Informatics in Primary Care 2006;14:203–9

Conference report

Routinely collected general practice data: goldmines for research?

A report of the European Federation for Medical Informatics Primary Care Informatics Working Group (EFMI PCIWG) from MIE2006, Maastricht, The Netherlands

Simon de Lusignan MSc MD MRCGP
Senior Lecturer, Primary Care Informatics, Department of Community Health Sciences, St George's, University of London, London, UK

Job FM Metsemakers MD PhD
Family Physician; Professor and Chair, Department of General Practice, Maastricht University, The Netherlands

Pieter Houwink MD
General Practitioner; Member of Board of GPs Amsterdam, ICT Portfolio; Chairman of Board of Advisors, OREGO User Group, Amsterdam, The Netherlands

Valgerdur Gunnarsdottir MSc MPH
Director of Health Informatics, Primary Health Care of the Capital Area, Reykjavik, Iceland

Johan van der Lei PhD MD
Professor, Department of Medical Informatics, Erasmus Medical Centre, Rotterdam, The Netherlands

ABSTRACT

Background Much of European primary care is computerised and many groups of practices pool data for research. Technology is making pooled general practice data widely available beyond the domain within which it is collected.

Objective To explore the barriers and opportunities to exploiting routinely collected general practice data for research.

Method Workshop, led by primary care and informatics academics experienced at working with clinical data from large databases, involving 23 delegates from eight countries. Email comments about the write-up from participants.

Outputs The components of an effective process are:

- the input of those who have a detailed understanding of the context in which the data were recorded
- an assessment of the validity of these data and any denominator used
- creation of anonymised unique identifiers for each patient which can be decoded within the contributing practices
- data must be traceable back to the patient record from which it was extracted
- archiving of the queries, the look-up tables of any coding systems used and the ethical constraints which govern the use of the data.

Conclusions Explicit statements are needed to explain the source, context of recording, validity check and processing method of any routinely collected data used in research. Data lacking detailed methodological descriptors should not be published.

Keywords: clinical records, general practice data, primary care informatics

Introduction

Much primary care research is based on pooled routinely collected general practice data.[1] Scandinavia,[2,3] The Netherlands[4] and the UK[5] have the longest tradition and highest level of computer use, though others are catching up.[6] Many countries have ambitious plans to integrate clinical records across all health providers.[7] Integrating clinical records should improve patient safety, avoid duplication of tests, provide data to research and audit the effectiveness of care.[8–10] This might be particularly important in improving the management of chronic diseases.[11,12]

Technology enables pooled data to be made widely available, but as yet there is no checklist of safeguards to help ensure that valid conclusions are drawn from these data. The strengths and potential weaknesses of these data have been known for some time, particularly the need to ensure data quality,[13,14] recognising that there might be gaps between the clinical record and actual performance.[15] However, there is a gap in our knowledge with no standardised approach to ensuring the quality of output from these databases.

We carried out this workshop to explore the opportunities and barriers to using routinely collected general practice data for research.

Workshop design

A full-day workshop was arranged for the day before the Medical Informatics Europe (MIE2006) conference in Maastricht. Invitations were sent to members of the European Federation for Medical Informatics Primary Care Informatics Working Group (EFMI PCIWG) and included in the conference programme.

An organising group (JM, SdeL and PH) designed the workshop and its objectives. The aim and design of the workshop was published within the conference programme for MIE2006. The planning for the workshop was largely carried out by email, with a final organising meeting immediately beforehand. The topic, large databases of pooled routinely collected general practice clinical data, was chosen as it is a topical issue as more and more primary care data are collected and also to reflect the strength of the 'registration networks' within The Netherlands. The location of the conference in Maastricht also provided the opportunity to visit practices that contribute data to the local Maastricht registration network and to see firsthand what impact this had on day-to-day use of computers in the practice.

The workshop ran as a single plenary session with opportunities provided for individual comments and questions posed to the group. PH, as a non-presenting organiser, chaired the meeting and facilitated discussion. Attendees at the group introduced themselves, described their use of computerised coding systems and their objectives for the workshop. Three short presentations were made with questions posed and discussions held during the talks.

