Security and confidentiality approach for the Clinical E-Science Framework (CLEF)

D. Kalra[1], P.Singleton[1,4], D. Ingram [1,] J. Milan[2], J.MacKay[3], D. Detmer [4], A. Rector[5]

[1] Centre for Health Informatics and Multiprofessional Education (CHIME) University College London

[2] Royal Marsden NHS Trust

[3] The Genetics Unit, Institute of Child Health, University College London

[4] Judge Institute, University of Cambridge

[5] Department of Computer Science, University of Manchester

Correspondence to:

Dr Dipak Kalra
Centre for Health Informatics and Multiprofessional Education (CHIME)
University College London
Holborn Union Building, Highgate Hill, London N19 5LW
email: d.kalra@chime.ucl.ac.uk
tel: +44 20 7288 3362
fax:+44 20 7288 3322

Summary

Objectives: CLEF is an MRC sponsored project in the E-Science programme that aims to establish methodologies and a technical infrastructure for the next generation of integrated clinical and bioscience research.

Methods: The heart of the CLEF approach to this challenge is to design and develop a pseudonymised repository of histories of cancer patients that can be accessed by researchers. Robust mechanisms and policies have been developed to ensure that patient privacy and confidentiality are preserved while delivering a repository of such medically rich information for the purposes of scientific research.

Results: This paper summarises the overall approach adopted by CLEF to meet data protection requirements, including the data flows, pseudonymisation measures and additional monitoring policies that are currently being developed.

Conclusion: Once evaluated, it is hoped that the CLEF approach can serve as a model for other distributed electronic health record repositories to be accessed for research.

Keywords

Electronic health records, data repository, pseudonymisation, grid computing, access control

Background: The CLEF Project

CLEF is a Medical Research Council sponsored project in the E-Science programme. It aims to establish methodologies and a technical infrastructure for the next generation of integrated clinical and bioscience research:

1. establishing a secure generic repository of structured and narrative patient records derived from operational clinical systems, capable of distributed interrogation via an intuitive query workbench;
2. developing novel language technology and software tools to analyse clinical narratives:
   a) to enable key clinical information to be extracted and encoded; and
   b) to assist in removing residual potentially identifying information;
3. establishing best practice in the pseudonymisation of clinical records, and the development of systematic methods and tools to do this on a scalable basis.

CLEF will provide an end-to-end solution for collecting and managing longitudinal data about cancer patients for both healthcare and biomedical research. It is designed to address the key problem of linking genomic information to the clinical course of patients' illnesses.

Objectives of the security and confidentiality policy

The key ethico-legal goal of CLEF is to provide mechanisms and policies to ensure that patient privacy and confidentiality are preserved while delivering a repository of medically rich information for the purposes of scientific research. This requires both policy and organisational safeguards and a multilevel technical framework.

There is a well-recognised need to establish a scalable methodology for deriving large numbers of longitudinal pseudonymised health records (de-identified, and with a protected link to the original patient's identity), in order to conduct clinical and bio-scientific research and recruitment for national clinical trials in ways not possible using current resources, *e.g.* cancer registries. To do so requires a managed and monitored (i.e. audited) framework for maintaining privacy and confidentiality, concealing patients' identities, managing authentication and auditing access so that risk to privacy is minimised. One key strand of the CLEF project, therefore, focuses on the development of rigorous generic methods to solve this problem using cancer care as an exemplar domain.

Requirements

There are strong legal protections on personal patient information, from the Data Protection Act 1998 [1] (following on from the European Directive 95/46/EC [2]), the Human Rights Act 1998 [3], as well as the common law of confidentiality. These generally require either the consent of the data subject or the pseudonymisation of the information. If the data were not pseudonymised, patient consent would be required for CLEF to acquire the data into its repository, and for each new kind of research access to the data.

Most research requirements do not need identifiable information, but they do require longitudinal records that link the various episodes for each patient, preferably derived from real EHR data sources, to enable them to observe patients' histories as they evolve. There is also sometimes a requirement to be able to re-identify specific patients in special circumstances, *e.g.* to warn patients of risks uncovered by research or in order to recruit patients for clinical trials.

