STATISTICS IN MEDICINE
Statist. Med. 2000; **19**:3095–3108

Empirical Bayes approach to estimating the number of HIV-infected individuals in hidden and elusive populations

Ying-Hen Hsieh[1,*,†], Cathy W. S. Chen[2] and Shen-Ming Lee[2]

[1] *Department of Applied Mathematics, National Chung-Hsing University, Taichung, Taiwan*
[2] *Department of Statistics, Feng-Chia University, Taichung, Taiwan*

SUMMARY

In this paper we estimate the numbers of intravenous drug users (IVDUs) and commercial sex workers (CSWs) in Thailand infected with human immunodeficiency virus (HIV) who have not developed acquired immunodeficiency syndrome (AIDS) directly from the semi-annual HIV serosurveillance data of Thailand from June 1993 to June 1995. We propose a 'generalized removal model for open populations' for estimating HIV-infected population size within a hidden, elusive, and perhaps high-risk population group, for all sampling time when capture probabilities vary with time. We apply empirical Bayes methodology to the generalized removal model for open populations by using the Gibbs sampler, a Markov chain Monte Carlo method. No assumption on the size of the hidden population in question is needed to implement this procedure. The statistical method proposed here requires very little computing and only a minimum of two sets of serosurvey data to obtain an estimate, thereby providing a simple and viable option in epidemiological studies when either powerful computing facilities or abundant sampling data are lacking. Copyright © 2000 John Wiley & Sons, Ltd.

1. INTRODUCTION

The explosive spread of the acquired immunodeficiency syndrome (AIDS) epidemic in Thailand in the 1990s has been well documented (see, for example, Reference [1]). While some reports on declining HIV prevalence have given us reason to be optimistic toward future prospects (for example, Reference [2]), other surveys reporting on the lingering high human immunodeficiency virus (HIV) seroconversion rate among high-risk groups (for example, Reference [3]) cautioned that more problems may still be ahead. The 1994 National Economic and Social Development Board of Thailand (NESDB) Working Group on HIV/AIDS Projection [4] reported that since 1991 the total number of new HIV infections has declined each year in Thailand. However, a behaviour survey [5] of young army conscripts from 1991 to 1993 has reported that, although

*Correspondence to: Ying-Hen Hsieh, Department of Applied Mathematics, National Chung-Hsing University, Taichung, Taiwan
†E-mail: hsieh@amath.nchu.edu.tw

Contract/grant sponsor: National Science Council of Taiwan
Contract/grant sponsor: Fogarty International Center/NIH; contract/grant number: 1 R03 TW00536-01

42 per cent of conscripts had not visited a commercial sex worker (CSW) in the year prior to conscription, most had at least one visit during their military service. Moreover, no difference by HIV-serostatus was evident in their patterns of visits to CSWs. Although recent studies indicate a definite declining trend of HIV infection in the general population, the 1997 HIV sentinel data still reports that the HIV prevalence among intravenous drug users (IVDUs) and CSWs remains high (see Reference [6]). Hence the extent to which the effect of the '100 per cent condom programmes' (see Reference [7]) and subsequent intervention programmes has had on the overall HIV prevalence in Thailand, especially among the high-risk and elusive groups (IVDUs, CSWs etc.), is yet unclear.

It is well known that the HIV infection in Thailand first emerged among the IVDUs, similar to many other countries in the world, but the speed with which the epidemic spread in the early 1990s has been attributed mainly to the large CSW population and their young male customers (see, for example, Reference [8]). A large amount of work has been done in recent years to study the sexual networking in Thai society (for example, References [9, 12]). However, much still remains unknown, including the actual size of the various high-risk groups and consequently the size of the infected population in each group. The lack of knowledge in this regard not only hinders theoretical study of the spread of epidemic, but also leads to uncertainty in the design of intervention policies and the implementation of health care. Given the added importance of budgetary concerns caused by the recent financial crisis in Asia, it is worthwhile obtaining theoretical estimates of the number of infected individuals in the high-risk populations in order for the policymakers of prevention programmes to have a fuller understanding of the spread of the epidemic and to make better use of a shrunken budget.

With the rapid growth of a world-wide AIDS epidemic in recent years, estimating the number of HIV-infected individuals in a certain population, for example, homosexuals, prostitutes, IVDUs etc., has become a major problem of public health concern in many countries. In a 1989 review article [13] on methods to estimate population size of high-risk groups for HIV infection, special attention was given to the potential use of the capture-recapture method (or multiple-record system method in dealing with human populations, see Reference [14]) for estimating populations of IVDUs and prostitutes. Subsequent work on estimating the number of drug users includes References [15–18]. Similar estimates for prostitutes in Glasgow using a multiple-capture method was also carried out in Reference [19]. For a discussion on problems in the estimation of hidden and elusive populations using the capture-recapture method see also Reference [20]. A lucid review of the historical development of the capture-recapture method and its applications to human diseases can be found in References [14, 21].

In all of the above-cited work, the emphasis has been placed on estimating the population size of drug users or prostitutes. However, a more direct question of epidemiological importance is the actual number of seropositives in a particular population. To that aim, Mastro *et al.* [18] combined their estimated number of IVDUs in Bangkok with the results from other HIV prevalence studies to yield an estimate of the HIV-infected IVDUs in Bangkok. Abeni *et al.* [22] also used data from four large testing sites in Lazio, Italy, to generate incomplete, partially overlapping lists of HIV-infected subjects with which they then estimated the population size of HIV-infected individuals in Lazio in 1990.

