61st Annual Conference of the New Zealand Statistical Association

held in conjunction with the

Statistical Methodologies Symposium celebrating the work of Professor Chin-Diew Lai

MASSEY UNIVERSITY

Palmerston North, New Zealand
Tuesday 29 June to Thursday 1 July, 2010

Jonathan Godfrey (Editor)
June 23, 2010

Tuesday 29 June

Time	Stream 1	Stream 2	Stream 3
9:00		**Welcome**	
9:20	Balakrishnan, N. (Bala): 12 "Some Cure Rate Models and Associated Inference and Application to Cutaneous Melanoma Data"		
10:20		**Morning Tea**	
11:00	Miller, Arden: 33 "MDS-optimal Supersaturated Designs"	Barton, James: 13 "Uncertainty in New Zealand's Greenhouse Gas Inventory"	Lin, Gwo Dong: 32 "Maximum Correlation for Baker's Bivariate Distributions with Fixed Marginals"
11:30	Kozek, Andrzej: 30 "Experimental Designs with Non-increasing Variances of the Ordinary Least Squares by Augmenting Heteroscedastic Data"	Bilton, Penelope: 15 "A Statistical Model to Characterize the Naphthenic Acid Component of Petroleum"	Wood, Graham: 52 "Normalization of Ratio Data"
12:00		**Lunch**	
1:00	Degnan, James: 21 "The Effect of Sequence Length on Estimated Distributions of Gene Trees"	Jones, Geoff: 28 "Inferring Infection History from Repeated Measures of Multiple Diagnostic Tests"	Ong, Seng Huat: 39 "Properties and Application of the Inverse Trinomial Distribution"
1:30	van Koten, Chikako: 47 "An Analysis of Power Approach to Time Series with a Known Periodic Input"	Betz-Stablein, Brigid: 14 "Disease Mapping Techniques Applied to Glaucoma Visual Field Datasets"	Ehlers, René: 22 "Triply Non-central Extended Bivariate Dirichlet Type I Distribution"
2:00	Zheng, Guan Yu (Fish): 54 "Empirical Study of Extreme Values on Seasonal Adjustments in Time Series Analysis"	Wang, Yuancheng: 49 "Assessing the Performance of Matrix Representation with Parsimony"	Alzaid, Abdulhamid: 8 "Binomial Difference Distribution"
2:30		Costilla, Roy: 19 "Estimating Cancer Survival in Subpopulations with a Small Number of Cases: A Parametric Approach"	Ardalan, Arash: 10 "A Generalized Normal-Laplace Distribution: Properties, Estimation and Applications"
3:00		**Afternoon Tea**	
3:30	Willink, Robin (1 of 2): 51 "Some Statistical Advances Originating in Measurement Science"	Young Statisticians' Session	Ng, Hon Keung Tony: 35 "Parametric Inference for System Lifetime Data with Signatures Available"
4:00	Turner, Rolf: 46 "Renyi's Theorem and Poisson Processes for the Uninitiated"	continued	Bodhisuwan, Winai: 16 "Bayesian Inference of Software Reliability Growth Model Based on Nonhomogeneous Poisson Processes"
4:30	Ali, Abdul: 7 "Ko te Anga te Hoto — Structure is the Link"	continued	Salehi, Ebrahim: 41 "On the Mean Residual Lifetime of Consecutive k-out-of-n Systems"

Wednesday 30 June

Time	Stream 1	Stream 2	Stream 3
9:00		**Welcome**	
9:10	Olkin, Ingram: 37 "Life Distributions in Survival Analysis and Reliability: Structure of Semiparametric Families"		
10:10		**Morning Tea**	
10:30	Kale, Hazel: 29 "Exploratory Data Analysis for Statistical Data Confidentiality"	Ganesalingam, Ganes: 25 "An Analytical Expression for the Misclassification Error Rates Associated with the QDF in Discriminating Two Normal Populations"	Wang, Ting: 48 "Markov-modulated Hawkes Process with Stepwise Decay"
11:00	Chen, Chen: 17 "Confidentiality for the 2011 Census: Statistical Thinking Applied"	Walker, Lyndon: 48 "Analysing Ethnic Partnership Matching Using a Grid-Based Evolutionary Algorithm"	Filus, Lidia: 23 "Weak Stochastic Dependence and Semi-pseudonormal Probability Distributions"
11:30	Haslett, Steve: 26 "Data Cloning for Confidentiality and Data Encryption"	Anwar, Nafees: 9 "Measurement and Visualization of Data Complexity for Classification Problem"	Kachapova, Farida: 28 "Population Monotony Coefficient"
12:00		**Lunch**	

