

REVIEW ARTICLE

Improving data collection and information retrieval for monitoring sexual health

R K W Lau BSc MD FRCP¹ and M Catchpole MB FRCP FFPHM²

¹Department of Genitourinary Medicine, St George's Hospital, London SW17 0QT and

²Communicable Diseases Surveillance Centre, London NW9 5EQ, UK

Summary: Recent advances in information technology (IT) offer a wealth of new tools to enhance data collection and information processing. Used in conjunction with established statistical techniques, these advances could be applied to develop a better understanding of demographic and behavioural factors in the acquisition and spread of sexually transmitted infections (STIs). Successful implementation of these technologies will, however, require significant investment of resources, closer working between clinical specialties, and dealing with the fundamental concerns many clinicians as well as patients have over collection of personal data. An important first step towards realizing some of these benefits for data on STIs would be the disaggregation of data collected from genitourinary medicine (GUM) clinics.

Keywords: Data collection, disaggregate data, data protection, sexually transmitted infections

INTRODUCTION

'If all you have is a hammer, then everything looks like a nail'

Knowledge of the epidemiology of STIs is central to developing an effective sexual health strategy. In England, GUM clinics have a statutory requirement to make a quarterly workload return (KC60 report) to the Public Health Laboratory Service, Communicable Disease Surveillance Centre (CDSC). These returns provide the core data for surveillance of STIs in England. Unfortunately, the KC60 report largely reflects clinical workload and the epidemiological inferences that can be drawn from the data are limited. Aggregation of the data makes it difficult to define or characterize the 'at risk' population, since it is not possible to distinguish individual patients each attending with a single episode of a specific genital infection from patients presenting with multiple infections, or the same patient(s) re-attending with the same condition. Basic data on sexual behaviour that would be central in understanding the epidemiology of STIs are conspicuous by their absence. Finally, by the time the annual report on national data is made available the information is already out of date.

To improve this situation a review of the way in which clinical data about STIs are collected, processed and interpreted is required¹. Part of the solution lies in disaggregating the dataset collected

from GUM clinics to allow record linkage of case reports and inclusion of a broader range of data items (fields) describing behavioural, geographic and ethnic traits. This would provide an opportunity to exploit some of the significant advances seen in IT over the last years, to enable data from GUM clinics to be processed in a more timely manner, and to enable the integration of these data with data from other sources. Meeting these objectives will require investment of resources, close collaboration between different clinical disciplines, and dealing with the fundamental concerns many clinicians as well as patients have over collection of personal data.

The purpose of this paper is to cover in some detail the opportunities for improved surveillance and epidemiological study that could flow from improved data capture and knowledge acquisition, and some of the information management concepts underpinning the processes involved. We also review some of the issues of confidentiality and personal privacy that will need to be addressed.

DISAGGREGATE DATA COLLECTION

One of the main obstacles to surveillance based on clinical reporting has been the purely pragmatic problem of finding the time or staff resource to record and collate data recorded in the clinical setting. A central theme of the National Health Service (NHS) Information for Health strategy is the creation of the electronic patient record and, ultimately, the electronic health record². Many parts of the health service have already introduced computerized database systems that store basic

demographic or contact information about patients, and increasingly these systems are also being used to store clinical data. With the introduction of computerized clinical records comes the opportunity for easier processing of those data for public health and audit purposes. Computerized data management systems allow larger volumes of data to be processed more rapidly and more accurately than older paper-based systems, giving rise to the potential for far more powerful surveillance. The need for rapid assimilation of data is particularly true for the surveillance of communicable disease, if outbreaks and the emergence of problems such as antibiotic resistance are to be detected in time for effective intervention to be mounted. In the case of STIs, the epidemiology of which is unique in its closeness of association with specific human behaviours, there is also a need for a relatively rich set of data on the demographic and behavioural characteristics of patients if trend and distribution data are to be interpreted. It is not logistically possible to collect and collate the multi-dimensional data sets that it is now recognized are required for STI surveillance through paper or aggregate statistical reporting systems.

