

REVIEW

Exploiting the potential of routine data to better understand the disease burden posed by allergic disorders

C. Anandan*, C. R. Simpson[†], C. Fischbacher[‡] and A. Sheikh*

*Division of Community Health Sciences: General Practice Section, University of Edinburgh, Edinburgh, UK, [†]Department of General Practice and Primary Care: Primary Care Clinical Informatics Unit – Research, The University of Aberdeen, Aberdeen, UK, and [‡]Public Health Medicine Information Services Division, NHS National Services, Scotland, UK

Clinical and Experimental Allergy

SUMMARY

The Department of Health and Scottish Executive are currently undertaking independent reviews of allergy services in England (and Wales) and Scotland. Each review will assess the disease burden posed by allergic problems, involving secondary analyses of routine National Health Service (NHS) datasets. Major suggestions for re-structuring and/or re-focusing the NHS efforts to better deal with allergic disease are anticipated. The UK has some of the best datasets of routine health data in the world, but despite their strengths, they have important limitations. These include gaps in data collection, particularly in relation to monitoring of Accident & Emergency and out-patient consultations, and in-patient prescribing, thereby resulting in considerable under-estimates of hospital workload. The current gaps in service monitoring are likely to under-estimate the burden and workload associated with allergic problems, particularly in secondary care. One major limitation of existing data sources is the general inability to link individual patient level data between different datasets. By unlocking this potential there are very considerable potential gains to be made. Data linkage techniques currently being developed in the UK offer exciting new possibilities of looking across the primary-, secondary- and tertiary-care interfaces and also assessing short- and long-term social and educational outcomes in relation to allergic disorders. The current reviews of allergy services being undertaken need to be cognisant of these inherent limitations of existing data sources and would do well to recommend strategic initiatives that could enhance the availability, accessibility and quality of these datasets. Ideally, this should include investment in central data repositories staffed by teams with the necessary technical and statistical expertise, which would also take responsibility for progressing data linkage capabilities.

Correspondence:

Aziz Sheikh, Division of Community Health Sciences: GP Section, University of Edinburgh, 20 West Richmond St, Edinburgh EH8 9DX, UK. E-mail: aziz.sheikh@ed.ac.uk

Background

The Department of Health and Scottish Executive are currently undertaking independent reviews of allergy services in England (and Wales) and Scotland, these being mounted in response to public, professional and UK Parliamentary Health Select Committee concerns about the quality of care being delivered to patients with allergic problems [1–3]. Both governmental reviews will be underpinned by an assessment of the disease burden posed by allergic problems, these involving secondary analyses of routine National Health Service (NHS) datasets. Major suggestions for re-structuring and/or re-focusing the NHS efforts to better deal with allergic disease are anticipated.

The United Kingdom has some of the best datasets of routine health data in the world and these are increasingly being used for important epidemiological studies (see, e.g., [4–10]). Analyses of some of these sources of data were, for example, used to estimate the United Kingdom disease burden posed by allergic disorders as reported previously in this journal [11]. Despite their many strengths, existing datasets remain under-exploited, this most probably reflecting a combination of lack of awareness of the types of data available, coupled with a dearth of expertise in handling and manipulating such data.

In this editorial, we aim to provide an overview of the main sources and types of data on allergic problems available to researchers and service planners, assess the relative

strengths and limitations of existing datasets and highlight important developments in data linkage techniques, which could, if strategically supported, greatly enhance the types of epidemiological and public health questions that these sources of routine data could help answer.

Main data sources for allergic diseases

The main types of routine data available on allergic conditions are outlined in Table 1. In England, some datasets are available from the Office of National Statistics (ONS), the government agency that holds detailed national economic, social and health information, but others need to be obtained from specific agencies. In Scotland, data can be obtained primarily from one resource: the Information Services Division (ISD) of NHS National Services Scotland.

Below, we summarize the salient features of some of the main datasets available.

Hospital admissions

Hospital Episode Statistics contain information on cause and duration of hospital stay, by financial year, on patients admitted to and treated in hospitals in England. Information is also recorded on the type of admission (emergency, routine or transfer) and discharge destination. All diagnostic data are coded using the International Classifications of Disease (currently ICD-10) and can be used, for example, to study geographical variations in allergic disease pattern, understand disease aetiology and investigate disease trends over time [12–14].

The ISD captures similar data for Scotland, but a key strength of the Scottish dataset is the internal linkage of hospital episode statistics, which allows individual episodes to be aggregated into continuous inpatient stays; this dates back to 1989 in Scotland, but was only nationally introduced in England last year. Multiple admission records for individuals can therefore be linked with each other and also with subsequent death data from General Register Office for Scotland (GROS).

