Some questions about big health data, and a project

Jeremy Wyatt DM FRCP ACMI Fellow

Professor of Digital Healthcare & Director, Wessex Institute of Health, University of Southampton

Clinical adviser on New Technologies, Royal College of Physicians, London

j.c.wyatt@soton.ac.uk
Big Health data: Asthmopolis
# Some sources of big health data

<table>
<thead>
<tr>
<th>Level</th>
<th>Examples of data collected</th>
</tr>
</thead>
<tbody>
<tr>
<td>Government</td>
<td>Open data, census data, fiscal / admin data, care.data</td>
</tr>
<tr>
<td>National organisation</td>
<td>Performance metrics; data assets (eg. HES, OS map data)</td>
</tr>
<tr>
<td>Supermarket</td>
<td>Purchases of fruit &amp; veg, alcohol, contraceptives, over the counter medicines; clothing sizes, kid’s ages, pregnancy…</td>
</tr>
<tr>
<td>NHS organisation</td>
<td>Service activity, adverse events, staffing,</td>
</tr>
<tr>
<td>NHS department</td>
<td>Case-mix, outcomes; drugs prescribed, radiology / pathology images; dictated text from reports, letters…</td>
</tr>
<tr>
<td>Professional</td>
<td>Coded patient data; procedures saved to logbook</td>
</tr>
<tr>
<td>NHS client</td>
<td>NHS encounters; personal monitoring data</td>
</tr>
<tr>
<td>Individual</td>
<td>Step count, quant. self data</td>
</tr>
<tr>
<td>Device</td>
<td>Bluetooth encounters; GPS trail; cell phone call record</td>
</tr>
<tr>
<td>Infomediary eg. Google, Experian</td>
<td>Online purchase history; credit worthiness; consumer category; investor category…</td>
</tr>
</tbody>
</table>
UK big data infrastructure relevant to healthcare

Farr Institute:
- (£40M) eHERCs: Scotland, HeRC Manchester, CIPHER Swansea, UCL
- (£39M) Medical Bioinformatics Centres: Leeds, Oxford Big Data Inst., MRC/UVRI Uganda, Warwick & Swansea, UCL Partners
- Working groups on capacity building, innovative governance, public engagement, e-infrastructure

Other:
- ESRC National Centre for Research Methods, Networks for Methodological Innovation
- (£10M) Innovate UK Open Data Institute Southampton / London / Leeds etc.
- (£42M + £25M) EPSRC Alan Turing Institute ("Make UK a world leader in analysis & application of big data and algorithm research"): BL, Cambridge, Edinburgh, Oxford, Warwick and UCL
Some key challenges

1. Data quality
2. Biases and methods for reliable inference
3. Confidentiality / trusted research platforms
4. Data analyst capacity
5. Communicating the results
Challenge 1: Data quality

• First law of health informatics: data collected for one purpose can rarely be used for another (Johan van der Lei, 1989)

• Numerator issues:
  – Local data definitions & conventions on clinical code usage
  – Completeness varies with time of day
  – Undocumented changes in data definitions, collection process, normal lab ranges, thresholds for payment...

• Denominator issues: changes in cohort type, coverage, follow up process; missing record = litigation...

• Lack of metadata
Leeds study of accuracy of Hospital Episode Statistics (HES)
Work of Rosy Tsopra

Methods:
• Obtain case notes from 105 consecutive hospital discharges from 5 chest medicine wards in Leeds hospitals Trust
• Consultant & coder work together to develop Gold Standard list of coded diagnoses

Preliminary results:
• Median of 11.5 gold standard diagnoses, HES (“official”) coding process: 7 diagnoses
• Correctness of official codes (TP rate): 88%
• Completeness of official codes (PVP): 43%

Implication: you will only find about half the cases you are looking for through HES
Diabetes prevalence in UK primary care? It depends on which database you check, and how...
Challenge 2: Biases and reliable inference

- If you carry out 20 analyses, at least one will show $p = 0.05$
- Association is not causation, even if we want it to be (Kahneman: Thinking Fast & Slow, 2012)
- Many biases when analysing health data: Simpson’s Paradox, confounding by indication, immortal time bias...
How do Mexican lemons prevent highway fatalities?

