Big health data: perspectives across the patient journey from linking multiple record sources

Harry Hemingway  FFPH, FRCP

Professor of Clinical Epidemiology

Director, Farr Institute of Health Informatics Research, UCL

15-17 September 2014, Birmingham
Outline

• What are big health data?

• Why me – personal journey

• What are big data good for?
  – Discovery
  – Trials
  – Outcomes, risk prediction and clinical decision making
  – Public health

• What is the role of the Farr Institute?
What are big data?
Big data

like teenage sex ‘everyone talks about it, nobody really knows how to do it, everyone thinks everyone else is doing it, so everyone claims they are doing it’

Dan Ariely
What is big data?
# How do we scale science in record linkages?

## National sources of health record data

### NCRS eg
- Personal Demographics Service
- Personal Spine Information Service
- Transaction Messaging Service
- Secondary Uses System
- NN48 / Central Issuing System
- Choose & Book, Payment by Results, GP2GP, etc
- Electronic Prescriptions

### NHS National Collections eg
- Commissioning Datasets
- Mental Health Minimum Data
- QOF / QMAS

### Specialist Collections eg
- Cancer / Diabetes / Renal Audit; waiting times; workforce

### Office of National Statistics
- Births, Deaths, Terminations & Marriages
- Census & Special Surveys
- Eg HSE, NDNS, GHS CEMACH, CEPOD, Infant Feeding, etc

### Unique Identifier
- @ Birth / Arrival in UK
- NHS Number
- Child, Mother, Father

### Primary care
- GPRD, EMIS, THIN et al.
- Child health records
- Immunisations [COVER]

### Hospital Care
- Records
- Hospital episodes
- Community ‘tail’ services
- Mental Health Minimum Data
- Mental Health Minimum data
- Fertility (NHS/private) Genitourinary medicine
- Infant Feeding, etc

### Diagnostic / Imaging
- Ultrasound / X-ray [PACS]
- Mammography
- Cytology / Pathology
- Haematology
- Chemical Pathology
- Virology / Microbiology
- Blood Transfusion
- Screening programs
- HPA data

### Prescribers / databases
- Cancer registers
- Diabetes registers
- Renal registers
- Congenital rubella syndrome register
- Congenital hypothyroidism registry
- Cerebral palsy registers
- Down syndrome register
- HIV database
- Newborn screening databases
- NICOR
- Juvenile chronic arthritis
- Inflammatory bowel disease
- Dysmorphology
- Database
- Rare disorders
- Public Health Observatories

### Cohorts / Biobanks
- 1946 1958 1970 Millennium
- ALSPAC, ELSA
- Midspan, Aberdeen, Walker
- Generation Scotland
- UK Biobank
- Newborn Biobank

### Environment
- UK Air Quality Archive
- Environmental Agency [Landfill]
- Drinking Water Inspectorate
- Geological Survey
- NIS data [mobile phone masts]
- Superoutput areas / small area microdata

### Social Care
- Child Protection, In Care / Adopted
- Elderly Care

### Income & Benefits
- Benefits, Housing, Income

### Education & Employment
- Preschool / day care
- Special Educational Needs
- Pupil Level Annual School Census [PLASC] eg SATS scores
- GCSE, GCE, Higher education
- Occupations and Employment

---

**Not available in US or Scandinavia**

But many UK national record sources are not linked
Multiple Record Linkages...needs expansion across NICOR registries

The CALIBER platform
Four nationwide EHR sources linked

Patient’s experience

Primary care
- Healthy, GP registration
- Stable angina
- Pneumonia hospitalization
- Myocardial infarction hospitalization
- See GP for follow-up
- Death

Hospitalization (HES)
- Admit / discharge dates.
- Primary diagnosis: Viral pneumonia, not elsewhere classified
- Admit / discharge dates.
- Primary diagnosis: Acute myocardial infarction
- Procedure: Percutaneous coronary intervention

Disease Registry (MINAP)
- ECG, cardiac markers.
- Diagnosis: STEMI

Death Census (ONS)
- Date of death. Cause: 1) Rupture of abdominal aortic aneurysm 2) Old myocardial infarction

