



UNIVERSITY
of
GLASGOW

Methodological Guidelines to Estimate the Prevalence of Problem Drug Use on the Local Level

CT.97.EP.05

Centre for Drug Misuse Research:

Gordon Hay
Neil McKeganey

EMCDDA:

Lucas Wiessing
Richard Hartnoll

Other Contributors:

Pierre-Yves Bello
Daniela D'Ippoliti
Antònia Domingo-Salvany
Martin Frischer
Ludwig Kraus
Filip Smit

Please use the following citation:

European Monitoring Centre for Drugs and Drug Addiction (EMCDDA).
Methodological Guidelines to Estimate the Prevalence of Problem Drug Use
on the Local Level. Lisbon: EMCDDA, December 1999.

Since the initial production of these guidelines, the preferred age ranges used
by the EMCDDA have changed to 15 - 24, 25 - 34 and 35 - 64 (or 15 - 64).
This report has been updated in the light of these changes, December 1999.

Contact Details

Centre for Drug Misuse Research, University of Glasgow
Glasgow, G12 8QG
United Kingdom.

European Monitoring Centre for Drugs and Drug Addiction
Rua Cruz de Santa Apolónia 23/25
1100, Lisboa
Portugal.

Further copies of this report can be obtained from the EMCDDA at the above
address.

Acknowledgements

This publication is based on an EMCDDA-commissioned project which examined methods for estimating the prevalence of problem drug use at the local level. During this and a previous project, the authors relied on the support of researchers throughout the European Union and within the EMCDDA. It would not have been possible to conduct local prevalence estimation projects without the support of various agencies and bodies which supplied data on drug use in the participating cities. The authors would like to acknowledge the following individuals who contributed to these projects:

Catherine Comiskey, Sofia Friere, Maria Moreira, Alvar Norén, Päivi Partanen, Kathy Politikou, Dan Seidler, Jaap Toet, Alfred Uhl and Ari Virtanen.

Contents

	Page No
1	Introduction 1
1.1	Why Estimate Prevalence? 1
1.2	Why This Publication?..... 2
2	Drug Misuse Prevalence Estimation 4
2.1	Case Definitions..... 4
2.2	Epidemiological Definitions 8
2.3	How to Estimate Prevalence 8
3	Methodological Guidelines 16
3.1	Identifying Data Sources..... 16
3.2	Negotiating Access to Data Sources 22
3.3	Data Collection 23
3.4	Data Collation 28
3.5	Data Analysis 31
3.6	Dissemination of Results 32
3.7	Summary 34
4	Statistical Analysis 35
4.1	Capture-recapture Data 35
4.2	Log-Linear Models 36
4.3	Model Selection 37
4.4	Confidence Intervals 38
4.5	Analyses using GLIM 39
4.6	Analyses using SPSS 44
5	Discussion and Conclusions..... 48
5.1	Methodological Issues 48
5.2	Refinements 50
5.3	Conclusions..... 51
6	References 53

APPENDIX I..... Analyses using GLIM

APPENDIX II..... Analyses using SPSS

1 Introduction

‘How many people use illicit drugs?’ is a question which is deceptively easy to ask but notoriously difficult to answer. To examine the tools which were available to answer this question a scientific seminar was organised by the European Monitoring Centre for Drugs and Drug Addiction (EMCDDA) and the Pompidou Group of the Council of Europe. This seminar came to be known as the Strasbourg Seminar and a collection of the presented papers formed the basis of a Scientific Monograph. (EMCDDA, 1997*a*). Methodological advances have been made since then, including an EMCDDA-funded comparative pilot study in which estimates of the prevalence of opiate use were obtained in seven European Union cities using the capture-recapture method (EMCDDA, 1997*b*). That study brought together a range of experts from throughout the European Union, and the experienced gained from undertaking prevalence studies in these, and other, locations has now been channelled into producing methodological guidelines.

1.1 Why Estimate Prevalence?

It is difficult to overestimate the importance of obtaining accurate information on the prevalence of illicit drug use. Such information is valuable both in terms of monitoring the impact of drug misuse at both national and local levels as well as in assessing the effectiveness of drug prevention efforts. Drug misuse, and the policy response to it, transcends national boundaries and therefore the EMCDDA needs to inform policy makers across Europe with information on the nature and extent of drug use in different settings. This can only be done using comparable methods across Europe, using similar definitions and applying the same methodology. It is only then that meaningful comparisons can be made between cities within a country, and also across countries.

There is a triangular nature to information on the nature and extent of drug misuse at the local level; information is needed to examine how well existing drug treatment agencies are providing services for their clients and how the needs of the unknown number of drug users can be met. Despite the importance of such information, data on drug misuse is often fragmented, ranging from local studies of problematic drug use, such as heroin injecting, through to national surveys providing information on levels of cannabis use, particularly by younger people. While such surveys have provided valuable information on the prevalence of certain kinds of drug misuse e.g. cannabis, ecstasy and hallucinogens, they are less well suited to estimating prevalence at the more problematic end of the spectrum, particularly at the local level.

The use of ‘heavier’ drugs such as heroin is a hidden and stigmatised activity, the details of which many individuals are unwilling to reveal to others, particularly those whom they perceive as representing official authority. Drug

use is also an activity which for the most part occurs outside the domains of the health care and criminal justice systems. It should therefore be remembered that, no matter what approach to prevalence estimation is taken, the picture produced by this process can only ever be an imperfect approximation of the real state of affairs. As a result, our knowledge of the world of illicit drug use and our ability to estimate the number of people using illicit drugs within a locality is less complete than we may judge to be desirable.

1.2 Why this publication?

In this publication we describe methods for estimating the prevalence of drug misuse at the local level, in particular the capture-recapture method. We begin in Section 2 by providing a summary of the prevalence estimation techniques that can be used, ranging from systematically counting drug users through to applying more advanced statistical techniques. We then describe the process that is undertaken during a capture-recapture study in Section 3 and present a series of guidelines which will be useful for those intending to carry out a capture-recapture study. At this stage we only touch on the more technical aspects of the methodology, leaving a more statistical description to Section 4. This section may only be of interest to those with a statistical background and it includes annotated output from two statistical packages that can be used to analyse capture-recapture data. A related publication provides a scientific review of the literature on drug misuse prevalence estimation methods at the local level.

Methodological guidelines, such as those presented within this publication, are necessary for several reasons. The range of expertise across Europe is broad; in some areas there are experienced research groups which, in addition to applying the methods in a range of settings, are contributing to the development of the methods. In others areas there are those with an interest in how estimates of the prevalence of drug misuse can be obtained but may not have the experience. All across Europe however, there will be policy makers and professionals who require accurate information on the nature and extent of drug misuse in their area and may be considering initiating prevalence research.

Much of what is presented within this publication arises from two consecutive EMCDDA-funded projects. The first was a pilot study in which the prevalence of opiate use was estimated in seven European cities; Dublin, Helsinki, Rome, Rotterdam, Setúbal, Toulouse and Vienna. The second project sought to make the experience gained within this methodological pilot study more accessible and understandable to those interested in carrying out a drug misuse prevalence estimation study in their local area. The three main objectives of that project were to produce a set of methodological guidelines of the use of capture-recapture methods, to produce a review of the related scientific literature, and to provide support to new and existing local prevalence estimation studies. The results of the pilot project have previously been

presented (EMCDDA, 1997*b*), however this publication provides a more in-depth discussion of local prevalence estimation.

It would be convenient if this publication could give precise answers to all the questions that may arise during a prevalence estimation project. Unfortunately, there are seldom clear answers in the field of drug misuse epidemiology. These methodological guidelines will however give the reader an insight into drug misuse prevalence estimation at the local level and will lead the reader through the process of obtaining prevalence estimates using capture-recapture methods. Some questions will remain unanswered, therefore it may be necessary to forge links with research groups who have experience in applying the method, or at least with statisticians versed in similar methods. In that sense, prevalence estimation methods are not a 'black box' which readily produce valid estimates.

It is hoped that this publication will appeal to a wide audience, ranging from policy makers interested in initiating a local prevalence study to those who are responsible for collecting, collating and analysing the data. Clearly this publication cannot comprehensively describe all aspects of epidemiological statistical analysis, however it is structured such that an understanding of the relevant issues can be gained by the non-expert.

2 Drug Misuse Prevalence Estimation

In this section we describe the different methods that can be used to estimate the prevalence of illicit drug use at the local level. Although we use the term ‘drug misuse prevalence estimate’ to encompass a range of tools for estimating the size of different drug using populations, we begin by exploring the various definitions, the drugs that are used throughout Europe and some of the related terminology.

2.1 Case Definitions

There are a range of terms that are applied, often interchangeably, to describe the illicit use of drugs. Some terms are medical in nature, for example drug injecting or drug dependence. It is clear that someone who injects drug can be classified as a drug injector and medical opinion can define what dependence is, for example the International Classification of Diseases (ICD10) codes refer to dependence syndrome as:

‘A cluster of behavioural, cognitive and physiological phenomena that develop after repeated substance use and that typically include a strong desire to take the drug, difficulties in controlling its use, persisting in its use despite harmful consequences, a higher priority given to drug use than other activities and obligations, increased tolerance and sometimes a physical withdrawal state. The dependence syndrome may be present for a specific psychoactive substance, for a class of substances or for a wider range of pharmacologically different psychoactive substances’

The term ‘drug addiction’ is often used interchangeably with ‘drug dependence’. In Spain, the SEIT information system additionally defines ‘drug abuse’ as

‘A non-adaptive model of psychoactive substance use, with continuous or recurrent use, which does not meet the criteria for the diagnosis of dependence of said substance’

Other terms can be more subjective, for example ‘problem drug user’. The Regional Drug Misuse Databases within the United Kingdom use the following definition of a problem drug user:

‘Any person who experiences social, psychological, physical or legal problems related to intoxication and / or regular excessive consumption and / or dependence as a consequence of his / her own use of drugs or chemical substances.’

This wider definition includes legal problems, albeit only in relation to intoxication, dependence or excessive use. Laws on illicit drug use vary across Europe relating to the consumption, possession or supply / trafficking of drugs, therefore the concept of a legal problem relating to drugs will vary across the continent. Indeed, the inclusion of supply, or even possession, within a 'legal problem' classification could encompass those that do not consume drugs, but only supply them to others.

Definitions such as problem drug use are often intrinsically linked to the nature of the drugs that are being consumed. We therefore detail some of the drugs that are used throughout Europe.

Heroin

The use of heroin appears to be the most problematic drug use in the majority of European settings. This opiate can be consumed in several ways, including heating the drug on foil and inhaling the fumes; 'chasing the dragon', or by dissolving the drug and injecting it. The most popular route of choice of consuming the drug may vary within and between countries. Tolerance and physical dependence can result from regular use of the drug, and withdrawal can be difficult. In some areas, the vast majority of drug users contacting services are heroin users, the most common form of treatment is often substitute prescribing with Methadone.

Methadone, other opioids and benzodiazepines

Methadone is a synthetic opiate (opioid) which can either be in tablet or liquid form. As it can be used as a substitute for heroin, it can be legally possessed with a prescription. Questions have arisen about the efficacy of substitute prescribing, with some people continuing to use heroin when they are using methadone. Prescribed methadone can also 'leak' onto the street and be used illicitly. Other opioids, manufactured for medical use, include dihydrocodeine (DF 118s), pethidine, Diconal, Palfium and Temgesic. Because of the severity of addiction to heroin, some users will resort to using any opiate when heroin is not available or they may use benzodiazepines such as diazepam or temazepam. These tranquillisers are often prescribed to treat anxiety and as sleeping tablets. They can be equally addictive, even when being prescribed medicinally.

Amphetamines and Cocaine

Amphetamines are synthetic stimulants which can be snorted or injected. In Nordic countries, particularly Sweden and Finland, the problematic use of amphetamines is more prevalent than that of heroin. Amphetamine use in some countries can be paradoxical; it may be one of the most popular drugs that are injected, but it can also be one of the most popular drugs used by

younger people within dance events with national surveys suggesting that up to 16% of younger adults have used amphetamines. Indeed amphetamine injectors may perceive themselves to be different to opiate injectors and therefore be less likely to attend treatment services or needle exchanges. Cocaine is also a stimulant that can be snorted, however it is more expensive than amphetamines and is therefore associated with a rich lifestyle. Crack is a smokeable form of cocaine, and has become a major problem in North America. The use of crack has not yet been seen to be as widespread in Europe.

Other synthetic drugs and cannabis

There are a collection of drugs, including MDMA (ecstasy), LSD and newer drugs such as ketamine which are sometimes described as ‘recreational’ drugs or hallucinogens. The use of these drugs is certainly not without problems however their use is rarely perceived to be as problematic as opiates or amphetamines. The use of cannabis also is also rarely seen as problematic. The legal response to these drugs varies throughout Europe, in particular there is a distinction between ‘soft’ and ‘hard’ drugs in the Netherlands.

This summary of the various drugs that are used within Europe has been brief, however necessary before we discuss prevalence estimation. The question ‘how many people use drugs’ needs to be refined and a case definition established before conducting a prevalence study. In many instances, the use of drugs such as cannabis or ecstasy will fall outside the remit of a prevalence study, partly due to the less serious nature of the use of these drugs but also because of the availability of information about the prevalence of use from other sources, such as national surveys. Developing a case definition, in relation to the use of more problematic drugs, may be more difficult.

There has to be a balance, in terms of the drugs included in a case definition, between the need to know how many people are users and the availability of information. While it may be useful to know precisely how many people are drug injectors, perhaps in relation to the spread of HIV or hepatitis through shared injecting equipment, it may not be possible from the available data to identify which drug users in contact with services or detained by the police actually inject. Indeed, the injecting of steroids may possibly be within the remit of some prevalence studies as steroid injectors can be at risk of blood borne diseases.

Sometimes this balance may be more of a compromise, particularly with a case definition such as problem drug use. For example, although the injecting of amphetamines may be as worrying as the injecting of heroin, the other role of amphetamine as a dance drug may mean that information on those detained by the police for possession of amphetamine may not always be useful. Even a wider definition, such as the use of opiates, cocaine or amphetamines, may cause difficulties within a prevalence estimation study due to the social and

behavioural variability of those that use amphetamines.

