e-Labs: Interoperable data, methods and people enabling population-based research

NCRI & caBIG Informatics Conference, 13th July 2007

Iain Buchan (Director)
Northwest Institute for BioHealth Informatics
www.nibhi.org.uk
Current position

• Hard to access basic clinical data

• Research fails to harness care data

• Artificially slow & expensive research

• Clinical creativity underused
Example of missing basics

• Hypothesis: Tamoxifen ineffective due to interactions at CyP-26D, e.g. with paroxetine

• Request to cancer registry →
  − Breast cancer recurrence data missing
  − Adjuvant therapy data patchy

• The data are in GP systems
Missing evidence

• Time-course
  – Early natural histories
  – Emergent risks

• Big picture
  – True clinical effectiveness
  – What-if/scenario planning
Predictors of treatment response:
- Fixed risks
- Lifestyle factors
- Co-morbidities
- Co-treatments
- Dose & delivery
- Treatment setting

Risks → Patient Population → Selection → Trial Population

% loss

True response

Efficacy & Safety

General Population → Risks
Givens for NHS data

• Increase in quantity & structure of data (biological; clinical; population)

• Complex errors (miscoding, gaming, change of assay etc.) → need meta-data

• Overloaded with irrelevant information & can’t find timely, relevant information
Opportunity of local collections

There is a trend towards *local* clustering of academic & NHS expertise around...

- tissue banks
- ‘omic facilities
- disease registers
Isolated research or care R&D federation?

Or...

Local NHS → E-Lab → Research Networks → Research

Service development
Building the Informatics Capacity

The Northwest Institute for Bio-Health Informatics (NIBHI) was formed in 2004 to build informatics capacity to leverage the discovery potential of bio-health data...

The stakeholders are:
What is *Bio-Health* Informatics?

- **inputs**
- **intellectual space**
- **outputs**

**Data & Information**

- Health
- Computational Thinking
- Statistics
- ‘omics

**Knowledge:**
- mechanisms
- interventions
- policies
NIBHI objectives

• Create a critical mass of trans-disciplinary informatics expertise

• Build the core e-Labs programme between the NHS and academia

• Deliver proof of principle outputs
  – Findings from novel uses of routine data
  – New software and analytical methods
Discovery themes and platforms

- Obesity
- Inflammation & lipids
- Psychosis
- Drug safety
- New: Cancer...

Population-based E-Labs & E-Epidemiology
Biostatistics & Translational Mathematics
Social BioHealth Informatics
Social Bio-Health e-Lab

- Research
- Service Development
- Academia Partners
- Health & Social Care Knowledge
- Industry Partners
- Social Contract & Governances
- Local Information Systems
- Population
e-Lab Processes

Raw Data
- GPs
- Hospital(s)
- Biobanks
- Surveys
- LA/Council
- National

Enhanced Data
- Link records
- Clean
- Organise
- Add metadata
- Share algorithms

Research
- Safety
- Real effectiveness
- Efficient trials
- Exploratory
- Public health
- Service models

Depersonalise
Integrating Records → Care in Salford

EMIS & Vision

GP

GP

GP

XML (HL7v3)

Integrating Records

Hosp. EPR

Integrated Database (at PCT)

Master Patient Index

Demographic Service

Research & Clinical Audit

XML Journal File

Pseudonymised Repository

Web Server

Web view of Patient Record

Web Forms

Optometrist

eye screening

Community nurses

Podiatry

Analysis Tools
Chronicles of care

![Graph showing trends in Total Cholesterol and Systolic BP over the years from 1993 to 2007. The graph indicates a decreasing trend in both parameters.]
New diagnoses of type 2 diabetes
Proportion of Diabetics Anaemic or Hb < 110

Chronic Kidney Disease Category

- eGFR >= 60
- eGFR 50-59
- eGFR 30-49
- CKD 4
- CKD 5

eGFR: Estimated Glomerular Filtration Rate
Hb: Hemoglobin
Research Examples

A) Iterative shaping of hypothesis between biology and medicine, plus enhanced data collection
African cows to Salford ICU

High cholesterol in African cattle identified as a protective factor against death from trypanosomiasis

Is high cholesterol a protective factor in humans undergoing extreme inflammation?