Workshop presentations

Presentation 1: Primary care data – navigating between Scylla and Charybdis

The first presentation, by JvL, described dilemmas with primary care records: a story of Scylla and Charybdis. Odysseus travelled between the monster and the whirlpool. Scylla is a metaphor for paper records – an unmanageable monster. Paper records require more and more space, often lack structure and require an army of people to manage them. Charybdis is the whirlpool into which all our routine clinical data is sucked as our records become computerised. JvL described the risks of making routinely recorded data widely available using historical examples: the impact of printing on Erasmus' work and Burnum's predictions about data from the medical record.

Erasmus did not realise how contentious remarks, acceptable in personal letters, caused offence when printing (a new technology) made them widely available. His biographer wrote:[16]

> Erasmus, who never realised how insulting he was, always gave cause for misunderstanding and conflict. Norms and values were not yet adapted to the art of printing that increased the publicity of the written word a thousand fold.

Burnum highlighted how routine data might become disinformation:[17]

> With the advent of the information era in medicine, we are pouring out a torrent of medical record misinformation.
>
> All medical record information should be regarded as suspect; much of it is fiction.

Primary care clinicians are sailing between Scylla and Charybdis: the limitations of paper and the potential misuse of our computerised data. We need to ensure those reusing our data understand the details of the origins of the data. 'Fishing trips' with no pre-defined hypothesis risk undermining the value of primary care data.

Presentation 2: An exemplar of a research data collection network

JM presented the principles that should underpin the use of routinely collected data for research using the Maastricht 'registration network' (RegistratieNet Huisartspraktijken – RNH) as an exemplar.[18–20] The components of an effective network are:

- Clear scope and objectives. This network extracts a limited list of data it collects; its focus is on high-quality diagnostic disease data, including cause of death. Its data has been used for epidemiology and longitudinal study and to provide a sampling frame for more complex studies.
- Technical infrastructure. The network collects from one system 'MicroHis'. Patients within the system have robust unique identifiers. The interface allows reminders about recruitment or management of trials; for example, case report forms (CRFs) 'pop up' as reminders to the general practitioners (GPs) in the participating practices. Data are collated on a 'one line per patient' basis.
- Ethics and anonymity. The network sits within an ethical research framework and ethical approval is required to use the data. The research data have no strong identifiers: practices are anonymous to researchers; patients within a practice can only be identified within that practice.
- Quality control. The denominator is defined by practice registration and compared with the national population. Training is carried out to ensure coding takes place; in addition the clinical system has automated coding reminders.
- Recognised limitations. A limited dataset (for example, no ethnicity data) and collection from a single brand of general practice system are limitations which are acknowledged.

Presentation 3: Getting inside the black box – describing data processing

SdeL described the importance of documenting the context within which any processed clinical data are recorded and the details of how they are processed. Often these informatics issues are the epiphenomenon in any research and inadequately or not described within published research (see Figure 1). These two elements should be important to primary care informaticians and standardised ways of describing them developed.

The context of data recording can vary between brands of clinical computer and healthcare system. Only healthcare professionals involved in use of these systems at the time any data were recorded can provide the necessary insights. Programmes of research should include either direct or simulated methods for validating routinely collected data. The former might include painstaking hand searches through records (for instance, a hand search of 500 records of people with chronic kidney disease to validate a larger study[21]), comparison with other studies and simulation of a clinical case to explore how it might be represented in the clinical record (see Box 1).

The processing of clinical data often occurs in a 'black box'. The methods used in the Primary Care Informatics Group to overcome these were described.[22] These emphasise the need to archive extraction queries, code look-up tables and original data extract and then to have a controlled process through to the final analysis.