It is sometimes methodologically easier to restrict the analysis to a homogeneous group of drug users, such as opiate users or opiate / benzodiazepine users. Members of this group could be perceived as the stereotypical 'junkie' – unemployed, drifting in and out of treatment and in trouble with the police. While stereotyping people in this manner can be both offensive and unconstructive, it removes the problem of having to distinguish between problematic and non-problematic use of drugs such as amphetamines.

Whatever case definition is used within a local prevalence study has to depend on both the nature of drug use in that locality and the availability of data. Drug misuse is a transient activity and it should be recognised that an individual's drug use may vary within a short time period. Many drug users will use more than one drug, particularly when the availability of their preferred drug is limited. Although it is easier to apply the methodology using some case definitions, it is, however, not until the analysis stage that a particular case definition is found to be the most amenable to analyse. One remedy would be to collect as much data as possible in as detailed a form as possible, perhaps by collating information on all non-cannabis offenders and all drug users contacting services and noting the drugs that these individuals use. At a later stage can those who use certain type of drugs be extracted and analysed. The decision as to which data should be collected can only be taken at the local level and will additionally depend on the resources available to collect data and the format it comes in.

There are different terms used to describe the illicit use of drugs

- **addiction**
- **problem drug use**
- **drug dependence**
- **drug use / drug injecting**

Many drugs can be used illicitly

- **heroin**
- **methadone**
- **amphetamines**
- **recreational drugs**
- **cannabis**

There should be clarity in what a prevalence estimate refers to

2.2 Epidemiological Definitions

Before going on to describe the different methods of obtaining information on the prevalence of substance misuse, it is worthwhile to discuss some of the

epidemiological concepts used in prevalence estimation,

The prevalence of a certain social attribute is defined as the proportion of people possessing that attribute. It is often expressed as a percentage, or sometimes as 'per thousand' or even 'per million' of the total population. The actual number of individuals is sometimes used instead of the prevalence, however without information on the baseline population, this number may be meaningless. If there are an estimated 1,000 drug users in one city and 2,000 in another city, without knowledge of the relative sizes of these two cities, the two figures cannot be directly compared.

When discussing percentages, care must also be taken in the different interpretation of the baseline population. Drugs are more commonly used by people in the 15 to 34 age group. That is not to say that children under 15 or people aged 35 or over do not use drugs, but when calculating a percentage, the age range of the baseline population is often one which includes those that are more likely to use drugs; using the total population as a baseline would therefore result in a lower prevalence value. This is often the case when a sample of people are asked about their drug use. If the sample are predominantly in the 15-34 age group, then the proportion currently using drugs would probably be greater than that in the 15-64 age group. When studies have been carried out in more than one location, differences in the age structure of the general population could distort comparisons.

The concept of prevalence also requires some indication of the time period that is being examined. Drug misuse can be a transient activity such that someone using drugs one month may not be using them the next, therefore the concept of current use may refer in different contexts to use within the past week, past month or past year. In addition, lifetime usage can be described, therefore care must be taken in comparing lifetime usage between different age groups, given that older people may have had a longer time to have used drugs.

2.3 How to estimate prevalence

Estimating the prevalence of drug use, particularly problem drug use such as opiate use or opiate injecting, is not always an easy task. What begins as a vast array of confusing, sometimes conflicting information can apparently be turned into a reliable estimate, often with a confidence interval giving added statistical credence. However, every estimate produced is subject to caveats or assumptions. If these assumptions are not valid, then the estimate may be inaccurate. The possibility of inaccuracies in prevalence estimates is often more obvious in more straightforward methods, for example when a researcher stands on a street corner with a clipboard, asking passers by if they take drugs. It is less obvious however, that an advanced statistical technique, such as the capture-recapture method, can produce erroneous estimates due to violation of some of the assumptions.

We begin by describing the use of surveys in prevalence estimation. We then discuss enumeration, which is a direct method of estimating the prevalence of drug use in which the number of known drug users is obtained by combining data from various sources and eliminating the double counting caused by overlaps between data sources. An indirect method known as capture-recapture can use these overlap data, along with the number of known drug users to estimate the size of the unknown drug using population. Other indirect methods which can be considered include network analysis techniques such as snowball sampling or multiplier techniques.

Surveys

General population surveys which provide information on the use of drugs at the national level have been carried out in several countries of the European Union, including Belgium, Denmark, Germany, Spain, France, Finland, Sweden and the United Kingdom. These population surveys can either be specific surveys on drug use, or more general health related or crime surveys. They can assess the nature and extent of drug misuse and often provide information on respondents' recent or lifetime drug use.

There is the methodological problem that general population surveys are less likely to include harder to contact people who may be more likely to be problem drug users, in particular sub-populations such as the homeless or those living in institutions. People may be reluctant to divulge information on matters which are deemed to be socially unacceptable, therefore general population surveys may be more appropriate in assessing the prevalence of recreational drug use, but not for problematic use. In addition very few people use drugs such as heroin therefore detecting the use of such drugs within a general population survey is quite difficult. There are, in addition, the problems relating to people wrongly filling in the questionnaires such that one or two incorrectly completed questionnaires may, in small surveys, exaggerate the use of certain drugs and in some cases the sample sizes needed to obtain worthwhile results often make the costs of sampling prohibitive.

The scope of a general population survey may mean that local variations in levels of drug use are not usually identified. Thus a survey may be useful in evaluating a national health promotion campaign about the use of drugs such as ecstasy, but will be of little use in deciding where to situate a needle exchange.

There can sometimes be benefits of using surveys. For example information on the use of drugs by younger people can be gathered by surveys of schoolchildren. National school surveys are more common within the European Union. It is often easier to get a representative sample, and repeated surveys or surveys of different age groups can describe trends in drug use.

Although general population surveys can suffer difficulties in reaching a representative sample, surveys of specific populations can be undertaken. These can either examine groups of people that may be more likely to be drug users, for example those detained for crimes such as theft, those in prison or those that attend emergency rooms. Such surveys will not, in themselves, provide estimates of the level of drug use in the more general population, but may provide data for other prevalence estimation methods. Specific surveys of drug users can also be undertaken, either to gain behavioural information such as route of administration, injecting or needle sharing, or medical information such as the prevalence of HIV or hepatitis.

Enumeration

There are a range of data sources which can be thought of as indicating the number of drug users in a particular locality. These include drug treatment agencies, methadone prescribers, needle exchanges, HIV treatment agencies, general practitioners, emergency rooms, mobile emergency units, hospitals, the police and prisons.

To use these sources within a prevalence study, it is necessary to obtain enough information from each data source to identify each individual drug user and thus to eliminate double counting. This requirement has to be weighed against the confidentiality requirements of each source. Hence initials, sex and date of birth are often used to sift out multiple occurrences. It should, however, be noted that collating information from data sources using this un-named identifier information can be imprecise, especially as the accurate collection of names and dates of birth may not be the highest priority for some agencies working with drug users. This exercise can be useful in describing the known drug using population.

Data from the sources listed above (usually with the exception of the police and prisons) are often required to collate registers, particularly those relating to drug users in treatment, for example the SEIT reporting system in Spain, the IVV/LADIS reporting system in The Netherlands, the RELIS/LINDDA reporting system in Luxembourg and the Regional Drug Misuse Databases in most of the United Kingdom. The police also collate information in many countries of the European Union, however just as any increase or decrease in the number of people entering treatment may be partly due to the increased availability of treatment or its effectiveness, any increase or decrease in the number of drug-related arrests may only reflect operational decisions of the Police. Even if these registers were highly correlated with the prevalence of drug use, by their definition they can only be taken to be describing the numbers of drug users in treatment or the level of detected drug-related crime. Clearly the spectrum of drug using behaviour is not restricted to those who are in treatment or are have been detained by the police.

Case-finding studies are a more systematic form of multi-source enumeration

in which agencies, such as those detailed above, contribute identifier data on drug users. The identifier data can again be used to sift out double counting and as total coverage of all possible sources is aimed for, the total number of known drug users can then be calculated. In many areas of Europe this total coverage is not feasible and other problems such as difficulties in identifying double counting serve to make this method impractical. Sweden, however, has a long tradition of case-finding (Olson, 1997).

Information on the nature and extent of the unknown population of drug users is also required to give the complete picture of drug use, and in particular the size of the unknown population of drug users is needed to obtain an estimate of the total number of drug users. There are several different ways of doing this, as described below.

Capture-recapture and multiplier techniques

A methodology, commonly referred to as capture-recapture, has increasingly been used to estimate the prevalence of drug misuse. As the name alludes to, this methodology was originally developed by ecologists who were interested in estimating the size of animal populations. Two analogous examples are presented here; one from ecology, and one from an early application of capture-recapture in estimating the prevalence of drug misuse.

An ecologist wants to estimate the number of fish there are in a lake: therefore a sample of fish are caught, counted, marked in some way and then released. We show this in Figure 2.1, where a sample of 15 fish have been caught and marked. At a later date, the ecologist returns to catch another sample, and by checking for marks (as demonstrated in Figure 2.2), the number of fish seen in both samples is discovered and thus the ratio of previously caught to previously uncaught fish in the second sample can be found. As it can perhaps be assumed that the ratio of caught to uncaught in the first sample is the same as the ratio of previously caught to previously uncaught in the second sample, the total population size can be estimated by multiplying the number seen in the first sample by the inverse of the ratio.

Figure 2.1

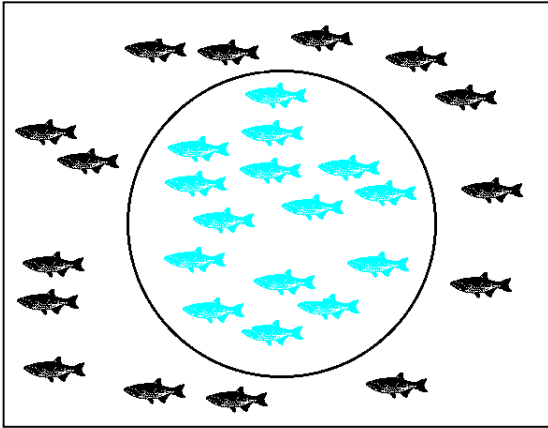
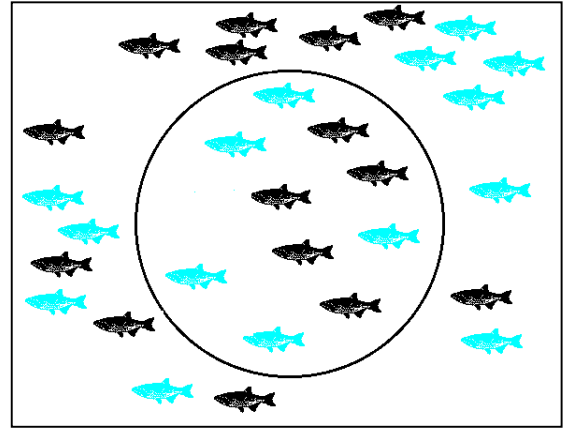


Figure 2.2



In the example shown in Figures 2.1 and 2.2, initially 15 fish were caught and marked in the first sample, and 10 were caught in the second. As five of the fish in the second sample bore marks from the first sample, we can perhaps assume the ratio found within in the second sample - $5 / 10$, is equal to the ratio if caught to uncaught fish in the first sample, therefore

$$\frac{5}{10} = \frac{15}{x}$$

Where x is the total number of fish in the lake – the value we wish to estimate. When we re-arrange this equation, we can show that

$$x = 10 \times \frac{15}{5} = 30$$

Therefore there are 30 fish in the lake. An example which seeks to estimate the size of a drug using population may further explain the method.

Hartnoll *et al.* (1985) applied the capture-recapture methodology when estimating the prevalence of opioid use in an area of London. They collected data concerning opioid users who had attended a drug clinic and those that had been admitted to a hospital for infectious diseases because of their drug use. By comparing these sources of data they found that 20 per cent, or a fifth, of the hospital sample had also attended the drug clinic. Thus the total number of opioid users could then be estimated to be five times the number who attended the drug clinic. Thus the size of the hidden population of drug users was estimated by merging two existing sources of data and examining the overlap between them.

These simple examples mask some of the problems of the capture-recapture

methodology. In Hartnoll’s case, if those who are attending in the clinic were more likely to have been admitted to the hospital then the resultant figure would be an overestimate. Thus, if there is some kind of relationship between data sources the estimate could be inaccurate. Unfortunately it is often unclear if such relationships are present and therefore the validity of estimates obtained when examining two data sources are often questionable.

The capture-recapture methodology can compensate for this problem by employing three or more sources. The extra information present in the third sample can be used to examine whether or not there are any relationships between data sources, and if they are, the estimate of the total population size can be adjusted accordingly. Frischer (1992) used three sources of data which held information on drug injectors in the City of Glasgow in 1989. These were combined data from treatment agencies, an HIV test register and the Police. In total, information was gathered on 1,738 individuals, and the overlap between the data sources can be described by Table 2.1.

Table 2.1. Presence or Absence from Three Data Sources of 1,738 Drug Injectors. *Source Frischer (1992).*

		Police			
		Present		Absent	
		Treatment Agencies			
		Present	Absent	Present	Absent
HIV	Present	5	15	109	389
	Absent	23	366	831	*

These data can be analysed using statistical packages. In short, the 7 pieces of information in this table can be used to predict the missing value, which would be an estimate of the number of drug users not present in any of the three data sources, or the unknown drug using population. This can be done by log-linear analysis. Different relationships between the data sources can be described using this analysis, for example if it was thought that those drug injectors attending treatment agencies were more likely to have been tested for HIV, then this relationship can be included. The decision to include any relationships between sources can be taken by examining how similar the observed overlap pattern is to what would be expected if such relationships were actually present. Thus different models can be fitted to the observed data and a preferred model would be one that closely fits the observed data.

