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
Holborn Union Building, Highgate Hill, London N19 5LW

[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

d.kalra@chime.ucl.ac.uk
www.clinical-escience.org

Abstract

CLEF is an MRC sponsored project in the EScience programme that aims to establish policies and infrastructure for the next generation of integrated clinical and bioscience research. One of the major goals of the project is to provide a pseudonymised repository of histories of cancer patients that can be accessed by researchers. Robust mechanisms and policies are needed to ensure that patient privacy and confidentiality are preserved while delivering a repository of such medically rich information for the purposes of scientific research. This paper summarises the overall approach adopted by CLEF to meet data protection requirements, including the data flows and pseudonymisation mechanisms that are currently being developed. Intended constraints and monitoring policies that will apply to research interrogation of the repository are also outlined. 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. .

Background: The CLEF Project

CLEF is an MRC sponsored project in the E-Science programme. It aims to establish policies and infrastructure for the next generation of integrated clinical and bioscience research. The project's core aims are:

1. to develop novel technology and software tools to analyse patient records. Language tools have been identified as a key technology in two areas:
   a) to enable information to be extracted from the text in clinical narratives; and
   b) to assist in removing residual potentially identifying information from clinical narratives;
2. to establish best practice in the pseudonymisation of clinical records, and the development of systematic methods and tools to do this on a scalable basis.

CLEF seeks to 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 policy/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, identifiable only by the originating health authority), in order to conduct the next generation of clinical and bio-scientific research and to recruit for national clinical trials in ways not possible using current resources, *e.g.* cancer registries. To do so requires a managed and monitored framework for maintaining privacy and confidentiality. It must

both conceal patients' identities and manage and monitor authentication and 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 (following on from the European Directive 95/46/EC), the Human Rights Act 1998, as well as the common law of confidentiality. These generally require either the consent of the data subject or the pseudonymisation of the information.

Most research requirements do not need identifiable information. What they require are longitudinal records that reliably link the various episodes for a single patient into a coherent "chronicle". A dynamically pseudonymised record that offers the ability to use real Electronic Health Records, and observe patients' histories as they evolve is highly attractive.

However, there is also 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. Some Research Ethics Committees (RECs) may even place such a requirement on research projects so that patients can directly benefit from the research where appropriate. Such re-identification must only be possible via the originating health care organisations.

Technical Approach

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

1. *pseudonymisation* – the removal of patient, geographical and organisational identifiers at source;
2. *depersonalisation* – methods of access via language extraction and generation that conceal or remove potentially identifying information;
3. *security* – policies and technical measures for the supervision and maintenance of the pseudonymous Electronic Health Record repository as if it contained identified patient records, in conformance with NHS and international standards including privacy enhancing technologies to reduce the risk of re-identification through queries;
4. *oversight* – specific 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 a current area of exploration within the project.

The high level view of the flow of information showing the points of control for privacy and confidentiality is given in Figure 1. The specific implementation of this scheme within the current state of the CLEF project is shown in Figure 2.