Technical Approach

The Electronic Health Record (EHR) at the Royal Marsden Hospital (RMH) is one of the main providers of patient records to the project. An approach has been developed by which real patient records (structured data sets and narrative letters and reports) can be suitably pseudonymised for removal from the RMH and included within the CLEF Repository. The process provides multiple layers for the protection of patient confidentiality and privacy:

1. *pseudonymisation* – the removal of patient, geographical and organisational identifiers;
2. *depersonalisation* – applying language extraction and generation to remove potentially identifying information;
3. *security* – policies and technical measures for the supervision and maintenance of the pseudonymous EHR repository as if it contained identified patient records, in conformance with NHS and international standards including privacy enhancing techniques and methods to reduce the risk of re-identification through queries;
4. *oversight* –policies for controlling access to CLEF repository and handling requests to link researchers back to real patients;
5. *monitoring* – organisational and technical measures to identify potential threats and intrusions.

The first four aspects of the approach are discussed in more detail below. The fifth is being explored within the project. Figure 1 shows the flow of information showing the points of control for privacy and confidentiality.

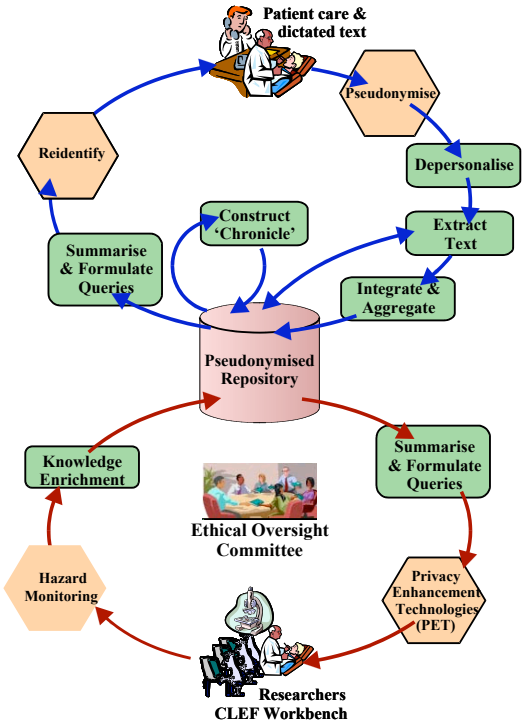

Figure 1: High-level view of CLEF information flow cycle with points of control for privacy indicated.

(1) Pseudonymisation

The CLEF pseudonymised repository is being established at University College London (UCL). UCL has been active in several EU projects over the past decade to investigate and specify the clinical and ethico-legal requirements, information models

and middleware services that are needed to underpin comprehensive EHRs [4,5]. UCL has designed and built a federated health record server based on these models, which has been evaluated in the Cardiovascular Medicine Department at the Whittington Hospital in London [6,7] and in the South West Devon ERDIP project.

The overall process is implemented and split amongst the different partners as shown in Figure 2. During the initial stages of the project until the methodology is proven, records will be restricted to those of deceased patients to minimise risk of harm to existing patients.

