European Journal of Epidemiology **19**: 1075–1083, 2004.

METHODS

Estimating the number of drug users in Bangkok 2001: A capture–recapture approach using repeated entries in one list

Dankmar Böhning[1], Busaba Suppawattanabodee[2], Wilai Kusolvisitkul[3] & Chukiat Viwatwongkasem[3]

[1]*Division for International Health, Institute of Social Medicine, Epidemiology, and Health Economy, Charité University Medical Center Berlin, Berlin, Germany;* [2]*Clinical Epidemiology Unit, Vajira Hospital, Dusit District, Bangkok, Thailand;* [3]*Department of Biostatistics, Faculty of Public Health, Mahidol University, Bangkok, Thailand*

Accepted in revised form 7 September 2004

Abstract. *Background*: Conventionally, capture–recapture techniques involving different lists such as police or hospitals are used for quantifying populations which are difficult to count, such as illicit drug user populations. Here, a novel approach is suggested based upon repeated entries in one list, which is less dependent on matching entries from different sources as in the conventional approach. *Methods*: For this purpose, a population-based study was conducted that utilizes all data on treatment episodes of drug users from all 61 health treatment centers in the Bangkok metropolitan region to estimate the size of drug use in the Bangkok metropolitan region. The data stem from the drug treatment surveillance system of the Office of the Narcotics Control Board (ONCB) and cover the period from October 1 to December 31, 2001. Based upon the frequency of treatment episodes of each patient, a count distribution arose which could be modelled well by means of a Poisson mixture model. Using this count model, an estimate for the number of unobserved drug users could be constructed. *Results*: From 11,222 drug users found during the period, 7063 (62.9%) were heroin users, 3346 (29.8%) metamphetamine users, and the remaining 813 (7.3%) distributed under 15 drug categories, none above 1%. The study concentrated on heroin and metamphetamine users who were predominantly male (96.2% for heroin and 91.8% for metamphetamine). Metamphetamine users were younger than heroin users (22.3 years 95% CI: 22.1–22.5 vs. 30.8 years 95% CI: 30.6–31.0). By using the truncated Poisson mixture model, an estimate of the unobserved frequency of drug users with zero treatment episodes could be constructed leading to an estimate of 11,296 (95% CI: 8,964–13,628) heroin users (completeness of identification: 38.42, 95% CI: 34.03–44.04%) and 32,105 (95% CI: 24,647–39,563) metamphetamine users (completeness of identification: 9.44, 95% CI: 7.79–11.97%) for the Bangkok metropolitan region. *Conclusions*: The proposed model showed excellent goodness-of-fit, unspecified for drug type and also if specified for the major drug types which allowed the prediction of the unobserved number of drug users in a realistic way, avoiding artefacts due to severe matching problems when using several, different sources. The technique is also easy to implement and can be used routinely to monitor drug user populations in space and time.

Key words: Capture–recapture, Heroin and metamphetamine use in Bangkok, Truncated Poisson mixture model

Introduction

Drug abuse has become a serious health problem for many countries including Thailand. Contrary to tobacco smoking or alcohol consumption which is not prohibited above a certain age limit, the consumption of other drugs like heroin or marijuhana is illegal in many countries. Due to this fact, the counting of human populations consuming illegal drugs is considerably more difficult than counting populations with health problems not faced with any legal sanctions. Since the reviewing papers were published by Hook and Regal [1] and the International Working Group for Disease Monitoring and Forecasting [2, 3], capture–recapture methods were applied frequently to these and similar health issues in epidemiology and public health in general. A MEDLINE-search with search term capture–recapture in public health resulted in only four publications found before 1980. From 1981 to 1990 19 papers were found, whereas from 1991 to 2002, 210 publications were detected. Some recent applications of capture–recapture techniques to illicit drug use were the estimation of injecting drug users in Edinburgh, Scotland, 1992–1994 [4], the size of opiate use in Barcelona, Spain 1993 [5], the estimation of opiate users in Amsterdam, Netherlands, 1997 [6], the amount of drug misuse in urban and non-urban settings in the north-east of

Scotland [7], and the estimation of prevalence of opiate use in Dublin, Ireland [8]. For Thailand, there have been only a few studies of this kind including a study on HIV-infected injection drug users in Bangkok 1991 [9]. Many capture–recapture contributions in public health use a modelling approach with two or more sources or lists. In the field of drug use, these sources are often a hospital-like institution and the police (if there are two sources [8, 9]) or treatment centers, surveys, family doctors etc. (if there are several sources [4–7]). If there are only two lists, an estimate for the population size can be constructed by multiplying the frequencies of persons identified by each list and dividing the result by the frequency of persons found on both lists, the Lincoln–Petersen estimate [10]. The estimator is only valid if the lists are statistically independent (a doubtful assumption for human populations as it has been pointed out by several authors) [2, 11]. An improved estimator was suggested by Chapman [12] by adding 1 unit to all three frequencies in the Lincoln–Petersen estimate and substracting 1 unit from the result. This is referred to as the Lincoln–Petersen–Chapman-, or just the Chapman-estimate. When three or more lists are used, the independence assumption can be relaxed and log-linear modelling is recommended to provide a more realistic estimate of the population size [2]. Another crucial assumption is that of *perfect* matching, e.g., a person entered on one list should be found on any other list if the person is entered on this list, and also should not be found on any other list if the person is not entered on this list. Matching has to be defined by criteria such as name, first name, date of birth, place of birth, etc. The stricter the criteria, the higher the estimate of the population size (danger of overestimation), and the less strict the criteria, the lower the estimate of population size (danger of underestimation). Using the two-sources Lincoln–Petersen–Chapman model, the effect of using different matching criteria is illuminated for mortality from road traffic accidents in Karachi, Pakistan, leading to rates from 9.7 to 31.5% [13].