In biological studies, it is often necessary to estimate the size of a population. Seber [23] classified populations into two categories, called 'closed' and 'open', depending on whether the population remains unchanged during the period of investigation, or changes through such processes as birth, mortality, emigration etc. In this work we wish to implement a procedure by which one can

estimate the number of HIV-infected individuals in a high-risk and hard-to-count population from two or more samples or serosurveys of the same population at different sampling times. In the 1950s, the removal model was proposed by Moran [24] and Zippin [25, 26] to estimate closed population size when each sampling results in the removal of captured animals. This model for a closed population has been studied subsequently by Otis *et al.* [27], Chaiyapong and Lloyd [28] and Yip and Fong [29]. In order to make a more precise inference, we propose a 'generalized removal model for open populations' which allows only recruitment (of new HIV-infected individuals) and deaths (removal of HIV-infected individuals due to development of AIDS) to occur during the experiment. We use the proposed method to estimate the number of HIV-infected IVDUs and CSWs in Thailand during the period of June 1993 to June 1995.

The rest of the paper is organized in the following manner. We describe the data used for our estimates in this paper in Section 2. Section 3 gives the proposed empirical Bayes procedure. In Section 4 we give the results obtained by applying our procedure to the data described in Section 2. Finally, in Section 5, we discuss the advantages of our method as well as certain limitations in applications.

2. HIV SENTINEL DATA OF THAILAND

The HIV serosurveillance data published by the Division of Epidemiology, Ministry of Public Health (MOPH) of Thailand [30] consists of serosurvey data from all 76 provinces of Thailand for IVDUs, CSWs (direct and indirect), male STDs, blood donors, and pregnant women in ANC centres. The 'direct' CSWs work in brothels, while the 'indirect' CSWs work in commercial establishments such as bars and massage parlours where sex can be available on request. For each half year from June 1989 to June 1995 and every year after June 1995, health workers in each province performed an HIV serosurvey for 100–200 individuals (if available) from each of the above-mentioned groups. Different sampling methods were employed for different groups. For example, cluster random sampling of various commercial sex establishments was used for testing CSWs on a voluntary basis while sampling for IVDUs took place during their visits to local drug users treatment centres run by the government. In all cases, the testing was mandatory with efforts to follow up the seropositive cases. Given that our aim is to estimate the size of HIV-infected individuals in a high-risk and elusive population, it would be of little practical use to estimate how many HIV-infected male STD patients there are in Thailand. Moreover, blood donors and pregnant women are by no means elusive and hard to count. Hence we only consider the three groups of IVDUs and CSWs (direct and indirect).

We wish to estimate the number of HIV-infected IVDUs and CSWs who have not progressed to AIDS for the time period June 1993 to June 1995 by directly using data of the Thai HIV Serosurveillance Round 9–13 taken semi-annually during that time period. Table I lists the resulting nation-wide seroprevalence data for these three groups for the five samples from June 1993 to June 1995. The province-by-province data is also available from the MOPH reports. However, the high mobility of these groups, especially the CSWs [11], renders the provincial data highly volatile from survey to survey and difficult to use in our estimates. We therefore confine ourselves to the estimates for national-wide totals. Also note that the separate numbers for the direct and indirect CSWs in June 1995 are not available due to a decision by MOPH after December 1994 to combine future surveys for direct and indirect CSWs on the ground that the trend of epidemics in these groups is well-established [31]. For every serosurveillance round, the prevalence rates for the direct

Table I. Thai sentinel data (Round 9–13) for intravenous drug users and commercial sex workers.

Date	IVDU			Direct CSW			Indirect CSW		
	HIV+	Total	%	HIV+	Total	%	HIV+	Total	%
06/93	1234	3515	35.11	2731	8979	30.42	608	7041	8.64
12/93	1276	3388	37.66	2412	8170	29.52	721	7793	9.25
06/94	1033	3234	31.94	2441	8653	28.21	703	8024	8.76
12/94	346	985	35.13	1313	4014	32.71	411	4186	9.82
06/95	1235	3585	34.45	—	—	—	—	—	—

— denotes not available.

CSWs obtained from the sentinel data are several times higher than the corresponding prevalence rates of the indirect CSWs – a reasonable result since the direct CSWs working in brothels would tend to have many more customers and be less selective when compared with their counterparts (indirect CSWs) working in bars and massage parlours. Subsequently we decide to use the sentinel data from June 1993 to December 1994, instead of the more recent data, to estimate the numbers of the direct and indirect CSWs separately in order to have more accuracy in our estimates of the CSWs. We also give estimates of the IVDUs for the same period plus the June 1995 Round, the last time the serosurvey was taken semi-annually.

3. STATISTICAL METHOD

First note that in describing the statistical procedure throughout this section, the term 'population' denotes the HIV-infected individuals among the CSWs and IVDUs whose size we wish to estimate. In our framework where the population size to be estimated is the number of HIV-infected individuals within a certain hard-to-count population, there is no recapture since it is reasonable to assume those tested positive will not be tested again. Hence the removal model is the appropriate choice of model to work with. In each sample, numbers of subjects are selected for testing. For example, in the 9th Round Thai sentinel data (June 1993 in Table I), 8979 subjects are selected from the direct CSW population for testing, and 2731 tested to be HIV-infected. Using the four sets of semi-annual data from June 1993 to December 1994 we estimate total population sizes of HIV-infected direct CSWs from June 1993 to December 1994. The generalized removal model for open populations proposed here can also be considered as one which gives estimates of HIV-infected population sizes for all sampling time when capture probabilities (that is, the probability of testing HIV-positive) vary with time. Moreover, since the sample taking would exclude anyone who has already developed AIDS symptoms, the estimate we obtain is the number of HIV-infected individuals who have not developed AIDS. It presents no hindrance in the assessment of the AIDS scenario, since the size of population with AIDS symptoms can be easily counted from clinical records.

3.1. Generalized removal model for open populations

We consider a sequence of s samples taken from the serosurvey data. Let t_j be the time when the jth sample is taken and B_j be the number of new HIV-infected individuals between time t_j and time t_{j+1}. Assume that all subjects in the HIV-infected population just before time t_j who have not been caught in the first $j-1$ samples have the same capture probabilities P_j in the jth sample.

