European Monitoring Centre
for Drugs and Drug Addiction

UNIVERSITA' DEGLI STUDI DI ROMA
TorVergata

Guidelines for Estimating the Incidence of Problem Drug Use

EMCDDA, February 2008

European Monitoring Centre for Drugs and Drug Addiction
Rua da Cruz de Santa Apolónia 23–25
PT-1149-045 Lisboa
Portugal
Tel: +351 21 811 30 00
Fax: +351 21 813 17 11
e-mail: info@emcdda.europa.eu
http://www.emcdda.europa.eu

Please use the following citation:

Scalia Tomba GP, Rossi C, Taylor C, Klempova D, Wiessing L. (2008) Guidelines for Estimating the Incidence of Problem Drug Use. EMCDDA, Lisbon.

Acknowledgements

Maria Grazia Calvani, Emanuela Colasante, Flavia Lombardo, Lucilla Ravà, Antonia Domingo Salvany, Siem Heisterkamp, Matthew Hickman, Ellen Amundsen, Sharon Arpa, Clive Richardson and Albert Sanchez Niubo have contributed to this work.

These Guidelines are the result of the EMCDDA financed project 'Project to stimulate the implementation of the EMCDDA guidelines on estimating incidence of injecting and problem drug use in the EU' (CT.06.EPI.150.1.0) and previous related projects. See Appendix 6 for a detailed description of work responsibilities and acknowledgements. This version of the Guidelines has not been formally copy edited.

1. Contents

2. Summary and overview

The present version of the Incidence Guidelines builds on previous versions of these Guidelines and on the many comments and suggestions received about useful contents and explanations. The main aim is to bring together the state of the art of drug use incidence estimation methods with explanations of these methods and related concepts, both in a non-technical style, for the beginning user, and with sufficient details for an effective use, where possible. The basic ideas that underlie estimation methods for drug use incidence based on indirect data, such as drug treatment admissions or overdose related deaths are presented in the form of statistical models for observations, a useful way to link a representation of the phenomenon of interest in terms of descriptive parameters with available observations.

In the present case, one main ingredient of the statistical models is the explicit consideration of the time delay between start of drug use and, in a proportion of cases, becoming a "known" drug user, e.g. by seeking treatment. This period of time is called the latency period (LP) and much attention is devoted to this concept and to the effects of delayed observation on inference.

A second important ingredient is understanding the role of observation schemes on the meaning and/or interpretation of observations. This effect can be seen in many circumstances. As an example, suppose that a point event occurs (e.g. exposure to an infective agent) and several individuals are affected, but that the effect only becomes apparent after a variable period of time (the incubation time, in this case). Then, if average incubation time is calculated the day after the exposure, on subjects falling ill between exposure and the first day after exposure, this average must be shorter than one day, because all the observed individuals have manifested incubation times shorter than one day. If the same calculation is performed on all data available after the second day, the average will certainly be longer but still shorter than two days. Only after all the exposed individuals have fallen ill, can a "correct" average incubation time be calculated. This shows that one cannot say that any average of observed incubation times gives a good estimate of average incubation time, but that the observation scheme must be considered before interpreting the observed data. It is usual to consider the estimates obtained during the first days as "truncated" or "biased", since they do not represent the quantity of interest, but it is often enough to understand that there is an observation scheme effect in order to be able to extract some useful, correct, information from such observations, even if the main quantity of interest cannot be estimated.

There are two principal estimation methods presented in these Guidelines, the Reporting Delay Adjustment method (RDA) and the Back-calculation method (BC). These methods are not alternatives to each other. They each have a certain range of applicability, depending on available data. Once again, the explanations are mostly meant to give some basic understanding of how these methods work and what their requirements, limitations and possibilities are as an introduction to start applying them. Some more technical subject matter is presented in the Appendices and for advanced details the reader is referred to the key scientific literature.

There is also a separate section devoted to the analysis of survival time data. This is a rich and complex field in Statistics and some basic knowledge about it is essential, since durations of various kinds (latency period, duration of drug use, etc) are important components of the world of ideas and concepts related to drug use incidence.

Some ideas about possible extensions of the presented methods to new kinds of data are also presented, as well as some ideas about possible new methods. The purpose of such excursions is to familiarize the non-statistical reader with the possibilities of method development, to perhaps inspire the statistically inclined reader to try something new and, last but not least, to convince the two kinds of readers that collaboration could be fruitful.

3. Introduction

The incidence of first drug use is an important epidemiological concept that helps understanding the diffusion of drug use in space and time. Incidence estimates can be used to evaluate current and future needs for, and effects of, services and interventions. In particular, incidence figures may provide an indication of whether the number of problem drug users is growing (epidemic phase), stable (endemic phase) or falling.

Incidence is defined as the number of new cases occurring within a given time period, usually a year. Compared to prevalence, which is defined as the number of existing cases at a certain point in time (or, for practical reasons, in a given period; in studies using administrative data this is often a calendar year), incidence gives more direct information on the recruitment of new cases. It is therefore usually a more sensitive indicator of prevention efforts than prevalence. Incidence is directly related to prevalence: adding up the incidence over a number of years, and subtracting the number of cases that have died or ceased to be cases for other reasons, gives the change in prevalence over the same time period.

It is acknowledged that accurate information on the incidence and prevalence of drug use and especially that of heroin and other opioid use is difficult to obtain. Evidence from national surveys and other sources indicates that the prevalence of heroin use in the general population is relatively low and has shown a decreasing trend in most European Union (EU) countries, although some new increases are being reported. Despite this low prevalence, most of the widespread drug-related health and social problems in EU countries are caused by the use of heroin and other opioids, although there is evidence that amphetamines, new synthetic drugs and in particular cocaine and are becoming increasingly important. The substances causing health and social problems as well as the route of administration of these substances vary across Europe. In Sweden, for example, the drugs causing most of the problems are amphetamines, whereas in Norway and Finland heroin and other opioids have been challenging the drug help system. The most common practice of using heroin in the Netherlands, Spain and Portugal is smoking ("chasing the dragon") whereas in Slovenia, Finland and the Czech Republic injecting drug use is still the common way of administration (see EMCDDA (2007a)).

These differences in substances and routes of administration make a common definition of the target group rather difficult, since different substances and practices of use are related to different health problems and subgroups of society. Nevertheless, comparisons across countries with regard to the trend of drug use call for a common definition of the target group, the use of equivalent data sources, and the application of the same methodology. In an attempt to find a common definition in spite of the differences between the EU countries we use the term problem drug use (PDU), which includes all different forms of problems due to the use of opioids, cocaine, and amphetamines irrespective of the route of administration.

The EMCDDA definition of PDU is ‘Injecting drug use or long duration/regular use of opioids, cocaine and/or amphetamines’. This includes all legal or illegal use of opioids with the exception of pain relief medication, all use of cocaine (both crack and powder cocaine) and all (meth-)amphetamines with the exception of ecstacy. Despite the fact that heavy use of cannabis can lead to health or social problems, the proportion of all cannabis use that is likely associated with health or social problems of similar severity as problematic heroin or cocaine/amphetamine use is likely so small that cannabis use is not included in the EMCDDA definition of PDU. (Despite this, cannabis users are highly prevalent in European drug treatment entry data due to the very high prevalence of cannabis use in the population, as well as - perhaps - their shorter and more frequently repeated treatments in comparison to for example opioid substitution clients, who may remain on treatment for many years without exiting and being registered at re-entering).

These guidelines describe statistical methods to estimate incidence from observed and observable data. They also try to explain the utility of formulating (statistical) models for observations, both as a way to gain better understanding of the assumptions that underlie the interpretation of data (and thus the knowledge gained from data) and as a way to find the most suitable statistical procedures for data analysis. Until now these methods have mostly been used to estimate incidence of first heroin use from drug treatment data only, however other sources can be used and incidence of first use of other drugs can also be estimated. In the following we will talk about estimating the incidence of drug use (DU) rather than of problem drug use (PDU), because even in the case of estimating first use of problem drug use that first use itself is usually not problematic. Thus, these methods can be used to estimate the incidence of that part of all drug use that will ultimately become problem drug use, which by that time may become visible by for example entry in drug treatment. It is highly relevant to estimate this part of the total incidence of first use of a drug, as it provides an early indicator of the bulk of the related future problems and (heath care or other) costs to society.

When a phenomenon is well defined and observable with good accuracy, there is usually little need for complicated statistical methods; data can be represented and summarized in a natural way to yield the desired insight into the phenomenon under study. Unfortunately, direct observation of incidence of drug use is very uncommon, and we will therefore have to rely on statistical methods and models to recover information about incidence from other types of observed data.

As will be explained, statistical methods allow the extraction of information from observations, but require various assumptions in order to work well. The validity of the conclusions does not only depend on the statistical method used, but also on the quality of data and on how well the assumptions fit reality. It is important to realise that to derive suitable information from observed data it is always necessary to know how they were collected and under what limitations. It is also very important to recognize and understand the limitations of the conclusions that can be drawn using a given statistical method with a given data set. At this point, we will only give one example, related to the concept of "relative incidence", introduced by Hickman et al (2001) to denote a fraction of total incidence relative only to individuals who fulfill a certain condition, in the specific case having sought treatment within 8 years of the start of drug use.

As we shall see, some methods of estimation are based on data from drug users who seek drug treatment for the first time (incidence of treatment). From these users,

information about when the drug use started is obtained and, by statistical means, the numbers of individuals who have started drug use but not yet sought treatment are estimated (the so called Reporting Delay Adjustment method). By summing the numbers of individuals in the data and the estimated numbers, we get an estimate of the numbers of individuals, per year, who have started to use drugs and who will end up seeking treatment. This is, however, not the total incidence of drug use, just that part that eventually seeks treatment. Without some external source of information on what percentage of all those who start using drugs problematically will eventually seek treatment, we will not be able to say anything about how many individuals have started using drugs but will never seek treatment (the reasons for not eventually seeking treatment may e.g. be death, spontaneous cessation of drug use or ability to seek treatment outside of the reporting system). In this case, we say that we have been able to estimate only the 'relative incidence', (in this context, the word 'relative' is seen as opposed to 'absolute' or 'total' incidence, which would include the cases of first use of a drug that will never progress to problem drug use, but which is impossible to estimate using these methods). I.e. relative incidence is the incidence of cases that will eventually (up to many years later) be observed (e.g. in drug treatment). This is thus an underestimate of, and gives a lower bound for, total incidence of first drug use. In addition, we usually assume that the relative incidence will indicate whether the total number of new cases is rising, stable or falling between years. Under this assumption, that relative incidence follows the trends in total incidence, relative incidence can be considered a very important indicator of the trend in the rate of all new cases of first use of a drug.

In addition to the statistical methods described, for which a certain level of statistical expertise and data quality/availability are needed, one might at least obtain some global impression of incidence (mainly whether it is increasing or not), by looking at simple descriptions of and parameters derived from observations e.g. treatment data. These include: a) comparing the age distribution of treatment cases at a certain moment in time with the age distribution of non-cases, or following trends in the age distribution of cases over time, b) comparing the proportion of first registrations (e.g. first treatment demands) among all registrations, or following trends of first registrations over time (see EMCDDA treatment demand indicator). Knowing that the average duration of problem drug use is at least more than 5 and possibly around 8 or even more years, one can easily deduce a rising epidemic of for example heroin use if these parameters are sufficiently "extreme". If treatment data, for example, contain 50% first treatment demands, or when half of the treatment cases are under age 25, then it is likely that there have been relatively high rates (incidence) of recent initiation. When results are not very marked however it is much better to correctly estimate the curve of relative incidence using the more sophisticated methods presented in these guidelines.

4. Definition of the target and the study population

In any estimation exercise, it is extremely important to be clear about what is being estimated, and what is not. It may be useful to distinguish between the "target population" (that part of the population whose size or other characteristics we wish to estimate) and the "study population" (what we eventually are able to estimate). The precise definition of the population targeted by any estimation procedure in the drug field is a difficult task.. A useful and operational target group characterization can be obtained by refering to the EMCDDA 2007 definition of "problem drug use" (PDU):

The EMCDDA defines problem drug use as 'Injecting drug use or long duration/regular use of opioids, cocaine and/or amphetamines' (EMCDDA 2007b).

Components of a detailed definition

Definitions of target group may combine a certain time period (e.g., a certain year), a specific substance group (e.g., opioids, amphetamines), the route of administration (e.g., injecting, smoking), frequency of use (e.g., experimental, occasional, habitual, regular, long duration), legal status (illicit, licit), and clinical diagnoses (dependence, abuse). As the utilised databases, e.g. treatment monitoring systems, usually report data for a calendar year it is natural to use this time frame also for the incidence estimation of problem drug use. Even when referring to the broadest possible target group, the "drug users", any definition should be confronted with the commonly used PDU definition.

Furthermore, there is often a geographical/administrative delimitation of the target group, such as "in the whole country" (national), "in a given region" or "in a given city" (sub-national). Care must then be exercised to discuss the problems with such definitions. Since the phenomenon of interest, drug use, is distributed over time, the locations of first use, present use and main part of a drug use career may be distinct and, even if known, a decision would be required as to which one of these is the one of interest. There may be discrepancies between the officially available individual data on location, such as place of birth, place of residence, and effective location. The definition of location is often coupled with the concept of nationality of individuals. In general, these problems decrease with the size of the concerned area. If "whole country" statistics are considered, only the possible immigration/emigration flows need be considered. These considerations about people, space and time are to some extent related to the concept of "closed population", often encountered in connection with e.g. capture-recapture methods. It may be reasonable to consider a given population as "closed", i.e. without changes (births, deaths, movements, entering, leaving, immigration, emigration...) and therefore consisting of the same individuals at two different time points, for a short interval of time. In a longer time perspective, either the changing nature of the population is directly acknowledged in the method of analysis of observations or else the effects of changes on the chosen method of study must be understood.

Usually, data is derived from the 'institutionally visible' population, i.e., only when an individual comes into contact with the legal, medical or social system do we know that he or she is a user. Any definition of the target population should therefore consider

the different interests, norms and values of these three systems because these may determine what part of the drug user population they are in contact with.

It would also possible to gather data from e.g street interviews or other "non institutional" settings, but questions about representativity (with respect to a larger PDU population) and reliability (of answers given in an essentially non controlled environment) should then be adressed before interpreting such data.

Pragmatic Definition of Main Substance

Not only does substance use vary between countries, birth cohorts, and even gender, also route of administration differs greatly between substances, countries and cohorts. Even if route of administration and frequency of use could clearly be related to a more or less hazardous consumption pattern, this information may not be directly available. Furthermore, substances are used in quite mixed, often chaotic patterns. Only very few opioid users do not use other drugs as well. Although in the past problem drug use was almost synomymous with heroin use and injecting, in more recent years the patterns of use of PDU have diversified and it is important to acknowledge the (sometimes huge) overlaps between users of different drugs and to recognise more clearly that PDU is currently much more than only heroin or opioid use. (Note that the EMCDDA PDU definition has not changed and has always included heroin, cocaine and amphetamines however the relative importance between these drugs has changed during the 1990's.).

A pragmatic definition for coding according to the concept of allowing for overlaps between users of different drugs could be summarised as follows:

If a user is using opioids he/she should be included in any estimate of opioid use, regardless of what other drug he/she uses. If the user uses cocaine then he/she should be included in any cocaine estimates regardless of other drugs used. Similarly for amphetamines. The consequence is that different incidence estimations can be made resulting in overlapping populations, but having the advantage that each estimate does completely describe the population using a certain drug. Obviously care should be taken that if sub-populations clearly have different contact-patterns with treatment they are not lumped together but estimated separately, e.g. this could be the case for crack-cocaine users versus powder cocaine users. Moreover, some of these groups should not be estimated at all if there are reasons to believe that the proportion eventually contacting treatment is very low, and/or the latency period distribution very long in relation to the observed data.

All possible groups of problem drug users are admittedly not covered by this definition (e.g. problematic users of cannabis or new synthetic drugs or injectors of substances different from those considered above), but it is not at present clear whether the methods discussed in these Guidelines would be appropriate for these groups. Although data on first drug treatment, used as a basis for many applications of the methods illustrated in this report, includes many other users than opioid users, for some substances it is not clear if a sufficient proportion of users eventually develop problems sufficient to seek treatment (e.g. cannabis users, powder cocaine users) or if the latency period is perhaps extremely long and thus not suitable for adjusting the observed data.

This logic of categorising patterns of drug use is summarised in Table 1 below.

Table 1: A possible classification of problem drug use categories by substance

Groups	Opioids	Cocaine	Amphetamines
Problem Opioid user	Yes	Yes/No	Yes/No
Problem Cocaine user	Yes/No	Yes	Yes/No
Problem Amphetamine user	Yes/No	Yes/No	Yes

Apart from defining the target group by substance used, is important, for the evaluation of related health risks, to specify, in all three situations above, whether the substances are injected or used in another way (smoking, snorting, chasing the dragon, eating etc.) and whether their use can fall under the 'problem drug use' definition – e.g. if they are not injected, then they have to be used long-term / regularly or at least this has to be assumed by the mere presence of the individual in the dataset (e.g. treatment). In addition it is important to derive overall estimates of the incidence of injecting drug use across different drugs. It is further important that the definition of the target group (e.g. incidence of heroin use) matches exactly the definition in the observed data (e.g. 'heroin users entering their first treatment') and that any latency period estimation to arrive at estimates is based on year of first heroin use and entry in treatment for heroin use. E.g. it is not acceptable to use 'year of first drug use' as a basis to estimate latency period in the estimation of incidence of heroin use, although it might be acceptable (provided this is made explicit) to alternatively use 'year of first heroin use' or 'year of first regular heroin use' depending on the data available.

When deciding on the definition of a target group special care should be taken with so-called 'time dependent variables' i.e. variables that may change over time. For example a heroin user might start smoking heroin but become visible in the drug treatment system as an injecting drug user. Latency period analysis by such variables might give artifactual results, e.g. it might seem that heroin users who smoke have a much shorter latency period than heroin injectors (suggesting a much faster progression to treatment entry) whereas in reality one is observing the effect that heroin users with a longer LP have had more time to switch to injecting. (In addition there may be calendar time or cohort effects in the population, e.g. smoking may have become more popular, that may complicate things further and have to be taken into account when interpreting results).

Conclusion

In summary:

- an estimation exercise should have a clearly defined target population;

- it is important to consider to what extent and under what assumptions available data, analysed with a certain statistical method, is able to give information about the chosen target group or, at least, in what way the studied population differs from the target population.

5. Data Sources

There are essentially two institutional systems capable of providing data for incidence estimation.. These are a) the clinical, medical and social system and b) the legal system. Even though the applications are, at present, only based on data coming from the first, characteristics of and problems associated with both are briefly discussed in this section.

Clinical, Medical, and Social System

The clinical, medical and social systems generally collect the most detailed information on drug users that come into contact with these systems. A drug user can come into contact with:

- Drug treatment agencies (inpatient versus outpatient, specialised on drug care versus general treatment agencies; e.g., drug counselling centres, general counselling centres, psychiatric hospitals, and specialised hospitals),
- Low threshold agencies (e.g., needle and syringe exchange and distribution programmes, drop-in centres),
- Substitution services (dependent on the regulation in each country these may be general practitioners, substitution ambulances, hospitals, treatment agencies),
- General practitioners (medical reasons),
- Emergency ambulances (mobile or stationary),
- HIV/hepatitis related services,
- Clinical psychologists (other than those working in the above services),
- Psychiatrists (other than those working in the above services).

In general a data source can be useful for incidence estimation using the methods described in these guidelines if:

1. It covers a reasonable number of calendar years (e.g. at least 8) as compared to the median latency period of the specific substance (for heroin use often between 4 and 6) in order to minimise left and right truncation problems
2. It reasonably well distinguishes first entry to the system from any repeat entry (e.g. it should be able to distinguish first ever drug treatments). If first entries are not distinguished but year of first entry is available then the data may still be used but analysis might be more difficult.
3. It provides individual data records for each treatment client which contain information on year of first use, or, it provides aggregate numbers of users showing up in treatment in each calendar year provided valid external information is available on the latency period distribution.

Problems associated with the above listed data sources are:

- in general, treatment monitoring systems do not cover all treatment facilities of a country. Moreover, the treatment facilities represented in the treatment monitoring system may differ from one country to another or within the same country in different areas.

- agencies usually collect their data for internal use, and in rather few countries a general treatment monitoring system covering most of the treatment centres has been established. Furthermore, double counting cannot be excluded, as many drug users will come into contact with a variety of treatment facilities several times each year. Due to privacy laws, utilising unique personal identifiers to prevent double counting is difficult in some European countries.

- drug users in urban areas have more options for receiving treatment in comparison to rural areas. On the other hand, the 'latency period' (time from first use to first treatment) seems to be shorter in less urbanised or rural areas which might be related to stronger social cohesion/control in those areas(EMCDDA (2000)). Therefore, the probability of coming into contact with treatment facilities during a given period may not be constant nationwide.

- some treatment agencies are interrelated and send patients to certain other centres. The capacity of treatment facilities is limited, drug users may be put on waiting lists.

- drug users, after contacting a treatment facility, may break off contact before any data can be collected or they may leave incomplete, unreliable or even wrong data.

- there exists a population of drug users not covered by the drug-related social and medical system, consisting both of drug users who have not yet progressed to problems severe enough to come into contact with drug services but are likely to do so later, and perhaps a group of drug users who may never contact these services.

- it is very likely that every drug user sooner or later comes into contact with, for example, a general practitioner, but this rather seldom utilised data base has manifold problems. It is not clear how many drug users will be recognised while visiting a doctor for medical reasons not obviously related to drug-use. Furthermore, in case of a non-fatal accident under influence of psychotropic substances it is almost impossible to distinguish between regular, occasional or experimental users.

Legal System

Data of drug users can be found in registers

- on convictions because of offences against laws on consumption, possession or supply of illegal drugs;

- on convictions because of secondary crimes, e.g. offences associated with obtaining money for the further acquisition of drugs (theft, shoplifting, prostitution, forgery);

- on convictions because of offences under the influence of psychotropic substances (e.g. driving, violence);

- on detainees in connection with the above mentioned categories;

- on mortality (e.g., all-cause deaths of registered drug users, drug-related deaths).

Problems associated with these data sources are:

- it may be difficult to distinguish between consumers and non-consuming dealers of drugs; on the other hand, some registers explicitly state whether individuals were under the influence of drugs in a certain occasion (driving, death, etc...).

- route of administration or frequency of drug use cannot easily be obtained, in order to separate regular from experimental users. Police records do usually not distinguish between minor experimentation with drugs, severe drug problems, mere drug dealing without consumption, and long-term or regular users. However, perhaps it can be assumed that in most cases these data will relate to regular users as these will have a much higher chance of getting caught by the police.

- secondary crimes may not be registered at all, and, depending on the drug policy of a country, no special attention may be given to the connection between the primary crime and drug use. Furthermore, legal activism regarding certain drugs may vary between countries and in time, for example police activity regarding the simple use of drugs (as opposed to dealing) differs between countries and, in some cases, even in the same country, between time periods.

- age at first use is not usually available.

6 .1 Introduction

Data exploration before actual incidence estimation: Qualitative analysis of age, first treatment, other indicators

This section presents ideas useful for exploration of data with respect to possible trends in incidence. The purpose of this data exploration is to formulate hypotheses about incidence trends, before the actual estimation, but also to visualize the effects of incidence trends on various observables, which might be of help in understanding, and interpreting the results of incidence estimation. It is also possible to skip this analysis and proceed directly to actual incidence estimation or, on the other hand, if no data are available for methods actually estimating incidence, use it as the only approach with all the limitations considered in interpretation of results.

Statistical Background

Important qualitative features of drug use incidence, such as steady decline or recent strong increase, may be inferred from time trends of epidemiological indicators based on secondary data, under usually reasonable assumptions. Secondary data can be defined as existing statistical and documentary information that is routinely collected, such as first treatment presentations or simply treatment episodes, drug seizures, infectious diseases indicators, or drug related deaths. Standard statistical summaries (statistical distributions, time series graphs and tables…) can be utilized to present the available information on time trends. However, it may be instructive to consult the Example box in Section 6.4, to see how the time trend in an assumed drug use onset incidence curve is modified by the latency period to yield a treatment incidence curve that seems to behave differently (in this particular case, because its full appearance over time cannot be observed) .

As an example, the observation that the population of drug users asking for treatment (either for the first time or in general), or being imprisoned, is ageing may be taken as evidence of a decreasing incidence of problem drug use, although, from a purely logical point of view, such a trend could also be the result of a (sudden) increased interest among older individuals for seeking treatment or committing crimes or, inversely, a decreased interest among younger users. Thus, an assumption of constant behaviour among drug users is necessary to motivate the desired conclusion. The conclusion can be given more likelihood if several sources of data show compatible trends. An example of this kind of analysis based on observable data can better clarify the approach.

Example: Description and interpretation of time trends of opioid users in treatment and police stations in Amsterdam 1985-1997

Data from the methadone register show a decreasing number of clients and an ageing treatment population. These time-trends are not limited to the treatment population, since data on methadone prescription to arrested opioid users at police stations show

similar trends. There is also an analysis of the number of arrested opioid users that shows a decreasing trend.

Figure 6.1.1 shows that the annual number of treatment participants of the methadone programmes in Amsterdam is decreasing from approximately 4900 in 1985 to 3100 in 1997. The average number of opioid users in treatment each week does not change, varying around 2000 during the whole period. This combination may suggest a decreasing prevalence of users combined with increasingly long average treatment periods (2000 in treatment, on average, each week means that 52x2000 = 104000 "treatment weeks" have been administered during the year and that the average number of treatment weeks/year per individual registered during a given year has increased from 104000/4900 = 21 in 1985 to 104000/3100 = 34 in 1997.