Another approach is based on a single source with repeated entries (at least one) in the list during the observational period. This approach is less dependent on matching criteria if a unique identifier such as an ID card can be used to identify repeated entries. Counting the repeated entries in the list leads to a frequency distribution of counts and, if an appropriate model for this count distribution can be found, an estimate of the (unobserved) frequency of zero entries in the list can be constructed [14]. This approach is common in ecological and wildlife applications [15], less frequent in public health applications. One of the earliest occurrences of this approach in public health goes back to McKendrick in 1926 who applied it to a cholera epidemic in a village in India. The data available were the counts of cholera cases in each house from which McKendrick constructed an estimate of the number of cholera-affected households with zero cases of cholera [16]. More recent applications were the estimation of number of dependent heroin users in Australia [17], drug injectors from needle exchange data in Scotland [18] or illicit drug users in Los Angeles County [19]. Another approach using repeated entries divides the sampling period in sub-intervals and considers each of these as a separate source [20].

The approach based upon counting repeated identifications was also briefly mentioned in Hook and Regal (p. 260) [1], emphasizing the frequent application of the method in genetic epidemiology. Count models differ in the way the count distribution is specified. The simplest model is the Poisson, which is frequently not flexible enough in capturing population heterogeneity. Parametric generalizations have been used including the Poisson-Gamma or Poisson-Normal distribution [21]. Semi-parametric generalizations have been of more recent nature [22], having the advantage of being distribution-free for the heterogeneity or random effect distribution.

There seems to be controversy as to whether the latter approach should be categorized as capture–recapture or something else. Clearly, the approach is building upon repeated entries in one list and developing a distributional model for the truncated counts of entries for a specific person or unit in this list. It is not building on the overlap of two or more different sources which is typical for conventional capture–recapture models. However, since most of the previous work [18, 19, 21–24] is categorizing the truncated count model under capture–recapture it appears adequate to consider this approach as a special capture–recapture model.

Methods

Data sources and characteristics

The study used all data on drug use from 61 health treatment centers in the Bangkok metropolitan region collected by the Office of the Narcotics Control Board (ONCB), Ministry of the Prime Minister, which occurred from October 1 to December 31, 2001. All private and public health treatment centers in the Bangkok metropolitan region licensed by the Ministry of Public Health to treat drug dependence were included in the study. According to the drug dependence treatment regulation of the Ministry of Public Health, all operating treatment centers have to report the admission of any drug dependent patient through a standard form to be submitted to central agencies including the ONCB. Since the launching of the surveillance system in 1979, this standard form has been revised several times. For this study, the latest revision of 1998 has been used, which includes about 27 items including information on demo-

graphics, social class and profession as well as on the form of drug dependency. Each patient received a unique identification code identical to the identification number of the Thai ID card. This code was used to enter information about the patient every time the patient initialized a new treatment episode. Repeated entries were determined using the identification code. If the code was missing on the ONCB-form, repeated entries were identified using first and second name and date of birth as matching criteria. Treatment episodes could last from a day to 45 days for heroin users and 5–7 days for metamphetamine users. From the available data source, it was possible to construct the information on the frequency of episodes for each patient in the sampling period, which served as the key element in the prediction process.

Statistical analysis

Count distribution of treatment episodes

Let $X_1, X_2, \ldots X_N$, denote the number of treatment episodes that occurred in the observational time window for the N drug users in the community under study. X_i is a count variable with values in $\{0, 1, 2, \ldots\}$ for each of the N drug users in the community. However, if $X_i = 0$ then there are no episodes recorded in any of the health treatment centers. Consequently, drug users are only observed if $X_i \geqslant 1$. The associated, observable count distribution is also referred to as the *truncated* count distribution [1]. This implies that N remains unknown and estimating N is the objective of the study. Let n denote the observed number of drug users, e.g. the size of the observed data set $\{X_i | X_i \geqslant 1, i = 1, \ldots, N\}$. Furthermore, let $p_x = Pr\{X = x\}$ be the probability of an arbitrary drug user having exactly x treatment episodes. Accordingly, p_0 is the proportion of drug users not observed in the sample. Since

$$\begin{aligned} N &= \text{number of observed} \\ &\quad + \text{ number of unobserved drug users} \\ &= n + Np_0 \end{aligned}$$

it follows that an estimate $\hat{N}$ of N can be constructed according to

$$\hat{N} = \frac{n}{1 - p_0}, \tag{1}$$

the Horvitz–Thompson estimator of the population size. (1) requires knowledge of p_0 and approaches differ in the way they accomplish estimating p_0.

