GATEway to the Cloud

Case study: A privacy-aware environment for Electronic Health Records research

Rob Smith, Professor Jie Xu, Saman Hima & Dr. Owen Johnson, School of Computing, University of Leeds, Leeds, UK

We describe a study in the domain of health informatics which includes some novel requirements for patient confidentiality in the context of medical health research. We present a prototype which takes health records from a commercial data provider, anonymises them in an innovative way and makes them available within a secure cloud-based Virtual Research Environment (VRE). Data anonymity is tailored as required for individual researchers' needs and ethics committee approval. VREs are dynamically configured to model each researcher's personal research environment while maintaining data integrity, provenance generation and patient confidentiality.

Keywords— privacy; health informatics; virtual research environment; anonymity; natural language processing

I. Introduction

Several branches of medical research focus on the study of patient health records. These records are generated by patient interactions with health professionals including but not restricted to General Practitioners (GPs) and Nurse Practitioners, nurses, hospital doctors and specialists.

Health records are a rich source of research data because they contain information related directly to illnesses alongside information providing potential context, such as lifestyle, risk factors and family history. Health records are usually documented as coded information (codes referring to conditions, medicines, interventions etc.) and unstructured free text.

In reality, practitioners do not always use codes when they ought to and do not always use them correctly, so much of the information available to research is hidden within free text. The type of information found in free text fields varies greatly: it depends on the details of discussions between patients and medical practitioners; the individual habits of those practitioners; and on what is considered relevant in the context of a particular consultation. As a result, health records are likely to contain a great deal of personal information about patients [1].

This might include data the patient did not expect to be recorded and data about other people, such as family members, who have not given permission for the storage or sharing of their data. Crucially, it might contain information which would allow the patient to be identified.

Some data recorded in health records is *obviously* personal (for example, names, locations, professions) whereas some is more subtly personal (such as natural language descriptions of relationships, activities etc.) [2]

It is a duty of those who provide researchers with health record data to ensure that the personal privacy (and especially the confidentiality) of patients is protected [3]. However, personal information is often important to researchers as it contextualises interventions. Such context information is often the reason researchers require access to health records in the first place. For example, family history and lifestyle might be used to identify statistical risk factors or relative effectiveness of interventions.

This type of correlation might be gleaned from analysis of records en masse but there are two problems of particular concern:

1. It is difficult to obtain such information from free text [4], and
2. Doing so might violate the privacy of patients and other individuals mentioned in the patient's records [5].

The challenge is to make available the information researchers need without violating the privacy of the patients who own that data. To achieve this goal, the JISC GATEway to the Clouds project has built a prototype of a cloud-based Virtual Research Environment (VRE).

VREs are self-contained environments which are preloaded with the data appropriate to the research to be conducted by individual researchers. A VRE can replicate a researcher's familiar personal research environment by incorporating their habitual tools and other data sources. This means that researchers can conduct experiments within their VRE rather than having to download the data and take them elsewhere, which constitutes a privacy risk. The project has also developed a process for using natural language processing (NLP) to deliver data into VREs at customisable levels of anonymity corresponding to individual researchers' needs and ethics committee approval.

II. Background

The case study centres on a longstanding collaboration between Leeds University and The Phoenix Partnership (TPP). TPP has developed a clinical information system, SystmOne, which connects different healthcare organisations. SystmOne provides a single interface for medical professionals to access and update patient data throughout that patient's lifetime of care.

With SystmOne, details of every appointment, medication, illness, allergy and contact a patient has ever had can be documented in a single location and made available to healthcare professionals within the context of a health consultation or intervention. It is a hosted solution, so data can be shared securely between a range of healthcare settings including GPs, child health, urgent care, palliative care, hospitals, mental health and social care with the emphasis on 'one patient, one record'.

TPP therefore has great deal of patient health data recorded by health professionals including approximately 23.5m patient records from 16m GP-registered patients. The dataset covers the social, primary and secondary NHS Electronic Health Records of a representative coverage of patients in England.

This data is very attractive to medical researchers and TPP has a longstanding collaboration with Leeds University within which it can share anonymised data under appropriate ethics committee approval. Ethics approval is a vital but time-consuming and often frustrating process, usually taking months to arrange.

