The use of routinely collected computer data for research in primary care: opportunities and challenges

Simon de Lusignan\textsuperscript{a} and Chris van Weel\textsuperscript{b}


\textbf{Introduction.} Routinely collected primary care data has underpinned research that has helped define primary care as a specialty. In the early years of the discipline, data were collected manually, but digital data collection now makes large volumes of data readily available. Primary care informatics is emerging as an academic discipline for the scientific study of how to harness these data. This paper reviews how data are stored in primary care computer systems; current use of large primary care research databases; and, the opportunities and challenges for using routinely collected primary care data in research.

\textbf{Opportunities.} (1) Growing volumes of routinely recorded data. (2) Improving data quality. (3) Technological progress enabling large datasets to be processed. (4) The potential to link clinical data in family practice with other data including genetic databases. (5) An established body of know-how within the international health informatics community.

\textbf{Challenges.} (1) Research methods for working with large primary care datasets are limited. (2) How to infer meaning from data. (3) Pace of change in medicine and technology. (4) Integrating systems where there is often no reliable unique identifier and between health (person-based records) and social care (care-based records—e.g. child protection). (5) Achieving appropriate levels of information security, confidentiality, and privacy.

\textbf{Conclusion.} Routinely collected primary care computer data, aggregated into large databases, is used for audit, quality improvement, health service planning, epidemiological study and research. However, gaps exist in the literature about how to find relevant data, select appropriate research methods and ensure that the correct inferences are drawn.

\textbf{Introduction}

The founding fathers of academic primary care conducted research using routinely collected practice data. William Pickles' description of infectious disease,\textsuperscript{1} Frans Huygen's 'Families with their illness',\textsuperscript{2} and John Fry's 'Common morbidity'\textsuperscript{3} were produced in an era of paper data collection and provide examples of general practice research that changed the face of medicine. One can only speculate as to what more Pickles, Huygen and Fry would have achieved, had they been able to work with the large multi-practice databases that computers make possible today. The first steps towards automated processing of data in general practice took place in the 1960s and 1970s with the creation of age–sex registers. These registers consisted of an individual card for each patient stored by gender and age. Cards could be punched with a hole to signify the administration of an immunization or that the person had a diagnosis of a chronic disease which could then be sorted with a knitting needle. This paper reviews how data are stored in primary care computer systems, and the opportunities and challenges for using routinely collected primary care data in research. After setting the stage of primary care informatics (PCI) three sections subsequently review:

(1) How data are stored in primary care computer systems;
(2) Current use of large databases of routinely collected data and

\textsuperscript{a}Primary Care Informatics, Division of Community Health Sciences, St George's Hospital Medical School, London SW17 ORE, UK and \textsuperscript{b}Chair department of family medicine, Department of General Practice, University Medical Centre Nijmegen, 229-HAG, PO Box 9101, 6500 HB Nijmegen, The Netherlands. Correspondence to Simon de Lusignan; Email: slusigna@sghms.ac.uk
The opportunities and challenges of conducting research based on routinely collected data.

PCI

Zuboff coined the term ‘to informate’ to describe the automation of process of using computers to automate information flows and create the capacity to derive new information. Its scientific study is termed informatics (Fig. 1). Health informatics is the study of data, information and knowledge and how to use this to improve health. Primary care is a distinct specialty, supported by its longitudinal medical records and decision making processes, justifying its own health informatics subspecialty. PCI is emerging as a scientific discipline whose importance has increased in parallel with the computerisation of primary care and the consequent availability of large volumes of routinely collected data.

The adoption of computerised medical records is accelerating worldwide. Many countries have ambitious plans to integrate clinical records across all health providers with general practice records at their core. Family practice clinical computer systems are increasingly linked to other systems for patient registration and receiving laboratory test results. Scandinavia, The Netherlands and the UK have the longest tradition and highest level of computer use in general practice; though others are catching up. Integrating clinical records should improve patient safety, avoid duplication of tests, provide data to research and audit the effectiveness of care. This may be particularly important in improving the management of chronic diseases. However, there are gaps in the PCI literature, particularly on methodologies of using large datasets of routinely collected clinical data. Although such data have been used for epidemiology and analysis of management of (chronic) disease, less is known about how to utilise databases most effectively.