Modelling the count distribution of treatment episodes

A simple count distribution is the Poisson distribution given as $p_x = e^{-\lambda}\lambda^x/x!$, for $x = 0, 1, 2, \ldots$, where $\lambda \geqslant 0$ is the unknown parameter in the Poisson distribution. The value of $S = X_1 + X_2 + \cdots + X_N$ is observed since those X_i which are 0 (and therefore not observed), will not affect the value of S. Suppose some initial value of $\hat{N}$ ($\hat{N} = n$, say) has been assigned. Then, the best estimate of λ is the mean

$$\hat{\lambda} = \frac{S}{\hat{N}}, \tag{2}$$

from which a new estimate of N according to (1) with $\hat{p}_0 = e^{-\hat{\lambda}}$ is constructed

$$\hat{N} = \frac{n}{1 - \hat{p}_0}. \tag{3}$$

From here, a new estimate of λ is built according to (2), then again a new estimate of N is constructed according to (3) with current value $\hat{p}_0 = e^{-\hat{\lambda}}$. This algorithmic procedure will cycle back and forth between steps (2) and (3) until convergence. The resulting estimate is the maximum likelihood estimate of λ [16, 25].

The Poisson distribution is typically not flexible enough to capture the heterogeneity in the observed distribution of counts. A powerful alternative is mixtures of Poisson distributions defined as $p_x = \sum_{j=1}^{k} e^{-\lambda_j}\lambda_j^x/x! q_j$, where λ_j is the Poisson parameter in the j-th component and q_j is the associated non-negative component weight $(q_1 + q_2 + \cdots q_k = 1)$. Suppose an estimate $\hat{N}$ of N is available. This implies that an estimate $\hat{n}_0 = \hat{N} - n$ of the frequency n_0 of drug users with 0 treatment episodes is readily available as well. Together with the observed set of frequencies $n_1, n_2, n_3, \ldots$ of drug users with 1, 2, 3, ... treatment episodes one can use the EM algorithm to find the maximum likelihood estimates $\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_k$ of the Poisson components as well as the component weights $\hat{q}_1, \hat{q}_2, \ldots, \hat{q}_k$ [26]. From here, a new estimate of $\hat{N}$ can be calculated according (4).

$$\hat{N} = \frac{n}{1 - \hat{p}_0}, \quad \hat{p}_0 = \sum_{j=1}^{k} \hat{q}_j e^{-\hat{\lambda}_j}. \tag{4}$$

Having an update of $\hat{N}$ available, an updated maximum likelihood estimate of the Poisson mixture is calculated. This process continues until convergence. The number of components k is chosen such that the likelihood is maximized. Typically, this is done iteratively, starting with $k = 1$, and then step-wise increasing the number of components by one, until no further increase in the log-likelihood is possible. The resulting estimate is also called the *non-parametric maximum likelihood estimate* of the mixing distribution and usually associated with Laird [27]. A software tool C.A.MUst, an extension of C.A.MAN [28], has been used for all calculations.

Model evaluation

Let $\hat{n}_1 = \hat{N}\hat{p}_1, \hat{n}_2 = \hat{N}\hat{p}_2, \ldots$ denote the fitted frequencies of the drug users with 1, 2, ... treatment episodes with associated probabilities $\hat{p}_1, \hat{p}_2, \ldots$,of observing a drug user with 1, 2, ... treatment episodes estimated under the model under consideration. These frequencies, fitted under the model, can be

compared with the observed ones to form the Neyman-χ-square

$$\chi^2 = \sum_{i=1}^{m} \frac{(n_i - \hat{n}_i)^2}{n_i}, \quad (5)$$

where m is the largest number of treatment episodes considered. The degrees of freedom are calculated as $df = m - 1 - p$, where p is number of estimated parameters in the model. For the simple Poisson model $p = 1$, for the Poisson mixture $p = 2k - 1$. To perform a valid model evaluation the degrees of freedom should at least be 1, implying that the number of components k in the Poisson mixture should not become larger than $m/2$, half of the largest number of treatment episodes considered. Count distributions in these settings have often few observations in the upper tail areas. To avoid artefacts, cells in the tail areas were collapsed starting with the first cell m having less than five observations. The Neyman-χ-square has been chosen in contrast to the Pearson-χ-square to avoid an artificial blow-up of the χ-square-value due to very small Poisson probabilities in the upper tail of the count distribution.

Confidence intervals

For confidence interval estimation, Bootstrap resampling techniques were used [1-3]. If p_0 in (1) would be known, the only source of variation in the estimator $\hat{N}$ is coming from sampling the n elements out of N. If p_0 is estimated using the truncated Poisson mixture model, there is a second source of random variation arising. To mimic both sources of variation, the Bootstrap was realized in the following fashion. Firstly, $n^{(b)}$ were sampled from a Binomial distribution with success parameter $p = n/\hat{N}$ and sample size parameter $\hat{N}$, where $\hat{N}$ was estimated from the original data set. Secondly, frequencies $n_1^{(b)}$, $n_2^{(b)}$,...were sampled from the truncated Poisson mixture model with parameters as estimated in the original data set, with $\sum_i n_i^{(b)} = n^{(b)}$. For each of these B resamples, $n_0^{(b)}$ was estimated, $b = 1, \ldots, B$, and these resample data were used to compute standard errors and confidence intervals. It was found that the statistics of interest stabilized beyond $B = 500$, so that $B = 1000$ was considered to be sufficient in all Bootstrap calculations.

