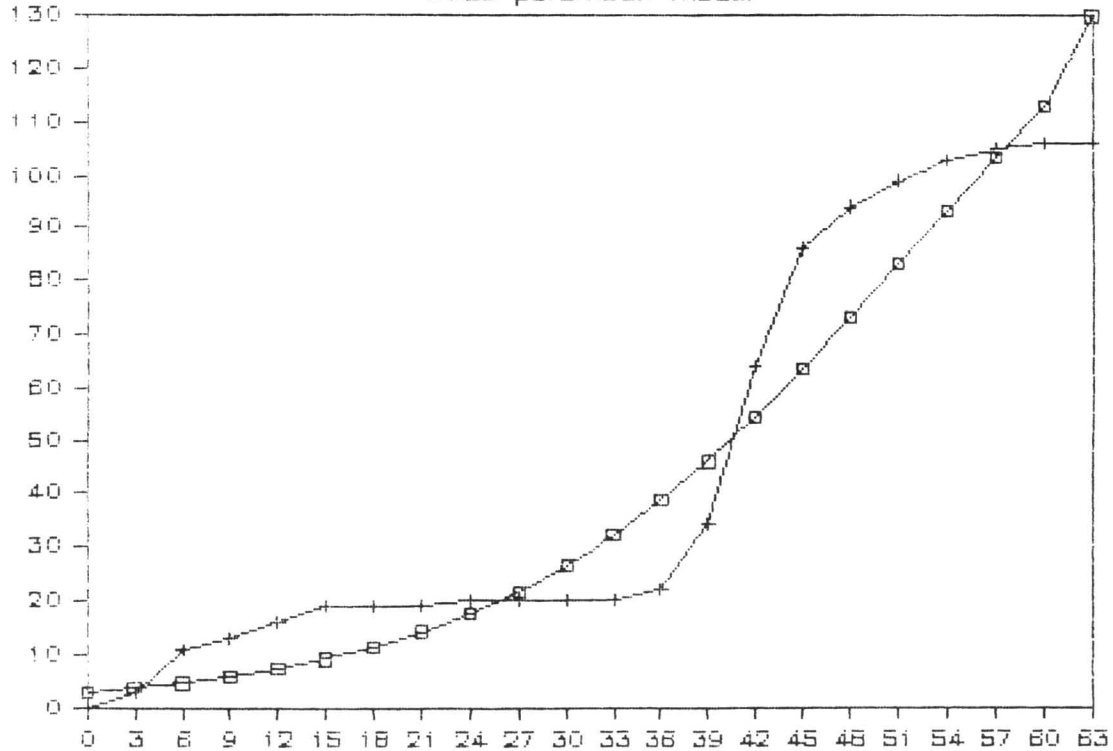
# Fifty percent sample

three parameter model



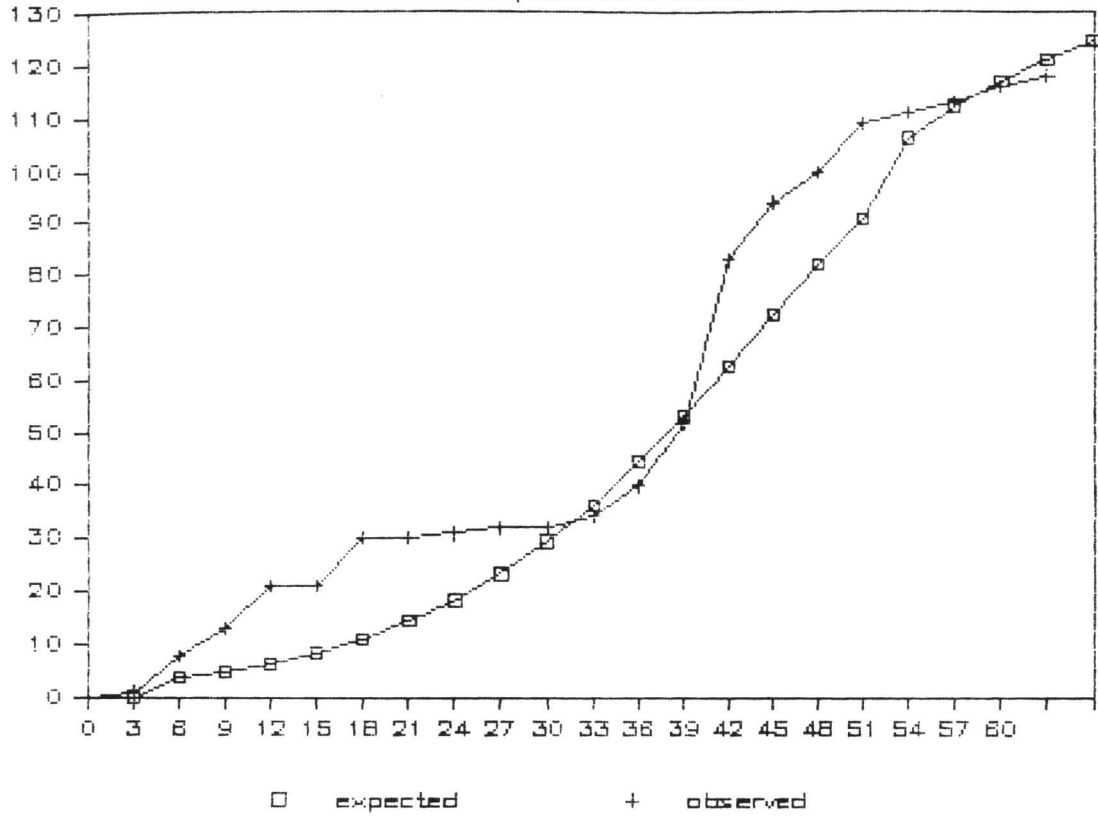
# six percent sample

three parameter model



# non-sample data

three parameter model



# sample of 30

three parameter model