How data are stored in primary care computer systems

Data in general practice computer systems are stored as either narrative or structured data

Clinical computer systems record data in two ways. Firstly, they allow the recording of ‘coded’ (also termed structured) data; this is usually done by selecting from a picking list or entering data into some sort of form or template. Secondly, most clinical computer systems also allow the entry of ‘free text’ or narrative. At

![Figure 1](https://academic.oup.com/fampra/article-abstract/23/2/253/527321)

The benefit of implementing IT systems goes beyond simple automation of information flows. To ‘informate’ is the process of deriving new information as a result of automation.
present ‘coded’ data are needed because there are so many ways that a clinical concept can be represented. For example: a patient with coronary heart disease can be represented by any of the following free-text labels: ‘Diagnosis of myocardial infarction,’ ‘Raised cardiac enzymes,’ ‘Myocardial ischemia,’ ‘triple vessel coronary artery disease,’ ‘three vessel coronary artery bypass grafting,’ and so on. When looking to identify people with coronary artery disease it may be important to design a search strategy that includes all the ways a concept might be represented.

Natural language processing has not yet developed to the point to replace ‘coded’ clinical data
Unfortunately natural language processing (NLP), the process of searching the narrative record, has not reached the point where free-text can be automatically turned into ‘coded’ data; though progress is being made. NLP has been used to classify problem titles, to improve searching of bibliographic and genetic databases, to combine data from multiple discharge summaries, to automate links from medical text to the literature and to attempt to code data in email messages from patients. Although prediction is hazardous in rapidly developing fields, it is likely to be at least a decade before NLP can obviate the need for clinicians to code clinical data.

Coding systems and terminologies have got progressively larger and more sophisticated
A code is a simple representation (or label) given to something that allows it to be processed within an information system. Classifications provide a method of ordering information within a defined area or domain. The World Health Organisation’s (WHO) International Classification of Disease (ICD), now in its 10th revision, is perhaps the best known classification; though many subclassifications and versions of it are in current use. A terminology should provide comprehensive labelling of all the concepts within a domain. For example, it will enable the inference that a person with diabetes is also a person with an endocrine disease. However, linking it to fasting blood sugar level as a diagnostic criterion is beyond the scope of a terminology. There is, as yet, no international consensus about the meaning of some of these labels; we have, therefore, provided definitions in Table 1. Readers who wish to explore this in more depth should compare the glossaries provided by ‘Open Clinical,’ WONCA (World Organisation of Family Doctors), and the Australian Standard.

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Definitions of Code, Classification, Terminology and Nomenclature</th>
</tr>
</thead>
<tbody>
<tr>
<td>Code</td>
<td>A representation applied to a term so that it can be more readily processed</td>
</tr>
<tr>
<td>Classification</td>
<td>Arrangements of all elements of a domain, into groups according to established criteria</td>
</tr>
<tr>
<td>Terminology</td>
<td>Language labels attached to a concept—all terms of a professional domain</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

There is no single standard system for recording structured data
There is no international standard approach to coding and classification. Many countries use ICPC (International Classification for Primary Care) developed by the WONCA international classification committee. It is included in the WHO family of classifications and has been translated into 20 languages and modified to meet different countries’ needs. Its alphanumeric coding is ordered anatomically and allows coding of reason for encounter, diagnostic and therapeutic procedures and diagnoses, directed at health problems with a prevalence of 1.0 or more per 1000 patients. For this 400 codes suffice.

UK primary care mainly uses Read at present and is due to migrate to SNOMED CT (Systematised Nomenclature for Medicine—Clinical Terms). An overview of the history of these systems is set out in Boxes 1 and 2.