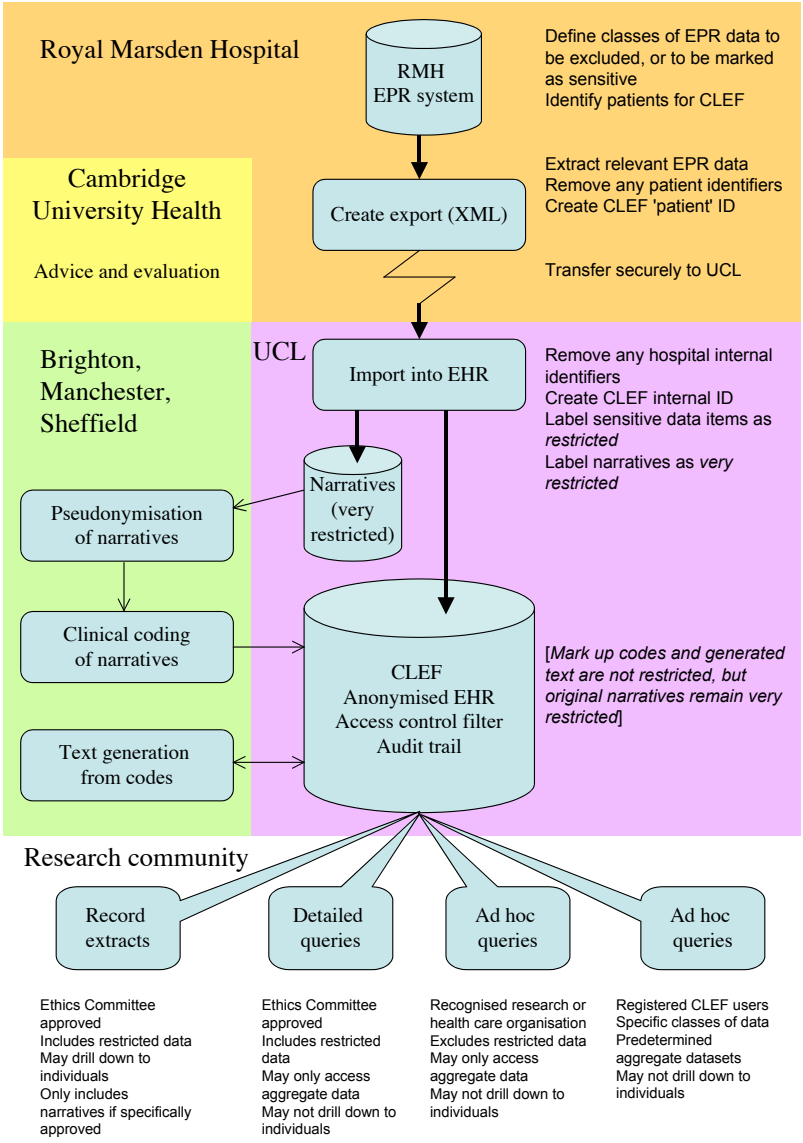

Figure 2: Data flow within current phase of CLEF project to generate the pseudonymised repository of EHRs

Records of patients at the RMH will be extracted from the main computer system and exported to the CLEF repository at UCL. Examples of the kinds of structured data to be exported from the RMH include the principal cancer clinical diagnoses, tissue (histology and cytology) diagnoses, laboratory investigation results, radiology

investigation results, surgical procedure codes, chemotherapy and radiotherapy treatment protocols, administered drugs, and death certificate information.

Narrative information sources will include: discharge summaries, clinic letters, histopathology reports, radiology reports, and surgical (operative procedure) notes.

These data will be subjected to a combination of computerised and manual de-identification on site before being sent via a secure communication to UCL.

1. At the Royal Marsden NHS Trust:
   a) any patient records flagged as not to be included in research (e.g. at the request of the patient) will be excluded;
   b) extraction will be limited to the particular data elements of the RMH EHR that are needed to support the anticipated research queries;
   c) key identifying fields, such as name, address, full postcode, NHS Number, will not be extracted;
   d) demographic and social history information that may be needed to support realistic research queries (such as age, postal district, occupation) will be marked as "sensitive";
   e) a secure "clef entry identifier" will replace the Royal Marsden Hospital patient ID field, so that there is no reference whereby a researcher could link back to the primary medical record and the patient's identity;
   f) all occurrences of the patient's name in text fields, for example in letters and reports, will be removed.
2. At UCL, the incoming data will be re-mapped into the CLEF EHR data-schema and the "clef-entry identifier" replaced by the internal clef-identifier, providing a second barrier between the identifiers in the repository and the original identifiers at the originating hospital.

Additional policies and procedures, which are still being defined, will be put in place:

1. at the time of querying: for monitoring and controlling queries;
2. for returning information to the Royal Marsden, ensuring that only the Royal Marsden can re-identify patients and only in appropriate circumstances; such data will almost certainly be limited to data originally provided by the RMH and exclude any data linked to each such patient from other provider sites;
3. an overall supervisory and regulatory framework, through responsibility to an oversight CLEF Ethics Board that will be established towards the end of the CLEF project, before any data is made available to external research groups.