Primary care consultations

Practice Team Information (PTI) is a Scottish-wide system for the collection of primary care data from a sample of Scottish general practices serving around 6% of the population. Information is collected from the whole general practice team, including general practitioners, and practice/community nurses. This system was developed from Continuous Morbidity Recording (CMR), which collected data only from contacts with general practitioners. Information on the number of primary care contacts for allergic conditions can be obtained and can therefore be

used to estimate the prevalence of allergic diseases at a national (Scottish) level [15]. For asthma and other allergic problems, data start from 2001.

Similar, but not identical, data resources for England and Wales include the Royal College of General Practitioners' Weekly Returns Service that captures mean weekly general practitioner episodes for asthma, hayfever/allergic rhinitis (AR) and atopic dermatitis-related conditions in England and Wales.

Additionally, recently available data from the Quality and Outcomes Framework (QOF) can be used to estimate the prevalence of asthma, but do not at present provide details on any other allergic conditions.

Aggregated primary care databases offering detailed individual patient data

Other database resources that could be used to explore allergic disease incidence and prevalence, and also disease aetiology and outcomes, include the General Practice Research Database (GPRD), QRESEARCH, Doctors Independent Network (DIN) and the Independent Medical Services (IMS) (see Table 1).

The GPRD division is a dedicated multi-disciplinary team based at the Medicines & Healthcare products Regulatory Agency (MHRA) in London who operate the GPRD; this represents the gold standard of longitudinal anonymized patient databases from primary care. There are several ways to access the Full Feature General Practice Research Database (FF-GPRD), either as datasets, through commissioned research and reports or through secure on-line access (eg. 16).

QRESEARCH is a relatively new UK-wide general practice database derived from the electronic health records of over 9 million patients throughout the United Kingdom [17]. Data come from 519 general practices using the EMIS clinical computer system and include information from patients who are currently registered with the practices as well as historical patients who may have died or left. Records extend back to the early 1990s, making it one of the largest and richest general practice databases in the world.

The DIN system has been accumulating individual (de-identified) UK patient records for 10 years, producing a database greater than 50 million patient years in total size. The database is a direct download of normal clinical activity [18].

IMS health specializes in the collection and interpretation of anonymized health information. They have secure databases that are used by governments and other public, professional and educational bodies. IMS have developed a secure intranet for community pharmacy, to which more than one-third of pharmacists are now connected, and also a hospital bed management system.

Table 1. Examples of UK databases containing allergic disease data

Data source	Description
England	
Office of National Statistics (ONS)	The government agency responsible for compiling and analysing the UK's key economic and social indicators. The site holds detailed data that can be downloaded free of charge. This includes the Longitudinal Study (LS) for England and Wales
Prescription Cost Analysis (PCA)	PCA data are based on information obtained from prescriptions sent to Prescription Pricing Authority (PPA) for payment. GPs, dentists and hospital doctors who write prescriptions are included provided they were dispensed within the community. These data do not cover items dispensed in hospital or on private prescriptions. Prescriptions written in England, but dispensed outside England, are not included
Hospital Episode Statistics (HES)	The Health and Social Care Information Centre (HSCIC) is responsible for the HES database, which contains personal, medical and administrative details on all patients admitted to, and treated in NHS hospitals in England
Royal College of General Practitioners Weekly Returns Service (RCGP WRS)	Records weekly consultations for selected diseases from 72 general practices in England and Wales
Scotland	
Information Services Division (ISD)	ISD is Scotland's national organization for health information, statistics and IT services. Data relevant to allergic diseases include: 1. Practice Team Information – data on consultations in a representative sample of Scottish Practices; 2. Prescription Cost Analysis; 3. Data on asthma from the General Practice Quality and Outcomes Framework (a quality initiative that produces information as a by-product); 4. Information on hospital discharges linked to information on deaths provided by the General Register Office for Scotland (GROS)
Primary Care Clinical Informatics Unit – Research (PCCIU)	PCCIU collects data from practices participating bi-annually in the Scottish Programme for Improving Clinical Effectiveness in Primary Care (SPICE) and currently has a database of more than 300 practices (about 1.8 million patients). A database of all data available in primary care (anonymized to patient month of birth, sector-level post code with names and addresses removed)
The Scottish Longitudinal Study (SLS)	The SLS is a large-scale linkage study that has been created by using data available from current Scottish administrative and statistical sources. These sources include census data, vital events data (births, deaths, marriages), NHS Central Register (NHSCR) data (migration in or out of Scotland) and NHS data (cancer registrations and hospital discharges)
GROS	A complete population registration system. Routine mortality data for all causes including asthma and other respiratory diseases are available by age and sex. The GROS produces annual routine statistics on numbers and causes of deaths classified using ICD-9 and ICD-10 codes
England and Scotland	
ONS population and mortality	ONS produces annual routine statistics on numbers and causes of deaths in England, Wales and (using data supplied by GROS) Scotland
General Practice Research Database (GPRD)	The GPRD Division is a dedicated multi-disciplinary team based at the Medicines & Healthcare products Regulatory Agency in London. This is an anonymised patient longitudinal database of primary care records
Doctors' Independent Network (DIN)	DIN comprises of over 1500 GPs in 350+ General Practices. The DIN system has been accumulating individual (de-identified) UK patient records for 10 years, producing a database of greater than 50 million patient years. The database is a direct download of normal working activity
International Medical Service (IMS) Health	The IMS specializes in the collection and interpretation of anonymized health information, often being the only source of information on aspects of medicine utilization within and outside the UK, as well as across Europe. These complex, secure, databases represent an important source of insight into the management of disease and are used by governments, drug safety organizations, public health institutions, universities, professional medical bodies, professional pharmacy bodies, industry and patient organizations across the world
Q-RESEARCH	A UK-wide general practice database. It is a large database derived from anonymized health records of over 9 million patients. Data come from 519 general practices using the EMIS clinical computer system

