Information in practice

Cross sectional survey of multicentre clinical databases in the United Kingdom

Nick Black, Marian Barker, Mary Payne

Abstract

Objectives To describe the multicentre clinical databases that exist in the United Kingdom, to report on their quality, to explore which organisational and managerial features are associated with high quality, and to make recommendations for improvements.

Design Cross sectional survey, with interviews with database custodians and search of electronic bibliographic database (PubMed).

Studies reviewed 105 clinical databases across the United Kingdom.

Results Clinical databases existed in all areas of health care, but their distribution was uneven—cancer and surgery were better covered than mental health and obstetrics. They varied greatly in age, size, growth rate, and geographical areas covered. Their scope (and thus their potential uses) and the quality of the data collected also varied. The latter was not associated with any organisational characteristics. Despite impressive achievements, many faced substantial financial uncertainty. Considerable scope existed for improvements: greater use of nationally approved codes; more support from relevant professional organisations; greater involvement by nurses, allied health professionals, managers, and laypeople in database management teams; and more attention to data security and ensuring patient confidentiality. With some notable exceptions, the audit and research potential of most databases had not been realised: half the databases had each produced only four or fewer peer reviewed research articles.

Conclusions At least one clinical database support unit is needed in the United Kingdom to provide assistance in organisation and management, information technology, epidemiology, and statistics. Without such an initiative, the variable picture of databases reported here is likely to persist and their potential not be realised.

Introduction

The need for high quality, multicentre databases that provide information about the use of health services has been thoroughly documented.[1–4] Briefly, they offer the opportunity to carry out evaluative research[5] and clinical audit,[6] to inform the planning and management of services,[7] and to provide individual clinicians with accurate estimates of the outcome of their care (assuming an accurate prognostic model exists) that can be shared with patients.[8] Despite their considerable potential, many multicentre clinical databases in Britain remain underused. This is because of a widespread lack of awareness of what databases exist and scepticism about their quality. Previous overviews of such databases have either not attempted to be comprehensive or not tried to provide independent reviews of quality.[9] In addition, little is known as how best to establish a good database, despite proposals that factors such as the involvement of epidemiologists and statisticians may be beneficial.[9]

In 2001 we created the Directory of Clinical Databases (www.docdat.org), which allows, for the first time, access to descriptions of the clinical databases that exist in the United Kingdom, including independent reports of their quality, and allows us to explore which organisational and managerial features of databases are associated with high quality and to make recommendations for improvements.[10 11]

Methods

Directory of Clinical Databases (DoCDat)

Inclusion is restricted to those databases that have information on the recipients of health care. We exclude databases in which information is limited to the provision of resources or services, such as a register of hospital bed provision, useful though such data are in studying and managing health services. We include databases that meet the following criteria (whether based on data collected retrospectively or prospectively):

- Provision of individual level data (whether or not users of the database are permitted to know patients' identities)
- Inclusion in the database is defined by a common circumstance (such as patient's condition or intervention required or undergone (which might be a diagnostic test, treatment, or a collection of interventions such as intensive care)); by an administrative arrangement (such as registered with a general practitioner, target for immunisation, subscriber to health insurance); or by an adverse outcome (such as maternal death)
- Data from more than one healthcare provider are included (usually many providers in a region or country).

We identified databases through inquiries to government health departments, royal colleges and specialist associations, and pharmaceutical companies; searches of previous reviews, research publications, and the internet; and word of mouth. Each entry was based on an in depth interview (usually by telephone) with the custodian of the database. Interviews lasted 30-60 minutes and followed a structured format. The principal aspects included were the geographical area covered, the length and periodicity of data collection, the number of patients or episodes collected to date, the use of nationally approved NHS codes, linkage to other databases, security and confidentiality of the data, the feasibility of ad hoc analyses, use of the data for audit and research, approval from professional bodies, the composition of any management team, and sources of funding.