Another issue is that approved anonymised health records are often delivered to researchers by insecure methods such as disk or email. Researchers import the data into their own research environment and are responsible for its security thereafter. This approach can lead to privacy problems. There are well-known examples [6] of large volumes of personal data being mistakenly left in public places when media or laptops are stolen or mislaid.

Similarly, it is unreasonable to assume in general that the environments in which the data are housed are adequately secure or that the data will not be misused [7].

A final problem is that anonymisation is an expensive, time-consuming and error-prone process [8]. There are two aspects to this: first, it cannot easily be guaranteed that all personal information is removed from a large volume of data during the anonymisation process; and second, it cannot be guaranteed that information vital to a particular research endeavour is not accidentally removed.

The JISC project *GATEway to the Cloud* has built a prototype solution to address some of these issues. On-going activity at Leeds University aims to develop an industrially-robust production system based on that prototype.

The remainder of this paper will describe the GATEway prototype and the novel privacy issues surrounding it.

III. VRE Requirements

The case study determined several novel requirements for deploying Virtual Research Environments to conduct health records research.

A guiding principle for a VRE in this instance is that it should resemble individual researchers' normal working environments as closely as possible. This means that it must be able to accommodate the tools and data the researcher would ordinarily use to conduct their research. It should be possible for standard tools to be pre-installed into VREs and managed by VRE administrators and for other tools to be installed and managed by users. Different OS options and versions should be available and customisable by administrators and researchers.

A VRE should contain the data a researcher is entitled to at an appropriate level of anonymisation. This should be customisable to individual researchers' needs: for example, some research endeavours require location, family or historical information, whereas others are concerned solely with treatments prescribed for certain conditions.

VREs and the workflows used to manage their lifecycles must be subject to audit. The audit trail should generate provenance related to the research processes and output. Provenance data can aid in the replication and verification of experimental results and the resolution of disputes about privacy. For example, it should always be possible to determine the specific dataset, level of anonymity and researchers involved in generating a particular set of results.

The computing resources available to a VRE should be dynamically customisable according to researchers' needs. If a researcher's environment is limited by resource, she is likely to remove patient data from the VRE and relocate it into an ungoverned environment, creating privacy risks.

Access to a VRE should be customisable to specific researchers' needs and ethics committee requirements. In some cases, data might be accessed only from a specific machine or from within a specific network (or VPN) or organisation. In others, it might be accessible over the Internet. The ultimate goal is that the system be sufficiently trustworthy that data providers and ethics committees agree that Internet access be the norm, but in the meantime, flexibility is vital.

The protection of patient privacy - and especially confidentiality - is paramount. In practice, this must place privacy management largely in the hands of patients themselves, who must be able to decide how and under what circumstances their records may be used. It also requires that management of privacy becomes a joint, co-built activity involving patients, medical practitioners, medical data providers and VRE administrators.

Some of these requirements are novel. For example, the concept of fine-grained customisation of anonymity levels, generated automatically by NLP as part of a mutually-managed research pipeline has not yet been attempted. Likewise, the idea of the co-management of privacy as an integral part of the research environment lifecycle alongside provenance has not been fully addressed.

IV. PRIVACY

The principal privacy concern in this domain is patient confidentiality [9, 10, 11]. This is achieved partly through anonymity and partly through the principles of notice and consent.

A. Anonymity

Records in a VRE are anonymised. This is the first line of protection for patients. Anonymisation means that a record should not contain personal information such as proper names.

Part of the anonymisation process is straightforward: remove personalised information from the name fields in the patient record. However, names also appear in free text fields within the record. For example, a GP might refer to the patient by name while discussing an appointment ("Mrs Hussain complained of migraine…") While humans are adept at identifying names within free text, it is time-consuming and error-prone. Given the volume of data involved, natural language processing (NLP) of free text data must be employed.

The problem is complex. For example, it is not always obvious what constitutes a person's name. Look-up tables of names are useful but not adequate since unusual names might be missed, names might be misspelled or used in unusual configurations. There is also the possibility of false positives, since many names are derived from place names, professions etc. (for example, Windsor can be either a surname or a place name; Smith can be a surname or a profession). If done without care, this can lead to important contextual information being identified as a name and erroneously removed from a VRE's dataset. For these reasons, a complex set of NLP rules is required to identify even the simplest and most direct personal information.

