Why anonymity fails

Ross Anderson
Cambridge University

Synopsis

- Health data are moving to the cloud, causing serious tussles over safety and privacy
- The extension of the open data idea to healthcare is now a slow-motion train wreck
- Everyone from drug companies to insurers want access to masses of personal data
- Yesterday: we learn that HSCIC gave hospital episode statistics data to over 1000 firms
- Patients can often be easily identified

Do patients want access to records?

- Google Health discontinued in 2011 after four years trying to build a health platform
- Microsoft equivalent turned into a business platform for hospitals etc
- Healthspace, a project to provide patient access to summary records, had only a handful of users
- The penny drops: most people are healthy 95% of the time and not interested in looking at record
- When sick they mostly have other priorities

Big Pharma certainly wants access

- In 1998 a startup (DeCODE) offered Iceland's health service free IT systems in return for access to records for research
- Funding was from Swiss drug company Roche
- Records to be 'de-identified' by encrypting the social security number, but would be linked to genetic, family data
- Icelandic Medical Association got 11% of citizens to opt out
- Eventually the supreme court ruled the system should be opt-in, and the scheme collapsed

European case law

- European law based on s8 ECHR right to privacy, clarified in the I v Finland case
- Ms I was a nurse in Helsinki, and was HIV+
- Her hospital's systems let all clinicians see all patients' records
- So her colleagues noticed her status – and hounded her out of her job
- Finnish courts wouldn't give her compensation but Strasbourg overruled them
- Now: we have the right to restrict our personal health information to the clinicians caring for us

Recent UK history

- Tony Blair ordered a “National Programme for IT” in the NHS in 2002
- Idea: replace all IT systems with standard ones, giving “a single electronic health record” with access for everyone with a “need to know”
- This became the biggest public-sector IT disaster in British history
- Billions wasted, suppliers dropped out, huge lawsuits, and the flagship software didn’t work

Scope Creep

- We've had big tussles over 'shared care'
- E.g. giving social workers access to GP records in Oxford has made young mums there reluctant to discuss post-natal depression
- Lobbying win: after the 2010 election, we killed the "childrens' databases" designed to share data between health, school, probation and social work ('Database State', Munro review)
- The NHS Information Centre now wants to revive the idea, but under its control

Public Opinion

- 2,231 adults asked October 2006 on central records database with no opt out:

strong support	12%
tend to support	15%
neither	14%
tend to oppose	17%
strongly oppose	36%
don't know	6%

- Several surveys since say the same: most don't want wide sharing, or research use without consent
- And there's the Catholic Bishops' Conference

Secondary Uses

- Cameron policy announced January 2011: make 'anonymised' data available to researchers, both academic and commercial, but with opt-out
- We'd already had a laptop stolen in London with 8.63m people's anonymised records on it
- In September 2012, CPRD went live – a gateway for making anonymised data available from (mostly) secondary care (now online in the USA!)
- From this year, GPES hoovering up GP stuff
- So: how easy is it to anonymise health records?

Advocating anonymisation

Transparency

Inference Control

- Also known as ‘statistical security’ or ‘statistical disclosure control’
- Started about 1980 with US census
- Before then only totals & samples had been published, e.g. population and income per ward, plus one record out of 1000 with identifiers removed manually
- Move to online database system changed the game
- Dorothy Denning bet her boss at the US census that she could work out his salary – and won!

Inference Control (2)

- Query set size controls are very common. E.g. in New Zealand a medical-records query must be answered from at least six records
- Problem: tracker attacks. Find a set of queries that reveal the target. E.g for our female prof's salary
  - 'Average salary professors'
  - 'Average salary male professors'
- Or even these figures for all 'non-professors'!
- On reasonable assumptions, trackers exist for almost all sensitive statistics

Inference Control (3)

- Contextual knowledge is really hard to deal with! For example in the key UK law case, Source Informatics (sanitised prescribing data):

	Week 1	Week 2	Week 3	Week 4
Doctor 1	17	21	15	19
Doctor 2	20	14	3	25
Doctor 3	18	17	26	17

