Listy Biometryczne - Biometrical Letters
Vol. 33 (1996), No. 2, 71-77

Surveillance of AIDS, a forecasting approach to adjusting for reporting delays

Antonino Salvaggio

Istituto di Igiene e Medicina Preventiva, Università degli Studi di Milano,
Via F.Sforza 35, 20122 Milano, Italy

SUMMARY

The author presents a forecasting model to adjust for reporting delays acquired immune deficiency syndrome (AIDS) surveillance data. This model allows adjustment as well as forecasting, and an easy treatment of situations in which calendar time of diagnosis and reporting delay for all incident cases are not available. It has been employed to analize AIDS incidence data reported from Lombardia, estimating inflating factors up to five years, and adjusting AIDS incidence counts.

KEY WORDS: AIDS incidence, forecasting, reporting delay, statistical model.

1. Introduction

AIDS incidence data derived from surveillance systems require adjustment for notification delays. Healy (1988) and Healy and Tillett (1988) considered this problem according to an empirical approach, by random sampling an empirical delay distribution based on past data. Conversely, a rather formal, strictly parametric method was proposed by Cox and Medley (1988), who modelled delay according to a mixture of two gamma distributions. Zeger, Lai-Chu and Diggle (1989), and Brookmeyer and Damiano (1989) introduced practical correction methods based on the Poisson regression and a product multinomial model, respectively. Rosenberg (1990) discussed a way to obtain maximum likelihood estimates for Brookmeyer and Damiano's model using a non iterative procedure. All these methods require counts obtained from cross-classification of incident cases according to calendar period of diagnosis and length of reporting delay.

In this paper, we present a model allowing adjustment and forecasting, and an easy treatment of incomplete cross-classified incident data. According to this model, AIDS incidence data reported from Lombardia have been adjusted by considering only few, recent reported series of incident cases (presented in Table 1).

2. Probability model and statistical methods

With respect to AIDS cases recognized during a (calendar) time interval i, we have defined *AIDS* $(i, i+j-1)$ as the number of cases reported up to the interval $i+j-1$. Also, having settled a horizon after n time intervals, beyond which we consider the probability of further reports to be negligible, we defined $\pi_0, \pi_1, ..., \pi_j, ..., \pi_n$ as the probabilities of reporting during each one of these relevant time intervals.

Thus, given the number of cases reported up to the interval $i+j-1$, the expected number of cases reported during the interval $i+j$ will be

$$\begin{aligned} \mathrm{E}[\Delta_{AIDS}(i, i+j) \quad &| \quad AIDS(i, i+j-1), \pi_0, \pi_1, ..., \pi_j] = \\ &= \quad AIDS(i, i+j-1) \cdot (\pi_j / \sum_{k=0}^{j-1} \pi_k) = \\ &= \quad AIDS(i, i+j-1) \cdot \beta_j, \end{aligned} \tag{1}$$

where $\Delta_{AIDS}(i, i+j)$ denotes the number of cases reported during the $(i+j)$-th interval (reports), and E[·] is the expectation operator. (We may substitute $\exp(\beta'_j)$ for β_j to ensure positive probability ratios.)

Furthermore, according to a forecasting, 'Bayesian' approach (Smith, 1979; West, Harrison and Migon, 1985; Smith and Miller, 1986), we have assumed the expected (total) number of AIDS cases, $\mathrm{E}[AIDS(i, i+n) \mid AIDS(i, i+m), \pi_0, \pi_1, ..., \pi_m]$, to be distributed according to the gamma model $G\{AIDS(i, i+m), \sum_{k=0}^{m} \pi_k\}$. This particular distribution was chosen because it is absolutely non-informative (degenerate) for $m < 0$, and gradually more informative up to the collection of all reports. Moreover, its expected value, $AIDS(i, i+m) \cdot (1/\sum_{k=0}^{m} \pi_k)$, and its variance, $AIDS(i, i+m) \cdot (1/\sum_{k=0}^{m} \pi_k)^2$, are suitable for inference about Poisson processes (as if data were collected during a fraction of a 'unit of time' numerically equal to $\sum_{k=0}^{m} \pi_k$).