What does linked record data look like?
To get at big data...need tools
How to define phenotypes using multiple EHR data sources?
1001, 2000-01-01, af_gprd=1
1231, 2012-03-03, af_hes=3
1121, 2013-05-04,
af_procs_gprd=1
1511, 1993-01-11,
heart_valve_gprd=2
9913, 2012-05-21, af_hes=1
67222, 1994-08-11, af_hes=1
682444, 1993-01-01,
heart_valve_hes=2
af=1,
af_diag_source="primary care" af_diag_date=2001-12-01
AF algorithm

1. Patient

2. AF diagnosis in GPRD or HES?
   - YES
   - NO

3. History or monitoring terms in record?
   - YES
   - NO

4. Only history or monitoring terms in record?
   - YES
   - NO

5. Recorded prior to diagnosis code?
   - YES
   - NO

6. Historical diagnosis
### Atrial Fibrillation

**Name**  
af

**Chapter**  
Circulatory disease/Atrial fibrillation

**Definition**  
Diagnosis of atrial fibrillation.

**Data Type**  
Categorical

**Data sources**  
GPRD, HES

**Dictionaries**  
Read, ICD10, BNF, Free text

**Authors**  
K. Morley (UCL), Shah A. (UCL), Patel R. (UCL), Liam Smeeth (LSHTM), R. Schilling (St Bartholomews & The Royal London Hospital), R. Hunter (St Bartholomews & The Royal London Hospital)

**Agreed**  
01/02/2013 (Revision 1)

**Category**  
Definition

<table>
<thead>
<tr>
<th>Source variables</th>
<th>Description</th>
<th>Source</th>
<th>Variable</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Atrial fibrillation diagnosis</td>
<td>Primary care</td>
<td>at_gprd</td>
</tr>
<tr>
<td></td>
<td>Atrial fibrillation diagnosis</td>
<td>Secondary care</td>
<td>af_hes</td>
</tr>
<tr>
<td></td>
<td>Atrial fibrillation procedures</td>
<td>Primary care</td>
<td>af_proc_gprd</td>
</tr>
<tr>
<td></td>
<td>Atrial fibrillation procedures</td>
<td>Secondary care</td>
<td>af_proc_opcs</td>
</tr>
<tr>
<td></td>
<td>AF medication</td>
<td>Primary care</td>
<td>at_drugs_gprd</td>
</tr>
<tr>
<td></td>
<td>warfarin or digoxin prescription</td>
<td>Primary care</td>
<td>at_warfarinDigoxin</td>
</tr>
<tr>
<td></td>
<td>Deep vein thrombosis</td>
<td>Primary care</td>
<td>dvt_gprd</td>
</tr>
<tr>
<td></td>
<td>Deep vein thrombosis</td>
<td>Secondary care</td>
<td>dvt_hes</td>
</tr>
<tr>
<td></td>
<td>Pulmonary embolism</td>
<td>Primary care</td>
<td>pe_gprd</td>
</tr>
<tr>
<td></td>
<td>Pulmonary embolism</td>
<td>Secondary care</td>
<td>pe_hes</td>
</tr>
<tr>
<td></td>
<td>ECG Text/Notes text mining</td>
<td>Secondary care</td>
<td>Algorithm</td>
</tr>
</tbody>
</table>
CALIBER Data Portal

• Online data discovery tool caliberresearch.org
• Access to all CALIBER phenotypes, algorithms and implementation details and scripts (SQL,R, Stata)
  – 45 users, 4 institutions, 538 phenotypes, >15,000 clinical diagnostic codes curated
• Standardization
  – Frontend is ICD10, backend becoming SNOMED-CT, LOINC
• A community rather than a static resource
  – Researchers contribute phenotypes and algorithms
  – Other researchers validate/enhance/correct them
Why me?
Why me?
What’s wrong that big data might help fix?
Cardiovascular diseases global #1

- Cause of death/premature death/disability adjusted life years
- So what has gone wrong?
  - wrong prevention
  - wrong treatments
  - wrong diagnoses / wrong names for diseases
  - wrong health systems (and too costly)
  - wrong relations to data, information and knowledge
  - wrong relations with patients

- wrong science! (done by the wrong people!)
Might big data help right these wrongs?
Yes!