Figure 6.1.1: Number of opioid users in methadone treatment

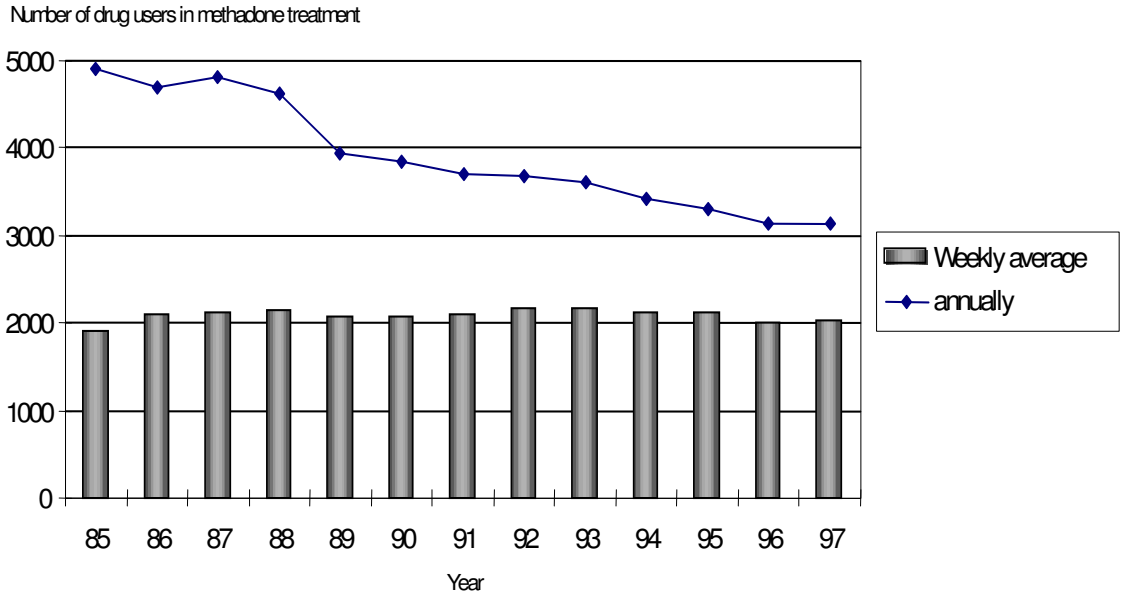
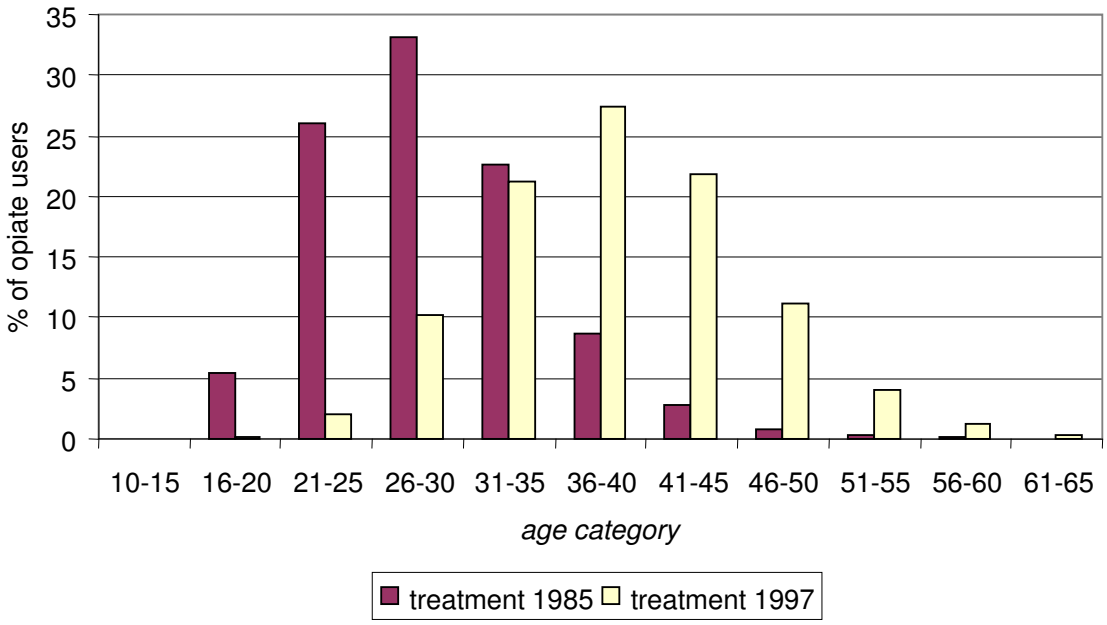


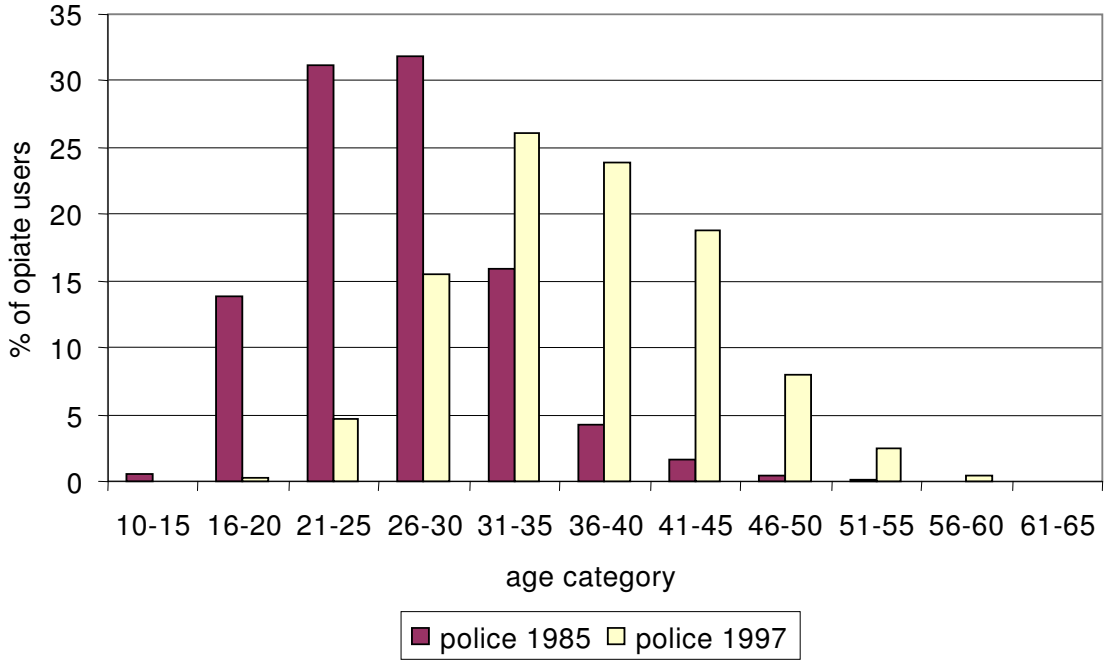
Figure 6.1.2 shows the age distribution of the treatment population in 1985 and 1997. It shows a clearly ageing population. In 1985 the majority (65%) was thirty years or younger, in 1997 the corresponding proportion was only 12.5%.

Figure 6.1.2: Age distribution of participants of methadone treatment



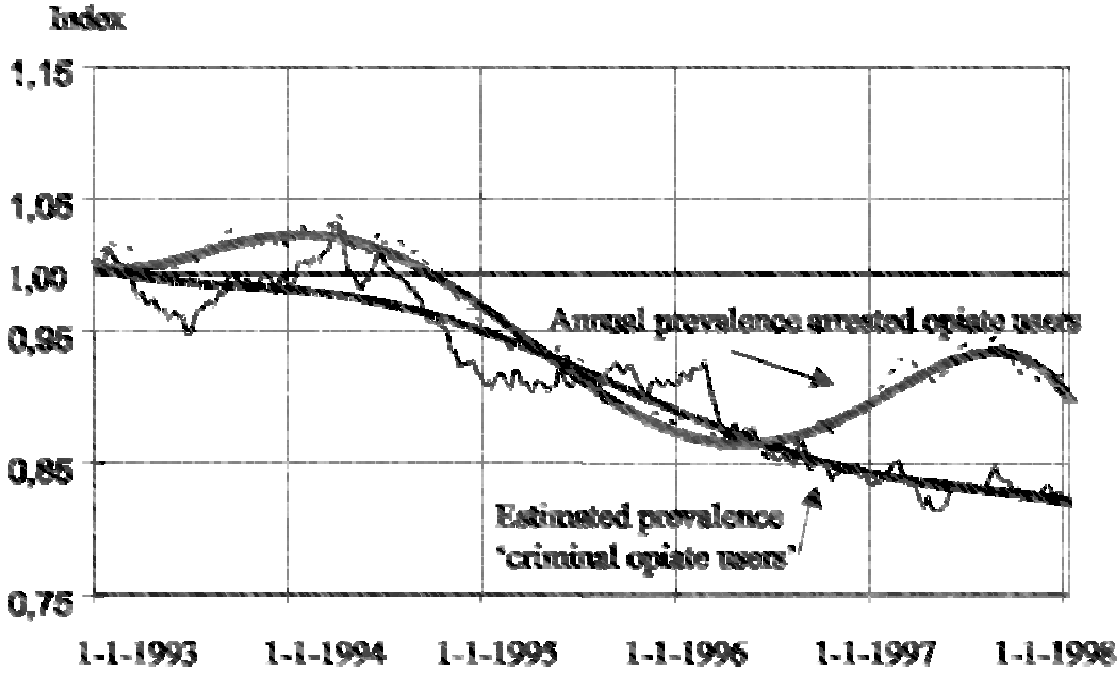
The decreasing number of opioid users in treatment and the ageing population could indicate that the incidence of opioid use is decreasing during the period of interest (thus the users will more and more consist of individuals who have started use "a long time ago" and who are now progressively ageing) but could also indicate that "new" and thus younger opiate users tend to avoid methadone programmes Therefore data of methadone prescriptions at police stations are studied. Methadone is prescribed at the police stations to prevent opioid abstinence syndrome episodes. These data are registered at the Central Methadone Register as well. Both in 1985 and in 1997 the arrested opioid users are younger than those participating in treatment programmes. However, a similar trend of ageing can be observed (Figure 6.1.3). This again confirms the progressive ageing of the drug users under study. If there had been a consistent number of young (and thus new) opiate drug users who avoided treatment programmes, these would however have been seen in the police stations and would have received methadone there. Then the age distribution of arrested drug users would not have shown the same shift to the right as the treatment distribution.

Figure 6.1.3: Age distribution of arrested opioid users receiving methadone in police stations



Analyzing the time trend of the prevalence of "criminal" opiod drug users is more difficult because the police activities vary over time. Figure 6.1.4 shows the indexed annual number of arrested users over time, i.e their number relative to the number in the base year 1993. It can be seen to decrease but also to oscillate. A clearer gradual decrease of the prevalence of 'criminal' opioid users is observed when these figures are adjusted for fluctuations in the police activity, as measured by the chance of being arrested. The variation of the number of arrests per arrested drug user per year has been used as an indicator of chance to be arrested. The relation between opioid users in treatment and the "criminal" opioid users is not clear, one might guess that treatment reduces the probability of being arrested and thus assume that the two groups are disjoint, to a large extent, but one might also doubt this assumption. However, both groups seem to decline with time and their components to grow older. Together, these observations are compatible with the hypothesis that the prevalence of drug users is decreasing, in particular because of a diminished incidence of opioid drug use.

Figure 6.1.4: Development of the number of arrested opioid users and 'criminal' opioid users



6.2 Statistical estimation methods and terminology

There are various ways to estimate the size of a given phenomenon, for instance drug use. Among these methods, the ones that are usually considered objective are based on data and statistical methodology. In order to use statistical methods, there is the need for a 'statistical model' that links observations to the underlying phenomenon of interest. This model is expressed in the language of probability distributions and parameters and, usually, incorporates a certain amount of schematizing and simplification. The whole purpose of the statistical exercise then becomes extracting as much information as possible from data about the parameters of the model, remembering that the assumptions underlying the model may or may not be good approximations of reality, and that our final conclusions must take such limitations into account.

A simple relation between incidence, duration and prevalence

Let us consider a simple, but very basic and instructive, example of "modelling reality".

Suppose we consider a rather long period of time during which we assume that incidence of drug use is constant, i.e. the same each year. We also assume that drug use can be considered as one interval in time, between start and end, with a duration that may be variable, from individual to individual, but whose average length remains constant during the considered period of time. All the assumptions related to things being the same over time are usually called 'stationarity assumptions'. It can now be shown that stationary incidence and duration imply stationary prevalence, i.e. the prevalence will be more or less the same each year, and furthermore that the average

prevalence will be equal to the average incidence multiplied by the average duration of drug use. Expressing this with symbols, we let average incidence be denoted by N (for New users/year), average prevalence by P and average duration by D (in years).

Then P=ND.

The rationale behind this formula (which mathematicians would call "a proof" or at least a sketch of a proof) becomes more obvious if we consider the concept of person-years of drug use during a e.g. one year. One way of calculating this number is P times 1. i.e. each prevalent person contributes 1 person-year each year. Another way is noticing that incident individuals, who are still active, which means that they have started , on average, up to D years back in time, contribute one year each during the year under consideration. These individuals number N times D. Thus ND = P.

An important consequence is that this formula allows estimation of one quantity, if the other two are known, or at least can be reasonably estimated. For instance, average incidence N could be estimated by P/D, i.e. average prevalence divided by average duration.

Let us discuss the usefullness of this result, considering the assumptions that have been made.

First of all, this simplified situation shows one important feature of prevalence as being rather simply related to incidence, with the relation depending of the duration of the condition that started at the moment of incidence. It also gives an immediate 'rule of thumb' to decide on reasonable values of the involved quantities. If, in a report, we see that the prevalence of drug users in a given region has been estimated as being around 10000 for a certain number of years in a row and that the duration of drug use can be believed to be around 10 years, the average incidence should be around 1000/year. Of course, incidence does not have to be constant over time, but we may still be satisfied by knowing its average order of magnitude.

More complicated models may be formulated in response to the aspects that seem too simplified in the above model. We might be satisfied by the assumption that duration of drug use is variable but that its average is constant over time (although this could be contraindicated if big changes in treatment access or repressive legislation or patterns of drug use have occured during the time period), but we usually think of incidence as a quantity having a trend over time and we are also interested in how this trend looks, since we would like to correlate it with interventions or other trends in society. Then the assumption of constant average incidence over a long time period does not appear very reasonable to start with. A model that contemplated the possibility of a time varying trend in incidence would be more suitable for our questions, and, if after estimating this trend, we discover that it is almost flat (i.e. constant over time), this will be a conclusion of our analysis, not an assumption that we cannot check. These kinds of considerations will lead to the more sophisticated models described later.

On the use of models and ways of expressing uncertainty

All statistical methods are used within the limits of some kind of model. In this context, model just means that we have decided that reality can be adequately described in simplified and at the same time precise terms. Very often, the model is formulated in mathematical terms, i.e. in a way suitable for quantitative assumptions and conclusions.

There are different kinds of uncertainty, when a mathematical model is formulated and used for statistical, analytical or predictive purposes. The two main varieties can be termed model uncertainty and parameter uncertainty. Model uncertainty is about the adequacy of the conceptual framework underlying the model or about the main qualitative features of the situation under study. Parameter uncertainty is the uncertainty about the relevant parameter values in a model that has already been accepted with regard to structure and features but where quantitative information is needed to completely specify the model. There are in fact two different kinds of parameter uncertainty, one where the parameter relates to an existing, well defined entity, such as the average duration of problematic drug use before treatment is requested, the other where the considered entity is unobservable or at least difficult to observe, either because it is hypothetical, like the effectiveness of a planned information campaign against drug use, or because it is hidden to normal observers, like the exchange activity levels of anonymous syringe sharing partners of an index person.

It is a matter of philosophical debate whether all these kinds of uncertainty are comparable, whether they all can be expressed in terms of probabilities and whether they can usefully be combined into a single assessment.

This is also the reason for the variety of techniques used to express uncertainty about the predictions or conclusions of modelling.

There is usually no direct assessment of model uncertainty. Some model features may be represented as parameter issues, e.g. the question whether the latency period, i.e. the time between start of problematic use and first treatment request, has constant average over a longer time interval. It may be possible to study this assumption if sufficiently detailed individual data is available. However, in most cases some features are taken for granted, because they appear natural, reasonable or simply usually accepted or commonly used.

Parameter uncertainty in measurable parameters is the simplest, and probably least important, part of uncertainty. This uncertainty is typically expressed in terms of standard errors of estimates, confidence intervals or equivalent Bayesian concepts.

Uncertainty about hypothetical parameters is more complicated. Usually, the most natural way of studying this kind of uncertainty is to define different possible and reasonable values for the parameters and then work out what happens with different value choices. The various outcomes are often presented separately, usually under the name of "scenarios" or "scenario analysis" (sometimes also called sensitivity analysis) and the weighting and the difficult probability assessments are left to the reader.

The perhaps most important conclusion about uncertainty and models is that, in any modelling exercise, the sources of uncertainty should be declared and discussed and the measures used to express uncertainty regarding the desired predictions should be clearly qualified, i.e. which kind of basic uncertainty they represent and, even more importantly, don't represent.

On the use of symbols, formulae and mathematical language

While most technical aspects of the methods that will be presented are discussed in the Appendices, and an effort has been made to explain assumptions and main ideas without too much formalism, there will inevitably be some formulae in the sequel, these being the best and most precise way to explain what the ingredients in the model mean. Thus, typically, we will use the following concepts and symbols:

- time in the models will be discrete, measured in years and denoted by 0,1,2,... up to a last year T. This usually means that our data and our quantities are measured in units

of one year, say, and that there is a conventional starting year called 0, but with a well defined meaning in real time (if this refers to heroin use, then it may be that we consider 1960 as an absolute starting year, before which the phenomenon was negligible, if the model is about Ecstasy use, then maybe 1980 is suitable, etc...). The last year T usually denotes the last year for which we have data in our analysis;

- since our focus is on incidence, we will have symbols for the incidence of drug use during the years we consider, say n(0), n(1),...,n(T). Thus, these symbols refer to the numbers of individuals who started their drug use during the corresponding year. Usually, these quantities are considered as not directly observable and we wish to estimate them, based on other, observable quantities. They will thus be among the parameters in our models;

- a basic component in statistical models is that observations will have a certain component of variability around whatever value they should have on average. This variability is modelled by assuming a probability distribution model. This situation can be exemplified by thinking about the simple experiment of throwing a balanced coin 10 times and recording the total number of Heads. The fact that the coin is well balanced will be expressed by saying that the probability of Head at each throw is ½. The expected number of Heads in a series is 5, but it is obvious that actual observations will yield results like 4, 6, sometimes 8 and even, but very infrequently, 10 or 0. In fact, all results between 0 and 10 are possible and the fact that these results will appear with more or less probability when carrying out the experiment, is usually summarized by assuming that the result X will follow the binomial distribution with parameters 10 and a ½ , denoted by the standard shorthand X ~ Bin(10,1/2). Many aspects of estimation can be discussed by considering the expected values of quantities, but sometimes further variability considerations are required to understand why there are technical problems in the estimation procedure but also what precision properties estimates will have;

- there is a common interpretation of probability, when there is a large group of individuals involved. What is seen as a probability for an individual, say the probability that he will experience death by overdose (OD) within 2 years of starting drug use or that he will seek treatment for the first time during his 5th year of drug use, will usually be interpreted as a corresponding proportion in a large population of individuals in the same conditions as the one mentioned above. Thus, the statements " the proportion of drug users that experience death by OD within the first two years is assumed to be 2%" and "the probability that a drug user will experience death by OD within the first two years is assumed to be 2%" will be used exchangeably. This correspondence is referred to as "the law of large numbers" in mathematical literature, has limitations, but is usually assumed to hold in large human populations.

- in many kinds of observations, some sort of individual time interval will be involved; for instance, in using first access to treatment data, the time elapsed between start of drug use and the time at which treatment is sought is such a quantity. There are several essentially equivalent terms: this time can be seen as a waiting time (waiting for something to happen), as a reporting delay or lag (in particular if the person seeking treatment on that occasion also indicates when drug use started), as a survival time (time within the initial condition until a change occurs, in this case being a DU not "known" to the system), as an incubation period or latency period (LP), LP is the terminology that will be used in this report (the terminology originates in medical circles, as the time, from the theoretical start of the condition, until a disease expresses itself; in this case, the drug use is "expressed" when the individual becomes observable by seeking treatment). Typically, such periods are variable from individual to individual and their properties are most easily described by a probability distribution. This distribution may be 'standard', i.e. of some type described in a Statistics book and with a given name (Weibull distribution, Gamma distribution, Log-normal distribution,

etc) or defined for the occasion, maybe on the basis of observations. In the latter case, symbols will be introduced, of the type p(0), p(1), p(2), etc, denoting the probabilities that the length of the time period will be (rounding off to whole years) 0 years (i.e. less than 6 months), 1 year (more than 6 months but less than 18 months), etc.

The basic mechanism that will complicate our incidence estimation is that the "incident event' (start of drug use) will not be observed directly when it happens; it will be observed because of a subsequent event in the "life career" of the drug user, such as seeking treatment for the first time or being arrested or admitted to hospital because of an OD, When the drug user thus becomes "known" to us, there are essentially two possibilities, namely that we will be able to register the true year of start of drug use or that we will not know this year, just that drug use must have started before the observation. Thus observation will be delayed, there will be a time lag between event and its observation. The methods that we will describe can therefore be generally classified as 'lag correction methods" or methods for analyzing "delayed observations". However, the amount of detailed information that we are able to obtain about our observations will be of great importance and different methods will be seen to be suited to different situations.

The basic "delay mechanism" can be looked upon in two different ways: starting from when the events really happen (drug use incidence or onset cohort; the term cohort is used here in the sense of a group of individuals who do something well determined at about the same time) or looking at when the events "become known". In the first case, the incidence during a given year will be spread out over succesive years, in proportions depending on the LP distribution. In the second case, cases "reported" during a given year (sometimes called an entry cohort) will come from previous years and each previous year contributes a number of cases depending on the incidence that year and the proportion of incident cases that will have the "correct" delay so that they are observed during our index year.

In both cases, it is important to note that all incident cases do not necessarily become observed. This may happen e.g. because some drug users will never seek treatment, they either die or stop spontaneously before. Considering that studies and analyses are performed at a certain moment in time, it may also happen that some users will seek treatment, but after the time at which the analysis is made; thus, for the purposes of the study in question, they have not yet been observed.

These considerations can usefully be translated into formulae. For instance, if we also define the quantities X(0), X(1), ...,X(T) as the numbers of individuals seeking treatment for the first time during the years 0,1,...,T, and use the standard mathematical notation E(..something..) for the expected value of ..something.., we can, as an example, write down the basic relations describing how the total of cases observed a given year are the sum of contributions from incidence during previous years:

$$E(X(k)) = n(0)p(k) + n(1)p(k-1) + ...+ n(k-1)p(1) +n(k)p(0), \text{ for } k=0,1,...,T$$

The meaning of this formula is that the expected value of the number of individuals seeking treatment in a given year during the considered period of time is the sum of contributions from each incidence or onset 'cohort' in the years up to the considered

year k (note the symbols n(0), n(1), etc, until the last n(k)), where each contribution is the fraction expected to wait exactly the number of years required to end up in year k (those starting in 0 have to wait k years, those starting in year 1 have to wait only k-1 years, etc, until those who started the same year, who have to have a delay on average shorter than 6 months in order to request treatment during the same year).

Although no consensus exists about the proper terminology, we will preferentially use the terms onset cohort (those who start using drugs a given year, the onset year), latency period (length of time period between onset and observation) and entry cohort (those who enter the study, or treatment, a given year).

We will now describe some different data situations and the related models and estimation methods. Which model and method is suitable for a certain situation dependsessentially on the data available. As mentioned above, there is an important difference between the situation when the onset year is known for all individuals in an entry cohort and the case when onset years are not known. In the first case, we can apply the method described in the next section, the RDA or lag correction method. In the second case, we will have to adopt the back calculation method, described in section 6.4.

6.3 The Reporting Delay Adjustment or Lag Correction method

The probably most favourable practically arising situation for incidence estimation is that for each individual in a data set, there is information about both year of onset of drug use and year of first entry into treatment, in addition to other possible individual level covariates, and that these data are available for the whole study population during a long uninterrupted sequence of years. The statistical problem becomes rather simple and an appropriate estimation method can be used. There are, of course, problems remaining, such as the relation between the individuals in the data set and the whole target population for the estimation, possible time trends and variations in access to treatment, etc...

Background

The RDA method requires individual data records or, at least, a crosstabulation by year of declared onset of drug use and year of first treatment

Historically, the Lag Correction method was first discussed, in the drug research field, by Hunt and Chambers, during the 1970's. Their work has recently been rediscovered and an essential paper describing their approach has been reprinted(see Hickman (2006) and Hunt (2006)). However, their work was not widely accepted at the time and the method was reproposed, in the AIDS reasearch field, by Brookmeyer and Liao (1990) under the name Reporting Delay Adjustment (RDA), together with the appropriate statistical theory and later reintroduced for estimation of the incidence of problem drug use by Hickman, Seaman and De Angelis (2001).

In the AIDS context, the problem of adjusting for reporting delay arises because of administrative delays in reporting the diagnosed AIDS cases, causing the last months of a reporting period to appear to have less diagnosed cases than previous months, simply because many cases from the last months have not yet been reported. This makes it difficult to assess how many cases have really been diagnosed during a year, say, at the end of that same year. If one could wait another year, for example, then the last reports from the previous year would certainly have come in and the data for that year would be complete, but then official reporting would be delayed by a year... To correct for this problem, the Reporting Delay Adjustment method was devised, allowing the estimation of the "missing" cases already at the end of the reporting period. Thus, an "estimated" correct total becomes immediately available. This "immediate estimate" can then gradually be substituted by the true numbers as they come in and time goes by.

The interval of time between "onset of drug use" and "presentation to treatment" can be seen as analogous to the time between "AIDS diagnosis" and "AIDS report", making the problem of estimating drug use incidence similar to that of estimating AIDS incidence. The methods developed for adjusting AIDS reports can thus be adapted to the drug use context, to estimate the lag between onset of heroin use and treatment presentation and, hence, the historical trends in heroin incidence.

Because of the important role played above by the Latency Period (LP), i.e. the interval of time between "onset of drug use" and " treatment entry", it is important that these two events have a strict definition, so that the LP acquires the same meaning for all individuals in the data set. While the "onset of drug use" usually has to be the answer

to a standardized question to each individual (for instance, "What year did you first use heroin?"), it is important that the "treatment entry" has the same meaning for all individuals, typically "the first time that the individual came into contact with a given type of treatment facility". It is, for instance, not sufficient to know that the individual was first registered in our data collection system a given year, if we don't know that this corresponds to the first time ever this individual had contact with the treatment system. (In practice even if year of first treatment is recorded there will always be some proportion of misunderstandings leading to potential data quality problems, it would be good to carry out a validation study on the quality of such routinely recorded information, if possible.)

A theoretical example and discussion of problems

In order to explain the application of the RDA for the estimation of DU incidence, let us define the following quantities:

- a given interval of years, denoted by 0,1,...,T.

- n(k): the onset incidence of first drug use of DUs (who present to treatment at least once between years 0 and T) at time k, k = 0,…,T. These individuals pass through a period of hidden drug use before they become visible by having their first contact with some health care service. These quantities are the parameters to be estimated.

- p(k): the probability distribution of the period between the first use of the drug, and the time of the first presentation for treatment, the “Latency period (LP) distribution”; p(k) is the probability that this presentation is “delayed” by k years, with respect to onset, for k=0,1,... This distribution must also be estimated from data, but the RDA will do this at the same time as DU incidence is estimated.

