SYSTEM DYNAMICS AND STATISTICS: RECOVERING THE AIDS INCUBATION TIME DISTRIBUTION FROM RIGHT-CENSORED DATA

Carole Roberts and Brian Dangerfield
University of Salford
SALFORD M5 4WT
U.K.

The identification of the AIDS incubation time distribution, together with its parameters, is a vital component of any mathematical model designed to portray scenarios concerning the future trends in reported AIDS cases. A dataset on Transfusion-Associated AIDS cases in the USA is available and can be utilised in this identification process. However, the task of achieving a best fit using either parametric or non-parametric statistical methods is hampered because, in particular, the data are right-censored and this leads to an extremely complex maximum likelihood estimation procedure. By employing an appropriate system dynamics software tool an optimising simulation approach to the fitting process is available as an alternative. This enables the resolution of a number of complications which hamper the maximum likelihood approach.

Introduction

The distribution of the incubation time for AIDS is a critical feature of the epidemiological projections of the spread of HIV-related disease. The purpose of our work is to analyse the fairly comprehensive Transfusion-Associated (TA-AIDS) dataset provided by the U.S. Centers for Disease Control (CDC) in Atlanta in order to determine a reasonable candidate distribution, together with its parameters, which models the incubation time adequately.

The CDC TA-AIDS dataset contains anonymised records on all AIDS cases in the USA where the subject has been infected by contaminated blood given by transfusion. Because of the relatively accurate identification of the date of primary infection this dataset offers a valuable resource in the search for the incubation time distribution.

Attempts have already been made to identify the distribution. Lui (1986), Medley et al (1987, 1988) and Kalbfleisch and Lawless (1989) have all used the CDC dataset. Costagliola et al (1989) have used French transfusion data while Rees (1987) has employed data derived from the CDC data. Bacchetti and Moss (1989) have instead employed cohort data from certain studies ongoing in San Francisco.

Rees utilises a heuristic method and has concluded that a Normal distribution with a mean around 15 years fits the data well. Most of the other authors take issue with Rees. They have used the statistical method of maximum likelihood to optimise the fit of expected to observed data. In general those on this side of the debate conclude that a Weibull distribution with a mean of between 8 and 9 years offers a best fit, although Bacchetti and Moss, using non-transfusion data, suggest the mean could be as high as 10 or 11 years.

There is a genuine controversy on this aspect of AIDS modelling and it is thus considered useful to offer another viewpoint using an approach and a software tool (DYSMOD/386) that none of the other authors have adopted.

The Data

The CDC Transfusion dataset - as with any data relating to the incubation time of AIDS - presents some major problems for analysis. Inevitably the data is right-censored in that there are many cases of people infected by blood transfusion who have not yet manifested symptoms of clinical AIDS. Added to this it is uncertain what proportion of infectives will ultimately convert to AIDS. Further, the incidence of infective transfusions in the U.S.A. is unknown particularly over the years to March 1985 after which all blood was screened for HIV. As is to be expected, the introduction of screening had an immediate effect. This is apparent from figure 1 which shows a sharp drop in the number of reported cases emanating from infective transfusions administered after 1985 quarter 1.

