Vault, cloud and agent: choosing strategies for quality improvement and research based on routinely collected health data

Simon de Lusignan BSc MBBS MSc MD(Res) FBCS CITP FRCGP
Reader in General Practice and Informatics, Division of Community Health Sciences, St George’s – University of London, London, UK

Frank Sullivan PhD FCGP
Professor of Primary Care Research and Development, University of Dundee, Dundee, UK

Paul Krause BSc PhD CMath FIMA
Professor of Software Engineering, Surrey University, Guildford, Surrey, UK

Introduction

The UK registration-based system for primary care collects some of the most complete and accurate longitudinal health data in the world. The reasons for this have been well described but include: an accurate denominator based on a national unique identification (ID; the NHS number in England and Wales or Community Health Index (CHI) number in Scotland); clinician coding of data at the point of care; linking or interoperability with other health service systems; and pay-for-performance based on routinely collected data. If practices agree to allow access to their records for a specific ethically approved research project there are a number of options on how patient identification and follow-up using their clinical records can be efficiently and safely achieved.

Datasets used for quality improvement (QI) and research are either derived from one or more clinical computing system, or involve linking primary care data with other datasets such as hospital data, or non-health data such as deprivation indices. Datasets derived from a single computer system vendor have produced research of international acclaim. For example, the GPRD (General Practice Research Database) is based on the INPS (In Practice Systems, Vision), and Q-Research is based on EMIS. Other groups have worked across different types of clinical computing system, mapping different data from different coding systems, local codes and the influence of very different methods of data entry. Probabilistic and deterministic record linkage methods allow those who undertake research and quality improvement to link clinical data to other datasets. The models for this linkage generally involve transient holding of strong identifiers such as full postcode, unique health service ID or name. However, new technical innovations allow this linkage to take place using fuzzy logic, without needing to hold strong identifiers.

Any method of holding data will need to meet necessary legal standards. In England there is a common law right to confidentiality, and a special exemption was made to this under Section 251 of the NHS Act 2006 (originally it was Section 60) to allow personal data to be used for medical purposes. Initially, use had to be approved by PIAG (Patient Information Advisory Group) but this has been superseded by the National Information Governance Board. Legislation is based on the European Human Rights Act within Europe, and on the Health Insurance Portability and Accountability Act (HIPAA) in the USA.

There are some well-established projects in the UK, and the Wellcome Trust and research councils have funded a number of new programmes of research that will be based on routinely collected data. Projects include a national service to extract general practice data for health service management and the intention to create a large network of practices willing to participate in research. This editorial urges those involved in working with large datasets to look carefully at the technologies they use. They may consider carrying on with data ‘vaults’. However, there are other options...
with significant potential benefits, such as ‘cloud’ computing, or avoiding the consent issues of holding date and using ‘agents’ to inform about numbers of cases held on computerised systems (Figure 1).

**Vault**

Currently, what we term the vault method is the only widely used approach. A ‘data dump’ is removed from the practice and stored in a remote secure server, the vault. The vault system is currently considered to be the most appropriate technology, as it allows physical security and controlled access to the dataset. Access is generally limited and often controlled by ethical committees. The downside of this approach is the reliance on physical security, the expense of running the vault, especially all its security features, and debate about whether such enormously complete and complex datasets should be held without consent. Many health service computing services are largely manned by contractors, some of whom may not fully appreciate all of the security issues. There is a steady trickle of cases in which people abuse their access rights to view records, and most vault solutions involve uncertainty about whether ‘opt-out’ (where patients can choose not to have their data extracted) or ‘opt-in’ (where only patients who consent are included) is needed.

**Cloud**

Cloud computing is an emerging approach to handling data. It offers the potential of cheaper and safer repositories for sensitive data than the vault approach. Data are encrypted and dispersed using special obfuscation applications and, contrary to the view of some current legislation, are potentially more secure than in a vault. The design and testing of a cloud system is complex and time consuming as it relies on a software solution rather than physical isolation for security. However, elements of what is required are emerging. For example, the SAPREL approach (used by SdeL) involves separate encryption to different servers of different strong identifiers, and then relies on fuzzy logic to link them without ever needing to hold patients’ strong identifiers; an approach highly commended by PIAG. Major international information technology companies are also starting to take cloud computing seriously, and it has even been used to hold the data for a cancer network. The upside of cloud computing is that the type of access can be determined at the design stage, with layers of indirection and encryption potentially providing better security. Its downsides are its newness and the need to build confidence in the techniques of security and recoverability that offer (perhaps surprisingly to some of those reading this) significant cost and performance advantages over data vaults, and the fact that the legal status of data held in the cloud are not fully defined.

**Agent**

Agents are autonomous encapsulated computer systems, which are capable of flexible interactions to achieve their objectives. Agents are one of the two types of ‘robots’ that exist within software engineering. Agents
carry out purposeful actions, interacting with other systems. The other type of robot is the ‘web crawler’ or ‘spider’ which collects materials from web pages for use in search engines. Agents have been around in computing for many years and more recently have been used in health care. This journal has published papers on their use to activate text message reminders to patients about their appointments. Agents have also been used to identify patients who are eligible for studies. This has become a standard approach in some of the American primary care research networks, and in this edition of the journal we carry a paper on how they have been used to identify people for a diabetes trial in the UK.\(^\text{14}\) Agents are an important alternative to centralised data repositories as they obviate the need for the researcher to hold practice data and sidestep any need for consent to examine records to see if people are eligible for a trial – avoiding the ‘consent for consent’ dilemma. An additional advantage of agents is they can quickly provide an estimate of the numbers of eligible people in a population to determine the feasibility of a proposed study. The downside is that whilst researchers may find eligible patients, other methodologies are required to determine how typical these cases are and how they compare with the rest of the practice population.\(^\text{15}\)

### Conclusions

The UK is in a unique position in which research can be conducted for the benefit of everyone, especially to improve the health care of our patients. Our national registration system provides unique opportunities for long-term follow-up of individuals and families. The accurate denominator; comprehensive use of computer systems in general practice at the point of care; and current and future increased linkage of systems further increase the value of our clinical records as a resource for research and quality improvement. This use will ultimately result in improved patient care.

Whilst some may argue we should not tinker with a system which is far from broken, our view is that cloud computing offers potential advantages over vault and that agent technologies may well turn out to be the technologies of choice for recruitment across large populations. This is notwithstanding the excellent recent Wellcome report on the use of data, which suggests that getting ‘opt-in’ consent may be a better long-term model.\(^\text{16}\)

### Recommendations

- Studies are urgently needed to compare vault, cloud and agent technologies using simulations and then incorporation into real studies.
- Do not expect there to be a single preferred solution. Rather each approach, and hybrids thereof, may provide the best solution in different circumstances.
- The consent dilemma will not go away and in the meantime we should be learning what methods and investment are needed to gain active consent from patients.

### REFERENCES

10. Electronic Patient Records and Databases in Research. www.wellcome.ac.uk/Funding/Biomedical-science/Past-funding/UK-funding/wtx055665.htm


**ADDRESS FOR CORRESPONDENCE**

Simon de Lusignan  
Reader in General Practice and Informatics  
Division of Community Health Sciences  
St George’s – University of London  
London SW17 0RE  
UK

*Accepted February 2010*