This methodology has been applied successfully in a range of European settings, including Glasgow, Liverpool and Dundee in the United Kingdom (Frischer *et al.*, 1993; Squires *et al.*, 1995; Hay and McKeganey, 1995), Barcelona, Spain (Domingo-Salvany *et al.*, 1995) and Toulouse, France (Bello and Chêne, 1997). The capture-recapture methodology is more comprehensively discussed in the next sections of this publication, while a description of the historical development and applications of this, and other

methods, is to be found in a related publication.

In a similar manner to the two sample capture-recapture method, the prevalence of drug misuse can be estimated by applying a multiplication factor to other indicators, in particular data on drug-related deaths. In-depth studies of drug use and mortality suggest that 1-2 per cent of drug injectors die per annum (Ghodse et al, 1985). Thus taking the upper level, an estimate of the number of drug injectors can be found by multiplying the number of deaths in the injecting population by 50. The inaccuracy of these estimates reflects that the multiplication factors will vary between different areas and between different times, leading some to suggest that these factors are little more than a guess.

Other methods

The capture-recapture method and the multiplier method use available data in estimating the prevalence of drug misuse. In some instances, the available data can be augmented by data gathered during fieldwork studies. Various terms are used to describe these methods, including nomination techniques or network analysis. The term 'snowball sampling' describes a method of obtaining the additional data, in this case by asking drug users to identify others that may be interviewed (Korf, 1997). In essence, these techniques are similar to multiplication methods, with one source being an existing source such as drug users in treatment. The other source would be the data collected during field work. Again these methods again rely on assumptions whose validity cannot readily be tested, and they may be more time consuming and expensive to undertake as opposed to other forms of prevalence estimation.

There are other advanced statistical techniques that can perhaps be used to estimate the prevalence of drug use at the local level, and these are discussed in a related publication. Many of them have not yet been applied in Europe, and most would require similar data to those used within a capture-recapture study.

Summary

Various methods for estimating the prevalence of drug use have been described above. It is important to note that there is not one 'best' method for obtaining information on the extent of drug use. The preferred method will depend on what way the information will be used. For example, a national general population survey would not be useful in determining the prevalence of drug injecting within a city. This information will be required by service providers at a local level in determining the need for control strategies such as needle exchanges or the provision of substitute prescribing. In contrast, a national advertising campaign highlighting the problems of ecstasy use will rely more heavily on information obtained from national surveys.

It is clear that questions relating to the prevalence of drug use cannot simply be answered by examining only a few sources of data or relying too heavily on one research method. Indeed, information on drug use has been compared to a jigsaw with several pieces missing. We have, however, described the pieces of the jigsaw that are often available and describe some of the methods that can be used to piece together the jigsaw and thus provide the required information on the prevalence of drug use.

There are a range of methods that can be used to estimate the prevalence of drug misuse

- surveys
- enumeration
- multipliers
- capture-recapture methods
- network analysis

Different methods are better used to examine different types of drug misuse

There may be no 'best' method, rather a combined approach

3 Methodological Guidelines

In this section we detail the process that is undertaken when conducting a drug misuse prevalence study at the local level. We begin by discussing the sources of data on drug misuse, then describe how such data can be collected and collated in a form which enables a description of the nature and extent of known drug use. The result of that stage of the process is a multi-source enumeration. We then briefly describe how this information can then be used to provide an estimate of the size of the hidden population of drug users using the capture-recapture methodology. Because the application of the methodology requires some understanding of statistical techniques and experience in the use of statistical packages, we leave some of the more statistical aspects of the methodology to the following section. It is therefore hoped that this section will be understandable to those without a statistical background.

There are several distinct stages in carrying out a prevalence estimation study and the skills needed to complete each stage are varied, ranging from the concise recording of identifier and other information within a manual data collection exercise, the manipulation of data using a spreadsheet, through to advanced statistical analysis. We shall now detail each stage of this process.

3.1 Identifying data sources

Despite the illicit use of drugs being a covert activity, it remains a legal, social and medical problem and results in drug users coming into contact with a range of agencies and services, including drug treatment agencies and the police. These agencies, to a varying extent, collate information on drug users that they have been in contact with. Sometimes this is a legal requirement, for example there are registers of drug users in treatment in many countries. In other instances the data collated on clients depends on the operational policies of the agencies. The reasons why and individual will appear within a data source will also vary between sources; it can be argued that those drug users in contact with drug treatment services are a self-selecting group, whereas those that have been arrested for drug offences may be more representative of the total drug using population.

Clearly, since prevalence estimation methods rely on obtaining accurate and relevant information from various data sources, a description of the various sources of information on drug misuse is important. We therefore describe the data sources that have been used in previous studies in Europe. It should be remembered that this is not an exhaustive list, rather a suggestion as to where a drug misuse prevalence project should begin to identify possible data sources. In the same vein, just because data from a particular kind of agency was easily accessible in a previous project, it may neither be possible nor practical to collect such information in every instance. Indeed, issues arise as to why agencies should co-operate within a drug prevalence study, bearing in mind the

confidentiality that they offer to their clients.

Just as many prevalence estimation methods rely on a minimum standard of data in terms of accuracy or relevance, there is often a minimum amount of data needed to provide a worthwhile prevalence estimate. Unfortunately, there are no definitive rules about the amount of data required by a prevalence study, this will depend in part on the nature of the sources and any relationships between them. Where possible, as many sources of data should be employed, collating as much data as available. This advice, however has to be balanced with the financial implications of collecting the data.

Drug treatment agencies

Counselling and support of those who have problems relating to drug misuse is available from a diverse range of agencies throughout Europe and the range of services offered by such agencies is equally diverse. Many drug treatment agencies will collate records on clients in similar manner to the way a hospital or a general practitioner would complete medical records. Often within these records, an assessment of a client's drug use will have been made, particularly when they have first contacted the service or returned to the service after a period of non-contact. This potentially could be the only information on a client's drug use that is available, and as such could be flawed. Only an ardent optimist would assume that a drug treatment service is so efficient that all of those presenting with a severe problem, for example heroin injecting, will rapidly alter the nature of their drug use. However, only a pessimist would assume that a drug treatment service is so inefficient that their clients' drug use would not change during their time in contact with the service.

The Surveillance System on Drug Addiction in Rome gathers information on drug users attending public treatment centres. It collects socio-demographic information on patients and the treatment offered. This data source covers the whole of the city and included over 6,000 individuals. It was used within a three-sample capture recapture study in 1997.

Low threshold agencies

There may be other agencies which operate in more relaxed manner, for example drop-in centres or needle and syringe exchange schemes. Such services often view drug misuse, and therefore the treatment of drug misuse as a social rather than a medical problem, and thus could be attracting a more representative group of drug misusers. These agencies may collate the same standard of information on their clients as the more formal drug treatment agencies described above, although in some instances they are more relaxed and some clients may only be known by a forename or an assumed name.

Intermède, in Toulouse, is a low threshold agency which opens each afternoon. Drug users can drink coffee, have a shower or use a laundry, while having access to a needle exchange and to social workers. The data collated by this agency was combined with other similar agencies within a three-sample capture recapture study in 1997.

Methadone prescribing services

In some areas the prescribing of methadone as a substitute to heroin will be organised separately from drug treatment services. As methadone is also a opioid and can be highly addictive, data concerning individuals being prescribed methadone are usually of a high quality. Even where more than one agency (including treatment agencies or general practitioners) are able to prescribe methadone, a central register may be in existence to prevent an individual obtaining methadone from more than one source. The nature of these data, in particular the serial nature of repeat prescribing, would reduce the problem of only being able to assume that such individuals are opiate users.

An additional point to note about substitute prescribing, particularly methadone prescribing, is that the legal problem which some users had previously faced should have been removed. This may not be a sufficient justification for removing methadone from the remit of a drug misuse prevalence study particularly because of the illicit use of methadone by those without prescriptions and the continuing use of other opiates by those in possession of a prescription. It can, however, cause difficulties in analysing data as some methadone users should be less likely to be in contact with the police.

The Central Patient Methadone Treatment List, in Dublin, is maintained by the Department of Health. It records those who receive methadone prescriptions from either a specialised clinic or a general practitioner. From this list it is possible to determine how many times, when and for how long each individual was on methadone. The data on over 3,000 individuals was used within a prevalence estimation study in 1997.

These last three headings have been a convenient way to summarise various types of drug treatment agencies that may exist in a city. In some areas, these three types of service may be provided by the one agency, in other areas similar services are provided by a range of agencies. While some of these agencies may only cover specific geographical areas of a city there may, in other instances, be duplication in the provision of service. Within a prevalence project it therefore may be necessary to combine data from different sources

into a single 'treatment' source. The decision to do this may be taken at the analysis stage of the project, therefore data from distinct agencies should be recorded as such during the data collection process.

Other Medical Services

Drug users, particularly drug injectors, are often are subject to higher rates of morbidity and mortality. Some problems can be dealt with at a drug treatment agency, or advice given that reduces the contact with other services that a drug user may have, such as information on safer injecting to reduce the risk of abscesses. Other problems may result in drug users contacting medical services such as emergency rooms, HIV / hepatitis related services, clinical psychologists, psychiatrists as well as their General practitioner.

One problem that exists in obtaining information from such sources, is that drug users are not the only clients such agencies or professional deal with. There may however be existing data collection processes that can be accessed, for example a psychiatric hospital admissions recording system which routinely records any drug-related admission. General practitioners may collate information on drug misusing patients, particularly if they are involved in substitute prescribing or are require to contribute to national monitoring system.

This problem can be addressed by setting up a data collection system, either prospectively or retrospectively. Either approach will have time and resource implications; a system in which the receptionist at an emergency room completes a brief form for each patient noting their drug use over a six month period would provide useful information within a drug misuse prevalence study. This approach would, however, extend the length of time the project takes to complete and would need a formal agreement for the emergency room to devote valuable time to fill in the forms, perhaps including a financial arrangement. To retrospectively examine the records of an agency such as an emergency room, could also be time consuming. This approach may be favourable as a dedicated data collector would soon become adept a sifting through records, perhaps with the aid of a formal screening instrument. What may initially appear as an insurmountable stack of records can be systematically worked through in a reasonable time period. As with drug treatment agencies, issues of confidentiality would have to be resolved prior to the data collection process.

Data sources on HIV or hepatitis may routinely record information such as the reason why someone has been tested or behavioural data such as drugs used or frequency of sharing injecting equipment. Although these data may be comprehensive and easy to collect, it should be remembered that the risk behaviour that led to an individual becoming infected may have ceased some time previous.

The following sources have been used within capture-recapture studies in different locations:

- Vienna – an Emergency Ambulance which deals with ‘acute opiate intoxication’
- Rome – a Mobile Emergency Unit which deals with overdose or withdrawals
- Helsinki – a Hospital Patient Discharge Register

Police

The legal response to drug misuse varies throughout Europe, with laws relating to the consumption, possession or supply of illegal drugs. The police within a city may collate these data themselves either for their own use or for national statistics or they may allow access to crime reports to enable someone to retrospectively collate such data. It should be remembered that there may be a difference between being detained in connection with the use of drugs and being convicted for a crime. Where available, information on detainees would be more useful and any delays in the legal system should not influence the data in a prevalence study.

Just because someone has been found in possession of a drug, it does not necessarily mean they use that drug. This point may be more pertinent when considering people detained for supplying drugs. Similarly, just because someone is detained in connection with one drug, e.g. cannabis, it does not mean they do not use more problematic drugs such as heroin. It should also be recognised that strategic decisions by the police may influence the chances that a drug user is detained for drug offences. Policies may vary within and between cities. Therefore although it may be convenient to assume that a police sample would be more representative than a treatment sample, it may not be.

Clearly offences relating to the use, possession or supply of drugs are not the only reasons why drug users come into contact with the police. To achieve a high enough income to support a serious drug problem, many users have to commit crimes, and there often exists a relationship between drug use and crimes such as theft or shoplifting. In many cases, information on drug users that commit these crimes is not collated by the police, although increasingly the link between crime and drugs is being examined, one example being the use of drugs while driving.

An individual’s drug use may be noted at other points in the legal system, for example when the punishment for the crime is determined. There may be a

system in which a report is completed on convicted individuals. This could be to establish whether or not a non-custodial sentence is appropriate, for example community service or attendance at a treatment agency. This function may be undertaken by various agencies, for example a probation service, a social service or even a drug treatment agency itself. If an individual does serve a prison sentence, the prison may assess whether or not they are a drug user, or offer substitute prescribing. Indeed the illicit use of drug often continues within a prison. Consideration has to be made as to whether the prison population is included within the remit of a prevalence study, although short term prisoners who are released back into the community may be of interest.

Police arrest data on possession or supply have been used in a variety of locations including Dublin, Helsinki, Toulouse and Vienna. In addition, police data in Dublin provided information on drug-related crime such as theft, and in Helsinki, information on those that had been arrested for driving under the influence of drugs has been used in a capture-recapture study.

Summary

We have described some of the sources of data on drug misuse, particularly those which may be of particular value in a drug misuse prevalence study. Other data on drug misuse may be available within a locality which may, or may not be useful, for example data from a youth service may not be useful in a project which specifically examines drug misuse in the adult population. When identifying data sources, a balance has to be made between gathering all possible information and the time or other resources available. While in some areas, case counting exercises have been undertaken which sought to count all drug users in contact with all agencies, in many other instances a compromise has to be made. Unfortunately, it will not become clear that enough relevant data have been collected until the analysis stage of a project.

There are a range of sources of information on drug misuse

- **drug treatment services**
- **methadone prescriptions**
- **police**
- **prisons**
- **emergency rooms**

These sources often provide information on different types of drug misuse

3.2 Negotiating access to data sources

A drug misuse prevalence study can only be undertaken with the co-operation of those who hold information on drug misuse. Each agency will have its own idea about the need or relevance of prevalence research, and each agency will have its own concerns about giving access to confidential data. Agencies which are not exclusively concerned with drug misuse may see requests for information on drug misuse as an additional burden which they may not be keen to take on.

They may also be more political obstacles to collecting data from some agencies. Drug misuse is a diverse phenomenon, and the social or medical response to it can be varied. Some agencies may not agree with the approach that other agencies have to drug misusers and may be apprehensive about contributing data to a prevalence project. Often this apprehension is increased because of the nature of a research project or its funding. For example a drug treatment agency may be uneasy about supplying data to a project undertaken by the local police. The independence of the researcher may result in easier access to data, particularly when the project is funded by a collection of policy makers or service providers which also fund the services which deal with drug users. It is also easier to persuade a manager of a service to give access to data when their funders have given their approval.