In addition, the interviewer independently assessed 10 aspects of database quality using a tested instrument.[9] Five of these relate to data quality (completeness of recruitment, completeness of data, use of explicit definitions for variables, independence of observations of outcomes, and extent to which data are validated). The interviews were supplemented by an analysis of a bibliographic database (PubMed) to ascertain the number of research papers in peer reviewed journals that had made use of the databases.

Analysis of databases

We included all 105 databases for which a full entry existed in DoCDat in August 2003. We analysed the databases to describe the clinical areas they covered, their organisation and management, how data security and confidentiality were managed, what the databases were used for, and the quality of the data, and to explore any associations between characteristics of databases and the five dimensions of data quality. We tested the statistical significance of any associations with the χ^2 test.

Results

Clinical areas covered by existing databases

The clinical subject most commonly covered by the 105 databases was cancer (see box). While the 11 general databases, mostly regional and national cancer registries, were well known, there were 14 less well known specialised ones. Like many databases in other clinical areas, these were mostly developed by enthusiastic individuals rather than by national organisations. The next most common subjects covered were surgery (15 databases), congenital anomalies (14), and trauma and intensive care (9). Among the 23 databases for medical conditions, diabetes and cardiovascular conditions predominated. In contrast, only three databases related to maternity care, and only four related to mental health.

Organisation and management of databases

Most of the databases (66%) covered one or more UK nations (England, Scotland, Wales, Northern Ireland) (table 1). The rest were restricted to a region of one of these countries. Most (81%) had collected data continuously since being established, while most of the rest had been set up for a one off period.

Although half the databases used nationally approved codes to identify patients (that is, the NHS number), less than a third used approved institutional codes, and only 16% used clinician codes. A third had obtained explicit approval from the relevant clinical or professional body, such as a royal college.

The group or team managing a database varied in size and in composition. While almost all included doctors (95%), only 42% included nurses, and 26% included allied health professionals (such as physiotherapists). Most recognised the need for technical and methodological input from epidemiologists (69%), statisticians (79%), and information technology specialists (63%). In contrast, only a minority saw a need for representation from managers (32%) and laypeople (24%).

Funding came from a variety of sources. Most received some funds from the public sector (mainly the Department of Health or the NHS), though sometimes this was only pump priming to get the database established. Three other sources—private sector, subscriptions from participating healthcare providers, and charities—each provided finance for 10-20%. A few databases (5%) reported no funding. The distribution of funding partly reflects the absence from DoCDat of databases owned by private companies, despite our attempts to include them.

Number and examples of clinical databases by clinical area

General (9 databases)
For example:
- Hospital Episode Statistics
- General Practice Research Database
- MRC National Survey of Health and Development

Cancer (general) (11 databases)
For example:
- National Registry of Childhood Tumours
- Northern Ireland Cancer Registry
- North West Cancer Registry

Cancer (specialised) (14 databases)
For example:
- Scotland and Newcastle Lymphoma Group
- Assessment of Stomach and Oesophageal Cancer
- British Association of Surgical Oncology—Breast Unit Database

Surgical procedures (15 databases)
For example:
- UK Hydrocephalus Shunt Registry
- National Adult Cardiac Surgical Database
- North West Arthroplasty Register

Infectious diseases (3 databases)
- National Prospective Monitoring Scheme on HIV
- Nosocomial Infection National Surveillance Scheme
- UK Register of HIV Seroconverters

Congenital anomalies (14 databases)
For example:
- Oxford Register of Early Childhood Impairments
- Trent Congenital Anomalies Register
- Glasgow Register of Congenital Anomalies

Maternity (3 databases)
For example:
- St Mary's Maternity Information System
- Confidential Enquiry into Maternal and Child Health

Trauma and intensive care (9 databases)
For example:
- Intensive Care National Audit and Research Centre—Case Mix Programme Database
- Trauma Audit and Research Network
- All Wales Injury Surveillance System

Diabetes (6 databases)
For example:
- UK Diabetes Information and Benchmarking System
- Quality Indicators in Diabetes Service
- National Paediatric Diabetes Audit