NHS, National Health Service; ICD, International Classifications of Disease.

In Scotland, additional data sources include the Primary Care Clinical Informatics Unit (PCCIU), the GROS and the Scottish Longitudinal Study (SLS) (see Table 1). The PCCIU collects data from practices participating bi-annually in the Scottish Programme for Improving Clinical Effectiveness in Primary Care (SPICE) and currently has a database of more than 300 practices (about 1.8 million patients). The data extraction process captures patient-anonymized data reflecting the computerized clinical record.

The GROS is a complete population registration system. Routine mortality data for all causes including asthma and other respiratory diseases are available by age and sex. The GROS produces annual routine statistics on numbers and causes of deaths classified using ICD-9 up to 2000 and then ICD-10.

The SLS is a large-scale linkage study that has recently been created by using data available from current Scottish administrative and statistical sources. It is hoped that this will become available for use in early summer 2006. The SLS is a replica of the England and Wales Longitudinal Study (LS), which has been running successfully for the past 30 years.

Community prescribing

In England and Wales, Prescription Cost Analysis (PCA; Table 1) includes data provided by the Prescription Pricing Authority (PPA) but does not include data for items dispensed in hospital, private prescriptions or those dispensed outside England and Wales. The ISD also provides data on prescriptions, which contain information on items dispensed by community pharmacists, dispensing doctors and appliance suppliers. Prescription data have details of the number of items and the gross ingredient cost of all prescriptions dispensed in the community. These data are useful for assessing overall community prescription costs for allergy and also for allowing comparisons with other disease areas such as cardiovascular conditions.

Survey data

Although not technically routine data, repeat national survey data also consist of useful information that can be

used to inform research into allergic disease (see Table 2). Several UK surveys provide online access to detailed information pertaining to lifestyle, household and individual characteristics, most notable of which are the English and Scottish Health Surveys, which are periodically repeated. More specifically, Morbidity Statistics in General Practice (MSGP) is a series of surveys of general practice, which can be used to assess overall general practice consultation rates for allergic problems and has, for example, recently been used to look at variations in consultation rate by migrant status [19].

Limitations of existing datasets

Despite the many strengths of existing UK datasets, there are a number of important limitations. These include obvious gaps in data collection, particularly in relation to monitoring of Accident & Emergency and outpatient consultations, and inpatient prescribing, thereby resulting in considerable underestimates of hospital workload [11]. Similarly, there are also deficiencies in relation to measure community workload, such as use of ambulance services, walk-in centres and out-of-hours consultations with NHS Direct and NHS 24. The fact that 'over-the-counter' sales of drugs are not routinely recorded makes it difficult to estimate the true burden posed by AR. There are also generic issues across datasets, such as the poor recording of ethnicity, making it very difficult to assess ethnic variations in disease epidemiology or outcomes.

There are practical difficulties in accessing data, particularly in England, as the absence of a central source means that datasets have to be obtained individually from each data source. For example, PCA, Hospital Episode Statistics and MSGP all need to be separately procured. Accessing data is more convenient in Scotland where the majority of data are stored centrally and can be obtained from the ISD.