In 1998 a survey of GUM clinics in England indicated that 85% of clinics had computer systems that could potentially provide anonymous case reports for surveillance (CDSC, unpublished data) electronically. Electronic reporting of disaggregate anonymous data has already been introduced in Scotland³, and has been successfully piloted in parts of England^{4,5}. A forthcoming review of collection of data from GUM clinics in England will assess the feasibility of replacing the current paper statistical return (form KC60) with electronic reporting.

Sexually transmitted infections do not only cause acute disease—several of these infections are associated with serious long-term sequelae. Examples include HIV infections and AIDS, human papillomavirus infection and cervical carcinoma, and chlamydial infection and pelvic inflammatory disease, infertility and ectopic pregnancy. A major part of the public health rationale for the control and prevention of these infections is the reduction in incidence of these long-term sequelae.

Surveillance of STIs, if it is to identify those groups most at risk of these sequelae, document the true burden of disease, and provide information on the impact of intervention, must encompass these long-term sequelae. The value of data on these sequelae, particularly for evaluation of the effectiveness of interventions, would be greatly added to if they could be linked to data on the primary infection. Many, if not most, of these sequelae are managed outside GUM services, and therefore surveillance of these conditions must rely on data from sources other than GUM clinics. Until such time as the electronic health record is realized, it will be necessary to bring together disparate data sets from different clinical and laboratory sources if

we are to create combined data on both the acute and the chronic consequences of STI. New information management technologies now make this a real possibility.

Data warehousing and centralizing resources

Although not an essential pre-requisite, the potential of new analytical techniques such as data mining can be enhanced if the appropriate data have been collected and stored in a data warehouse—a system for storing and delivering massive quantities of data⁶. The promise of data warehousing is that data from disparate (clinical) databases can be consolidated and managed within one single database. Once data about the organization's processes become readily available, it becomes easy and therefore economical to mine it for new and potentially meaningful relationships (Figure 1).

DATA MINING

'We are drowning in data but starved of knowledge'

Data mining is a set of automated techniques used to extract buried or previously unknown pieces of information from large databases^{7,8}. Successful data mining makes it possible to unearth patterns and relationships, and then use this 'new' information to formulate new hypotheses or even make proactive knowledge-driven decisions. The role of data mining of clinical data sets would probably primarily be one of hypothesis generation, which would then be tested through formalized regression analysis (see below) or *de novo* directed studies. The central tenet of data mining is the automated discovery of new facts and relationships in data. The raw material may be any data, be it business or clinical, and the data mining algorithm is the excavator, sifting through the vast quantities of raw data looking for the valuable nuggets of information. The validity and the value of the relationships highlighted by such techniques will depend on the quality and the range of the base data. It will only be worth investing in sophisticated data mining techniques if sufficient resource is also invested in ensuring the capture of sufficiently high-quality data on possibly related events.

MULTIVARIATE ANALYSIS AND MULTIPLE LOGISTIC REGRESSION

Along with data mining, a number of statistical methods could be applied to the data to generate predictions and trends. Foremost among these would be the large family of multivariate statistical techniques, which deal with numerous variables at the same time. In essence, multiple regression is a straightforward generalization of simple regression,