(2) Depersonalisation - Extraction of data elements from narratives

In the real world, much medical information is transferred through exchange of letters between clinicians. Hence much of the data that is available in electronic form is in free-text format.

The text fields, particularly narratives, will be parsed by routines developed at the University of Sheffield to extract only clinically structured data – in doing so any extraneous socially significant information would be left behind. The original narratives will then be securely separated from the rest of the CLEF Repository.

The processing of the free text data to identify clinically relevant information and to extract this into a structured and codified format will greatly increase the value of such data to researchers, even if some fine detail is lost. This will be done through semantic analysis and extensive use of clinical vocabularies and ontologies. One positive effect of this data extraction is that by focusing solely on medical facts much of the social context is omitted.

(3) Security policy and technical measures

The information to be held in the CLEF repository might still be considered 'sensitive personal data' under the definition of the Data Protection Act 1998, so the general approach taken by CLEF will be to treat these records as if they still retain some (albeit hypothetical) risk of re-identification. There is always a chance that some unusual or unique characteristics of an individual clinical journey portrayed within an EHR might suggest which real person it concerned.

A security policy has been proposed for the CLEF project teams, covering storage, access and data flows, that would meet many of the requirements of data protection that would pertain to the control of access to real and identifiable patient records.

This challenge will in the long term be greater when records from multiple provider sites are combined in the CLEF repository, possibly through a trusted third party with access to the original patient NHS numbers and the responsibility for managing LCEF repository person identifiers.

The approach for safeguarding research query access to the final CLEF repository includes:

1. limiting the majority of research queries to the return of aggregate data (e.g. frequency tables) and not the findings in individual patients;
2. limiting access to the individual pseudonymised records to clinical research projects that have themselves obtained ethical approval for the queries they intend to run.

The main risk to patients would be through a mechanism of inferential data-mining (whereby known information about a person's medical history are used to identify a unique set of records which might then reveal more about that individual). In order to limit such risks the following restrictions are placed on access:

1. only individuals registered with Research Ethics Committee approved projects may have access to the system and this will be time limited to the project;
2. projects and researchers will only be allowed access to specific fields or ranges of records relevant to their project;
3. generally, only aggregate data will be provided unless ethical approval permits access to individual record-level data
4. there will be checks on query criteria to identify possible inferential attacks, either through overlapping queries or highly specific queries;
5. where individual record data is to be provided with a facility for longitudinal linking, a project-specific re-mapping of the unique identifier will be used so that data-sets provided to different projects cannot be re-linked.

There is a growing body of literature investigating the risks of person re-identification through data mining and probabilistic techniques [8], and a similarly expanding set of algorithmic techniques proposed for profiling and monitoring serial queries and result sets to detect attempts to triangulate towards unique person characteristics [9,10,11]. This and other work in the field is being reviewed within the project to determine the kinds of audit trails that need to be built in and constraints that ought to apply to the specification of queries by the CLEF workbench tools.

A series of requirements have been drafted that will apply in general to all research communities accessing the final live CLEF services, via GRID networks.

1. Reliable identification and traceability of any GRID users accessing CLEF
2. Assignment of GRID access control levels

3. Authentication of users during sessions to ensure that sessions cannot be hi-jacked
4. Security of data transmissions
5. Non-repudiation of query requests
6. Local decryption of data packages
7. Local screen security, both for user entry of passwords, and to ensure that potentially sensitive data is not displayed without user presence and knowledge.

The technical approach to certain aspects of GRID security, such as authorisation and privilege management, is still being refined, drawing on the use of X509 certificates (as demonstrated in the PERMIS project [12]) and new standards such as Security Assertions Markup Language [13]. The CLEF approach will be to utilise these infrastructure components through collaboration with other UK e-Science projects. However, work remains to be done to define appropriate user roles, privileges and other relevant metadata to reflect CLEF's authorisation policies.

A separate concern is the process of protecting the population the CLEF repository from diverse and distributed healthcare provider sites. These data will be partly but not wholly de-identified, since some parts of the de-identification and depersonalisation processes will only be possible once the data has been acquired by CLEF. It is as yet unclear if the levels of security proposed for the grid will satisfy requirements within the NHS for the communication of partly de-identified health data.