1:00	Willink, Robin (2 of 2): 52 "A Confidence Interval for the Bounded Normal Mean from a Small Sample: an Adaptation of the Feldman-Cousins Interval"	Aubry, Jean-Marie: 11 "Large Deviations for Quasi-arithmetically Self-normalized Random Variables"	Zitikis, Ričardas: 54 "Weighted Distributions, Insurance Premiums, and Extreme Events"
1:30	McGirr, Rebecca & Hawkes, Tim: 32 "A New Output Geography for Offence Statistics"	Davis, Walter: 20 "Fisher's Rank & Order Conditions and Instrumental Variables: Connections and Implications"	Yee, Thomas: 53 "Parameter Estimation in Many Statistical Distributions"
2:00	Namay, Rico: 34 "Population-preserving Propensity Score Stratification for Survey Nonresponse Modelling"	Tularam, Anand & Roca, Eduardo: 45 "An Investigation of the Relationship Between Socially Responsible Investment Markets Based on the Dynamic Conditional Correlation Methodology"	Bebbington, Mark: 13 "Analyzing Treatment Effects on Distributions with Complex Structure"
2:30	Scott, Alastair: 42 "Pseudo Likelihood-Ratio Tests for Survey Data"	Torkashvand, Elaheh: 44 "Applications of Bayesian Point Null Hypothesis Testing Via the Posterior Likelihood Ratio"	Nagatsuka, Hideki: 33 "Consistent Method of Estimation for Distributions with Unknown Origin"
3:00		**Afternoon Tea**	

3:30	Haywood, John: 26 "Improved Multi-step Forecasting via a New Test of MLE Robustness"	Chee, Chew-Seng: 17 "Mixture-based Nonparametric Density Estimation: Maximum Likelihood vs. Least Squares"	Tang, Boxin: 44 "Robust Designs Through Partially Clear Two-Factor Interactions"
4:00	Rohan, Maheswaran: 40 "Using Finite Mixtures to Compute Robustified Statistics for Regression Parameters"	Fernando, Sarojinie: 22 "Spatio-temporal Modelling of Relative Risk"	Khoo, Michael: 30 "Univariate Synthetic Control Charts for Variables Data: A Review"
4:30	Rodado, Armando: 40 "Selecting Central American Volcanoes for an Empirical Bayes Analysis"	Hazelton, Martin: 27 "Shape Constrained Semiparametric Regression"	Chan, Ping Shing Ben: 16 "Optimal Allocation for Multi-level Stress Testing with Extreme-value Regression"
5:00		NZSA AGM	

Thursday 1 July

Time	Stream 1	Stream 2
9:00	**Welcome**	
9:05	Cheng, Ching-Shui: 18 "Multistratum Fractional Factorial Designs"	
10:05	**Morning Tea**	
10:30	Sampson, Allan: 42 "Multivariate Modelling Issues for Multiple Outcomes in Post-mortem Tissue Studies"	Cook, Len: 19 "Various Stages in the Evolution of Official Statistics in Britain, their Influences on New Zealand, and Differences in the Development of Official Statistics Here"
11:00	Ong, Hong Choon: 38 "Modelling the Aids Epidemic in Penang, Malaysia"	Noble, Alasdair: 36 "Using Statistical Models to Combine Existing Data Sources to Produce Sounder, More Detailed, and Less Expensive Official Statistics"
11:30	Lai, Chin-Diew: 31 "Distributions for Late Life Deceleration Phenomenon"	Richens, Andrew: 39 "Using Score Functions to Identify Key Firms in Statistics New Zealand Surveys"
12:00	Wrap-up	
12:15	**Lunch**	
1:15		Westbrooke, Ian: 50 "R with Menus — R Commander Software for Teaching Statistics to Non-statisticians"
1:45		TBA
3:15	**Afternoon Tea**	

Modelling the Aids Epidemic in Penang, Malaysia

Hong Choon Ong
School of Mathematical Sciences, Universiti Sains Malaysia

Lay Fong Sin
School of Mathematical Sciences, Universiti Sains Malaysia

Li Ling Tan
School of Mathematical Sciences, Universiti Sains Malaysia

This study briefly look at some of the statistical methods that have been developed to model the HIV/AIDS epidemic and also use the back calculation method to estimate the HIV infection rate in Penang. The back calculation program has been chosen to model the underlying HIV/AIDS epidemic in Penang, Malaysia because it makes use of the AIDS incidence data which is more reflective of the epidemic as compared to the number of HIV infected recorded which is known only if tests are conducted. The AIDS incidence data collected however have some limitations and uncertainties due to censoring of the data, reporting delay, under reporting and the changes in the AIDS surveillance definition. The back calculation method is used to reconstruct and estimate the number of people who have been infected previously in Penang using the AIDS incidence data obtained and an estimate of the incubation period distribution. The simulation shows the presence of under and delay reporting in the AIDS incidence data obtained especially in the early stages.