- X(j,k): the number of DUs enrolled in treatment during year k, k=0,1,...,T (these numbers can, somewhat confusingly, but correctly, be called treatment incidence) who have started DU in year j (this information is now part of the data). These are the observations on which inference will be based.

The RDA (a simple example, using the notation introduced above)

For the purpose, we will now consider a constructed situation, where we "know" the real incidence and LP distribution, but where we imagine that a statistician, who doesn't know these things, tries to estimate them from observations.

True incidence of DUs (assumed unknown to the analyst), the object of our estimation exercise:

Year	Onset incidence (n)
0	10
1	30
2	20
3	10

Distribution of the latency period (LP) (also assumed unknown to the analyst):

p(0)=1/5

p(1)=1/5

p(2)=2/5

p(3)=1/5

The various onset cohorts are expected to present for treatment according to the formula

$E(X(j,k)) = n(j)p(k-j)$, for $k=0,1,...,T$ and $j=0,1,...,T$

This means that various proportions of a given start cohort present for treatment in subsequent years, stating the year of DU start, so that all bold entries in the table below can now be observed.

In numbers, the situation can be presented as follows: (the numbers in bold type are observed, the ones in slightly shaded fields are hidden to the observer, but follow the postulated model...)

This is what the statistician is expected to see after observing entry into treatment during years 0,1,2 and 3 and collecting information about onset year.

Onset year	Cohort size (n)	Year of entry in treatment				Observed per onset cohort	Not yet observed
		0	1	2	3		
0	?	**2**	**2**	**4**	**2**	**10**	?
1	?	**-**	**6**	**6**	**12**	**24**	?
2	?	**-**	**-**	**4**	**4**	**8**	?
3	?	**-**	**-**	**-**	**2**	**2**	?
# entering treatment		**2**	**8**	**14**	**20**	**44**	?

This is the real situation, that the statistician doesn't know, but that he wishes to estimate.

Onset year	Cohort size (n)	Year of entry in treatment				Observed per onset cohort	Not yet observed
		0	1	2	3		
0	10	**2**	**2**	**4**	**2**	**10**	0
1	30	**-**	**6**	**6**	**12**	**24**	6
2	20	**-**	**-**	**4**	**4**	**8**	12
3	10	**-**	**-**	**-**	**2**	**2**	8
# entering treatment		**2**	**8**	**14**	**20**	**44**	26

Because we assume that the individuals in the right hand column (Observed...) have

all correctly answered the question about their start of drug use, thus giving us the numbers in the central part of the table, the only problem remaining is that a number of individuals from each onset cohort have not yet shown up by the last year of observation (the ones in the last column). If these "missing" individuals could be estimated, these estimates could be added to the numbers in the Observed column to yield the total number of individuals in each onset cohort, i.e. incidence. If the LP distribution were known, this would be an easy task. Take, for instance, the row in the table corresponding to onset year 2. Eight individuals from that cohort have been observed by the end of year 3. If it was known that during the 2 years that have elapsed for that cohort 40% of individuals have presented for treatment (1/5 + 1/5 = 2/5), then 8 must be equal to 40% of the total incidence that year, i.e. incidence must be equal to 20. In reality, the LP distribution is usually not known to the analyst, but if it may be assumed that at least one row in the table is complete, i.e. that all individuals in the corresponding cohort have entered into treatment or, in other words, that the maximum delay is shorter than the observed time period, then the LP distribution may be estimated from a "complete" row. In our case, if we assume that the reporting delay is at most 3 years, then the first row must be complete and then the probabilities in the LP distribution can be estimated by the proportions formed by the numbers in that row divided by the row total (since the numbers in our table are "exact", since they correpond to expected numbers, we get exactly the proportions (1/5, 1/5, 2/5, 1/5)...)

In reality, there is partial information about the LP distribution also in other rows, even if not complete. This leads to a more complex, but better, estimation method, that, at the end, will produce an estimate both of the LP distribution (or of a time constrained version, if the observation period is short) and of the onset incidence (or again, of a partial incidence, conditioned to being reported within the length of the observation period, if this period is short). On the other hand, numbers will typically not be exactly as expected from the model, but will contain some natural variability around the expected values. The statistical method must then try to correct for this variability.

Some comments are in order:

Importance of data structure

- the important point is that data must be of the type shown in bold in the box above, i.e. both year of onset of DU and entry into treatment must be known for each individual (one then counts the number of individuals with each possible combination of onset and entry years and fills in the table); all relevant individuals should be included in the data set, not just a sample (however, see comment below about missing data or cases) and the data should be complete for a time period longer than most probable values of the LP (say, 8-10 years of consecutive observation years, assuming a LP with 5 years median). This last requirement of a long complete time period derives from the explanation of how the RDA works, in the box above: it must, in principle, be possible to observe the complete entry into treatment of a whole drug use onset cohort to be able to estimate the correct proportions of a cohort that will present with different delays and this, again in principle, requires an uninterrupted period of sufficient length;

Relative incidence because of short observation period

- the point about maximum delay being less than the observation period (T) is important. Otherwise, one can only estimate a form of "relative incidence", namely that part of incidence that will enter treatment within T years, without knowing how this part

is related to the whole. Under the assumption of same distribution of LP for all cohorts under study, this relative incidence will however be a constant, although unknown, proportion of total incidence, thus revealing the "true trend" but not the absolute level of incidence. This proportion is the proportion of those that will ever seek treatment who do so within T years and this number can not be estimated from the data under consideration, since that would require a longer observation period than actually available. However, if one is willing to make a guess about this value, based on other studies of LP distributions, then one can "adjust" the relative incidence estimates up to a probable absolute level. Whether this is wise or not depends on the adequacy of the assumption that the multiplier which could not be estimated from the data could instead be assumed equal to the corresponding multiplier in another, better known situation;

Relative incidence because some DU never enter treatment

- the problem with "relative incidence" vs absolute or total incidence doesn't really disappear even if the assumption about maximum delay being less than the observation period seems reasonable. There is still the problem that there may be drug users who will never enter treatment and who can therefore not be estimated from the data;

Stability of the LP distribution

- there is however a potentially important problem, typically connected with long (backward) observation periods, namely that we assume that the LP distribution has stayed the same during the whole period, while there may have been drastic changes in treatment supply or general conditions (the HIV/AIDS epidemic) that could have affected the LP distribution. This problem may even exist in the shorter term because of changes in treatment capacity/waiting list policy;

'Left truncation problem'

a related problem is the time period for which the LP distribution is well defined, the problem arising because there was no treatment system during part of the period. Suppose for instance that treatment facilities have been available since 1990, but that drug use has been going on since 1970. Then data have only been collected on individuals entering treatment since 1990. These individuals are of two kinds, those belonging to onset cohorts prior to 1990 and those belonging to onset cohorts from 1990 and onwards. The latter category poses no special problems, since treatment has been offered during the whole of their drug use careers and the LP distribution may be assumed to be the same for different onset cohorts if availability of treatment has been the same all the time. But what should we assume about the onset cohort of 1985? According to the model, these individuals should follow the same LP distributions as those starting after 1990 and that means that the same proportion of the 1985 cohort "dropped out" of drug use during the first 5 years, although there was no treatment available, as from the 1990 cohort during the 1990-95 period, when there was a treatment offer. If this assumption feels uncomfortable, it could be wise to restrict analysis to individuals reporting onset year within the time frame of existence of the treatment/reporting system and thus limiting incidence estimates to the same time period. Otherwise, some special model for drug use and termination before treatment became available must be formulated. This problem has been referred to as the "left truncation problem" (where to truncate the data set to get a reasonably time homogeneous model...) and is discussed in more detail in e.g. Hickman et al (2001);

Missing data, missing individuals...

- another important point is that all individuals of interest must be included in the data. If, for instance, only one individual out of two in the data set gives information about year of onset, but all individual have the year of entry into treatment, the table above can only be based on those having complete information and the resulting incidence estimate will only include those that give complete information... If we are willing to assume that those who gave onset information and those who did not are essentially identical as groups, then total incidence (with respect to the original target population; for estimation based on treatment data, this is usually that part of all PDU that will eventually seek treatment...) might be estimated, in our case, by twice the incidence estimated from the complete data, but often, there may be suspicion that individuals without complete data may be different from those with complete data and then the adjustment is not straightforward. In the same spirit, if there is data only from a sample of all treatment facilities in a given area, the estimates derived from the data will cover only the treatment facilities included in the sample. An external estimate of the relation between the sample and the whole population will be needed to "multiply" the obtained estimates in order to get a total estimate;

Accuracy of data

- inaccuracy in the data may introduce biases. Whilst dates of birth are generally accurate, "age at first use" may be less reliable. Overstatement or understatement of age at first use could lead to under-estimation or inflation of trends in incidence. Most heroin users start their drug use between the ages of 17 and 20, but there is a significant number that start drug use earlier or later, and the true distribution of age at first heroin use is not known;

Data detail and level of measurement

- as the RDA method requires individual data records or, at least, a crosstabulation by year of declared start of drug use and year of first treatment, the field of application will usually be in local or multi-local studies where data may be expected to be more detailed and more homogeneous than at, say, national level.

What if the LP distribution has already been estimated in some other way?

- of course, if the LP distribution is known or has been estimated in some other way, which is believed to be more reliable than the described approach, then one might simply proceed like in our example with the row in the table above for onset year 2 and directly estimate incidence from each row, using the appropriate proportion from the LP distribution. This is not likely to be a common situation. See section on Back-calculation for more information.; anticipating the contents of that section, it is important to realize that even assuming a 'known" LP distribution, the simple approach outlined above is possible if the "row totals" (see example tables above) are known, meaning that the information about onset year is available. If only "column totals" are available (i.e entry in treatment), then the problem is more complicated and the BC method has to be used.

Relation with Generalized Linear Models (GLM) or log-linear modelling

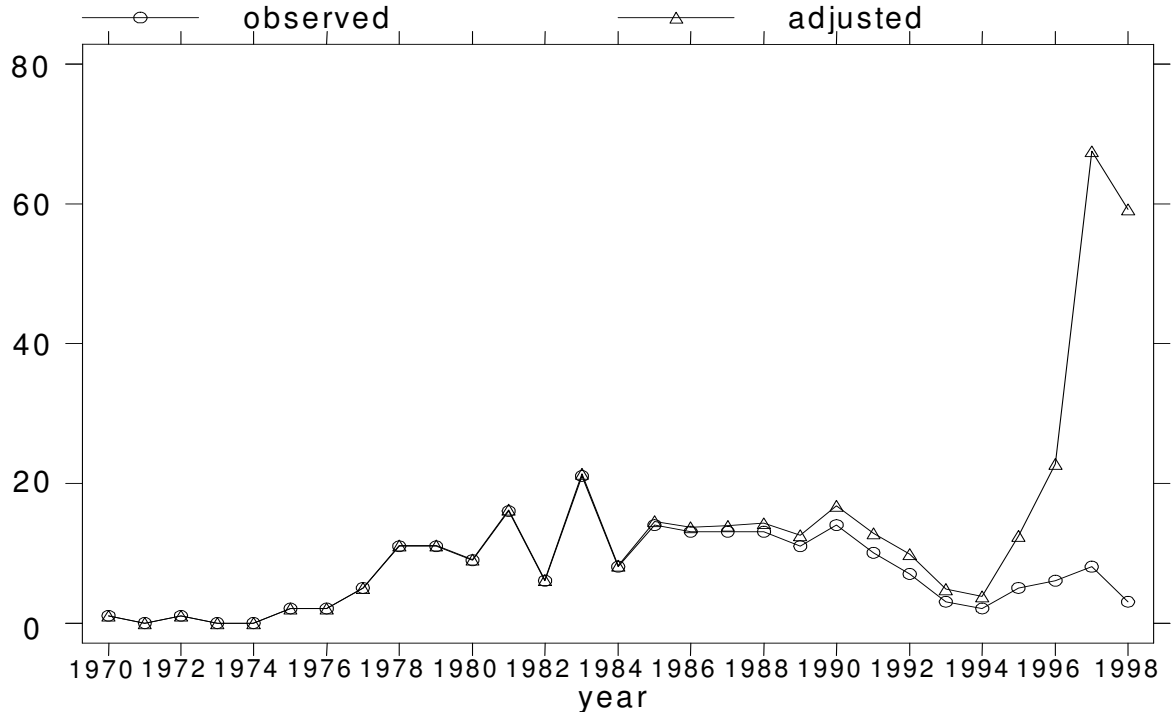
- technically, the method used for the RDA estimation is the same as would be used in a Generalized Linear Model (GLM) with a log-linear link function and Poisson distributed response. That kind of estimation procedure is available in most statistical software. The GLM framework also allows for covariates, such as sex, age, etc, if data are available with such subdivisions. Furthermore, it is possible to test some alternatives to stationary LP distribution by allowing for interaction between length of delay and calendar time. Some further discussion of these more advanced topics is presented in the Appendix 1.

The RDA method has been applied to various data sets. The results obtained for the province of Ferrara (Italy) are shown in the following example. No covariate was included. Other applications of the method include Hickman et al (2001) with London data.

Example: Estimating incidence of heroin use in Ferrara 1970-1998

Data of the kind discussed above were available for the time period 1970-1998 for the province of Ferrara (Italy). The curve labelled "observed" in Fig. 6.3.1 corresponds to the numbers in the next to last column in the Box above, i.e. to the total number of individuals in the data set having declared a given starting year for their drug use. By adjusting for the "reporting delay" , as estimated from the detailed data (not shown here), the adjusted numbers are obtained. These are the observed ones plus those that have been estimated as having started in a given year but not yet having requested treatment, although assumed to eventually do so. It can be seen that the adjusted number is always higher than the observed one, which is natural, but also that there is visible difference only about 10 years back from 1998. This shows that the "maximum reporting delay" seems to be of that magnitude (since previous numbers do not seem to need any reporting delay correction) and thus certainly less than the 28 year long observation period. Therefore it can safely be assumed that the reporting delay distribution is complete. The rather high estimates for the last four year period are slightly surprising. They arise because the probabilities of short reporting delay have been estimated to be low in the total data set. (As an example, suppose that the observed 1998 value is 2, i.e. 2 individuals have reported that they started use in 1998 and requested treatment within the same year; suppose also that the probability of requesting treatment within the same year as the start of drug use has been estimated to be 3%. Then the 2 individuals should be 3% of the total incidence of the year, which however forces the estimate of this incidence to be 66). Apart from an explanation based on bad luck or unusual variation, it is also possible that the LP distribution has changed during the last years, making DUs request treatment earlier than before. This change will concern such a small part of data so as to not be able to influence the global estimates, but the consequence would be that the estimates in the Fig. would be much higher than reality.

Figure 6.3.1: *Incidence curve estimated by the RDA method for Ferrara.*



6.4 The Back-Calculation Method

It may happen that the available treatment incidence data is not as detailed as supposed in the previous section about the RDA. In particular, only yearly totals of numbers of individuals initiating treatment may be available (thus not the cross-classification according to reported onset year of starting drug use). This is a typical situation at, say, the national level, where reports from the various treatment facilities have been sent in to a central authority, where some summary tables have been compiled. This situation is less favourable than in the previous section, because the information about reporting delays that could be inferred from the rows of the data table in the previous section has now been lost (if it ever existed...). Nevertheless, if an "external" estimate of the distribution of reporting delay, i.e. latency period (LP), can be provided, then the estimation can still be done. The name "back-calculation" has arisen from one particular way of viewing the estimation problem: one thinks of the incidence events (somebody starting to use drugs) as points along a timeline.These points are not directly observable, but each point generates another time point at some distance of the original time point, representing the time when the individual will seek treatment for the first time. These events are assumed to be observable. Thus we observe a version of the original set of points where each point has been moved forward by some quantity (in our case, the LP of the individual). Our estimation task is now to estimate the original set of points, i.e. to put the observed points "back in time" in a reasonable way. i.e. to do "back-calculation".

Background

The Back-Calculation method (BC) was developed to describe the dynamics of the AIDS epidemic (see Brookmeyer & Gail (1994) for a historical perspective). Based on AIDS incidence data and on medical knowledge as well as on statistical assumptions regarding the latency period (LP; or incubation time) between HIV infection and appearance of AIDS, the HIV prevalence and incidence in the past years is estimated. The results were sometimes used for projections of the AIDS incidence in the (near) future, by assuming that incidence would continue in the future at the estimated level for recent years. It is an important feature that most AIDS patients were not able to define with precision when their HIV infection occured, but separate studies of the LP had been conducted on special data sets (cohort studies) where this information was available.

The parallel that allows this method to be adapted to the DU incidence situation is that incidence means start of the condition (infection with HIV, start of DU) and that observation of an individual becomes possible after a latency period (LP) (from HIV to reported AIDS, from start of drug use to first request for treatment). However, it is assumed that information on when the individuals seeking treatment started their DU is not available (if this information is available then one should use the appropriate model and estimation method, the RDA or log-linear model, see the previous section) but that some "external" estimate of the LP distribution is available. This may happen because detailed data were available on a regional level, but now estimation is to be extended to national data, which are only known as totals; one may also take an LP distribution established in some other country or simply a hypothetical one based on widely accepted average values. Of course, the more assumptions have to be made in order to interpret a data set, the more reservations must be made about the validity of the conclusions.

Is there a choice between RDA and BC?

This question really contains two different questions, related to the detail of available data. With reference to the example data tables above and below, if the only data available is "column totals", i.e. numbers of drug users entering treatment, without information about drug use onset year, then there is no choice, only the BC method may be applicable. If data is in the format discussed in the preceding section, such that that RDA method could be applied, i.e both "row totals" and 'column totals" (in fact, also detailed "cell" information) are available, then one could, at least in theory, ask whether there could be any advantage in using only the "row totals" together with some external estimate of LP distribution, or even only the "column totals" and the external LP estimate with the BC method presented in this section.

The answer is no, for a variety of reasons. It can be argued that there will almost never be any good reason to believe that the knowledge one assumes to have about the relevant LP distribution for the data set under analysis is better than the information about this distribution inherent in the data set. This is because, whatever the source of the supposed knowledge, the time and place and circumstances under which this knowledge was derived were probably different from the ones related to the data set under consideration. The use of an LP distribution as correct as possible is essential for the correctness of the incidence estimates and using a "wrong" LP distribution will result in "wrong" incidence estimates. Also, whatever choice is made about how to go about the estimation of incidence from data of the type described in the previous section about the RDA method, it will never be a good idea to throw away the information inherent in the data table about the incidence cohorts, as reflected by the "row totals", which are "exact" knowledge about a part of each cohort. This is in fact what would happen if one decided to apply the BC method to this data, as the example box below will show, since one would then use only the "column totals", thus disregarding an important part of the information in the data set.

In conclusion, if there is data suitable for the RDA method, then this method should be used. If only total yearly treatment incidence data is available (the "column totals" referred to above), then there is no choice: supposing a reasonable choice of LP distribution can be made (or a suitable set of alternatives explored), the BC method can be applied, but not the RDA method.

Technicalities...

There is no standard software for the BC method, nor is there, in fact, a standard BC method. While the basic model, i.e. the relation between the observable treatment entry incidence and the underlying drug use onset incidence is the same, there are different back-calculation methods which offer alternative numerical procedures to find the estimates of incidence (penalized maximum likelihood, Empirical Bayes, MCMC), which differ in assumptions about how the latency period distribution is described (parametric vs nonparametric) or in how to model the incidence trend (parametric vs nonparametric). The immediately important concept here is "parametric", which denotes a shape of time trend that can be specified by a few "parameters". This is usually an apparent advantage because there are few numbers to estimate, their apparent precision is usually good, but the price to pay for this is that the different shapes that can be obtained by different choices of parameters are limited and thus that the fitting of such a curve to data may be misleading if reality does not look the way the possible curves can. Typically, the possible curves may have only one possible peak, while "reality" has two, or reality might have a flat phase while the curves are not able to reproduce such a feature. The alternative is "nonparametric", which paradoxically means with a lot of parameters, typically so that incidence each

year may have any value, independently of other values. Since such curves, while being able to represent any shape feature, often look quite jagged, a "smoothness" restriction is then employed to reflect an underlying continuity of the incidence process.

More technical details are given in Appendix 2. Below, we instead present some examples to illlustrate principles and applications of BC.

A theoretical example illustrating how the BC method works

In order to explain the application of the BC for the estimation of DU incidence, we will again consider the same constructed example as in the previous section. this time however, we will assume that the statistician only has information about first entry into treatment and not about DU onset.

We will use the same symbols:

- a given interval of years, denoted by 0,1,...,T.

- n(k): the onset incidence of DUs (the size of yearly drug use onset cohorts) year k, k = 0,…,T. These individuals pass through a period of hidden drug use before they become visible to the system by having their first contact with some health care service. These quantities are the parameters to be estimated.

p(k): the distribution of the period between the time of the first use of drug, and the time of the first presentation for treatment, the “Latency period (LP) distribution”; p(k) is the probability that the "delay" between onset and observation is k years, for k=0,1,... This distribution is now assumed known or reliably estimated or estimable from a different but pertinent data set. Anyway, the numbers {p(k)} will be assumed known and fixed for the purposes of estimation using the BC method (see section 6.6 for a discussion of methods of estimation of LP).

X(k): the number of DUs enrolled in treatment during year k, k=0,1,...,T (these numbers can, somewhat confusingly, but correctly, be called treatment incidence). These are the observations on which inference will be based.

How the Back Calculation method works (a simple example, using the notation introduced above)

Once again, we assume the numbers below to represent the real situation, that the statistician of course doesn't know, but wishes to estimate:

True incidence of DUs (assumed unknown to the analyst):

Year	Onset incidence (n)
0	10
1	30
2	20
3	10

Distribution of the latency period (LP) , in years:

p(0)=1/5

p(1)=1/5

p(2)=2/5

p(3)=1/5

Observe that the sum of these probabilities is 1. This means that only delays between 0 and 3 years are possible and that all the incident DUs will eventually present for treatment (or else, that the incidence defined above is to be interpreted in a "relative" sense, namely as those that will eventually present for treatment)

The various incidence cohorts are expected to present for treatment according to the formula

$E(X(k)) = n(0)p(k) + n(1)p(k-1) + ...+ n(k-1)p(1) +n(k)p(0)$, for $k=0,1,...,T$

which means, explicitly,

$E(X(0)) = n(0)p(0)$

$E(X(1)) = n(0)p(1) + n(1)p(0)$

$E(X(2)) = n(0)p(2) + n(1)p(1) + n(2)p(0)$

$E(X(3)) = n(0)p(3) + n(1)p(2) + n(2)p(1) +n(3)p(0)$

In numbers, the situation can be presented as follows: (only the numbers in bold type can be observed, all other (in shaded fields) are hidden to the observer...). It should be observed that the numbers in the table are the same as in the RDA example, but that we assume that we have much less information in the present case, since we are able to observe only the column totals, while we could observe almost everything, except the cohort sizes, in the previous case. In fact, it would not be possible to draw any conclusions from these data alone, we must now have an externally derived LP distribution estimate in order to proceed. We will therefore assume that the statistician has an estimate of the LP distribution available.

What the statistician sees:

Onset year	Cohort size (n)	Year of entry in treatment				Observed per onset cohort	Not yet observed
		0	1	2	3		
0	?	?	?	?	?	?	?
1	?	-	?	?	?	?	?
2	?	-	-	?	?	?	?
3	?	-	-	-	?	?	?
# entering treatment		**2**	**8**	**14**	**20**	**44**	?

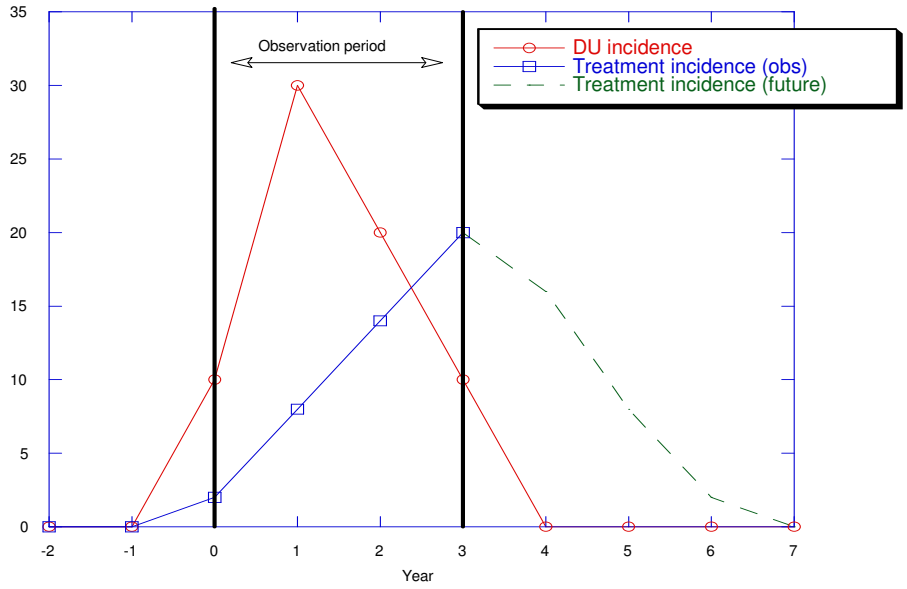
The statistician also has (approximate) knowledge about the LP distribution...

What the statistician doesn't know, but wishes to estimate...

Onset year	Cohort size (n)	Year of entry in treatment				Observed per onset cohort	Not yet observed
		0	1	2	3		
0	10	2	2	4	2	10	0
1	30	-	6	6	12	24	6
2	20	-	-	4	4	8	12
3	10	-	-	-	2	2	8
# entering treatment		**2**	**8**	**14**	**20**	**44**	26

It is interesting to note that the assumed DU incidence peaks in year 1 and then decreases, but that the treatment incidence is steadily rising during the 4 years of observation. As can be seen from Fig. 6.4.1 below, the treatment incidence gives a misleading impression about the DU incidence trend during the observation period, which is the reason for trying to estimate the "correct" DU incidence from the treatment data. It can also be seen, by considering the "future" development of the treatment incidence, that this indicator is a delayed and smoothed out version of the DU incidence, in general.

Figure 6.4.1 Comparison between DU incidence and typical treatment incidence in a simplified example.



Furthermore, by considering the "hidden" numbers in the table, it is seen that DU incidence from later years (in this case, e.g. the two last years) contributes rather little to the observed total treatment incidence numbers. This makes estimation of the DU incidence corresponding to these years more difficult. The estimation problem consists of two parts: the first is estimating the number of DUs that have not yet been observed (the total DU incidence during the 4 years has been 70 individuals, but only 44 of these have been observed, the others will be observed in the "future"), the second is that the treatment incidence, both observed and estimated future, must be "sorted out" to the different DU incidence years.

