Methods for protection of privacy in the 2015 Charter for Safe Havens in Scotland for handling unconsented data from NHS patient records: a critical look

Paul McKeigue

Usher Institute of Population Health Sciences and Informatics

Concerns about using unconsented deidentified health care records for research

- Davidson (2012): study of public acceptability
  - participants doubted that removal of identifiers would protect against "anyone with the necessary know-how".
  - participants were prepared to trust academic researchers but not industry ones
- NHS England (2016): plans to create an opt-out database accessible to industry rearchers (*care.data*) were abandoned in the face of opposition
  - reidentification had been described as "a small residual risk", noting that "such an attack would be illegal".
  - Presser (2015): "dysfunctional chronology: care.data was initially poised to be launched under the radar without democratic consultation or diverse viewpoints, then was subjected to multiple bodies of regulation in order to stay afloat, then came under increasingly greater scrutiny due to distrust."

2015 Charter for Safe Havens in Scotland: recommendations

- Unconsented datasets should be available only through "safe haven" data warehouses.
- Linked datasets should be kept only for the "minimal time necessary"
- Analytical outputs should be manually checked for "statistical disclosure" before the user is allowed to download them

Problems

- reduced productivity
  - blocking transfer of data from host to client entails blocking clipboard.
  - manual checks for disclosure control cause delay and add to costs
- Long-term research platforms based on linked data cannot be constructed on the fly from raw data
  - examples: cohort studies of chronic disease such as diabetes, drug safety studies
  - cleaning the linked dataset and deriving new variables takes years of work by researchers with domain-specific expertise
- Does it protect privacy against attacks that might be realistically be mounted?

How can privacy be protected when making health care data available for research

- Technical security of platform
  - authentication, encrypted connections, limits on data transfer from host to client (can be bypassed).
- Trust in the integrity and competence of the researcher
  - accreditation of researchers, compulsory training courses on data security
- Deidentification of the data
  - Typically removes identities, geographic identifiers with population less than 10^5, day and month of birth
  - free text records and images may require additional scrubbing
  - more stringent rules require removal of day and month of all dates: makes studies of short-term effects such as adverse drug reactions impossible

Other measures for protecting privacy: adding noise

- Apply "differential privacy" algorithm (Dwork 2006) to queries
  - differential privacy means that the information that attacker can gain about an individual who was included in the dataset is not appreciably more than it would have been if that individual had been excluded
  - implementation is based on adding noise to the results of each query
  - standard algorithms protect against arbitrary side information - unnecessarily stringent for health records

Other measures that transform or encrypt the data

Problem with all these is that researchers need to be able to inspect and clean the raw data before analysis

- Synthetic datasets that have the same joint distribution of variables as the original dataset
- secure distributed computing - user does not have direct access to linked data
- homomorphic encryption - analysis is done on encrypted data (computationally burdensome)

Risks to privacy in deidentified datasets

- *Reidentification attacks*
  - user with access to the individual-level data may be able to reidentify individuals
- *Attribute disclosure attacks*
  - aggregate data released for publication may leak information about attributes of the target individual.

Both types of attack rely on using *side information*: the attacker knows, for the target individual, the values of some variables that are in the deidentified dataset.

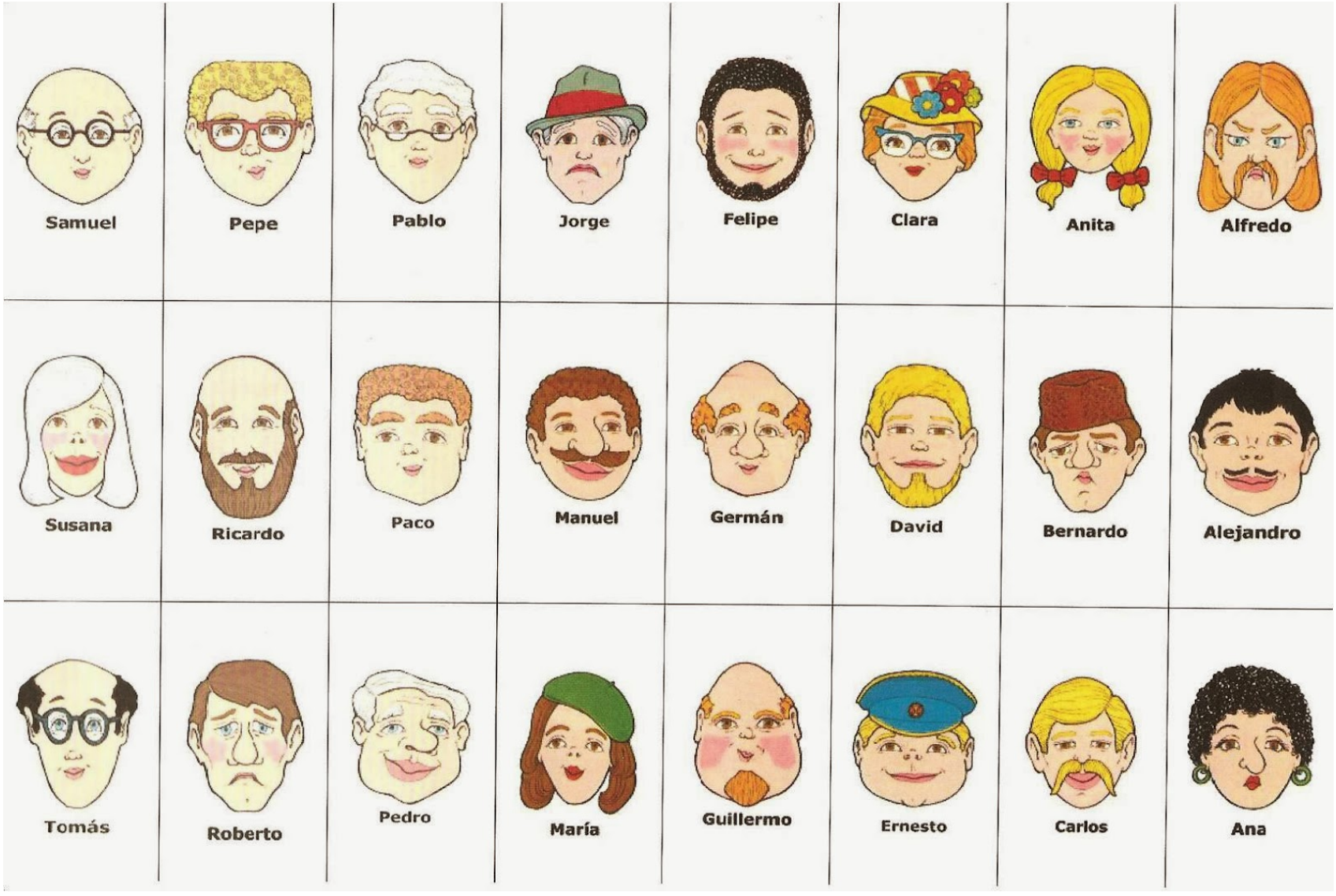
Using side information for a reidentification attack: a real example