Workshop discussion

Delegates described very different arrangements for collecting routine clinical data in general practice. There was a spectrum of responses ranging from a

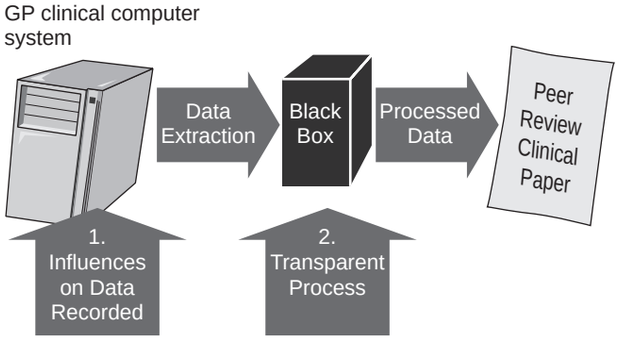

Figure 1 Informatics is the epiphenomenon in the processing of routine general practice data

Box 1 Simulated consultations to explore possible data recording – the same clinical history produces dissimilar histories and data

Patient history

- Mrs B is a 33-year-old woman, married with two teenage children. For the last two weeks she has been coughing at night and sometimes wheezing at night and after exercise. Her mother had asthma and her father (a heavy smoker) died three years ago of lung cancer.
- There is no history of asthma, eczema or hay fever. She has never smoked.

Clinical records made by Drs A + B

- Dr A records the following:

Problem title:	Asthma
History:	Night cough, wheeze
Examination:	Chest clear, peak flow 400
Prescription:	Salbutamol inhaler

- Dr B records the following:

Problem title:	Cough
History:	Worried she might have cancer as father started a cough and had lung cancer
Comment:	Never smoked, reassured

Coded data entry for Drs A + B

- Dr A:
  - Asthma NOS
  - Peak Flow = 400
  - Salbutamol inhaler 2p qid
- Dr B:
  - [D] Cough
  - Never smoked tobacco

single national system to a large number of inconsistently used systems.

- Iceland has a single clinical system used throughout primary care. The data are collected into regional databases, largely unexploited for research. Different coding systems are used for different data elements collected in the primary care consultation.
- In Croatia, approximately 60% of general practices use computers, but there is little standardisation of coding or clinical computer system.
- Czech GPs largely make free-text records.
- Germany has over 200 different clinical computer systems deployed in primary care.
- The Netherlands and the UK have almost complete coverage across primary care with five or six systems covering the country.

All population denominators have limitations. Countries where people can register with only one GP have advantages over countries where they do not have this restriction. Some factors, like high turnover of population, lead to medical records being less complete and to possible overestimates of the population denominator. Migration and illegal immigration exist to some extent in all populations. These individuals might have more health problems but might not appear in the population denominator. Central computerised registration systems reduce the number of so-called 'ghost' patients, but never completely eliminate them.

'Coding' of clinical data was seen as a distracter from the clinical consultation, but clinical coding was not seen to be an entirely negative process. A number of mechanisms for overcoming some of the barriers to clinical coding were identified, including: knowing that your data were contributing towards research; linkage of data to guidelines and to provide information to out-of-hours doctors; improved quality of care especially in the management of chronic disease; and to achieve financially incentivised quality targets. Problem-orientated records might help linkage of diagnosis or problem to investigation, therapy and referral.

Recording structured data was also associated with perverse incentives and some gaming was reported; these effects appearing to be greatest when financial incentives were included. Recording items in free text avoided triggers associated with recording structured data. Free text records might also be 'lost' within the

medical record if not coded and made into a problem. Financially incentivised targets make clinicians more wary of using codes associated with them.

There was little standardisation in approach to clinical coding and studies using routinely collected data should quote the degree of inter-practice variation in data recording. Practices differed as to what was summarised as a 'problem' in general practice records. Some practices liked to restrict the problem list to chronic diseases, others recognised that very often minor problems like 'cough' can end up on the problem list. Differences were described between practices, regions and countries as to how data were recorded.

Larger and more complex coding systems (such as Read 2–5-byte version) might generate more complex datasets than smaller, more compact coding systems (such as ICPC – International Classification of Primary Care), as the latter generates fewer potential pseudonyms for the same clinical context.