However, since we know the theoretical relation between the expected observed numbers and the quantities of interest (n), we might try to solve the equations above for values of n(0),n(1),n(2) and n(3) that would yield the observed numbers. This is the basic idea behind the BC method.

In our case, this amounts to solving the following linear system for {n(k)} for given {X(k)}:

$X(0) = n(0)x1/5$

$X(1) = n(0)x1/5 + n(1)x1/5$

$X(2) = n(0)x2/5 + n(1)x1/5 + n(2)x1/5$

$X(3) = n(0)x1/5 + n(1)x2/5 + n(2)x1/5 + n(3)x1/5$

If the observed numbers happen to coincide with the expected numbers, as given in the table above, i.e. (2,8,14,20), it is easily seen by substituting in the equations that

the solution turns out to be exactly (10,30,20,10) and thus our "estimation procedure" has worked perfectly.

However, if a little "natural variation" causes the observations to be (3,7,15,20) instead of (2,8,14,20), the solution becomes (15,20,25,20), an estimate that still has an almost correct general tendency, but with values quite different from the correct solution, although the difference between the data sets (2,8,14,20) and (3,7,15,20) is quite small.

If instead (2,8,16,18) is observed, the solution becomes (10,30,30,-10), which is rather unacceptable, because of the negative number during the fourth year.

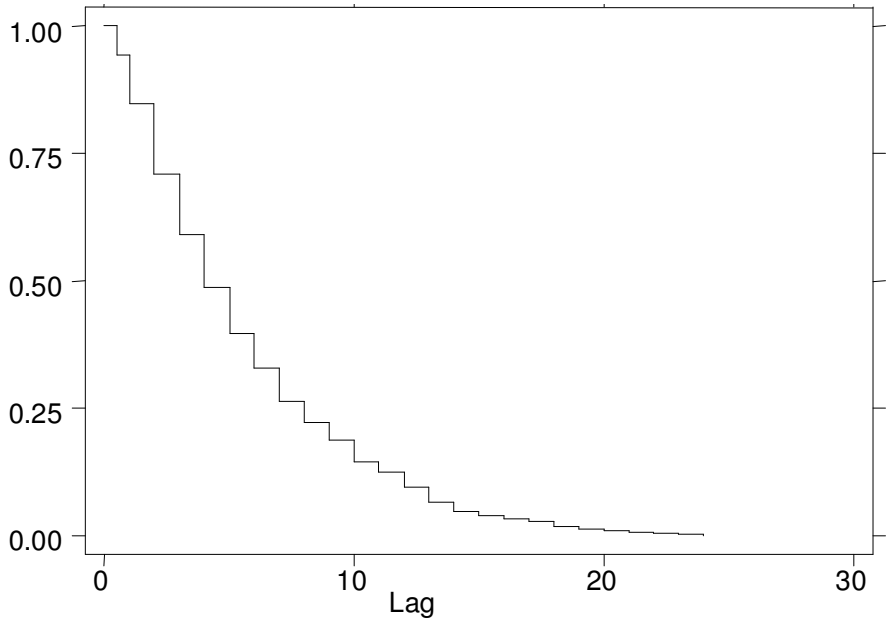
It is this unfortunate tendency of the equation system above to yield large differences in estimates in the presence of small (and quite probable) variations in observed numbers that creates the need for a sophisticated mathematical machinery to "smooth" the solutions. Smoothing means to avoid large changes in the values over short time intervals. In the example above, there is variation in the observed data, from year to year, but the largest change equals 8; in the corresponding solution, the largest change equals 40. Smoothing is achieved by considering a gain from being as exact an estimate as possible, but at the same time a penalty if the obtained estimate exhibits too much variation and then finding a balance between these two factors. It is , at least to the present date, not possible to do this in a simple way with simple statistical instruments, such as Excel spreadsheets or standard statistical software.

Example: Estimating incidence of heroin use in Emilia-Romagna (Italy)

Heroin by injecting caused the majority of drug related consequences, during the '80s and '90s, for both the health and the criminal justice systems in Italy (EMCDDA, 2000). The number of people between the ages of 15 and 54 who have used heroin at some time in their life is estimated to be, in 1999, at least 300,000 (Ravà and Rossi, 1999). Thus, in this example, the application of the BC method is focussed on estimating the onset incidence curve of just heroin use. The estimate of the LP distribution that was used in this example results from a wide ranging series of analyses of data sets within an Italian multicentric longitudinal study of DUs in treatment (the VEdeTTE study). More details about the LP related analyses carried out on these data sets are given in the section 6.6

The LP distributions were estimated for different regions and provinces, with rather similar results. The LP distribution (Kaplan-Meier estimate) for the region Emilia-Romagna is shown in Fig. 6.4.2 The LP distribution in Emilia-Romagna has a median of approximately five years. There seems, however, to be a tendency for LP to be longer in individuals who start their drug use early.

Figure 6.4.2 Latency period distribution for Emilia-Romagna (in years).



The BC method was applied to data provided by the Ministry of Health, who routinely collects data on treatment in public services (see Country report EMCDDA 2000). Data on DUs in private services were excluded as these are mostly referred from public centers and double counts would have been high. Data include only clients who enrolled in treatment for the first time and do not include those who only contacted services but did not receive treatment. Treatment refers to any therapeutic and rehabilitation procedure - either pharmacological or not - offered by the service, even outside the premises (prisons, therapeutic communities, hospitals). The data are shown in Fig. 6.4.5 (the points marked Observed).

In this case, the Bayesian/MCMC method was used to obtain the BC estimates of drug use incidence (see Appendix 2).

The following figures show the results obtained for the region Emilia Romagna. In Fig. 6.4.3, the estimated DU incidence is shown together with a pointwise 95% confidence

band. Fig. 6.4.4 shows the corresponding estimate of cumulative incidence and in Fig. 6.4.5, the observed data is compared to the corresponding fitted points, obtained by using the estimated DU incidence and the assumed LP distribution to calculate the "predicted" treatment incidence.

Figure 6.4.3: Incidence of DUs estimated for Emilia Romagna with confidence intervals

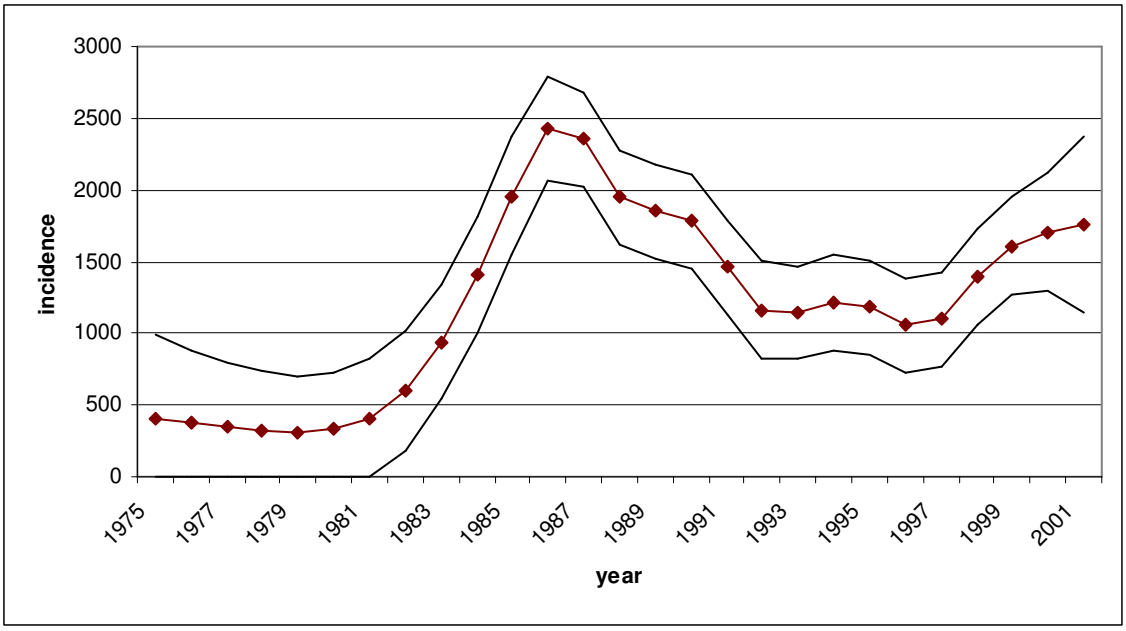


Figure 6.4.4: Cumulative incidence of DUs estimated for Emilia-Romagna

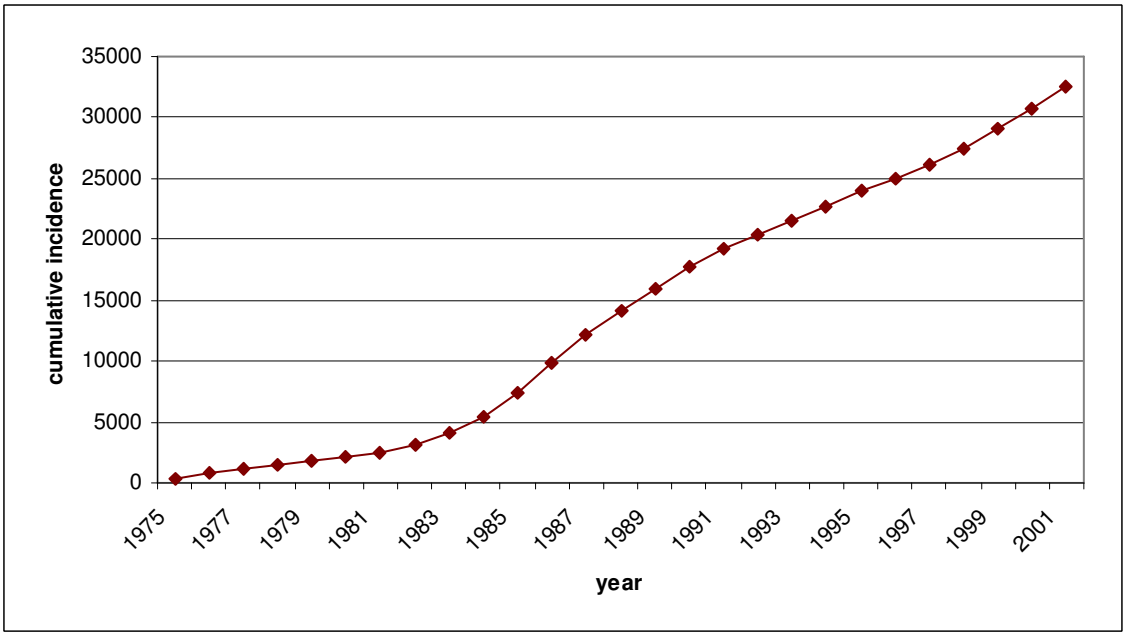
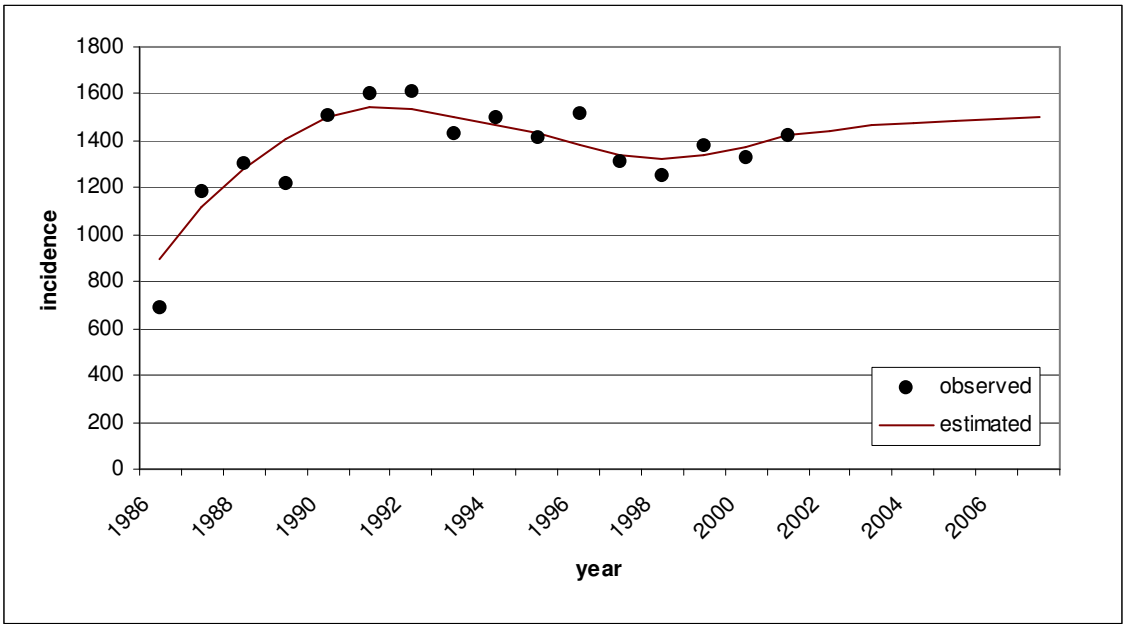


Figure 6.4.5: Incidence of DUs in treatment in Emilia-Romagna



Possible problems with data and estimates

Bias resulting from truncation of LP observations

Estimates of the latency period using treatment/surveillance data may under-estimate the true distribution because the data are right truncated (i.e. long LPs are not observed because of the reporting horizon). This bias is larger for recent drug use "epidemics" whereas it will be smaller for older (stabilised) epidemics. A bias in the LP distribution results in a bias in the incidence curves estimated by BC. As a rule, assuming an LP distribution which specifies times shorter than they should be will result in incidence estimates which are smaller than they should be. Thus, a sensitivity analysis is always appropriate (see next Example).

Use of covariates

In this example, the same LP distribution has been used for all individuals in the data set. If some covariate was believed to be influential with regard to the LP distribution (age, sex, location, etc...) and data are sufficiently detailed to allow a classification according to this covariate, the BC method can be modified to incorporate this subdivision of data. However, there is no standard software available for BC, so the corresponding programmes would have to be written (see Appendix 2)

Double counting of individuals in treatment entry data

An important problem related to the treatment incidence data used with the BC procedure is represented by double counting of individuals which causes the observed incidence of DUs presenting to treatment to be higher than the actual one, and, as a consequence, a bias in the BC estimates. An attempt to overcome this problem could be made by reducing the reported observed incidence data, on the basis of some information about the amount of double counting, if available.

BC estimates are more imprecise for recent incidence than for past incidence

Due to the rather long latency period, therapy data provide little information on recent DU incidence. Therefore, estimation of recent incidence of drug use is more imprecise than estimation of historical incidence. For the same reason, BC should not be applied when the time series of aggregated therapy incidence data is too short. As a rule of thumb, a series of at least 8-10 years of observation should be available.

Checking assumptions and parameters...

Back-calculation involves several important assumptions and parameters that need to be estimated for drug users (e.g. the shape of the LP distribution, the influence of covariates such as age, sex, education level…..). These are based on external information coming both from observational secondary data and from surveys. There may be need for detailed analysis of LP data (see next section) and investigations about the functioning of the treatment registration system, the method chosen to classify treatment episodes as "first treatment", etc...

The smoothing parameter

The BC incidence estimate also depends on the choice of a smoothing parameter, which determines the balance between accurate estimation and acceptable variability of estimates. In principle, this parameter can be estimated, a possibility which is discussed in Appendix 2, but this would make the computational time longer and requires various assumptions of its own. It can also be kept fixed at various values and then, by comparing the results, from one extreme were all variation in the incidence curve is essentially removed (too much smoothing) to the other where the incidence curve appears jagged and irregular (too little smoothing), a suitable intermediate value can be chosen.

Sensitivity analyses

For all the reasons discussed above, uncertainty about the LP distribution, about the influence of covariates, about the choice of smoothing parameter, sensitivity analyses are recommended. This means redoing the DU onset estimation with different choices of assumptions and comparing the resulting estimates to identify possible major sources of uncertainty.

Two examples of sensitivity analysis for BC estimates

Application to Italian Data 1988-2000

In order to evaluate the sensitivity of the BC method to the inclusion of the covariate age in the model and the choice of the latency period distribution, a sensitivity analysis, was performed (EMCDDA, 1999) (in this case, a slightly different BC method, called Empirical Bayes BC, was used, compared to the previous example, where an MCMC (Monte Carlo Markov Chain) BC was used; the use of different estimation methods had mainly technical reasons, one of the corresponding programmes being able to handle covariates, the other being able to deliver confidence intervals; see Appendix 2).

Italian data about treatment incidence in public institutions during the period 1988 to 2000 was used; 8 different estimates of latency period distribution (4 Gamma and 4 Weibull models), as provided by the latency period analysis performed for 4 Italian cities (Rome , Milan , Frosinone , Latina) were considered. The BC was performed both with and without the age-covariate.

Figures 6.4.6 shows the 16 different estimated incidence curves of problem drug use as being rather similar, allowing the conclusion that estimation is not very sensitive to the differences in assumptions about the LP distribution and whether or not the covariate age is included or not in the model. This impression is strengthened by the comparison between fitted treatment incidence values and observed data in Fig. 6.4.7. It can be noticed, in the same Figure, that predicted values for the years 2001-2006 have been computed as well, simply assuming that the latest estimated DU incidence (year 1999-2000) is valid also for subsequent years and then computing estimates of the treatment incidence also in the future, relative to data.

Figure 6.4.6. Incidence curves for the total population estimated by the Empirical Bayesian Back-Calculation procedure with different estimates of the latency period distribution and with and without the covariate age.

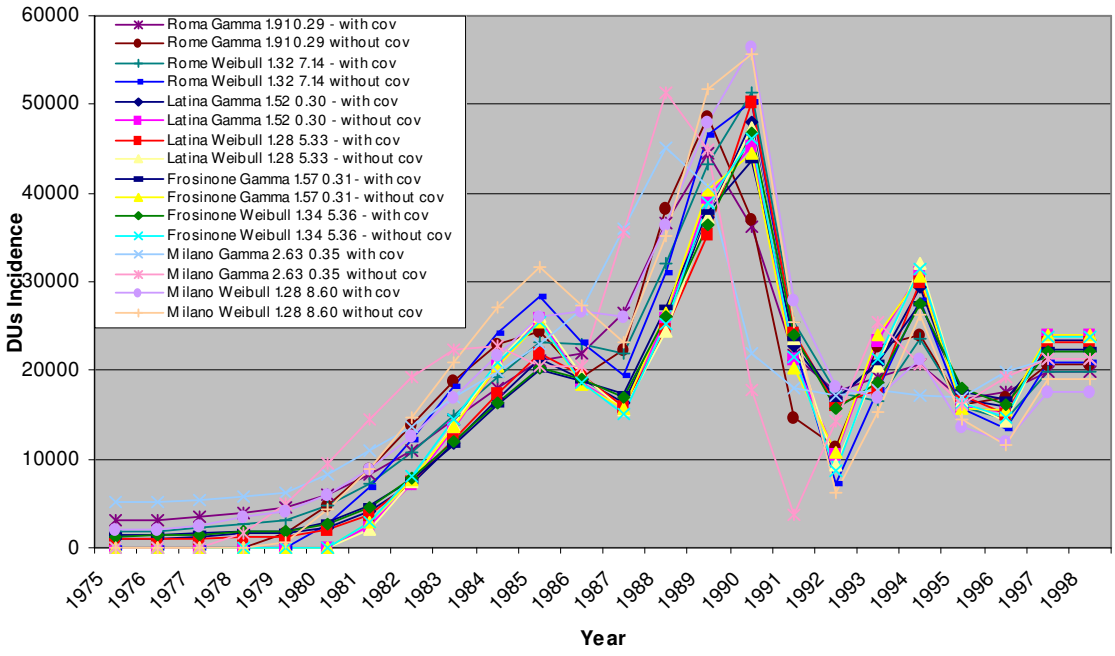
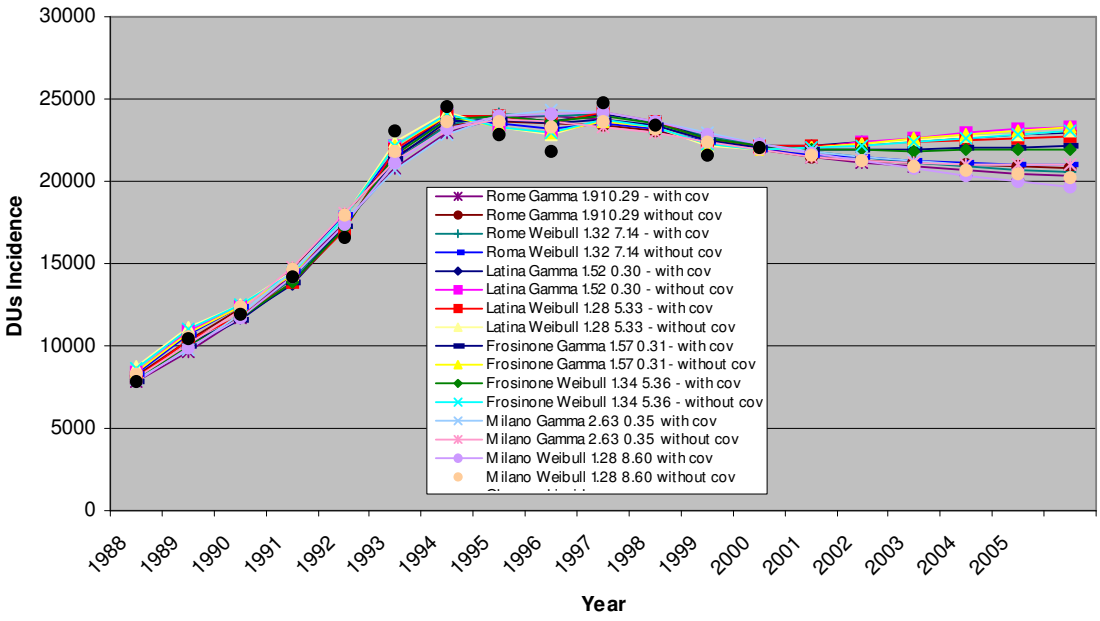


Figure 6.4.7. Incidence curves for the presentation to therapy estimated by the Empirical Bayesian Back-Calculation procedure with different models of the latency period distribution and observed data.



Application to Spanish Data 1991-2001

Since 1991, the Spanish Treatment Demand Indicator project provides information on new admissions to treatment. The register covers all public treatment centres and private ones with public financing. The register considers a treatment as new admission if previous treatment was more than 6 months before (or if no previous treatment is registered). There is a self reported variable for each treatment admission, stating whether that admission is the first approach, or not, to treatment.

To apply Back-Calculation, it is not necessary to have individual data, as aggregated data on treatment incidence in given time periods is sufficient. However, the estimation of the LP distribution is necessarily based on individual data. Such individual data was made available through a detailed study of a 10% sample of opioid user treatment admissions from three years (1992, 1996 and 2001) (see Table 6.4.1).

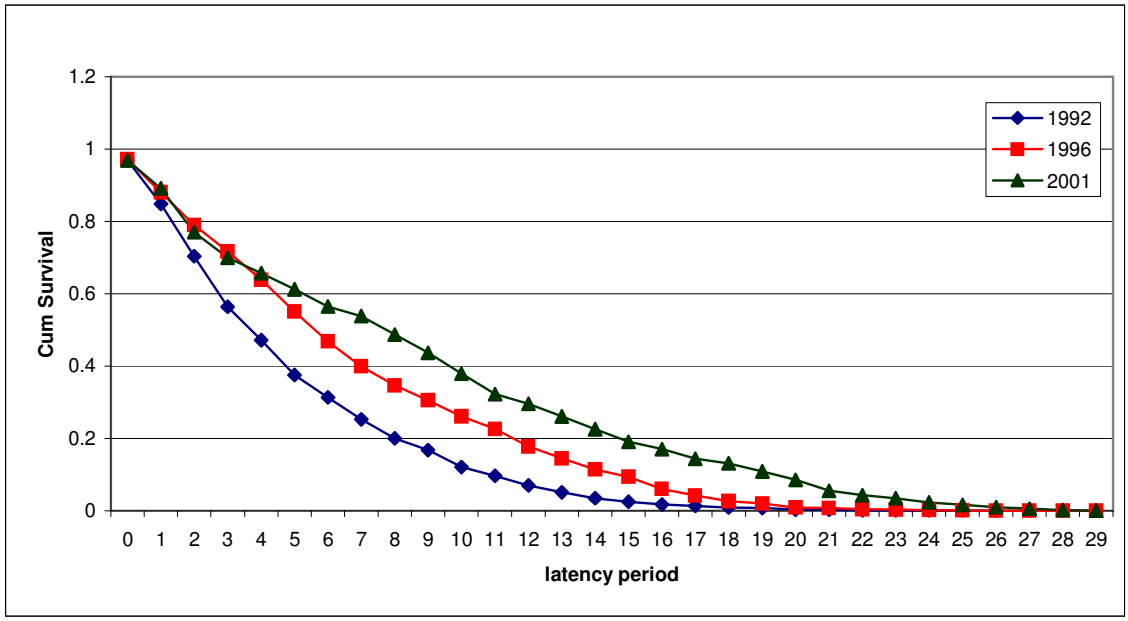
Table 6.4.1 Treatment admissions for opioid use in Spain.

10% sample	Previously treated	First treatment	Total
1992	1553	1906	3459
1996	2716	1577	4293
2001	2432	728	3160
Total	6701	4211	10912

From this sample, providing data on a total of 10912 opioid users, 4211 were classified as treated for the first time. it is interesting to observe that the proportion of first time treatments among treatment admissions decreases with time, indicating a possibly decreasing recruitment of new DUs during the observed period.