We define N_j be the total number of subjects in HIV-infected population just before time t_j, and $N_j = B_0 + \cdots + B_{j-1}$. The likelihood function can be obtained as follows:

$$L(\boldsymbol{B}, \boldsymbol{P}|\mathscr{D}) \propto \left\{ \prod_{j=1}^{s} \binom{N_j - M_j}{u_j} P_j^{u_j} (1 - P_j)^{N_j - M_{j+1}} \right\} \tag{1}$$

where $\mathscr{D} = \{u_1, \ldots, u_s\}$, $\boldsymbol{B} = (B_0, \ldots, B_{s-1})$ and $\boldsymbol{P} = (P_1, \ldots, P_s)$; u_j is the number of distinct HIV-infected individuals captured in the jth sample. Therefore, $M_{j+1} = u_1 + \cdots + u_j$ is the number of distinct HIV-infected individuals captured in the first j samples. We call this model a generalized removal model for open populations, due to the removal of the observed HIV-infected individuals. We extend the removal model of Otis *et al.* [27] for a closed population to allow recruitment to occur between samples. Note that it is reasonable to assume that individuals tested to be HIV-infected in the jth sample will not be caught after jth sample. This implies that once identified in the survey, individuals will not be captured again.

The proposed model involves more parameters than the minimal sufficient statistic. Consequently, all parameters cannot be estimated without additional restrictions, and maximum likelihood estimation of the population size proves to be impossible. In order to make the population size N (for a closed population) an identifiable parameter under maximum likelihood estimation, Otis *et al.* [27] suggests letting $P_1 = \cdots = P_s = P$ or $P_{s-2} = P_{s-1} = P_s$. As it is not possible to obtain valid estimation of the HIV-infected population by using maximum likelihood estimation for an open population, we propose a Bayesian estimation procedure. Bayesian inference of a population size for various models has been proposed in the literature (see, for example, References [32–34]). In the Bayesian setting we would wish to give prior distributions to the unknown parameters of the model, $\boldsymbol{N}$ and $\boldsymbol{P}$. We assume the prior of $\boldsymbol{N}$ is constant (vague prior) which is also used by Castledine [35]. It is appropriate in cases where we only have vague prior knowledge about N_j. Moreover, we assume that the priors of P_j's are *a priori* independent and follows a beta distribution $\mathrm{Be}(\gamma_1, \gamma_2)$. The posterior distribution of $\boldsymbol{N}$ given $\boldsymbol{P}$ is a truncated negative binomial. The complete conditional posterior distributions are given in Appendix A. Since there are AIDS-related deaths during the process, we define the semi-annual survival rate specific to an HIV-infected individual between the $(j-1)$th and jth sample to be ϕ. The conditional expectation of M_{j+1} and N_{j+1} for the $(j+1)$th sample given M_j (the number of distinct HIV-infected individuals captured in the first $j-1$ samples) and N_j (the total number of subjects in HIV-infected population just before time t_j), respectively, are

$$E(M_{j+1}|M_j) = \phi M_j + u_j \quad \text{and} \quad E(N_{j+1}|N_j) = \phi N_j + B_j \tag{2}$$

The detailed derivation of (2) is given in Appendix B. This assumption is appropriate since the majority of the IVDUs and CSWs in question are in their prime years when the natural mortality is rather low, and also because the samples in this work span a relatively short period of time (two and a half years). Hence we assume that the natural death rate of IVDUs and CSWs during this time period is negligible.

3.2. Markov chain Monte Carlo approach

We utilize an empirical Bayes analysis in the proposed model by using the Gibbs sampler, a Markov chain Monte Carlo (MCMC) method. The Gibbs sampler is a Markovian updating scheme enabling one to obtain samples from a joint distribution via iterated sampling from full conditional

distributions. Detailed discussions can be found elsewhere [36, 37]. Interested readers are also referred to References [38] and [39] for a comprehensive review of the Gibbs sampler.

In order to utilize the empirical Bayes analysis and to implement the Gibbs sampler, the hyper-parameters of P_i, namely (γ_1, γ_2), are needed. We describe the procedure for choosing γ_1 and γ_2 in Appendix C. The Bayes estimates are based on Monte Carlo samples from the Gibbs sampler run of 10 000 iterations after 2000 burn-in, and selecting every 20th sampled value. The MCMC method is assured to converge using a procedure developed in Reference [40]. The details are omitted to save space.

4. RESULTS

We are interested in making an inference about the population size of the HIV-infected population for each sample. We choose the 6-month survival rate ϕ for HIV-infected individuals (that is, the mean probability that an infected individual will not develop AIDS during the six months between samples) to be 95 per cent and 90 per cent. The time from HIV infection to symptomatic AIDS (when patients usually die within a year) is approximately 8 to 10 years for gay men in the West [41]. However, studies have indicated a much shorter time for HIV-infected individuals in the developing countries, with reports ranging from mean incubation time of 3.5 years for a study on infected female prostitutes in Uganda [42] to mean survival time from diagnosis to death of 7 months for patients in a hospital in suburb of Bangkok, Thailand [43]. However, many factors influence the results from these studies, the most prominent being that the diagnosis of infection usually occurs at advanced stages of disease in many developing countries. For our study, a 6-month survival rate of 90 per cent would result in 3.5-year survival rate of 47.83 per cent (since $0.9^7 = 0.4783$), while 95 per cent survival rate for 6 months implies that survival rate after 6.5 years is 51.13 per cent ($0.95^{13} = 0.5113$), resulting in median survival time of approximately 3.5 and 6.5 years, respectively. Our results will show only minor differences in the estimates using the different survival rates.

The results are given in Table II. For each case, Table II lists median, mean, standard error and a 95 per cent credible interval for N_j obtained from 2.5 per cent and 97.5 per cent quantiles. Note that in the June 1995 sentinel data the direct and indirect CSWs are combined in reporting due to a recommendation by the Division of Epidemiology of Ministry of Public Health. In each case the 95 per cent credible interval becomes smaller for each succeeding estimate. This is due to the underlying feature of our method that when the difference between the succeeding estimates tends to get smaller, the corresponding standard error would also become smaller as a result.