Mike Lauer, Director Division of Cardiovascular Sciences National Institutes of Health

“US models are being eclipsed by non-US studies that are much larger, yet considerably less expensive”
Yes!
Eric Topol ‘wireless and genomic medicine’
Pace and scale of translation

Discovery

Trials

Outcomes & quality research

Public health and clinical decisions ➔ health gain

Big data / Health records

Discovery

Trials

Outcomes & quality research

Public health and clinical decisions ➔ health gain

Big data / Health records
Discovery
Genomics

500k participants, 47 baseline biomarkers and custom gene array data available in 2014, cardiac and brain imaging in 100k underway

Open access

Scalable approaches to disease phenotypes (startpoints or endpoints) based on linked electronic health record resources

- cardiac
- diabetes
- stroke
- cancer

Example of Farr Institute working across Wales, Scotland and England
Discovering new risk factor associations: CVD aggregates vs specific diseases
Are the risk factors the same?
To answer this question reliably we need

- **Scale:** e.g. >1 million adults followed for 5 years

- **Phenotypic resolution:**
  - Baseline risk factors
  - Follow up for disease outcomes

Cost to research funder of such data collection?
The research costs are substantial

Information governance
Store, share, harmonise, analyse EHR data.....with scalable tools

And develop pool of clinical expertise
Blood pressure and incidence of twelve cardiovascular diseases: lifetime risks, healthy life-years lost, and age-specific associations in 1·25 million people

Eleni Rapsomaniki, Adam Timmis, Julie George, Mar Pujades-Rodriguez, Anoop D Shah, Spiros Denaxas, Ian R White, Mark J Caulfield, John E Deanfield, Liam Smeeth, Bryan Williams, Aroon Hingorani, Harry Hemingway

Summary

Background The associations of blood pressure with the different manifestations of incident cardiovascular disease in a contemporary population have not been compared. In this study, we aimed to analyse the associations of blood pressure with 12 different presentations of cardiovascular disease. [A: we have added a study aim here. Please amend if you wish]

Methods We used linked electronic health records from 1997 to 2010 in the CALIBER (CArdiovascular research using Linked Bespoke studies and Electronic health Records) programme to assemble a cohort of 1·25 million patients, 30 years of age or older and initially free from cardiovascular disease, a fifth of whom received blood pressure-lowering treatments. We studied the heterogeneity in the age-specific associations of clinically measured [A: OK?] blood pressure with 12 acute and chronic cardiovascular diseases, and estimated the lifetime risks (up to 95 years of age) and cardiovascular disease-free life-years lost adjusted for other risk factors at index ages 30, 60, and 80 years. This study is registered at ClinicalTrials.gov, number NCT01164371.
Higher resolution epidemiology: blood pressure and 12 cardiovascular diseases

Cohort N ≈ 2 million adults, >100,000 disease events

Myocardial infarction

Abdominal aortic aneurysm

Confirms what we know from combining multiple expensive studies

New knowledge

...a challenge for experimental medicine

Rapsomaniki et al, CALIBER The Lancet 2014
Years of life lost to CVD

Rapsomaniki et al, CALIBER Lancet 2014;383(9932):1899-911
Cumulative life time risk of 12 cardiovascular diseases

Rapsomaniki et al, CALIBER Lancet 2014;383(9932):1899-911
Inverse, null, weak and strong…what’s the ‘risk factor’?

‘Higher resolution’ approaches: implications

- Disease mechanism
- Trial design
- Screening and risk prediction
Discovery

Trials
Developing informatics platforms for stratified trials

- **Rapid feasibility**
  - EHR-based eligibility counts

- **Recruiting**
  - EHR randomisation
  - UCLP eConsent

- **Following up & safety**
  - Real-time outcome dashboards
  - Embedded eCRF

Protocol -> EHR data sources -> EHR-based eligibility counts
Thrombus Aspiration during ST-Segment Elevation Myocardial Infarction

Ole Fröbert, M.D., Ph.D., Bo Lagerqvist, M.D., Ph.D., Göran K. Olivecrona, M.D., Ph.D., Elmir Omerovic, M.D., Ph.D., Thorarinn Gudnason, M.D., Ph.D., Michael Maeng, M.D., Ph.D., Mikael Aasa, M.D., Ph.D., Oskar Angerås, M.D., Fredrik Calais, M.D., Mikael Danielewicz, M.D., David Erlinge, M.D., Ph.D., Lars Hellsten, M.D., Ulf Jensen, M.D., Ph.D., Agneta C. Johansson, M.D., Amra Kärengren, M.D., Johan Nilsson, M.D., Ph.D., Lotta Robertson, M.D., Lennart Sandhall, M.D., Ivar Sjögren, M.D., Ollie Östlund, Ph.D., Jan Harnek, M.D., Ph.D., and Stefan K. James, M.D., Ph.D.