Perhaps the greatest limitation of existing data sources, however, is the general inability to link individual patient-level data between different datasets and it is through unlocking this potential that there are very considerable potential gains to be made (see Table 3).

Table 2. Examples of survey data available

Survey	Details
Scottish & English Health Surveys	A series of national surveys of the Scottish/English population aged between two and 74 living in private households. The Scottish Health survey produces national-level estimates, but provides only limited local information
Morbidity Statistics in General Practice (MSGP)	A decennial series of surveys of general practice. MSGP4 was a national survey of 60 general practices in England and Wales in 1991/1992 covering 1% sample of the population (over 500 000 patients)
Economics and Social Data Service (ESDS)	Available via the UK Data Archive (UKDA): the General Household Survey Annual Population Survey and The Labour Force Survey. Detailed data coverage on household, family and individual information

Table 3. Examples of the types of questions it may be possible to answer through linking individual patient data across routine datasets

How does risk of allergic disease vary by type of delivery?
Does risk of allergic problem vary by ethnic group and/or country of birth?
How many days off school/work are taken due to allergic problems?
What is the impact of hayfever on exam performance?
What are the referral rates between primary and secondary care for allergic disease?
What are the average life-time costs of treating patients with one or more allergic problems?

Future directions: data linkage

Data linkage techniques that are currently being developed in the United Kingdom offer exciting new possibilities of looking across the primary-, secondary- and tertiary-care interfaces and also assessing short- and long-term social and educational outcomes in relation to allergic disorders.

Linkage using personal identifiable information needs careful consideration as there is an inherent potential risk for breaching confidentiality. Important developmental activities and demonstration projects, both in Scotland and England, however, clearly show that such linkage is feasible and can, with appropriate safeguards and protocols, maintain the anonymity of patients.

The recently reported Scottish demonstration project involving the linking of the ethnic code in the census and health-care data exemplifies the steps involved in ensuring safe and accurate data linkage [20]. With approval and permission from the Privacy Advisory Committee, the Scottish Multi-Centre Research Ethics Committee and the GROS, ethnicity data from the census were linked with unique Community Health Index (CHI) numbers and then with the Scottish Morbidity Record, which had previously been augmented with mortality data. In order to convert the CHI number into a code that ensures it is impossible to reverse the process, a one-way encryption algorithm was used. Consequently, relevant hospital and death datasets were linked indirectly via the encrypted CHI number, after personal identifiers and other information not needed were discarded. This project found that it was possible to link 94% of 4.9 million census records to a CHI record.

A similar project is currently underway in England investigating the relationship between asthma severity, ethnicity and school achievement, which involves linking geo-referenced routine datasets from primary care, education, housing and local authorities [21]. Patient-level data from general practices are linked with other data via address, age and gender to produce household-level data. This process uses a purpose-written address-matching algorithm for linking data, a by-product of which is the creation of a consistent set of addresses, which serve as identifiers for linking pupils to households, their school and general practitioner. All person-identifiable data are removed once datasets are matched and geo-coded. If

successful, this work will offer considerable opportunities for exploiting hitherto almost completely under-utilized data sources held by a range of agencies to quantify the social/non-clinical impact of an important chronic health condition and public health problem.

Conclusions

Routine data offer considerable potential for understanding the disease burden and public health implications of allergic problems, and also important opportunities for aetiological investigation using both case-control and, increasingly, cohort study designs. There is also considerable opportunity for looking at long-term outcomes (such as disability and death) of experimental studies through the flagging of enrolled patients' notes. A comprehensive understanding of the disease burden posed by allergic problems does, however, require comprehensive datasets covering all NHS contacts. The current gaps in service monitoring are, however, likely to under-estimate the burden and workload associated with allergic problems, particularly in secondary care. Perhaps, the most exciting opportunity, however, lies in the possibility of linking between routine datasets to allow a range of hitherto unanswered questions in relation to allergic disease aetiology and outcomes to be assessed. The current reviews of allergy services being undertaken need to be cognizant of the inherent limitations of existing data sources and would do well to recommend strategic initiatives that could enhance the availability, accessibility and quality of these datasets. Ideally, this should include investment in central data repositories staffed by teams with the necessary technical and statistical expertise, which would also take responsibility for progressing data linkage capabilities.

Acknowledgements

Potential conflict of interests: A. S. is part of the Department of Health Allergy Review Stakeholder Group and together with C. F. serves on the Scottish Executive's Review of Allergy Services in Scotland Working Group. The authors have been supported by funding from the Chief Scientist Office (CSO) ref: CZG/2/252.