Finally, we have assumed the number of reports, $\Delta_{AIDS}(i, i+j)$, conditional on its expected value, to be distributed according to a Poisson distribution, whose continuous parameter, i.e. the expected number of reports $\mathrm{E}[\Delta_{AIDS}(i, i+j) \mid \cdot]$, is related to the expected (total) number of AIDS cases $\mathrm{E}[\Delta_{AIDS}(i, i+n) \mid \cdot]$ according to the expression

$$\begin{aligned} \mathrm{E}[\Delta_{AIDS}(i, i+j) \quad &| \quad AIDS(i, i+j-1), \pi_0, \pi_1, ..., \pi_j] = \\ &= \quad \mathrm{E}[\Delta_{AIDS}(i, i+n) \mid AIDS(i, i+j-1), \pi_0, \pi_1, ..., \pi_{j-1}] \cdot \pi_j. \end{aligned} \tag{2}$$

From the properties of thė gamma distribution (Feller, 1966, p. 47-48), it follows that also the expected number of reports, $\mathrm{E}[\Delta_{AIDS}(i,i+j) \mid AIDS(i,i+j-1),\pi_0,\pi_1,...,\pi_j]$, will follow a gamma model: $G\{AIDS(i,i+j-1),(\sum_{k=0}^{j-1}\pi_k)/\pi_j\} \equiv G\{AIDS(i,i+j-1),\beta^{-1}\}$.

Thus, being $\Delta_{AIDS}(i,i+j)$ distributed according to Poisson, and its expected value according to a gamma distribution (distribution that resumes the information collected up to the interval $i+j-1$), the 'unconditional' distribution of $\Delta_{AIDS}(i,i+j)$, for $j \geq 1$, will be negative binomial (Smith, 1979; West, Harrison and Migon, 1985; Smith and Miller, 1986), according to the expression

$$\begin{aligned} \mathrm{p}[\Delta_{AIDS}(i,i+j) \quad & \mid \quad AIDS(i,i+j-1),\pi_0,\pi_1,...,\pi_j] = \\ & = \mathrm{p}[\Delta_{AIDS}(i,i+j) \mid AIDS(i,i+j-1),\beta_j] = \\ & = \binom{AIDS(i,i+j-1)+\Delta_{AIDS}(i,i+j)-1}{\Delta_{AIDS}(i,i+j)} \cdot \\ & \cdot(\beta_j^{-1})^{AIDS(i,i+j-1)} \cdot [1+\beta_j^{-1}]^{-[AIDS(i+j-1)+\Delta_{AIDS}(i,i+j)]}, \end{aligned} \tag{3}$$

where the first multiplicative term on the right is a binomial coefficient. The first equality in (3) stresses the independence of the distribution from future events: even the arbitrary chosen horizon is uninfluential, being report probabilities introduced only as ratio of one another, $\beta_j \equiv \pi_j / \sum_{k=0}^{j-1}\pi_k$.

Given the cases reported during a first time-interval, say i, the probability of a series of subsequent reports, depending on $\beta_j \equiv \beta(i,j)$, for $j = 1,...,m(m \leq n)$, will be the product of m probabilities like the one defined in (3). The probability of multiple series will simply be the product of the probabilities of each component series. Therefore, if we consider β_j to be independent of i, i.e. $\beta(i,j) \equiv \beta_j$, it will be easy to obtain estimates of β_j according to the maximum likelihood method.

3. Example

We have adjusted for reporting delay the numbers of AIDS cases notified to the Regione Lombardia (from 1 January 1983) through 31 December 1992. Lombardia, with 8,886,402 residents on 1 January 1988, is the Italian region most seriously affected by AIDS. Table 1 presents, according to the calendar time of diagnosis, AIDS cases reported through 31 December 1990, 1991 and 1992, respectively.