The main issue which agencies see as a reason for not giving access to their data is confidentiality. There could indeed be two distinct issues; the fact that a researcher may thus be free to sift through clients records and the fact that data can then be used to identify the clients of the agency at a later stage in the study. Medical records can be quite detailed, and during their contact with a service a drug user may disclose extremely personal information. The confidentiality of such information is imperative and drug treatment agencies may have reservations about allowing a researcher to have access to the records, particularly if there is a possibility that the researcher may know some of the agency's clients. The second issue can be almost entirely resolved by

only collating limited information on each client of the agency, as described below.

3.3 Data Collection

The collection of data within a prevalence study needs to be considered as important as any other stage in the project particularly as any inaccuracies within the data collection process can unduly influence the resultant estimate. For this and other reasons it would be preferable to have the same person collecting, collating and analysing the data. Clearly this is not always possible, therefore those collecting the data should have a knowledge of the methodology and how any violations of the assumptions may affect the estimate. It may therefore be better to have a small number of people collecting data. This may also help in fostering a good working relationship between the data collectors and the contributing data sources. In an attempt to obtain consistency between the information gathered by different data collectors, a screening instrument may be used. This could, for example, provide a checklist for inclusion within the study in terms of drugs used or length of drug use. In some situations, it may be easier for the data collectors to be in regular contact with the researcher who will be analysing the data, particularly as some of the problems relating to data collection may not be anticipated as they are peculiar to an individual agency.

As a drug user may be in contact with more than one agency, and therefore be included in the data from more than one source, sufficient information is needed on each individual to identify multiple occurrences. Matching records between data sources can be complex, and within the area of record linkage, it is recognised that problems exist even when several different fields of data on each individual has been collected.

To alleviate concerns about confidentiality, many studies have restricted the data collection process to the following fields:

- forename initial
- surname initial
- date of birth
- sex

These fields of data should be sufficient to identify those that are in contact with more than one source, but there are many problems that may have to be addressed.

To begin with, the possibility that errors in the recorded data exist should be acknowledged, for example dates of birth can be easily recorded erroneously, perhaps by transposing the month and day of the month. In some cases, the sex of an individual may not be obvious and as full names and addresses are collated at source; this may be a problem that is peculiar to the prevalence estimation project.

The names recorded by agencies may differ under different circumstances. This can partly be due to the culture of a country, for example people having a proper name that they would be known as by the police, but maybe a middle name that they are more commonly known as at a low threshold treatment agency. Married women may interchangeably use their married or maiden name and people may be able to use an assumed name at some services. The data collection process will also be subject to missing data, for example an age recorded instead of a date of birth.

Even if all data were accurately recorded in the same manner at every contributing agency, there exists the possibility that two or more distinct individuals have the same identifier information. An obvious example would be twin brothers who have the same forename initial, however some combinations of initials may be more common than others.

There may be other ways of identifying which individuals are in contact with more than one source, for example Soundex codes which convert a surname into a code which can then be matched across agencies. In some areas, a common identifier number may be used by more than one agency, e.g. a social security number or a health service number. When a single agency undertakes more than one service, such as counselling and a needle exchange, they may themselves have a method of linking records together which can be used within a prevalence study.

In addition to the identifier information detailed above, information on an individual's drug misuse is required. As previously discussed, there may be inaccuracies inherent in these data such as someone found in possession of a drug they do not use by the Police. It may be easier to code the drugs used into a category, both to save time within the data collection process and to assist in the analyses.

A reference date for the information is useful within a prevalence study to assist in analysing the data. For example the date when someone was arrested or the date that an initial assessment at a treatment agency was undertaken. Care must be taken when an individual appears more than once in a single source. A good example would be a person who was detained for drug offences during the study period. Care must be taken in sifting out this multiple occurrence, particularly if they were detained for possession of cannabis on one occasion and possession of heroin later. Sifting out the heroin

instance may remove relevant data.

Finally, some indication of the area of residence of an individual is useful. These geographical data can be used to match individuals, however the accuracy of such data may not be sufficient. Rather, the information is useful in splitting the study area into different localities and to check that an individual actually resides in the study area. This process can often be more vague than the matching process, particularly when geographical information is not routinely recorded. Once the area has been divided up into various sub-areas then separate analyses made be performed for each of the smaller districts. An additional, more pertinent reason which will be discussed in greater detail is that prevalence estimation methods need to assume that those being studied are relatively similar, particularly in relation to their contact with services. If, for example, drug users for one part of a city were less likely to attend a service based at the other side of town, then this assumption would be violated.

Postal code data can also be used as geographical identifier information, although a system of identifiers used to speed up the delivery of mail may not be the most practical when examining drug use. Often the geographical information stored by one agency is in a different format to that of others.

In many cases, data has to be collected manually, either by pen and paper or being typed straight into a laptop computer. This can be a laborious process, not particularly due to the amount of data that is required, but more because of the logistics of extracting a large file from a filing cabinet, finding the page with the relevant information, then replacing the file. There may also be the problem of lack of space, with the person collecting the data having to seek accommodation near to filing cabinets thus getting in the way of agency staff. The work of an agency cannot be put on hold during the data collection process, and the files on some clients may be in use during the data collection period and thus cannot be accessed.

Computerised data

In many instances, the data that are required have already been computerised by the agency, although rarely in the form described above. Names need to be converted into initials, and the relevant information needs to be extracted from a larger computer file. There are many ways of doing this, depending on the computer package used to store the data within the agency and the package that will be used to collate and analyse the data by the researchers. It is often easier for the researcher to obtain a file that includes names and addresses and then delete out the names after the initials have been obtained and remove the address after postal codes or area of residence have been assigned. This is often contrary to the confidentiality requirements of the agency supplying the data and therefore staff within the agency may be required to convert their data

into a satisfactory format.

Confidentiality is a key concern to agencies which collate information on drug misuse

Un-named identifier information may be collated

- **forename initial**
- **surname initial**
- **date of birth**
- **sex**

This allows identification of overlaps between data sources but preserves the anonymity of clients

As different data sources collate different information and have different types of client, it is important to have consistency within the data collected in the course of a capture-recapture research project. Several definitions need to be employed, relating to the scope of the project:

Time Period

Although the capture-recapture method assumes that the number of drug users is relatively stable over the time period that is being studied, this time period needs to be long enough to give sufficient numbers from each source. In many studies, a 12 month period is used, however in some areas where patterns of drug use are more fluent, shorter periods perhaps should be employed, such as 3 months or 6 months. Even when a 12 month period is being examined, it is prudent to record drug users identified from different periods within the year, for example those who are identified only from the first six months, only from the later six months, and those that would be identified from both semesters. Thus if there are not a sufficient number of sources available to analyse the data using the traditional multi-sample capture-recapture method using distinct sources, the two semesters can be modelled as separate sources. This would then need to be recognised within the analysis, as it may be assumed to be some relationship between the two semesters.

Age Group

Drug use is often more prevalent within younger age groups, however as the method concentrates on data from sources such as drug treatment agencies which do not usually cater for younger drug users, it is sensible to impose a lower limit on the age of those included within a study. As population figures within cities are usually presented within 5 year age groups, capture-recapture studies often stratify the collated data into 15-24, 25-34 and 35-64 groupings.

In practice, it is often easier to collect all possible data then extract the data for the required age groups in the data cleaning stage of the analysis. As dates of birth are often collected to allow identification of overlaps the choice of cut-off ages for the groups may be taken at a later stage if necessary.

Area

Local prevalence estimation studies are usually undertaken in cities, with the expectation that the resultant estimate refers to the number of drug users resident within the city. Care must therefore be taken to use a consistent definition of a city, particularly when the same city name is used interchangeably to describe a local government area, a health authority area or a conurbation. As previously noted, some degree of geographical information should be collected from each contributing data source. This may not be necessary if it is clear that each source only collects data from the city, and uses the same definition of the city as the prevalence estimation project.

Similarly, care must be taken that each of the contributing data sources covers the whole of the city. If some data sources only cover specific areas of a city, it is important to merge similar sources to gain a larger coverage, for example, if a city has two drug treatment agencies covering different areas, the data from these two agencies should be merged together to obtain a single treatment source covering the whole city. If data, even from only one of the contributing sources, are only available for one part of the city, then the definition of the local area that is being studied needs to be altered.

We have described above some of the definitions that need to be considered when undertaking a local prevalence study. There may be additional issues that need to be addressed. For example, we have described how it is important that all contributing sources cover the same area and supply information over the same time period and use a common age range. In the same vein, a data source that only has data on females may not be of immediate use in the analysis.

The above definitions are usually easy to implement; either an individual is within the defined age range or they are not. Other definitions are harder to implement, such as those relating to the drugs used by the individual and how problematic their use is. But it is just as important that all contributing agencies use similar definitions, for example a needle exchange may have clients that only use amphetamines, or even clients that only use steroids, therefore it would be wrong to use such data within a study which is examining opiate use. We will discuss later how even when a strict definition (such as opiate use) is common to all sources, variation in the severity of use within and between sources may cause problems in analysing the data.

Comparability of definitions between sources is required

- **time period**
- **age range**
- **geographical area**
- **drugs used**
- **nature of drug use**

3.4 Data collation

Data cleaning

An initial step in collating the data is to clean the data from each source. This process involves removing erroneous data and checking that all individuals meet the study criteria, i.e. they use the specified drug(s) and reside in the city. It is possible at this stage to produce summary statistics from each source, such as the number of clients, the mean age of the clients, the percentage that are male or the area of the locality that they reside in. When cleaning the data, a decision has to be made concerning records with missing fields, or those that have alternatives for a single field, such as a maiden and a married name for a female. It may be easier to eliminate all records that are incomplete, however where data are scarce, this may reduce the likelihood of obtaining a prevalence estimate. It is possible that a single data source may have multiple occurrence of the same individual, either in error or because of the nature of the data. If a record with missing fields refer to an individual who is already within the data source and a decision is made not to delete the record then the total number of people identified from within that source will be inflated. When eliminating multiple occurrences from within a single source, care must be taken not to eliminate potentially useful data. The example of deleting out a record where someone is detained in possession of heroin by the police instead of an instance when they were found in possession of cannabis has already been discussed, however if one record includes the person's area of residence and another does not, the geographical information should not be forgotten.

Identification of overlaps

Once the data from each contributing data source have been collected and cleaned, particularly in relation to the case definitions, a picture of the extent of drug use in a locality begins to be formed. The next stage is to fit the pieces of information together to identify the number of known users thus resulting in a multi-source enumeration. We have previously described the information that can be used to identify multiple occurrences, we will discuss the practical details of identifying these overlaps.

The information from each source must be stored in the same format, not only in terms of the computer packages used, but also in the structure of the

database or the spreadsheet. Although possible, we discount the manual identification of overlaps as being impractical in anything but the smallest of prevalence studies. A typical spreadsheet storing information from a prevalence study could take the format shown in Figure 3.1

Figure 3.1 Typical Spreadsheet Storing Data from a Capture-recapture study.

Source	Ref.	Fore Initial	Sur Initial	Date of Birth	Sex	Area	Drugs Used	Date
1	1	A	A	1/1/70	M	N	H	1/2/99
1	2	B	C	12/12/74	F	E	A, H	1/3/99

In this example, the first record refers to a male with initials AA, born on the 1st of January 1970, who lives in the north of the locality and was noted as using heroin on the 1st of February. The second record refers to a female who was noted as using amphetamines and heroin. They are indicated as being in the same source (in this case Source 1), but have a unique reference number from within that source. Other auxiliary data may be stored, some of which would be particular to a specific source. For example, within a police source it may be useful to note if the person was detained for possession or supply, or within a treatment source it may be useful to have some indication about the length of time an individual has been using drugs. As long as the core fields in each record are stored in the same format and in the same column, it does not matter how much additional information is recorded within a spreadsheet.

Once the data from each source is in the same format, with a different number indicating the source, there are several approaches to identifying overlaps. The method used could be subject to the individual preference of the researcher, however several issues should be noted.

There are two main types of framework for identifying overlaps; deterministic and probabilistic. Within a deterministic framework, two records either match or do not match. Under a probabilistic framework, the probability that two records refer to the same individual is calculated, and this information is used to decide if it is a match or not. The former method is straightforward, however small inaccuracies within the data may affect the matching process. The later method may be too complicated to implement, therefore it is often a compromise between the two methods that is used.

An understanding about the fields within a record is needed before matching commences. In some areas of Europe, either the forename or the surname of an individual may be practically any of the 26 letters of the alphabet. But some initials will be more common than others. Clearly the probability of someone

having the initials AB will more greater than the initials XV. The popularity of initials will vary throughout Europe, taking into account regional variations. Due to the age distribution of drug users, dates of birth corresponding to younger ages will be more common than those for example, in their sixties. This information on the relative probability of an individual having a given set of identifier information may be little more than common sense, however using it within a matching process is rarely systematic, rather it can be used subjectively in deciding what is or is not a match in some instances.

One method of identifying overlaps would to merge the data from all sources into a single data file of a spreadsheet package then sort it by one (or more) of the identifiers. To minimise the effort in matching, it is prudent to sort first by the field which has the most permutations such as the date of birth. If a secondary sort is by surname initial then visual scanning of a file will quickly identify matches. It is also useful to have sorted by source number, particularly if there are three or more sources. This assists in categorising matches, e.g. by not differentiating between a '12' match and a '21' match which could both notify an individual as being in sources 1 and 2. Although a computerised matching procedure can be devised, perhaps using the macro facilities of a spreadsheet or its formulae, visual inspection and identification of overlaps may be the easiest way of accounting for inaccuracies in the data, e.g. by finding two records in which the initials are the same but the first two parts of the date of birth have been transposed. Although judgements about what will constitute a match in these cases can be subjective, it is important to be a systematic as possible in the process.

