Digital Curation for the Public’s Health
Ethics, Security & Trust

DCC/RIN Research Data Management Forum
“Dealing with Sensitive Data: managing ethics, security and trust”

10th March 2010

Prof. Iain Buchan
University of Manchester
Digital Curation for the Public’s Health

• Where does the public’s health need digital innovation?

• How can research curators promote this innovation?
  – Ethics
  – Security
  – Trust

• Is a framework required?
  – Social contract
  – Digital & operational infrastructure: e-Lab
Public Health Needs for Digital Innovation: I

CITIZEN-LED PREVENTION & EARLY INTERVENTION

Approaches to Healthcare

• Clinical model:
  – Rescue the ill
  – Resource $\propto$ illness
  – Specialise to optimise

• Public Health model:
  – Rescue the at-risk
  – Resource $\propto$ disease/risk
  – Generalise to optimise
Healthcare Digital Innovation is Mostly Clinical

- Specialist-driven
- Excellent in niches
- Incremental

...while public health technologies may be left to perverse market-forces.
A life-course view of personal health

- Birth
- Instability
- Hip fracture

Personal cost: healthy years of life lost
Societal cost: large proportion of healthcare resource used

+ structured exercise programme

Death

Amount of Potentially Healthy Years

Lifecourse
State-led: Health Promotion

Manchester Public Health Development Service

Citizen-led: Health & Wellbeing

Know → Act

Feel → Interact

Persuasive Technology

Social Marketing

Social Networks
Re-wire the brain to resist over-malnutrition?

Burning Fat

Depositing Fat

(ketone/other) molecules on skin

Active Polymers in wristband +/- other signals & data

Frequent Choices
Care costs escalate without prevention

- Public Health
- Prevention
- Wellbeing

Citizen > State

Unified Modelling
Digital Coherence

State > Citizen

Clinical Care
Social Care
Science

Life-course view
(barely visible investment)
Life of a person

Sickness view
(fire-fighting investments)
Life of a parliament
DYNAMIC LINKAGE OF RESEARCH INTO PRACTICE AND PRACTICE INTO RESEARCH
Healthcare Problem: **Gaps in Communication & Organisation**

Self Care  
Primary Care  
Hospital A  
Clinical Care  
Secondary Care  
Hospital B
Digital Bridges Since 1990s: Integrated Care **Pathways (Disease-specific)**

Self Care

Clinical Care

Primary Care

Secondary Care

Hospital A

Hospital B
**Missing**: Patient & Community

‘**Big-picture**’ Across Disease / Specialist Pathways

Self Care

Diabetology: Glucose control
Ophthalmology: Diabetic eye care
Nephrology: Chronic kidney disease

Primary Care

Hospital A

Hospital B

Clinical Care

Person: Experiencing n of 1 trials

Future Health ‘Avatar’: Fuse research evidence about the average person with a longitudinal model of the person represented
Health Records & Knowledge Bases

Electronic Health Records (eHR)

Open Unifying Modelling: Across mechanisms and contexts

Data-intensive Paradigm -shift

Health Avatars & Dynamic Models

\[ \cup \text{models} = \text{Avatar} \]

- Multi-scale & Multi-system Health:
  - Research
  - Policy
  - Care

Unified Graphical Model

Expertise

Large scale inference

Model refinement

Data-intensive Paradigm shift

e.g. Chronic obstructive pulmonary disease

e.g. Lung cancer

\[ \land \text{mechanisms and contexts} \]
Trans-disciplinary Analysis of Clinical Research Data

Machine Learning in Epidemiology...
Introducing Health Sciences Signal Paths to Physical Scientists & Engineers

Problem Space

Observation Space

Data Space

\[ y = b_1x_1 + b_2x_2 + b_3x_3 + c \]

...like resolving an image through a prism through a doyley
Hypothesis-driven Epidemiology: Sieving Associations