All estimates indicate an increase from the previous sample (of six months before). However, the size of increase decreases for each of the succeeding samples for all three groups studied. This result seems to confirm an earlier report [4] on the decline of the number of new infections in recent years. Table III gives the percentage increases from the previous half-year, that is

$$\text{percentage increase} = \frac{N_{j+1} - N_j}{N_j} \times 100\%$$

for all three groups. For all sampling periods studied, the percentage increases are less than the previous one. That is, the percentage increase of the number of HIV-infected individuals from the previous half-year period decreases for each of the half-year periods studied. Figures 1(a)–(c) give plots of the percentage increases for these three groups (IVDUs, direct CSWs, and indirect CSWs) at $\phi = 90$ per cent and $\phi = 95$ per cent.

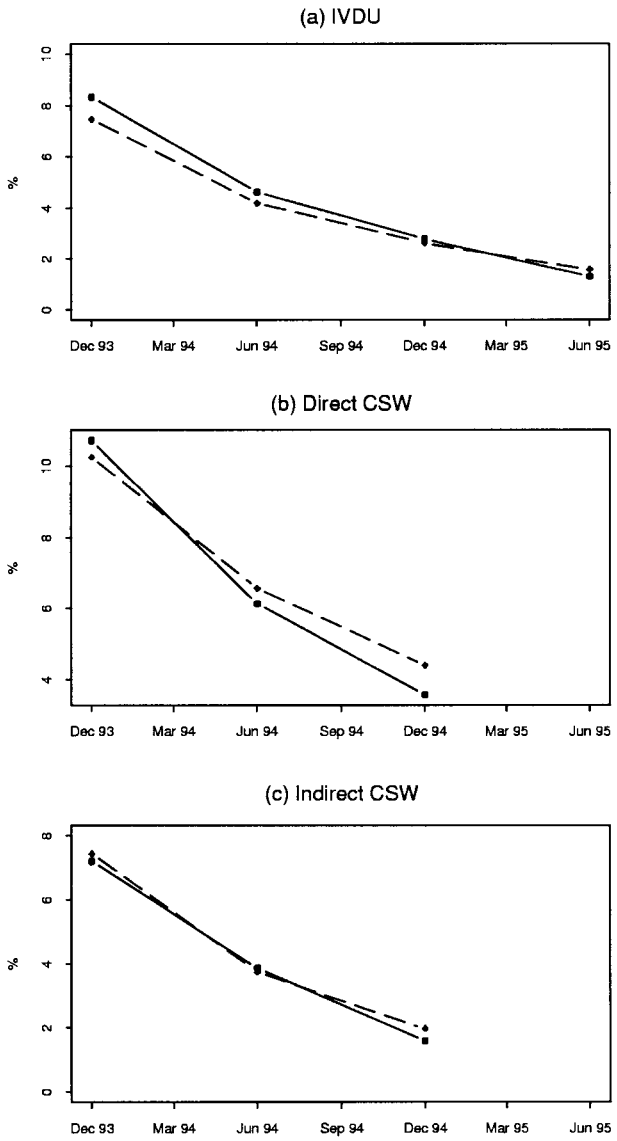

Figure 1. The percentage increases of the estimated number of HIV-infected IVDUs and CSWs from the previous half-year. The solid line is for $\phi = 90$ per cent and the broken line is for $\phi = 95$ per cent. (*a*) intravenous drug users (IVDUs); (*b*) direct commercial sex workers (CSWs); (*c*) indirect commercial sex workers (CSWs).

Table II. Results of estimates for HIV-infected IVDUs and CSWs.

	$\phi = 90\%$					$\phi = 95\%$				
	Median	Mean	SE	95% CI		Median	Mean	SE	95% CI	
IVDU										
06/93	29 067	28 960	1008	26 784	30 494	29 488	29 326	1125	26 754	31 148
12/93	31 488	31 409	490	30 332	32 290	31 690	31 639	595	30 404	32 630
06/94	32 945	32 887	293	32 197	33 333	33 018	32 976	438	32 019	33 691
12/94	33 864	33 819	169	33 362	34 004	33 884	33 843	314	33 160	34 322
06/95	34 301	34 281	86	34 059	34 375	34 415	34 387	158	34 051	34 625
Direct CSW										
06/93	54 595	54 461	2922	48 379	59 399	53 647	53 433	2943	46 979	58 559
12/93	60 452	60 177	2034	55 640	63 480	59 152	59 039	1894	54 844	62 208
06/94	64 157	63 994	1356	61 133	66 161	63 039	62 901	1401	59 529	65 069
12/94	66 445	66 363	776	64553	67485	65811	65 721	886	63 777	67 237
Indirect CSW										
06/93	15 181	15 067	519	13 855	15 850	15 062	15 044	482	14 014	15 874
12/93	16 275	16 249	213	15 807	16 602	16 183	16 167	250	15 674	16 600
06/94	16 903	16 871	112	16 609	17 018	16 788	16 778	160	16 434	17 062
12/94	17 171	17 155	60	17 020	17 299	17 119	17 111	90	16 910	17 264

Table III. Percentage increase of numbers of HIV-infected IVDUs and CSWs from previous half-year June 1993 to June 1995.

	IVDU ϕ		Direct CSW ϕ		Indirect CSW ϕ	
Date	90%	95%	90%	95%	90%	95%
06/93	NA	NA	NA	NA	NA	NA
12/93	8.33%	7.47%	10.73%	10.26%	7.21%	7.44%
06/94	4.63%	4.19%	6.13%	6.57%	3.86%	3.74%
12/94	2.79%	2.62%	3.57%	4.40%	1.59%	1.97%
06/95	1.29%	1.57%	—	—	—	—

NA denotes not applicable.
— denotes not available.

5. CONCLUDING REMARKS

A generalized removal model for open populations is proposed to estimate the number of HIV-infected individuals in a hidden and elusive population directly from two or more sets of serosurvey data. The estimate does not include those who have already developed AIDS. However, this presents no obstacle in public health policymaking since the latter data can be easily obtained from hospital records.