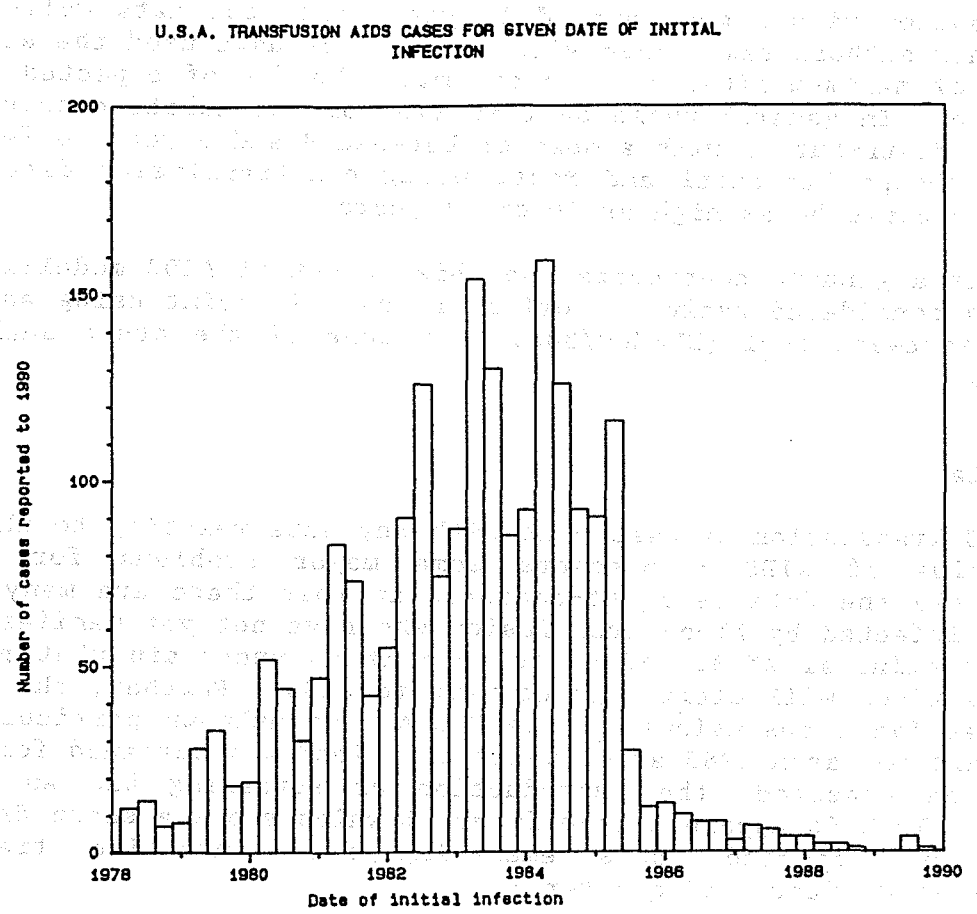

Figure 1: Numbers of Transfusion Associated AIDS cases arising from initial infection on the date shown (taken from 1990 Q2 dataset)

Allied to the effect of blood being screened, there is a belief that people were refraining from giving blood in the year or so up to the point of introduction of screening, arising either from a desire to prevent further infection (by those who knew their HIV status or suspected themselves to be seropositive) or because of a fear of acquiring HIV infection in the process of donating their blood.

The dataset as received from the CDC contained N=3514 records. It was refined by, firstly, removing all those aged 12 years and under (234). Also removed were records for cases diagnosed prior to 1982 (2), those with an uncertain transfusion date (157), an unknown transfusion date (870), and where the transfusion was given pre-1978 (51). Finally, 93 cases showing a diagnosis date prior to the transfusion date were extracted as were 9 cases giving a report date prior to the diagnosis date. The data cleaning exercise left N=2098 usable cases. For the optimisation runs the final six quarters' data (1989 Q1 - 1990 Q2 inclusive) was also deleted (N=5) since it was forcing an unrealistically large number of expected infective transfusions in

1989/90, a result of the algorithm attempting to achieve a fit to such a short run of reported incidences. The final total of reported AIDS cases used in the analysis was, therefore, 2093.

Estimation of the reporting delay

For all individuals recorded on the dataset the time is given (month and year) when CDC received notification that they were confirmed as presenting with clinical AIDS. For most reports the date of diagnosis is also given. Delays in reporting, however, can have a substantial impact on tabulated numbers of cases diagnosed, particularly in recent time periods.

The model attempts to reflect the observed situation. Having estimated the incidence of AIDS cases each quarter derived from one particular quarter's infective transfusions, these estimates are then smoothed using a reporting delay parameter. Thus the model generates the expected incidence of reported cases for each successive quarter following the quarter of infection. A separate analysis was undertaken to estimate a reporting delay parameter before its incorporation into the main model.

The numbers diagnosed and the numbers reported in each quarter were plotted and compared. The numbers diagnosed were smoothed and the smoothing parameter optimised by minimising an objective function such that the smoothed incidence of diagnosed cases gave a good fit to the quarterly reported incidence. The resultant value for the smoothing parameter was then input into the main model.