Results

Demographic characteristics

A total of 11,222 drug users were identified in the Bangkok metropolitan area by means of the ONCB during the time interval. These are predominantly metamphetamine (3346 or 29.8%) and heroin (7063 or 62.9%) users on which attention is focussed in the following. Table 1 shows that metamphetamine users are considerably younger than heroin users (mean age is 22.3 years with 95% CI: 22.1–22.5 for the former vs. mean age 30.8 years with 95% CI: 30.6–31.0 for the latter). Heroin users show three times higher mean of treatment episodes count than metamphetamine users. Heroin users are more often married than metamphetamine users, but they are also more often separated and divorced/widowed. However, this difference is mainly due to the difference in the age structure of heroin and metamphetamine users and is strongly diminished if considered within the four age groups defined by the age quartiles.

Table 1. Number of drug users identified by ONCB, mean age with standard deviation, proportion of males in percentage, and mean treatment episodes with standard deviation

Drug	Count	Mean age (SD)	Percentage of males	Mean TE (SD)
Heroin	7048	30.8 (8.2)	96.0	2.9 (2.52)
Metamphetamine	3334	22.3 (5.9)	91.8	1.1 (0.66)
Other	812	34.3 (11.5)	94.7	2.6 (2.68)

SD, standard deviation; TE, count of treatment episodes.

Estimating the number of heroin and metamphetamine users

Here, the modelling of the distribution of the counts of the treatment episodes is considered. Figure 1 shows three curves for the group of heroin users: the distribution of observed counts, the distribution of predicted counts under the homogeneous Poisson model and under the mixed Poisson model. The homogeneous Poisson has a bad goodness-of-fit value $\chi^2 = 3245.20$ with 13 df. The Poisson mixture gives an acceptable goodness-of-fit value with $\chi^2 = 5.65$ and 2 df. Figure 2 shows the analogous curves for the group of metamphetamine users. Again, good per-

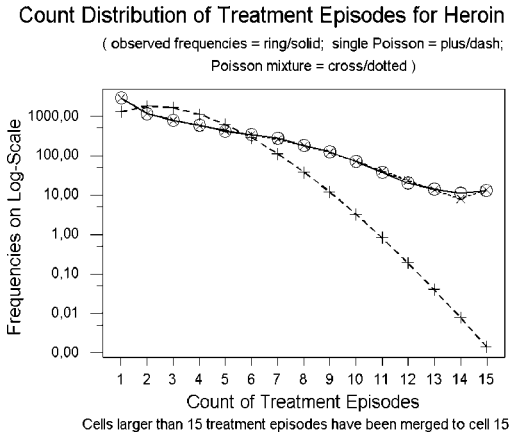

Figure 1. Count distribution of treatment episodes for heroine users.

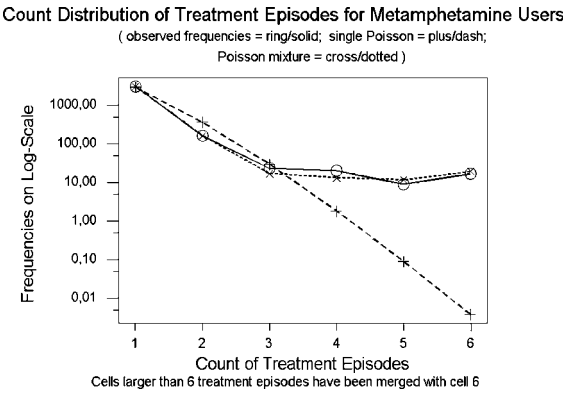

Figure 2. Count distribution of treatment episodes for metamphetamine users.

fomance values are found for the Poisson mixture model with $\chi^2 = 4.98$ and 2 df vs. $\chi^2 = 297.23$ with 4 df for the homogeneous Poisson model.

Number of heroin users

The estimate of the unobserved number of heroin users is provided by the Poisson mixture model as 10,219 with 95% CI: (7046–13,392), (see Table 2). The count distribution of treatment episodes changes

Table 2. Number of unobserved heroin users with 95% CI estimated using the poisson mixture model

		Discrete mixture		
Age group	$\hat{n}_0$ 0(95% CI)	λ_j	p_j	k
Unstratified	10,219	0.214	0.705	4
	(7046–13,392)	2.130	0.187	
		5.850	0.105	
		12.200	0.003	
I	754	0.384	0.736	4
	(0–1530)	2.967	0.173	
		7.008	0.089	
		14.563	0.003	
II	3685	0.157	0.704	4
	(2493–4877)	1.975	0.184	
		5.755	0.106	
		11.631	0.005	
III	4607	0.122	0.766	4
	(3360–5854)	2.084	0.156	
		5.719	0.074	
		11.524	0.004	
IV	2250	0.362	0.701	3
	(886–3614)	2.459	0.182	
		5.936	0.117	
Age-stratified	11,296[a]			
	(8964–13,628)[b]			

[a]sum of estimates of n_0 for the four age groups.
[b]CI was constructed as asymptotic normal interval using as estimated variance the sum of the Bootstrap variances from the four age groups.