Routinely collected general practice data are already widely being used and it is not feasible to stop this. Rather than banning epidemiologists' 'fishing trips' through data looking for associations, the participants thought it more important to define principles that should apply to the processing of routinely collected data.

Discussion

The principal finding of the workshop was that unless a systematic approach is used to define the context of the data recording and its method of processing, then conclusions drawn from it might not be valid.

The learning from the participants has been synthesised into a list of recommendations that should apply for those involved in collecting and processing routinely collected data. The ten Maastricht rules:

1 State the research question or purpose for which the data will be used before starting the study or collecting the data.
2 Define the population denominator and its limitations; the population denominator and the unique identifier used to link patients to their data should be identified.
3 Record the characteristics of the practices involved in the study and how they might vary from 'usual' practice. International studies should take account of different cultural characteristics. This would include special training or payments.
4 Describe the context of data recording. Where feasible, include in the research project team at least one member who has consulted using the clinical computer system from which any routinely collected computer data are derived. Consider factors that might influence data recording. These include personal, cultural, technical, health system and financial factors and changes in disease definition or evidence base.
5 List all the coding systems in use at the time of the study and type of record system. Problem-orientated records encouraging linkage between problem and investigation, treatment and referral might be easier to interpret. Integrated systems that include comprehensive laboratory, referral, social and other information might offer richer data to the researcher. Use look-up tables contemporary to the data recording in analysing the data. Report any quirks of the coding system that might influence data quality.
6 Check the data quality of key variables. This can be done directly by hand searching notes or indirectly by comparison with other populations and simulation. Present data about inter-practice variation in data recording and how this might be explained.
7 Archive the queries and the original data extract as part of the governance process.
8 Describe the data processing in detail, especially any cleaning process.
9 Ethical approval, governance policy, adherence with data protection and any potential conflicts of interest should be clearly stated. Datasets should be anonymised. Practice identity should be invisible to investigators; individual patient identities should only be accessible within the general practice with which they are registered.
10 Audit trail: an audit trail should exist between the original data extracted and the final data.

Other literature has reported the value of problem-orientated medical records and comparing the outputs from different large databases[23] and lessons about the role of users in system development may be transferable to the domain of collecting and interpreting research data.[24] JvL, when stating the first law of informatics,[25] suggested that it was wrong to reuse data; this workshop report moves us on from that position, suggesting an extensive range of safeguards that must be in place if clinical data are to be reused.

The limitations of these findings are that they are based on a small sample of people attending an informatics conference. Further research is needed to test the assertions made as a result of this workshop.

Conclusions

This workshop has proposed that the processing of routinely collected data must include the input of primary care professionals who understand the context in which it was recorded and that data processing should be more transparent. We conclude that only routinely collected general practice data processed using these guidelines should be published.

ACKNOWLEDGEMENTS

We thank workshop participants, members of the EFMI PCIWG, the organising committee of MIE2006 and Dr Juergen Stausberg for his helpful comments.

