Journal of Public Health Medicine

Vol. 21, No. 3, pp. 299–304
Printed in Great Britain

Clinical information for research; the use of general practice databases

Ross Lawrenson, Tim Williams and Richard Farmer

Summary

General practice computers have been widely used in the United Kingdom for the last 10 years and there are over 30 different systems currently available. The commercially available databases are based on two of the most widely used systems – VAMP Medical and Meditel. These databases provide both longitudinal and cross-sectional data on between 1.8 and 4 million patients. Despite their availability only limited use has been made of them for epidemiological and health service research purposes. They are a unique source of population-based information and deserve to be better recognized.

The advantages of general practice databases include the fact that they are population based with excellent prescribing data linked to diagnosis, age and gender. The problems are that their primary purpose is patient care and the database population is constantly changing, as well as the usual problems of bias and confounding that occur in any observational studies. The barriers to the use of general practice databases include the cost of access, the size of the databases and that they are not structured in a way that easily allows analysis. Proper utilization of these databases requires powerful computers, staff proficient in writing computer programs to facilitate analysis and epidemiologists skilled in their use. If these structural problems are overcome then the databases are an invaluable source of data for epidemiological studies.

Keywords: computers, databases, general practice, epidemiology

Introduction and background

Observational research relies heavily on routine data. Increasingly general practitioners (GPs) in the United Kingdom are using computers to help manage their practices. These computers are used to generate prescriptions, to keep patient records and to support the financial management of the practice. Over 90 per cent of practices in the United Kingdom are computerized[1] and about 8 per cent are believed to be paperless – that is, the practices use the computer record as the primary source of information about their patients.[2] Combining the anonymized records from a number of practices allows researchers to use the collected information for medical research. There are over 30 different software systems designed for use by GPs that are currently available in the United Kingdom.[1] Two of the most commonly used systems – the VAMP Medical system and the Meditel system – have been used to provide commercially available databases.[3] There are other smaller 'privately owned' general practice databases that have been used for research. These are often based on a small number of practices that have combined information for a specific research project.[4–7]

The three commercially available databases are the General Practice Research Database (GPRD), the MediPlus database and the Doctors Independent Network database (DIN). The GPRD is owned by the Department of Health and managed by the Office for National Statistics (ONS). It currently receives data from over 400 practices from throughout the United Kingdom. All the contributing practices use a system known as the VAMP Medical system. This system uses its own coding system known as OXMIS medical codes and for prescriptions uses the Prescription Pricing Authority (PPA) codes. Practices have to undergo intensive training in the use of the system and the quality of the data is checked before practices can be considered 'up to standard'. Data from such practices are available from 1988 onwards, involving approximately 7 million patients. The MediPlus database is based on information from 150 practices that use the Meditel system. It has information on 1.8 million patients, again from practices throughout the United Kingdom. All the diagnostic records, prescriptions and notes are coded. The Meditel system uses Read codes, the coding system adopted by the Department of Health for general practice.[8] Doctors using either the VAMP Medical or the Meditel system can attach free text to the appropriate code; for reasons of confidentiality this information is not available to researchers using the databases. The DIN database collects data from over 200 practices, again from throughout the United Kingdom, and has over 2 million patient records. It is owned by the contributing doctors and is a registered charity. It also uses Meditel system and Read codes,

European Institute of Health and Medical Sciences, Stirling House Campus, Stirling Road, Surrey Research Park, Guildford GU2 5RF.

Ross Lawrenson, Director Primary Care Research

Tim Williams, Research Fellow

Richard Farmer, Professor of Pharmaco-epidemiology and Public Health

Address correspondence to Dr Ross Lawrenson.

© Faculty of Public Health Medicine 1999

and there is some overlap with the MediPlus database because some doctors contribute their data to both.

Each of the three commercially available databases has a similar core structure; this includes a registration file, a prescription file, a file containing problems or diagnoses and a note file. An individual patient will have data recorded in each of the four files. These files have to be linked to allow the patient's complete record to be examined. The registration record contains the patient's year of birth (the day and month are removed to retain anonymity), gender and marital status. Information on occupation and ethnicity is rarely recorded. The GPRD registration file also includes the date the patient registered with the general practice, the registration status, and, if the patient has left the practice, the leaving date. This is extremely useful because the file will also note deregistration as a result of death and so allow identification of deceased patients who have no record of death elsewhere. Information about the doctor and the practice including the region or country in which the practice is situated can also be obtained. More detailed information on locality is not available, to ensure the anonymity of the patients and the practices. Most practices in the UK are now linked to the local health authority and receive payment based on the number of registered patients. In the past, practice lists could become inflated because patients who had died or moved away remained registered.[9,10] Since the advent of links to health authorities the list of patients registered is a much more accurate record of the practice population.