According to the model proposed, a point estimate of the expected number of AIDS cases recognised during any i-th interval, on the basis of reports recorded up to time $i+j-1$, has been computed as

Table 1. Calendar period of diagnosis (year) vs annual reports summation for AIDS cases from Lombardia, through 31 December 1992

Calendar period of diagnosis	Evaluation of reporting			Adjusted AIDS count	
	31-12-90	31-12-91	31-12-92	estimate	c.i. 95%*
1983	2	2	3	3	(3-3)
1984	11	11	12	12	(12-12)
1985	82	85	85	85	(85-85)
1986	181	182	182	182	(182-182)
1987	365	370	371	371	(371-378)
1988	558	565	566	569	(566-576)
1989	817	831	834	843	(836-860)
1990	786	907	924	940	(930-963)
1991	-	944	1140	1182	(1158-1218)
1992	-	-	918	1125	(1054-1207)

*according to the confidence limits of the corresponding inflation term (obtained by the bootstrap, percentile method)

$$\begin{aligned} \mathrm{E}[AIDS(i, i+n) \quad &| \quad AIDS(i, i+j-1), \pi_0, \pi_1, ..., \pi_{j-1}] = \\ &= \quad AIDS(i, i+j-1) \cdot (1/\sum_{k=0}^{j-1} \pi_k), \end{aligned} \tag{4}$$

where the rightmost multiplicative 'inflation' term was obtained as a function of the coefficients β_j, $1/\sum_{k=0}^{j-1} \pi_k = \prod_{k=0}^{n-j}(\beta_{n-k} + 1)$. Confidence limits for each inflation term were calculated according to the bootstrap, percentile method (DiCiccio and Romano, 1988), by sampling with replacement the standardized residuals from expected values, normalised with respect to the expected variances (Hinkley, 1988, in particular p.330-332).

Table 1 presents, besides base data, adjusted AIDS incidence counts according to the estimated inflation terms, and their lower and upper confidence limits.

Table 2 presents the estimates of the coefficients β_j, b_j(± s.e.), as well as the estimated inflation terms to be used for reporting delay adjustment (and their bootstrap calculated confidence intervals).

4. Discussion

It is possible to demonstrate that, given the total number of AIDS cases reported for each interval i, the present model and the multinomial one proposed by Brookmeyer and Damiano (1989), and Rosenberg (1990), produce equivalent results. (In fact, by

Table 2. Estimates of one year multiplicative coefficients and inflation terms (to fifth year multipliers), according to the year of reporting evaluation (the latest year whose reports are added up, with respect to the year of diagnosis)

Year of reporting evaluation*	One year multiplier $b_j \pm$ s.e.(b_j)†	Inflation term, $\prod_{k=0}^{5-j}(b_{5-k}+1)$ (c.i. 95%) ‡
$j-1=0$	0.1832 ±0.0111	1.225 (1.148 – 1.314)
1	0.0180 ±0.0032	1.036 (1.015 – 1.068)
2	0.0072 ±0.0023	1.017 (1.006 – 1.042)
3	0.0065 ±0.0026	1.010 (1.002 – 1.031)
4	0.0036 ±0.0025	1.004 (1.000 – 1.017)

* with respect to the year of diagnosis, $(i+j-1)-i$;

† the mean residual deviance was 2.92 and this been allowed for in standard errors;

‡ 95% confidence intervals by the bootstrap, percentile method; 10,000 replications.

factoring the multinomial distribution into binomial terms (McCullagh and Nelder, 1989, p. 170), the differences between the two models result to be only in binomial coefficients, which in both models do not depend on the unknown parameters).

However, the present model, as it is formulated so as not to depend on future information, is naturally apt to be used not only to estimate parameters, but also to monitor reports. The expected distribution for the number of cases reported during the time interval $(i+j)$, conditional on the number of cases reported up to the interval $(i+j-1)$, will be negative binomial, according to the probabilities definite in expression (4). The expected value and the variance of the number of reported cases will be

$$\begin{aligned} \mathrm{E}[\Delta_{AIDS}(i,i+j) \quad &| \quad AIDS(i,i+j-1),\pi_0,\pi_1,...,\pi_j] = \\ &= \quad AIDS(i,i+j-1)\cdot(\pi_j/\sum\nolimits_{k=0}^{j-1}\pi_k), \end{aligned} \tag{5}$$

$$\begin{aligned} \mathrm{Var}[\Delta_{AIDS}(i,i+j) \quad &| \quad AIDS(i,i+j-1),\pi_0,\pi_1,...,\pi_j] = \\ &= \quad AIDS(i,i+j-1)\cdot(\pi_j/\sum\nolimits_{k=0}^{j-1}\pi_k) + \\ &\quad\quad AIDS(i,i+j-1)(\pi_j/\sum\nolimits_{k=0}^{j-1}\pi_k)^2. \end{aligned} \tag{6}$$