Finally, when matching between data sources, care must be taken when allocating a match to each individuals. For example, it is sometime possible to note an overlap as an individual as being in source 1 and source 2, and still note that individual as being in source 1. One method would be to delete out the multiple occurrence so that only one record per unique individual remains within the data file.

The result of this data collation process, whether done as described above or by using other methods, will be a large file which indicated the sources that each individual has been identified from. These data now need to be converted into a format to allow analyses by extracting the number of people identified as being solely in each of the sources and being in each combination of more than one source. This can be done either by counting the number of each different type of overlap, by using the statistical functions of the spreadsheet or by exporting the data to a statistical package. The sex, geographical area, the drugs used and the age (as calculated from the date of birth) should also be used to crosstabulate the data allowing the overlap pattern to be separately determined for males and females, younger and older drug users, opiate use / amphetamine users etc.

The nature and quality of data will vary between sources,

**therefore the data cleaning and data collation
may be a complex process**

**Identifier information is subject to error, therefore the
identification of overlaps is subject to error**

**Errors in matching between sources may influence any
prevalence estimates**

3.5 Data Analysis

Once the size of the overlap between the different data sources have been found, then the data can be presented in the following manner, which is known as a contingency table.

Table 3.1 Typical Contingency Table Summarising Data from a Capture-recapture Study

		Source 1			
		Present		Absent	
		Source 2			
		Present	Absent	Present	Absent
Source 3	Present	6	79	34	389
	Absent	13	162	116	x

In this table, we use data which were used to provide an estimate for the size of the opiate using population of Toulouse, see EMCDDA (1997b) for additional information on the three sources. Here six individuals were identified from all three sources, and 79 individuals were identified from Source 1 and Source 3, but not from Source 2. The hidden population is denoted as x , and it is this quantity that the capture-recapture method aims to estimate, and thus give an estimate of the total population size. The analysis of these particular data is presented as a worked example in an Appendix to this document.

Statistical packages can be used to fit different models to the data and thus provide an estimate of the size of the hidden population. These different models will include the different relationships that may be between the three sources. If, however, the drug users identified from the three sources are substantially different from the drug users that have remained hidden, then there would be a three-way relationship between the sources. If this was the case, then the analysis would not be able to offer a reliable estimate.

To measure how close each model fits the data, a value known as the deviance can be calculated. Thus each estimate will have an associated deviance and the estimate which has the lowest deviance may be the most accurate. A

confidence interval can be attached to each estimate to give an indication about how accurate the estimate will be.

As the analysis, and also the interpretation of the results, requires some statistical expertise, we leave an extended discussion to the next section. It is, however, recommended that those considering using capture-recapture methods to estimate the size of a drug using population should establish contact with those who have experience in using the methodology.

The capture-recapture method may be used to obtain a prevalence estimate using data on the overlap between sources of information on drug misuse

This statistical method is not straightforward therefore statistical assistance may be required within a prevalence estimation project

There are methodological issues that need to be considered when using the method to estimate the prevalence of drug misuse

3.6 Dissemination of results

Although the capture-recapture method may appear complicated, there are several points that need to be considered when describing the research and writing up the results for dissemination. Many of the points raised below are common to any scientific study; the description should be transparent enough to allow others to re-analyse the data and to judge whether or not the conclusions made, particularly the prevalence estimates, are valid. Thus when writing about the research, the method should be more fully described, including a critical assessment of its applicability within the project.

Within a report, each of the data sources should be described, along with summary statistics such as the age distribution and the sex breakdown. If geographical information from the various sources has been collected, then a description of the spatial distribution of the 'known' drug users can be illuminating. It may not be possible to completely describe how the case definitions of the project, in particular in relation to type and severity of drug use of the agency's clients, are influenced by the choice of contributing sources, however a description of each source would be useful.

A description of the matching process used to identify overlaps should be included, stating whether or not exact matching has been used or if a more subjective process has been followed, perhaps allowing for close matches to be identified as overlaps. Some indication about how accurate the process has been would aid the reader.

In order that other people may analyse the data themselves, the overlap pattern should be presented in a contingency table, and where possible contingency tables of the stratified data should also be produced. A table detailing the results of the analysis should also be presented, including some of the statistical output such as the deviance and the 95% confidence interval from each of the possible models. A reference should be given to the method used to calculate the confidence estimate.

The resultant estimate should be placed in context with other information about the nature and extent of drug use within the locality within a discussion section. The estimate needs to be critically assessed, particularly in relation to the interactions that have been identified within the modelling process.

The results from a prevalence research project should not only be disseminated to policy makers and service providers, but also back to the agencies from which the data were obtained. Not only should this been done out of courtesy, but showing the end result from the research project may help in gaining future access to similar data.

It should be recognised that in producing a prevalence estimate, unwanted or inaccurate reporting in the media may follow. Indeed others with a vested interest in the results may seek to make political capital. Often there is little to compare a prevalence estimate with and what may appear to be a high prevalence of drug misuse may not be put into context or realistically compared with other areas. Unless comparable studies have been carried out over the same time period and using the same methodology and case definitions, there can be no valid direct comparisons made. That, however, may not stop people portraying the results as dramatic.

Drug misuse prevalence estimates, particularly at the local level, may be open to interpretation

Various people may have a vested interest in the estimates

The prevalence estimates may be discussed, out of context, by sections of the media

3.7 Summary

We have detailed the process that is undertaken when using the capture-recapture methodology to estimate the size of a drug using population. There are various stages, from identifying data sources, collecting the data and identifying overlap cases through to analysing the data using statistical methods. Each stage is important, however the skills needed differ. Many of the stages rely on a clear and systematic approach to working with the data, however the analysis stage may require additional statistical knowledge. In the

next section we describe the statistical analysis in greater depth.

4 Statistical Analysis

In this section we explore further the capture-recapture methodology, which can be used to estimate the size of hidden populations such as drug users. While it has been important in the preceding section to describe the method and the related assumptions in a manner accessible to those without a statistical background, it is also important that the methodology is discussed more rigorously. It is therefore assumed that the reader of this section is experienced in basic statistical analyses and has experience of using a statistical package such as GLIM, SPSS, SAS or BMDP. Those readers without a statistical background are advised to turn to the discussion in the following section. The following examples assume that three sources of data are used, however the analysis can easily be extended to four or more sources.

4.1 Capture-recapture data

Once the overlaps cases have been identified, using a method such as that described above, a contingency table can be constructed, as shown below.

Table 4.1 Typical Contingency Table Summarising Data from a Capture-recapture Study

		Source 1			
		Present		Absent	
		Source 2			
		Present	Absent	Present	Absent
Source 3	Present	<i>a</i>	<i>b</i>	<i>e</i>	<i>f</i>
	Absent	<i>c</i>	<i>d</i>	<i>g</i>	<i>x</i>

In this table, the number of people who were identified from all three sources is denoted as *a*, those that were identified from Sources 1 and 2, but not Source 3 are denoted as *c*, and those that were only identified in Source 3 are denoted as *f*. The hidden population is denoted as *x*, and it is this quantity that the capture-recapture method aims to estimate, and thus give an estimate of the total population size.

The number in each found in each source can be confirmed as follows:

$$N_1 = a + b + c + d$$

$$N_2 = a + c + e + g$$

$$N_3 = a + b + e + f$$

Where N_i is the total of individuals identified from Source *i*. The observed

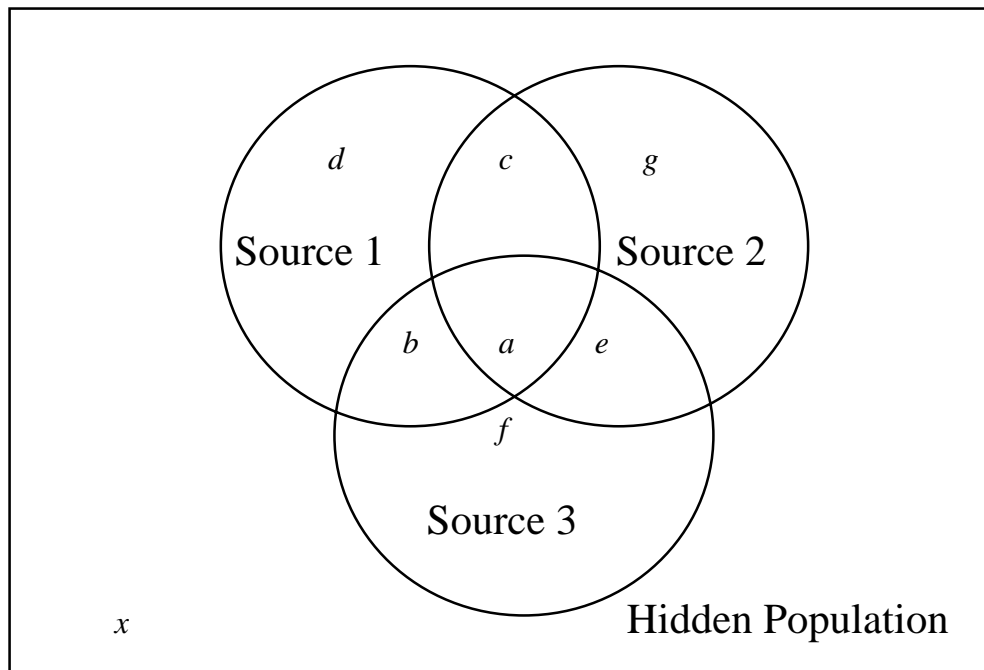
population can also be found:

$$N = a + b + c + d + e + f + g$$

Thus N is the total number of drug users found from the multi-source enumeration.

The data can alternatively be presented as a Venn Diagram, as shown below.

Figure 4.1 Venn Diagram Summarising Data from a Capture-recapture Study



4.2 Log-linear models

The capture-recapture data, as summarised by a contingency table, can be analysed using a statistical package such as GLIM or SPSS. A log-linear regression model can be fitted to the data and this model can be used to provide an estimate of x , the size of the hidden population. An introduction to log-linear modelling, in particular its application in estimating the size of hidden populations is to be found in Bishop, Fienberg and Holland (1975). Different log-linear models can be fitted to the data to include interactions between data sources. For example, if drug users attending a treatment agency were more likely to have been admitted to hospital, then there would be positive interaction between these two sources. A decision as to whether such interactions are present can be made by examining a value known as the deviance (G^2). This measures how closely the observed data agree with the model and is similar to the χ^2 values commonly used to examine contingency

tables. As there are only seven cells in the contingency table instead of the eight that would usually be in such a table, it is only possible to fit interactions between two data sources (although three of these different interactions can be fitted), along with fitting pairs of interactions.

It is not possible to include a simultaneous interaction between all three data sources therefore it is a necessary within the analysis to assume that this ‘three-way’ interaction is zero. This assumption may not be valid, and the validity will be related to the size of the deviance when fitting the best model to the data. If this deviance value is large, even after including all possible interactions, then the resultant estimate may be biased. Once an estimate has been obtained, then a confidence interval can be produced. Several methods can be used to produce this confidence interval, depending on the statistical package that has been used to analyse the data.

4.3 Model Selection

Hook and Regal (1997) discuss the validity of methods for model selection and weighting for model uncertainty in capture-recapture estimation. In addition to describing the Akaike Information Criterion (AIC), which can be used to assess whether interactions should be included in models and therefore can be of use in choosing the ‘best’ model (Akaike, 1985), they discuss the Bayesian Information Criterion (BIC) which can take two different forms, the original as proposed by Schwarz (1978), the other an alteration proposed by Draper (1995). Hook and Regal denote these as SIC and DIC respectively. The formulae for these criteria are as follows:

$$\begin{aligned} \text{AIC} &= G^2 - 2(\text{df}), \\ \text{SIC} &= G^2 - (\ln N_{\text{obs}})(\text{df}), \\ \text{DIC} &= G^2 - (\ln (N_{\text{obs}}/2\pi))(\text{df}), \end{aligned}$$

where G^2 is the deviance and df is number of degrees of freedom associated with the model, N_{obs} is the ‘known’ population of opiate users and \ln denotes the natural logarithm function. When using any of these criteria the model with the lowest value of the criterion would be the favoured one.

Hook and Regal go on to discuss the use of the weighted Bayesian Information Criteria, in which both the SIC and the DIC can be used to obtain a weighted average of the different estimates from the three-sample capture-recapture method; these weighted averages can also be applied to the upper and lower values of the 95% confidence intervals associated with the eight estimates. These weighted estimates can be considered as ‘Bayesian’ as they combine an estimate with a measure of how likely it is, to produce the weighted estimate. Hook (personal communication) recommends using all possible three-sample estimates to construct the weighted estimate, including the model which includes all two-way interactions between the three data sources and has an associated deviance of zero. There does, however, appear to be some justification for not including this ‘saturated model’. Similarly, it may be

questionable to include estimates in which the associated deviance suggests that the model clearly does not fit the data.

The selection of the models used to describe the overlap pattern and to produce an estimate of the size of the hidden population is, on one level, intrinsically linked with the discussions about case definitions and the data sources that are employed. On another level the concept of parsimony is also important such that simple models without complex interactions may be preferable. It is not always possible to fit certain models to the stratified data due to zero values in some cells. Related to the problem of structural zeros, where a mechanism exists which ensures that no overlap can occur between certain sources is the problem of artificial overlaps where inclusion in one source will automatically mean inclusion in another.

An additional benefit of the three-sample capture-recapture method, as opposed to the two-samples, is that a specific interaction can be included in the model, even when a criterion such as the AIC does not make the suggestion. In many instances, interactions can be included which are consistent with what professionals in the field would expect, i.e. 'medical' sources being dependent on each other but independent of a 'legal' source. Without returning to the discussion the case definitions, this is sensible when you consider the differences between those who have 'medical' problems due to drugs and those who only have legal problems.

The debate between including interactions because they seem sensible, and being led by statistics such as the change in deviance or the AIC, is made more interesting because log-linear modelling, which capture-recapture modelling essentially consist of, is the preferred method of examining categorical data, and thus discovering relationships between sources. It is also possible that the sample sizes, in this case the data from the sources may not be large enough to show significant interactions, even when they are present. If these interactions are thought to be present between data sources then they should perhaps be included.