Inference Control (4)

- Perturbation – add random noise (e.g. to mask small values)
- Trimming – to remove outliers (the one HIV positive patient in Chichester in 1995)
- We can also use different scales: practice figures for coronary artery disease, national figures for liver transplants
- Random sampling – answer each query with respect to a subset of records, maybe chosen by hashing the query with a secret key

Inference Control (5)

- Modern theory: differential privacy (pessimistic)
- Practical problem in medical databases: context
- 'Show me all 42-yo women with 9-yo daughters where both have psoriasis'
- If you link episodes into longitudinal records, most patients can be re-identified
- Add demographic, family data: worse still
- Active attacks: worse still (Iceland example)
- Social-network stuff: worse still
- Paul Ohm's paper: "Broken Promises of Privacy"

CPRD

- The clinical practice research datalink, run by the MHRA, makes some data available to researchers (even to guys like me :-)
- Freedom of information request for the anonymisation mechanisms
- Answer: sorry, that would undermine security
- Never heard of Kerckhoffs?
- Search for me, cprd on whatdotheyknow.com

Next problem – care.data

- The PM promised in 2011 our records would be anonymised, and we'd have an opt out
- The Secretary of State for Health, Jeremy Hunt, assured us in March 2013 that existing opt-outs would be respected
- In July this was reversed by the NHS England CIO
- NHS opt-outs are like Facebook's: the defaults are wrong, the privacy mechanisms are obscure, and they get changed whenever too many people learn to use them

The row over HES

- Hospital Episode Statistics (HES) has a record of every finished consultant episode going back 15 years (about a billion in total)
- Mar 13: formal complaint to ICO that PA put HES data in Google cloud despite many rules on moving identifable NHS data offshore
- Apr 3: HSCIC reveals that HES data sold to 1200 universities, firms and others since 2013

The row over HES

- Some HES records have name, address, …
- Some have only postcode, dob, …
- Some have this removed but still have "HESID" which usually contains postcode, dob
- Even if the HESID were encrypted, what about cardioversion, Hammersmith, Mar 19 2003?
- Yet the DoH describes pseudonymised HES data as "non-sensitive"!

HES data bought by …

- 40–42, 46–47, 62–66, 95–98, 159–162 … : selling data outside the NHS
- 191–2, 321, 329, 331, 362 … drug marketing
- 329, 336, … medical device marketing
- 408: Imperial College with HESID, date of birth, home address, GP practice: still marked "non sensitive”
- Many: market anaysis, benchmarking, efficiency…

The big tussle in Europe

- Data Protection Regulation currently making its way through the Europarl
- Attempt to exempt medical data (art 81, 83)
- You'll be deemed to consent to secondary use and forbidden to opt out retrospectively, or even claim that consent was coerced
- Most lobbied ever law in Europe with 3000+ amendments from big pharma, researchers …
- Looks like it will be stalled till after election
- But I v Finland is still case law

Now add DNA

- The UK Department of Health is launching a '100,000 genomes' project to use genetic analysis in both direct care and research
- All sequence data centralised; if you don't consent to unlimited research use (including sharing with 23andme) then no treatment
- The FDA just stopped 23andme from offering health advice to new customers
- In the UK, a Nuffield Bioethics Council inquiry

Take-away

- Think safety and privacy, not 'security'
- Scale matters! A national system with 50m records is too big a target (even 5m)
- Governance failure has real safety costs
- Privacy failings limit access to healthcare, especially for the vulnerable
- Similar debates in the USA, Norway, Austria…
- Above all we need honesty – we need to stop pretending that pseudonyms protect privacy

Snowden?

‘When you discover that a paraplegic Canadian woman was denied entry to the USA after a border-guard accessed a database that revealed she'd once been suicidally depressed, it's easy to see how you – or someone you love – might suffer far-reaching consequences even from accurate data used for the purpose it for which it was intended.’ – Cory Doctorow, Guardian