Cardiovascular disease (7 databases)
For example:
- National Sentinel Audit of Stroke
- Myocardial Infarction National Audit Programme
- National Pacemaker Database

Respiratory disease (3 databases)
- UK Cystic Fibrosis Database
- Chronic Obstructive Pulmonary Disease 2001 Audit
- Scottish Asthma Management Initiative

Other medical conditions (7 databases)
For example:
- Northern Region Haematology Register
- Scottish Motor Neurone Disease Register
- UK National Renal Registry

Mental health (4 databases)
For example:
- Functional Analysis of Care Environments
- Carers and Users Expectations of Mental Health Services
- National Drug Treatment Monitoring System

Table 1 Organisation and management of 105 clinical databases with full entry in DoCDat, August 2003

Features	No (%) of databases
Geographical area covered:	
Two or more UK nations	45 (43)
One nation	24 (23)
Regional	36 (34)
Time frame for data collection:	
Continuous	85 (81)
Periodic	4 (4)
One off	16 (15)
Use of nationally approved codes:	
Patient codes	50 (48)
Clinician codes	17 (16)
Institution codes	30 (29)
Routinely linked to other databases	49 (47)
Approval by clinical or professional body	36 (34)
Composition of database management team:	
Doctors	100 (95)
Nurses	44 (42)
Allied health professionals	27 (26)
Epidemiologists	72 (69)
Statisticians	83 (79)
Information technologists	66 (63)
General managers	34 (32)
Laypeople	25 (24)
Funding:	
Public sector	76 (72)
Private sector	13 (12)
Provider subscriptions	12 (11)
Charity	19 (18)
None	5 (5)

Five of the databases were established over 50 years ago (table 2), four being cancer registries and one a longitudinal birth cohort. However, most databases were much more recent, with over half starting since 1990. While there is evidence that the establishment of new databases has accelerated in recent years, our figures refer only to those still functioning in 2003. It is likely that some databases established in earlier decades have since stopped, giving an underestimate of the incidence of new ones in that period. Information for the most recent years is also likely to be an underestimate because of delays in identifying new databases.

Table 2 Age, size, and rate of growth of 105 clinical databases with full entry in DoCDat, August 2003

Characteristics	No (%) of databases
Year established	
Before 1960	5 (5)
1960-9	6 (6)
1970-9	16 (15)
1980-9	14 (13)
1990-4	17 (16)
1995-9	29 (28)
2000 onward	8 (8)
Size (No of patients or episodes):	
<2000	15 (14)
2000-9999	27 (26)
10 000-49 999	25 (24)
50 000-1 million	26 (25)
>1 million	12 (11)
Growth rate (new patients or episodes/year)*:	
<100	5 (6)
100-499	15 (18)
500-1999	23 (27)
2000-9999	16 (19)
10 000-49 999	17 (20)
50 000-499 999	6 (7)
>500 000	3 (4)

*Based on 85 databases with continuous recruitment.

Table 3 Security and confidentiality of 105 clinical databases with full entry in DoCDat, August 2003

Characteristics	No (%) of databases
Data storage:	
Stand alone computer	30 (29)
Connected to external network	75 (71)
Back up data storage:	
Stand alone computer	5 (5)
Connected to external network	32 (30)
Disks or CDs	68 (65)
Paper forms stored	62 (59)
Confidentiality:	
Irreversibly anonymised	13 (12)
Reversibly anonymised	34 (32)
Identifiable	58 (55)
Patients informed of data collection:	
Individually informed	23 (22)
Collectively informed	6 (6)
Not informed	76 (72)
Patient consent for data collection:	
Signed consent	12 (11)
No signed consent but opt out	10 (10)
No signed consent or opt out	83 (79)

The databases varied considerably in size, from data on a few hundred patients or episodes of care to over 130 million (table 2). Not surprisingly, smaller databases tended to have been established more recently: the five smallest had all started in the previous five years. The three largest databases recorded hospital admissions for an entire country (Hospital Episode Statistics, Scottish Morbidity Record, and Patient Episode Database for Wales). The growth rate of databases with continuous recruitment also varied considerably. Five databases acquired fewer than 100 patients or episodes of care a year. These tended to cover rare conditions (such as acromegaly or motor neurone disease). In contrast, the national hospital inpatient databases accumulated more than half a million new episodes a year.

