Data Linking and Integration for Health Applications

eResearch Australasia, Brisbane, 27 June 2007

Dr Anthony Maeder
Research Director
E-Health Research Centre / CSIRO ICT Centre
Brisbane, Queensland, Australia

Overview

- E-Health and Health Data
- The HDI Software Tool
- Current Projects
- Future Directions
E-Health and Health Data

Scope of e-Health

- Contemporary health care has adopted evidence-based medicine delivered by multi-disciplinary, multi-party health care teams in a patient-centred approach
- e-Health encompasses the broad application of Information and Communication Technologies, in support of health care needs
- The main e-Health domains of activity are:
  - Health Information Systems (data and software tools)
  - Health Services Delivery (work practices and processes)
Australian Health Systems Scene

- **State-based systems**, with different structures for management of hospitals and community health
- National and local organisation of **private providers**
- Australian Government provides **financial resource**
- Many independent legacy software systems: specialised, non-interoperable, unsupported
- Safety, Quality and Efficiency issues driving **reform of work practices** and wider sharing of information
- Key problem is **lack of universal health identifier**: National e-Health Transition Authority aims at this

Need and Benefits of Data Linking

- Currently patient data resides across numerous different databases which are unconnected and owned separately
  - Different information systems and reporting systems
  - Government vs Hospital vs GP vs Allied health systems

- **Health care improvement opportunities flow from linking this data**
  - higher levels of patient care due to fuller information
  - extension of evidence-based practice
  - better planning or decision making for specific cases
  - improvements to training and education, safety and quality
Privacy Issues

- National Privacy Principles
  - Use of health data for treatment purposes
  - Secondary use of data for research purposes

- State Acts related to Health Data:
  - Health Records and Information Privacy Act (2002) - NSW
  - Health Records Act (2001) – Victoria
  - Many others

- Organisation principles
  - Other policies applicable site by site
  - Access to data governed by ethics compliance

---

Privacy & Related Legislation in Australia

**Consent:***
- Explicit
- Implied
- Statutory
- At the point of collection
- Reasonable measures

**收集:**
- mf, IT, telehealth

**Health Records:**
- Health Records and Information Privacy Act (2002) - NSW
- Health Records Act (2001) – Victoria
- Many others

**Organisation principles:**
- Other policies applicable site by site
- Access to data governed by ethics compliance

---

Source: Victorian Privacy Commissioner

Prepared by Michelle Fisher, Manager (Policy)
Health Data Sources

- **National level**
  - Medicare – cost codes identify treatments
  - PBS - Pharmaceuticals Benefits Scheme

- **State level**
  - Health department hospital admissions data
  - State based disease-specific data collections
  - Pathology reports & results
  - Radiotherapy reports
  - Radiology reports & images
  - Registries

- **Hospital level**
  - Hospital Information system
  - Hospital pharmaceuticals database

- **Hospital units**
  - Clinical information systems
  - Unit specific data sources

- **Clinical areas**
  - Clinician based data sources

- **External sources**
  - General Practitioners
  - Emergency Services
  - Allied health enterprises

Data Utilization

- Data collection management and organisation
- Knowledge discovery (population or cohorts)
- Understanding and comparison of cases
- Pre-processing and reshaping
- Statistical correlation and analysis
- Data aggregation and integration
- Identify events and trends
- Health awareness and promotion
Problems to Overcome for Data Linking

- Major practical impediments exist for data linking
  - Patient security and privacy restrictions
  - Diversity and independence of databases
  - Complexity of data formats
  - Sophistication of aggregation methods

- Existing solutions tend to adopt a heavy approach
  - Manual processing to achieve one-off linking
  - Data repositories or warehousing
  - Trusted third party units offering linking services
  - Full scale integration and interoperability of all systems

The HDI software tool
The Health Data Integration Project

- Aims to provide **novel methods for linking** multiple databases, by allowing data custodians to retain control of data
- Processes queries and reporting operations **remotely** at the data locations (in situ)
- Allows **privacy and security** restrictions to be met
- Enables **audit trails** of data access operations and users to be produced
- Fully **software engineered application product** has been produced at EHRC after about 20 person-years of effort
HDI: Data Custodian Control