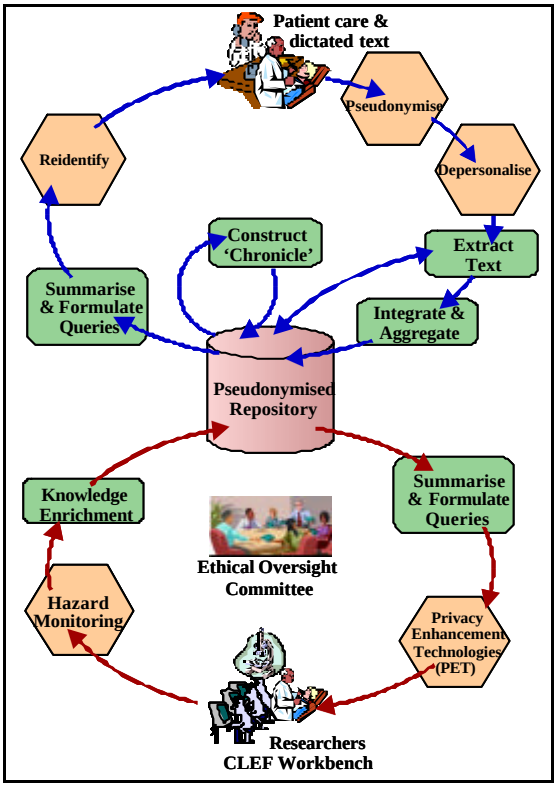

Figure 1: High level view of CLEF information flow cycle with points of control for privacy indicated.