The diagnostic records contain significant diagnoses or problems in coded form. Thus a record may include a diagnostic code for 'myocardial infarction' or an alternative code might be 'chest pain'. The notes files contain measurements such as height, weight, smoking habits, blood pressure, biochemical results, etc. In a recent GPRD study of women of childbearing age prescribed oral contraceptives we found 85 per cent of 1098 control patients had their body mass index (BMI) recorded, 92 per cent had their smoking habits recorded and 96 per cent had a record of their blood pressure. The notes file will also contain administrative events such as specialist referrals, in-patient episodes or referrals to other primary care workers. A recent validation, again of the GPRD, showed 85 per cent of a group of patients referred to hospital had their hospital referral recorded on the database, whereas in a similar validation of the MediPlus database the finding was only 65 per cent. The recording of these data is probably more complete in the practices contributing to the research databases than would be found in other UK practices. There is a certain amount of overlap between diagnosis and note records in terms of the type of data recorded. Thus chest pain may appear in either of the two files. It is important to remember this when conducting research. A strategy that only looks for the outcome of interest in the diagnosis file will miss those patients where it is recorded in the note file.

The prescription file is one of the most useful attributes of general practice databases and includes a record of over 95 per cent of all prescriptions issued by the GPs. The records include the product prescribed, the dose and duration, and whether it was a repeat prescription. The drug coding systems will identify drugs at a specific dosage and preparations uniquely. Usually the number of tablets or amount of medication prescribed is recorded. In the Meditel system there is also a code to indicate that a course of treatment has been stopped prematurely. Unfortunately, in all the databases the instructions on how to take the medication are written in free text. This may include dozens of variations. For example, the instruction 'take one tablet twice a day' may be written 'twice a day', 'twice daily', '2× daily', 'B.D.', etc. To calculate a daily dosage of medication for an individual patient from these data requires the ability to write a suitable program so that these instructions are all converted to a single code.

Uses of general practice databases

Opportunities for conducting observational studies using general practice databases include epidemiological studies of the prevalence and incidence of different diseases,[11,12] studies of health services research[13] and studies on pharmacoepidemiolgy.[14–16] The commercially available databases can provide data that have been subject to quality checks, are representative of the whole United Kingdom and are of sufficient size to give useful information about rare diseases. Before these databases are used, however, researchers need to be aware of not only the advantages but also some of the barriers and potential pitfalls.

Generalizability of studies using the databases

When considering the results of studies from general practice databases and before deciding their relevance to our own patients we need to consider the generalizability of the database. In other words, are the patients included on the database similar to other patients in the UK attending a GP and are the doctors that contribute to the database similar to other UK GPs? Over 95 per cent of the UK population is registered with a GP and individuals can only be registered with one doctor at any given time.[17] Thus general practice databases are population based and include patients of all ages. This means that the data are able to supply the denominator for incidence and prevalence studies in a way that hospital data cannot. When considering numerator data the completeness depends on the disease being considered. With some diseases such as meningococcal meningitis almost all the cases are likely to present to the health services. With others such as acne the presentation or not to a GP depends on many other variables. The database can only describe the incidence of presentation to the GP or the prevalence of diagnosed disease in the population. The interpretation of the findings requires detailed knowledge

of the influences on general practice and is one reason why studies using these databases are best done in collaboration with a practising UK GP.

All three include practices from throughout England, Scotland, Wales and Northern Ireland. There is a selection of urban and rural, large and small practices. Those patients missing from the database – the 5 per cent of patients not registered with a GP – are often the homeless and groups of patients who choose not to register. These include a higher proportion of young males and a higher proportion of people from ethnic minorities. The health of these groups will in many circumstances differ from the health of those registered with a GP, and investigation of their health needs by extrapolating the findings from general practice database studies would not be appropriate. However, we believe that the databases are highly representative of most other groups. This can be checked by comparing the outcomes and characteristics of the patients on the database with data from other sources such as Census, birth, mortality or prescribing data.