The most recent CDC dataset gives reported cases up to end-June 1990. However, because of delays in submission of case notifications, the reporting delay parameter was estimated using only cases reported as diagnosed up to and including 1988 quarter 4. Fewer outstanding notifications are likely for time points at least one and a half years prior to the release date of the dataset. Both the complete CDC public dataset on all reported AIDS cases and the TA-AIDS dataset were analysed separately and it was found that the mean reporting delay of 0.57 years for transfusion cases was rather higher than that for the CDC public dataset (0.45 years). Other researchers have already pointed out this feature. Because our reporting delay distribution is exponential, a figure of 0.57 years for the mean implies that 35.9% of cases are reported within 3 months and 83.1% within one year. This compares with figures of 50% and 85% quoted by the CDC (CDC, 1991).

Description of the model

The model projects the quarterly incidence of AIDS from an assumed incidence of infective transfusions. The incidence of infection is modelled using an exponential function, typical for the beginning stages of an epidemic. The functional form employed is y= a exp(bt), the parameters a and b being two of those to be optimised. This exponential function is multiplied by a table function which incorporates a set of factors whose range is between 0 and 1. It operates from the beginning of 1983 through to the end of 1988 at which point it is assumed that the incidence of infective transfusions has fallen to zero. The factors are designed to capture the extent of the reduction in infective transfusions towards and after quarter 1 1985. The expected number of infective transfusions per quarter is estimated by successively integrating the exponential incidence function to produce separate quarterly totals, one each for every quarterly cohort of initial infectives, 44 cohorts in all: the model yields in excess of 450 equations in total.

Having established the number of infective transfusions for each quarter, this figure is then multiplied by the candidate incubation time density function and by SER, the Symptoms Emergence Ratio - being the proportion of infectives who will ultimately progress to AIDS. Thus we can estimate the incidence of AIDS cases each quarter derived from one particular quarter's infective transfusions. This estimated incidence of AIDS diagnoses is then smoothed according to the previously determined reporting delay parameter to generate the expected incidence of reported AIDS cases for each successive quarter following the quarter of infection. These expected numbers are then compared with the observed numbers from the CDC dataset, input in the form of a set of 44 table functions. The comparison of expected and observed is done only at the discrete quarterly points and is terminated by LENGTH, the duration of the simulation, chosen to coincide with the date to which the latest CDC data is available. Thus the right-censoring of the reported data is assimilated automatically by ceasing the simulation!

The objective function, which in these preliminary experiments is restricted to the sums of squared deviations between the expected and observed quarterly incidence of reported AIDS cases arising from all cohorts of initial infectives, is minimised over 1000 iterations of the model. This takes approximately two hours on a 386-SX PC equipped with a maths co-processor. In the process the model simultaneously optimises the following unknown parameters:-

(i) those relating to the form of the curve describing the early incidence of infective transfusions

(ii) a set of multipliers (a table function) which captures the extent of the reduction of infective transfusions towards and after quarter 1 1985

(iii) the proportion of infectives actually converting to AIDS

(iv) the parameters of the candidate incubation distribution.

Some Results

The optimisation experiments are at an early stage and presented here are some initial results. Four distributions have so far been considered as possible candidate incubation time density functions: the Erlang types 2 and 3, the Weibull and the Normal.

The density functions for the distributions are:

1. The Erlang

```
DF.K=((TYPE/MIP)**TYPE/2)*(TIME.K)**(TYPE-1)*EXP(-(TYPE/MIP)*TIME.K)
```

where MIP is the mean incubation period in years. For a type 2 Erlang TYPE/2 is replaced by TYPE.

2. The Normal

```
DF.K=(1/(SQRT(2*PI)*SIGMA))*EXP(-0.5*ABS((TIME.K-MIP)/SIGMA)**2)
```

were MIP is as above. SIGMA, the standard deviation, was set at MIP/3. This allows for very short incubation times but virtually rules out the possibility of negative incubation times occurring.

3. The Weibull

```
DF.K=SHAPE*SCALE**SHAPE*TIME.K**(SHAPE-1)*EXP(-(SCALE*TIME.K)**SHAPE)
```

where SHAPE is the shape parameter and SCALE the scale parameter.

Figure 2 illustrates the fit resulting from the use of each of the optimised incubation distributions.