The removal of personal information from patient records is more complex still. For example, consider the following as part of a free text field compiled by a GP:

```
Ms Hutchinson's father, Allen, aged 82
was admitted to the Royal Victoria
Infirmary in July 2009 suffering from
heart disease.
```

There are four potential privacy issues for patients:

1. Ms Hutchinson is mentioned by name and is therefore identifiable.

2. Even if her name were removed, Ms Hutchinson might be identified from the information about her father.

3. It might be possible to make inferences about Ms Hutchinson's health based on her father's medical history.

4. The record contains personal information about Ms Hutchinson's father, who has likely not given permission for it to be shared.

The anonymisation process must be able to cope with scenarios such as this, which is why complex, context-dependent and domain-specific NLP is required.

B. Levels of anonymity

Different medical research scenarios require different views of the same dataset. Personal information can contextualise data and fully anonymised records might not be suitable for some purposes. In the example of Ms Hutchinson above, her father's medical history might be medically relevant for some studies, but not others. As a matter of privacy principle, researchers should not be granted access to identifiable data they do not need and do not have ethics approval to use.

For this reason, different *levels* of anonymity are required, each with a different set of potentially identifiable data elements. To achieve this, natural language processing is used to tag data according to contextually relevant factors and certain tagged information redacted for particular researchers according to an anonymisation schema.

By providing different levels of anonymity according to individual researchers' needs, we can protect the confidentiality of patients on an individual basis and potentially streamline the process of gaining ethics committee approval.

V. CONSENT AND NOTICE

Consent and notice are important principles for privacy preservation [12]. *Consent* requires that data owners have some meaningful choice over how their data is shared and used. *Notice* requires a mechanism with which users can extract information about how their data has been used and about any relevant changes in privacy policies. Within this application, consent and notice equate to the following:

A. Consent

Patients must be able to opt in or out of participation in medical research depending on the details of local legislation. Patients should be able to choose what types of medical research their records can be used for (for example, some patients might wish to prevent their data from being used in research that involves experimentation on animals).

Patients must be able to choose a maximum level of information that they are prepared to share for the purposes of certain types of medical research. For example, they might not wish to share family history in trials that involve mental health medicine.

B. *Notice*

It must be possible for patients to determine when their records have been used, by whom and for what purpose. They must be able to find out detailed information about the projects their data has been used in.

Patients must be informed if there are changes to the policies governing their requirements and should be informed about how the governance has changed. They should be able to modify their privacy requirements accordingly. For example, if privacy policies or anonymisation schemas change, patients should be informed so that they can modify their consent.

C. *Ethics*

One of the purposes of the GATEWay project was to work toward streamlining the process of ethics committee approval for research projects. This might be achieved by combining a secure environment for conducting research with levels of anonymity controlled by anonymity schema. The idea is that anonymity schemas represent standard uses of data and once a project employing a schema has been approved, it should be easier to gain approval for other projects using the same schema. Conversely, individuals must be able to determine how their data has been shared and used.

VI. NATURAL LANGUAGE PROCESSING AND ANONYMISATION SCHEMA

In the medical domain, data sets contain protected health information (PHI) that can identify individuals. Anonymisation is the removal of PHI, which is a 2-step process:

1. Identification of PHI and its classification with PHI categories.
2. Anonymisation of identified PHIs by replacing them with their respective PHI categories.

The data used in this research contain 2534 PHIs which are classified into following PHI categories: Patient Name; Doctor Name; Place Name; Other Name; and Risky Behaviour.

For example, consider the following excerpt from a medical record:

```
Mrs Ward has health risks due to
excessive alcohol consumption. Her
husband, Derek Ward, may be at risk too.
```

In this example, “Derek Ward” should be identified as PHI and classified as 'Other Name'. This is achieved by specifying rules using the NLP software GATE (http://gate.ac.uk/). The rule for identifying “husband, Derek Ward” in this case is as follows:

```
Rule:OtherNames
(
     (
     {Lookup.majorType == other}
       // Dictionary of relations, roles,
occupations
     {Token.kind==punctuation}
     (SPACE)
     (NAME)                         //'NAME'  is
Macro rule for identifying proper names
     (SPACE)
     (NAME)
     )
)
:label
-->
:label.OtherName={Rule=OtherNames}
```

This record will be tagged as XML as follows;

```
<patientname>Mrs Ward</patientname> has
health risks due to <risky
behaviour>excessive alcohol
consumption</riskybehaviour>. Her
<othername>husband, Derek
Ward</othername>, may be at risk too.
```

This tagging reveals important semantic data that might otherwise remain hidden within free text in a health record.