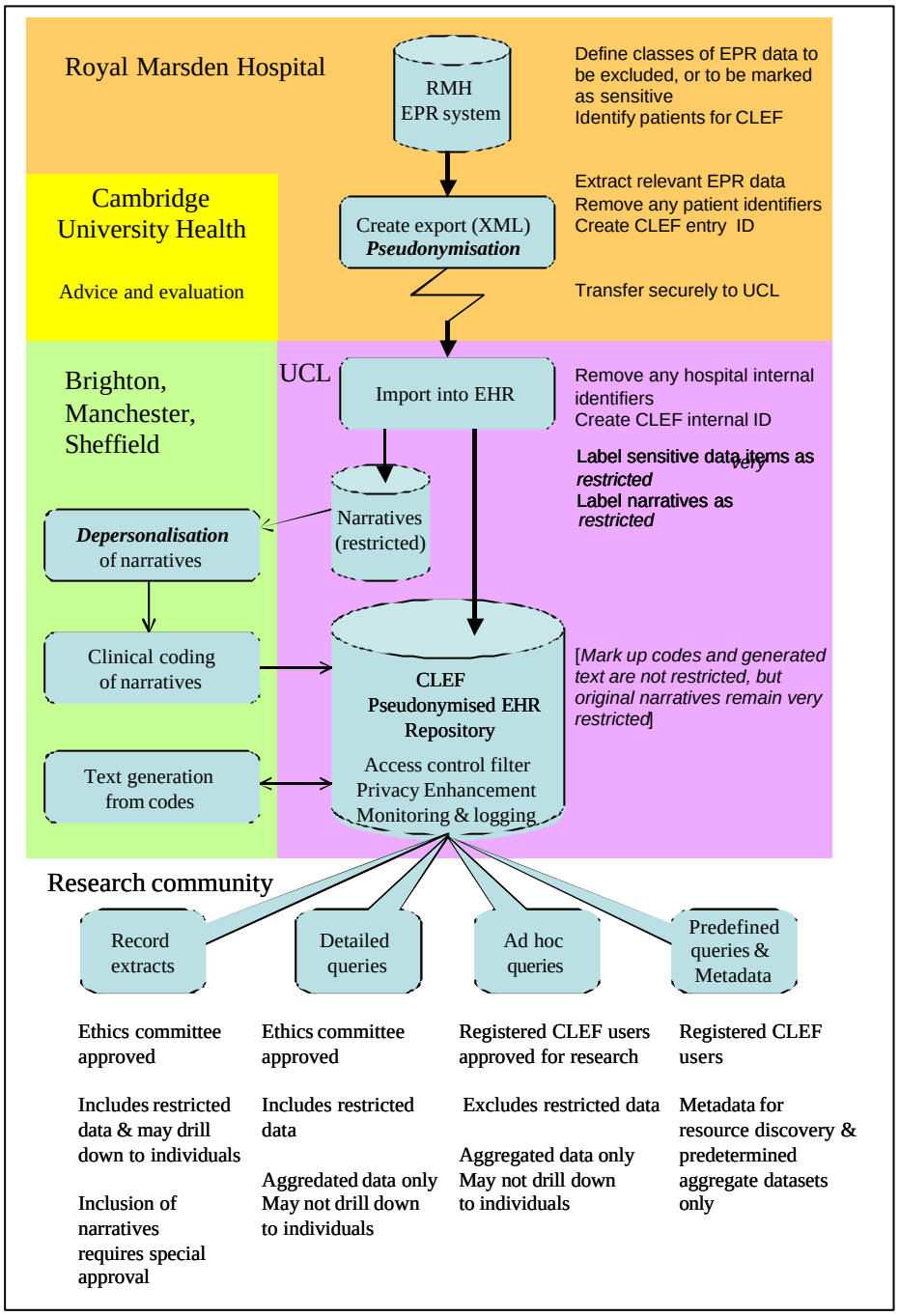

Figure 2: Data flow within the current phase of CLEF project to generate the pseudonymised repository of EHRs.

Pseudonymisation (1)

The CLEF pseudonymised repository of electronic health records (EHRs) will be established at University College London (UCL) using the results of European research into the design of Electronic Health Record systems, meeting established clinical and ethico-legal requirements. UCL has been active in several EU projects over the past decade to investigate and specify the requirements, information models and middleware services that are needed to underpin comprehensive multi-professional electronic health records [e.g. Ingram 1995, Grimson et al 1998]. UCL has

designed and built a federated health record server based on these models, which has been evaluated in the Department of Cardiovascular Medicine at the Whittington Hospital in north London [Kalra et al 2001, Kalra 2002] and in the South West Devon ERDIP project[1].

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.

The overall process is implemented and split amongst the different partners as shown in Figure 2.

Records of patients at the ROYAL MARSDEN HOSPITAL will be extracted from the main computer system and subjected to a combination of computerised and manual de-identification on site before being sent via a secure communication to UCL. These steps will include:

1. limiting extraction to the particular structural data elements of the Royal Marsden Hospital Electronic Health Record that are needed to support the anticipated research queries for the CLEF Electronic Health Record;
2. the exclusion of principal patient identifiers such as name, address, next of kin and GP information;
3. marking as 'sensitive' any demographic and "social history" information that may be needed to support realistic research queries (such as age, postal district, occupation).

The various confidentiality-enhancing measures are:

1. At the Royal Marsden NHS Trust:
   a) any patient records flagged as not to be included in research (at the request of the patient and/or consultant) will be excluded from data extraction;
   b) key identifying fields, such as name, address, full postcode, NHS Number, will not be extracted
   c) 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;
   d) all occurrences of the patient's name in text fields will be removed.
2. Prior to transfer to the CLEF Electronic Health Record system at UCL, the extracted data-set will be further depersonalised by running a number of procedures to remove other potential identifiers. These procedures will be developed through the various phases of the project (see below) and applied particularly to narratives which are considered to have the highest risk of containing identifiable information.
3. 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.

Identifiable patient information will not be released by the Royal Marsden Hospital under any circumstances. The Royal Marsden Hospital paper record systems are not accessible by this project.

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;
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.

Depersonalisation - Extraction of data elements from narratives (2)

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 removed.

In the real world, much medical information is transferred through exchange of letters between clinicians, through default of proper work-flow systems to support clinical care. Hence much of the data that is available is perforce in free-text (or quasi-free-text) format. On the plus side, such correspondence usually references only key relevant information, filtering out much of the chaff generated by individual laboratory

[1] Please see http://www.swdhis.nhs.uk/erdip/public/index.shtml

reports and tests, which may not actually be pertinent to the condition.

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 recall and precision is lost (following the principle that half a loaf is better than no bread, and that inaccessible data trapped in free text format is virtually useless). This will be done through semantic analysis and extensive use of clinical vocabularies and ontologies.