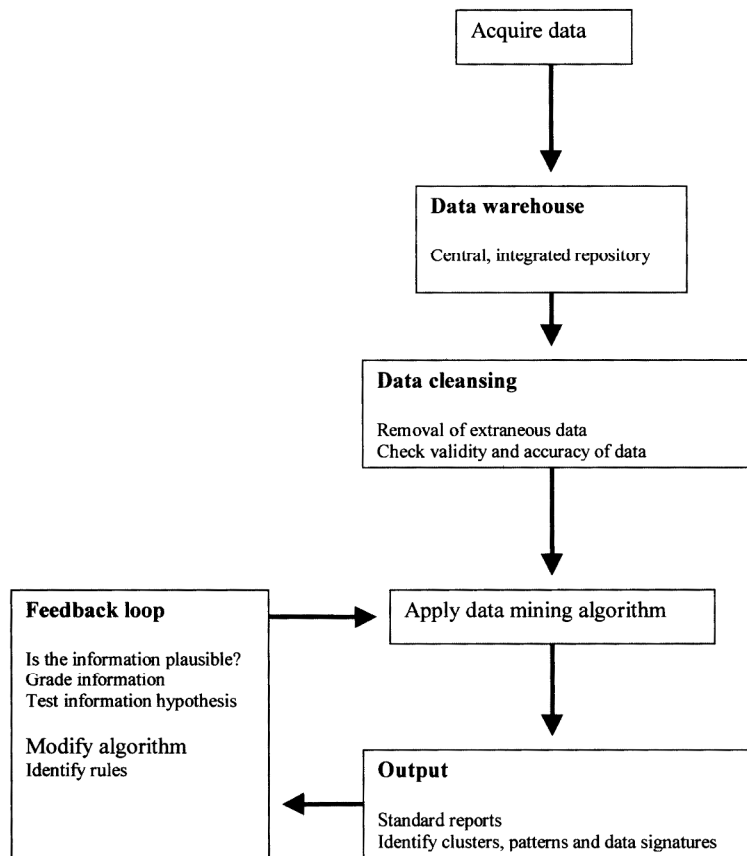


Figure 1. Processes involved in data warehousing and data mining

the process of fitting the 'best' straight line through the dots on an x - y plot or scattergram.

Multiple regression models are particularly suited in observing 'natural' (as opposed to experimental) phenomena where there may be one dependent variable (the number of cases of gonorrhoea) and many independent variables (age, ethnicity, area of residence, etc.). It is, however, inherently a correlational technique; it cannot of itself tell anything about the causalities that may underlie the relationships it describes.

Although logistic regression finds a 'best fitting' equation just as linear regression does, the principles on which it does so are rather different. Instead of using a least-squared deviations criterion for the best fit, it uses a maximum likelihood method, which maximizes the probability of getting the observed data given the fitted regression coefficients^{9,10}.

DIFFERENCES BETWEEN DATA MINING AND OLDER ANALYTICAL TOOLS

Data mining differs from other analytical tools in the approach used in exploring the data relationships. Traditional database queries can answer questions like 'how many cases of gonorrhoea did my clinic diagnose in 1998?' Other analyses, often

called multi-dimensional or online analytical processing (OLAP), allow users to do more complex queries, such as comparing gonorrhoea rates per head of adult population by quarter and region for the prior 2 years. In both cases, however, the results are simply figures extracted from the data or an aggregate of existing data. The relationship among these data is already known to the user, who, by framing the proper question, obtains the desired answer.

Data mining however, uses discovery-based approaches in which pattern-matching and other algorithms are used to discover key relationships in the data, previously unknown to the user. The data are sifted in search of frequently occurring patterns, trends, and generalizations about the data without intervention, bias or guidance from the user. For example, we might want to discover a model which best describes the 'typical' person attending a clinic, grouped according to a primary diagnosis (such as first-episode genital warts), sex-age range, district of residence or ethnicity. The data are searched with no hypothesis in mind other than for the system to group the patients according to the common characteristic found. Box 1 summarizes the 5 possible types of information derived from data mining. Once a new association is uncovered, more traditional approaches to data

analysis are required to assess the plausibility, strength and significance of that association.

APPLICATIONS TO GENITOURINARY MEDICINE

If applied to data generated by and held within GUM clinics, electronic disaggregate reporting, data mining and warehousing could not only generate the current KC60-type of quarterly workload returns more rapidly, but also help identify:

- Patterns of clients attending specific clinics and even particular sessions in a given clinic; associations among clients' demographic characteristics; predictions on which clients might have particular genitourinary conditions.
- Patterns of clinic usage and profiles of clients likely to default from follow-up.
- Behaviour patterns of 'at risk' client groups; expectations of risk behaviour and behaviour modification.
- Prediction of therapeutic outcomes based on type of medical condition, therapy offered and patient demographic grouping.

Extending the data set to include information from other clinical settings that provide sexual health services could further increase the added value that would accrue from these approaches to data management.