The proposed model for open populations involves more parameters than the minimal sufficient statistic and therefore all parameters are not estimable by using maximum likelihood estimation without additional restrictions. Our Bayesian approach enables us to estimate more parameters than observations at hand. Therefore, the non-identifiability can be resolved in the proposed approach.

The model assumes that the number of HIV-infected individuals removed due to AIDS (by onset of AIDS or AIDS-related death) between each sample is less than newly infected individuals during

the same time interval. This is a reasonable assumption for HIV/AIDS due to the long incubation period of HIV, but might not be applicable in diseases with short incubation time. Moreover, at the time of the serosurveys (1993–1995), the HIV epidemic in Thailand was at its early stages. All studies have shown the numbers of HIV-infected population at that time to be increasing (see, for example, Reference [4]). Clearly this method might not be appropriate in the case of an epidemic which has reached saturation.

We also implicitly assume the number of individuals detected to be HIV-positive but not included in this HIV sentinal data to be negligible since this is the comprehensive national HIV serosurveillance programme carried out by the Thai government.

The survival rate ϕ is assumed to be constant, although in reality it varies with each individual's detected time since infection. If we assume ϕ to be dependent on the time since infection, then ϕ_i is the survival rate for the ith individual. However, we would then need detailed information regarding each individual's time of infection, which is not available. Moreover, the stochastic nature of each individual's progression to AIDS and death also requires a much more complicated model which is beyond our scope. There are, of course, ways by which one could possibly improve upon the assumption of constant survival rate. For example, our model can be easily modified to allow the survival rate to change from sampling period to sampling period (that is, replace ϕ by ϕ_j in equation (2)), thus taking account of the time-varying nature of the survival rate. However, an estimate of the average survival rate of all infected individuals at each sampling period is also difficult to obtain, if not impossible.

The model assumes no natural (unrelated to AIDS) deaths between the samplings. Although the populations under study here, namely the IVDUs and CSWs, are in general adults with generally low natural mortality, they are also at risk for other diseases (for example, sexually transmitted diseases) which tend to increase mortality. In applications of this model one should always keep the time interval in which the samples are taken reasonably short. This is one reason that in this work we did not make use of the complete serosurvey data in Thailand which started in 1989. In applications where the intended population might have higher mortality, even shorter time intervals would be advisable to avoid large errors in the estimates.

Back-estimating the number of elusive population from our estimate for the HIV-infected individuals in that population is also possible, when a more precise estimate for the population size is unavailable. However, one needs to exercise caution in this endeavour as our estimate is only a rough approximation at best. To illustrate how this can be done, we know of no reported census or estimate of numbers of HIV-infected IVDUs or CSWs in Thailand. However, Mastro *et al.* [18] estimated the number of HIV-infected IVDUs in Bangkok to be approximately 12 000 by first using their 1991 data on IVDUs in Bangkok to obtain an estimated number of IVDUs in Bangkok of 32 574. For the purpose of comparison, we use our national median for HIV-infected IVDUs in June 1993 with $\phi = 90$ per cent 29 067, which is closest in time to the 1991 estimate of [18]. Dividing 29 067 by the seroprevalence of 35.11 per cent, we arrive at an estimate of 82 789 IVDUs in Thailand in June 1993. For further comparison with the result of Reference [18], in fiscal year 1989 there were 60 323 admissions for treatment at 138 registered heroin/opiate detoxification centres in Thailand, out of which 27 056 admissions are in Bangkok (see Reference [1]). Assuming that the number of IVDUs in Bangkok maintains roughly the same proportion when compared with the nation-wide total in June 1993, we obtain an estimate of 37 133 IVDUs in Bangkok. Moreover, we obtain an estimate of 13 038 HIV-infected IVDUs in Bangkok for June 1993. Note, however, that by combining Bangkok data with the rest of the country in our estimates tend to cause an underestimate of the true number.

One should also note that, as in all problems of estimation, the manner in which the sampling was conducted also could have a great effect on the accuracy of estimates. In the case of the Thai serosurvey data, cluster random sampling was used for CSWs while samples for IVDUs were taken from IVDUs seeking treatment at local clinics.

Finally, due to the high degree of variation among the 76 provinces, point estimate of the national total has only limited accuracy. However, to estimate the provincial totals separately would lead to other (perhaps more severe) problems, one being the high mobility of the CSWs (see Reference [11]), resulting in inaccuracy of the data from one sample to another. Hence we are limited in our choice of estimation by these very practical considerations.

Information regarding hidden and elusive populations are difficult to obtain (see Reference [20]). The dilemma has proved to be even more challenging in the context of the HIV epidemic. In this work we have developed a statistical method by which one can extract information regarding the size of the HIV-infected population within a certain high-risk and hard-to-count group. Our results give an estimate of the HIV-infected individuals in the IVDU and CSW groups at the end of each sample. This allows the policy makers to set public health policies with a clear understanding of the current direction of the epidemic. It is also worthwhile pointing out that our method is simple to run on a personal computer and requires only a minimum of two sets of serosurvey data in order to obtain an estimate. It provides an easily implemented and useful alternative to estimate the magnitude of the HIV/AIDS epidemic, especially when either detailed serocensus data or sophisticated computer hardware is not readily available.