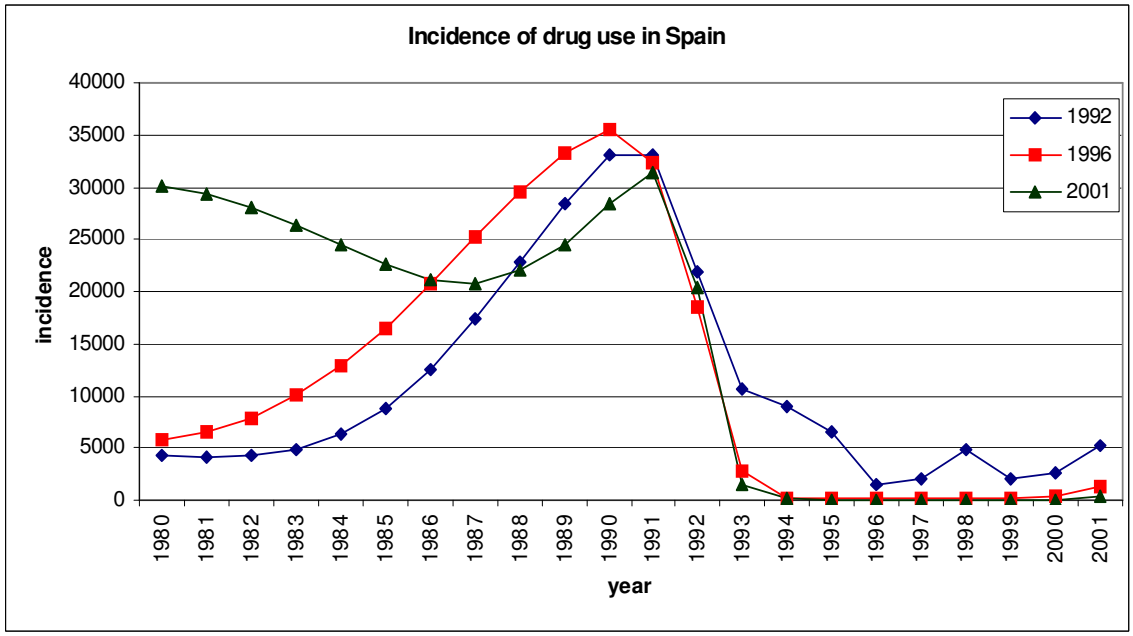
The survival functions obtained from these 4211 subjects, considering data from each treatment year separately, are displayed in figure 6.4.8. The mean latency time for the 1992 data was 5 years, for the 1996 data 7 years and for the 2001 data 9 years. This progressive lengthening of the average LP is discussed in more detailed form in the next section and in Appendix 3, together with the bias problems connected with "backwards" and "forwards" estimation of LP distributions. For the purpose of this Example, we will simply take the three distributions below as three possible candidates and study the sensitivity of BC estimates of DU incidence to the the three alternatives.

Figure 6.4.8 Survival functions computed from 3 different first treatment year cohorts of Spanish DUs.



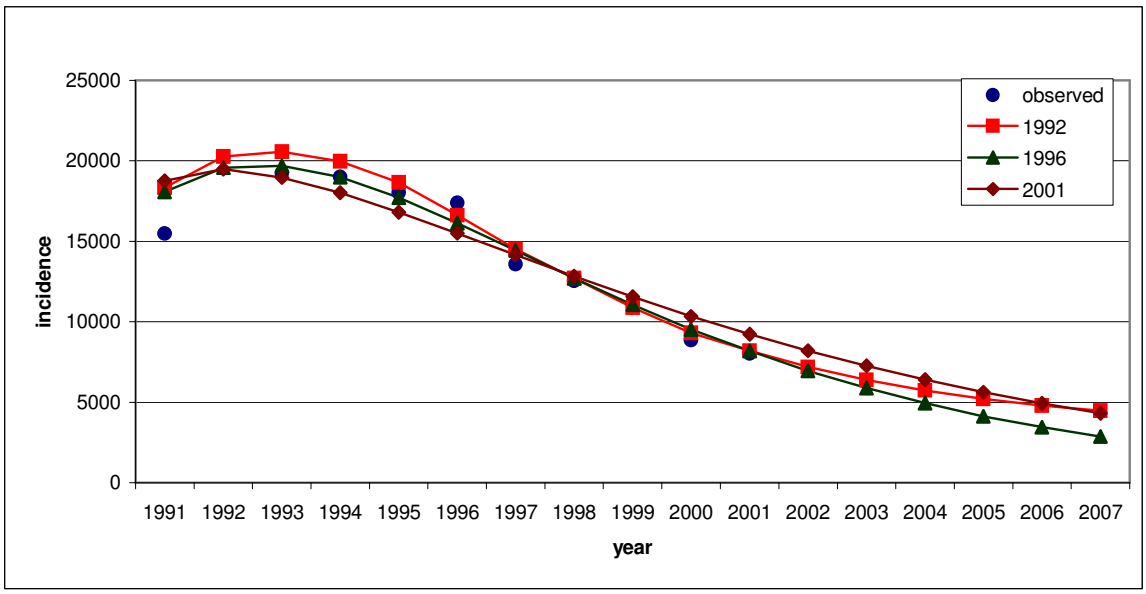
The aggregated treatment incidence data spanned 11 years and contained 163611 first treatment admissions of (self-reported) opioid users. The information was organised in 6 month periods from 1991-1 to 2001-2 (22 periods). The DU incidence was then estimated, this time using the Bayesian formulation of the BC algorithm. The results of the DU incidence estimation can be seen in Fig. 6.4.9, while the observed data (now aggregated to annual numbers) and the fit derived from the three estimated DU incidences, including a six year forecast of treatment incidence, assuming that the DU incidence estimated in the last years of the observation period remains adequate for the future period, is shown in Fig. 6.4.10.

Figure 6.4.9 Estimated DU incidence in Spain with 3 different LP distributions.



As mentioned previously, the three LP distributions, in the order 1992, 1996, 2001, imply progressively longer latency periods. In order to explain a treatment incidence (see Fig. 6.4.10) that peaks in 1992 and then decreases rather markedly, it is reasonable that a large part of the DU incidence must be attributed to the period before the peak, and also at a longer distance from that peak the longer the average latency period is assumed to be. This behaviour can be clearly seen in the three estimated DU incidence curves, using the three different LP distributions.

Figure 6.4.10 Yearly observed and fitted treatment incidence in Spain with 3 different LP distributions.



It is noteworthy that assuming the "longest" LP distribution implies that a high incidence peak as far back in time as 1980 is estimated It is interesting that the large

difference between the 2001 incidence estimate and the other two estimates is not seen in Fig. 6.4.10, indicating that rather different DU incidence models may give rather similar fits to observed treatment incidence data. This means that the observed data is not able to discriminate very well between the three DU incidence alternatives, and that the main conclusion of the analysis should be that the marked decrease in treatment incidence since 1992-93 must correspond to a similar strong decrease in DU incidence, that most DUs started before 1992, but that it is quite assumption dependent to decide how this incidence was distributed over time.

6.5. Other methods

In section 6.3, the available data was assumed to be rather detailed, essentially individual level data for a whole population during a long time period. In section 6.4, the data was supposed to be aggregated by treatment entry year, but still covering a long period of time, without information on year of DU onset, and an external LP distribution was then required in order to produce DU incidence estimates. What can then be done if even less data, or different data, are available?

Using RDA and BC with other data sources other than first treatment

In theory, it is not necessary to focus on "first treatment request" as the endpoint of a latency period. It is sufficient that the endpoint is well defined, possibly only related to drug use, possibly happening to a large part of the DU population, and that it is possible to estimate the relevant LP distribution (it doesn't really matter if the event does not happen to all DUs; either it is known, together with the distribution of LP, what proportion of all DUs is finally reached, in which case estimation of the absolute incidence can be made or, if this final proportion is not known, only relative incidence will be estimable). One interesting example can be found in De Angelis, Hickman & Yang (2004), where the event "death by opioid overdose" is used to estimate DU incidence in England 1968-2000. Other possibilities could include first contact with the judicial system or first appearance of a drug use related health condition, but care should be taken that the studied phenomenon is reasonably constant over time and potentially equally "available" to all DUs (problem with changes in law enforcement or resources, temporary epidemics of a certain disease, etc...).

Using prevalence estimates to estimate incidence

One or several succesive prevalence estimates are not in themselves sufficient to determine or estimate incidence. It is enough to think of the following situation "the prevalence is the same at two points in time"; this observation can be explained by "zero incidence and long duration of the condition" or "incidence = prevalence and short duration of the condition". Essentially, some assumptions will have to be made and conclusions will depend on how reasonable these assumptions are in the specific situation. An example from a different but related field is an attempt to reconstruct past hepatitis C incidence in the US by Armstrong et al. (2000), based on an age-specific prevalence at one point of time and a sample study of the age-specific incidence at about the same time. The assumption that is made is that past incidence may have different levels in different periods of time but that the age-specific incidence remains the same. With this assumption, it becomes possible to estimate the variation in the general level of incidence in the past. It should however be remembered that hepatitis C has a duration structure which is very different from that of drug use. The example shows, however, that it is possible to construct models which use only available data and, hopefully, reasonable assumptions, to estimate quantities of interest.

In another recent paper, Nordt & Stohler (2006) use the RDA method on treatment data from Zurich (CH) to estimate DU onset incidence, but also estimate prevalence from incidence (in order to compare with existing prevalence estimates) by adopting what they call a yearly cessation rate (in this case = 4%), estimated from treatment data as the average proportion of individuals who drop out of treatment but do not reappear within 10 years and interpreted as the percentage of drug users who effectively stop belonging to the population. Such an estimate (which is equivalent to

an average period of 25 years of drug use, possibly including some stops and relapses, within the 10m year limit) would allow incidence to be estimated from two reasonably close prevalence estimates, since Prevalence2 = Prevalence1 - those that have quit during the period + incidence during the period.

Combining first treatment data with prevalence estimates or other information

A further possibility is to combine the treatment incidence data used with the BC method with other information to provide better estimates of the drug use onset incidence. Two examples of this, from the HIV/AIDS field, are the works by Glad et al (1998) and De Angelis et al (1998) about estimating HIV incidence using AIDS incidence data and an estimated incubation period but, in the first case, adding some estimated HIV prevalence data in one or two time points or, in the second case, considering time information about first time HIV positivity tests, which are supposed to be proportional to the yet "unknown HIV positives", i.e. infected individuals who have not yet tested positive nor shown AIDS symptoms. As discussed in Section 6.4, the similarity between the HIV and the DU incidence inference problems is evidenced by identifying incidence of DU with incidence of HIV, the AIDS incubation period with the LP and first DU treatment with the AIDS diagnosis. Continuing in this fashion, HIV prevalence at a certain time would then correspond to prevalence of drug use (= DUs who have not yet entered treatment + DUs who have entered treatment but are still "active"; this inference will require some notion of the distribution of DU activity period after first treatment), while first time HIV positivity (something that precedes or, sometimes, coincides with AIDS diagnosis) would correspond to information about drug users before or at first treatment. this kind of information could arise from linking police, social and treatment registers or drug overdose recoveries or deaths again linked with treatment registers, so as to distinguish individuals already known to the treatment system from "new" individuals and also previously reported individuals from first time reports.

Modeling the probability of being in treatment

A very recent development in methods to estimate DU incidence is the method proposed by Nordt & Stohler (2008). The method is based on the idea to model the probability of being in treatment at a given time after DU onset as a function of the time distance between the given time and onset. This "general inclusion function", as the authors call the probability model, is estimated by fitting a statistical model to treatment data for a long period (1992-2004) in the canton of Zürich, Switzerland. The authors then document the apparent stability of this "general inclusion function" over time in their data and the compatibility of the resulting incidence estimates with other estimates of prevalence and OD mortality rate and then propose 'incidence estimation from data on a single treatment day" as a method that may be extended to other settings.

Once a "general inclusion function" is available, the subsequent estimation of incidence from treatment data is rather straightforward, since it would proceed according to what is commonly called a "multiplicator method" in e.g. DU prevalence estimation. As an example, suppose that 10 individuals stating that they started DU 10 years ago, with respect to the study date, are found in treatment on a given day and that the model for probability of being found in treatment 10 years after DU onset yields a value of 20% for this probability, the incidence of DU 10 years ago can be estimated by dividing 10 by 0.20, i.e. 50 individuals, since the observed individuals should correspond to 20% of the total to be estimated.

Further developments, in particular with respect to generalizability to other settings and data types and interpretation of the estimated "general inclusion function", will show the actual potential of this very interesting approach.

General conclusion about methods

In general, the less data is available, the more assumptions, essentially non-verifiable, will be needed in order to estimate the past DU incidence. At some point it is probably better to admit that there is not sufficient information for incidence estimation, rather than producing a very hypothetical incidence estimate. One may be better off by qualitative conclusions of the type "drug use in one country is similar to that in another country, with better known incidence history", "workers in the area agree that the phenomenon is increasing/decreasing", etc. It may also be argued that good prevalence statistics are more important than incidence statistics for the immediate prevention work, but as soon as one gets interested in the question, natural questions about trends in incidence will arise...

The examples above of methods and data other than the "main" RDA and BC methods used with first treatment data also show that there is room for innovation and that one could, and maybe should, try to do one's best with the available information. Prevalence of DU is a most important indicator, but incidence reflects the dynamics of the phenomenon and other aspects, such as duration of drug use, LP until first contact with treatment or probability of being in treatment, all contribute to a more nuanced knowledge about the DU dynamics.

6.6. Analyzing Latency Period data

Survival times, truncation and censoring, observation schemes

The analysis of survival time data is an important and still developing part of Statistics (see Collet, 1994; Marubini & Valsecchi, 1995). The classical applications are industrial (lifetime of products, machines, etc...), medical (survival of patients with and without treatments) and demographic/economic (duration of various stages of human life, applications to insurance and pensions).

The basic situation is that it is assumed that a given phenomenon has a duration well described by a probability distribution, that a sample (representative and independent observations of the phenomenon) is available and that one wishes to estimate characteristics (i.e. mean, median, variance) of the underlying distribution, the shape itself of the probability distribution or, in the case the distribution is assumed to be of a certain analytical form with some unknown parameters, to estimate these parameters, which will then characterise the distribution. This situation is common to most applications of Statistics.

What makes the analysis of survival time data a field of its own, is that certain complications, related to how durations can be observed, are frequent and, if these complications are not recognized and proper remedies/methods adopted, that the estimates produced in some standard way risk being biased, i.e. not represent the "true" distribution correctly.

Rather than giving mathematical definitions, some examples will illustrate the standard situations:

Right censoring and the importance of a correct analysis

- one starts up 100 lamps in order to see how many days they will work before failing. During the first month, 60 lamps fail at different times. At the end of the month, it is decided to end the experiment. Thus one has 60 observed failure times, all shorter than one month and the observation that 40 lamps were still burning after a month, indicating that their failure times must be larger than one month, but without information on when these 40 lamps would have failed. This situation is called "censoring from the right" (or right censoring), because we know that 40 lamps had values "to the right of" one month (i.e. larger) but we are not allowed to see these values (they have been "censored").

It is quickly realised that we can not e.g. estimate the average duration of a lamp correctly from this data set. If we just take the mean of the 60 "really observed" values, this will underestimate the true average duration, because we have disregarded the 40 observations that would certainly have had larger values, if these had been observed, than the 60 observed ones. Not even giving the value "one month" to these and then taking the mean of the 100 values thus produced really helps, since the attributed value is again certainly too small. A little reflection shows that we have only acquired knowledge about how the probability is distributed in that part of the distribution that concerns values up to one month and, if no assumption is made about how this part of the distribution should provide information about the unobserved part, then no further conclusions can be drawn about the shape of the unobserved part, only the conclusion that his part of the distribution stands for approximately 40% of cases.

If, however, the distribution had been assumed to belong to a certain parametric family (Gamma, Weibull, Lognormal, etc.), then, essentially, knowing the shape of one part

already determines the parameters and thus, implicitly, the shape of the other parts as well, even if unobserved. It would thus seem to be an advantage to rely on parametric models in this case, but one must not forget that one is relying on a probably unmotivated and effectively unverifiable assumption in order to make some more "definite" conclusions...

Left censoring

- the same situaton may be verified, but "to the left" (left censoring). Suppose, for instance, that we set free 100 animals of some type and of the same age in an enclosed forest area and that we come back after one month count how many of the animals are still alive (there is the problem of tracing all the living animals, but we may have radio transmitters...). Supposing that we find 80 still living, we can now conclude that 20 had "durations" shorter than a month, but we do not know what these durations were, exactly.

Interval censoring

- this "censoring" mechanism, that allows us to know that a survival time has a value satisfying some condition (larger than a month, smaller than a month) but not its exact value, can be further generalized to "interval censoring" (this would have happened if we had returned to the forest above after a further month and then only found 60 animals alive, with the conclusion that 20 animals had survival times between one and two months, without furher details...

Different censoring patterns

- in fact, censoring does not have to occur according to the same mechanism for all individuals. Consider, for instance, patients admitted to hospital for a well defined problem. They then receive some treatment and the survival time from this moment is studied. Suppose that this study starts a given day and that follow up of patients will be performed for one year after the starting date. Then a patient arriving after one month can only be followed up for 11 months, one arriving after 4 months has a maximum follow up of 8 months, etc. If some of these patients survive longer than their respective follow up times, they will have different right censoring times. This scheme is called "staggered entry within a fixed follow up period in calendar time".

Difference between censoring and truncation

- the censoring problem is thus that we have a well defined number of subjects and for some of these we may obtain precise data, for the other censored data, but we have some information for every subject. This situation is different from another important complication, called truncation. Essentially, there may be certain kinds of survival values (short ones, long ones, ...) that we will not be able to observe, but we will not even know how many such values we have not observed. Consider e.g. "mad cow disease"(BSE, bovine spongiform encefalopathy) and the (possibly) related vCJD (variant Creutzfeld-Jacob disease) disease in humans.The BSE epidemic may occur at a more or less well defined time and an unknown number of individuals are exposed in such a way as to (possibly) cause vCJD. When cases of vCJD are found, the duration of the period between exposure and diagnosis can be established and, after some time, there will be a data set of durations. We can also conclude that durations longer than our observation period can not have been registered yet, but in this situation we do not know how many these will be. We then say that our observations are right

truncated. The conclusions that can be drawn from the data are different from the right censored case, in this case we will only have information about the "distribution of durations, conditional on the duration being shorter than L", say, where L is the length of the observation period. We will know the shape of a part of the distribution, but not the corresponding level or total corresponding probability. Also in this case, assuming a parametric form for the unknown distribution, might, in apparence, solve the problem, since even this kind of data may be enough to identify the parameters of the distribution, but even more faith than in the corresponding censoring situation has to be had in the assumptions, since here both shape and absolute level of what is unobserved will depend, through the assumptions, on what has been observed.

Covariates and survival times

Just as in other statistical situations, one may be interested in studying the influence of covariates on the survival times of interest. The term covariate can be given a very wide interpretation: it may simply be the classification of individuals in two or more well defined groups (the group membership is then the covariate of interest, i.e. a label that can be attributed to each individual), it may be a continuous phenomenon, such as age at first drug use, it may be something that may or may not happen during the survival time, such as an arrest or an overdose. Again, the peculiarity of survival data that makes special statistical methods necessary, is the presence of censoring and truncation and other possible complications in the observation plans. This is the reason why terms like "Cox proportional hazard model", "log-rank test" or "generalized Wilcoxon tests" will be encountered, instead of perhaps more familiar statistical terms like "2 by 2 table", t-test" or "multiple linear regression". For somebody not wishing to dwelve too much on technical matters, the simplest way of thinking about survival times and covariates is to group data into a few groups with similar or equal, if possible, covariate values within each group, then estimate the probability distribution within each group in a graphical form and then compare the results from the different groups to see whether there are marked differences between the resulting shapes. This may however not always be possible or desirable and then more technical methods will have to be employed.

Latency periods between onset of drug use and first treatment request

Finally coming to latency period observations, the usual situation is that there is a data set consisting of individuals registered in treatment centres, for which the time of first drug use is known. (The question answered may be about the age at first use or the year of first use and the latency time is then computed from knowledge about the year in which the data was collected or the age of the respondent at that time.) Then the following considerations are of importance for the interpretation of the data:

- it is important to ensure that only first treatment request or admissions are included in the data set. Data must represent a homogeneous phenomenon to be useful; if some individual are at their first visit and others at some subsequent visit, any subsequent conclusion will be difficult to interpret. It might even be important to distinguish between the two words used above, request and admission, if these two things do not always coincide. Also possible changes in the way treatment has been available over time may influence the meaning and comparability of latency times. It is not clear how one should model changes in the treatment availability in an analysis of latency time: it seems reasonable to use DU onset cohort as a covariate, implying that individuals starting DU at different oiunts of time may have different LP distributions; one could also represent treatment availability as a real time effect, i.e. acting directly on the "hazard" of starting treatment at a given point in calendar time (note that it is not

calendar time directly that is the covariate but some measure of treatment availability and/or attractiveness in calendar time). Finally, the definition of the start of the latency time should be made as precise as possible, e.g. by making clear in used questionnaires whether first substance use ever or start of problematic use is intended, and this definition should be based on the same substance as the one leading to first treatment entry.

- in many ways the data will be similar to the data discussed in connection with the RDA method, although, for purposes of latency time estimation, there is no need to have full coverage of all treatment requests in a given area, a (representative) sample is sufficient.

- in the previously presented sections on RDA and BC, the latency time distribution was used in our models for the data with the following implicit interpretation: a certain number of individuals start their drug use a given year and to each one of these a latency time will be "assigned" from the LP distribution. Thus the natural way to organize data is to group together all those who have the same starting year (drug use onset time) and then try to estimate the LP distribution from such a group. The problem is that data from such a group is necessarily truncated (e.g. onset in year 1999, study containing admissions between 1990 and 2005, truncation = 6 years...), different onset cohorts have different truncation times and some way must be found to merge the information from the different groups. This problem has already been faced by the RDA method, where it was concluded that if the longest observation period (in our hypothetical example above, 15 years) is sufficiently long to exceed the "maximum possible delay", then the RDA method can be used to estimate the delay distribution (in our case, the LP distribution) without truncation. In fact, using the RDA method on data is among the least complicated ways of estimating the LP distribution from "RDA-like data". The estimated incidences may however be meaningless, because they refer only to the subgroup of DUs "that gave rise to the sample", if the data do not contain data on all DUs. Furthermore, the LP distribution will be rather coarse, represented by a series of delay probabilities, but in exchange the analysis is quite easy to perform with standard statistical software. This way of organizing the estimation problem is sometimes called "forward estimation", because the LPs are measured "forwards" from a given drug use onset time.

Difference between DU onset cohorts and treatment entry cohorts for the estimation of the LP distribution

It is very important to realize that the grouping of data by drug use onset cohorts is not equivalent to grouping data by treatment onset cohorts (sometimes referred to as "backward estimation"). In order to understand the problem better, we again consider the synthetic example presented in the RDA section. There, it was assumed that the LP distribution was specified by

$p(0)=1/5$

$p(1)=1/5$

$p(2)=2/5$

$p(3)=1/5$

i.e. only delays between 0 and 3 are possible. Then with given DU onset cohorts, the following table shows the expected numbers of individuals from the different cohorts that will enter treatment, together with the situation after year 3, if treatment admissions have been observed from year 0.

Onset year	Cohort size (n)	Year of entry in treatment				Observed per onset cohort	Not yet observed
		0	1	2	3		
0	10	**2**	**2**	**4**	**2**	**10**	0
1	30	**-**	**6**	**6**	**12**	**24**	6
2	20	**-**	**-**	**4**	**4**	**8**	12
3	10	**-**	**-**	**-**	**2**	**2**	8
# entering treatment		**2**	**8**	**14**	**20**	**44**	26

The previously discussed forward estimation will find 10 subjects in DU onset cohort 0 and these have delays in proportions 1:1:2:1, which of course are the correct proportions from the LP distribution; there will be 24 subjects in cohort 1, with delays in proportions 1:1:2:0 (the longest delay has been truncated); 8 subjects in cohort 2, with proportions 1:1:0:0 (two longest delays truncated), and the last cohort delivers little information. However, although truncated, proportions from all cohorts are correct with respect to the "true" LP distribution, which has proportions 1:1:2:1.

If we instead look at the corresponding proportions calculated for the delays reported by treatment onset cohort 3 (column marked 3), say, we find 1:2:6:1 (remember that the proportions are ordered according to delay=0,1,2,3, which means reading the column from the bottom to the top), indicating that delay=2 is 6 times more common than delay=0, for instance, which is not true. Treatment onset cohort 2 yields the proportions 2:3:2:0; again there is a truncation problem, but more interesting, the proportion information does not coincide with that from the previous cohort 3.

Some reflection shows that these effects are caused by a mixture of the LP distribution with the DU incidences. If, as in the case of cohort 3 above, there was a peak in incidence "long ago" (year 1), then "long" LPs (LP=2) will be overrepresented in the treatment cohort. A consequence is that the information in the different treatment cohorts would be the same and also correct only if the DU incidence had been constant over the whole period.

In conclusion, analysing LP data from a treatment onset cohort ("backwards estimation") will yield correct results only if it can be assumed that DU incidence has remained constant over a sufficiently long period (more or less corresponding to the maximum probable delay). The reason for considering the possibility to use "backwards estimation", despite its potential problems and biases, is that the alternative requires data collected over a long time period and special methods of analysis, while just one year, say, would be enough to define a treatment onset cohort. It is also the case that the most recent treatment cohorts suffer less from truncation problems, since the truncation is now measured from the beginning of the DU phenomenon. Therefore, under the assumption of constant incidence, data can be analysed as if it were not truncated, with standard survival data analysis methods.

In the example below, based on an analysis of Italian data, the reasonability of carrying out a "backwards" analysis is first verified and then data are analysed as if they collectively represented the LP distribution, without truncation problems. A certain quantity of survival data analysis related terms is also introduced and explained.

Example: The latency period analysis for the Lazio region (Italy)

The data comes from a multicentric longitudinal study named VEdeTTE. This study was carried out between 8 January 1998 and 18 October 2000 in 13 Italian regions to evaluate the efficacy of treatments for heroin addiction.

As mentioned before, the latency period can be analysed by “entry” cohort (i.e. by year of first treatment) or by “onset” cohort (i.e. by year of first use). If incidence is reasonably constant, they produce the same estimates. However, if incidence changes over time, analyses by entry cohort may be biased as they tend to produce decreasing observed periods when incidence is increasing and viceversa.

As a first step, two such estimates were therefore computed and compared. The entry cohort is denoted by Trtfst98 and includes all the individuals starting their first treatment in 1998 (onset since 1985). The onset cohort is denoted by Usefst87 and includes all the individuals starting first drug use in 1987.

The results of Kaplan-Meier (estimation of the survival distribution, i.e. for each time the probability that the LP will exceed that time, in a nonparametric, i.e. without specific model assumptions, with a method that also works with right censored data) estimation for the two data-sets are shown in the following graph. As can be seen, the two curves are not very different.

In the following table summary statistics show that also the main statistical indices are not so different.

Figure 6.6.1 Comparison of survival distributions derived from backwards and forwards estimation within the same data set.