4.4 Confidence Intervals

The estimates obtained from the application of a method such as capture-recapture need to be assessed in conjunction with the statistical uncertainty that is inherent in any estimate. Those that successfully undertake capture-recapture studies are in some way fortunate in that the estimates produced can usually only be compared with ball-park estimates derived by those working in the field. The number of 'current' drug users within a city will be variable and, as trends in drug use can fluctuate, the best any prevalence study can hope for when using retrospective data is a historical estimate of drug use; this usually is still of relevance to policy makers at the present time. Additionally, any statistical estimate needs to be interpreted with caution for the reasons detailed

below.

The estimate produced from a capture-recapture project is the end point of a process in which error can be introduced at several stages. Mistakes can be made during the collation of an individual data source; identifiers such as the date of birth can be recorded erroneously either by mistake or because falsified information has been recorded, particularly in a low-threshold agency. The matching process is not always guaranteed to correctly identify all matches, or eliminate any possible false matches. It is only once a contingency table has been produced that the statistical modelling can commence, and even then, the relevance of the confidence interval depends on how accurately the model portrays the relationships between the data sources.

Various methods for producing a confidence interval can be used, however a method favoured by Cormack (1992) and by Regal and Hook (1984) which recognises that the estimate for the hidden population is derived from an asymmetric distribution has commonly been used. Thus the possible problem of producing confidence intervals, in which the lower bound of the total drug using population is less than that which has been identified from the multi-source enumeration, does not arise. Sometimes, however, the upper limits of the confidence intervals reported are simply not feasible, and these usually occur when the model has included more than one interaction. That is not to say that a preferred estimate should be chosen because of its small confidence interval.

The three-sample capture-recapture analysis can be carried out on a range of statistical software, often with the aid of pre-written macros or subroutines. While the point estimates derived are usually exactly similar, the methods for obtaining confidence intervals sometimes differ therefore there needs to be a consensus about which packages are of most use. We can now demonstrate the use of two different packages, GLIM and SPSS in analysing capture-recapture data.

4.5 Analyses using GLIM

This section describes how the GLIM statistical package (Francis et al, 1994) can be used within the analyses of data from multiple sources within a capture-recapture analyses. The GLIM package (GLIM4) was developed by the Royal Statistical Society (UK). It is described as 'The Statistical System for Generalized Linear Interaction Modelling' and can be used in many types of statistical analysis of data, including model fitting. More information about the package can be found at the NAG website <http://www.nag.co.uk> It should be noted however that the package operates on PCs through an interface to MS DOS, and both the installation and operation of the package are less straightforward than traditional Windows based packages.

Capture-recapture data

To analyse the data using GLIM, we assume that the data has been converted into the following format.

Table 4.2 Data Table Summarising Data from a Capture-recapture Study

Source 1	Source 2	Source 3	Count
1	1	1	<i>a</i>
0	1	1	<i>e</i>
1	0	1	<i>b</i>
0	0	1	<i>f</i>
1	1	0	<i>c</i>
0	1	0	<i>g</i>
1	0	0	<i>d</i>
0	0	0	<i>x</i>

The hidden population in all of the above is denoted as x , and it is this quantity that the capture-recapture method aims to estimate, and thus give an estimate of the total population size. Within the data table, the presence in each source is denoted by 1, and absence by 0.

$$N_1 = a + b + c + d$$

$$N_2 = a + c + e + g$$

$$N_3 = a + b + e + f$$

The observed population can also be found:

$$N = a + b + c + d + e + f + g$$

Analyses

The three sources are known as factors and the 7 counts (and the missing value) are known as the response variable. Annotated output from a GLIM session is presented within this document. Once a GLIM session has started, commands are inputted and they appear in the annotated output preceded by [*i*]. Commands in GLIM are always preceded by the dollar sign \$, and a command line is terminated by \$. Only the first four letters of a command are recognised, therefore abbreviations can be used. To repeat a command the ':' can be used. The first command is to set the standard length of the data, using the command

```
$units 8$
```

Thus when any data are entered, GLIM expects 8 values. The series of counts can then be entered, using the `$data r$` and `$read` commands. To input the data line by line, we can omit the final `$` to get prompted for data by GLIM. Thus the seven counts will be stored in a variable *r*. By convention the missing value is set to be zero.

We need to input the factors (presence or absence from each source) and other information about the modelling process that will be undertaken. By convention, the factors are named *p1*, *p2* and *p3*, and a weighting variable *w* is assigned. This weighting variable lets GLIM know that the 8th value of *r* is missing. To assist in fitting different models, it is useful to assign names to the different interaction terms that can be fitted. By convention, an interaction between source 1 and source 3 is named *i13*. To shorten these names, the interaction between source 1 and source 2 is named *i1* instead of *i12* and the other interaction term is named *i2*. This is done by multiplying the factors together, using the commands such as

```
$calc i1=p1*p2$
```

Some other information is needed for GLIM to fit different models; an indication of which variable is the response variable, what distribution the values of the count variable will take (usually Poisson), and what weight should be used

```
$yvar r$
```

```
$err p$
```

```
$weight w$
```

Fitting different models is done using the `fit` command. For example

```
$fit p1+p2+p3$
```

will fit the basic model with no interactions. GLIM responds with two values, the scaled deviance and the residual df. The scaled deviance, known as G^2 , is comparable to values compared to a χ^2 when examining contingency tables. G^2 is a measure of how well the model fits the data, and thus we wish it be low. There are 3 degrees of freedom (df) when fitting this basic model.

Different interactions terms can be added:

```
$fit +i1$
```

or removed

```
$fit -i1$
```

such that

```
$fit p1+p2+p3$
```

followed by

```
$fit +i1$
```

would fit the model with an interaction between source 1 and source 2. Alternatively this can be done directly by

```
$fit p1+p2+p3+i1$
```

The benefit of adding and removing the interaction terms is that GLIM will give the change in deviance as well. As a check to see what model has actually been fitted by **d**isplaying the **m**odel.

```
$d m$
```

would display the model. To see what GLIM has estimated the missing value to be there are two options. One is to **d**isplay a table which presents the observed and fitted values, along with the **r**esiduals:

```
$d r$
```

The other would be to **l**ook at the **8th** fitted **v**alue:

```
$look 8 %fv$
```

It should be noted that this is the estimated size of the hidden population, the known population (the sum of the other seven cells) must be added to get the total population size estimate. When there are three data sources, there are essentially only eight models that can be fitted. The independence model with not interactions, the three models that have one interaction, the three models that have two interactions, and the model that has all three interactions, which is known as the saturated model. As the saturated model uses all available data, the degrees of freedom should be zero, as will be the deviance.

To examine the significance of each of the interactions that are included in the model, the GLIM can **d**isplay the **e**stimates:

`$d e$`

It is quite useful to know the size of the interaction and the standard errors associated with them, particularly when the sample size is small or when a lot of interaction terms are included in a model.

So far, only point estimates (along with a measure of how good the associated model is) have been obtained. It is also necessary to obtain confidence intervals. Although Bishop, Fienberg and Holland (1975) suggest a method for obtaining symmetrical confidence intervals, the approach, independently proposed by Cormack (1989) and Regal and Hook (1984) is preferable. Both methods aim to find a value, sufficiently far from the point estimate to increase the deviance by 3.84 (the 5% significance point of the χ^2 distribution). Two further point estimates, both known as weighted Bayesian estimates, can be produced, along with confidence intervals. In short, these estimates use either the DIC or SIC, along with the corresponding estimates, to obtain a weighted estimate.

Finally, the command to end a GLIM session is `$stop$` and two files should have been produced `glim.jou`, which records the commands that have been entered in the preceding session, and `glim.log` which additionally records the output. As the next session of GLIM would overwrite these files, it may be useful to rename them if they are to be needed again.

Macros can be written for GLIM, and the analysis of capture-recapture data can be simplified by using a macro, however it is useful to have a basic knowledge of some of the commands used in GLIM. For example, most of the commands used in describing the analysis, apart from setting the unit length and entering the count data, can be undertaken by running a macro. A macro can also be created to obtain and output the estimate for the size of the hidden population for each possible model. Such a macro can also easily calculate the two sample estimates (e.g. only using source 1 and source 2 and looking at the traditional two-sample estimate). Following on from Hook and Regal (1997), criteria for selecting the 'best' model such as the AIC, DIC or SIC can also be calculated using a macro and the method for obtaining a 95% confidence interval proposed by Cormack can also be implemented using a macro.

4.6 Analyses using SPSS

This section describes how the SPSS statistical package can be used within the analyses of data from multiple sources within a capture-recapture analysis. It may be easier to analyse capture-recapture data using the GLIM software, partly because these packages are more suited to the type of analysis. A description of the analyses using SPSS is presented more for those who are

acquainted with this statistical package or for those who only have access to this package.

SPSS is a powerful software package for data management and analysis. Various versions will be in use on different computer systems, however the analyses described in this report requires the equivalent of Version 6.1 for Microsoft Windows or a more recent version. The examples presented within this report were obtained by using Version 7.5 for Microsoft windows using the General Loglinear analysis commands which are contained within the Advanced Statistics add-on enhancement to the base system. Further information on the SPSS package can be found at <http://www.spss.com>

The more recent versions of SPSS make use of drop down menus as opposed to syntax commands that some more experienced SPSS users may be more familiar with. The following example describes the use of the menus, however the syntax is also presented.

Capture-recapture data

We have previously shown how the data can be presented visually within a Venn diagram. We now need to show how this data can be converted into a format required by SPSS.

Table 4.3 Data Table for Use with SPSS

Source 1	Source 2	Source 3	Count
1	1	1	<i>a</i>
2	1	1	<i>e</i>
1	2	1	<i>b</i>
2	2	1	<i>f</i>
1	1	2	<i>c</i>
2	1	2	<i>g</i>
1	2	2	<i>d</i>
2	2	2	<i>x</i>

The hidden population in all of the above is denoted as x , and it is this quantity that the capture-recapture method aims to estimate, and thus give an estimate of the total population size. Within the data table, the presence in each source is denoted by 1, and absence by 2. It should be noted that this terminology differs from the conventional ‘1 for presence and 0 for absence’ which is used when analysing the data using GLIM. This is to enable the analyses in SPSS.

The number in each found in each source can be confirmed as follows:

$$N_1 = a + b + c + d$$

$$N_2 = a + c + e + g$$

$$N_3 = a + b + e + f$$

The observed population can also be found:

$$N = a + b + c + d + e + f + g$$

Analyses

To analyse this type of data, it must be in the format of a data table as described above. The three sources are known as factors and the 7 counts (and the missing value) are known as the response variable. Annotated output from an SPSS session is presented as within this document.

The first step is to enter the data into SPSS. The data presented here is the same as that presented in the GLIM analyses to enable a more direct comparison. The data can be entered in the following format:

Table 4.4 SPSS Spreadsheet

p1	p2	p3	w	r
1	1	1	1	6
2	1	1	1	34
1	2	1	1	79
2	2	1	1	389
1	1	2	1	13
2	1	2	1	116
1	2	2	1	162
2	2	2	0	0

It may be convenient to define each variable p1, p2, p3, w and r as type 'Numeric' with 0 decimal places. We show here the data that are analysed as a worked example in an Appendix.

The variable which stores the seven counts, r, needs to be identified as such. This can be done by using the

Data
Weight Cases...
Weight cases by

commands, where the 'Frequency Variable' is selected to be r.

In order to find the best fitting model, a variety of approaches can be taken, however one of the most convenient would be to fit the model with all 2-way interactions and use the back stepping method to test which interactions can be

removed from the model (Smit, Brunenberg and Van der Heijden, 1996).

To do this, the

Statistics
Loglinear
Model Selection

commands are used. The three factors; p1, p2 and p3, are entered into the **Factor(s)** box and the range of these factors are defined to have the minimum value 1 and maximum value 2. The **Cell Weights** are set to be w. The **Model...** window can be used to include all 2-way interactions (p1*p2, p2*p3 and p1*p3) as the model which the backward elimination starts at. Using this method should result in the most parsimonious model which SPSS then displays.

To then go on to find the estimate of the hidden population size, the parameter estimates for that model must be obtained. This is done by using the

Statistics
Loglinear
General

commands. Here again p1, p2 and p3 are entered into the **Factor(s)** box and w is entered into the **Cell Structure** box. The **Distribution of Cell Counts** should be set to be Poisson and the **Model...** window should be used to enter the specific model which SPSS has suggested is the most parsimonious (or for that matter any other model for which an estimate is required). Use the **Options...** window to specify that the **Estimates** are required. The plots that SPSS suggests are not required.

The estimate of the hidden population size can be calculated as the exponential of the value of the constant parameter. An asymptotic 95% confidence interval for the size of the hidden population can be obtained by calculating the exponentials of the vales SPSS gives as a confidence interval.

Instead of using the menu-driven commands, the following SPSS syntax can be used, after the data has been entered in the above format.

```
Weight by r.  
Hiloglinear  
  p1(1,2) p2(1,2) p3(1,2)  
  /cweight w  
  /method backward  
  /print freq resid  
  /design p1*p2 p1*p3 p2*p3.  
Genlog  
  p1 p2 p3  
  /cstructure = w  
  /model = Poisson  
  /print estim  
  /plot none
```

/design p1 p2 p3 p1*p3. (or whatever the suggested model is)

The annotated output from such an analyses is to be found at the end of this report.

Although SPSS can be used to fit various models to capture-recapture data, and thus provide an estimate for the size of a hidden population, it does not appear capable of calculating non-symmetric confidence intervals based on the likelihood approach such as that proposed by Regal and Hook (1984) or Cormack (1992).

5 Discussion and Conclusions

In the preceding sections we have described how data on drug misuse can be collected from a range of sources and can then be used to provide an estimate of the prevalence of drug misuse using capture-recapture methods. In this section we provide a methodological discussion, focussing on aspects of the prevalence estimation exercise that need to be addressed when the reliability and validity of an estimate is considered.

5.1 Methodological issues

The capture-recapture methodology was developed to estimate the size of animal populations and has been used in epidemiological studies to estimate the size of under-reporting in disease registers. A review of the theoretical development of the methodology is to be found in a related publication, however a discussion of the methodological issues, particularly in relation to estimating the size of drug using populations is now warranted.