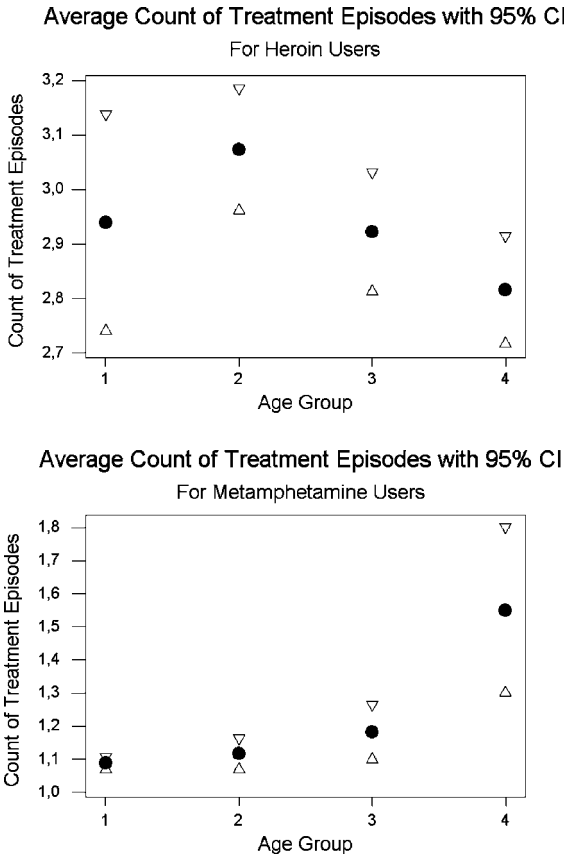

Figure 3. Mean count of treatment episodes for heroin (top) and metamphetamine (bottom) users.

with age (see also Figure 3, top), so that also an age-adjusted estimate of the unobserved number of heroin users is computed leading to 11,296 with 95% CI: (8964–13,628), (see Table 2 again). Together with the observed number of 7048, the total number of heroin users in Bangkok Metropolis is estimated as 18,344 with 95% CI: (16,006–20,710). Using the Bangkok Metropolis Census of 2000, this estimate leads to a prevalence of 4.79 heroin users per 1000 residents aged 15–44 years with 95% CI: (4.18–5.41).

Number of metamphetamine users

The estimate of the unobserved number of metamphetamine users is provided by the Poisson mixture model as 30,730 with 95% CI: (25,130–36,329) (see Table 3). The count distribution of treatment episodes shows a significant trend increasing with age (see also Figure 3, bottom). This requires an age-adjusted estimate of the unobserved number of metamphetamine users leading to 32,105 with 95% CI: (24,647–39,563), (see Table 3 again) which is considerably higher than the unstratified estimator. Together with the observed number of 3346, the total number of metamphetamine users in Bangkok Metropolis is estimated as 35,439 with 95% CI: (27,966–42,913). Using the Bangkok Metropolis Census of 2000, this estimate leads to a prevalence

Table 3. Number of unobserved metamphetamine users with 95% CI estimated using the poisson mixture model

		Discrete mixture		
Age group	$\hat{n}_0$ (95% CI)	λ_j	$\hat{p}_j$	k
Unstratified	30,730 (25,130–36,329)	0.101 4.443	0.990 0.010	2
I	22,354 (15,802–28,906)	0.088 3.220	0.998 0.002	2
II	6316 (3296–9336)	0.100 4.44	0.998 0.002	2
III	2445 (1159–3731)	0.142 5.050	0.996 0.004	2
IV	990 (0–2376)	0.147 5.574	0.984 0.016	2
Age-stratified	32,105[a] (24,647–39,563)[b]			

[a]sum of estimates of n_0 for the four age groups.
[b]CI was constructed as asymptotic normal interval using as estimated variance the sum of the Bootstrap variances from the four age groups.

of 9.26 metamphetamine users per 1000 residents aged 15–44 years with 95% CI: (7.31–11.21).

Estimating completeness of identification

For health surveillance systems such as the one provided by the ONCB of Thailand, it has become important to provide an estimate of its *completeness*. *Completeness* is defined as $n/N \times 100\%$ which can be estimated as $n/\hat{N} \times 100\%$. For heroin users, the completeness of identification is $7048/18{,}344 \times 100\% = 38.42\%$ (95% CI: (34.03–44.04%)), whereas for metamphetamine users the completeness of identification is considerably lower with $3346/35{,}439 \times 100\% = 9.44\%$ (7.79–11.97%). Due to the severeness of symptoms, the completeness of identification is higher for heroin users than for users of metamphetamine.