CONFIDENTIALITY, CONSENT AND PRIVACY

The proposal to process electronic disaggregate patient information, particularly if that information is to be collated from a variety of services, within a wide area computer network has a number of far-reaching implications. It would require developing a viable regional and potentially, national, business plan to cover such a move, funding arrangements to develop and maintain it, and re-defining professional relationships between clinical specialties. But most of all, there will be a need to ensure that the proposal would not impact on the legal framework which underpins the practice of GUM

in the United Kingdom or breach statute or Common Law on confidentiality, data protection or privacy (human rights).

While data mining has the potential to identify 'hidden' relationships in personal data and the epidemiology of STIs, it also has the potential for abuse and misunderstanding. There is the notion of an Orwellian 'Big Brother', overseeing not just patients and their medical conditions but also the clinicians providing the service. There is specific legislation in England and Wales with respect to confidentiality of information on patients with STIs, in the form of the *National Health Service (Venereal Diseases) Regulations 1974* and the *National Health Service Trusts (Venereal Diseases) Directions 1991*. Neither of these documents precludes the disclosure of data, even patient identifying data, on patients with sexually transmitted diseases so long as that disclosure is to a suitably qualified or suitably employed individual and is in the interests of the control or prevention of disease. Public health surveillance clearly falls within the definition of the purpose of disease control and prevention.

More important to the issue of use of patient data, whatever the disease or the service attended, is interpretation of current European and UK legislation on data protection and human rights and of the Common Law on confidentiality. Of these, the Common Law on confidentiality is the key, since the Data Protection Act largely superimposes on the Common Law of confidentiality a framework of rights and principles governing the processing of data that have been lawfully obtained. In its draft guidelines on 'Personal Information in Medical Research'¹¹, the Medical Research Council (MRC), notes the following:

'Common Law recognises that it can be in the public interest for doctors to disclose confidential information, and that the nature and scale of the disclosure has to be balanced against the benefits to society'

'While the law establishes some common principles, it does not specify when confidential information should

Box 1. Types of information gathered from data mining

Associations are occurrences linked in a single event. For example, a study of teenage girls attending a clinic might reveal that when gonorrhoea is diagnosed, genital chlamydial infection is also detected 65% of the time.

Sequences are events that are linked over time. For example, if a homosexual man presents with ano-rectal gonorrhoea, then 15% of the time he will present with another new infection within one month.

Classification might help to discover certain characteristics of patients presenting with a particular condition, or *vice versa*. For example, what are the personal and demographic characteristics of gonococcal infection compared with genital wart virus infection? Or, what are the 3 commonest presenting conditions in a white single man over 40 years living in London SW17?

Clustering discovers different groupings with the data. This might help to identify geographic areas/post code prefixes with high rates of particular sexually transmitted infections.

Forecasting estimates the future value of continuous variables, such as numbers of new cases of first episode genital herpes, etc.

and should not be disclosed to others, in research or most other activities'

In this context, it is essential that there is clarity as to the communicable disease surveillance purposes for which identifiable data can be used. It is to be hoped that the recently announced National Confidentiality and Security Advisory Body¹², that is to advise Ministers, the Information Policy Unit of the NHS Executive, the wider Department of Health, and the NHS Information Authority on how key confidentiality and security issues including ethical considerations should be addressed, will provide the required clarifications. In the meantime, the most recent relevant case law is the Judgement in the Court of Appeal in respect of the case between Source Informatics Limited and the Department of Health¹³, which includes the statement that:

'it is clear from the information before us that for certain limited purposes patient information is used in identifiable rather than anonymised form . . . For present purposes, I say no more than that, provided, as I understand to be the case, the use of such identifiable data is very strictly controlled, there appears to be no reason to doubt that this is acceptable—whether because it falls within the public interest defence or as is perhaps the preferable view, because the scope of the duty of confidentiality is circumscribed to accommodate it, it is not necessary to decide on this appeal.'

One of the core principles of the 1998 UK Data Protection Act is that personal data shall be obtained only for one or more specified and lawful purposes, and shall not be further processed in any manner incompatible with that purpose or those purposes. While there is clearly a public health interest case for using patient data primarily collected for the purposes of direct clinical care for public health surveillance, we need to be more circumspect where data are being used for research. Aspects of the use of data mining techniques that are more akin to research than surveillance include their use to identify previously unknown associations between data in a database, which might highlight the need for new interventions. There is, again, limited guidance on the use of patient data for such purposes, although the draft MRC guidelines 'Personal Information in Medical Research'¹¹ state that 'most international and national ethical codes do not require explicit, informed, consent for research based only on records that will not directly affect the individual'.