REFERENCES

1 de Lusignan S and van Weel C. Routinely collected computer data: opportunities and challenges. *Family Practice* 2006;26:253–63.
2 Hasvold T. A computerized medical record: the 'Balsfjord system'. *Scandinavian Journal of Primary Health Care* 1984;2:125–8.
3 Krogh-Jensen P. Electronic records for general practice: the Danish system. *Scandinavian Journal of Primary Health Care* 1984;2:121–3.
4 Knottnerus JA. Role of the electronic patient record in the development of general practice in The Netherlands. *Methods of Information in Medicine* 1999;38:350–4.
5 Benson T. Why British GPs use computers and hospital doctors do not. *Proceedings of the AMIA Annual Fall Symposium* 2001: 42–6.
6 Taylor H and Leitman R. European physicians, especially in Sweden, Netherlands and Denmark, lead US in use of electronic routinely collected computer data: opportunities and challenges. *Harris Interactive: Healthcare News* 2002;2:1–3. www.harrisinteractive.com/news/newsletters/healthnews/HI_HealthCareNews2002vol2_Iss16.pdf.
7 Ash JS and Bates DW. Factors and forces affecting EHR system adoption: report of a 2004 ACMI discussion. *Journal of the American Medical Informatics Association* 2005;12:8–12.
8 Berner ES, Detmer DE and Simborg D. Will the wave finally break? A brief view of the adoption of electronic medical records in the United States. *Journal of the American Medical Informatics Association* 2005;12:3–7.
9 Tomlin A and Hall J. Linking primary and secondary healthcare databases in New Zealand. *New Zealand Medical Journal* 2004;117:U816.
10 Palomo L, Gervas J and Garcia-Olmos L. The frequency of illnesses attended and its relationship with the maintenance of the family doctor's skill. [Article in Spanish.] *Atencion Primaria* 1999;23:363–70.
11 Mitchell E, Sullivan F, Grimshaw JM, Donnan PT and Watt G. Improving management of hypertension in general practice: a randomized controlled trial of feedback derived from electronic patient data. *British Journal of General Practice* 2005;55:94–101.
12 de Lusignan S, Hague N, Brown A and Majeed A. An educational intervention to improve data recording in the management of ischaemic heart disease in primary care. *Journal of Public Health* 2004;26:34–7.
13 Pringle M and Hobbs R. Large computer databases in general practice. *British Medical Journal* 1991;312:741–2.
14 Black NA. Developing high-quality clinical databases: the key to a new research paradigm. *British Medical Journal* 1997;315:831–2.
15 Rethans JJ, Martin E and Metsemakers J. To what extent do clinical notes by general practitioners reflect actual medical performance? A study using simulated patients. *British Journal of General Practice* 1994;44:153–6.
16 Johan Huizinga, *Erasmus*, 1924.
17 Burnum JF. The misinformation era: the fall of the medical record. *Annals of Internal Medicine* 1989;110: 482–4.
18 Metsemakers JF, Hoppener P, Knottnerus JA, Kocken RJ and Limonard CB. Computerized health information in The Netherlands: a registration network of family practices. *British Journal of General Practice* 1992;42:102–6.
19 van den Akker M, Metsemakers J, Limonard C and Knottnerus J. *General Practice: a goldmine for research.* Maastricht: University of Maastricht, 2004.
20 University of Maastricht. *RegistratieNet Huisartspraktijken – RNH.* www.hag.unimaas.nl/rnh/ [Website in Dutch.]
21 Anandarajah S, Tai T, de Lusignan S *et al.* The validity of searching routinely collected general practice computer data to identify patients with chronic kidney disease (CKD): a manual review of 500 medical records. *Nephrology Dialysis Transplantation* 2005;20:2089–96.
22 de Lusignan S, Hague N, van Vlymen J and Kumarapeli P. Routinely collected general practice data are complex, but with systematic processing can be used for quality improvement and research. *Informatics in Primary Care* 2006;14:59–66.
23 Carey IM, Cook DG, de Wilde S *et al.* Implications of the problem-orientated medical record (POMR) for research using electronic GP databases: a comparison of the Doctors Independent Network Database (DIN) and the General Practice Research Database (GPRD). *BMC Family Practice* 2003;4:14.
24 Houwink P. The role of users and user groups in the continuing development of medical record systems for family physicians. *Medinfo*1995;8(Pt 1):310–12.
25 van der Lei J. Use and abuse of computer-stored medical records. *Methods of Information in Medicine* 1991;30:79–80.

CONFLICTS OF INTEREST

None.

ADDRESS FOR CORRESPONDENCE

Simon de Lusignan
Senior Lecturer, Primary Care Informatics
Department of Community Health Sciences
St George's, University of London
6th Floor, Hunter Wing
Cranmer Terrace
London SW17 0RE
UK
Tel: +44 (0)20 8725 5661
Fax: +44 (0)20 8725 3584
Email: slusigna@sgul.ac.uk

Accepted September 2006