Discussion

One or several sources

Typically, for estimating the population size, capture–recapture procedures employ several sources such as hospitals, police stations, or practicing physicians. In a previous study on the site of the drug user population in Bangkok [9], health treatment centers were used as one source and police stations as the other source. One of the difficulties in this approach is to achieve a perfect matching of the two lists. Typically, matching criteria involve variables such as name, surname, gender, age, type of drug, place of residence. If each of these variables have the two values "same" and "similar", then a variety of different matching types are possible. In a recent study [29], using again the two sources of health treatment centers and police stations, the estimate of the drug user Bangkok population size was considered dependent on of six different matching types. The strictest rule required all of the four matching variables name, surname, gender, and age to be the same, whereas the least strict one required only gender to be the same with name, surname, and age only being similar. (The similar-rule for name and surname can be considered as a potential adjustment for the easy occurrence of spelling errors in the names in the Thai language). The estimates provided a range from 430,000 (weakest matching rule) to the unrealistic estimator of 3,300,000 (most rigorous rule) [29] which would make more than half of the Bangkok metropolis look like drug users. This shows the severe dependence of two source estimates (even if the independence assumption for the two sources is valid) from the matching criteria.

On the other hand, as one reviewer pointed out, the matching problem is not necessarily solved when using repeated entries in a single source. For example, different centres might generate different identifiers. However, in this study, the identification number from the Thai ID card was used to determine repeated entries which is a unique number and every Thai person older than 14 years must have an ID card. Nevertheless, errors due to matching could not completely be ruled out since missing entries for the identification number did occur, and for those cases first and second name as well as date of birth were used to match repeated entries.

The sampling interval was set as 3 months for this study. This length of time interval is not untypical for drug user studies. Hey and Smit [18] used sampling intervals from 1 to 12 months. Hser [19] used a sampling interval of 1 year. For conventional capture–recapture procedures, Pal et al. [34] used a 5-months period as sampling interval, while Debrock et al. [33] used a sampling interval of 2 months and Razzak and Luby [13] used a 10 months sampling period. For estimating opiate users in Amsterdam with three sources, Buster et al. [6] argue that a 3-months sampling interval is best suitable for meeting the closure assumption of the population. Similarly, using truncated count models to estimate the number of injecting drug users in Scotland, Hay and Smit [18] write: "Keeping the study period short (say, 1 month) is one way of meeting the closure assumption. It is hard to see how the population size of opiate users can change dramatically in a single month". A potential drawback of choosing a short study period is that high frequent treatment episodes are missed and the true variation of the counting distribution underestimated. Whereas the simple homogenous Poisson model is sensitive to large counts, the Poisson mixture is less sensitive to the occurrence or non-occurrence of large counts. In particular, the estimate of the population size is largely dependent on lower

component values in the Poisson mixture, determined by the lower observed frequencies. Nevertheless, in this study, the occurrence of frequent treatment episodes for heroin users could be observed (which occurred in other, similar studies as well [19, 23]. For metamphetamine users, treatment episodes are less frequent (see also Figure 3).

An estimate of 36,600 opiate users is given for 1991 with an approximation of 12,000 for the HIV-infected injections drug users [9]. The approach taken here (which is less dependent on matching entries in different sources) provides an estimate of 18,300 for the heroin users. Since 83% of the observed heroin users are injecting, the number of heroin injecting drug users is estimated as 15,200 for the year 2001 which seems of similar dimension as the numbers given in the before-mentioned study [9].

Using a truncated count distribution to predict the population size avoids two crucial assumptions of a two sources approach: the independence assumption and the assumption of perfect matching of members on two or several lists (though it has to be assumed/guaranteed in the truncated count distribution approach that repeated occurrences are not treated as different entries).

Hook and Regal [1] note that the count distribution also occurs as the marginal distribution of counts identified by one source, two sources, three sources, etc., though it is less appealing in situations where several sources are available (since the available information in the contingency Table is only partially utilized with the truncated count distribution). Nevertheless, in an internal validity analysis [30] which compared log-linear and parametric truncated count distributions, it was demonstrated that the parametric count distribution provided surprisingly good estimates with error rates of not more than 10%. The counting distribution occurs as the marginal count of all sources involved in the capture–recapture study. *Vice versa*, given only one source, several sources can be constructed from this source by subdividing the given sampling interval into several sub-intervals, creating several sources [20].

Using truncated count distributions for the number of treatment episodes has be done previously. Wickens [23] reviews the simple Poisson model for a data set of treatment episodes of intravenous drug users in Los Angeles taken from Hser [19]. Hay and Smit [18] discuss the estimation of the number of drug injectors from needle exchange data in Scotland which utilize information on the number of attendances of individual drug users. However, these previous approaches maintain a simple modelling solution.

Adjusting for unobserved heterogeneity and Poisson mixture

Consider a parametric model like the Poisson distribution with a specific parameter. This model is frequently unable to capture all the variability in observed data. Instead, different parameters will be required to fit different parts of the observed data. However, it is not observed which parts of the data belong to which parametric part of the model: there is *unobserved heterogeneity*. It might be even helpful to think of unobserved heterogeneity as a *missing covariate* [26]. Ignoring it will lead to model misspecification and the prediction of population size will be biased, in fact, it will *underestimate* the population size. Consequently, accounting for unobserved heterogeneity is important. The associated model accounting for it leads to a mixture model, mixing over the parametric part, here the Poisson distribution [31]. There is freedom in choosing the mixing distribution. It is quite common to think of the Gamma distribution as potential parametric mixing distribution as described in the context of estimating the size of a criminal population by Rossmo and Routledge [24].