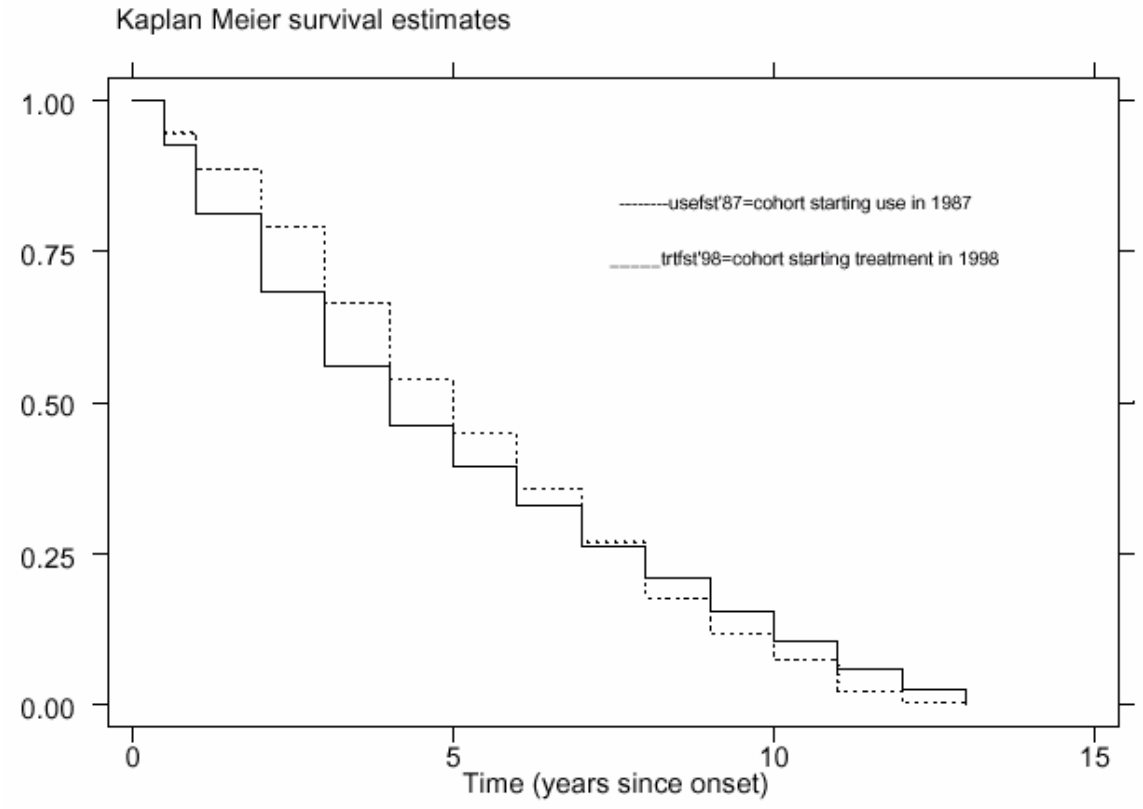


Table 6.6.1 Some statistical summaries of the two estimated distributions

	Observed cases	Mean survival	C.I 95%	25% quantile	50% quantile (median)	75% quantile
Trtfst98	880	5.01	(4.78, 5.25)	2	4	8
Usefst87	749	5.31	(5.09, 5.54)	3	5	8

A t-test was performed to compare the two mean estimates in a more formal way and confirmed that the means are not significantly different. The distributions also do not differ significantly.

For both data-sets “Age at first use” is the only covariate which has a significant influence on LP. This is checked by using the Cox proportional hazard model, a method to estimate and test the effects of covariates on survival times. The effect is modelled as a constant increase or decrease of the hazard, i.e. the short term risk of experiencing the terminal event (in this case treatment entry) - essentially the question is if large values of the covariate tend to make the survival times shorter or longer than average. The method is used to test for significant covariate effects but it is quite difficult to translate the corresponding effect estimates to changes in probabilities or averages. The results of the stratified analyses - which are much easier to interpret, but grouping according to "homogeneous" covariate values is required - are reported in the tables 6.6.2 and 6.6.3.

Comparing the statistical summaries and the confidence intervals, it can again be seen that the backward analysis gives results quite similar to the forward analysis. The means and quantiles shown in the tables (average and median less than 6 years) indicate that the time period covered by the study (DU onset since 1985) is sufficiently long to reduce truncation problems, at least for recent treatment cohorts.

Table 6.6.2 Statistical summaries of LP distributions stratified by "age at first use" for the data set Usefst87 (DU onset cohort)

Dataset Usefst87 (onset)		Observed cases	Mean survival	C.I 95%	25% quantile	50% quantile (median)	75% quantile
Age at first use	9-17	247	5.84	(5.46,6.23)	3	6	8
	18-22	389	5.28	(4.98,5.60)	3	5	8
	23-28	97	4.37	(3.8,4.94)	2	4	6
	≥29	16	3.5	(1.9,5.1)	0.5	3	6

Table 6.6.3 Statistical summaries of LP distributions stratified by "age at first use" for the data set Trtfst98 (treatment entry cohort)

Dataset Trtfst98 (entry)		Observed cases	Mean survival	C.I 95%	25% quantile	50% quantile (median)	75% quantile
Age at first use	9-17	233	6.02	(5.55,6.49)	3	6	9
	18-22	414	5.14	(4.79,5.49)	2	4	8
	23-28	170	3.97	(3.5,4.44)	1	3	6
	≥29	63	3.24	(2.51,3.98)	1	2	5

For exploratory purposes, we now assume that incidence has remained reasonably constant and that (most of) the data is not affected by truncation problems. We then perform some analyses on the whole Lazio data set. The total number of reports for this region was 1735 (overall sample), but only 1555 records are complete with first use and first treatment age. Different sample sizes, with the same order of magnitude, were used for the various stratifications studied, due to the elimination of cases with missing data.

Table 6.6.4 and Figure 6.6.2 show the estimated survival function (Kaplan-Meier curve) with a median of 4 years and quartiles at 2 and 8 years.

Table 6.6.4: Survival function of the latency period estimated by the Kaplan-Meier method for the total sample from the Lazio region.

Time (years)	Survival function	(standard error)	Time (years)	Survival function	(standard error)
0	1		11	0.10	(0.0075)
1	0.87	(0.0084)	12	0.08	(0.0067)
2	0.73	(0.0113)	13	0.06	(0.0060)
3	0.60	(0.0124)	14	0.04	(0.0053)
4	0.48	(0.0127)	15	0.03	(0.0045)
5	0.39	(0.0124)	16	0.02	(0.0039)
6	0.31	(0.0117)	17	0.02	(0.0033)
7	0.25	(0.0110)	18	0.01	(0.0028)
8	0.19	(0.0100)	19	0.009	(0.0024)
9	0.16	(0.0092)	20	0.005	(0.0018)
10	0.13	(0.0086)	>20	0.008	

Table 6.6.5: Means and standard errors, medians and quartiles for the whole sample and for the sample stratified by sex, educational level and age at first use.

Grouping		# Obs.	Mean	Std.Error of mean	25% quartile	Median	75% quartile
Whole sample		1555	5,51	0,11	2	4	8
Sex	women	241	5,15	0,27	2	4	7
	men	1314	5,57	0,11	2	4	8
Ed. level	lower	1071	5,27	0,12	2	4	7
	higher	477	6,06	0,21	3	5	8
Age at first use							
	9-17	508	6,26	0,19	3	5	8
	19-22	753	5,54	0,15	2	4	8
	23-28	233	4,36	0,24	2	3	6
	29-52	77	3,63	0,40	1	2	5

The influence of possible covariates (sex, educational level and age at first use of drug) on the latency period distribution is reported in Table 6.6.5. P-values for hypothesis testing are reported in the box below. There is no highly significant effect of sex, meaning that the latency period has approximately the same distribution for both sexes (see also Figure 6.6.3). The other two covariates are highly statistically significant (p=0.000 << 0.05), as can be seen in Figs 6.6.4 and 6.6.5.

Log-Rank and Wilcoxon test P-values.

Covariate:	log-Rank	Wilcoxon
Sex	0.185	0.047
Educ. level	0.000	0.005
Age1°use	0.000	0.000

Figure 6.6.2: Kaplan-Meier estimate of the survival function of the latency period for Lazio

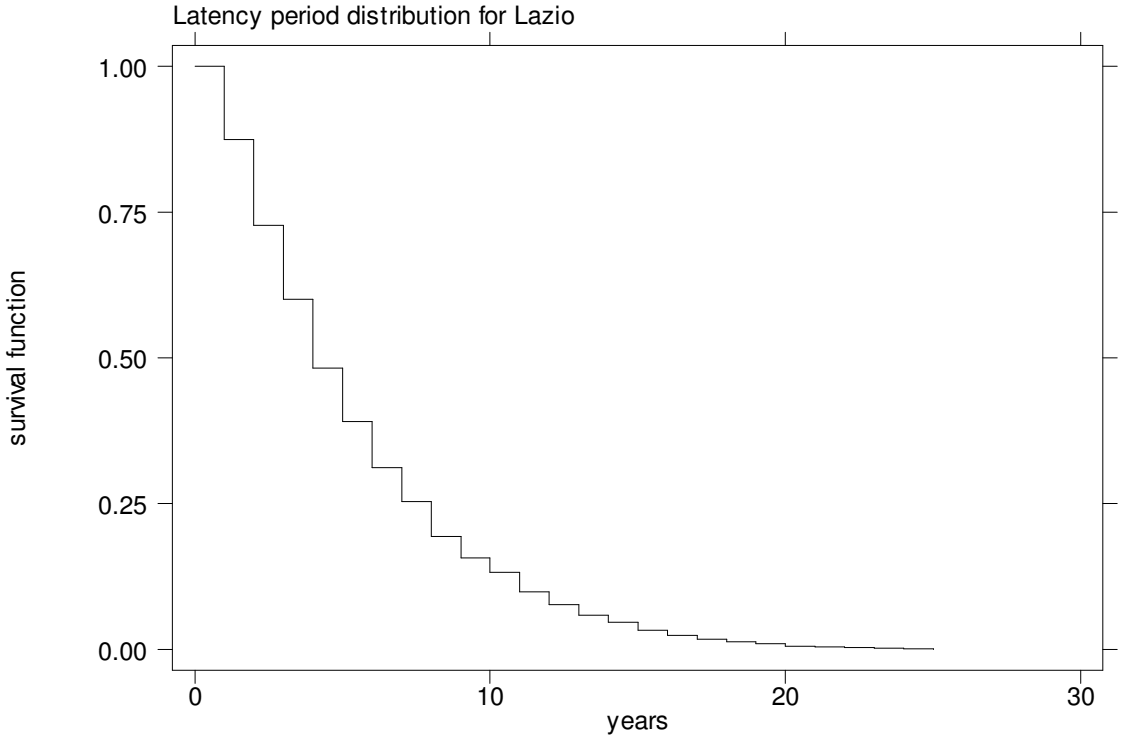


Figure 6.6.3: Kaplan-Meier estimate of the survival function of the latency period (Lazio region), for the sample stratified by sex

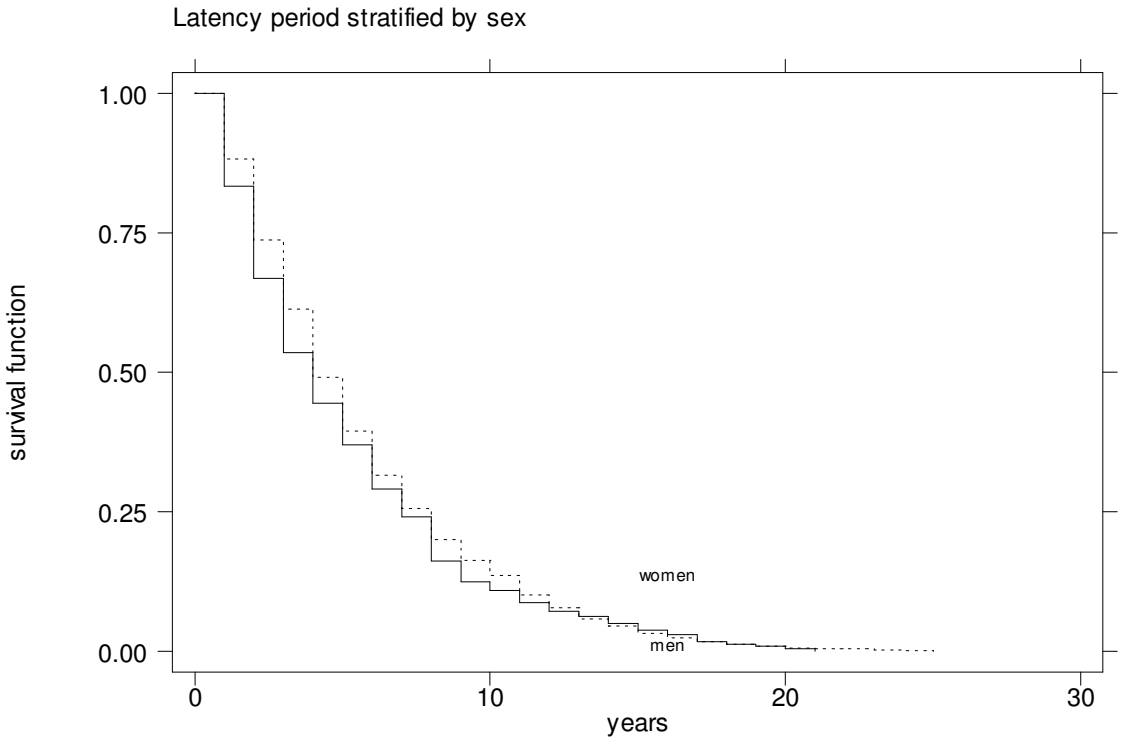


Figure 6.6.4: Kaplan-Meier estimate of the survival function of the latency period (Lazio region), for the sample stratified by educational level

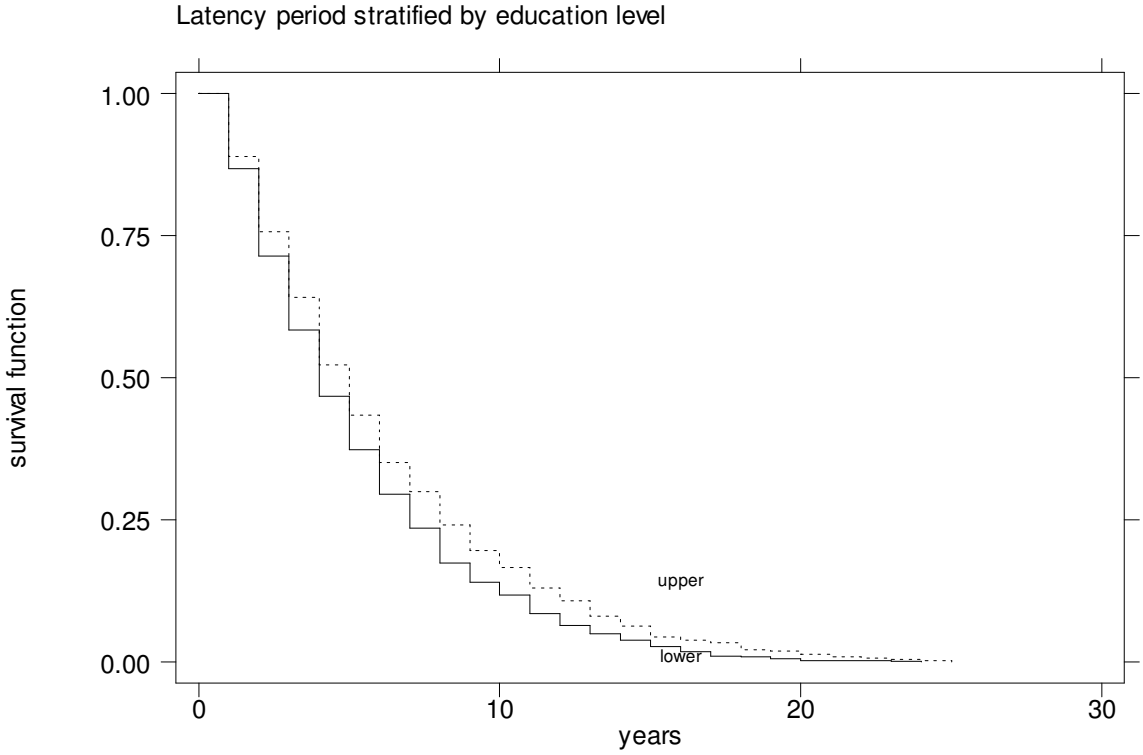
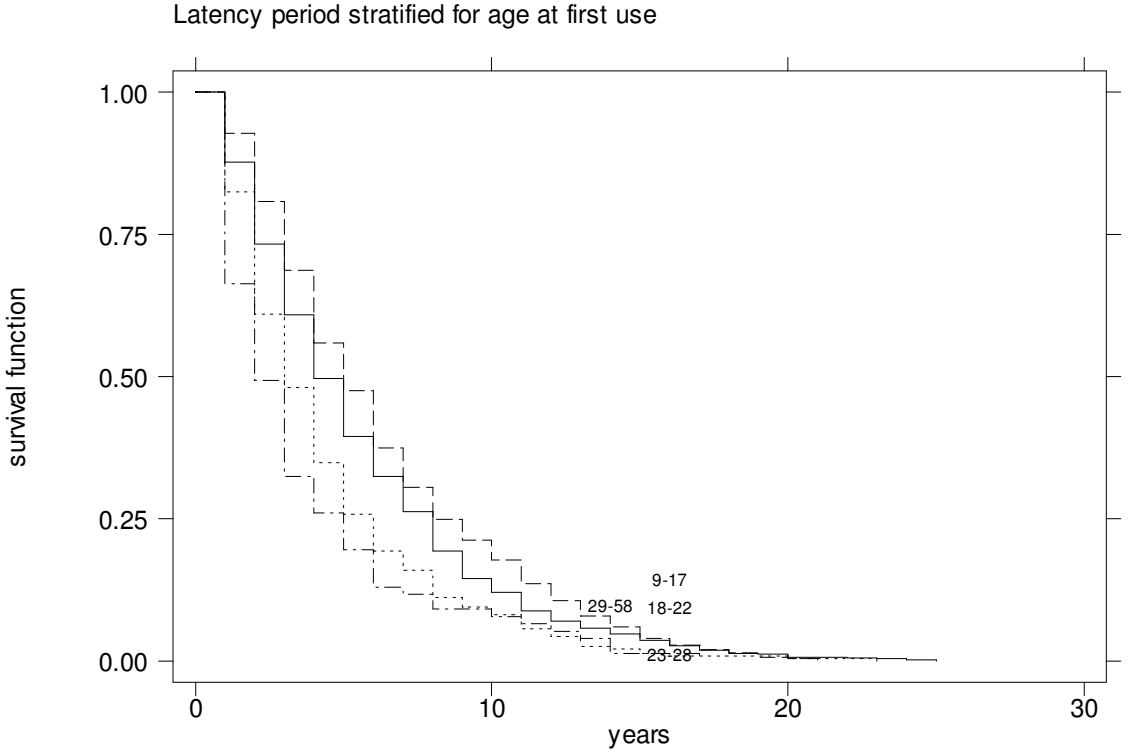


Figure 6.6.5: Kaplan-Meier estimate of the survival function of the latency period (Lazio region), for the sample stratified by age at first use (four classes)



The same analysis for effects of covariates was also performed using the Cox proportional hazard model, thus without stratification. The main results are reported below. The influence of the covariates is measured by the hazard ratios, with values near to 1 indicating no effect, values lower than 1 indicating longer survival times with increasing value of the covariate and values above 1 shorter times with increasing value of the covariate. The significance of the effect cannot be evaluated directly from the magnitude of the estimated effect, it is necessary to consult the appropriate p-value.

Covariates	Haz.Ratio	Std.Err.	z	P >|z|	[95% Conf.	Interval]
Sex	0.8541	0.06106	-2.21	0.027	0.7424	0.9825
Ed. level	0.7837	0.04439	-4.30	0.000	0.7014	0.8758
AgeFirstUse	1.0494	0.00589	8.59	0.000	1.0379	1.0610

All three covariates have statistically significant effects, although sex less than the other two. By interpreting the hazard ratio as relative risk (and knowing the coding that has been used in the data set), we can say that, among drug users eventually entering treatment:

- Females have a 15% higher chance at any time during their drugs career than males to enter treatment, as males have on average longer latency periods to first treatment;
- Drug users with a lower educational level have a 22% higher chance than those with a higher educational level to enter their first treatment, at any time during their drugs career;
- Drug users who were one year older when they started using heroin have an about 5% higher chance at any time during their drugs career to enter first treatment, as starting at an earlier age yields longer latency periods, on average.

7. Final remarks

In these guidelines, two main different drug use incidence estimation methods (Back-Calculation and Reporting Delay Adjustment method) are presented. Both methods are based on observing all first treatment episodes in the target population over an extended period of time. In both methods, the concept of LP (latency period and its distribution) plays an important role. However, each method is adapted to a certain data structure: the RDA to the availability of individual level data on first DU and first treatment, the BC to having only aggregated first treatment data (requiring, in addition, a separate estimate of the LP distribution). The separate estimation of the LP distribution is also discussed. In addition, many explanatory comments on the presented methods are presented.

The main theme of this report is however the usefulness of statistical modelling, both as a tool for better understanding of assumptions and relations between observation and quantities of interest and as an efficient way to define statistical procedures (and their properties) for the analysis of data. The methods presented are not the only ones applicable for the estimation of DU incidence, just those that apply to commonly available data and that have already been used in the literature. New methods, maybe applied to different kinds of data, will certainly appear in the future and modifications of the old methods as well. Therefore it is hoped that these guidelines have also presented the bases for understanding new developments, at least in a practical (as opposed to mathematical/technical) sense.

Another important message, maybe not explicitly emphasized but implicitly present all the time, is that estimating DU incidence is worthwhile as a tool to understand the DU dynamics. Prevalence of DU may be the most important indicator for immediate resource planning, but understanding incidence is fundamental for policy planning. Thus, efforts should be made to locate data sources suitable for inference about DU incidence and to allocate resources to statistical elaboration/research in order to make the best possible use of the available data. It may also be that the information content of available data can be improved by linking, maybe with subsequent anonymization, different registers or introducing small modifications in existing data collection routines. It is an old statistician's saying that "Using available data is OK, but wishing for better data is better" :-)

It is acknowledged that the mathematical formulation of statistical procedures seems complicated to most non-statistical/mathematical people. The proper reaction should then be to make a little effort to understand, but also to call in a statistician to collaborate with. Another saying, this time attributed to the great statistician Fisher is "Calling in a statistician only at the end of an experiment, usually results in him being able to diagnose only what the experiment died of", meaning that statistical expertise is needed in all the stages of estimation work, from the start (planning, definition of target population, formulation of questionnaires, etc) to the end (the statistical analysis of data, the formulation of the proper conclusions, etc).

8. References

Armstrong GL, Alter MJ, McQuillan GM, Margolis HS. (2000) The past incidence of hepatitis C virus infection: implications for the future burden of chronic liver disease in the United States. Hepatology 31: 777-782.

Brookmeyer R, Gail MH (1994) AIDS Epidemiology. A quantitative approach. OUP, New York&Oxford.

De Angelis D, Gilks WR, Day NE (1998) Bayesian projection of the AIDS epidemic. Applied Statistics 47,4, with discussion.

De Angelis D, Hickman M, Yang S.(2004) Estimating long-term trends in the incidence and prevalence of opioid use/injecting drug use and the number of former users: back-calculation methods and opioid overdose deaths. Am J Epidemiol 160, 994-1004.

EMCDDA, “Pilot project to estimate time trends and incidence of problem drug use in the European Union”, Final report, Lisbon, 1999a.

EMCDDA, "1999 Annual Report on the State of the Drugs Problem in the European Union", Lisbon, 1999b.

EMCDDA, “Study on incidence of problem drug use and latency time to treatment in the European Union”,(CT. 99.EP.05), Lisbon, 2000.

EMCDDA, "Trends in injecting drug use 2001-2005", Lisbon, 2007a, <http://www.emcdda.europa.eu/html.cfm/index39640EN.html>

EMCDDA, "Annual report 2007", Lisbon 2007b, <http://www.emcdda.europa.eu/html.cfm/index419EN.html>

Glad IK, Frigessi A, Scalia Tomba G, Balducci M, Pezzotti P (1998) Bayesian back-calculation with HIV seropositivity notifications. Technical rep. no. 4/98, Dept. of Mathematics, Univ. of Oslo, Norway.

Hickman M., Seaman S., De Angelis D.(2001) Estimating the relative incidence of heroin use: application of a method to adjust observed reports of presentations at specialist treatment agencies, American Journal of Epidemiology, 153(7), 632-641.

Hickman M. (2006) The diffusion of heroin epidemics: Time to re-visit a classic. Int J Drug Policy, 17(3), 143-144.

Hunt L (2006) Reprinted classic: Recent spread of heroin use in the United States. Int J Drug Policy, 17(3), 145-153.

Nordt C, Stohler R (2006) Incidence of heroin use in Zurich, Switzerland: a treatment case register analysis. The Lancet 367, 1830-1834.

Nordt C, Stohler R (2008) Estimating heroin epidemics with data of patients in methadone maintenance treatment, collected during a single treatment day. Addiction, doi:10.1111/j.1360-0443.2007.02055.x

Ravà L., Rossi C., “Estimating the size of a hidden population involved in the HIV/AIDS epidemic: a method based on Back-Calculation and dynamical models”, in Simulation in the Medical Sciences, Anderson & Katzper eds., The Society for Computer Simulation, San Diego, California, 57-62, 1999.

Further relevant literature

Brookmeyer R and Damiano A. Statistical methods for short-term projections of AIDS incidence. Statistics in Medicine 1989; 8: 23--34.

Brookmeyer R., Gail H.G., A Method for Obtaining Short-term Projections and Lower Bounds on the Size of the AIDS Epidemic. J. of the American Statistical Association 83, 301-308, 1988.

Brookmeyer R., Gail H.G., Minimum size of the acquired immunodeficiency syndrome (AIDS) epidemic in the United States, Lancet, 2, 1320-1322, 1986.

Brookmeyer R., Liao J., "The analysis of delays in disease reporting: methods and results for the acquired immunodeficiency syndrome", American Journal of Epidemiology, Vol. 132, No. 2, 355-365, 1990.

Brookmeyer R., Reconstruction and Future Trends of the AIDS Epidemic in the United States. Science Articles, 37, 37-253, 1991.

Collet D. "Modelling survival data in medical research", Chapman and Hall, London, 1994.

Efron B, Tibshirani R.J., "An Introduction to the Bootstrap, Chapman and Hall, London, 1993.

EMCDDA Modelling drug use: methods to quantify and understand hidden processes", EMCDDA Scientific Monograph Series, No 6, 2001.
http://www.emcdda.europa.eu/html.cfm/index428EN.html

Marubini E., Valsecchi M.G., "Analysing survival data from clinical trials and observational studies", Wiley, NY, 1995.

Menard S. Longitudinal Research, Sage University Paper on Quantitative Applications in the Social Science, series no. 07-076. Newbury Park, CA: Sage, 1991.

Pompidou Group Project on Treatment Demand: Tracking long-term trends. Guidelines for the estimation at local level of incidence of problem drug use (prepared by Carla Rossi), August 2002.

Pompidou Group Project on Treatment Demand: Tracking long-term trends. Final Report, 2003, P-PG/Epid (2003) 37.

Ravà L., Calvani M.G., Heisterkamp S., Wiessing L., Rossi C. "Incidence indicators for policy making: models, estimation and implications", UN Bulletin on Narcotics, Vol. LIII (1-2), 135-155, 2001.