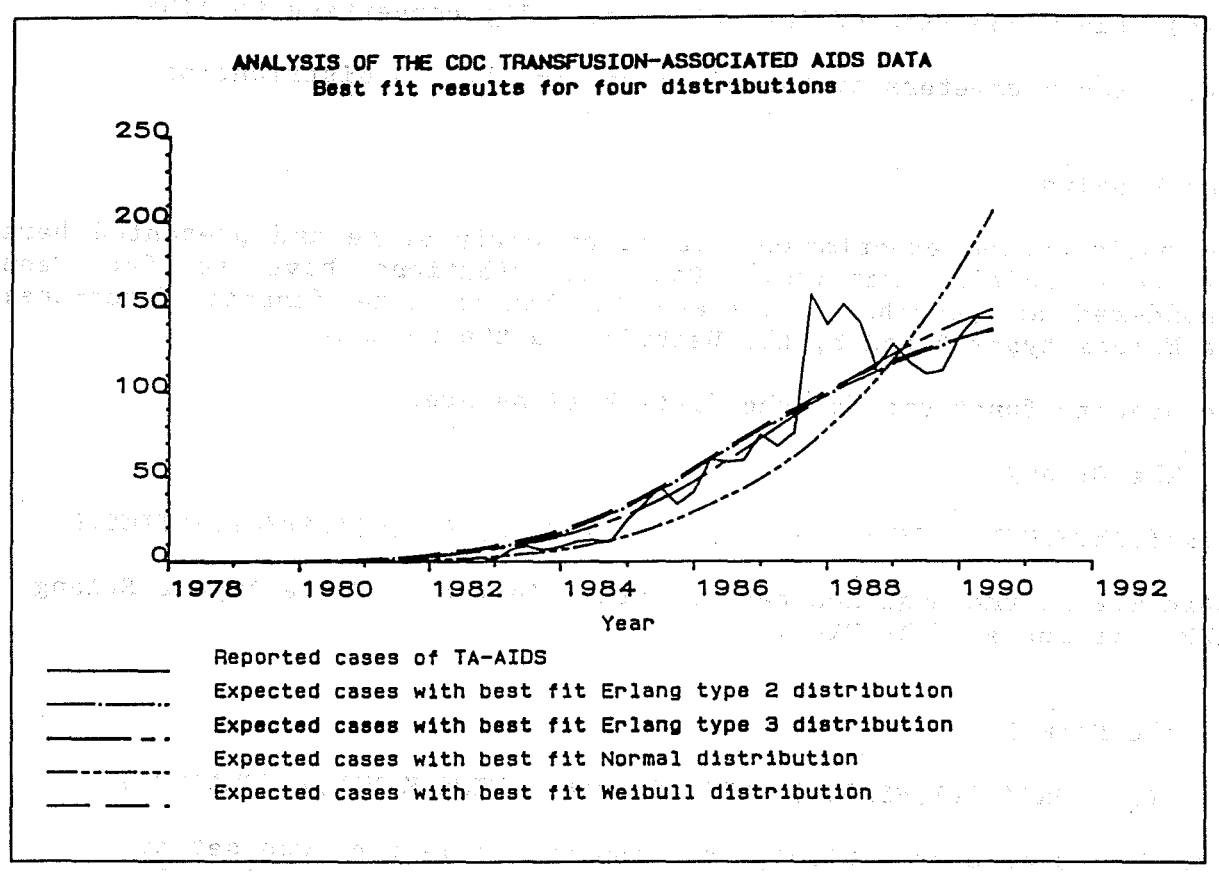

Figure 2: The reported quarterly incidence of TA-AIDS and the fitted expected number of cases arising from the optimised distributions

Some numerical results from the optimisations are shown in table 1 below.

Candidate Distribution	Optimised Parameter values	Minimum sums of squares	Expected number of infected transfusions up to 31 Mar 1985
Erlang Type 2	MIP=19.80 SER=0.994	3334	3506
Erlang Type 3	MIP=10.92 SER=0.960	3436	1890
Normal	MIP=14.32 SD = 4.77 SER= 1.0	4250	8715
Weibull	SHAPE=1.97 SCALE=0.10 SER=0.718 (MIP=8.9)	3315	2016

Table 1: Optimised parameter values and associated results for four distributions

As a comparison, the number of infected transfusions reported so far prior to the end of quarter 1 1985 (when screening of blood was introduced) is 1,986 in the dataset we are using.