When considering the GPs it may be that the doctors who contribute to the database behave differently from others in the UK. We have no evidence that this is true and, again, comparison with other sources of data such as that obtained from the Prescriptions Pricing Authority suggests that there are more similarities than differences.[18,19]

Accuracy of the data

A key attribute concerns the presence of information about individual diagnoses. As with mortality and hospital episodes data the accuracy of diagnostic data is crucial. This involves both the validity and the completeness of the data. The validity relates to the likelihood that a patient with a diagnosis recorded on the database truly has that disease. This depends on the condition being diagnosed, the diagnostic aids available and the clinical acumen of the doctor. When considering a computer diagnosis such as diabetes, looking for further information can increase the likelihood that the record is valid. For example, evidence of regular diabetic monitoring such as raised HbA1c levels increases the likely validity of the code, as does the presence of a prescription for insulin or an oral hypoglycaemic medication. Further confirmation can be obtained by going back to the GP and asking for information particularly from hospital records. Because both the GP and the patient's identity are anonymized to maintain confidentiality this process of obtaining further information from the practice has to be carried out through an intermediary. It can become an expensive and time-consuming extra step and is not always necessary particularly when similar validations have been carried out in the past and the specificity of computerized data is already known. Examples of validation studies on general practice records include the studies by Nazareth *et al*. on psychosis[20] and Pringle *et al*. on diabetes.[4] The other component determining accuracy is the completeness of the recording or the sensitivity of the database; for example, if a patient has been diagnosed as having a myocardial infarction, what is the likelihood that this will be recorded on the computer record? Jick *et al*. reported a sensitivity for the GPRD of 87 per cent.[21] For some diagnoses such as known diabetes the sensitivity of the recording by GPs exceeds 90 per cent and our experience is that the specificity of the data, especially when considering relatively rare conditions, is close to 100 per cent. For these reasons we believe that these data are of sufficient quality for research. Both the GPRD and the MediPlus database reimburse the GPs for their time in providing quality data. The data being submitted to both these databases are checked for completeness and accuracy by running set queries on the data. Practices that fall below the expected recording standards are given the opportunity to improve their performance but repeated failure to meet the standards set will result in the removal from the database.

Pharmacoepidemiology

Pharmacoepidemiology is the use of observational studies to determine the benefits and hazards of pharmaceutical treatments within a population. The databases include excellent prescribing data. We estimate that in both the GPRD and in MediPlus over 95 per cent of prescriptions written by the GPs are recorded on the database. This is because the practices use the computer to generate their prescriptions. Examples of when prescriptions are not recorded on the database can include those provided during a home visit, private prescriptions and those for controlled drugs such as methadone.[22] Prescriptions from secondary care centres and specialized clinics, such as family planning clinics, are also not recorded on general practice systems. Because of the way pharmaceutical prescriptions are reimbursed in the UK, the vast majority are written by GPs. Consequently, these data on prescribing can be an excellent resource for pharmacoepidemiology. As well as the formulation prescribed, the dose and length of the prescription are also recorded for individual patients and can be linked to relevant outcomes (including death). Because the age and gender of all patients is known, adjustment for these important confounding variables can be carried out. There is also, as mentioned above, often useful information on height and weight, blood pressure and smoking habits. Although the missing data can cause a problem in multivariate analyses, useful inferences can often still be made.

The general practice databases are ideal for conducting rapid case–control studies particularly when exposure to a medication is the exposure of interest. However, because of the potential problems of accuracy and completeness mentioned above, care must be taken in designing algorithms for the identification of potential cases. Blinding the observers to the exposure of interest when selecting cases can be achieved by printing out records with the exposure encrypted or removed from the record. Longitudinal studies can also be carried out

using the databases. Methodologically, a major problem with all cohort studies is the issue of patients joining and leaving the study population, i.e. left and right censoring. Although left censoring can be dealt with by only using inception cohorts, or taking cohorts present on the database at a particular point in time, there is still the problem of losing patients to follow-up, or right censoring. All longitudinal studies have this problem. Researchers often try to reduce the problem by selecting cohorts with special characteristics to make follow-up easier (for example, doctors or those belonging to an insurance scheme). However, this introduces another form of selection bias because these groups are not representative of the general population. The general practice databases are more representative of the whole population and can quickly and efficiently give results based on large numbers. For mortality studies the death certificates of individual patients can almost always be obtained (again through an intermediary) and linked to a clinical history. Population movements and the inability to track patients from one contributing practice to another lead to difficulties in ascertaining previous health episodes. This can be problematic where the existence of prior events, or the identification of first ever prescriptions of a drug, are important. Any cohort study using a general practice database will have to be conducted within these structural limitations. Despite this, many useful pharmacoepidemiology studies have been carried out, using in particular the GPRD.