After the identification and classification of PHIs, the anonymisation is completed by replacing identified PHIs with their respective PHIcategories (XML tags) according to an anonymisation schema such as the following:

```
{-patientname
+riskybehaviour
-othername}
```

Resulting in a record with the data contained within the tags redacted, indicating that an anonymous patient has health risks due to excessive alcohol consumption and that another - unknown - person might also be at risk:

```
Patient has health risks due to excessive
alcohol consumption. Other Person may be
at risk too.
```

Other schemas will represent different research objectives and different risks.

The NLP tagging and redaction based on anonymisation schema help ensure that as much semantic information as required can be easily recovered, even from free text fields, but that information a researcher is not entitled to will be removed.

VII. The Prototype

The GATEway project developed a prototype of the VRE and the data anonymisation process.

A. *The data anonymisation pipeline*

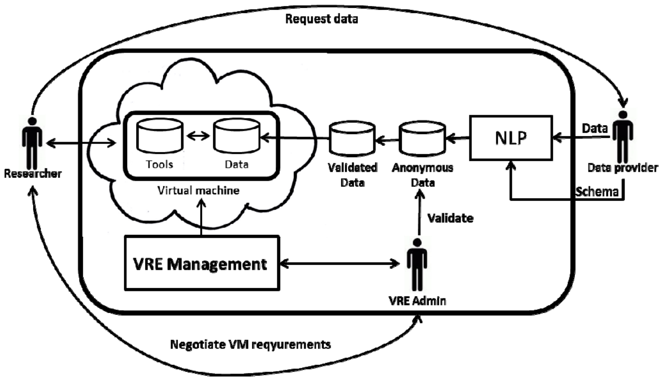

Figure 1: the pipleline from raw data to anonymised data in a researcher's VRE

1. A researcher requests data from a data provider, specifying needs and the use to which the data will be put.

2. The data provider prepares the data set and an anonymisation schema and uploads the data to the VRE.

3. The NLP module tags the data and uses the anonymisation schema to remove any tagged data the researcher is not entitled to see.

4. A VRE administrator performs a risk assessment on the anonymised data, examining it for personal information that has been missed by NLP and identifying possible false positives.

5. The VRE administrator negotiates with the researcher over VM requirements including what operating systems, applications and resources are needed.

6. The VRE administrator creates a VM for that researcher (or modifies an existing one) and uploads the validated dataset into it.

7. The researcher customises her VM if necessary by uploading tools and additional data, then can begin to conduct research.

The data provider is responsible for creating datasets and anonymisation schemas for researchers. VRE Administrators are responsible for the anonymisation of datasets, the creation and management of VMs and the assignment of validated anonymous datasets to particular VMs.

We have developed a module for the GATE

VIII. VRE Architecture

The VRE is built according to the following (high-level) architecture:

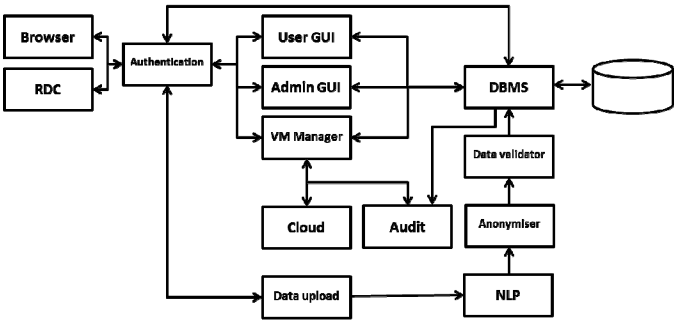

Figure 2: VRE architecture

A. *Virtual machines*

Each researchers or group of researchers is assigned a VM which contains the data and tools they need to conduct their research. The resources available to a VM are dynamically customisable and reclocatable, hosted on a private cloud.