Security and confidentiality of data

Most databases (71%) stored their data on a computer connected to an external network, albeit with robust firewalls to prevent intruders (table 3). Back up versions of the database were usually on disks or CD Roms (65%), although 30% of databases backed up to another computer with external connections. Almost 60% also retained data on paper forms.

Ideally, data should be reversibly anonymised (by using key codes), so as to minimise the risk of disclosing individual identities[12] but to maximise possible use of the database.[13] This was true for only 33%. Of the remainder, 12% were irreversibly anonymised and 55% contained patient identifiers. For most databases (72%), the patients had not been informed that personal data were being collected, and for 88%, signed consent was not obtained.

Uses of databases

Irrespective of the purposes for which the databases were established, most could be used for all four of the principal applications cited in the introduction. Most (75%) allowed ad hoc analyses to be conducted centrally, where the data were aggregated and stored, but only 58% allowed such analyses to be

Table 4 Uses of 105 clinical databases with full entry in DoCDat, August 2003

Uses	No (%) of databases
Ad hoc analyses for providers:	
Possible locally (n=103)	61 (58)
Possible centrally (n=105)	79 (75)
Audit reports:	
Provider-specific (n=95)	42 (40)
Multicentre (n=95)	28 (27)
Research (No of papers* based on database):	
None	28 (27)
1-4	24 (23)
5-9	12 (11)
10-29	21 (20)
30-99	12 (11)
≥100	8 (8)

*Papers published in peer reviewed journals.

Table 5 Quality of data in 105 clinical databases with full entry in DoCDat, August 2003

Quality criteria	No (%) of databases
Completeness of recruitment:	
Few (<80%) or unknown	36 (34)
Some (80-89%)	9 (9)
Most (90-97%)	16 (15)
All or almost all (>97%)	44 (42)
Completeness of data:	
Few (<80%) or unknown	44 (42)
Some (80-89%)	19 (18)
Most (90-97%)	23 (22)
All or almost all (>97%)	19 (18)
Use of explicit definitions of variables:	
None	41 (39)
Some (<50%)	10 (10)
Most (50-97%)	16 (15)
All or almost all (>97%)	38 (36)
Independence of observations of primary outcome:	
Outcome not included	11
Observer neither independent nor blinded	23/94 (24)
Independent observer not blinded	11/94 (12)
Independent observer blinded or outcome is objective	60/94 (64)
Extent of data validation:	
No validation	3 (3)
Range or consistency checks	5 (5)
Range and consistency checks	69 (66)
Range and consistency checks plus external check	28 (27)

conducted locally by the providers of the data (table 4). Of the 95 databases that identified the healthcare providers, 40% produced audit reports on individual providers, but only 29% provided multicentre comparative reports.

The use of databases for research was also patchy. About a third were unable to provide a bibliography of peer reviewed journal articles based on use of their database. Our search of PubMed revealed that about a quarter of the databases had not been used for any articles and a further quarter had been used in fewer than five articles each. In contrast, eight databases had each been the basis of over 100 articles. These data probably underestimate the research output as PubMed does not facilitate searches by database name and some authors may fail to mention the database they used. In addition, some articles may have appeared in journals not indexed by PubMed, and some recently established databases would not be expected to have generated any research output yet. However, we may also have overestimated the output as some articles, while authored by database custodians and associates, may not have made use of the database in question.