References

- 1 Holgate ST, Ewan PW, (eds). *Allergy: the unmet need*. London: Royal College of Physicians of London, 2003.
- 2 House of Commons Health Committee. The provision of allergy services. HC 696_I 2004. London: The Stationery Office. Available from: <http://www.publications.parliament.uk/pa/cm200304/cmselect/cmhealth/696/69604.htm#a3>
- 3 Levy ML, Price D, Zheng X, Simpson C, Hannaford P, Sheikh A. Inadequacies in UK primary care allergy services: national survey of current provisions and perceptions of need. *Clin Exp Allergy* 2004; **34**:518–9.
- 4 Peng MM, Jick H. A population-based study of the incidence, cause, and severity of anaphylaxis in the united kingdom. *Arch Intern Med* 2004; **164**:317–9.
- 5 Court CS, Cook DG, Strachan DP. Comparative epidemiology of atopic and non-atopic wheeze and diagnosed asthma in a national sample of english adults. *Thorax* 2002; **57**:951–7.
- 6 Court CS, Cook DG, Strachan DP. The descriptive epidemiology of house dust mite-specific and total immunoglobulin E in england using a nationally representative sample. *Clin Exp Allergy* 2002; **32**:1033–41.
- 7 Saxena S, George J, Barber J, Fitzpatrick J, Majeed A. Association of population and practice factors with potentially avoidable admission rates for chronic diseases in london: cross sectional analysis. *J R Soc Med* 2006; **99**:81–9.
- 8 Cassell JA, Mercer CH, Sutcliffe L *et al*. Trends in sexually transmitted infections in general practice 1990–2000: population based study using data from the UK general practice research database. *BMJ* 2006; **332**:332–4.
- 9 Tata LJ, West J, Smith C *et al*. General population based study of the impact of tricyclic and selective serotonin reuptake inhibitor antidepressants on the risk of acute myocardial infarction. *Heart* 2005; **91**:465–71.
- 10 Smeeth L, Boulis M, Hubbard R, Fletcher AE. A population based case-control study of cataract and inhaled corticosteroids. *Br J Ophthalmol* 2003; **87**:1247–51.
- 11 Gupta R, Sheikh A, Strachan DP, Anderson HR. Burden of allergic disease in the UK: secondary analyses of national databases. *Clin Exp Allergy* 2004; **34**:520–6.
- 12 Sheikh A, Alves B. Age, sex, geographical and socio-economic variations in admissions acute anaphylaxis: analysis of four years english hospital data. *Clin Exp Allergy* 2001; **31**:1571–6.
- 13 Alves B, Sheikh A. Age specific aetiology of anaphylaxis. *Arch Dis Child* 2001; **85**:348.
- 14 Gupta R, Sheikh A, Strachan D, Anderson HR. Increasing hospital admissions for systemic allergic disorders in england: analysis of national admissions data. *BMJ* 2003; **327**:1142–3.
- 15 Simpson CR, Lee AJ, Taylor MW, Helms PJ. Changing incidence of respiratory presentations in primary care fact or artefact? *Arch Dis Child* 2005; **90**:982–3.
- 16 Bremner SA, Carey IM, DeWilde S *et al*. Early-life exposure to antibacterials and the subsequent development of hayfever in childhood in the UK: case-control studies using the general practice research database and the doctors' independent network. *Clin Exp Allergy* 2003; **33**:1518–25.
- 17 Hippisley-Cox J, Stables D, Pringle M. QRESEARCH: a new general practice database for research. *Inform Prim Care* 2004; **12**:49–50.
- 18 Carey IM, Cook DG, De Wilde S *et al*. Developing a large electronic primary care database (doctors' independent network) for research. *Int J Med Inform* 2004; **73**:443–53.
- 19 Netuveli G, Hurwitz B, Sheikh A. Ethnic variations in incidence of asthma episodes in England & Wales: national study of 502,482 patients in primary care. *Respir Res* 2005; **6**:120.
- 20 Bhopal R, Fischbacher CM, Steiner M *et al*. Ethnicity and health in Scotland: can we fill the information gap? A demonstration project focusing on coronary heart disease and linkage of census and health records. University of Edinburgh, 2005.
- 21 Asthma UK. Is poor educational attainment in white and bangladeshi children related to asthma? A study linking health, education and social care datasets http://www.asthma.org.uk/applications/research_database/search.cgi?template=project_detail.html&dbname=nacdata&key2=819&taction=searchdbdisplay