Rosenberg P., Gail M.H., Backcalculation of Flexible Linear Models of the Human Immunodeficiency Virus Infection Curve. Applied Statistics 40, 269-282, 1991.

Rosenberg PS. A simple correction of AIDS surveillance data for reporting delays. JAIDS 1990, 3(1): 49-54.

Rossi C., Monitoring drug control strategies: hidden phenomena, observable events, observable times, International Journal of Drug Policy, 10-1, 131-144, 1999.

Tanner M.A., Tools for Statistical Inference, Springer-Verlag New York, 1996.

Zeger SL and Lai-Chu S. Statistical methods for monitoring the AIDS epidemic. Statistics in Medicine 1989; 8: 3--21.

Appendix 1: The RDA method and log-linear modelling

There is a simple relation between the model formulated in Section 6.3 and the models commonly called Generalized Linear Models (GLM) and this has the advantage of allowing standard statistical software to be used (with a small amount of adaptation of data and results) for the estimation and even to extend the RDA method to include covariates such as sex, age categories, etc and also some kinds of exception from the rule of complete observation of a series of years longer than th maximum delay..

If we assume that the variation around the average values E(X(j,k)) = n(j)p(k-j), for k=0,1,...,T and j=0,1,...,T is of Poisson type, the estimation problem can be restated as follows:

- first reparametrize, setting d= k-j, i.e. expressing the observation (entryyear=j, treatmentyear=k) as (extryyear=j, delay=d); we will thus have the observations {X(j,d)} with both j and d between 0 and T, but with $0 \leq d \leq T-j$, i.e. the data fills only half of a square data table;

X(j,d) have independent Poisson distributions, for k=0,1,...,T and j=0,1,...,T, with means $\mu(j,k)$ that satisfy $\log(\mu(j,d)) = \alpha(j)+\beta(d)$, i.e. a classical GLM with log-link and Poisson distribution and categorical covariates j and k.

The original parameters are recovered by setting

$p(i) = \exp(\beta(i))/(\text{sum of all } \exp(\beta(d))$

$n(j) = \exp(\alpha(j))x(\text{sum of all } \exp(\beta(d))$

i.e. the original parameters are proportional to the exponential of the estimated ones (because of the log-linear formulation) and the proportionality is resolved by setting the sum of {p(i)} equal to one (thus assuming that the delay distribution is complete, i.e. that the maximum possible delay was less than the time span observed or accepting that the {p(i)} estimated represent the distribution of delay conditional on being less than or equal to the maximum delay observed, even if this was less than the maximum delay possible and, in this case, only estimating "relative incidence".

Several comments can now be made.

- if we are in the situation where maximum observed delay is shorter than maximum possible delay, we set the sum of estimated [p(i)} equal to one if we do not know what the true value of this sum is. By setting it to one, we get the conditional distribution. However if we knew, or could estimate from external data, what the true sum was (say that we only have observations up to 5 years of delay, but that we are willing to assume that the "true" probability of entering treatment within 5 years of DU onset is 60%, based on other studies) then we can set the sum above equal to the assumed value and then obtain "absolute incidences" instead of relative ones. Of course, we better be right about our assumption...

- it is interesting to consider uncertainty estimates (i.e. standard errors) in addition to parameter estimates. The software will deliver such estimates for the parameters $\{\alpha(j)\}$ and $\{\beta(d)\}$. The standard errors of our parameters of interest can then be worked out, approximately, using the so called Delta-method, i.e. by extending the usual univariate approximation var(g(X)) = (g'(E(X))^2)Var(X) to the multivariate case. The relation to

be applied can symbolically be stated as: the covariance matrix of {{p(i)},{n(j)}} equals $M^T CM$, where C is the covariance matrix of {{α(j)},{β(d)}} and M is the matrix of partial derivatives of the transformation between the two set of parameters.

- the GLM framework allows for extra variability or over-dispersion in observations, typically by specifying a negative-binomial distribution instead of the Poisson distribution, but more general forms are available.

- the model allows data missing at random without special fitting procedures; the model also allows "non-square" data structures, for instance more delay years and onset cohorts than observation years, as long as parameters are identifiable (it becomes our responsability to decide that the same set of delay parameters is applicable to all the onset cohorts in the model...)

- it is quite simple to add covariates (if these are known for all subjects in the data), such as sex or age class at onset, to the model; it then becomes important to realize how the (log-)linear model works, i.e. understand the difference between main effects and interactions in the specification of models...

- it is also possible to test some simple forms of changes in the delay distribution, e.g. by introducing an indicator for two or more classes of onset years and then specifying an interaction effect with delay.

All these options require some more-than-basic understanding of the log-linear model and related statistics.

Appendix 2: On the Back-calculation method

There is no standard software for the back-calculation method. Furthermore, there is no standard method either, in a technical sense. One could probably find agreement about the basic formulation of the problem, for instance

E(X(k)) = n(0)p(k) + n(1)p(k-1) + ...+ n(k-1)p(1) +n(k)p(0), for k=0,1,...,T

with {X(k)} Poisson distributed, {p(d)} assumed known and {n(i)} to be estimated, under non-negativity and some smoothness restrictions,

Three different methods have, to our knowledge, been used with DU data: empirical Bayes (Heisterkamp, Rava', Rossi), Bayes MCMC (Scalia Tomba, Rossi) and smoothed EM (De Angelis (2004)). More approaches have been formulated in the field of HIV/AIDS research (see Brookmeyer & Gail (1994)).The differences are mostly technical, i.e. related to how estimation and smoothing is carried out.

We will briefly describe one of these approaches, the Bayesian/MCMC formulation (for more details see Glad et al (1998)).

- the model is defined as above, in a Bayesian perspective, meaning that we attribute an a priori distribution to {n(i)}, in this case a so called non informative smoothing prior. Concretely,

the log- likelihood function will be

$$\sum_{t=0}^{T}\left(X_t \ln(\mu_t) - \mu_t\right), \text{ with } \mu_t = \sum_{i=0}^{t} n(i)\, p(t-i),$$

and the log-prior will be

$$-\frac{1}{\sigma^2}\sum_{i=1}^{T}\left(n(i) - n(i-1)\right)^2,$$

with σ^2 considered as a given smoothing parameter (in reality, various values will be tested, until visual inspection of the output reveals an appropriate degree of smoothing; in theory, this parameter could also be estimated together with the other parameters, but this has not yet been implemented).

- estimation is now performed used the standard Metropolis-Hastings MCMC algorithm, i.e. a series of parameter values is generated by proposing new values according to a given proposal distribution, then deciding whether to keep the previous value also as a new value or to accept the new value according to an acceptance probability calculated at current and proposed values. This series of parameter values constitutes, according to theory, a Markov chain whose stable distribution is the desired a posteriori in the model, i.e. the distribution of parameter values given the observations and from which a posteriori inference can be made. Thus, in practice, one generates a long series, throws away a first part (burn-in, representing the transient part of the series, before stationarity has set in) and then considers the distribution of the remaining values as an approximation to the true a posteriori.

Since a detailed explanation at this point would be either too technical or too schematic, according to taste, we now reproduce a piece of standard C-code that implements this procedure. In theory, the code is self-explaining, but comments are supposed to guide the reader.

```
#include <stdio.h>
#include <math.h>
#include <stdlib.h>

#define T 28                    /* no of periods=years */
                                /* 1=1980, 2=1981, ....., 22=2001 */
#define FNA 13                  /* first non-zero treatment period */
#define SIGMA1 10.0             /* param. of local unif. proposal periods 1...17 */
#define SIGMA2 10.0             /* param. of local unif. proposal periods 18-50 */
#define VARL 1000000            /* used in prior of L: SMOOTHING PARAMETER */
#define START 500.0                     /* initial value for L:s */
#define BURNIN 1000             /* no of initial iterations not used */
#define NITER 21000                     /* no of iterations of T cycles */

#define alfa 0.05642833         /*parameters used to generate the p-vector, called f in this
program, according to a Weibull distribution*/
#define beta 1.26

#define ris "BC-bayes.txt"      /*output file*/

int main(void)
{
static double a[T+1] = {0,0,0,0,0,0,0,0,0,0,0,0,7880,10507,11972,14156,16644, /*the data,
i,e, counts of first treatments/year*/
                        23028,24590,22786,21844,24738,23410,21600,22092,19718,
                        22056,23044}; /* Italia , T=27 and FNA=12 */
static double f[T];
static double S[T+1],sigma[T+1],sumccol[T+1],L[T+1],mu[T+1],newmu[T+1];     /*the
parameters of interest are contained in vector L*/
static double mean[T+1],quad[T+1];
double newL,diff,test,minsigmaL,sigmaval;
static int index[T+1],accept[T+1];
int i,j,iter,ind,seed;
FILE *outf;
extern double ran0(int *);      /*routine that improves the standard C random numer
generator*/

outf=fopen(ris,"w");
seed=12241;

/*The parameters used are alfa (Weibull scale) and beta (Weibull shape)*/
for(i=1;i<=T;i++) S[i]=exp(-alfa*pow(i,beta));
f[0]=(1.0-S[1])/2.0;
f[1]=(1.0-S[2])/2.0;
for(i=2;i<=T-1;i++) f[i]=(S[i-1]-S[i+1])/2.0;
for(i=1;i<=T;i++) fprintf(outf,"%f\t",f[i]);

fprintf(outf,"\n\n");

for(j=1;j<=T;j++)
        {
        if(j<=17) sigma[j]=SIGMA1;
        else sigma[j]=SIGMA2;
        }
```

```
for(i=1;i<=T;i++)
        {
        index[i]=((i>=FNA)?i:FNA);
        accept[i]=0;
        sumccol[i]=0.0;
        mean[i]=0.0;
        quad[i]=0.0;
        for(j=index[i];j<=T;j++) sumccol[i]+=f[j-i];
        }

for(i=1;i<=T;i++) L[i] = START;

for(iter=1;iter<=NITER;iter++)
        {
        for(i=FNA;i<=T;i++)
                {
                mu[i]=0;
                for(j=1;j<=i;j++) mu[i]+=f[i-j]*L[j];
                }
        for(j=1;j<=T;j++)
                {
                sigmaval=sigma[j];
                minsigmaL=((L[j]>sigmaval)?sigmaval:L[j]);
                newL=(sigmaval+minsigmaL)*ran0(&seed)+L[j]-minsigmaL;
                 ind=index[j];
                diff=newL-L[j];
                for(i=ind;i<=T;i++) newmu[i] = mu[i]+diff*f[i-j];
                test=0.0;
                for(i=ind;i<=T;i++) test+=a[i]*log(newmu[i]/mu[i]);
                test-=diff*sumccol[j];
                if(j==1)
                        test+=(2.0*L[2]-L[1]-newL)*diff/VARL;
                else if(j==T)
                        test+=(2.0*L[T-1]-L[T]-newL)*diff/VARL;
                else
                        test+=2.0*(L[j+1]+L[j-1]-L[j]-newL)*diff/VARL;

        test+=log((sigmaval+minsigmaL)/(sigmaval+((newL>sigmaval)?sigmaval:newL)));
                if(((test>=0)?1:(test>=log(ran0(&seed)))))
                        {
                        L[j]=newL;
                        for(i=ind;i<=T;i++) mu[i]=newmu[i];
                        if(BURNIN<iter) accept[j]++;
                        }
                } /*end of one T parameter update*/
        if(iter%1000==0) printf("%6d\n",iter);           /*to screen, for check of progress*/
        if(BURNIN<iter)
                {
                fprintf(outf,"%6d\t",iter); /*output of selected parameters to file*/

        fprintf(outf,"%.1f\t%.1f\t%.1f\t%.1f\t%.1f\t%.1f\n",L[10],L[15],L[20],L[30],L[40],L[50])
;
                 for(i=1;i<=T;i++) /*computations for summary means and variances at the
end*/
```

```
                        {
                        mean[i]+=L[i];
                        quad[i]+=L[i]*L[i];
                         }
                }
        } /*end of iterations*/

for(i=1;i<=T;i++)fprintf(outf,"%5.2f",((double)  accept[i])/(NITER-BURNIN));  /*  proportion
accepted proposals,measure of good choice of proposal*/
fprintf(outf,"\n\n");

for(i=1;i<=T;i++)
        {
        mean[i]/=(NITER-BURNIN);
        quad[i]/=(NITER-BURNIN);
        quad[i]=sqrt(quad[i]-mean[i]*mean[i]);
        fprintf(outf,"\n%6d\t%.1f\t%.1f\t%.1f",i,a[i],mean[i],quad[i]);
        }
return(0);
}


double ran0(idum)
int *idum;
{       static double y,maxran,v[98];
        double dum;
        static int iff=0;
        int j;

        if(*idum < 0|| iff==0)
                {
                iff=1;
                maxran=RAND_MAX+1.0;
                srand(*idum);
                *idum=1;
                for(j=1;j<=97;j++)dum=rand();
                for(j=1;j<=97;j++)v[j]=rand();y=rand();
                }

        j=1+97.0*y/maxran;
        if(j>97||j<1)printf("RAN0:Impossible");
        y=v[j];
        v[j]=rand();
        return (y+1.0)/(maxran+1.0);
}
```

Appendix 3: Some topics related to survival data analysis for latency times

LP can be approximately estimated on the basis of aggregated routine age data, easily available in various countries, using the machinery of BC. Necessary assumptions are (approximate) stationarity of incidence and negligible influence of covariates, in particular of age.

Let us consider, for example, age distribution at onset and at treatment. Let us denote the first by g(y) (assumed known), where y is the age at onset, and by f(x) (also assumed known) the second one, where x is the age at first treatment (x=y+t, where t is the latency time). Such distributions are linked by the following equation:

$$f(x) = \int_{t=0}^{x} g(x-t)\,d(F(t))$$

where F(t) is the latency period distribution. The equation is mathematically identical to the fundamental equation of Back-Calculation, hence deconvolution methods can be easily applied to obtain the unknown function F(t). The estimate is analogous of the LP distribution estimated backwards, thus it is affected by the same possible biases which must be taken into account. However, it allows to estimate the LP distribution even for those countries who cannot provide individual data to apply standard survival analysis methodologies.The adequacy of the approximation can be judged by examining the similarity of estimates over different time periods, effectively testing the (approximate) constancy of incidence. The influence of covariates could be included once coefficients of Cox regression models are known or estimated form an external sample similar to the population of interest.

For example, it is known from previous projects that the influence of age at first use is similar in different sites. The coefficients of Cox model are not significantly different in the various samples analysed. Thus such coefficient can be used to modulate the basic survival function obtained on the basis of aggregated data by using the deconvolution method on the equation shown above.

Let us consider the data from Spain provided by A. Domingo. The various age distributions are shown in the Figures 1 and 2.

In particular, Figure 1 reports the graph of age at onset observed for the three samples available to be analysed. It is evident that the three distributions are not significantly different, showing that the onset age distribution can be considered constant in the period analysed.

Figure 2 shows the distribution of age at treatment for the same samples. As can be seen those entering treatment in later calendar years are older than those entering in previous years. This apparent trend in the age at first treatment only depends on the decreasing behaviour of the incidence curve in recent years which produces longer LP estimates (see section 6.6, backwards estimation bias). This effect also makes it possible to use the trend of age at treatment distribution as a qualitative indicator of the behaviour of incidence (see Section 6.1).

If we apply the BC deconvolution on the three couples of distributions (each two with a fix calendar year), we obtain the same LP distribution estimates as obtained directly on the basis of individual data, but without covariates. The three distributions are shown in Figure 3.

This approach can be particularly useful as most countries seem to lack raw data for estimating the latency period, but more countries can provide results of estimation of

the distributions of age at onset and age at treatment, as can be seen from EMCDDA tables in the appendix 4 “Opiates TDI data”. Such data, provided in more disaggregated form, could be used to produce estimates of incidence curves by BC and correct for biases due to truncation.

Figure 1: age at onset observed for the three samples form Spain.

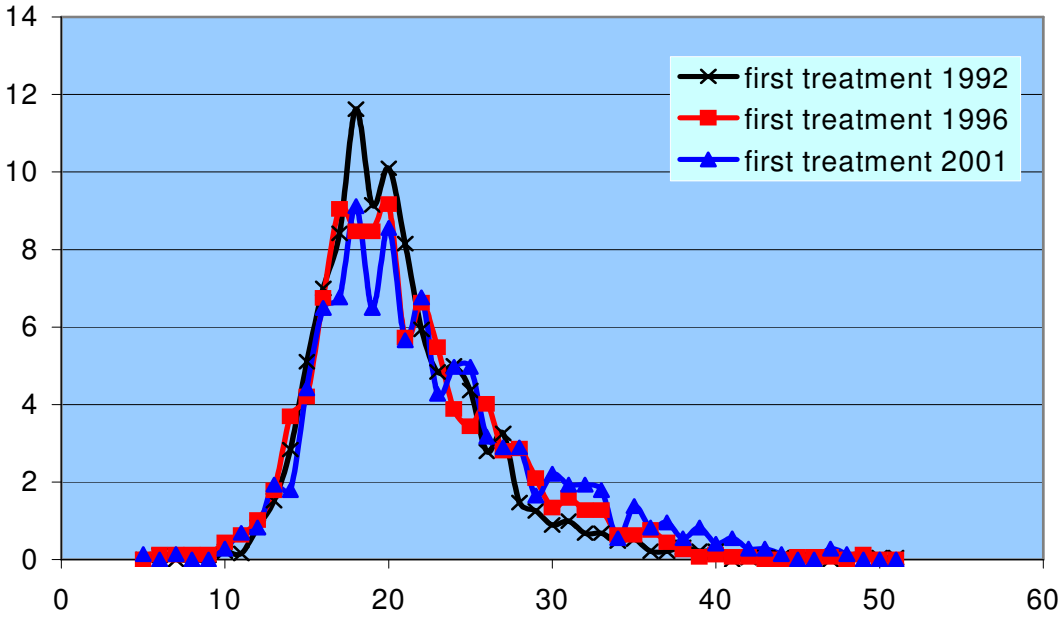


Figure 2: age at treatment observed for the three samples from Spain.

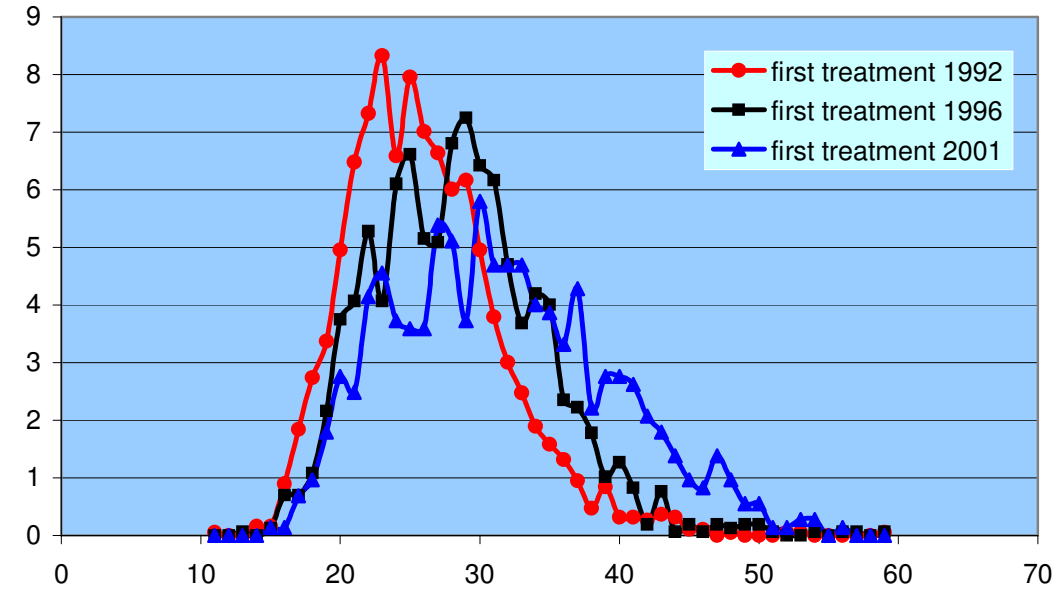
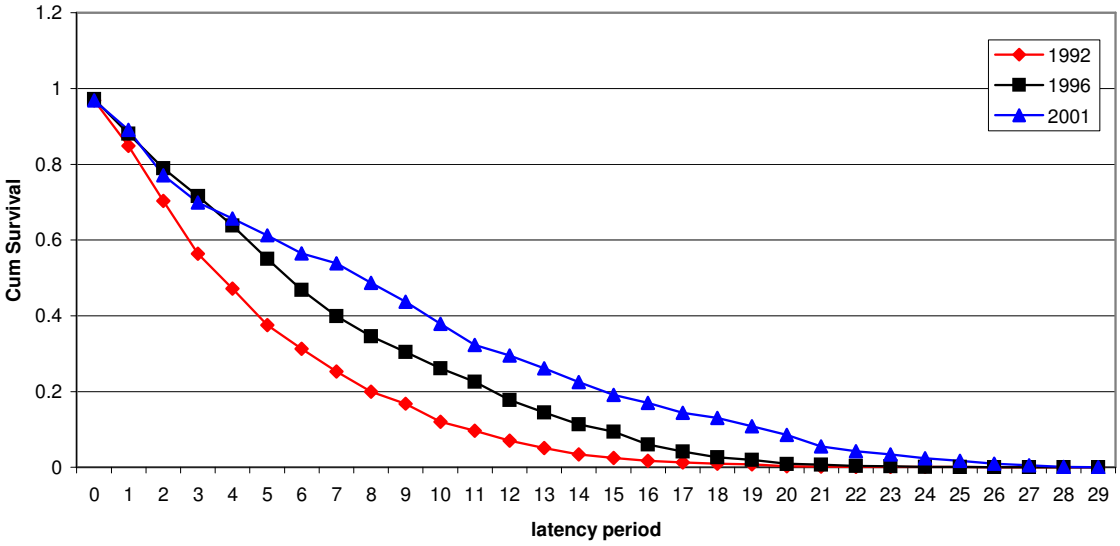


Figure 3: survival functions of BLP for the three samples from Spain.



Appendix 4: Standard methods in survival analysis with SPSS

The purpose of this appendix is to show how statistical software may appear to the user, how it can be used, how to interpret the output and how data is usually coded.

An example of survival data set

The survival data analyzed in this example refer to latency times from first heroin use to first treatment. For the following analyses, it is assumed that the data are representative and not subject to relevant truncation problems (see discussion in section 6.6). The number of the cases is 561 and the first five records of the data set are shown in Figure 1. Here the endpoint of interest is not death, as the word "survival" would suggest, but first treatment demand.

Figure 1: First five cases of the data set

Status	Latency	Age_Her	Gender	Year_Tr	Route
1	3.00	1	0	1993	0
1	5.00	2	1	1993	0
1	1.00	2	1	1994	0
1	2.00	2	1	1994	0
1	1.00	2	1	1994	1

Status is a variable identifying subjects for whom the terminal event (first treatment demand) has occurred and it takes value 1 for all cases. This means that there are no censored cases (which would be coded 0);

Latency is the latency time in years;

Age_Her is the age at first heroin use grouped in the seven classes 0-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40+, that are coded with the numbers from 0 up to 6;

Gender is coded 0 if female and 1 if male;

Year_Tr is the year of first treatment;

Route is coded 0 if subjects are injectors and 1 if they are not injectors.

SPSS Kaplan-Meier Procedure

From the menu choose:

Statistics > Survival > Kaplan-Meier

First, select the time variable:

⇒ Time: latency

Select a status variable to identify cases for which the terminal event has occured. In this example the variable status is selected:

⇒ Status: status

Click Define Event and then enter the value 1 indicating that the terminal event has occurred.

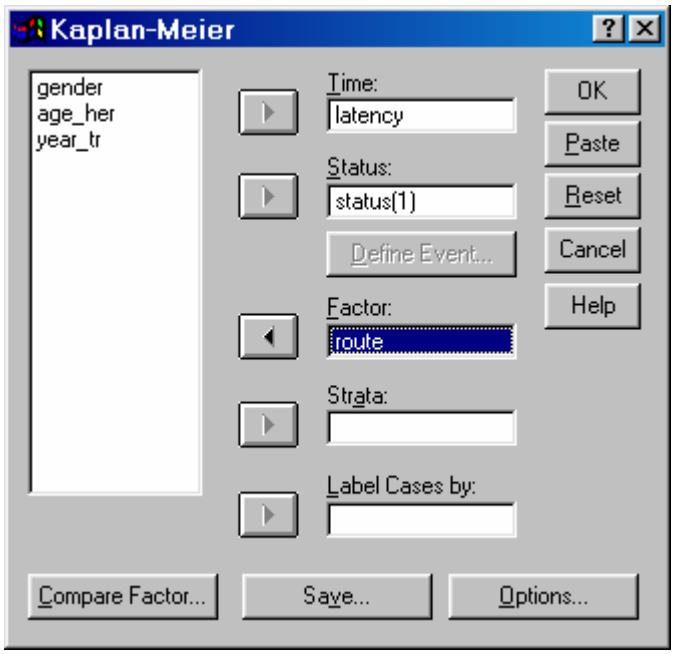
⇒ Define Event

Single value: 1

Optionally, you can select a factor variable to examine group differences. For example, if covariate Route is chosen, the procedure produces a separate survival table for the group of injectors and the group of non-injectors. If no factor is declared, the procedure produces one survival table for all subjects.

⇒ Factor: route

Figure 2: Kaplan-Meier dialog box



You can request different output types from Kaplan-Meier analyses selecting the Options button in the Kaplan-Meier dialog box:

⇒ Options

⇒ Statistics …

⇒ Plots: Survival

You can select statistics for the computed survival functions, such as survival tables, mean and median survival, quartiles.

Plots to examine the survival, hazard, log-survival, and one-minus-survival curves graphically are available.

If you have included factor variables, separate statistics are generated and separate functions are plotted for each group.

Figure 3: Kaplan-Meier Options dialog box

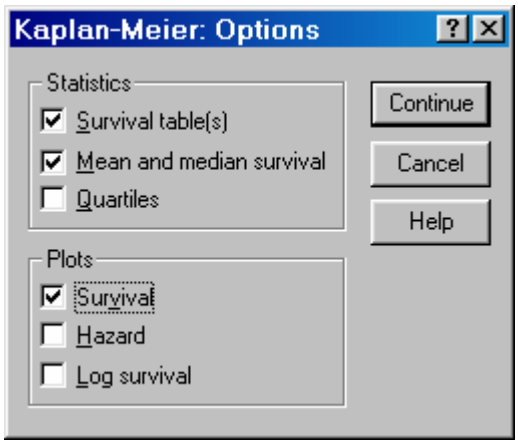


Figure 4 shows part of the output that results from submitting the data set to the SPSS Kaplan-Meier procedure.Only the survival table for the group with covariate Route = 0, that is for the group of injectors, is shown here.