Quality of the data

We measured the quality of the data in the databases against five criteria (table 5). Over half (57%) of the databases recruited at least 90% of eligible people or episodes. This underestimates the true prevalence of good databases, because up to a third of database custodians did not know their recruitment proportion. Similarly, over 40% did not know the level of completeness of their data. Of those that did, about 70% reported high levels of completeness.

About half the databases did not use explicit definitions for most of the variables collected. Of the 94 databases that included an outcome variable, 64% either used an independent, blinded observer (that is, one who was unaware of the intervention undergone) or this was unnecessary as the outcome was objective (usually survival). The data in almost all databases (92%) were subjected to range and consistency checks. In 27% some form of external validation (such as comparison with medical records) was also conducted.

Associations between characteristics of databases and data quality

Only one of the eight organisational characteristics we examined was significantly associated with database quality (table 6): databases that informed patients they were being included were more likely to have complete data. Given that 40 associations were tested, one statistically significant one (at $P=0.016$) may have occurred by chance. Seven other associations were significant at the 10% level ($P=0.1$): national databases were associated with better validation but poorer assessment of outcome than regional databases; approval from a professional body was associated with better validation; routine linkage was associated with better recruitment and validation; and having an epidemiologist or statistician in the management team was associated with better recruitment and validation.

Discussion

We found multicentre clinical databases in all areas of UK health care, though their distribution was uneven—cancer and surgery were better covered than mental health and obstetrics. They varied greatly in age, size, growth rates, and the geographical areas they covered. Their scope (and thus their potential uses) and the quality of the data collected also varied. The latter was not associated with any organisational characteristics. Despite impressive achievements, many faced considerable financial uncertainty.

Considerable scope exists for improvements: greater use of nationally approved codes; more support from relevant professional organisations; greater involvement by nurses, allied health professionals, managers, and laypeople in database management teams; and more attention to data security and ensuring patient confidentiality (something that most database custodians are currently addressing in meeting the requirements of the Patient Information Advisory Group[14]). With some notable exceptions, the audit and research potential of most of the databases was not being realised.

Study limitations

Although this review provides critical information on UK clinical databases for the first time, it is limited in three ways. Firstly, we may have missed some key databases. In particular, those included were restricted to the public sector, despite our attempts to gain access to information about databases held by

Table 6 Association between eight organisational characteristics of 105 clinical databases with full entry in DoCDat, August 2003, and five quality criteria. (Values are relative risk of characteristic being associated with good quality*)

	Quality criteria				
Organisational characteristics	**Recruitment**	**Completeness of data**	**Explicit definitions**	**Independent outcomes**	**Validation**
National coverage (*v* regional)	0.98	0.77	1.15	0.80	1.88
Continuous or periodic data collection (*v* one off)	1.16	1.72	0.91	1.25	1.47
Patients informed of data collection (*v* not informed)	1.30	1.79†	0.83	1.09	1.24
Approval by professional body (*v* none)	0.82	1.08	1.04	1.03	1.64
Linkage to other databases (*v* none)	1.30	0.80	1.24	1.16	1.75
Membership of management team (*v* none):					
Epidemiologist or statistician	1.58	0.85	1.13	0.86	1.87
General manager	1.31	0.93	1.04	1.08	0.96
Layperson	0.97	0.74	1.02	0.79	0.53

*Definitions of good quality. Recruitment: ≥90% of eligible population recruited to database. Completeness of data: 80% of variables 95% complete. Explicit definitions: 50% of variables have explicit written definition. Independent outcomes: independent observer blinded to intervention or objective outcome. Validation: range and consistency checks.
†P=0.016.

What is already known on this topic

High quality clinical databases can support evaluative research, audit, clinical management, and planning of services

There is a widespread lack of awareness of what databases exist in the United Kingdom, scepticism about their quality, and uncertainty as to the extent to which they are used

What this study adds

UK clinical databases exist in all areas of health care, varying greatly in age, size, geographical area covered, scope, and quality

The audit and research potential of many databases is not being realised

Considerable scope exists for improvements, which could be facilitated by a dedicated national support unit

pharmaceutical companies. Secondly, our information was largely self reported by database custodians. Despite careful interrogation, some accounts may be unjustifiably favourable. Thirdly, the lack of statistically significant associations between organisational characteristics and data quality (with one exception) may reflect the small sample size.