One positive side-effect of this data extraction is that by focusing solely on medical facts much of the social context is omitted. While social context may be critical in certain areas, *e.g.* mental health, removing it reduces the likelihood that extraneous information might identify an individual, *e.g.* ' … <the patient> attended the <clinic> accompanied by her partner, a well known politician. …'. The text extraction process will aim only to record that the patient attended a clinic on a certain day (and even the exact date might be blurred to limit the risk of identification still further.)

Security policy and technical measures (3)

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. Whatever the precautions, there is always some chance that some unusual or unique characteristics of an individual clinical journey in an EHR might make the patient recognisable to someone with sufficient knowledge from other sources.

A draft security policy has been proposed for the CLEF implementation that would meet many of the requirements of data protection, Caldicott-Guardian responsibilities and other published requirements that would pertain to the control of access to real and identifiable patient records. This includes local security policies for each CLEF partner site that needs to access or process data from the repository. The approach for 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 REC 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 [e.g. Sweeny L 2002], 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 [e.g. Ferris et al 2002, Murphy and Chueh 2002]. 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.

Oversight – policies for access (4)

A series of requirements have been drafted that will apply to 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 security levels as access to medical data sets may impose restrictions (e.g. not undergraduate students)
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

Four 'Use Cases' are envisaged for research access to repository data, which will be managed via the CLEF workbench and by attribute certificate services within the EHR repository middleware.

1. *Open, Aggregated data – all users*

   CLEF may make available to GRID users generally aggregated data-sets which are fully pseudonymised and approved for release to *bona fide* researchers.
   CLEF would expect reliable **identification of users** making requests for such data, both to develop performance statistics to justify ongoing funding, and to be able to vet for any unusual activity that may indicate security or confidentiality breaches.
   CLEF may require some measure of **secure links** to such users to be able to meet assurances to Ethics Committees that data is only being provided to *bona fide* researchers, e.g. SSL links as standard. This may also have implications for general **access controls** within GRID, and the passing of a general GRID access level to CLEF to permit access for even pseudonymised medical data.

2. *Aggregated queries on individual records – CLEF registered users only*

   CLEF will allow most CLEF registered users to run queries on the pseudonymised data-sets to extract aggregated statistics (possibly with cell-size restrictions to limit identifiability). Privacy enhancement and monitoring techniques will be used to blur results so as to minimise the risk that queries, singly or together, might allow individuals to be identified.

   Access to such aggregated queries would require a high-level of **identification and authentication** of the individual making access, including **session control**. This would include **non-repudiation of query requests** and acknowledgement of data-set delivery (to manage any re-requests of data which might be spoofed).

   Secure links would definitely be required, probably at least SSL 128-bit.

3. *Access to disaggregated pseudonymised record sets – special approval only*

   Specific projects (and hence specific users within that project group, but possibly only the Principal Investigator) would be permitted access to the individual pseudonymised data-sets (though nearly always with restrictions on the table columns that could be accessed; almost never access to the entire record).

   This would require an even greater emphasis on **identification and authentication**, as well as security of the data provided through a session.

   The data provided may need to be encrypted to a higher level than SSL 128-bit, and hence may require some form of local processing to de-crypt the data, and secure its delivery at the user workbench.

4. Downloading of subsets of the repository – *special approval only*

   Some projects may be allowed to download sub-sets of the CLEF database (subject to approval by a Research Oversight Committee) which will have a specific encryption of identifying fields for each specific project (to prevent linking of separate approved extracts to re-create the CLEF database in whole or in part).

   The required mechanisms will need to be explored more fully, and may be covered within the requirements 1-3 above.