However, since there is neither a theoretical agrument for choosing a parametric mixing density like the Gamma, nor any empirical way to find evidence for or against this *unobserved heterogeneity distribution*, it is left non-parametric here and estimated with a discrete mass distribution. For the untruncated Poisson mixture it was shown [32] that the unspecified mixing distribution is identifiable and the maximum likelihood estimate is a unique, discrete mass distribution, which Laird [27] called the *non-parametric maximum likelihood estimator*. The non-parametric maximum likelihood estimator of the mixing distribution also provides an estimate of the number of components K which is used in all analyses here. Following the arguments elsewhere [32] identifiability and uniqueness can be deduced for the truncated Poisson mixture as well.

It is clear from the preceding discussion that the simple truncated Poisson model will underestimate, potentially severely, the population size. An interesting alternative was suggested by Zelterman [21] that was frequently used recently to estimate the size of a drug user population [18]. The estimator of Zelterman is defined as

$$\hat{N}_Z = \frac{n}{1 - \exp(-2n_2/n_1)}$$

using the fact that the Poisson parameter can be written as $\lambda = (j+1)Po(j+1,\lambda)/Po(j,\lambda)$,where $Po(i,\lambda) = e^{-\lambda}\lambda^i/i!$ This expression can be estimated using $(j+1)n_{j+1}/n_j$ for any $j = 1, 2, \ldots$, though typically lower count values will have the highest frequencies. The choice of $j = 1$ leads to the estimate $\hat{N}_Z$. The interesting point here is that the estimator avoids using the simple Poisson model directly and should be more robust against model misspecification. Indeed, if the components of the Poisson mixture are well separated, the Zelterman estimator gives reasonable results: $\hat{N}_Z = 33{,}664$ for the metamphetamine users in comparison to $\hat{N} = 34{,}076$

using the mixture model. Here, the two components are quite apart from each other (see Table 3), which explains the good performance of the Zelterman estimator. For the heroin users, the two estimators differ more substantially: $\hat{N}_Z = 12{,}796$ in comparison to $\hat{N} = 17{,}267$ using the mixture model. Here, the mixture components are not well-separated (see Table 2), and the Zelterman estimator underestimates the population size considerably.

Adjusting for observed heterogeneity

To achieve an unbiased estimator, it is important to adjust for *observed* heterogeneity as well, which is typically described in the form of covariates. In this study, as important covariates, demographic variables were considered such as study age, gender, marital status, and others. Adjusting for observed heterogeneity can simply be accomplished by means of stratification. An age dependence of the observed distribution of treatment episodes was established, in particular for metamphetamine users, and, consequently, age-adjusted estimates of the population size were provided. For gender, no significant differences in the distribution of treatment counts could be found. For marital status, observed differences in the episode count distributions vanished when stratified by age. It appears that the major demographic variable have been accounted for, and remaining, unobserved confounders are adjusted for by means of the Poisson mixture model.

Conclusions

The proposed model shows excellent goodness-of-fit which appears to be one of the major benefits compared to the simpler (and not well fitting) models used previously [23] and which should allow the prediction of the unobserved number of drug users in a more realistic way. Accounting for unobserved heterogeneity is essential and full modelling of the heterogeneity distribution is important to avoid underestimating the size of the drug user population. The suggested methods provide the tools for doing so and also give numerical algorithms for applying these tools. The technique is also easy to implement and can be used routinely to monitor drug user populations in space and time.

Finally, it should be mentioned that even good fitting count distributional models will *not* guarantee unbiased prediction of the population size. Here, only validation studies (in which populations with known size are used for comparison) can help to gain further insight into the predictive abilities of the suggested approach.

Acknowledgements

The work of Dankmar Böhning and Chukiat Viwatwongkasem is supported by the German Research Foundation and the National Research Council of Thailand. The authors wish to express their thanks to the following institutions in Thailand: (1) The Office of the Narcotics Control Board, (2) The Drug Abuse Prevention and Treatment Division, Health Department, Medical Service Department, (3) Vajira Drug Service Center, Bangkok Metropolitan Administration, and (4) The Ethical Committee Board of the Bangkok Metropolitan Administration. The authors are grateful to the Editor and Reviewers for their most helpful comments.