<table>
<thead>
<tr>
<th>Association</th>
<th>Bias</th>
<th>Type</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>C → M</td>
<td>Cause-effect</td>
<td>Real</td>
<td>Cause-effect</td>
</tr>
<tr>
<td>MI → C</td>
<td>Reverse</td>
<td>Real</td>
<td>Effect-cause</td>
</tr>
<tr>
<td>C ↔ ? → MI</td>
<td>Confounding</td>
<td>Real</td>
<td>Effect-effect</td>
</tr>
<tr>
<td>C ↔ MI</td>
<td>Random error</td>
<td>Spurious</td>
<td>Chance</td>
</tr>
<tr>
<td>C ↔ MI</td>
<td>Systematic error</td>
<td>Spurious</td>
<td>Bias</td>
</tr>
</tbody>
</table>

C = caffeine, MI = myocardial infarction (heart attack)

Disciplined approach to causal inference, Bradford-Hill: Criteria (temporality, strength, dose-response, consistency, plausibility, consideration of alternatives, open to experiment, specificity, coherence)
Need to Address Complexity & Scale

Problem 1: Dwindling hits from tools to detect independent “causes”

Problem 2: Knowledge can’t be managed by reading papers any more

The big public health problems e.g. Type 2 Diabetes have “complex webs of causes”

The “data-set” and structure extend beyond the study’s observations
A Graphical Model of Asthma

1. Maximal attainable lung functions
   - Developmental narrowing – maternal smoking, maternal atopy, maternal age
   - Nutritional deficiency
   - Prematurity
   - Obesity
   - Pregnancy related factors

2. Repair
   - Remodelling

3. Reversibility

4. Bronchial hyperresponsiveness

5. Airway Inflammation
   - Outdoor air pollution
   - Allergic rhinitis
   - Acquired immune response
   - Allergic sensitisation
   - Allergen exposure
   - Mucus hypersecretion

6. Immune response to virus
   - epitelium
   - Viral infection
   - NANC

7. PERIPHERAL AIRWAY OBSTRUCTION
   - Smooth muscle
   - Asthma therapy

8. Microbial products
   - Innate immune response
   - Daycare
   - Siblings
   - Antibiotics
   - Dog
   - Cat
The Atopy Model

Unsupervised Clustering

Switch between Sens Class

Acquired Sensitization at age 1

Acquired Sensitization at age 3

Acquired Sensitization at age 5

Acquired Sensitization at age 8

Skin Test at age 1

Skin Test at age 3

Skin Test at age 5

Skin Test at age 8

IgE Test at age 1

IgE Test at age 3

IgE Test at age 5

IgE Test at age 8

Prob Pos Skin Test given sensitized

Prob Pos IgE Test given sensitized

Prob Pos Skin Test given NOT sensitized

Prob Pos IgE Test given NOT sensitized

To Infer
Atopy Classes and Asthma

Multiple Early Sensitisations
odds ratio (OR) [95% CI]: 30.5 [12.4-75.3]
c.f. Ever atopic group:

% with asthma (symptomatic on challenge)

- No atopy
- "Atopy 4"
- "Atopy 3"
- "Atopy 2"
- "Atopy 1"
Trans-disciplinary Simulation of Public Health Impacts

Digital Assets ➔ Policy Decisions...
Physical Activity
Diet
Smoking
Deprivation
Obesity (BMI)
Diabetes or IGT
Cholesterol (LDL & HDL)
Blood Pressure
Combined CVD Risk
Unstable Angina
Chronic Angina
Acute MI
Early Heart Failure
Recurrent MI
Severe Heart Failure
MI survivors
Stroke PAD etc
Additional CVD Risk Factors

Population Policies & Behaviours
Biological Risk Factors
Combined CVD Risk
CVD Patient Groups

Outputs: Population-based incidence, prevalence; Deaths prevented; Life-Years; Life expectancy; Costs; Cost-effectiveness ratios

Pulling evidence together into one, realistically-complex model: e.g. MRC IMPACT II
Recombine evidence around probabilistic graphical models of disease & care-services

Ask ‘what if’ invest in statins vs. smoking cessation etc.
Open Unified Models for Health Policy

Current

Separate Thinking

Researchers
Modellers
Policy Makers
Practitioners

Future

CLAHRC Systems

Expert Interfaces

Unified Open Model

Co-thinking

Leading Methodology
Local Scenario Planning
Open to Scrutiny

Future

Health Economists
Public Health Workers
Peer Reviewers
Clinicians
Epidemiologists
Healthcare Managers
Software Engineers
Statisticians
Can models be built from literature?

