Clinical e-Science Framework
De-identifying the EHR: building a resource for research
All Hands Meeting - BOF session
Dr Dipak Kalra, UCL on behalf of the CLEF Consortium

CLEF’s Goals
- Collect clinical information from multiple sites
- Analyse, structure and integrate it
- Make it available using GRID tools (e.g. myGrid)
- To authorised clinicians and e-Health scientists
- In a secure and ethical collaborative framework

The CLEF records resource
- A repository of longitudinal cancer clinical records
  - that has been analysed and semantically indexed
  - to provide a summary of what happened and why at each point in a patient’s evolving story of care
  - that can be queried across substantial populations of similar patients through an intuitive query workbench

The CLEF repository has to be:
- scalable to populate
  - capable of incorporating large numbers of fine grained personal health records
  - from many different clinical systems in primary, secondary and tertiary care
  - each longitudinally linked so that the CLEF record can grow as each actual patient’s care progresses
- widely accessible to distributed research teams across the UK and ultimately internationally
- conformant to ethical and legal requirements

All use of personal health data is regulated
- In the UK:
  - Common Law of Confidentiality
  - Data Protection Act 1998
  - Human Rights Act 1998
  - Section 60 of Heath & Social Care Act 2001
  - BMA Guidance Oct 1999
  - GMC Guidance Sept 2000
- At a European Level
  - Council of Europe Recommendation R(97)5 (1997)

Personal data
- The Data Protection Act defines "personal data" as: "data which relate to a living individual who can be identified (a) from those data, or (b) from those data and other information which is in the possession of, or is likely to come into the possession of, the data controller"
- This is likely to apply to any clinically useful information about living patients
- 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

This is non-negotiable
Anonymised research data

- If legitimately processed for research or statistical purposes, "can be kept indefinitely and are exempt from the subject access rights if the results of the work are not made available in a form from which data subjects can be identified."
- If CLEF can make sure the data is anonymous, consent is not required and the data may be used for any reasonable research purpose.
  - This is the only scalable approach.
- But... no anonymisation can be perfect.

The CLEF ethics approach

1) De-identification
- Replace real patient identifiers with random keys
  - Done securely by the clinical site providing the data
  - Consistent longitudinally within an enterprise
  - Not known to CLEF
  - But it remains possible for the original site to re-identify a patient if it is warranted.
  - "One-way key encryption".

2) Depersonalisation

Medical narratives (letters, reports, summaries) are
- Rich in useful clinical data
- Most likely to reveal something personal about the patient.

2a) CLEF tools will lexically analyse all such narratives
- To remove occurrences of personal names or references, locations, highly-specified occupations etc.
- To extract and code the key features of the clinical story and care process.
- Records will be stored within a standards-based architecture.
  - Incorporating formal access control measures.

2b) These original depersonalised narratives will not be accessed directly by the query workbench.
- Access will be limited to the extracted coded data.

Pseudonymisation at hospital

- Data acquisition cycle
  - Integration & aggregation
  - De-personalise
  - Pseudonymise in hospital

- Extract information
  - Construct 'chronicle'
  - Summarise & formulate queries
  - Release identity by hospital

- Data access cycle
  - Knowledge enrichment
  - Hazard monitoring
  - Privacy enhancement technologies

- Data acquisition cycle
  - Architecture outline

Pseudonymisation

- Individual summaries & queries
- Data access cycle

Pseudonymisation technologies

- Incorporating formal access control measures.
- CLEF entry key
- Consistent longitudinally within an enterprise.
- Done securely by the clinical site providing the data.
- Regulatory, restricted and monitored access.

Pseudonymisation

- Medical narratives (letters, reports, summaries are rich in useful clinical data.
- Most likely to reveal something personal about the patient.

2a) CLEF tools will lexically analyse all such narratives
- To remove occurrences of personal names or references, locations, highly-specified occupations etc.
- To extract and code the key features of the clinical story and care process.
- Records will be stored within a standards-based architecture.
  - Incorporating formal access control measures.

2b) These original depersonalised narratives will not be accessed directly by the query workbench.
- Access will be limited to the extracted coded data.

1b) Exclude highly identifying data elements from the record extraction
- E.g., demographics (except postal district, gender, year of birth).