Modelling the Aids Epidemic in Penang, Malaysia

[1]Hong Choon Ong, [2]Lay Fong Sin, [3]Li Ling Tan

[1, 2&3]*School of Mathematical Sciences,*
Universiti Sains Malaysia,
11800 USM, Penang, Malaysia

Abstract

This study briefly look at some of the statistical methods that have been developed to model the HIV/AIDS epidemic and also use the back calculation method to estimate the HIV infection rate in Penang. The back calculation program has been chosen to model the underlying HIV/AIDS epidemic in Penang, Malaysia because it makes use of the AIDS incidence data which is more reflective of the epidemic as compared to the number of HIV infected recorded which is known only if tests are conducted. The AIDS incidence data collected however have some limitations and uncertainties due to censoring of the data, reporting delay, under reporting and the changes in the AIDS surveillance definition. The back calculation method is used to reconstruct and estimate the number of people who have been infected previously in Penang using the AIDS incidence data obtained and an estimate of the incubation period distribution. The simulation shows the presence of under and delay reporting in the AIDS incidence data obtained especially in the early stages.

Keywords: Back calculation method, AIDS modelling, HIV infection

1. Introduction

Methods commonly used method to model the HIV/AIDS epidemic in such countries as USA and UK are based on one of three methods. These methods are similar in that they all fit some form of calendar time to the incidence of AIDS but they differ in the degree to which the mechanisms that generate the data are incorporated into the model. The first method attempts to fit a function of the calendar time like polynomial or other mathematically convenient curves to the AIDS incidence curve and extrapolating into the future. This method is not efficient, as it makes use of less information (Healy & Tillett, 1988).

At the other extreme, the second method attempts to model the full dynamics of the transmission of the epidemic in the population, providing much insight into the qualitative evolution of the epidemic and identifying the key variables that determince the future number of cases but this method has unverifiable assumptions and contain many unkown parameters (Arca et al., 1992).

The third method which is intermediate between the first and second methods is the method we choose to model the HIV/AIDS epidemic in Penang – back calculation method (Ong & Soo, 2006). This method is applied to estimate the past HIV infection cases from the AIDS incidence data and an estimated incubation period. The back calculation method is used on the AIDS incidence data in Penang by using the back calculation program from Bacchetti, Segal & Jewell (Bacchetti et al., 1993) and the incubation period distribution from Brookmeyer and Gail (Brookmeyer & Grail, 1988) for estimating the HIV infection rates in Malaysia. The number of HIV+ cases, on the other hand, is dependent on the test made and is unreliable as a trend. For example, a steep rise in the number of HIV+ cases may be due to the mandatory testing of all intravenous drug users in drug rehabilitation centres and increase in detection through aggressive case finding. AIDS/HIV was nonexistent in

Malaysia until 1986. The first case of AIDS in Malaysia was reported in December, 1986 in an American of Malaysian origin. The American of Malaysian origin had returned home for a visit and died due to Pneumocystis carinii pneumonia. During the same year, there were another three persons who were found to be HIV positive.

The bulk of the AIDS incidence cases in Penang are injecting drug users and both heterosexually and homosexually transmitted. Mariotto et al. (1992) also reported the three modes of transmission are similar in their incubation period distribution and there is not much statistical difference between them. In this study, the incubation period distribution is taken from Brookmeyer (1991), which is based mainly on homosexuals in the USA.

HIV incubation period is the random time between the inception of the HIV infection and the onset of clinical AIDS or time when AIDS is diagnosed. Incubation period of HIV is very long and highly variable within and between cohorts. Distribution of this non-negative random variable is known as HIV incubation period distribution. Estimation of this distribution is hard because the time of infection in the risk groups is usually unknown, except in the blood transfusion associated cases where the time of infection is from the date of transfusion.

2. Methodology

2.1. The method of Back calculation

The back calculation method uses the AIDS incidence data with the estimate of incubation period distribution to reconstruct the number of people who have been infected previously. Then, the incubation distribution is applied to these estimated numbers to predict the AIDS incidence data.