B. *User interface*

Researchers connect to their VMs via either a browser or Remote Desktop connection. The browser interface is restricted but customisable: researchers can run common queries on data, build their own queries, import tools into the environment and dynamically change resource allocations. The Remote Desktop interface gives full access to the VM (according to access control) for more complex tasks.

C. *Data upload*

The data pipeline is discussed in the previous section. The Data Upload module is used by data providers to submit data from their clinical information systems. The NLP module takes raw datasets and produces semantically tagged datasets.

The Anonymiser takes tagged datasets and removes information according to the associated anonymisation schema. The Data Validation module enable enables risk assessment of anonymous datasets and manages their upload to the appropriate VMS.

D. *Audit*

The audit module records the creation and assignment of users, VMs, anonymisation schema and datasets to provide a record of what data is accessible under what circumstances to which researchers.

IX. Conclusions and Future Work

The case study describes some novel requirements which arise from the increased ubiquity and availability of data associated with changing attitudes to data collection, management and use; increased ability to index, process and visualise data; and increasing public awareness of the need to protect one's own privacy (and particularly in this case, confidentiality).

We have built a prototype to demonstrate proof of concept of many of these ideas. However, it does not address all the requirements of privacy in this environment. For example, it does not fully implement the issues of consent and notice. Consent might be implemented through an additional layer of privacy policy and related protocols. In this scenario, patient privacy policies could be matched with anonymity schema to ensure that only patients who opt in to (or do not opt-out of) the conditions of a schema are included in particular datasets. This is the subject of on-going research. We anticipate that audit be expanded to include provenance generation to aid replication and validation of experiments as well as the policing privacy.

Notice could be implemented using notification or syndication generated by events in the audit module in conjunction with appropriate protocols and anonymity schema to inform patients about the use of their data or changes in policies. This is also a subject of on-going research.

We are currently working to expand the prototype beyond proof of concept to a production-ready environment in conjunction with Leeds Information Systems and Services department and our industrial partners. This will involve migrating the service from the private cloud to the White Rose Grid, which is a large-scale computing resource shared by the universities of Leeds, York and Sheffield. We also plan to extend the service to data providers other than TPP and to applications other than health records research.

X. References

[1] Hoerbst, A., and E. Ammenwerth. "Electronic Health Records." *Methods Inf Med* 49 (2010): 320-336.

[2] Iakovidis, Ilias. "Towards personal health record: current situation, obstacles and trends in implementation of electronic healthcare record in Europe." *International journal of medical informatics* 52.1-3 (1998): 105.

[3] Terry, Nicholas P., and Leslie P. Francis. "Ensuring the privacy and confidentiality of electronic health records." *U. Ill. L. Rev.* (2007): 681.

[4] Meystre, Stéphane M., et al. "Extracting information from textual documents in the electronic health record: a review of recent research." *Yearb Med Inform* 35 (2008): 128-44.

[5] Demuynck, Liesje, and Bart De Decker. "Privacy-preserving electronic health records." *Communications and Multimedia Security*. Springer Berlin/Heidelberg, 2005.

[6] Kitteringham, Glenn. "Lost laptops= lost data: Measuring costs, managing threats."*Crisp report, ASIS International Foundation*. 2008.

[7] Kaufman, Lori M. "Data security in the world of cloud computing." *Security & Privacy, IEEE* 7.4 (2009): 61-64.

[8] Szarvas, György, Richárd Farkas, and Róbert Busa-Fekete. "State-of-the-art anonymization of medical records using an iterative machine learning framework."*Journal of the American Medical Informatics Association* 14.5 (2007): 574-580.

[9] Schoenberg, Roy, and Charles Safran. "Internet based repository of medical records that retains patient confidentiality." *Bmj* 321.7270 (2000): 1199-1203.

[10] Anderson, Ross. "NHS-wide networking and patient confidentiality." *BMJ: British Medical Journal* 311.6996 (1995): 5.

[11] Simon, Gregory E., et al. "Large medical databases, population-based research, and patient confidentiality." *American Journal of Psychiatry* 157.11 (2000): 1731-1737.

[12] Langheinrich, Marc. "A privacy awareness system for ubiquitous computing environments." *UbiComp 2002: Ubiquitous Computing* (2002): 315-320.