References

1. Hook EB, Regal R. Capture–recapture methods in epidemiology: Methods and limitations. Epidemiol Rev 1995; 17: 243–264.
2. International Working Group for Disease Monitoring and Forecasting. Capture–recapture and multiple record systems estimation I. History and theoretical development. Am J Epidemiol 1995; 142: 1047–1058.
3. International Working Group for Disease Monitoring and Forecasting. Capture–recapture and multiple record systems estimation II. Applications in human diseases. Am J Epidemiol 1995; 142: 1059–1068.
4. Davies AG, Cormack RM, Richardson AM. Estimation of injecting drug users in the City of Edinburgh, Scotland, and number infected with human immunodeficiency virus. Int J Epidemiol 1999; 28: 117–121.
5. Domingo-Salvany A, Hartnoll RL, Maguire A, et al. Analytic considerations in the use of capture–recapture to estimate prevalence: Case studies of the estimation of opiate use in the metropolitan area of Barcelona, Spain. Am J Epidemiol 1998; 148: 732–749.
6. Buster MCA, van Bussel GHA, van den Brink W. Estimating the number of opiate users in Amsterdam by capture–recapture: The importance of case definition. Eur J Epidemiol 2001; 17: 935–942.
7. Hay, G. Capture–recapture estimates of drug misuse in urban and nonurban settings in the north east of Scotland. Addiction 2000; 95: 1795–1803.
8. Comiskey CM, Barry JM. A capture–recapture study of the prevalence and implications of opiate use in Dublin. Eur J Public Health 2001; 11: 198–200.
9. Mastro TD, Kitayaporn D, Weniger BG, et al. Estimating the number of HIV-infected injection drug users in Bangkok: A capture–recapture method. Am J Public Health 1994; 84: 1094–1099.
10. Bishop YMM, Fienberg SE, Holland, PW. Discrete multivariate analysis: Theory and practice. Cambridge, MA: MIT Press, 1975; 229–256.
11. Tilling K. Capture–recapture methods – useful or misleading? Int J Epidemiol 2001; 30; 12–14.
12. Chapman CJ. Some properties of the hypergeometric distribution with applications to zoological censuses. U California Public Stat 1951; 1: 131–160.

13. Razzak JA, Luby SP. Estimating deaths and injuries due to road traffic accidents in Karachi, Pakistan, through the capture–recapture method. Int J Epidemiol 1998; 27: 866–870.
14. Grogger JT, Carson RT. Models for truncated counts. J Appl Economet 1991; 6: 225–238.
15. Dorazio RM, Royle JA. Mixture models for estimating the size of a closed population when capture rates vary among individuals. Biometrics 2003; 59: 351–364
16. Meng X-L. The EM algorithm and medical studies: A historical linik. SMMR 1997; 6: 3–23.
17. Hall WD, Ross JE, Lynskey MT, et al. How many dependent heroin users are there in Australia? MJA 2000; 173: 528–531.
18. Hay G, Smit F. Estimating the number of drug injectors from needle exchange data. Addiction Res Theory 2003; 11: 235–243.
19. Hser YI. Population estimation of illicit drug users in Los Angeles County. J Drug Issues 1993; 23: 323–334.
20. Domingo-Salvany A, Hartnoll, RL, Maguire A, Suelves, JM, Antó. Use of capture–recapture to estimate the prevalence of opiate addiction in Barcelona, Spain, 1989. Am J Epidemiol 1995; 141: 567–574.
21. Zelterman, D. Robust estimation in truncated discrete distributions with application to capture–recapture experiments. J Stat Plann Inf 1988; 18: 225–237.
22. Mao CX, Lindsay BG. A Poisson model for the coverage problem with a genomic application. Biometrika 2002; 89: 669–681.
23. Wickens TD. Quantitative methods for estimating the size of a drugusing population. J Drug Issues 1993; 23: 185–216.
24. Rossmo DK, Routledge R. Estimating the size of a criminal population. J Quant Criminol 1990; 6: 293–314.
25. Dempster AP, Laird NM, Rubin DB. Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). J R Stat Soc [B] 1977; 39: 1–38.
26. Böhning D. Computer-assisted analysis of mixtures and applications. Meta-analysis, disease mapping and others. Boca Raton: Chapman & Hall/CRC, 2000.
27. Laird NM. Nonparametric maximum likelihood estimation of a mixing distribution. J Am Stat Assoc 1978; 73: 805–811.
28. Böhning D, Dietz E, Schlattmann P. Recent developments in computer-assisted analysis of mixtures. Biometrics 1998; 54: 367–377.
29. Suppawattanabodee B. Estimating the number of drug users in Bangkok: A capture–recapture method. Master Thesis in Biostatistics, Faculty of Graduate Studies, Mahidol University, Bangkok, 2003.
30. Hook EB, Regal RR. The validity of Bernoulli census, log-linear, and truncated binomial models for correcting for underestimation in prevalence studies. Am J Epidemiol 1982; 116: 168–176.
31. Collet D. Modelling binary data. Boca Raton: Chapman & Hall/CRC, 1999.
32. Simar L. Maximum likelihood estimation of a compound Poisson process. Ann Stat 1976; 4: 1200–1209.
33. Debrock C, Preux P-M, Houinato D, et al. Estimation of the prevalence of epilepsy in the Benin region of Zinvië using the capture–recapture method. Int J Epidemiol 2000; 29: 330–335.
34. Pal DK, Das T, Sengupta, S. Comparison of key informant and survey methods for ascertainment of childhood epilepsy in West Bengal, India. Int J Epidemiol 1998; 27: 672–676.

Address for correspondence: Prof Dr Dankmar Böhning, Division of International Health, Institute of Social Medicine, Epidemiology, and Health Economy, Charité University Medical Center Berlin, Fabeckstr. 60–62, Haus 562, 14195 Berlin, Germany
E-mail: boehning@zedat.fu-berlin.de