(4) Oversight – policies for access

An Ethical Oversight Board, to be established, will approve the kinds of organisations, teams and purposes for which the CLEF repository may be queried, and defining the appropriate security measures to be taken e.g. for authentication, authorisation and encryption.

The final ethical and security related results of CLEF will be:

- a validated approach, accepted by research ethics committees, patient representative groups, and other stakeholder groups (professional bodies and health services);
- an Ethical Oversight Committee, and exemplar policies and procedures, on issues such as depersonalisation and access control;
- open source tools: mechanisms to support security and tools for the active monitoring of use.

Conclusions

CLEF is exploring a pseudonymisation approach to parallel the 'consent' approach of the BioBank initiative. CLEF will identify processes and procedures that are both technically feasible as well as politically and socially acceptable to permit efficient access to health records to further medical research.

CLEF may give rise to an ongoing research database if there is continuing funding. Equally, the policies and methods developed may serve to inform other projects nationally or internationally.

An important objective of the project methodology is to establish best practice in pseudonymisation and in the security policies that should pertain to such a repository. A formal evaluation of the proposed approach will be carried out and published.

Acronyms

CLEF: Clinical E-Science Framework
EHR: Electronic Health Record
ERDIP: Electronic Record Demonstration and Implementation Programme
MRC: Medical Research Council (UK)
NHS: National Health Service
UCL: University College London
RMH: Royal Marsden Hospital

Acknowledgements

The CLEF project is partially funded by the MRC under the E-Science Initiative, grant number G0100852. Special thanks are due to its clinical collaborators at the Royal Marsden and Royal Free hospitals, to colleagues at the National Cancer Research Institute (NCRI) and NTRAC and to its industrial collaborators – see www.clinical-escience.org for further details.

References

1. Data Protection Act 1998, The Stationery Office Limited London 1998, www.hmso.gov.uk/acts/acts1998/19980029.htm
2. EU Directive 95/46, http://europa.eu.int/comm/internal_market/privacy/law_en.htm
3. Human Rights Act 1998, The Stationery Office Limited London 1998, www.hmso.gov.uk/acts/acts1998/19980042.htm
4. Ingram D. The Good European Health Record Project in: Laires, Laderia Christensen, Eds. Health in the New Communications Age. IOS Press: Amsterdam; 1995; pp. 66-74
5. Grimson J., Grimson W., Berry D., Stephens G., Felton E., Kalra D., Toussaint P., and Weier O.W. A CORBA-based integration of distributed electronic healthcare records using the synapses approach. IEEE Trans Inf Technol Biomed. Sep 1998; 2(3):124-38
6. Kalra D, Austin A, O'Connor A, Patterson D, Lloyd D, Ingram D. Design and Implementation of a Federated Health Record Server. Toward an Electronic Health Record Europe 2001, Paper 001: 1-13. Medical Records Institute for the Centre for Advancement of Electronic Records Ltd.
7. Kalra D. Clinical foundations and information architecture for the implementation of a federated health record service. PhD Thesis. Univ. London. 2002.
8. L. Sweeney. k-anonymity: a model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10 (5), 2002; 557-570
9. De Moor G., Claerhout B., De Meyer F. Privacy Enhancing Techniques: the Key to Secure Communication and Management of Clinical and Genomic Data. Methods Inf Med. 2003; 42(2): 148-153.
10. Ferris TA., Garrison GM, M, Lowe HJ. Proposed Key Escrow System for Secure Patient Information Disclosure in Biomedical Research Databases. Procs AMIA 2002 Annual Symposium 245-249
11. Murphy S, Chueh H. A Security Architecture for Query Tools used to Access Large Biomedical Databases. Procs AMIA 2002 Annual Symposium 552-556

12. The PERMIS Project. [http://www.permis.org/] Last accessed February 2004
13. Security Assertions Markup Language. OASIS Security Services TC. [http://www.oasis-open.org/committees/tc_home.php?wg_abbrev=security] Last accessed February 2004