The basic calculation of the back calculation method is based on the relationship between the cumulative numbers of new AIDS cases from the time of the onset of the epidemic, 0 to t (designated Z(t)) and the number of the new HIV infection g(s) at time s since the start of the epidemic (s=0). Then, let u be the time spent between the initial infection and the eventually diagnosis of AIDS (u = t-s) and F(u) be the distribution function of this incubation period.

The basic convolution equation is:

$$Z(t) = \int_{o}^{t} g(s)\ F(t-s)\, ds$$

2.2. An Application of the Back Calculation Method to Penang AIDS Incidence Data

The back calculation method is applied on the AIDS incidence data of Penang. This method is chosen because it makes use of the AIDS incidence data which is quite representative of the trend of the epidemic rather than the number of HIV cases recorded which is not reliable. We input the Penang incidence data (refer Table 1) into the back calculation program in FORTRAN by Bacchetti et al. (4). The incubation period distribution in this program is taken from Brookmeyer (7), which based on homosexuals in United States. Majority of AIDS cases in Penang are transmitted through IDU and the incubation period distribution is similar to that of both heterosexual and homosexual transmission (6).

Using the AIDS incidence data Z(t) and an estimate of the incubation period distribution F(u), we estimate the past HIV infection rates g(s) in Penang in the equation above using a deconvolution technique. This estimated past HIV infection rates can then be compared with the recorded past HIV infection rates to draw inferences about the AIDS epidemic and the nature of the HIV/AIDS data recorded.

Table 1: Total number of HIV/AIDS cases and related deaths reported in Penang from year 1988 to 2006

Year	HIV infection	AIDS	
		Cases	Death
1988	4	1	0
1989	6	1	1
1990	26	1	1
1991	37	1	0
1992	63	4	1
1993	64	4	8
1994	113	9	9
1995	174	14	13
1996	194	24	24
1997	215	27	27
1998	223	52	35
1999	225	105	55
2000	194	70	52
2001	215	69	34
2002	268	88	45
2003	266	49	20
2004	308	59	34
2005	251	70	50
2006	228	81	47
Total	**3074**	**729**	**456**

Source: AIDS/STI Unit, Ministry of Health Malaysia

3. Result

The result of the back calculation program applied is the estimated number of HIV infection from January 1991 to December 2006. The cumulative number of recorded and estimated HIV+ cases, and the recorded number of AIDS incidence cases is shown in Table 2.

Table 2: The recorded and estimated HIV/AIDS cases in Penang from year 1986 to 2006

Year	Cumulative HIV recorded	Cumulative HIV estimated	AIDS recorded
1988	4	-	1
1989	10		1
1990	36		1
1991	73	33.74	1
1992	136	85.36	4
1993	200	145.11	4
1994	313	373.98	9
1995	487	578.11	14
1996	681	855.87	24
1997	896	1322.40	27
1998	1119	1955.52	52
1999	1344	2761.67	105
2000	1538	3534.56	70
2001	1753	4074.35	69
2002	2021	4473.09	88
2003	2287	4810.00	49
2004	2595	5139.57	59
2005	2846	5795.06	70
2006	3074	7874.34	81