ICU data and physicians in Salford ‘E-Lab’ accessible, and physiological ‘clamping’ reduces confounding

Data cleaning, meta-data capture, analysis
Change in total cholesterol

- Day 1 to 2
- Day 2 to 3

Died
Survived
Cholesterol & inflammation

• Discovery theme growing
  – Shaping and testing hypotheses around causal vs. reverse causal models
  – Speculative trials in mouse model
  – Metabolomic proposal submitted

• Enhanced data collection and coding because it was of BENEFIT TO CARE
  = data quality by-product of research
Example 2) Discovery from routine data in a long series

Surveillance of obesity in Wirral children whilst the epidemic took hold...
Child Deprivation (2001 Census)

**Fifths of IDAC 2004**
Proportion of households with children claiming benefits

Red (light) = most deprived
Red (dark)
Purple
Blue (dark)
Blue (light) = most affluent
Adiposity of 3 yr olds
1988 - 1989

Fifths of adiposity
SDS BMI fifth
Red (light) = fattest
Red (dark)
Purple
Blue (dark)
Blue (light) = thinnest
Adiposity of 3 yr olds
1990 - 1991

**Fifths of adiposity**
SDS BMI fifth

Red (light) = fattest
Red (dark)
Purple
Blue (dark)
Blue (light) = thinnest
Adiposity of 3 yr olds
1992 - 1993

Fifths of adiposity
SDS BMI fifth
Red (light) = fattest
Red (dark)
Purple
Blue (dark)
Blue (light) = thinnest
Adiposity of 3 yr olds
1994 - 1995

Fifths of adiposity
SDS BMI fifth
Red (light) = fattest
Red (dark)
Purple
Blue (dark)
Blue (light) = thinnest
Adiposity of 3 yr olds
1996 - 1997

Fifths of adiposity
SDS BMI fifth
Red (light) = fattest
Red (dark)
Purple
Blue (dark)
Blue (light) = thinnest
Adiposity of 3 yr olds 1998 - 1999

Fifths of adiposity
SDS BMI fifth

Red (light) = fattest
Red (dark)
Purple
Blue (dark)
Blue (light) = thinnest
Adiposity of 3 yr olds
2000 – 2001

Fifths of adiposity
SDS BMI fifth
Red (light) = fattest
Red (dark)
Purple
Blue (dark)
Blue (light) = thinnest
Adiposity of 3 yr olds
2002 - 2003

Fifths of adiposity
SDS BMI fifth

Red (light) = fattest
Red (dark)
Purple
Blue (dark)
Blue (light) = thinnest
Taller, faster-growing children have carried more of the obesity epidemic
Faster infant growth and slower early child growth

- Dotted line: 6 weeks to 8 months
- Solid line: 8 months to 2 years

Mean (95% CI) length/height gain (cm)

Birth Year

- Mean length/height gain increases from 13.5 to 17 cm over the years.
- Growth rates are faster in the first 8 months compared to the following 2 years.
Obesity and linear growth

• Direct public health / policy use

• New quest to identify the links between growth and obesity - ? targets within individuals or just shared environment

• Mouse equivalent of ‘tall-fat’ child phenomenon identified with knock out of gene on putative pathway - ? target
Discipline-based E-Labs

e.g. PsyGrid
PsyGrid Connectivity Architecture
PsyGrid

- Aim to ‘e-enable’ the world’s largest cohort study of first-episode psychosis

- 2 years in
  - Cohort study running well
  - Two ‘e-enabled’ trials mounted on cohort
  - Extreme care with security
  - Focus on rapid, systematic data capture
  - Adopted by Mental Health Research Network
  - Translating to other networks
Linkage:
Data-data
Researcher-data
Researcher-computation

Engineer a factory,
not just a warehouse...
Variations in DNA sequence, or **Single Nucleotide Polymorphisms (SNPs)** between individuals can contribute to disease conditions. Many diseases result from complex interactions of multiple SNPs.