METHODS

We conducted a multicenter, prospective, randomized, controlled, open-label clinical trial, with enrollment of patients from the national comprehensive Swedish Coronary Angiography and Angioplasty Registry (SCAAR) and end points evaluated through national registries. A total of 7244 patients with STEMI undergoing PCI were randomly assigned to manual thrombus aspiration followed by PCI or to PCI only. The primary end point was all-cause mortality at 30 days.

RESULTS

No patients were lost to follow-up. Death from any cause occurred in 2.8% of the
Discovery

Trials

Outcomes

research/real

world evidence
Temporal resolution

....with ‘big data’ can study both

Onset

Prognosis

Healthy → Onset of first cardiovascular disease → Second cardiovascular disease, death
Outcomes research: capturing clinically meaningful complexity
one startpoint to many types of endpoint

PROGRESS 4 article series in BMJ / PLoSMed 2013
Acute myocardial infarction: a comparison of short-term survival in national outcome registries in Sweden and the UK

Sheng-Chia Chung, Rolf Gedeborg, Owen Nicholas, Stefan James, Anders Jeppsson, Charles Wolfe, Peter Heuschmann, Lars Wallentin, John Deanfield, Adam Timmis, Tomas Jernberg, Harry Hemingway

UK: 30-day risk of death 10.5% (95% CI 10.4-10.6)
Sweden: 30-day risk of death 7.6% (95% CI 7.4-7.7)

Number at risk:
- UK: 390,948, 370,883, 362,830, 357,954, 354,475, 351,919, 349,845

Figure 3: Kaplan-Meier curves for cumulative mortality at 30 days after admission with acute myocardial infarction in Sweden and the UK
‘Real world’ prognosis of stable CAD (n=102,023) and 5 yr risk of coronary death + non-fatal MI (n=8,856)

A ‘gold standard’ for estimating relevant risks?

Rapsomaniki et al, CALIBER, EHJ 2014;35(13):844-52
Prognostic models using linked EHR: Which clinically recorded factors add to discrimination?

Origin of data is EHR therefore implementation of risk prediction models in decision support tools (with evaluation) is feasible

Rapsomaniki et al. CALIBER, EHJ 2014;35(13):844-52
Discovery

Trials

Outcomes research/real world evidence

Public Health
Outcomes assessment: importance of linking multiple record sources

Death registry

Disease registry

Hospital admissions

Primary care

All four data sources

Crude annual incidence of myocardial infarction per 100,000

Herrett et al, CALIBER, BMJ 2013;346:f2350
How does CVD first present?
In the real world, today

- MI/Fatal CHD: 18%
- Ischaemic stroke: 13%
- Stable angina: 12%
- Heart failure: 12%
- Transient ischaemic attack: 11%
- Peripheral arterial disease: 11%
- CHD: 10%
- Unstable angina: 5%
- Abdominal aortic aneurysm: 2%
- Ventricular arrhythmia/sudden cardiac death: 3%
- Intracerebral haemorrhage: 2%
- Subarachnoid haemorrhage: 1%
- Abdominal aortic aneurysm: 2%

N=1.93 million patients
>110K CVD events
5 year median follow-up

CALIBER 2014, under review
What is the role of the Farr Institute?
Drought

• **Data**
  – Need much wider national record linkages – CPRD-NICOR-HES
  – Need to liberate ‘submerged’ deeper hospital phenotypes
  – Need to converge EHR, omics and imaging

• **Tools**
  – Health informatics ‘20 years behind bioinformatics’
  – And UK 20 yrs behind US?