In studies which aim to estimate the size of an animal population (for example see Seber, 1982), there are a series of assumptions or conditions that must be true for the estimates to be valid. Most of the assumptions translate into our epidemiological application, for example:

- The population is closed; there is no movement into or out of the population in the period that is being studied;
- Those that are present in more than one source - the overlap cases - are identified as such.
- Being present in one source does not effect the probability of being in another source;

We can now look at these assumptions in closer detail and discuss their validity with respect to drug misuse prevalence estimation.

For the first assumption to be true, then the population of drug users must remain constant over the study period and nobody starts or ceases to fall under the remit of the study. There are two main ways that this can be violated, either by individuals moving away from or into the study area, or by people either ceasing or starting to use drugs. In some areas, the movement in and out of a city by drug users can cast doubts on the validity of the assumption, for example by drug 'tourists' visiting areas where drugs are more available. This problem can be partially addressed by stipulating in the study definition that persons must be resident in the city, or have been resident for some time prior

to inclusion in the study. In other areas, the small amount of migration of drug users would do little to affect the analysis.

The second problem with the 'closed population' assumption is often harder to address. It would be unrealistic to assume that people would not begin using drugs during the study period, and clearly the provision of drug treatment services may indeed help people to stop using drugs. The increased mortality of drug users should not be ignored either. One approach to making the assumption more valid would be to shorten the study period. The associated cost may however be to reduce the amount of available data on drug users such that it is not possible to fit the most realistic model. This may result in either a totally unrealistic estimate or a plausible, but false, one.

The second assumption requires that the identification of overlaps is not subject to error. As in other areas, there exists the possibility of false positives and false negatives. Due to errors in the recording of data, and because full identifier information is seldom available with which to match, people may be wrongly identified as being in more than one source, or more likely that an overlap case is not identified as such.

The third assumption is often harder to describe as it relates to more than one problem. The first problem is that the study population is heterogeneous; drug users, for whatever reason, are not all equally as likely to be present in a particular source. The second problem is that those present in one source may be more, or less, likely, to be in another. Drug users, and their lifestyles, are diverse, and this diversity can be reflected in many ways. Steps can be taken to make the first part of the assumption valid. It may perhaps be possible that male drug users have a different probability of being arrested than female drug users, or that younger drug users are less likely to contact services than older drug users. This variability, if it exists, should therefore be accounted for. One method would be to stratify the population, perhaps by age or sex, to give a more homogenous population on which to undertake the analysis, however the comment above about reducing the amount of data on which to perform the analysis is equally valid.

There is a related problem in that spatial variation may also exist. This problem may not be so pertinent when an entirely urban location is being studied, but if there are areas within the study location where perhaps access to drug treatment services are reduced, or the police may be less likely to detain drug users then a model fitted to the data for the entire city may not be valid for that area. Again a solution may be to stratify the data by geographical area.

There also exists the problem such that, although it may be possible to stratify the population by age or sex, a more worrying source of heterogeneity may be the severity of a persons drug problem. Clearly those with a bigger drug problem would be more likely to contact services, therefore a prevalence study which only used sources of data which cater for more problematic drug user

should result in an estimate of the total number of problematic drug users in that area. However, other drug users may be better classified as having only a legal problem with drug use, either due to laws restricting the possession or the supply of drugs, or the need to commit crime to finance their drug use. These drug users may be unlikely to contact medical services and if the contributing sources therefore include both problematic and non-problematic drug users, the 'equal probability' assumption may be violated.

The second part of the 'equal probability' assumption is also difficult to resolve. An example of how this assumption may be violated would be a system where those who have been detained for a drug offence by the police would be required to attend at a drug treatment service, or perhaps because someone is being prescribed methadone from one treatment source then they would not be attending the local needle exchange.

The effect of this, and the other violations of the assumptions is often unclear. Methodological studies can begin to explore the possible effects on prevalence estimates any violation of the assumptions may have, unfortunately the dynamics of drug misuse is often far too complex to be completely understood. While the violation of one assumption may result in an underestimate, another violation may result in an overestimate. It cannot however be assumed that these would cancel each other out.

It is often not until the analysis stage that the effect of any violation of the assumptions may be recognised, perhaps by the inability to obtain a simple model that fits the data. Even when a model appears to accurately describe the overlap data, it cannot be guaranteed that a violation of the assumptions has rendered the estimate unreliable. It is because of this that some regard the capture-recapture method itself as being unreliable and would not recommend its use in prevalence estimation (Papoz *et al.*, 1996). Others take a more pragmatic approach, preferring to undertake the analysis and then to judge the estimate's reliability, particularly in relation to the assumptions and the model that has been fitted to the data (IWGDMF, 1995; Hook and Regal 1995). There is no right or wrong answer, the estimate produced is just that; an estimate. Whether or not that is of use in determining how many people use drugs in a specific locality is a decision that can only be made after the analysis is completed.

5.2 Refinements

There are several refinements to the basic analyses that can be made, lessening the effect possible violations of assumptions and thus making the resultant estimate more valid. While one approach would be to increase the number of data sources, perhaps by collecting additional data, it may be more sensible to effectively decrease the number of sources that are used within the analyses.

If many sources are available, it may be that some of them have a similar remit or that relationships exist between them to justify merging them together. Thus a capture-recapture study which initially had four or five sources may benefit from merging similar sources together and re-analysing the data. Another approach may be to examine combinations of sources selected from the total number of available sources. For example, if five sources were initially available, systematically removing one or two sources and re-analysing the data may uncover relationships between the sources and aid the decision as to what constitutes a valid estimate. An extreme example of this would be to perform two-sample analyses on subsets of the data. It would be unlikely if any of these estimates in themselves would be reliable, however the combination of estimates may then give an insight as to what the true population size may be.

There is a related method, known as the truncated Poisson model which can be used to obtain a prevalence size estimate from one source of data. It requires information on the number of times an individual has been in contact with that particular source, and therefore may be more suited to some sources rather than others. It works on the principle that if the number of people attending once, twice, three times and so on is known, then this information can be used to estimate how many people have attended zero times, i.e. the hidden population. In a similar manner to the three-sample capture-recapture method assuming a log-linear model to describe the contact pattern between sources, the truncated Poisson model would model the number of contacts within a single source as a Poisson distribution. As no information can be drawn about the number of zero contacts, the Poisson distribution is truncated. A fuller description of this methodology is to be found in the report from the previous EMCDDA project (EMCDDA, 1997*b*), however it should be noted that the estimates are again subject to some of the assumptions present in the capture-recapture analysis: the closed population assumption, the homogeneous population assumption and the constant probability of being observed more than once assumption, in this case over time.

Other refinements to the basic model may theoretically be possible, for example by implicitly modelling an open population, however such models would require more data than the closed population capture-recapture method. Thus this method has rarely been used within Europe.

5.3 Conclusions

We have described above some of the more pertinent issues that need to be considered when carrying out a local prevalence study using the capture-recapture method. The section has been drawn from the authors experience in applying the methodology in a variety of settings, however it would be over-optimistic to assume that all possible problems that a prevalence study may encounter have been described. Local variations in the nature and extent of drug misuse, along with factors such as the availability of data may serve to

present additional problems that can only be addressed at the local level.

The estimation of the prevalence of drug misuse at the local level should be put into context. In some instances the availability of data may result in the costs of initiating a study becoming prohibitive, particularly when the success of a study may be limited because of the scarcity of data. In other areas, data on drug misuse is already being collated, and a capture-recapture study may be an inexpensive way of interpreting the data and giving an estimate of the total population size. Thus the costs of a prevalence study need to be weighed against the possible benefits, particularly in relation to other the different methods that can be used. The fact that a capture-recapture study can use existing data may mean that it can be one of the quickest methods of providing a prevalence estimate, particularly if it is incorporated into a monitoring system. Although repeat capture-recapture studies have been rare, there appears no reason why they cannot be systematically applied, perhaps in conjunction with a monitoring system to provide information on trends.

Thought must also be given as to the uses of prevalence estimates and the proper interpretation. As an initial step within a capture-recapture study is to find the number of known users, the ratio of known to unknown users can be obtained and considered. Care must be taken however in making direct comparisons between the known population, the estimated population and the number of people contacting treatment services. There may be many valid reasons why only a proportion of the total population of drug users are in contact with services. Some drug misusers may not wish to seek treatment for their drug problems, indeed they may not even perceive themselves to have a problem. Other drug misusers may feel that their problems may not be catered for by existing services.

Finally, care must also be taken to distinguish what the prevalence estimate does not encompass. Capture-recapture estimates typically refer to drug misuse at the more problematic end of the drug using spectrum, and can often disregard the use of particular drugs such as amphetamines. When using any prevalence estimate, it must be stressed that other drug problems may exist, and just because one type of drug problem has been enumerated, it does not mean that others are irrelevant. This warning may be particularly pertinent when using the results of a capture-recapture study to plan services. There may be a vicious circle in that data from treatment agencies which mainly provide treatment for opiate use are used to show that opiate use is prevalent in that area and therefore more opiate misuse based services should be provided, perhaps at the expense of providing services for other forms of drug misuse.

6 References

- Akaike, H. (1974) A new look at the statistical model identification. *IEEE Transactions on automatic control*, **19**:716-723.
- Bello, P.-Y., and Chêne, G. (1997) A capture-recapture study to estimate the size of the addict population in Toulouse, France In EMCDDA Scientific Monograph Series, Number 1, *Estimating the prevalence of problem drug use in Europe*. EMCDDA, Lisbon
- Bishop, Y., Fienberg, S., and Holland, P. (1975) *Discrete Multivariate Analysis*. MIT Press, Cambridge MA.
- Cormack, R. M. (1992) Interval estimation for mark-recapture studies of closed populations. *Biometrics*, **48**:567-576.
- Domingo-Salvany, A., Hartnoll, R. L., Maguire, A., *et al.* (1995) Use of Capture-Recapture to Estimate the Prevalence of Opiate Addiction in Barcelona, Spain, 1989. *American Journal of Epidemiology*, **141**:567-574.
- Draper, D. (1995) Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society, Series B*, **57**:78-79.
- EMCDDA (1997a) Scientific Monograph Series, Number 1. *Estimating the prevalence of problem drug use in Europe*. EMCDDA, Lisbon.
- EMCDDA. (1997b) Methodological pilot study of local level prevalence estimates. EMCDDA, Lisbon.
- Francis, B, Green, M, Payne, C. eds. (1994) *The GLIM System, Release 4*, Oxford: Clarendon.
- Frischer, M. (1992) Estimated prevalence of injecting drug-use in Glasgow. *British Journal of Addiction*, **87**:235-243.
- Frischer, M., Leyland, A., Cormack, R., *et al.* (1993) Estimating the population prevalence of injection drug use and infection with human immunodeficiency virus among injection drug users in Glasgow, Scotland. *American Journal Of Epidemiology*, **138**:170-181.
- Ghodse, A, Sheehan, M, Taylor, C, and Edwards, G. (1985) Deaths of drug addicts in the United Kingdom, 1967-1981. *British Medical Journal*, **290**, 425-428.
- Hartnoll, R., Mitcheson, M., Lewis, R., *et al.* (1985) Estimating the prevalence of opiod dependence. *Lancet*, **i**:203-205.

Hay, G., and McKeganey, N. (1996) Estimating the prevalence of drug misuse in Dundee, Scotland: An application of capture-recapture methods. *Journal of Epidemiology and Community Health*, **50**:469-472.

Hook, E. B., and Regal, R. R. (1995) Capture-recapture methods in epidemiology - methods and limitations. *Epidemiologic Reviews*, **17**:243-264.

Hook, E. B., and Regal, R. R. (1997) Validity of methods for model selection, weighting for model uncertainty, and small sample adjustment in capture-recapture estimation. *American Journal of Epidemiology*, **145**:1138-1144.

International Working Group for Disease Monitoring and Forecasting. (1995) Capture-recapture and multiple-record systems estimation II: Applications in human diseases. *American Journal Of Epidemiology*, **142**:1059-1068.

Korf, D. (1997) The tip of the iceberg: Snowball and nomination techniques, the experience of Dutch studies. In EMCDDA Scientific Monograph Series, Number 1, *Estimating the prevalence of problem drug use in Europe*. EMCDDA, Lisbon.

Olson, B. (1997) Problems of definition and other aspects of case finding: the Swedish nationwide studies. In EMCDDA Scientific Monograph Series, Number 1, *Estimating the Prevalence of Problem Drug Use in Europe*. EMCDDA, Lisbon

Papoz, L, Balkau, B and Lellouch, (1996) Case counting in epidemiology – limitations of methods based on multiple data sources. *International Journal of Epidemiology*, **25**, 474-478.

Regal, R. R., and Hook, E. B. (1984) Goodness-of-fit based confidence intervals for estimates of the size of a closed population. *Statistics in Medicine* **3**:287-291.

Schwarz, G. (1978) Estimating the dimension of a model. *The annals of statistics*, **6**:461-464.

Seber, G. (1982) *The estimation of animal abundance and related parameters*. Charles Griffin, London.

Smit, F, Brunenberg, W, Van der Heijden, P. (1996) Het schatten van populatiegroottes: toespansingen en een voorbeeld. *Tijdschrift Sociale Gezondheidszorg*, **74**, 171-176.

Squires, N. F., Beeching, N. J., Schlecht, B. J. M., *et al.* (1995) An estimate of the prevalence of drug misuse in Liverpool and a spatial analysis of known addiction. *Journal Of Public Health Medicine*, **17**:103-109.