![Graph showing the correlation between fresh lemons imported from Mexico and US highway fatality rate.](source)

Source:
www.cqeacademy.com/cqe-body-of-knowledge/continuous-improvement/quality-control-tools/
### Simpson’s Paradox: mortality in diabetes

<table>
<thead>
<tr>
<th></th>
<th>Type 1</th>
<th>Type 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall mortality</td>
<td>64% of 358</td>
<td>97% of 544</td>
</tr>
</tbody>
</table>

Data from Poole Diabetes cohort, cited by Julious et al BMJ 1994
Confounding by indication

You are given a large cancer database to analyse:

• 40% of patients treated with a new drug survive 5 years versus 30% of patients treated with the old drug.
• Difference persist despite taking into account differences in age, baseline cancer severity, genetic markers...
• Conclusion: the new drug reduces mortality by 10%.

However:

• Maybe allocation to the new drug depends on the doctor’s intuition on who will survive (subtle predictive feature not recorded in any database).
• So, receipt of new drug is marker of better outcome - not cause.
Estimating survival benefit of Ezetimibe in 2233 all-cause deaths in heart attack survivors using routine GP data

Eg. First incident MI; missing cholesterol levels; medication covariates

Some solutions

• Understand data flows & quantify the biases
• Include the *source of data* in the model
• Expertise in analytical methods: life course epidemiology, multi-level modelling, functional data analysis for episodic frequent data…
• Explore “novel” [for health] non-experimental study designs: case-crossover, regression discontinuity, instrumental variable analysis…
Regression discontinuity design

- Some drugs / procedures are applied according to a test result or predictive model
- People just above & just below an allocation threshold are very similar
- If you have enough people to compare, you can *estimate* the impact of the intervention
- Eg. chemotherapy on older women – RCT failed to recruit

Thistlethwaite & Campbell, 1960
Challenge 4: Data analyst capacity

Big data is a new health discipline, requiring **both** fluency in a range of technical areas:

- Data warehouses, safe havens & software (I2B2, Briskit...)
- New analysis methods and tools
- Data coding, quality and biases
- Information governance and research ethics...

And good domain understanding – or analysts may misunderstand questions, miss serendipitous findings...

“We’re creating great datasets, but don’t have enough scientists to analyse them”

(NASA asteroid data contest, 11-3-14)
Predictive models are drifting away from reality

* Euroscore predictions

* European System for Cardiac Operative Risk Evaluation
So, what kinds of question can Big health Data safely answer?

Descriptive questions:
• Rates of symptoms, diseases, investigations, treatments [Google Flu debacle]
• Severity of illness, results of tests, doses of therapy etc.
• Distribution of services, diseases, risks etc. [Asthma map]

Questions about association:
• Predictive modelling to aid targeting of services & drugs

Causative questions?
Asthma as a health problem

What is it?
Reversible narrowing of small airways in lung causing wheeze, cough or shortness of breath.

What causes it?
Genetic predisposition + allergy, viral infection, air pollution eg. vehicle emissions, soldering fumes (colophony asthma).

How common is it?
- 5.4 million people in UK currently receive treatment for asthma.
- It affects 1.1 million children (1 in 11) and 4.3 million adults (1 in 12).

Why does it matter?
- Every 10 seconds someone is having a potentially life-threatening asthma attack in the UK.
- Every day, the lives of three families are devastated by the death of a loved one from an asthma attack.
- Tragically, two thirds of these deaths are preventable.
The project: data analytics for web intelligence and content personalisation

Collaborator: a major UK asthma charity attracting 150,000 website visitors per month (70% new, 30% repeat)

Project aim: to understand and improve user experience and offer each user a tailored selection of the rich content (inc. videos) from its website

Data: anonymised Google Analytics data based on 2 year longitudinal cookies

Methods: explore page pathways, traffic sources & destinations, trends and other insights

Confidentiality: no ethics required but students will sign a non-disclosure agreement (results of interest to both voluntary and commercial sectors)

Project outputs: report, analytical tools, algorithms or software for charity staff

Patient benefits: improved knowledge about asthma, confidence in making self-management decisions

Summary: this project provides teams with excellent introduction to web analytics in voluntary sector, with strong potential to improve human health.