“Not fully”

Natural Experiments & Hypothesis-generating Patterns of Healthcare

Healthcare Evidence Base
Anaemia at lower levels of kidney impairment than commonly thought.

Clinical (audit) question leading to scientific finding: required local metadata (assay change) not in national datasets.
Framework for Digital Research in Healthcare: I

DIGITAL & OPERATIONAL INFRASTRUCTURE: E-LAB
Digital Dust (data deposit > use)
Data supply is not the bottleneck

Methods/Models ↑

Contextual expertise ↔

Data ↑↑↑↑
Cloud of millions of care messages in the local health economy

Structured Data

Organise

Transform & Examine

Structured Data & Metadata
NHS e-Lab: Salford Pilot

Data Repository in PCT

Real-time

Person-identifiable and sensitive information removed

FIREWALL

24-hourly updates

Anonymised Data Repository in PCT

Sense-making software & support

Outputs

Trusted person poses question(s)

Optometrist
Eye screening
Community nurses
Podiatry

Deaths, Demographics etc.

Biomics Data

Hosp.

Link on NHS number

GP

GP

GP

GP

GP

GP

GP

GP

GP

GP

GP

GP

GP
Federation: More local use $\rightarrow$ better quality data

Community development

Health Community 1
Health Community 2
Health Community 3

e-Lab

e-Lab

e-Lab

Research Networks

Research

North West e-Health

10 developers
Work/Research Object

Services

Workflows

Packages of Outputs

Reproduce
Share
Morph
Inherit

Research Object

Data/Workflow
Analysis Scripts
Annotations
Slides
Manuscripts
References

Encapsulated \rightarrow (DAG) discovery?
Socially-stimulating science, in-silico

Research Object

Find
Share
Reuse

Research protocol
Data-sources
Data-preparation scripts
Working datasets
References

Statistical analysis scripts
Analysis-logs & notes
Figures/Graphics
Manuscripts
Slides
Prototype NHS e-Lab

Life Expectancy In Salford

Project Details

What is life expectancy for wards in Salford

This is to answer the question of life expectancy in Salford with deprivation information for wards

- Data
  - Life Expectancy
  - Life Tables

- Documents
  - LifeExpectancy.csv

- Data Exploration
  - You have not added any explorations to your project. Add an exploration here.

- Notes
  - What is life expectancy for wards in Salford

- Snapshots
  - You have not added any snapshots. You can add a snapshot by browsing your Data and creating charts and maps.

- People
  - gary
## Life Expectancy in Salford

### Data

#### Life Expectancy

- **WordCode**: Locality codes.
- **WordName**: Locality names.
- **LifeExpectancy**: Life expectancy.
- **DeprivationSc**: Deprivation score.
- **IsHighDepriv**: Indicator for high deprivation.
- **IsHighDiabetes**: Indicator for high diabetes prevalence.
- **DiabetesPrev**: Diabetes prevalence.
- **CvdPrevalence**: CVD prevalence.
- **DiabetesPrevC**: Diabetes prevalence category.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>00BRFA</td>
<td>Barton</td>
<td>73</td>
<td>3.342892</td>
<td>false</td>
<td>true</td>
<td>6</td>
<td>10</td>
<td>9</td>
<td>7</td>
</tr>
<tr>
<td>00BRFB</td>
<td>Blackfriars</td>
<td>70</td>
<td>7.474104</td>
<td>true</td>
<td>true</td>
<td>5</td>
<td>6</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>00BRFC</td>
<td>Broughton</td>
<td>70.1</td>
<td>8.241365</td>
<td>false</td>
<td>true</td>
<td>10</td>
<td>15</td>
<td>11</td>
<td>9</td>
</tr>
<tr>
<td>00BRFD</td>
<td>Cadishead</td>
<td>72.3</td>
<td>1.498242</td>
<td>true</td>
<td>false</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>00BRFE</td>
<td>Claremont</td>
<td>74</td>
<td>0.4376208</td>
<td>false</td>
<td>false</td>
<td>3</td>
<td>5</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>00BRFF</td>
<td>Eccles</td>
<td>73.4</td>
<td>2.245419</td>
<td>false</td>
<td>false</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>00BRFG</td>
<td>Iriam</td>
<td>70.2</td>
<td>0.8358688</td>
<td>false</td>
<td>false</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>2</td>
</tr>
</tbody>
</table>
Next: Merge with visualisation research
Digital Curators Promoting Healthcare Innovation: I