Other regions and countries have developed their systems to meet their needs. For example Australia has developed its own modification of ICD, ICD-10-AM (Australian Modification) and extended ICPC-2 to produce a nomenclature that included chronic conditions routinely managed in Australian primary care ICPC-Plus. In 2005, development of ICPC-3 got underway, seeking to cross-link to ICD-10, the ICF functional status, ATC drugs classification and others. It will also fundamentally revise the chapter on social problems. This way, ICPC will become the basic structure of health information.

The main difference between the different coding systems is their level of granularity. ICPC is the smallest (offering fewest coding alternatives), ICD-10 comes somewhere in the middle, and Read Clinical Terms (CT v3) provides a complex system. The impact of using...
different coding systems is illustrated in Box 3, where a Dutch GP using ICPC, compares their choice of codes with an Icelandic GP using ICD-10, and an English GP using CT v3.

Using large databases of routinely collected primary care data

Large GP datasets used in UK primary care research

Aggregated practice data presented in research have usually been collected with a single brand of computer system. Contributing practices may have had special training or feedback, and only those who have data above a certain quality level may be included. Reliance on a subset of practices can over-state data quality when generalising to routine practice.
Relatively little research has been done in comparisons between different computer systems and their advantages and disadvantages.44,45 The principal networks used for research in the UK are shown in Table 2.

Another approach to the acquisition of general practice data is to use data extraction tools which extract the data required for a specific study. Generally, studies using this approach extract data from a range of different computer systems.20 Hybrid solutions are also possible with a wider sample of data of general practices being extracted to answer a specific research question.20,46

Examples of large GP databases worldwide
Internationally, there are a large number of practice networks and databases available. They vary greatly in methodology, size and type of data collected. At the smaller end of the scale the Nijmegen academic family practice research network, founded in 1971, comprises four practices (11 practitioners) collecting a wide range of longitudinal data.47–50 At the larger end data from the Veterans Association are collected from American military establishments and from medical facilities which support ex-servicemen around the globe.51,52

Some GP data are still collected using paper forms. One of the best established is the National Ambulatory Medical Care Survey (NAMCS) in the USA, which collect data from sample practices about their case-mix and workload every year.53 Also BEACH (Bettering the Evaluation And Care of Health) in Australia uses paper to collect continuous data from 20 practices.54 Paper is also used in New Zealand, though the data are then transmitted centrally electronically.55 Examples of computerised data bases are summarised in Table 3.

The opportunities and challenges of routinely collected primary care data
These are summarised in Fig. 2.

Opportunities
Growing volumes of data are routinely recorded. More and more data are routinely collected using fewer coding systems. Telematics, patient held records, all add to the increasingly accessible volume of data.56

Some data, notably repeat prescribing, blood pressure and major morbidities are complete and accurate but other data less so.57 On their own, such data make little sense. Data items need to be related to individual patients’ characteristics: sex, age, social class, use of health care facilities, medical family history, life events etc.—the ‘thick and rich’ of primary care research.58

Improving data quality. Feedback,44 incentives,59 and evidence-based guidelines, all contribute to improvements in data quality. Once quality procedures are in place, it is possible to achieve high data quality standards.60,61 A complete and accurate dataset is important in clinical research as missing data are hard to

---

**Box 3 Comparing the coding of pneumonia in general practice using ICPC, ICD-10 and CTv3**

(1) Coding pneumonia with ICPC (International Classification of Primary Care)
A Dutch GP using this system is presented with a choice of only two options
(i) Bronchopneumonia OR (ii) Other pneumonia.
If this GP wishes to code in more detail they must use ICD.
(2) Coding pneumonia with ICD-10 (International Classification of Diseases and Health Related Problems)
An Icelandic GP will code problems using this system. They are presented with a choice of 80 codes. These relate to different causes of pneumonia including congenital cause, infections that may result in pneumonia and the type.
(3) Coding pneumonia with CT v3 (Read Clinical Terms version 3)
A UK GP using this system will be presented with a choice of 182 alternatives. These alternatives have far finer granularity than those offered by ICD, e.g. Right middle zone pneumonia. Many of these have lower levels with further codes available.