Many would argue that data mining poses a threat to personal privacy in that the large amounts of personal data coupled with powerful new technologies will reveal an individual's associations (e.g. sexual contacts) and patterns of (sexual) behaviours which could be identifiable to others¹⁴. There is also a danger that a central repository of personal information could be the target for

computer 'hackers' and blackmailers. Clearly, if this proposal is to succeed, there must be confidence in the system, and its safeguards and guarantees in respect of data protection and current legislation covering this^{15,16}. There must also be in place administrative accountability and openness in dealing with public concerns about systems security in relation to electronic healthcare records¹⁷.

CONCLUSION

This paper will have achieved its objective if, by describing the processes involved, readers develop an understanding of the potential that new information technologies offer public health surveillance and the tensions that these raise in respect of upholding the privacy safeguards of the individual. In the new millennium we will need to urgently redress the inefficiencies of an outdated data collection system on the one hand, with an improved information system, mindful of the enormous potentials and dangers offered by these new technologies.

Acknowledgements: We thank our colleagues, Drs Fiona Davidson and Helen Maguire, for reading the paper and offering their helpful suggestions.

References

- Catchpole MA, Harris JRW, Renton A, Hickman M. Surveillance of sexually transmitted infections: fit for purpose? *Int J STD AIDS* 1999;**10**:493–4
- NHS Information Authority. *Information for Health: Contents*. www.nhsia.nhs.uk/strategy/full/contents.htm (accessed 1 September 2000)
- Noone A, Chalmers J, Young H. Surveillance of sexually transmitted infections in Scotland. *Eurosurveillance* 1998;**3**:65–8
- Catchpole M, Connor N, Brady A, *et al.* Behavioural and demographic characteristics of attenders at two genitourinary medicine clinics in England. *Genitourin Med* 1997;**73**:457–61
- Hickman M, Judd A, Maguire H, *et al.* Incidence of gonorrhoea diagnosed in GUM clinics in South Thames (west) region. *Sex Transm Inf* 1999;**75**:306–11
- Berson A, Smith SJ. *Data Warehousing, Data Mining, and OLAP*. New York: McGraw-Hill, 1997
- Lavrac N. Selected techniques for data mining in medicine. *Artif Intell Med* 1999;**16**:3–23
- Adriaans P, Zantinge D. *Data Mining*. New York: Addison-Wesley, 1996
- Johnson AM, Wadsworth K, Bradshaw S, Field J. Sexual lifestyles and HIV risk. *Nature* 1992;**360**:410–12
- Hughes G, Catchpole M, Rogers PA, *et al.* Comparison of risk factors for four sexually transmitted infections: results from a study of attenders at three genitourinary medicine clinics in England, 1994–1997. *Sex Transm Inf* (in press)
- MRC Working Group on Personal Information in Medical Research. *Personal Information in Medical Research: Draft MRC Guidelines*. London: Medical Research Council, 1999
- NHS Executive. *Recruitment to the National Confidentiality & Security Advisory Body—Applications invited for Board Members*. <http://www.doh.gov.uk/confiden/recruitment.htm> (accessed 1 September 2000)

- 13 Richards T. Court sanctions use of anonymised patient data. *BMJ* 2000;**320**:77
- 14 Miller M, Cooper J. Security considerations for present and future medical databases. *Int J Biomed Comput* 1996;**41**:39–46
- 15 Hodge JG Jr, Gostin LO, Jacobson PD. Legal issues concerning electronic health information: privacy, quality, and liability. *JAMA* 1999;**282**:1466–71
- 16 Barrows RC Jr, Clayton PD. Privacy, confidentiality, and electronic medical records. *J Am Med Inform Assoc* 1996;**3**:139–48
- 17 Smith E, Eloff JH. Security in health-care information systems—current trends. *Int J Med Inf* 1999;**54**:39–54

(Accepted 2 October 2000)