- Data custodian retains control and security
  - No warehousing of data
  - All patient-identifying data is encrypted
- Databases are added to a HDI installation by the data custodian
  - The data custodian specifies who can use the data and how they can use it
- Metadata layer linked to industry standards to provide a common language and across repositories – increasing usability

HDI: Delineation of Responsibility

- HDI Domain concept
  - Provides demarcation of roles and responsibilities
    - Domain Administrator
    - Data Custodian
    - Project Administrator
    - Project member
  - Supports existing ethics committee approval process
Query on linked CRC surgical data and chemotherapy data to get a dataset of information on CRC patients, their current status and their chemotherapy treatments.

Query results show CRC surgical data and chemotherapy data for each de-identified patient.
HDI: Performing an Analysis

- Generate the survival chart, for example, Kaplan Meier.
- Analyse survival outcomes by stage.

HDI: Generating a report

Data from multiple databases can be used for reporting.

Eg… linked data from surgical databases, chemotherapy records and cancer registries.
Current Projects

Data Linking Projects

Some Current Clinical Applications of HDI

- Queensland Health: Queensland Oncology On Line
- Royal Melbourne Hospital: Colorectal Cancer
- Sydney South West Area Health Services: Colorectal Cancer
- Automated Cancer Staging: Lung Cancer
Royal Melbourne Hospital: Colorectal Cancer - Overview

Objective
- Integration of CRC screening, surgery and family history information:
  - For research on sensitivity and specificity of the Faecal Occult Blood Test (FOBT)
  - For research on surgical outcomes including factors such as adjuvant therapy and co-morbidities

Cohort
- Surveillance program patient data over 25 years (approx 3000 records)
- Surgical data of approx 4000 records; Family history approx 4000 records

Databases
- 3 databases (CRC Screening Program (administrative and outcomes data), Surgical and Clinical CRC data, Family History data)

Data Quality
- Significant cleaning work required on databases to regularise recorded data and remove data entry, data format, or historical information change errors
Royal Melbourne Hospital: Colorectal Cancer - Outcomes

- Data cleaning effort and data analysis gave access to 25 years of surveillance information
  - FOBT sensitivity published

Surgical Outcomes
- By factors such as smoking history, adjuvant therapy or diabetes status

CRC Services Research Group (NSW) - Overview

- Objective
  - Demonstrate the ability to gather patient information across databases and across hospitals to provide summary information on quality and safety of patient care, adherence to clinical guidelines and comparison across hospitals
  - To be able to provide this information to clinicians for access to their own data for case management AND for monitoring as above

- Cohort
  - Approx 1000 ‘dummy’ patient records with CRC and corresponding administrative and treatment information

- Databases
  - Different databases (Registry, Administrative, Surgical, Chemotherapy) for 3 different hospitals

- Data Quality
  - Constructed data included deliberate errors in patient demographics and different formats of clinical information to demonstrate HDI’s ability in inexact matching and transformation of data entries to a selected standard
Hospital A: Large teaching hospital, included in Cancer Institute NSW Clinical Cancer Registry Pilot

Hospital B: Private hospital, not included in Cancer Institute NSW Clinical Cancer Registry Pilot

Hospital C: Non-teaching hospital, not included in Cancer Institute NSW Clinical Cancer Registry Pilot

Hospital A - All chemotherapy (if required) performed by Hospital A

Hospital B - All chemotherapy (if required) performed by Hospital B

Hospital A - All chemotherapy (if required) performed by Hospital A or B

Hospital C – less CRC cases in general, and some harder cases referred to Hospital A or B. All chemotherapy (if required) performed by Hospital A or B.