---

**Table 2 Large single computer system databases currently in use in the UK**

<table>
<thead>
<tr>
<th>Large database</th>
<th>Clinical system</th>
<th>Previous names</th>
<th>URL</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>General Practice Research Database (GPRD)</td>
<td>In-practice Systems (IPS)</td>
<td>VAMP, Reuters Vision</td>
<td><a href="http://www.gprd.com">www.gprd.com</a></td>
<td><a href="http://www.inps.co.uk">www.inps.co.uk</a></td>
</tr>
<tr>
<td>Mediplus database</td>
<td>iSoft-Torex</td>
<td>Torex Meditel</td>
<td><a href="http://www.ims-global.com/index.html">www.ims-global.com/index.html</a></td>
<td><a href="http://www.isoftplc.co.uk">www.isoftplc.co.uk</a></td>
</tr>
<tr>
<td>The Doctors Independent Network (DIN)</td>
<td>Torex Meditel</td>
<td><a href="http://www.dinweb.org/dinweb">www.dinweb.org/dinweb</a></td>
<td><a href="http://www.isoftplc.co.uk">www.isoftplc.co.uk</a></td>
<td></td>
</tr>
<tr>
<td>Q-research</td>
<td>Egton Medical Information Systems (EMIS)</td>
<td>None</td>
<td><a href="http://www.qresearch.org">www.qresearch.org</a></td>
<td><a href="http://www.emis-online.com">www.emis-online.com</a></td>
</tr>
<tr>
<td>Database</td>
<td>Organisation</td>
<td>Coding system</td>
<td>URL</td>
<td></td>
</tr>
<tr>
<td>----------</td>
<td>--------------</td>
<td>---------------</td>
<td>-----</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(2) Maastricht database</td>
<td>(2) Department of GP Maastricht University</td>
<td>(2) ICPC</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(3) Nijmegen academic family practice research network</td>
<td>(3) University of Nijmegen 1</td>
<td>(3) ICPC</td>
<td></td>
</tr>
<tr>
<td>Finland</td>
<td>Stakes statistical databases</td>
<td>National Research and Development Centre for Welfare and Health (STAKES, Finland).</td>
<td>ICD-9, now ICD-10</td>
<td><a href="http://info.stakes.fi/nettihilmo/english/default.htm">http://info.stakes.fi/nettihilmo/english/default.htm</a></td>
</tr>
<tr>
<td>Denmark</td>
<td>Continuous morbidity registration and quality development combined by using the Extended Danish International Classification of Primary Care</td>
<td>University of Odense, Dep. of General practice</td>
<td>Extended Danish ICPC (ICPC-E). A small part of ICD-10 relevant to primary care included in ICPC</td>
<td><a href="http://www.ulb.ac.be/esp/emd/dk_falkoe.htm">www.ulb.ac.be/esp/emd/dk_falkoe.htm</a></td>
</tr>
<tr>
<td>Norway</td>
<td>Data retrieval in general practice in Norway</td>
<td>Department of Community Health and General Practice University of Trondheim</td>
<td>All problems labelled using ICPC</td>
<td><a href="http://www.ulb.ac.be/esp/emd/no_grimsmo.htm">www.ulb.ac.be/esp/emd/no_grimsmo.htm</a></td>
</tr>
<tr>
<td></td>
<td>(2) CSerMEG: General Practice Research Group</td>
<td>(2) Centro Studi e Ricerche in Medicina Generale</td>
<td></td>
<td></td>
</tr>
<tr>
<td>USA</td>
<td>(1) Department of Veterans Affairs</td>
<td>(1) Department of Veterans Affairs</td>
<td>(1) SNOMED-CT</td>
<td><a href="http://www.hhs.gov/healthit/attachment_2/iii.html">http://www.hhs.gov/healthit/attachment_2/iii.html</a> <a href="http://www.musc.edu/PPRNet/index.htm">http://www.musc.edu/PPRNet/index.htm</a></td>
</tr>
<tr>
<td></td>
<td>(2) Practice Partner Network (PPRNET)</td>
<td>(2) Medical University of South Carolina (MUSC) + Physician Micro Systems, Inc. (PMSI)</td>
<td>(2) ICD- CM (Clinical modification)</td>
<td></td>
</tr>
</tbody>
</table>
As practices ‘informate’, procedures become routine so data quality improves; for example, computer data are used to screen for depression, recall patients and provide reminders to physicians.