Audit and the use of enquiry software

Data from individual general practices can be used for research, especially when examining common conditions. More commonly, practices want to audit their data against agreed standards of practice. Most of the computer systems being used in general practice have a facility for identifying groups of patients and prescribing records. There is also available software for examining the data from individual practices and comparing audits between practices. One such enquiry software is called MIQUEST. The Department of Health has supported its development and use. Currently, it can only be used in practices with either the Meditel or EMIS systems. MIQUEST Interpreter software can be easily installed as an 'add on' and is free. Independent programmers have developed other software and these are able to make inquiries of other general practice computer systems. The benefit of using enquiry software is that the same data can be examined from a number of practices and compared.[23] This has been one of the purposes of the NHS Executive funded project, using MIQUEST and called 'Collecting health data from general practice'. This project has focused on improving the recording of data in general practice. It is also possible to aggregate the data from an enquiry into a database, which can be used for research. This, however, is time consuming – the same enquiry has to be run separately in each practice and then aggregated, rather than combining all the practice data into a database and running the query once. The writers of MIQUEST queries have to have an array of knowledge and skills, and these are at present in short supply. The main advantage of MIQUEST is that a small number of practices can be compared locally as part of an audit or local data can be collected for health needs assessment.

Outcomes research

Another field of interest for general practice databases involves outcomes research, particularly using economic analyses. In the United Kingdom the GP is the main point of access to the National Health Service (NHS). Consequently, general practice records contain considerable information on health services utilization, both of primary and secondary care. Discharge summaries from hospitals and other specialist centres are sent to the doctor as a matter of routine and important clinical information can be entered onto the computer (at present usually manually). Work is needed on validation of such data and better techniques are needed for attributing costs. One advantage of the Meditel system is that all records – notes, prescriptions and test results – can be linked under a problem number. Problem numbers are generally synonymous with a particular diagnosis or problem. This feature is especially useful for Health Service resource utilization studies. Because of the peculiarities of the NHS the generalizability of the findings from economic analyses to other countries is less valid than generalizing, say, drug reactions to different populations. As mentioned above, almost all United Kingdom practices are linked to the local health authority for registration purposes and the majority are linked for some item of service payments. Whereas these administrative links are extremely useful, the clinical links with hospitals are less well developed, with less than 10 per cent of practices contributing to the databases being linked for laboratory or other purposes.[22] However, when these links are established, outcomes research will become much more reliable.

Structural issues in using databases

Before undertaking research using these databases there are a number of structural problems that potential users need to address. The cost of paying all the contributing practitioners, the collection, quality checks and aggregation of the data means that the price of access is beyond the means of most researchers. These databases are enormous and need a computer powerful enough to hold the records of several million patients. Their use requires the employment of someone with the programming skills to manipulate large and poorly structured data sets. For example, researchers will often want to calculate the number of patient years at risk, or to estimate daily doses of therapy. As the databases are currently structured, these calculations are not possible without restructuring the data. When designing studies the help of a skilled statistician is also invaluable. With access to the appropriate statistical packages, it is also possible to

model longitudinal data. Clinical expertise in interpreting the data is also required. This includes an understanding of how GPs work and also specialist clinical knowledge of the area being examined. We believe the best use of these databases is likely to be achieved by academic units that have specialist knowledge in a particular area, expertise in epidemiology and statistics, and an understanding of general practice.

Conclusions

In conclusion, general practice databases are a rich resource for epidemiological research. They are population based and can quickly produce large samples of patients, and matched controls. The databases can be used to conduct cross-sectional, case–control and cohort studies rapidly and efficiently, and they provide good information on demographic characteristics, prescribing and diagnoses. The major barriers in using the databases are the cost of access and the costs associated with their analyses. Other disadvantages include the problems of bias, the handling of missing demographic data and the need in many circumstances to validate the database. Concerns have also been raised about the generalizability of the data. Comparisons with other sources suggest that the databases are, because of the number of patients included and the way the practices are selected, genuinely representative of the United Kingdom population. The generalizability is one of the databases' strengths. Proper utilization of the databases requires powerful computers, staff proficient in writing computer programs that facilitate analysis, and epidemiologists skilled in their use. These resources are expensive and would benefit from being centralized in a number of properly financed academic units. If the structural problems are overcome then the databases can provide an invaluable insight into the health and medical care of patients in the community. Their viability is dependent on their continued use and we believe more researchers should consider using them to help answer their research questions.