APPENDIX A: CONDITIONAL POSTERIOR DISTRIBUTIONS

For the likelihood function (1) and prior distributions described in Section 3.1, the conditional posterior distributions are given by

$$\pi(\boldsymbol{P} \mid \boldsymbol{N}, \mathscr{D}) = \prod_{j=1}^{s} \mathrm{Be}(u_j + \gamma_1, N_j - M_{j+1} + \gamma_2) \tag{A1}$$

$$\begin{aligned}\pi(N_j \mid \boldsymbol{N}_{(-j)}, \boldsymbol{P}, \mathscr{D}) &= \frac{\binom{N_j - M_j}{u_j} P_j^{u_j} (1-P_j)^{N_j - M_{j+1}}}{\sum_{N_j = \max\{N_{j-1}, M_{j+1}\}}^{N_j+1} \binom{N_j - M_j}{u_j} P_j^{u_j} (1-P_j)^{N_j - M_{j+1}}} \\ &= \frac{\binom{(N_j - M_{j+1}) + (u_j+1) - 1}{u_j} P_j^{u_{j+1}} (1-P_j)^{N_j - M_{j+1}}}{\sum_{N_j = \max\{N_{j-1}, M_{j+1}\}}^{N_j+1} \binom{(N_j - M_{j+1}) + (u_j+1) - 1}{u_j} P_j^{u_{j+1}} (1-P_j)^{N_j - M_{j+1}}}\end{aligned} \tag{A2}$$

where $\boldsymbol{N}_{(-j)}$ denotes the vector $\boldsymbol{N}$ with the N_j deleted. $(N_j - M_{j+1})$ follows a truncated negative binomial with parameters u_{j+1} and P_j and $N_{j-1} \leqslant N_j \leqslant N_{j+1}$. Subsequently one can easily implement the Gibbs sampler to generate $(N_j - M_{j+1})$ from the truncated negative binomial in equation (A2), and therefore the estimates of N_j can be obtained. We could also use Jeffreys' prior (see References [35, 34]), $\pi(\boldsymbol{N}) = \prod_{j=1}^{s} (1/N_j)$, in which case the conditional posterior of N_j becomes

$$\pi(N_j \mid \boldsymbol{N}_{(-j)}, \boldsymbol{P}, \mathscr{D}) = \frac{\frac{1}{N_j}\binom{N_j - M_j}{u_j} P_j^{u_j} (1-P_j)^{N_j - M_{j+1}}}{\sum_{x = \max\{N_{j-1}, M_{j+1}\}}^{N_{j+1}} \frac{1}{x}\binom{x - M_j}{u_j} P_j^{u_j} (1-P_j)^{N_j - M_{j+1}}}$$

However, this prior leads to a much more complicated posterior form than that of equation (A2) (and those used in References [35, 34]), therefore it is not readily implementable in our scheme. It remains an open question to consider other types of priors and to investigate the sensitivity of the posterior distribution of N_j.

Estimation of the marginal posterior densities for the $(\boldsymbol{N}, \boldsymbol{P})$ is then achieved by repeated sampling from (A1) and (A2) alternately, conditional upon current estimates of other unknown parameters, until convergence is achieved.

APPENDIX B: DERIVATION OF EQUATION (2)

It is assumed that no natural (unrelated to AIDS) deaths occurred in the process in a births-only model considered in Reference [33]. That is

$$M_{j+1} = M_j + u_j \tag{A3}$$

$$N_{j+1} = N_j + B_j \tag{A4}$$

where B_j is the number of new HIV infections between the $(j-1)$th and jth sample. This assumption is plausible when the time span between the samples is short, mainly due to the fact that the expected survival time of an uninfected individual is significantly longer than that of an HIV-infected individual. However, since there are AIDS-related deaths during the process, equations (A3) and (A4) are no longer valid. Hence we define the semi-annual survival rate specific to an HIV-infected individual between the $(j-1)$th and jth sample to be ϕ.

Suppose that $M_j^{(s)}$ and $N_j^{(s)}$ denote the respective numbers of survivals of M_j and N_j between the $(j-1)$th and jth sample. It follows that $M_j^{(s)} \mid M_j \sim \text{Bin}(M_j, \phi)$ and $N_j^{(s)} \mid N_j \sim \text{Bin}(N_j, \phi)$, where Bin denotes a binomial distribution. Moreover, $N_{j+1} = N_j^{(s)} + B_j$ and $M_{j+1} = M_j^{(s)} + u_j$. We assume that $N_j \leqslant N_{j+1}$, that is, the number of AIDS-related death is less than the number of new HIV infections. In particular, $N_j^{(s)}$ and $M_j^{(s)}$ are random variables and are unobservable. Given the values of M_j and N_j, we can estimate $M_j^{(s)}$ and $N_j^{(s)}$ by their conditional expectations. That is, ϕM_j and ϕN_j are estimates of $M_j^{(s)}$ and $N_j^{(s)}$, respectively. It follows that the conditional expectation of M_{j+1} and N_{j+1} for the $(j+1)$th sample given M_j and N_j, respectively, are

$$E(M_{j+1} \mid M_j) = \phi M_j + u_j \quad \text{and} \quad E(N_{j+1} \mid N_j) = \phi N_j + B_j$$

APPENDIX C: THE HYPERPARAMETERS (γ_2, γ_2) of P_i

In order to choose the value for γ_1 and γ_2, we assume that $P_j = Pe_j$, where P is a constant and e_j is the sample size of the jth sampling. That is, the capture probability in jth sample is proportional to the jth sample size.

We adopt the idea of Reference [44] for estimating the population size in a closed population in a capture-recapture model. If P_j follows Beta(γ_1, γ_2), then the expectation and coefficient of variation of P_j are $\gamma_1/(\gamma_1 + \gamma_2)$ and $\sqrt{\{\gamma_2/(\gamma_1(\gamma_1 + \gamma_2 + 1))\}}$, respectively. Moreover

$$E(u_1 \mid P_1, \ldots, P_s, N_1) = N_1 P_1$$

$$E(u_2 \mid P_1, \ldots, P_s, N_1) = (N_1(1-P_1)\phi + B_1)P_2$$
$$= N_1\left((1-P_1)\phi + \frac{B_1}{N_1}\right)P_2 \tag{A5}$$

Under the assumption of $P_j = Pe_j$, we have

$$\frac{E(u_2 \mid P_1, \ldots, P_s, N_1)e_1}{E(u_1 \mid P_1, \ldots, P_s, N_1)e_2} = \left((1-P_1)\phi + \frac{B_1}{N_1}\right)$$

When $\phi = 1$ and $B_1 = 0$

$$N_1 = \frac{E(u_1 \mid P_1, \ldots, P_s, N_1)}{1 - \frac{E(u_2 \mid P_1, \ldots, P_s, N_1)e_1}{E(u_1 \mid P_1, \ldots, P_s, N_1)e_2}}$$

Therefore

$$E(P_1) = \frac{E(u_1)}{N_1}$$
$$\sim 1 - \frac{E(u_2 \mid N_1)e_2}{E(u_1 \mid N_1)e_1}$$

Since $P_j = Pe_j$, P_j and e_j have the same coefficient variation. We can therefore solve for γ_1 and γ_2.