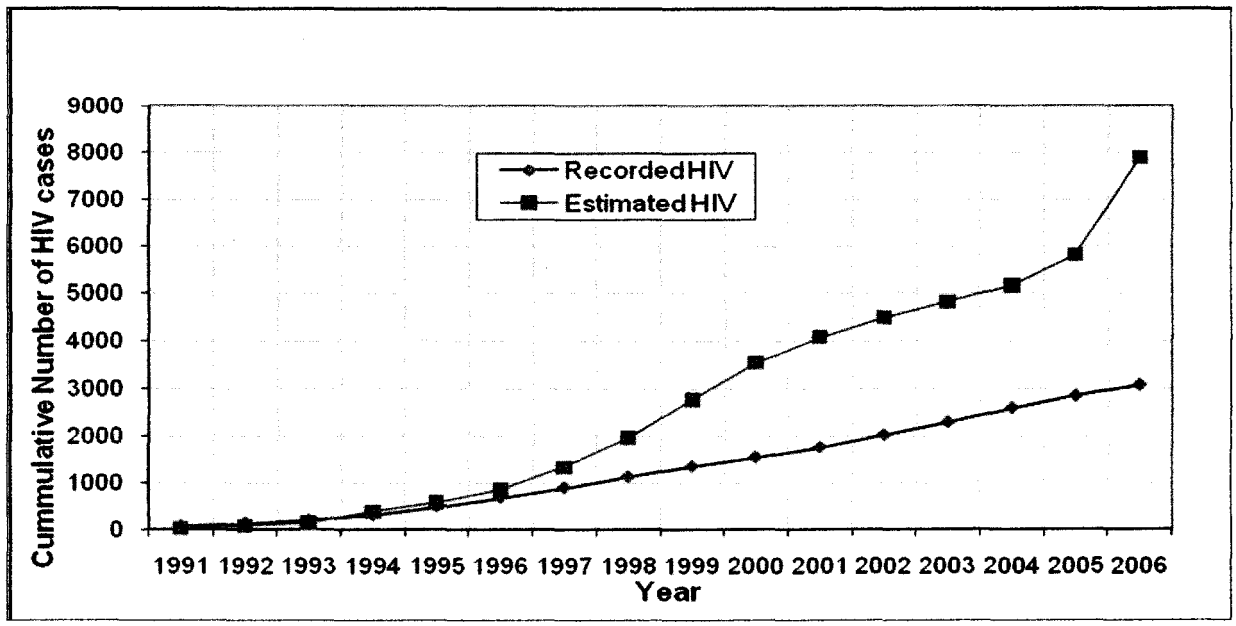

Figure 1: The comparison between the cumulative recorded HIV and the cumulative estimated HIV in Penang from year 1991 to 2006

Figure 1 shows the comparison between cumulative recorded HIV and the cumulative recorded estimated HIV cases in Penang from the year 1991 to 2006. Technically the estimated HIV should be higher than the recorded HIV because of the probable existence of untested cases. However, for the year 1991 to 1994, the estimated value is slightly lower than the recorded value. This may happen because of delay reporting from the previous years and changing of the definition of AIDS surveillance system. According to the year 1993 AIDS surveillance definition, anyone with CD4 lymphocyte count below 200/mm^3 is considered as a condition for AIDS.

There is a slowdown in the increase of cumulative estimated HIV value in the year 1999 to 2004 in Figure 1. This trend is supported by a slowdown in the increase of the number of AIDS cases. This is because the back calculation method depends on the inputted AIDS incidence data which is back calculated and according to the AIDS data, there is a sudden drop in the number of AIDS cases recorded in year 2003. There is a big difference for the year 2006 between the estimated and recorded value. This could show that there might have some cases of under reporting during that period.

5. Conclusion and Discussion

Every method used for modeling the AIDS epidemic has its own advantages and disadvantages. We choose back calculation because this method makes use of AIDS incidence data which is the most readily available indicator of the epidemic. The incubation period of homosexual (available from the United States) and intravenous drugs users are very similar. Based on this, we can calculate the past HIV infection rate.

The back calculation method provides a simple conceptual framework for modeling the progression of the AIDS epidemic. This method is able to reconstruct and estimate the underlying HIV infection rate which is not possible if we were to use the first method which is the curve fitting method. However, back calculation is heavily dependent on accurate AIDS incidence data.

The main problem in most developing countries is the incomplete reporting data and lack of the information on the AIDS epidemic. Delay reporting and under reporting is one of the major uncertainties.

Acknowledgement

We would like to thank the Ministry of Health Malaysia for the permission to use the Penang state AIDS incidence data in our study. The research is funded by the Universiti Sains Malaysia (USM) Research University Grant No. 1001/PMATHS/811154.

References

[1] Healy M. J. R. and Tillett H. E., 1988. "Short-term extrapolation of the AIDS epidemic", *Journal of the Royal Statistical Society* A, 151, pp. 50-61.

[2] Arca M., Perucci C. A. and Spadea T., 1992. "The epidemic dynamics of HIV-1 in Italy: Modelling the interaction between intravenous drug users and heterosexual population", *Statistics in Medicine*, 11, pp. 1657-1684.

[3] Ong H. C. and Soo K. L., 2006. "Backcalculation of HIV infection rates in Malaysia" , *The Medical Journal of Malaysia*, 61, pp. 613-617.

[4] Bacchetti P., Segal, M. R. and Jewell, N. P., 1993. "Backcalculation of the HIV infection rates." *Statistical Science*, 8(2), pp. 82-119.

[5] Brookmeyer R. and Gail M. H., 1988. "A method for obtaining short-term projection and lower bound on the size of the AIDS epidemic", *Journal of the American Statistical Association*, 83(402), pp. 301-308.

[6] Mariotto A. B., Mariotti S., Pezzotti P., Rezza G. and Verdecchia A., 1992. "Estimation of the acquired immunodeficiency syndrome incubation period in intravenous drug users: A comparison with male homosexuals", *American Journal of Epidemiology*, 135(4), pp. 428-437.

[7] Brookmeyer R., 1991. "Reconstruction and future trends of the AIDS in the United States", *Sciences*, 235, pp. 37-42.