Implications of results

Considerable effort and resources are expended by clinicians, methodologists, computing staff, and others to create and maintain clinical databases in Britain. Many database custodians work in isolation with little contact or support from others engaged in similar activities. The need for improvements in database organisation, data quality, and data use is widely recognised among these highly committed individuals and groups. The inception, in March 2003, of a forum for database custodians to exchange experiences and help one another has both highlighted these needs and started to meet them. However, without resources to promote and support the development of databases, such measures can have only limited impact.

Just as clinicians are not expected to be able to organise and carry out randomised trials without the support of clinical trials units, so we should not expect high quality databases to be developed without some dedicated specialised support. This could be met by the establishment of at least one clinical database support unit. This could provide advice and assistance in organisation and management, information technology, epidemiology, and statistics. Without such an initiative, the variable picture of databases reported here is likely to persist, and their potential will not be realised.

We thank all the database custodians for contributing to DoCDat, and Faith Hummerstone and Anton Buxton for data collection and entry.

Contributors: NB conceived and designed the study, wrote the paper, and acts as guarantor. MB and MP carried out the analyses and commented on the paper.

Funding: DoCDat was funded by the National Centre for Health Outcomes Development.

Competing interests: The authors created and maintain DoCDat.

Ethical approval: Not needed.

1 Pryor DB, Califf RM, Harrell FE, Hlatky MA, Lee KL, Mark DB, et al. Clinical databases. Accomplishments and unrealized potential. *Med Care* 1985;23:623-47.
2 McGlynn EA, Damberg CL, Kerr EA, Brook RH. *Health information systems: design issues and analytic applications*. Santa Monica: RAND Health, 1998.
3 Black NA. High-quality clinical databases: breaking down barriers. *Lancet* 1999;353:1205-6.
4 Smith R. Is the NHS getting better or worse? We need better data to answer the question. *BMJ* 2003;327:1239-41.
5 Williams JG, Cheung WY, Cohen DR, Hutchings HA, Longo MF, Russell IT. Can randomised trials rely on existing electronic data? A feasibility study to explore the value of routine data in health technology assessment. *Health Technol Assess* 2003;7:iii, v-x, 1-117.
6 Scottish Audit of Surgical Mortality. www.show.scot.nhs/sasm/ (accessed 16 Oct 2003).
7 Department of Health. *Comprehensive critical care: a review of adult critical care services*. London: DoH, 2000.
8 Lundin J, Lundin M, Isola J, Joensuu H. A web-based system for individualised survival estimation in breast cancer. *BMJ* 2003;326:29.
9 Newton J, Garner S. *Disease registers in England*. Oxford: Institute of Health Sciences, University of Oxford, 2002.
10 Black N, Payne M. Improving the use of clinical databases. *BMJ* 2002;324:1194.
11 Black N, Payne M. Directory of clinical databases: improving and promoting their use. *Qual Saf Health Care* 2003;12:348-52.
12 Lowrance WW. *Learning from experience: privacy and the secondary use of data in health research*. London: Nuffield Trust, 2002.
13 Black NA. Secondary use of personal data for health and health services research: why identifiable data are essential. *J Health Serv Res Policy* 2003;8(suppl 1):36-40.
14 Higgins J. The Patient Information Advisory Group and the use of patient-identifiable data. *J Health Serv Res Policy* 2003;8(suppl 1):8-11.

(Accepted 4 April 2004)

bmj.com 2004;328:1478

Health Services Research Unit, Department of Public Health and Policy, London School of Hygiene and Tropical Medicine, London WC1E 7HT
Nick Black *professor of health services research*
Marian Barker *research assistant*
Mary Payne *research fellow*
Correspondence to: N Black nick.black@lshtm.ac.uk