• **People**
  – (re) building **public trust** (care.data)
  – Not nearly enough **clinicians** with the training and opportunity to drive improvements in care (and research) through data (cf new US sub-specialty)
  – Careers for technical staff
  – **Interdisciplinarity**
Strengthening health informatics research

- MRC coordinated 10-partner £19m call for e-health informatics research centres across the UK

Cutting edge research using data linkage capacity building

- Additional £20m capital to create Farr Institute

- UK Health Informatics Research Network

Coordinate training, share good practice and develop methodologies

Engage with the public, collaborate with industry and the NHS

- Farr London
- Farr Scotland
- Farr at Swansea, Wales
- Farr N8, Manchester
What are the aims?
= research along the translational pathway

Discovery
Proof of concept (Experimental medicine)
Clinical Trials
Quality and outcomes research
Public health and Health gain

Cardiometabolic  Infection  Mother and Child Health  Phase 2: Cancer, Neuroscience, Eyes, Musculoskeletal

Farr Tools: Informatics methods
Farr People: Capacity development
Farr Curated data: Research-ready cohorts
Rapid evolution of initiatives: emphasis on infrastructure

**JULY 2013**
- Farr Institute of Health Informatics Research
- Genomics England

**JULY 2013**
- Medical Bioinformatics Awards

**OCTOBER 2013**
- Health Informatics Collaboration
  (sharing hospital data across 5 Biomedical Research Centres)

**February 2014**
- Medical Bioinformatics Awards
Who was William Farr?

“Diseases are more easily prevented than cured and the first step to their prevention is the discovery of their exciting causes.”

1807-1883

Compiler of Scientific Abstracts at General Register Office
...aka ‘Big data health’

Gave us cause of death and International Classification of Disease

Local actions e.g. Victoria Park
Conclusion

• Most of what we know about mortality and morbidity has come from much ‘smaller’ data than is currently available to researchers

• Personalisation is a secular phenomenon across multiple sectors in society: Medicine offers vanguard and laggard examples!

• If informatics is about data, tools and people – then it is the people which need most urgent development.
Farr London (original) Investigators

CARDIOVASCULAR
- Mike Barnes, Director of Bioinformatics
- James Carpenter, Professor of Medical Statistics
- John Deanfield, Professor of Paediatric Cardiology
- Mark Caulfield, Professor Clinical Pharmacology
- Spiros Denaxas, Health Informatics Senior Research Associate
- Nicholas Freemantle, Professor of Clinical Epidemiology and Biostatistics
- Harry Hemingway, Professor of Clinical Epidemiology
- Aroon Hingorani, Professor of Genetic Epidemiology
- Steffen Petersen, Reader in Advanced Cardiovascular Imaging
- John Robson, GP, Clinical lead for the Clinical Effectiveness Group
- Liam Smeeth, Professor of Epidemiology
- Adam Timmis, Professor of Clinical Cardiology

INFORMATICS
- Anne Blandford, Professor of Human–Computer Interaction
- Peter Coveney, Professor of Physical Chemistry
- James Freed, Head of Health Intelligence and Standards
- Dipak Kalra, Professor of Health Informatics
- John Shawe-Taylor, Professor of Computing
- Paul Taylor, Reader in Health Informatics
- Alan Wilson, Professor of Urban Regional Systems

MOTHER & CHILD
- Peter Brocklehurst, Professor of Women's Health
- Tito Castillo, Chief Operating Officer, LIFE Study
- Carol Dezateux, Professor of Paediatric Epidemiology
- Ruth Gilbert, Professor of Clinical Epidemiology
- Irene Petersen, Senior Lecturer Epidemiology and Medical Statistics
- Judith Stephenson, Professor of Reproductive and Sexual Health
- Phil Koczan, Chief Clinical Information Officer
- Irwin Nazareth, Professor of Primary Care and Population Science
- Max Parmar, Director of MRC Clinical Trials Unit

INFECTION
- Mike Catchpole, Head of Epidemiology and Surveillance
- Andrew Hayward, Senior Clinical Lecturer in Infection
- Richard Pebody, Head of the Seroepidemiology Programme
- Deenan Pillay, Professor of Virology

PHASE 2 CLINICAL WORKSTREAMS
- Andy Goldberg, Senior Lecturer in Trauma and Orthopaedics
- Anthony Moore, Professor of Ophthalmology
- Kathy Pritchard-Jones, Professor of Paediatric Oncology
- Martin Rossor, Professor of Neurology & Director of DeNDRON