If $\phi \neq 1$ and $B_1 \neq 0$, then

$$\frac{E(u_2 \mid P_1, \ldots, P_s, N_1)e_1}{E(u_1 \mid P_1, \ldots, P_s, N_1)e_2} = \left((1-P_1)\phi + \frac{B_1}{N_1}\right)$$
$$= \left((1-P_1) + \left\{(1-P_1)(\phi - 1) + \frac{B_1}{N_1}\right\}\right)$$

Let $\theta^* = [(1-P_1)(N_1(\phi-1)+B_1) + P_1B_1]/N_1$. Under the assumption of $N_j \leqslant N_{j+1}$, we have $(1-P_1)(N_j(\phi-1)+B_j) \geqslant 0$ and it follows that $\theta^* \geqslant 0$.

Moreover

$$1 - \frac{E(u_2 \mid P_1, \ldots, P_s, N_1)e_1}{E(u_1 \mid P_1, \ldots, P_s, N_1)e_2} = 1 - \left((1-P_1)\phi + \frac{B_1}{N_1}\right)$$
$$= 1 - \left((1-P_1) + (1-P_1)(\phi-1) + \frac{B1}{N_1}\right)$$
$$= P_1 - \frac{(1-P_1)\{N_1(\phi-1)+B_1\} + P_1B_1}{N_1}$$
$$= P_1 - \theta^*$$

Hence

$$\frac{\gamma_1}{\gamma_1 + \gamma_2} < 1 - \frac{E(u_2 \mid N_1)e_2}{E(u_1 \mid N_1)e_1}$$

Therefore when the ratio of θ^* to $\gamma_1/(\gamma_1+\gamma_2)$ is small, we can use the values of γ_1 and γ_2 computed with equation (A5) as an approximate choice of γ_1 and γ_2.

ACKNOWLEDGEMENTS

The authors would like to thank A. Chao for valuable discussions and the anonymous referees for comments and suggestions which greatly improved this paper. We are also grateful to Dr Kumnuan Ungchusak of the Ministry of Public Health of Thailand for making available to us the various reports on Thai serosurveillance data cited in the References. This research is supported by grants from National Science Council of Taiwan for which the authors are grateful. Y.-H. Hsieh is also supported by Fogarty International Center/NIH grant (1 R03 TW00536-01).

REFERENCES

1. Weniger B, Limpakarnjanarak K, Ungchusak K, Thanprasertsuk S, Choopanya K, Vanichseni S, Thongcharoen P, Wasi C. The epidemiology of HIV infection and AIDS in Thailand. *AIDS* 1991; **5**(suppl 2):s71–s85.
2. Mason CJ, Markowitz LE, Kitsiripornchai S, Jugsudee A, Sirisopana N, Torugsa K, Carr JK, Michael RA, Nitayaphan S, McNeil JG. Declining prevalence of HIV-1 infection in young Thai men. *AIDS* 1995; **9**:1061–1065.
3. Sawanpanyalert P, Ungschusak K, Thanprasertsuk S, Akarasewi P. HIV-1 seroconversion rates among female commercial sex workers, Chiang Mai, Thailand: a multi cross-sectional study. *AIDS* 1994; **8**:825–829.
4. Brown T, Gullaprawit C, Sittitrai W, Thanprasertsuk S, Chamratrithirong A. *NESDB Projections for HIV/AIDS in Thailand*: *1987–2020*. Thai Red Cross Society Program on AIDS: Bangkok, 1994.
5. Celantano DD, Nelson KE, Suprasert S. Dynamics of risk behavior for HIV infection among young Thai men. *Journal of Acquired Immune Deficiency Syndromes* 1995; **10**:477–483.
6. Phoolcharoen W. HIV/AIDS prevention in Thailand: success and challenges. *Science* 1998; **280**:1873–1874.
7. Rojanapithayakorn W, Hanenberg R. The 100% condom program in Thailand. *AIDS* 1996; **10**(1):1–7.
8. Morris M, Pramualratana A, Podhista C, Handcock MS. Bridge populations in the spread of HIV/AIDS in Thailand. *AIDS* 1996; **10**:1265–1271.
9. Sittitrai W, Phanuphak P, Barry J, Brown T. *Thai Sexual Behavior and Risk of HIV infection*: *A Report of the 1990 Survey of Partner Relations and Risk of HIV infection in Thailand*. Program on AIDS, Thai Red Cross Society, and Institute of Population Studies: Chulalongkorn University, Bangkok, 1992.
10. Napaporn H, Bennet A, Knodel J. *Sexual Networking in a Provincial Thai Setting*. AIDS Prevention Monograph Series Paper No. 1: Bangkok, 1992.
11. Bhassorn L, Noppavan C, Penporn T, Wattana A. *The Demographic and Behavioral Study of Female Commercial Sex Workers in Thailand*. Institute of Population Studies: Chulalongkorn University, Bangkok, 1993.
12. Thongthai V, Guest P. Thai sexual attitudes and behaviors: results from a recent national survey. Report for Gender and Sexuality in Modern Thailand Conference, 11–12 July, 1995.
13. Taylor R. *A Review of Methods for Estimating the Size of Subgroups Particularly At Risk of Infection with HIV and Development of Proposals Which Could be Used to Enumerate These Populations in the Field*: *With Particular Reference to the Use of Capture-Recapture Methods for Estimating Populations of IV Drug Users and Prostitutes*. World Health Organization, Global Programme on AIDS: Geneva, 1989.
14. International Working Group for Disease Monitoring and Forecasting. Capture-recapture and multiple-record systems estimation I: history and theoretical development. *American Journal of Epidemiology* 1995; **142**(10):1047–1058.
15. Frischer M, Bloor M, Finlay A, Goldberg D, Green S, Haw S, McKeganey N, Platt S. A new method of estimating prevalence of injecting drug use in an urban population: results from a Scottish city. *International Journal of Epidemiology* 1991; **20**(4):997–1000.
16. Frischer M, Green ST, Goldberg DJ, Haw S, Bloor M, McKeganey N, Covell R, Taylor A, Gruer LD, Dermot K, Follett EA, Emslie JA. Estimates of HIV infection among injecting drug users in Glasgow, 1985–1990. *AIDS* 1992; **6**(11):1371–1375.
17. Frischer M. Estimated prevalence of injecting drug use in Glasgow. *British Journal of Addiction* 1992; **87**(2):235–243.
18. Mastro TD, Kitayaporn D, Weniger BG, Vanichseni S, Laosunthorn V, Uneklabh T, Uneklabh C, Choopaya K, Limpakarnjanarat K. Estimating the number of HIV-infected injection drug users in Bangkok: A capture-recapture method. *American Journal of Public Health* 1994; **84**(7):1094–1099.
19. Bloor M, Leyland A, Barnard M, McKeganey N. Estimating hidden populations: a new method of calculating the prevalence of drug-injecting female street prostitution. *British Journal of Addiction* 1991; **86**(1):1147–1483.
20. Neugebauer R, Wittes J. Annotation: Voluntary and involuntary capture-recapture samples-problems in the estimation of Hidden and elusive populations. *American Journal of Public Health* 1994; **84**(7):1068–1069.
21. International Working Group for Disease Monitoring and Forecasting. Capture-recapture and multiple-record systems estimation II: applications in human diseases. *American Journal of Epidemiology* 1995; **142**(10):1059–1068.
22. Abeni DD, Brancato G, Perucci CA. Capture-recapture to estimate the size of the population with human immunodeficiency virus type 1 infection. *Epidemiology* 1994; **5**(4):410–414.
23. Seber GFA. *The Estimation of Animal Abundance*, 2nd edn. Griffin: London, 1982.
24. Moran PAP. A mathematical theory of animal trapping. *Biometrika* 1951; **38**:307–311.