A special requirement is that it should be possible to link back to the original patient id for ethically approved reasons, for example to contact high risk to patients or as part subsidiary research project, or to identify patients for recruitment to trials. This process for direct access to patient records is proposed to be as follows:

1. researchers will be required to submit a request to the CLEF Ethics Board administration. The request will be assessed for ethical appropriateness, including a check against the original Research Ethics Committee approval);
2. the CLEF administration will then submit the request to the repository holders, in this case UCL, to identify which CLEF repository records are required. Such

access requires unusual authorisation and will be strictly logged and monitored

3. the repository holders (UCL) will then have to make a similar request to originating hospital, Royal Marsden, subject to the same vetting processes, and accompanied by the original entry IDs for the patients involved. If the originating hospital, the Royal Marsden, agrees, only then can it trace the entry IDs back to the original Royal Marsden Hospital ID that allows the hospital to reidentify the patient. The originating hospital can then contact the patient either directly or via their general practitioner, in order to gain consent to further research access to their full medical records, to participate in a trial, or to be recalled if at risk.

This three-stage process across three separate organisations and identifiers should ensure that identification is only possible when appropriate and duly authorised.

Progress to date

One-way key encryption is now being used to create a CLEF "patient" identifier that is distinct from any health service issued numbers, permitting longitudinal growth of the CLEF record through a non-reciprocal link from the RMH to CLEF. The mapping between these sets of identifiers is held securely at the Royal Marsden Hospital.

Parts of the clinical records are now being extracted from the Royal Marsden Hospital Electronic Health Record system, initially only for deceased patients, for transfer to the CLEF Electronic Health Record server at UCL, beginning with the narrative case notes and correspondence parts of the records. In parallel, the research teams at the Universities of Sheffield, Brighton and Manchester have begun analysing a number of manually de-identified sample narratives to design the target templates and data structures that are anticipated to be derived from the complete corpus. A clinical advisory group has been active throughout the project in proposing the kinds of clinical queries and data elements that are likely to be of greatest value to the research community, as well as contributing to the ethical and security approaches described in this paper.

Conclusions

CLEF explores options and policies concerning a pseudonymisation solution to parallel the 'consent' approach underpinning the BioBank initiative. CLEF will identify processes and procedures that are both technically feasible as well as politically and socially acceptable to permit continuing and more efficient access to medical records to further medical research.

CLEF may give rise to an ongoing research database if there is continuing funding and sufficient subscribing organisations are prepared to provide data under this approach. Equally, the policies and methods developed may serve to inform other projects within the UK (e.g. the current NHS National Programme for IT) or around the world.

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.

Acknowledgements

CLEF is supported in part by the grant G0100852 from the MRC under the E-Science Initiative. 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 & Extended Bibliography

Legal aspects

Data Protection Act 1998, The Stationery Office Limited London 1998, www.hmso.gov.uk/acts/acts1998/19980029.htm

Human Rights Act 1998, The Stationery Office Limited London 1998, www.hmso.gov.uk/acts/acts1998/19980042.htm

Regina v Department of Health, Ex parte Source Informatics Ltd (CA) [2001] QB 424, [2000] EuLR 397, [2000] 1 AllER 786, The Times 18 January 2000 (use of anonymised patient information), www.lawreports.co.uk/source.htm

EU Directive 95/46, http://europa.eu.int/comm/internal_market/privacy/law_en.htm

Confidentiality Guidance

British Medical Association, Confidentiality and disclosure of health information, BMJ Publishing Group, London 1999, www.bma.org.uk/public/ethics.nsf

General Medical Council, Seeking patients' consent: the ethical considerations, GMC London 1999, www.gmc-uk.org/standards/consent.htm

Medical Research Council, Personal Information in Medical Research, MRC London 2000, www.mrc.ac.uk/pdf-pimr.pdf

Privacy enhancing techniques

L. Sweeney. k-anonymity: a model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10 (5), 2002; 557-570

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

Murphy S, Chueh H. A Security Architecture for Query Tools used to Access Large Biomedical Databases. Procs AMIA 2002 Annual Symposium 552-556

Electronic Health Records

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

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

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.

Kalra D. Clinical foundations and information architecture for the implementation of a federated health record service. PhD Thesis. Univ. London. 2002.

NHS National Programme for IT, www.doh.gov.uk/ipu/programme (accessed 29/07/03)