**Technical progress.** Technical barriers to processing data are diminishing. Hardware can readily process larger volumes of data. Disk size is growing faster than text-based records. Communications technology allows integration of clinical records and data derived from home monitoring.

The importance of unique patient identifiers is increasingly recognised. Unique identifiers are critical if data relating to the same individual is to be linked. The ‘community health index number’ used in Scotland is a good example of a simple effective solution. The NHS number could fulfil the same role in England and Wales.

Proprietary and open source tools are making data extraction easier. Interoperability between systems is enabled by standards for exchange, management and sharing of information. For example, Health Level 7 (HL7) provides a framework for achieving this at the application level—e.g. allowing physicians using a clinical system to request an X-ray appointment on the radiology computer system.

**Opportunities exist to link GP data with genetic data.** A great deal of post-genomic research is generic, rather than ‘disease specific’ which makes generic primary care databases attractive for genetic research. Exciting opportunities exist to link primary care data with genetic data. As our knowledge about genetics grows computed data will allow exploration of the genetic–environmental balance in the causation of a wide range of diseases and even eventually enabling pharmacotherapy to be personalised. Iceland has developed a genetic database linked to clinical computer systems. UK Biobank will also link genetic and primary care data, emphasising the relevance of primary care databases in this research enterprise.

**International PCI community.** There is an international network of academics and practitioners with a special interest in PCI. Health informatics also has its own journals and conferences. Groups can be found in general practice family practice organisations as well as in the international informatics associations. The major organisations, their associated journal and conferences are shown in Table 4.

**Challenges**

**Identify appropriate research methods for analysing data from primary care databases.** Details about the research methods employed by researchers using primary care databases need to be described in much greater detail. Descriptions are often inadequate to allow other researchers to replicate the studies performed. Comparisons between databases and about the validity of data from well equipped practices have been made. However, it is often unclear how decisions were made to include or exclude practices, patients or individual data items from studies.

**How to infer meaning.** Research often requires an accurate denominator which is most easily achieved when health systems provide primary care listing of patients. Problems with health systems’ administration is likely to result in inaccurate reporting of morbidity nevertheless. Without a reliable denominator it is more difficult to calculate or compare incidence or prevalence. Surrogate markers (e.g. using prescriptions of thyroxine as a marker for the diagnosis of myxoedema) can be used to infer the incidence and prevalence of disease.

The architecture of the computerised record is also important; data linkage and problem orientation can facilitate research. Data-linkage, where treatments and interventions are linked to the relevant problem, can help with data interpretation. Unfortunately it is not a feature of all computer systems. In problem-orientated medical records data are readily grouped by ‘problem’ rather than being a simple chronological record.
Computers readily lend themselves to problem-orientated displays—so long as the data is appropriately coded. For example, in many systems it is possible to display all the consultations for hypertension; and the drugs prescribed for that problem.79

End points can be difficult to gather from primary care data and linking ICPC to functional status classification might present a vital development here. Capturing data about death or reasons that patients stop attending a physician can be difficult; though the French Sentinel network appears to have overcome many of these problems capturing reliable suicide data80 and in The Netherlands they have achieved the same in diabetes and heart disease.81

A recent systematic review concluded that lack of a gold standard for completeness and accuracy of data hampers assessment of quality.82 There is often marked inter-practice variation in data quality.83,84,85 For example in the Health Search Database (HSD) in Italy has shown that for some diseases, patient-reported prevalence is the same as the primary care record, in others it is lower;86 though the French Sentinel system found good correlation between hospital and practice diagnoses.85