**The scientific problem**

Variations in DNA sequence, or **Single Nucleotide Polymorphisms (SNPs)** between individuals can contribute to disease conditions.

**The data**

Genotype at each position in genome is encoded by 2-bits. Disease status (+/-) of each individual recorded.

Person 1 +

11 12 22

...ATTAGGACAAATAGCTT...  
...ATTAGGACAAATAGCTT...

Person 2 -

11 22 12

...ATTAGGAGCAATAAGCTT...  
...ATTAGGAGCAATAAGCTT...

**The computational problem**

Modern technology can measure 0.5M SNPs on cohort of 5k individuals. Compute association between disease status and genotype. Significance computed by comparing to 10k random data sets.

**The solution**

**Parallelization for Genome-wide Screening of Common Complex Human Diseases**

University of Manchester (NIBHI, ARC), Microsoft Technology Centre (Thames Valley, UK) & Melandra Ltd.

**Current technology**

- Single locus: 1 day
- Pairs of loci: 120 yrs

**The solution**

- Easy access to analysis services by remote collaborating scientists
- Workflows: User-friendly creation of bespoke analysis routines
- Rapid downstream mining of analysis results: Rapid biological knowledge discovery
- Increased compute power. Pair interactions = 1½ days (24 Opteron cores)

**Parallelization**

- Single locus: Cost = 0.5M x 5k x 10k
- Pair of loci: Cost = ½ x 0.5M x 0.5M x 5k x 10k

**Parallelization**

- SharePoint Services
- Portal Server
- WF .NET Framework 3.0
- MS SQL
- MS CCS

**Analysis speed**

**Barriers to use**

- Drip feed of scientific discovery
 shares Genomics’ with Microsoft

- Proof of concept goals achieved
  - enabled more powerful SNP interaction studies
  - made the research tools more accessible

- Speed
  - Test set: 100K SNP, 600 samples, 10K permutations
    - 1 - way
      PLINK = 4-5hrs
      NIBHI = 65 sec
    - 2 - way
      PLINK = 38hrs
      NIBHI = 11 mins

- Full project: from Oct 2007
  - Web-delivered genetic epidemiology tools
  - Fundamental mathematical and software work
  - Employ more powerful computers
Future: Big, Fast Exploration

Problem Question (Hypothesis)

Drill down

Population
Individual
Molecular

Factors

Model

Findings

Mine
Making research more transparent, sharable, reproducible, and easily preserved
Unclear Public Good

- Research
- Health Records
- Depersonalise

Clear Public Good

- Research Objects
- e-Lab: Population
- Local Ownership
- Asset Enrichment
E-Lab Local Sustainability

- Standards
  - Governance
  - Technical
- Easy, open tools
- Local use & ownership
- Compelling findings

Interoperable local e-Labs = confederation
## Northwest e-Lab Confederation

<table>
<thead>
<tr>
<th>Technology</th>
<th>Experience</th>
<th>Local ownership</th>
</tr>
</thead>
<tbody>
<tr>
<td>Salford e-Lab pilot: 230k</td>
<td>e-Lab model adopted within key NHS Trusts 1.5M</td>
<td>Regional adoption 3.5-5.0M</td>
</tr>
</tbody>
</table>
Learning from e-Lab pilots

R&D pipelines from/for healthcare =
Data +
Local expertise $\rightarrow$ meta-data +
Easy collaborative research platforms

Federate local ownership & enrichment
Cancer e-Lab recommendations

• **Co-invest** to create a critical mass of Social Bio-Health Informatics

• Embed spokes of e-Labs in cancer centres

• Set cancer interoperability tests for wider health intelligence systems
  – don’t isolate cancer
Future cancer intelligence tests

- Tamoxifen question answered within 3h
- Kaplan-Meier update within 14d of event
- Routine pre-diagnosis signals detected
- Phase IV+: natural control groups identified
- Real-time feasibility analysis
- Real-time workflow-based recruitment
- Long-term content-feeds on trial participants
- >90% of tissue samples registered in e-Lab
- ...
Thanks for your attention

www.nibhi.org.uk