- Naraynan (2008): individuals in the Netflix Prize dataset of movie rentals could be identified by the dates of reviews they had posted in the Internet Movie Database.
  - 6-8 movie ratings were enough for reidentification
  - movie viewing history is sensitive information

Information theory and entropy

- Uncertainty is quantified by *entropy*
  - information is quantified as reduction in entropy
- Privacy can be quantified as the entropy of a probability distribution, measured in *bits*.
  - If there are 2^H records equally probable as matches to the target individual, the entropy of that individual's identity is H bits
  - for adult in Scotland, entropy of an identity is $\sim$22 bits
  - To maintain protection against re-identification, entropy of the probability distribution over identities should be at least 3 bits ($2^3 = 8$ equally probable matches to the target individual).

A guessing game with 4.6 bits of entropy



Side information in deidentified health records

- Variables most likely to be used for side information - gender, geographical location (health board), and year of birth - contain about 12 bits of information.
  - attacker would need to gain an extra 10 bits of information for successful reidentification.
  - A single hospital admission date, which might be available to an attacker, contains about 10 bits of information.
- Linkage of health records to social data (e.g. arrest records) opens up more possibilities for an attacker to exploit side information

Adding noise to reduce information leaked

- Adding a random offset to all dates on a given individual would help to protect against attribute disclosure attack.
- For identity X, and a side variable Y, the information about X leaked by knowing Y is the mutual information $\mathrm{I}(X, Y)$
- if we replace Y by a random variable Z obtained by adding noise to Y, the information leaked is reduced by $\mathrm{I}(X; Y|Z)$: the mutual information between identity X and and side information Y given proxy variable Z$
- to reduce the information leaked by a date by 4 bits, we need to add an average perturbation of about ± 5 days to all dates in each individual's record
    - discrete Laplace distribution maximizes entropy for given mean absolute value of offset.

NHS National Services Scotland Statistical Disclosure Control Protocol (2017)

- For frequency tables of a "sensitive" variable, cell values less than 5 may not be shown
- this rule does not take account of whether the other variables used for classification are likely to be available to an attacker.
    - for instance a frequency table of disease status is likely to leak information if the other classifying variables are age and gender, but not if they are lab test results.
- Where does the "rule of five" come from?
    - NSS protocol is based on rules used by Office of National Statistics (England)
    - Statistics of Trade Act (1947): published tables should not contain any cells with fewer than five businesses.

Where statistical disclosure control can fail: an extreme example

- Homer (2008): summary statistics from genetic case-control studies, based on thousands of individuals, can leak information allowing an attacker to establish the disease status of an individual who was in the study if the attacker has access to the individuals genotypes
  - genetic case control study with N cases and N controls
  - for each variant, the summary effect size estimate leaks about $0.7/N$ bits of information for discriminating case-control status
  - usually $N < 10,000$ and number of typed variants $>$ 300,000
  - Unlikely that attacker would have access to individual's genotypes unless that individual has provided samples to a personal genome company.
- General point: simple rules about minimum numbers in table cells cannot substitute for quantifying information leak.

Conclusions: beyond the "one-size-fits-all" approach

Table 1: Example of a flexible framework that distinguishes low-risk from high-risk linkages, and trusted from untrusted researchers

	Health only	Health + social
Core/trusted	Direct access	Locked-down platform
Less trusted	Locked-down platform	Synthetic data, distributed computing, homomorphic encryption