ETHICS
Ethical Principles

- Respect for autonomy
- Beneficence
- Non-maleficence
- Justice

*Beauchamp and Childress; Principles Biomedical Ethics, OUP, 5th edition 2001*
Respect for Autonomy

• Patient/subject
  – Consent
    • Opt-in
    • Opt-out
  – Right to participate or not
  – Advocates where appropriate
    • Clinical; carer; guardian; data custodian

• Investigator
  – Access to patients/citizens?
Beneficence

• Duty to deliver good for the data donors

• Under-use is unethical

• Audit the context of use

• Measuring good
  – Research quality
  – Clinical utility
  – Patient/citizen involvement
Linked Health Records

Audit; Research; Intelligence

De-identify

Unclear Public Good

Clear Public Good

Work Objects

e-Lab for a defined community

Local Ownership

Asset Enrichment

Linked Health Records
Non-maleficence

• Part of NHS & University contracts

• Part of clinical and research information governance protocols

• Not dealt with by restricting access to data
Justice

• Knowing the uses of your health records
• Knowing how your practice is measured
• Fair access to data
• Fair access to methods
• Fair access to models
• Intellectual property protection
• Fair networking opportunities for investigators
Data Curation Example: Obesity e-Lab

The e-Science target:
Fragmented understanding of public health problems such as obesity

...data, methods/models and expertise split across disciplines (e.g. Social vs. Biomedical) and settings (e.g. Academia vs. Healthcare)
Wirral (0.3M), UK

Child poverty map
(households with children: % on benefits in 2001-3)

Fifths of IDAC 2004
Red (light) = most deprived
Red (dark)
Purple
Blue (dark)
Blue (light) = most affluent
BMI of 3 yr olds
1990 - 1991

Fifths of BMI
SDS BMI fifth

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

Fifths of BMI
SDS BMI fifth

Red (light) = fattest
Red (dark)
Purple
Blue (dark)
Blue (light) = thinnest
Fifths of BMI
SDS BMI fifth

Red (light) = fattest
Red (dark)
Purple
Blue (dark)
Blue (light) = thinnest
BMI of 3 yr olds
1996 - 1997

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

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

Fifths of BMI
SDS BMI fifth
Red (light) = fattest
Red (dark)
Purple
Blue (dark)
Blue (light) = thinnest
Women and not men from low-income households are fatter in England

Data from Health Survey for England
Women from low-income households and men from high-income households are fatter in Greater Manchester.

Data from Health Survey for England
Beware Discipline Clouds

**Social** Research:
Data, methods & people

*Previous slides show social-biomedical signals about obesity from under-used datasets*

**Biomedical** Research:
Data, methods & people
Obesity e-Lab Aim

..to increase the sharing and reuse of
data sources & extracts
and data processing methods
in one in-silico environment (‘e-Lab’)
shared by social and health researchers
Focus

• Health Surveys for England
  – Large-scale (participants * variables)
  – Annual since early 90s
  – Under-used by NHS who fund it
  – Key barrier: extracting a research-ready subset of data
  – Data archive → playground = e-Lab
Supporting and developing interdisciplinary understanding