Figure 4: Typical Kaplan-Meier output

```
Survival Analysis for LATENCY

Factor ROUTE = injectors
Time     Status       Cumulative     Standard     Cumulative      Number
                       Survival       Error         Events       Remaining

 .10       1            .9978         .0022          1            454
 .20       1                                         2            453
 .20       1                                         3            452
 .20       1                                         4            451
 .20       1            .9890         .0049          5            450
 .25       1            .9868         .0053          6            449
 .30       1                                         7            448
 .30       1                                         8            447
 .30       1            .9802         .0065          9            446
...(omissis)...

Number of Cases:  455   Censored:    0      (   .00%)   Events: 455

          Survival Time    Standard Error   95% Confidence Interval

 Mean:         2.26                .08     (     2.11;       2.41 )
 Median:       2.00                .04     (     1.92;       2.08 )


Survival Analysis for LATENCY
```

		Total	Number Events	Number Censored	Percent Censored
ROUTE	injectors	455	455	0	.00
ROUTE	non-injectors	106	106	0	.00
Overall		561	561	0	.00

The output shown in Figure 4 contains the following items:

Time: is the time of occurrence of the terminal event of interest, that is the first treatment demand;

Status: indicates whether the subject has experienced the terminal event or has been censored. In this case it takes on the value 1 for all the cases and there is no censored case;

Cumulative Survival: is an estimate of the probability of surviving longer than the time listed in the Time column;

Standard Error: is the standard error of the Cumulative Survival estimate;

Number Remaining: is the number of cases the have not still asked for first treatment after the specified time;

Mean Survival Time: is not the arithmetic mean, it is equal to the area under the survival curve for the uncensored cases. The survival curve is shown in Figure 5;

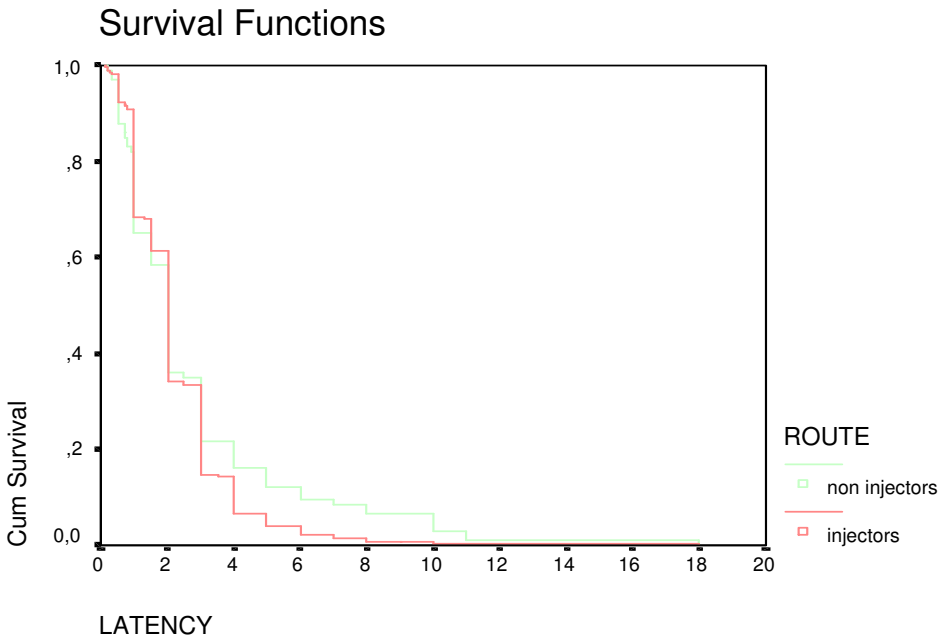
Median Survival Time: is the first event at which cumulative survival reaches 0.5 (50%) or less;

The output finally reports summary comparisons for the two levels of Route.

The output that produced the mean and the median time for the group of injectors also produced a mean and a median time for the group of non-injectors. By comparing these values for the two groups it is possible to examine whether the group with Route = 0 has longer survival times than the group with Route =1. The mean and the median for non-injectors are 2.80 and 2.00, respectively. The mean and the median for injectors are 2.26 and 2.00. The larger mean for non-injectors suggests longer survival times.

The differences between the two survival functions can be analysed graphically. Figure 5 shows a plot with two cumulative survival curves, one for each group, included in the output of the Kaplan-Meier procedure.

Figure 5: Cumulative Survival Functions



The survival function for non-injectors seems to assume larger values (larger survival probability = longer survival time...) for most of the times and this is consistent with the mean observed.

The commands used to produce the output above can be saved, by writing them in a SPSS syntax file, for later re-use, maybe with different data sets.

First, click on the File item on the menu bar at the top of the SPSS window and in the pop-up menu that appears click on New. In the next pop-up you simply create a new Syntax Editor window by selecting SPSS Syntax.

Then type the following instructions in the Syntax Editor

```
KM
  latency  BY route
  /STATUS=status(1)
  /PRINT TABLE MEAN
  /PLOT SURVIVAL .
```

After the KM command you have to specify the variable name that refers to the length of time to the occurrence of the event of interest, for instance latency. A factor variable can be specified after the BY command in order to examine group differences (we used route variable in the present application). Finally you need to specify several options:

STATUS refers to the time variable and the code for the occurrence of an event, in order to determine whether the event has occurred for a particular observation. The code must be enclosed in parentheses after the variable name;

PRINT TABLE MEAN simply instructs the program to produce a KM table for each level of the factor variable;

PLOT SURVIVAL produces a graphical display of the estimated cumulative survival distribution for each level of the factor.

To run the program, select the lines you have written and then click on the Run button at the top of the syntax window Icon Bar. Alternatively, you can run the block of syntax at the bottom of a window by moving your cursor to the beginning of the starting line and clicking on Run.

Cox Regression Analysis

From the menu choose:

Statistics > Survival > Cox Regression

Select the time variable:

⇒ Time: latency

Select a status variable to identify cases for which the terminal event has occurred:

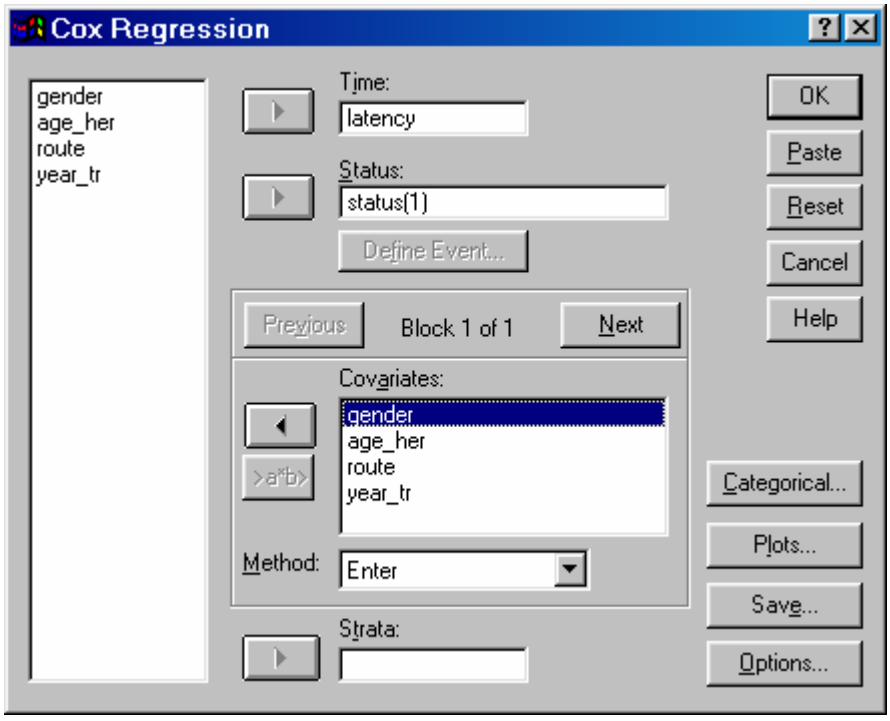
⇒ Status: status

Click Define Event and then enter the value 1 indicating that the terminal event has occurred.

⇒ Define Event

Single value: 1

Figure 6: Cox Regression dialog box



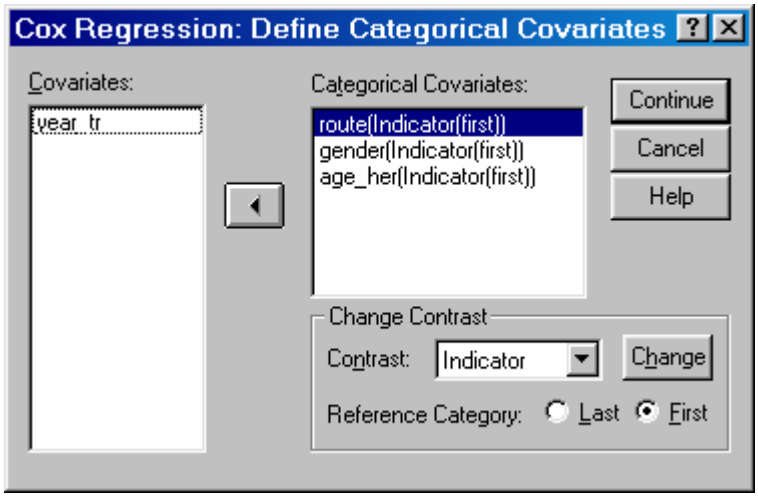
You have to specify the list of covariates to be included in the model:

⇒ Covariates: Gender and Age_her and Route and Year_Tr

Click Categorical to choose details of how the Cox Regression procedure will handle categorical variables. First, you have to define the list of variables identified as categorical in the Define Categorical Covariates dialog box (Figure 7). String variables (denoted with the symbol < after their names) if present will already be specified in the Categorical Covariate list. Select any other categorical covariates from the Covariate list and then move them into the Categorical Covariate list:

⇒ Categorical Covariates: Gender and Age_her and Route

Figure 7: Cox Regression Define Categorical Covariates dialog box



For each selected variable in the Categorical Covariates list it is possible to choose the contrast method to compare the effects of the different levels of the variable. The alternative methods are: Indicator, Simple, Difference, Helmert, Repeated, Polynomial and Deviation.

A common choice is to select the Indicator method, for which contrasts indicate the presence or absence of category membership. In this case the reference category is represented in the contrast matrix as a row of zeros. If you select Indicator, Simple or Deviation, you have to choose either First or Last as reference category. The method is not actually changed until you click Change:

⇒ Contrast: Indicator

⇒ Reference Category: Last or First

⇒ Change

If the model includes more than a covariate, as in the present example, it is possible to choose the method to include and exclude variables. In the Cox Regression dialog box (Figure 6) you can select:

⇒ Method: Enter or Forward or Backward

If you select Enter method, all of the variables are forced into the model in one step.

In Forward selection, the model begins as the baseline model without any variables in it. The variables are considered one at a time and are included if they meet the selection criterion based on the p-value (the default value for inclusion is 0.05). As

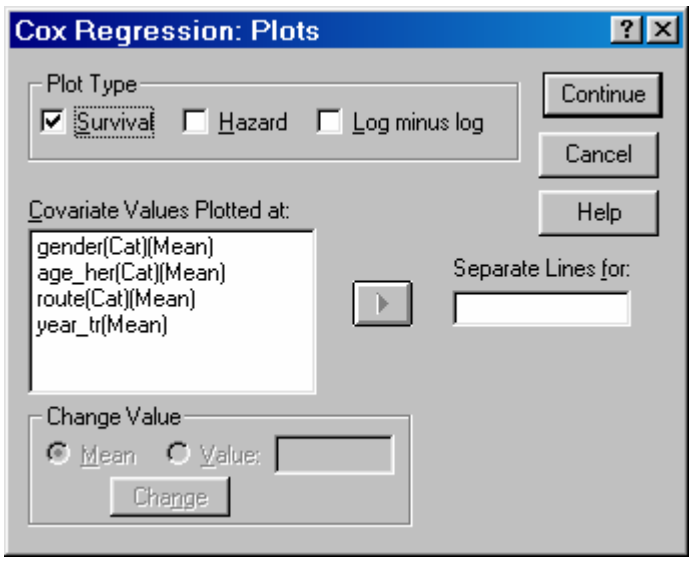
each new variable is included the variables already present are evaluated for removal. When no more variables meet entry or removal criteria, or when the last model is identical to a previous model, the algorithm ends.

In Backward selection, all of the selected variables are entered into the model at the first step. Each variable is then considered for removal. All of the variables that meet removal criteria are removed. Then the excluded variables are reconsidered for inclusion. When no more variables can be entered or removed, the algorithm ends.

It is possible to plot the survival, hazard, log-minus-log, and one-minus-survival functions: in the Cox Regression dialog box (Figure 6) click on the Plot button to get the Cox Regression Plots dialog box (Figure 8) and select the functions you are interested in:

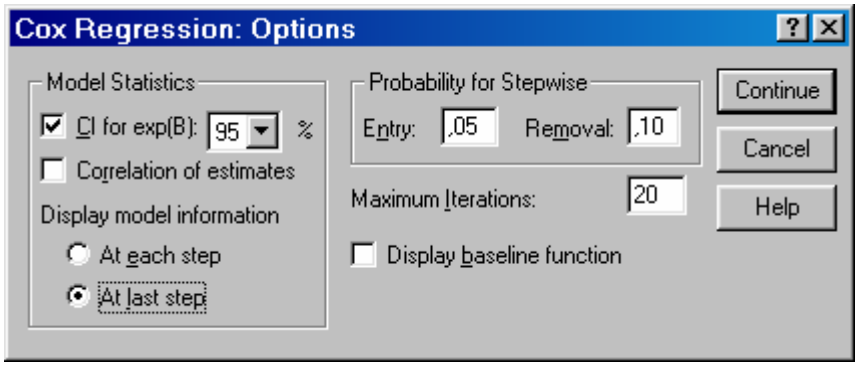
⇒ Plot type: Survival

Figure 8: Cox Regression Plots dialog box



You can control various aspects of your analysis and output: click on the Options button in the Cox Regression dialog box to get the Cox Regression Options dialog box (Figure 9).

Figure 9: Cox Regression Options dialog box



It is possible to obtain confidence interval for the estimated relative risk exp(B):

⇒ Model Statistics: CI for exp(B)

Figure 10 shows the first part of the output for Cox Regression related to the model with four variables included: Gender, Age_Her, Route, Year_Tr. The output has been produced forcing all of the variables into the model in one step that is with Enter selection method.

Figure 10: Cox Regression Output

Indicator Parameter Coding

	Value	Freq	(1)
GENDER			
	0	110	.000
	1	451	1.000

Indicator Parameter Coding

	Value	Freq	(1)	(2)	(3)	(4)	(5)	(6)
AGE_HER								
	0	21	.000	.000	.000	.000	.000	.000
	1	218	1.000	.000	.000	.000	.000	.000
	2	211	.000	1.000	.000	.000	.000	.000
	3	71	.000	.000	1.000	.000	.000	.000
	4	23	.000	.000	.000	1.000	.000	.000
	5	11	.000	.000	.000	.000	1.000	.000
	6	6	.000	.000	.000	.000	.000	1.000

Indicator Parameter Coding

	Value	Freq	(1)
ROUTE			
	0	455	.000
	1	106	1.000

Abbreviations for Terms in the Regression Model

```
Abbrev.   Full Name

Trm2(1)   AGE_HER(1)
Trm2(2)   AGE_HER(2)
Trm2(3)   AGE_HER(3)
Trm2(4)   AGE_HER(4)
Trm2(5)   AGE_HER(5)
Trm2(6)   AGE_HER(6)


     561  Total cases read
         0  Cases with missing values
         0  Valid cases with non-positive times
         0  Censored cases before the earliest event in a stratum
       0  Total cases dropped
     561  Cases available for the analysis


Dependent Variable:  LATENCY

          Events  Censored

             561         0 (0%)


Beginning Block Number 0.  Initial Log Likelihood Function
-2 Log Likelihood    6205.401
Beginning Block Number 1.  Method:  Enter
Variable(s) Entered at Step Number 1..
    GENDER
    AGE_HER
    ROUTE
    YEAR_TR
-2 Log Likelihood    6198.493

                   Chi-Square    df    Sig
Overall (score)         6.449     9  .6943
Change (-2LL) from
 Previous Block         6.908     9  .6467
 Previous Step          6.908     9  .6467
```

The output contains the following information:

Coding system for the categorical covariates: dummy variables (0-1 variables) are used to indicate different levels of each factor. If there are N levels then N – 1 dummy variables are to be used. In the easiest case of a variable such as Gender with only two levels, a single dummy variable is adopted. The reference level used in the contrasts is that with all the dummy variables set to 0, in this case the first level of each factor.

A report summary with the number of the cases available for the analysis and the cases excluded;

Block Number 0 refers to the baseline model with all coefficients B's set to 0. It is reported minus 2 times the log likelihood for the baseline model to be compared with -2 log likelihood for the model of interest with the covariates included, to test whether all population B's can considered 0;

Block Number 1 refers to the model with the predictor variables included with Enter selection method;

Changes from previous Block is the difference between -2 log likelihood for block 0 and block 1. This is the likelihood - ratio (LR) test;

Changes from previous Step refers to stepwise regression procedures (forward or backward selection);

Overall (score) is another test to check whether all the coefficients B in the model can be considered 0 in the population. It is distributed as chi-square and it is an approximation of the LR test;

df (degrees of freedom) is the number of variables in the present model minus the number of variables in the previous model;

Sig is based on the chi-square test. Both the score and -2 log likelihood are distributed as chi-square for large samples. The level of significance is fixed at 0.05.

The change in -2 log likelihood from the previous block in which all B's were set to 0, is 6205,401 – 6198,493 = 6,908. Since the significance of the chi-square is more than 0.05 we can consider the baseline model with all coefficients B's set to 0. If the significance were less than 0.05 we could conclude that at least one B was different from zero.

The score test is coherent with the likelihood ratio test.

Figure 11 contains the second part of the output of the Cox Regression procedure:

Figure 11: Continuation of the Cox Regression Output

```
---------------- Variables in the Equation -----------------

Variable            B       S.E.       Wald  df      Sig        R

GENDER           .0552      .1076      .2630   1    .6081    .0000
AGE_HER                                3.7065   6    .7163    .0000
 AGE_HER(1)     -.0293      .2338      .0157   1    .9004    .0000
 AGE_HER(2)     -.0041      .2343      .0003   1    .9862    .0000
```

AGE_HER(3)	.0708	.2510	.0797	1	.7777	.0000
AGE_HER(4)	-.3817	.3102	1.5142	1	.2185	.0000
AGE_HER(5)	.0671	.3791	.0313	1	.8595	.0000
AGE_HER(6)	-.1511	.4709	.1030	1	.7482	.0000
ROUTE	-.1515	.1133	1.7866	1	.1813	.0000
YEAR_TR	.0051	.0166	.0932	1	.7602	.0000

		95% CI for Exp(B)	
Variable	Exp(B)	Lower	Upper
GENDER	1.0567	.8558	1.3049
AGE_HER(1)	.9712	.6142	1.5356
AGE_HER(2)	.9959	.6292	1.5764
AGE_HER(3)	1.0734	.6563	1.7556
AGE_HER(4)	.6827	.3717	1.2539
AGE_HER(5)	1.0694	.5087	2.2481
AGE_HER(6)	.8597	.3416	2.1636
ROUTE	.8594	.6882	1.0732
YEAR_TR	1.0051	.9729	1.0383

B's are the four estimated coefficients (one for each variable included in the model). It is interpreted as the predicted change in log hazard for a unit increase in the predictor;

S.E. is the standard error of the estimated coefficients, B's;

Wald is the Wald statistic used to test whether the estimated coefficients B's are different from 0 in the population and it is distributed as a chi-square;

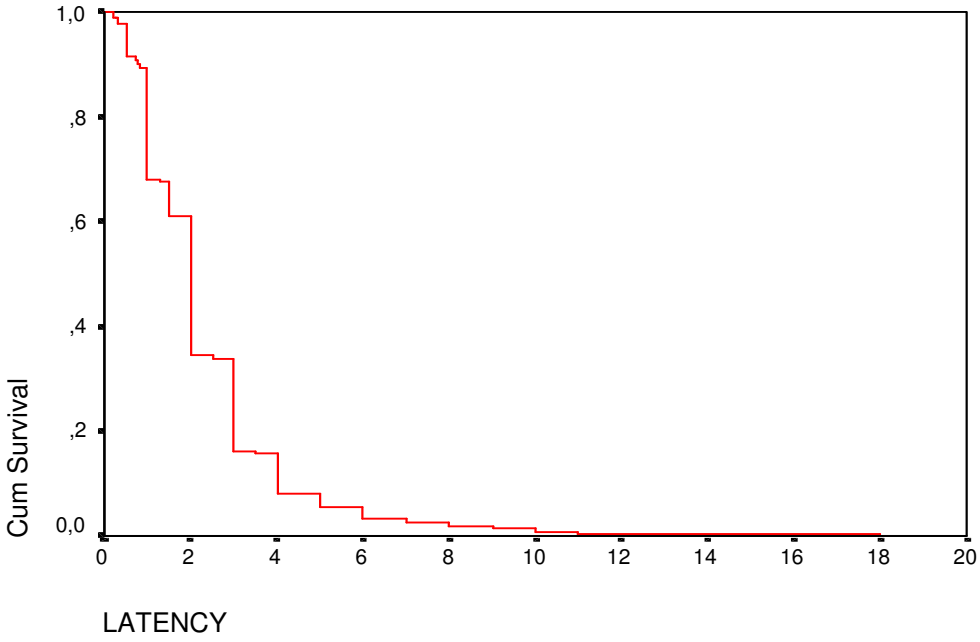
df is degrees of freedom of the Wald statistic;

Sig is the significance level for the Wald statistic;

Exp(B), for a dichotomous variable X, such as Route, in which there are two levels, 0 and 1, it is the relative risk, that is the ratio of the risk with X = 0 compared to the risk with X = 1. For a multiple level variable X, such as Year_Tr, Exp(B) estimates the percentage change in risk with each unit change of the covariate value.

The output also provides the plot of the survival function, as reported in Figure 12.

Figure 12: Cumulative Survival Function



The SPSS syntax file corresponding to this application would be the following:

```
COXREG
  latency
  /STATUS=status(1)
  /CONTRAST (route)= INDICATOR (1)
  /CONTRAST (gender)=INDICATOR(1)
  /CONTRAST  (age_her)= INDICATOR (1)
  /METHOD=ENTER gender age_her year_tr route
  /PLOT SURVIVAL
  /PRINT=CI(95)
  /CRITERIA=PIN(.05) POUT(.10) ITERATE(20) .
```

As in the Kaplan-Meier syntax file, the COXREG command must be followed by the time variable (latency). Then you have to specify the following options:

STATUS refers to the time variable and the code for the occurrence of an event, in order to determine whether the event has occurred for a particular observation. The code must be enclosed in parentheses after the variable name;

CONTRAST specifies the type of contrast used for the categorical covariates. The interpretation of the regression coefficients depends on the contrasts used. The default is DEVIATION (deviation from the overall effect) but in the present application we used INDICATOR that indicates the presence or absence of category membership. By default, the reference category would be the last category (represented in the contrast matrix as a row of zeros). To choose a reference category other than the last you

simply have to specify the sequence number of the category in parentheses after keyword INDICATOR (for instance, we choose the first category).

METHOD specifies the order of processing and the manner in which the covariates enter in the model. The default is ENTER meaning that all variables are entered in a single step. Alternatively you could choose FSTEP or BSTEP to get forward stepwise selection or backward stepwise selection;

PLOT SURVIVAL produces a graphical display of the estimated cumulative survival distribution;

PRINT produces a regression report. In this case we used the command to request as a specific output CI, that is the confidence intervals for the estimated relative risk exp(B). You have to specify the confidence level in parentheses and the default is 95%.

CRITERIA controls the statistical criteria used in building the Cox Regression model. How these criteria are used depends on the method specified on the METHOD subcommand (in this case ENTER). PIN indicates the probability of the score statistic for variable entry: a variable whose significance level is greater than PIN cannot enter in the model (the default is 0.05). POUT indicates the probability of a suitable statistic (Wald, Likelihood Ratio, Conditional Likelihood Ratio) to remove a variable: a variable whose significance is less then POUT cannot be removed (the default is 0.1). ITERATION specifies the maximum number of iterations (the default is 20).

Appendix 5: Abbreviations

AIDS	Acquired Immuno Deficiency Syndrome
B-BC	Bayesian Back Calculation method
BC	Back Calculation method
DU	Drug use/user
EB-BC	Empirical Bayesian Back Calculation method
E-M	Expectation Maximization (algorithm)
EMCDDA	European Monitoring Centre for Drugs and Drug Addiction
EU	European Union
HIV	Human immunodeficiency virus
IDU	Injecting drug use/user
LP	Latency period
MCMC	Markov Chain Monte Carlo
PDU	Problem drug use/user
RDA	Reporting Delay Adjustment
UK	United Kingdom

Appendix 6: History and acknowledgements

These Guidelines are the result of the EMCDDA financed project 'Project to stimulate the implementation of the EMCDDA guidelines on estimating incidence of injecting and problem drug use in the EU' (CT.06.EPI.150.1.0) and previous related EMCDDA projects. The present version, which essentially is the 2nd edition of the Guidelines, has been scientifically edited and in part written by Gianpaolo Scalia Tomba, University of Rome Tor Vergata, Italy. To a large extent, content and style are derived from the 1st edition of the Guidelines (2004) (CT.02.P1.55), which was produced by Carla Rossi and coworkers also from the University of Rome Tor Vergata, Italy, in close collaboration with the EMCDDA (Lucas Wiessing). Maria Grazia Calvani, Emanuela Colasante, Flavia Lombardo and Lucilla Ravà, under the guidance of Carla Rossi, carried out most of the data analyses presented in sections 6.3, 6.4 and 6.6. Important inputs and suggestions from Antonia Domingo Salvany, Siem Heisterkamp, and Matthew Hickman to the 2004 version of the Guidelines are also gratefully acknowledged.

The writing and revision of this updated version of the Guidelines has benefitted from the continuous guidance of Lucas Wiessing and Colin Taylor, EMCDDA, since the start of the revision project, and from Danica Klempova, EMCDDA, during the final phase of the work. Also, many suggestions by participants at the yearly EMCDDA meetings on Problem Drug Use are acknowledged, where preliminary ideas about the contents and style of the Guidelines have been presented and discussed. In particular, comments and suggestions by Ellen Amundsen, Sharon Arpa, Antonia Domingo Salvany, Clive Richardson and Albert Sanchez Niubo have been of great assistance for the completion of the present version of the Guidelines.