In the current version of the model, the array of multiplication factors handling the reduction in infective transfusions has been allowed to fall to values < 1 from 1983. Hence our model outputs substantially lower numbers of estimated infective transfusions up to 31 March 1985 than those estimated by other authors. Future work will include investigations into alternative forms for modelling the decay in the rate of infective transfusions since our current model may be reducing the numbers too quickly.

The Weibull gives the "best fit" to the data as measured by the sums of squares objective function. However, both the Weibull and the Erlang type 3 optimisations are not entirely satisfactory from the point of view of the associated number of infected transfusions generated by the model. In the case of the Erlang type 3 the number generated by the end of March 1985 does not even reach the numbers actually reported so far; the Weibull barely does. However, as noted above, the number of infective transfusions may be decaying too rapidly.

Nevertheless, on the basis of our current results, attention moves to the optimisations on the Erlang type 2 and the Normal which more realistically project the incidence of infective transfusions. Of these the Erlang type 2 achieves the smaller sum of squares so on this criterion would appear to be the stronger candidate. Examination of figure 2 shows clearly that the Normal distribution yields a poor fit to the reported cases of TA-AIDS so far. It is worth noting that both the Erlang type 2 and the Normal models produce longer mean incubation periods which in turn implies a higher incidence of infected transfusions than the Erlang type 3 and the Weibull. Hence they do present a more pessimistic view of the overall size of the epidemic. Over all the optimisations, the lowest value of SER obtained was 0.72 .

Figure 3 shows a comparison of the density functions for the four optimised distributions. These have been plotted with SER=1 and illustrate how the incubation distribution applies to a single cohort all infected during the same quarter.