Anonymised research data

- If legitimately processed for research or statistical purposes, "can be kept indefinitely and are exempt from the subject access rights if the results of the work are not made available in a form from which data subjects can be identified."
- If CLEF can make sure the data is anonymous, consent is not required and the data may be used for any reasonable research purpose.
  - This is the only scalable approach.
- But... no anonymisation can be perfect.

The CLEF ethics approach

1) De-identification
- Replace real patient identifiers with random keys
  - Done securely by the clinical site providing the data
  - Consistent longitudinally within an enterprise
  - Not known to CLEF
  - But it remains possible for the original site to re-identify a patient if this is warranted.
  - "One-way key encryption".

2) Depersonalisation

Medical narratives (letters, reports, summaries) are
- Rich in useful clinical data
- Most likely to reveal something personal about the patient.

2a) CLEF tools will lexically analyse all such narratives
- To remove occurrences of personal names or references, locations, highly-specified occupations etc.
- To extract and code the key features of the clinical story and care process.
- Records will be stored within a standards-based architecture.
  - Incorporating formal access control measures.

2b) These original depersonalised narratives will not be accessed directly by the query workbench.
- Access will be limited to the extracted coded data.

Pseudonymisation at hospital

- Data acquisition cycle
  - Integration & aggregation
  - De-personalise
  - Pseudonymise in hospital

- Extract information
  - Construct 'chronicle'
  - Summarise & formulate queries
  - Release identity by hospital

- Data access cycle
  - Knowledge enrichment
  - Hazard monitoring
  - Privacy enhancement technologies

- Data acquisition cycle
  - Architecture outline

Pseudonymisation technologies

- Incorporating formal access control measures.
- CLEF entry key
- Consistent longitudinally within an enterprise.
- Done securely by the clinical site providing the data.
- Regulatory, restricted and monitored access.

Pseudonymisation

- Medical narratives (letters, reports, summaries are rich in useful clinical data.
- Most likely to reveal something personal about the patient.

2a) CLEF tools will lexically analyse all such narratives
- To remove occurrences of personal names or references, locations, highly-specified occupations etc.
- To extract and code the key features of the clinical story and care process.
- Records will be stored within a standards-based architecture.
  - Incorporating formal access control measures.

2b) These original depersonalised narratives will not be accessed directly by the query workbench.
- Access will be limited to the extracted coded data.

1b) Exclude highly identifying data elements from the record extraction
- E.g., demographics (except postal district, gender, year of birth).
Depersonalisation by CLEF Language Technology...

Non-obvious identifying information removed using language technology.

3a) Regulation and restriction of access

- Ethical Oversight Board will approve the kinds of organisations, teams and purposes for which the CLEF repository may be queried
  - defining the appropriate security measures to be taken
  - e.g. for authentication, authorisation and encryption
- A research project specific identifier will be used for data extracts to prevent cross-linking
  - the approval process will determine the extent to which longitudinal access to records is required, and the extent of drill-down permitted

Privacy Enhancement & authorisation

Queries logged, threats to confidentiality monitored.

With special authorisation, researchers may examine individual records in anonymised form.

Gainin ethical approval

- This depersonalisation process has MREC approval as a valid candidate methodology but as it has not yet been validated, CLEF cannot yet use live patient data without consent
- However, the project has been approved to use the records of deceased patients as an initial step towards developing, refining and evaluating the depersonalisation approach
- If successful, CLEF hopes to be permitted to migrate to live patient's records next year

Extraction of key information from text

Information Extracts identifies events and relationships between them from the text, based on templates & knowledge resources.

3b) Monitoring of access

- All accessed will be logged in an audit trail database
- Published algorithms will be used to help detect attempts to combine queries maliciously
- Selected research clients will be requested to help spot personal characteristics that slip through the net
  - the process of depersonalisation is still early R&D

Gaining ethical approval

- This depersonalisation process has MREC approval as a valid candidate methodology but as it has not yet been validated, CLEF cannot yet use live patient data without consent
- However, the project has been approved to use the records of deceased patients as an initial step towards developing, refining and evaluating the depersonalisation approach
- If successful, CLEF hopes to be permitted to migrate to live patient's records next year
Intended final security results

• A validated approach
  – accepted by MREC, PIAG, and other stakeholder groups (BMA, GMS, NHS, etc.)
• Exemplar policies and procedures
  – Ethical Oversight Committee
  – employee/researcher contracts
  – safe data extraction
  – access controls
• Open source tools
  – mechanisms to support security
  – active monitoring of use, limiting risk of inferential attack