These last expressions suggest that we may consider the model in an iterative weighted regression framework. In fact, when the variance depends on the expected value and on no other unknown parameter, negative binomial models may be treated as members of the family of generalised linear models (McCullagh and Nelder, 1989). So, using software able to treat these classes of problems, the implementation of the model will be particularly easy. (According to a notation by now classic (McCullagh and Nelder, 1989), omitting the index i, and naming x_j the (known) number of cases reported up to the time interval $(i+j-1)$, the expected number of cases reported during the interval $(i+j)$ will be $\mu_j = \beta_j x_j$, and the variance will be $V(\mu_j) = \mu_j + \mu_j^2/k_j$, where k_j (known) is, again, $AIDS(i, i+j-1)$.)

The parametrisation we have chosen, i.e. according to the coefficients β_j, allows us to maintain a direct relation with the available observations, even if cross-classification of incident cases according to the calendar period of diagnosis and to the length of reporting delay is incomplete. The parameters depend only on ratios of cases recorded in subsequent intervals. So, we will be able to easily and efficiently utilize the information available in many data structures.

References

Brookmeyer, R., Damiano, A. (1989). Statistical methods for short-term projections of AIDS incidence. *Statistics in Medicine* **8**, 23-34.

Cox D.R., Medley, G.M. (1988). A maximum likelihood method of prediction in the presence of reporting delay. In: Short-term prediction of HIV infection and AIDS in England and Wales. London: Her Majesty's Stationery Office, 58-59.

DiCiccio, T.J., Romano, J.P. (1988). A review of bootstrap confidence intervals. *Journal of the Royal Statistical Society*, Series B, **50**, 338-354.

Feller, W. (1966). An introduction to probability theory and its applications. Vol. II. 2nd ed. New York, Wiley.

Healy, M.J.R. (1988). Extrapolation forecasting. In: Short-term prediction of HIV infection and AIDS in England and Wales. London: Her Majesty's Stationery Office, 60-61.

Healy, M.J.R., Tillett H.E. (1988). Short-term extrapolation of AIDS epidemic. *Journal of the Royal Statistical Society*, Series A, 50-61.

Hinkley, D.V. (1988). Bootstrap methods. *Journal of the Royal Statistical Society,* Series B, **50**, 321-337.

McCullagh, P., Nelder, J.A. (1989). *Generalized linear models.* 2nd ed. London: Chapman & Hall.

Rosenberg, P.S. (1990). A simple correction of AIDS surveillance data for reporting delays. *Journal of Acquired Immune Deficiency Syndrome* **3**, 49-54.

Smith, J.Q. (1979). A generalization of the Bayesian steady forecasting model. *Journal of the Royal Statistical Society,* Series B, **41**, 375-387, Smith, R.L., Miller, J.E. (1986). A non-Gaussian state-space model and application to prediction of records. *Journal of the Royal Statistical Society,* Series B, **48**, 79-88.

West, M., Harrison, P.J., Migon, H.S. (1985). Dynamic generalized linear models and Bayesian forecasting. *Journal of the American Statistical Association,* **80**, 73-97.

Zeger, S.L. Lai-Chu, S., Diggle, P.J. (1989). Statistical methods for monitoring the AIDS epidemic. *Statistics in Medicine* **8**, 3-21.

Received 5 December 1995

Prognozowanie częstości AIDS z uwzględnieniem przypadków opóźnionych zgłoszeń

Streszczenie

Autor prezentuje model prognostyczny który dopuszcza możliwość opóźnionego zgłaszania przypadków AIDS. Model pozwala zarówno na dostosowanie obliczeń do występujących opóźnień jak i na przewidywanie liczby zachorowań. Umożliwia także obliczenia w sytuacji gdy data diagnozy i opóźnienie nie są znane dla wszystkich zachorowań. Model został wykorzystany do analizy przypadków AIDS zanotowanych w regionie Lombardia (Włochy).

Słowa kluczowe: częstość AIDS, prognozowanie, opóźnione zgłoszenia, model statystyczny