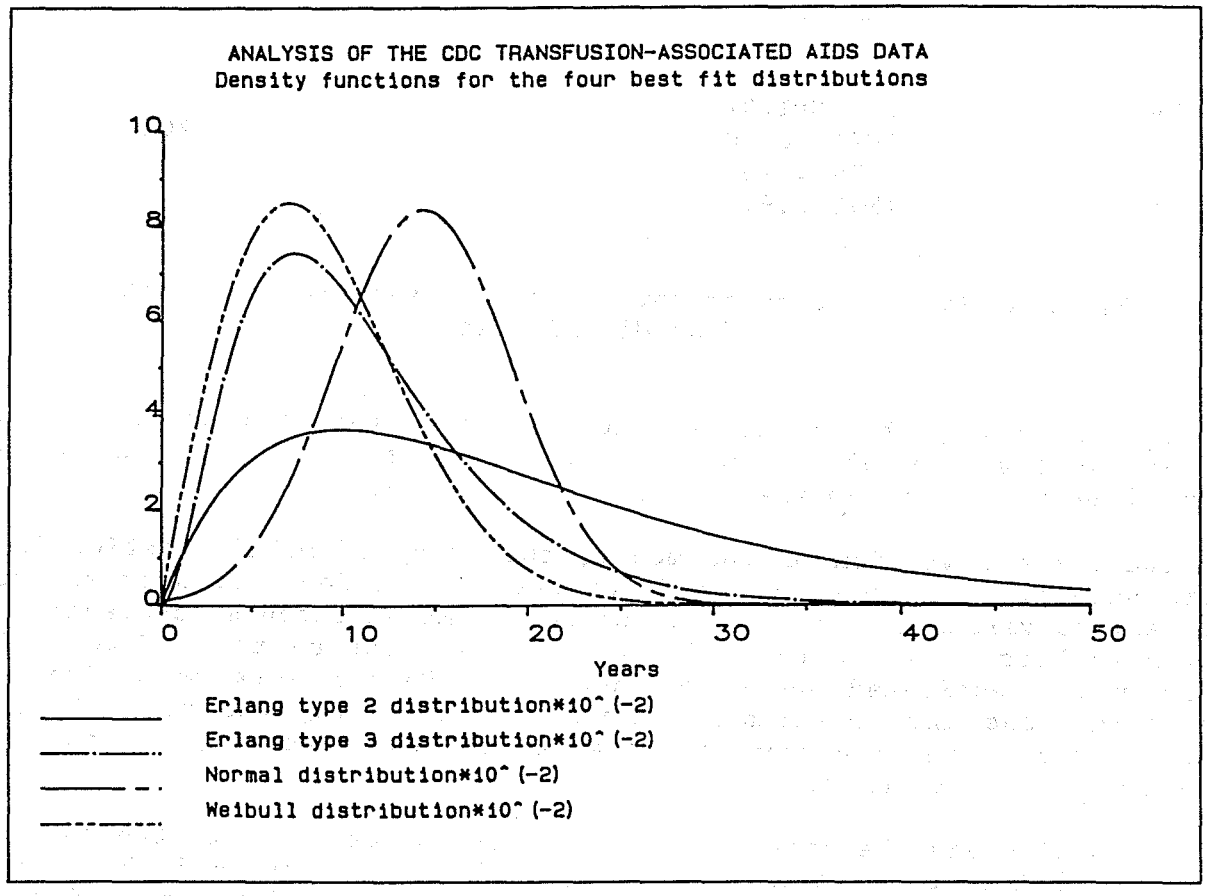

Figure 3: Comparison of the density functions of the optimised distributions

In common with all search routines used for exploring multidimensional

parameter space, there is the possibility that a local rather than a global optimum for the objective function is obtained. One way of investigating this is to change the starting values for the parameters to be optimised. A change in starting values for the Normal distribution for example, produced a mean equal to 8.03 years, an objective function of 4567 and an expected number of infected transfusions up to the end of quarter 1 1985 of 1,693 cases. This result also serves to illustrate the identifiability problems and arbitrary constant of proportionality as discussed, for example, by Kalbfleisch and Lawless (1989). The transfusion data themselves are unable to discriminate between very high infection rates and longer incubation times on the one hand and low infection rates and shorter incubation times on the other. As more data becomes available ultimately the identifiability problem will disappear.

It can be seen that with the limited number of optimisation experiments undertaken so far, our results imply that distributions with low mean incubation times are associated with an unrealistically low incidence of infective transfusions. Whilst it is true that it makes little difference to the projected course of the epidemic which actual distribution is used for the incubation time, the projection is significantly affected by the value chosen for its mean. If, as Rees (1987) has proposed, this mean is of the order of 15 years rather than the currently more popular belief that it is between 8 and 10 years, the numbers already HIV-positive and the overall size of the AIDS epidemic are of an order of magnitude larger than that presently predicted. So far, our results appear to be more supportive of the view of the mean incubation time taken by Rees although not of the Normal distribution as he has proposed, but it must be emphasised that our research on this aspect is at an early stage.

Conclusion

Even from the results obtained thus far, it seems clear that a system dynamics software tool and model format offers a genuinely useful alternative approach to the statistical problem of estimating best fit probability distributions when the underlying data is right-censored. The authors intend to conduct further experiments to encompass other distributions and other best fit criteria as well as investigating constrained optimisations and the effects of different starting values for parameters, together with alternative forms which capture the reduction in infective transfusions and with variations to the step multiplier (a system parameter of the DYSMOD search routine). Provision of confidence intervals on estimates is also envisaged. This will provide extra information in the choice between candidate distributions.

REFERENCES

Bacchetti, P. and A.R. Moss. 1989. Incubation Period of AIDS in San Francisco. *Nature*. 338: 251-253.

Centers for Disease Control. 1991. HIV/AIDS Surveillance.

Costagliola, D., J.Y. Mary, N. Brounard et al. 1989. Incubation time for AIDS from French Transfusion-Associated Cases. *Nature*. 338: 768-769.

Kalbfleish, J.D. and J.F. Lawless. 1989. Inference Based on Retrospective Ascertainment: an Analysis of the data on Transfusion-Related AIDS. *Journal of the American Statistical Association*. 84(406): 360-372.

Lui, K.J., D.N. Lawrence, W.M. Morgan et al. 1986. A Model-based Approach for Estimating the Mean Incubation Period of TA-AIDS. *Proceedings of the National Academy of Science (USA)*. 83: 3051-3055.

Medley, G.F., R.M. Anderson, D.R. Cox et al. 1987. Incubation Period of AIDS in Patients Infected via Blood Transfusion. *Nature*. 328: 719-721.

Medley, G.F., L. Billard, D.R.Cox et al. 1988. The Distribution of the Incubation Period for the Acquired Immunodeficiency Syndrome. *Proceedings of the Royal Society of London, Biol*. 233: 367-377.

Rees, M. 1987. The Sombre View of AIDS. *Nature*. 326: 343-345.