References

1 NHS Management Executive. *Computerisation in GP practices 1993 survey*. Leeds: Department of Health, 1993.

2 Purves IN. The paperless general practice. *Br Med J* 1996; **312:** 1112–1113.

3 Jick H. A database worth saving. *Lancet* 1997; **350:** 1045.

4 Pringle M, Ward P, Chilvers C. Assessment of the completeness and accuracy of computer medical records in four practices committed to recording data on computer. *Br J Gen Pract* 1995; **45:** 537–541.

5 Lancaster T, Surman G, Lawrence M, *et al*. Hormone replacement therapy: characteristics of users and non-users in a British general practice cohort identified through computerised prescribing records. *J Epidemiol Commun Hlth* 1995; **49:** 389–394.

6 Fleming DM, Watson JM, Nicholas S, Smith GE, Swan AV. Study of the effectiveness of influenza vaccination in the elderly in the epidemic of 1989–90 using a general practice database. *Epidemiol Infect* 1995; **115:** 581–589.

7 Wilkes S, Jones K. Retrospective review of the prevalence and management of infertility in women in one practice over a five year period. *Br J Gen Pract* 1995; **45:** 75–77.

8 Saint Yves IF. The Read Clinical Classification. *Hlth Bull Edinburgh* 1992; **50:** 422–427.

9 Fraser RC, Clayton DG. The accuracy of age–sex registers, practice medical records and family practitioner committee registers. *J R Coll Gen Pract* 1981; **31:** 410–419.

10 Sheldon MG, Rector AL, Barnes PA. The accuracy of age–sex registers in general practice. *J R Coll Gen Pract* 1984; **34:** 269–271.

11 Lawrenson R, Tyrer F, Pierce M. Mortality rates in diabetic patients from a community based population compared to local and age and sex matched controls. *Diabetic Med* 1997; **14**(10): 895.

12 Turnbull S, Ward A, Treasure J, Jick H, Derby L. The demand for eating disorder care. An epidemiological study using the general practice research database. *Br J Psychiat* 1996; **169:** 705–712.

13 Gibbs RG, Todd JC, Irvine C, *et al*. Relationship between the regional and national incidence of transient ischaemic attack and stroke and performance of carotid endarterectomy. *Eur J Vascular Endovascular Surg* 1998; **16:** 47–52.

14 Farmer RD, Lawrenson RA, Thompson CR, Kennedy JG, Hambleton IR. Population based study of risk of venous thromboembolism associated with various oral contraceptives. *Lancet* 1997; **349:** 83–88.

15 Garcia RL, Ruigomez A. Gastric acid, acid-suppressing drugs, and bacterial gastroenteritis: how much of a risk? *Epidemiology* 1997; **8:** 571–574.

16 Meier CR, Jick H. Drug use and pulmonary death rates in increasingly symptomatic asthma patients in the UK. *Thorax* 1997; **52:** 612–617.

17 Office of Population Censuses and Surveys. *General practice morbidity statistics; fourth national study 1991–2*. London: OPCS, 1995.

18 Bashford JN, Norwood J, Chapman SR. Why are patients prescribed proton pump inhibitors? Retrospective analysis of link between morbidity and prescribing in the General Practice Research Database. *Br Med J* 1998; **317:** 452–456.

19 Donoghue JM, Tylee A. The treatment of depression: prescribing patterns of antidepressants in primary care in the UK. *Br J Psychiat* 1996; **168:** 164–168.

20 Nazareth I, King M, Haines A, Rangel L, Myers S. Accuracy of diagnosis of psychosis on general practice computer system. *Br Med J* 1993; **307:** 32–34.

21 Jick H, Jick SS, Derby L. Validation of information recorded on general practitioner based computerised data resource in the United Kingdom. *Br Med J* 1991; **302:** 766–768.

22 Lawrenson RA, Coles G, Walton K, Farmer RDT. Characteristics of practices contributing to the MediPlus database and the implications for its use in epidemiological research. *J Informatics Primary Care* 1998(May); 14–18.

23 Pearson N, O'Brien J, Thomas H, *et al*. Collecting morbidity data in general practice: the Somerset morbidity project. *Br Med J* 1996; **312:** 1517–1520.

Accepted on 31 March 1999