<table>
<thead>
<tr>
<th>Sharing resources – tools, methods, data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sharing expertise – discussions and reuse around shared resources</td>
</tr>
<tr>
<td>Developing interdisciplinary understanding – language, tacit assumptions, methods</td>
</tr>
<tr>
<td>Promoting interdisciplinary working</td>
</tr>
</tbody>
</table>

First step - sharing of resources

Shared resources provide the basis for discussion

Discussions lead to deeper interdisciplinary understanding

Understanding of other domains promotes more effective interdisciplinary working
MethodBox users include NHS Public Health analysts and Department of Health Public Health Observatory analysts, as well as academic social scientists and epidemiologists.
Download methods and scripts shared by other Methodbox users

The owner of the script can create links back to other methods and scripts used in the development of this script.
Digital Curators Promoting Healthcare Innovation: II

SECURITY
Major Issues with Clinical Studies

• Bias & generalisability
• >50% over-run
• >30% don’t hit recruitment targets
• Unrealistic feasibility assessment
• Consent-management confusion
openCDMS

An Open Source System for:

• multi-centre remote electronic data collection;
• highly configurable security system employing Role Based Access Control;
• fully customisable data set definition including data elements, validation rules and scheduling;
• fully configurable online randomisation with email and SMS text message notification;
• project management reporting including recruitment, completion and UKCRN accrual;
• on-line and off-line data collection;
• flexible query system for identifying eligible trial participants and nested case-control studies;
• designed for compliance with 21 CFR part 11; EMEA GCP; ISO 27001
openCDMS in use

• PsyGrid study – cohort of 700+ schizophrenics followed from first episode for 18 months

• Running numerous mental health trials

• ADDRESS – Type I+II diabetes 10y cohort study

• DARE – Diabetes cohort (phenotype and genotype)
Investigator-shaped data capture
Clean Data
± Link External Data
± Collect Observations
Recruit Participants
Test Feasibility
Hypothesis
Test
Governance
(Helsinki & local)
Collect Observations
Recruit Participants
Test Feasibility
Seek Ethical Approval
Design Study
Generate Hypothesis
Publish
Explain Results
Fuse Data
Publish
Managing clinical studies

Making sense of local healthcare

+ community

North West e-Health

openCDMS - clinical data management system

e-Lab
Consent-for-consent

...is the consent to

search an individual’s health record
to determine whether or not they should be invited
to take part in a clinical study.
FARSITE

Feasibility Assessment

and Recruitment System

for Improving

Trial Efficiency
Study Protocol (sponsor)

+/- Tweak

Refined Search

Recruitment Estimate (researcher)
[de-identified records]

Email to attending clinicians
“Click to identify your eligible patients”
[health e-record]

Study Management (identifiable)

Study Management (identifiable)

Rapid, Consistent Feasibility & Audit

FIREWALL
Realistic Recruitment Estimates
Digital Curators Promoting Healthcare Innovation: III

TRUST
Trust & Benefit in Research across Health Records

Now
Database-centred

x Health Agencies

Clinical

Clinical Information Governance

Research Governance

Ethical Oversight

x Research Agencies

Future?
e-Lab: Community-centred

Anonymised e-Lab

Integrated Health

x Health Communities
Public Involvement

• Patient / Citizen Scientist

• Social network ↔ investigator

• Early & mobile signals beyond clinical reach

• Relevant outcomes
Framework for Digital Research in Healthcare: II

SOCIAL CONTRACT
Digital Curation for the Public’s Health: What is the Social Contract?

• UK has much already in place in the laws and governance across NHS, higher education and allied public services

• Law, standards, regulations and some infrastructure work nationally

• Trust and capacity-building to provide the best data and analyses works locally ← more attention
Open Unifying e-Lab

∪ (open models)

∪

Easy computation =

∪

Abstract reasoning and motivation of domain experts

More complete insights

More reusable evidence

Better management across diseases

Earlier intervention

Greater citizen involvement