Hospital A: Cancer Institute NSW Clinical Cancer Registry

Hospital B: Oncologist 5: Chemotherapy Records DB

Hospital C: Surgeon 4: Procedural Records DB (no chemo data)

We have demonstrated that with HDI it is possible to report on indicators that require information to be linked within and across hospitals
Staging of a cancer requires access to all available data:
- Radiology and histology text reports
- Information extracted from other forms of data, for example, radiological images

Cancer Stage Interpretation System (CSIS)
- Cancer staging is necessary to determine effective care for individual patients, as well as to design and evaluate health programmes at a population level
- Develop improved ways to access and analyse stored medical images and reports to better facilitate the staging of cancer patients
Future directions

Clinical to Genomic

- Integrating biomarker data sources and patient data can provide information on the efficacy, safety and toxicity of drugs
  - Advanced non-small-cell lung cancer - Iressa effective in only 10% to 15% of patients. Scientists pinpointed mutations in a gene within some tumor cells that allows Iressa to work


- Micro Array experiments may be useful in finding disease biomarkers
  - Linking micro array results with known clinical outcomes will allow new biomarkers to be found
EHRs and Virtual Registry

- **Pre-population of Electronic Health Record**
  - Use data linking to gather as much information as possible (to required health record specification) about an individual to pre-populate their complete health record
  - Outcomes:
    - Reduced manual workload for initial set up of health record

- **Virtual registry**
  - Use data linking to gather registry data set information about patients from existing sources (rather than building brand new sources with paper based information submission) – and present back a view of the registry for administration and analysis
  - Outcomes:
    - Significantly reduced manual and paper based workload for clinicians and registry officers resulting in more timely registry information, and greater compliance with registry requirements if data capture is at source
Contemporary ICT Advances

- Web Services
  - Easy(ier) federation of data and services

- Ontologies
  - Relating concepts using semantic properties
  - Discipline based
    - Need domain expertise to define the ontology of data sources and fields

- The semantic web
  - Using ontologies in the web applications

Using SNOMED CT

- Use SNOMED CT for mapping of data to terms
  - Structured data sources
  - Natural language
    - Reports
    - Discussions

- To make this possible
  - Subsets for particular domains
  - Augmenting the domain specific subsets
  - Fast querying
SNOMED-CT Scope

- Clinical Terms
- Comprehensive, not specialty or domain
  - Human, veterinary, drugs, social, disease, observations, interventions and wellness
- ~400,000 concepts (fully specified names)
- ~1M descriptions (synonyms etc)
- ~1.4M relationships (900,000 defining)
- URU principles
  - Useable, Repeatable, Understandable

SNOMED CT Top-level Concepts

- 138875005 | SNOMED CT Concept
  - 123037004 | body structure |
  - 404684003 | clinical finding |
  - 243798009 | context dependent category |
  - 308916002 | environments and geographical locations |
  - 272379006 | event |
  - 106237007 | linkage concept |
  - 363787002 | observable entity |
  - 410607006 | organism |
  - 373873005 | pharmaceutical / biologic product |
  - 78621006 | physical force |
  - 260787004 | physical object |
  - T1388002 | procedure |
  - 362981000 | qualifier value |
  - 419891008 | record artifact |
  - 48176007 | social context |
  - 370115009 | special concept |
  - 123038009 | specimen |
  - 254291000 | staging and scales |
  - 105590001 | substance |
Expressions, equivalence and subsumption

- 85189001|acute appendicitis
  - subsumed by
    - 74400008|appendicitis|
    - 64572001|disease|:
      116676008|associated morphology|=23583003|inflammation|
      363698007|finding site|=66754008|appendix structure|
  - equivalent to
    - 74400008|appendicitis|:
      260908002|course|=53737009|acute|

Querying

- Simple queries can use subsumption/equivalence, for example
  - find all cases of 74400008|appendicitis| with a particular treatment or outcome
  - finding those also classified as
    - 85189001|acute appendicitis
    - or even as a 64572001|disease| with 23583003|inflammation| of the 66754008|appendix structure|
HDI Platform Technology

- De-identified linked data for analysis
- HDI integrates data
- Custodial controlled data

HDI HUB

- Statistical Packages e.g. R, SPSS
- Reporting Tools e.g. Crystal Reports
- Custom Applications

De-identified virtual linked data set

HDI Data Source

http://ict.csiro.au/