Payments to GPs can distort coding practice. When GPs receive payments for specific diagnoses (pneumonia but not upper respiratory tract infection), interventions (for prescribing, but not for advice or wait-and-see) or performances (home visits, but not for nurse-led clinic) it is likely that GP records will report antibiotics-treated pneumonia in a home visit, rather than common cold that was advised by the nurse to wait-and-see. Medical certification for sickness absence can also skew data recording.87

Pace of change. The capacity and capability of computers to process data and the rate of change in health services appears to be accelerating. There are particular problems which can arise when changes occur.88,89 Data loss can be a problem when changing computer system or coding system. The Nijmegen family practice research network has overcome the risk of data loss by continuing the same e-book system for more than 30 years.90,91 This is a luxury not afforded to other data collecting networks which must change in line with reforms in their health system. This may become more of a problem as the pace of change accelerates.

Integrating systems. Problems with the lack of a reliable unique identifier for patients make linkage with other systems challenging. Where links have been made between morbidity data and socio-economic data these comparisons have proved useful.92

Ethical issues: data ownership, security, confidentiality and privacy. Ownership of data is a complex issue.
GPs collect data that are often shared by third parties (Departments of health, insurance companies) and to what extent should third parties be placed in a position to access these data for (their) research remains unclear. Clarification is needed about when identifiable or anonymised data can be used in public health. When and how consent should be obtained and the degree of anonymisation required also need better definition.93 However, it appears that patients may be in blissful ignorance about the risks of secondary use of their data than physicians are on their behalf. It appears that a minority of patients have clear opinions about whom they wish their records to be shared.94 Most (79%) patients are happy to see their data used for research by not-for-profit organisations.95

Discussion

Computerisation provides cheap and relatively easy access to large volumes of data; and data-quality is improving all the time. Informatics opens possibilities for research on a scale unprecedented in its history. This is an opportunity for innovation which should be capitalised on.

The challenges faced by researchers working with primary care data are the lack of tools to explore a narrative record, and inferring meaning when working with incomplete data. When studying the natural history of a disease, it is essential that all patients studied actually have the disease—we have not only to consider whether or not the GP did make a correct diagnosis but also whether the data were recorded correctly in the computer. When using computer data to study the use of health care resources and quality of care: the computer is likely to have an accurate record of the biomedicine that was practiced—whether or not this was relevant to the patient’s underlying health problem. As coding systems change we will need to work out how to incorporate historic data into longitudinal records. Ethical issues about data ownership and permissions remain problematic.

It is inevitable that routinely collected data will be used more and more in research. For research into health recourses use and quality of care, often ‘simple’ service data may suffice. However, for clinical research more in-depth supervised data are required. This needs investment into (research) practices and this will only be possible in a selected number of practices, linking to other databases (death certificates and secondary care) to further improve the quality of the data. Research in primary care requires large numbers of data, in particular when assessing the odds of serious, rare diseases and complications behind every-day symptoms and signs. Given the numbers of general practice databases available, an innovative scientific development might be linking international data bases and research networks.

Conclusions

Technological developments, the commitment of health services to implement them, improvements in clinicians’ skills and willingness to use IT have all contributed toward raising data quality to the point where routinely collected clinical data can be used for research. Routinely collected primary care computer data, aggregated into large databases, can be used for audit, quality improvement—especially chronic disease management, health service planning, epidemiological study and research. As our understanding of genetics grows the data can also be used to study the environmental–genetic balance in disease.

However, there are gaps in the literature about how to find relevant data; select appropriate research methods for its analysis, and ensure that the correct inferences are drawn. This paper sets out to fill some of these gaps and develop the health informatics evidence base.

Acknowledgements

The series editors: Professors Frank Sullivan and Azeem Majeed suggested the idea for the paper and have made helpful suggestions throughout its development.

References

14 Taylor H, Leitman R. European physicians especially in Sweden, Netherlands and Denmark, Lead U.S. in use of electronic


37. NHS Information Authority. SNOMED CT (Systematized Nomenclature of Medicine—Clinical Terms.) Available at: www.nhsia.nhs.uk/snomed/.


58. Herbert, CP. Future of research in family medicine: where to from here? Ann Fam Med 2004; 2(Suppl. 2): s60–s64.