25. Zippin C. An evaluation of the removal method of estimating animal populations. *Biometrics* 1956; **12**:163–169.
26. Zippin C. The removal method of population estimation. *Journal of Wildlife Management* 1958; **22**:82–90.
27. Otis DL, Burnham KP, White GC, Anderson DR. *Statistical Inference for Capture Data on Closed Animal Populations*. Wildlife Monographs, v. 62, University of Kentucky: Louisville, 1978.
28. Chaiyapong Y, Lloyd CJ. Improved inference from recapture experiments with behavioural response through modelling and auxiliary experimentation. *Australian Journal of Statistics* 1996; **38**:317–331.
29. Yip P, Fong YT. Estimating population size from a removal experiment. *Statistics & Probability Letters* 1993; **16**:129–135.
30. Ministry of Public Health. HIV sentinel serosurveillance in Thailand, Round 9–13. Division of Epidemiology, Ministry of Public Health, Thailand, June 1993–June 1995.
31. Ungchusak K, Saengwonloey O, Junsiriyakorn S, Rujiviput V, Thepsittha K, Thonghong A. Result of HIV serosurveillance, 12th round, December 1994. Division of Epidemiology, Ministry of Public Health, Thailand, 1995.
32. Lee SM, Chen CWS. Bayesian inference of population size for behavioral response models. *Statistica Sinica* 1998; **8**:1233–1247.
33. Chen CWS, Lee SM, Hsieh Y-H, Ungchusak K. A unified approach to estimating population size of births only model. *Computational Statistics and Data Analysis* 1999;**32**:29–46.
34. George EI, Robert CP. Capture-recapture estimation via Gibbs sampling. *Biometrika* 1992; **79**:677–683.
35. Castledine BJ. A Bayesian analysis of multiple-recapture sampling for a closed population. *Biometrika* 1981; **67**: 197–210.
36. Gelfand AE, Smith AFM. Sampling-based approach to calculating marginal densities. *Journal of American Statistical Association* 1990; **85**:398–409.
37. Gelfand AE, Hills SE, Racine-Poon A, Smith AFM. Illustration of Bayesian inference in normal data models using Gibbs sampling. *Journal of the American Statistical Association* 1990; **85**:972–985.
38. Casella G, George EI. Explaining the Gibbs sampler. *American Statistician* 1992; **46**(3):167–174.
39. Gilks WR, Richardson S, Spiegelhalter DJ. *Markov Chain Monte Carlo in Practice*. Chapman and Hall: London, 1996.
40. Raftery AE, Lewis SM. How many iterations in the Gibbs sampler? In *Bayesian Statistics 4*, Bernardo JM, Berger JO, Dawid AP, Smith AFM (eds). Oxford University Press: Oxford, 1992; 765–776.
41. Garnett GP, Anderson RM. Factors controlling the spread of HIV in heterosexual communities in developing countries: pattern of mixing between different age and sexual activity classes. *Philosophical Transactions of the Royal Society of London, Series B* 1993; **342**:137–159.
42. Anzala A, Wambugu P, Plummer FA. Incubation time to symptomatic disease and AIDS in women with known duration of infection. Abstract TUC103, VII International Conference on AIDS, Florence, 1991.
43. Kitayaporn D, Tansuphaswadikul S, Lohsomboon P, Pannachet K, Kaewkungwal J, Limpakarnjanarat K, Mastro TD. Survival of AIDS patients in the emerging epidemic in Bangkok, Thailand. *Journal of Acquired Immune Deficiency Syndromes* 1996; **11**:77–82.
44. Lee SM, Chao A. Estimating population size via sample coverage for closed capture-recapture models. *Biometrics* 1994; **50**:88–97.