Appendix I Typical GLIM session which provides an estimate of the size of a hidden population using data from 3 sources.

```
[o] GLIM 4, update 8 for IBM etc. 80386 PC / DOS on 27-Jun-1998 at 14:38:07
[o] (copyright) 1992 Royal Statistical Society, London
[o]
[i] ? $units 8$           Sets the size of the data columns. 8 cells in the contingency table.
[i] ? $data r$           Input the data from the contingency table
[i] ? $read
[i] $REA? 6
[i] $REA? 34
[i] $REA? 79
[i] $REA? 389
[i] $REA? 13
[i] $REA? 116
[i] $REA? 162
[i] $REA? 0               This is the missing cell which we aim to estimate. It should be set to zero.

[i] ? $data p1$          This factor describes presence or absence from source 1
[i] ? $read
[i] $REA? 1
[i] $REA? 0
[i] $REA? 1
[i] $REA? 0
[i] $REA? 1
[i] $REA? 0
[i] $REA? 1
[i] $REA? 0

[i] ? $data p2$          This factor describes presence or absence from source 2
```

```
[i] ? $read
[i] $REA? 1
[i] $REA? 1
[i] $REA? 0
[i] $REA? 0
[i] $REA? 1
[i] $REA? 1
[i] $REA? 0
[i] $REA? 0
```

```
[i] ? $data p3$
[i] ? $read
[i] $REA? 1
[i] $REA? 1
[i] $REA? 1
[i] $REA? 1
[i] $REA? 0
[i] $REA? 0
[i] $REA? 0
[i] $REA? 0
```

This factor describes presence or absence from source 3

```
[i] ? $calc i1=p1*p2$
[i] ? $calc i2=p2*p3$
[i] ? $calc i13=p1*p3$
```

Calculates the interaction terms

```
[i] ? $data w$
[i] ? $read
[i] $REA? 1
[i] $REA? 1
[i] $REA? 1
[i] $REA? 1
```

This variable is a weight used to denote which is the missing cell

```
[i] $REA? 1
[i] $REA? 1
[i] $REA? 1
[i] $REA? 0
```

```
[i] ? $look p1 p2 p3 w r$      Just to check the structure of the variables
```

```
[o]            P1            P2            P3            W            R
[o] 1    1.000    1.000    1.000    1.000       6.00
[o] 2    0.000    1.000    1.000    1.000      34.00
[o] 3    1.000    0.000    1.000    1.000      79.00
[o] 4    0.000    0.000    1.000    1.000     389.00
[o] 5    1.000    1.000    0.000    1.000      13.00
[o] 6    0.000    1.000    0.000    1.000     116.00
[o] 7    1.000    0.000    0.000    1.000     162.00
[o] 8    0.000    0.000    0.000    0.000       0.00
```

```
[i] ? $yvar r$                      Tells GLIM that r is the response variable in the regression model
```

```
[i] ? $err p$                        Tells GLIM that the error structure is Poisson
```

```
[i] ? $weight w$                    Tells GLIM to ignore the cells that we have weighted out when fitting models
```

```
[i] ? $fit p1+p2+p3$                This fits the basic model with no interactions between the data sources
```

```
[o] scaled deviance = 6.5681 at cycle 3
[o]     residual df = 3            from 7 observations
[o]
```

The output here presents the deviance (G^2) value, which can be compared with a χ^2 distribution with 3 degrees of freedom.

```
[i] ? $look 8 %fv$
[o]      %FV
[o] 8    1009.
```

Here we can check the estimated value (the 8th fitted value)

The estimated *hidden* population would be 1,009 if this model was the best fitting

```
[i] ? $d r$
[o]   unit  observed   fitted  residual
[o]     1      6       6.827   -0.317
[o]     2     34     40.653   -1.043
[o]     3     79     66.219    1.571
[o]     4    389    394.301   -0.267
[o]     5     13     17.474   -1.070
[o]     6    116    104.046    1.172
[o]     7    162    169.480   -0.575
[o]    (8)      0   1009.168    0.000
[o]
```

Alternatively we can display the residual values

Again this shows the size of the hidden population

```
[i] ? $fit +i1$
```

Here we include the interaction between sources 1 and 2

```
[o] scaled deviance = 4.7612 (change = -1.807) at cycle 3
[o]   residual df = 2      (change = -1      ) from 7 observations
```

We can see how including the interaction slightly improves the fit

```
[i] ? $fit -i1$
```

We can take this interaction back out again to return to the basic model

```
[o] scaled deviance = 6.5681 (change = +1.807) at cycle 3
[o]   residual df = 3      (change = +1      ) from 7 observations
[o]
```

```
[i] ? $fit p1+p2+p3+i1$      Alternatively this command could have been used to go direct to the model with the
                               interaction
```

```
[o] scaled deviance = 4.7612 at cycle 3
[o]   residual df = 2      from 7 observations
[o]
```

We can include / take out the different interactions as follows. We have only displayed the size of the estimate for the model that we previously know is the best fitting.

```
[i] ? $fit p1+p2+p3$
[o] scaled deviance = 6.5681 at cycle 3
[o]   residual df = 3      from 7 observations
[o]
[i] ? $fit +i2$
[o] scaled deviance = 3.9020 (change = -2.666) at cycle 3
[o]   residual df = 2      (change = -1      ) from 7 observations
[o]
[i] ? $fit -i2$
[o] scaled deviance = 6.5681 (change = +2.666) at cycle 3
[o]   residual df = 3      (change = +1      ) from 7 observations
[o]
[i] ? $fit +i13$
[o] scaled deviance = 0.13241 (change = -6.436) at cycle 3
[o]   residual df = 2      (change = -1      ) from 7 observations
[o]
[i] ? $fit -i13$
```

```

[o] scaled deviance = 6.5681 (change = +6.436) at cycle 3
[o]   residual df = 3      (change = +1    ) from 7 observations
[o]
[i] ? $fit +i1+i2$
[o] scaled deviance = 0.69978 (change = -5.868) at cycle 3
[o]   residual df = 1      (change = -2    ) from 7 observations
[o]
[i] ? $fit -i2+i13$
[o] scaled deviance = 0.011611 (change = -0.6882) at cycle 3
[o]   residual df = 1      (change = 0     ) from 7 observations
[o]
[i] ? $fit -i1+i2$
[o] scaled deviance = 0.096012 (change = +0.08440) at cycle 3
[o]   residual df = 1      (change = 0     ) from 7 observations
[o]
[i] ? $fit p1+p2+p3$
[o] scaled deviance = 6.5681 at cycle 3
[o]   residual df = 3      from 7 observations
[o]

[i] ? $fit p1+p2+p3+i13$
[o] scaled deviance = 0.13241 at cycle 3
[o]   residual df = 2      from 7 observations
[o]

```

```

[i] ? $look 8 %fv$           As this is the model that is judged to be the best fitting we again look at the estimate
[o]   %FV
[o] 8   1379.

```

```

[i] ? $d r$
[o]   unit   observed   fitted   residual
[o]   1       6         6.596   -0.232

```



```
[o]      2      34    32.824    0.205
[o]      3      79    78.404    0.067
[o]      4     389   390.176   -0.060
[o]      5      13    13.580   -0.157
[o]      6     116   116.000    0.000
[o]      7     162   161.420    0.046
[o]     (8)       0  1378.868    0.000
[o]
```

```
[i] ? $stop
```

We finish the GLIM session.

Appendix II Typical SPSS session which provides an estimate of the size of a hidden population using data from 3 sources.

* * * * * H I E R A R C H I C A L L O G L I N E A R * * * * *

DATA Information

Information about the
analysis

8 unweighted cases accepted.
0 cases rejected because of out-of-range factor values.
0 cases rejected because of missing data.
799 weighted cases will be used in the analysis.

FACTOR Information

Factor	Level	Label
P1	2	
P2	2	
P3	2	

* * * * * H I E R A R C H I C A L L O G L I N E A R * * * * *

Backward Elimination (p = .050) for DESIGN 1 with generating class

The model with all two-way interactions is fitted

P1*P2
P1*P3
P2*P3

Likelihood ratio chi square = .20682 DF = 1 P = .649

If Deleted Simple Effect is	DF	L.R.	Chisq	Change	Prob	Iter
P1*P2	1			.000	1.0000	20

Step 1

The best model has generating class

The interaction p1*p2 is removed and the change in χ^2 is found

P1*P3
P2*P3

Likelihood ratio chi square = .10429 DF = 2 P = .949

If Deleted Simple Effect is	DF	L.R.	Chisq	Change	Prob	Iter
P1*P3	1			3.798	.0513	10
P2*P3	1			.029	.8644	12

Step 2

Other interactions are
examined

The best model has generating class

P1*P3
P2

Likelihood ratio chi square = .13346 DF = 3 P = .988

If Deleted Simple Effect is	DF	L.R. Chisq Change	Prob	Iter
P1*P3	1	6.435	.0112	10
P2	1	574.104	.0000	2

—

* * * * * H I E R A R C H I C A L L O G L I N E A R * * * * *

Step 3

The best model has generating class

P1*P3
P2

Likelihood ratio chi square = .13346 DF = 3 P = .988

* * * * * H I E R A R C H I C A L L O G L I N E A R * * * * *

The final model has generating class

The most
parsimonious model
is displayed

P1*P3
P2

The Iterative Proportional Fit algorithm converged at iteration 0.
The maximum difference between observed and fitted marginal totals is .320
and the convergence criterion is .389

Observed, Expected Frequencies and Residuals.

Factor	Code	OBS count	EXP count	Residual	Std Resid
P1	1				
P2	1				

P3	1	6.0	6.6	-.62	-.24
P3	2	13.0	13.6	-.63	-.17
P2	2				
P3	1	79.0	78.4	.60	.07
P3	2	162.0	161.4	.58	.05
P1	2				
P2	1				
P3	1	34.0	32.9	1.05	.18
P3	2	116.0	115.8	.20	.02
P2	2				
P3	1	389.0	390.2	-1.18	-.06
P3	2	.0	.0	.00	.00

Goodness-of-fit test statistics

Likelihood ratio chi square = .13346 DF (UNADJUSTED) = 3 P = .988
DF (ADJUSTED) = 2 P = .935
Pearson chi square = .13152 DF (UNADJUSTED) = 3 P = .988
DF (ADJUSTED) = 2 P = .936

—

>Note # 13864
>UNADJUSTED DF have NOT been adjusted for structural or sampling zeroes.
>For ADJUSTED DF one degree of freedom is subtracted for each cell with an
>expected value of zero. The UNADJUSTED DF are an upper bound on the true
>DF, while the ADJUSTED DF may be an underestimate.

—

GENERAL LOGLINEAR ANALYSIS

Data Information

Information about the
analysis

8 cases are accepted.
0 cases are rejected because of missing data.
799 weighted cases will be used in the analysis.
8 cells are defined.
1 structural zeros are imposed by design.
0 sampling zeros are encountered.

Variable Information

Factor	Levels	Value
P1	2	
		1.00
		2.00

P2	2	1.00
		2.00
P3	2	1.00
		2.00

Model and Design Information

Model: Poisson
Design: Constant + P1 + P2 + P3 + P1*P3

Correspondence Between Parameters and Terms of the Design

Parameter	Aliased	Term
1		Constant
2		[P1 = 1.00]
3	x	[P1 = 2.00]
4		[P2 = 1.00]
5	x	[P2 = 2.00]
6		[P3 = 1.00]
7	x	[P3 = 2.00]
8		[P1 = 1.00]*[P3 = 1.00]
9	x	[P1 = 1.00]*[P3 = 2.00]
10	x	[P1 = 2.00]*[P3 = 1.00]
11	x	[P1 = 2.00]*[P3 = 2.00]

-

Note: 'x' indicates an aliased (or a redundant) parameter.
These parameters are set to zero.

- - - - -

Convergence Information

Maximum number of iterations: 20
Relative difference tolerance: .001
Final relative difference: .0008

Maximum likelihood estimation converged at iteration 2.

- - - - -

Table Information

Factor	Value	Observed Count	%	Expected Count	%	Observed and expected values in the contingency table
P1	1.00					
P2	1.00					
P3	1.00	6.00	(.75)	6.60	(.83)	
P3	2.00	13.00	(1.63)	13.58	(1.70)	
P2	2.00					
P3	1.00	79.00	(9.89)	78.40	(9.81)	
P3	2.00	162.00	(20.28)	161.42	(20.20)	

P1	2.00		
P2	1.00		
P3	1.00	34.00 (4.26)	32.82 (4.11)
P3	2.00	116.00 (14.52)	116.00 (14.52)
P2	2.00		
P3	1.00	389.00 (48.69)	390.18 (48.83)
P3	2.00	.00 (.00)	.00 (.00)

—

 GENERAL LOGLINEAR ANALYSIS

Table Information

Factor	Value	Resid.	Adj. Resid.	Dev. Resid.
P1	1.00			
P2	1.00			
P3	1.00	-.60	-.26	-.24
P3	2.00	-.58	-.19	-.16
P2	2.00			
P3	1.00	.60	.26	.07
P3	2.00	.58	.19	.05
P1	2.00			
P2	1.00			
P3	1.00	1.18	.35	.20

P3	2.00	.00	.00	.00
P2	2.00			
P3	1.00	-1.18	-.35	-.06
P3	2.00	.	.	.

Goodness-of-fit Statistics

	Chi-Square	DF	Sig.
Likelihood Ratio	.1324	2	.9359
Pearson	.1309	2	.9367

GENERAL LOGLINEAR ANALYSIS

Parameter Estimates

Parameter	Estimate	SE	Z-value	Asymptotic 95% CI	
				Lower	Upper
1	7.2290	.1705	42.41	6.89	7.56
2	-2.1450	.1781	-12.04	-2.49	-1.80
3	.0000
4	-2.4754	.1430	-17.31	-2.76	-2.20
5	.0000
6	-1.2624	.1684	-7.49	-1.59	-.93
7	.0000

← This estimate will be the log of the size of the hidden population

8	.5403	.2141	2.52	.12	.96
9	.0000
10	.0000
11	.0000

Covariance Matrix of Parameter Estimates

Parameter	1	2	4	6	8
1	.0291				
2	-.0275	.0317			
4	-.0204	.0189	.0204		
6	-.0275	.0260	.0189	.0284	
8	.0275	-.0317	-.0189	-.0284	.0459

Aliased parameters are not shown.

Correlation Matrix of Parameter Estimates

Parameter	1	2	4	6	8
1	1.0000				
2	-.9048	1.0000			
4	-.8387	.7404	1.0000		
6	-.9568	.8669	.7829	1.0000	
8	.7526	-.8318	-.6158	-.7866	1.0000

Aliased parameters are not show