Can Anonymous data exist in an era of massive information availability?

Prof. Chris Dibben
University of Edinburgh

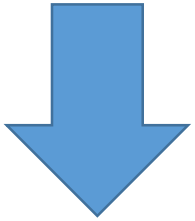
THE UNIVERSITY
of EDINBURGH

Probably not –

but are we thinking about anonymisation in a sensible way?

Anonymisation

Jane is a 39 year old female, suffering from diabetes, who presented herself for treatment on 24th May.



A 39 year old female, suffering from diabetes, presented herself for treatment on 24th May.

In Law

- US, ‘personally identifiable information’ (PII)
- The European Union’s (EU) data protection regime incorporates a category of ‘personal data’,

defined as data from which the subject of the data is identifiable, either on its own or in tandem with auxiliary pieces of data

Importance of anonymous data

- The principles of data protection should therefore not apply to anonymous information, namely information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable

‘Broken Promises’

- The value of anonymisation as a concept and a practical process has been questioned in recent years
- Ohm 2010, Rubinstein and Hartzog 2016, Narayanan and Shmatikov 2010

Simple model of anonymisation is somewhat unrealistic

- 2006, AOL released the search histories of 650,000 users
- Users were pseudonymised by replacing user names with numbers
- *The New York Times* investigated the data to discover the identities of some of the subjects, and one such person agreed for her identity to be published in a story

"A Face is Exposed for AOL searcher No. 4417749"

- "numb fingers" "60 single men" "dog that urinates on everything"
- "landscapers in Lilburn, Ga" "homes sold in shadow lake subdivision gwinnett county Georgia"
- Thelma Arnold, a 62-year old widow Lilburn, Ga.

Source: New York Times, 2006 Aug 9th

Information available to the adversary

- Ohm asserts, ‘data can be useful or perfectly anonymous but not both’ (Ohm 2010, 1704)
- If Ohm’s claim that data are either useful or anonymous but not both were true, then it would seem that sharing data to extract their potential value requires a reckless attitude to risk.

Where does this leave us?

- *formal anonymisation* - clear the original data of direct identifiers - **inadequate**
- *absolute anonymization* – likely to render a whole class of research impossible

- Is there another option?

Where to go from here?

- focus on the ***empirically plausible*** rather than the ***logically possible*** ones.
- required of us by law: 'To determine whether a natural person is identifiable, account should be taken of all the means **reasonably likely** to be used' (GDPR Recital 26).

Functional anonymisation

- Other data
- Data users
- Governance processes
- Infrastructure

Hazard vs Risk

Fundamental issue

- A trade-off to be made
- between the social (or commercial) value of sharing data,
- and some risk of identifying people, even if that trade-off has consequences for personal privacy

- Publically acceptable

Functional anonymisation

- Other data
- Data users
- Governance processes
- Infrastructure

The risk of re-identification is a function of several components

- The *motivation* of an adversary wishing to attack anonymised data in order to re-identify somebody within it (this will affect what happens and how).
- The potential *consequences* of disclosure (which will affect the motivations of an individual to attempt a re-identification, and the cost-benefit analysis of the data controller).
- How a disclosure might happen without malicious intent (the issue of *spontaneous identification*).
- The *governance structures*, *data security* and other *infrastructural* properties surrounding the release/sharing of the data (this will affect the risk).
- The *auxiliary data/knowledge* that could be linked to the data in question (without which disclosure or identification is impossible).
- *Divergence* between the data in question and the other data/knowledge (even if they overlap in content, there may be differences and lack of fit due to alternative semantic encoding, error, quality, differences in measurement, differences in calculation, and so on).

Two “man” rule

Whether data are anonymous or not (and therefore personal or not) is a function of the relationship between those data and their environment.

Black and white guarantee

Risk minimisation balanced against benefit

Motivation

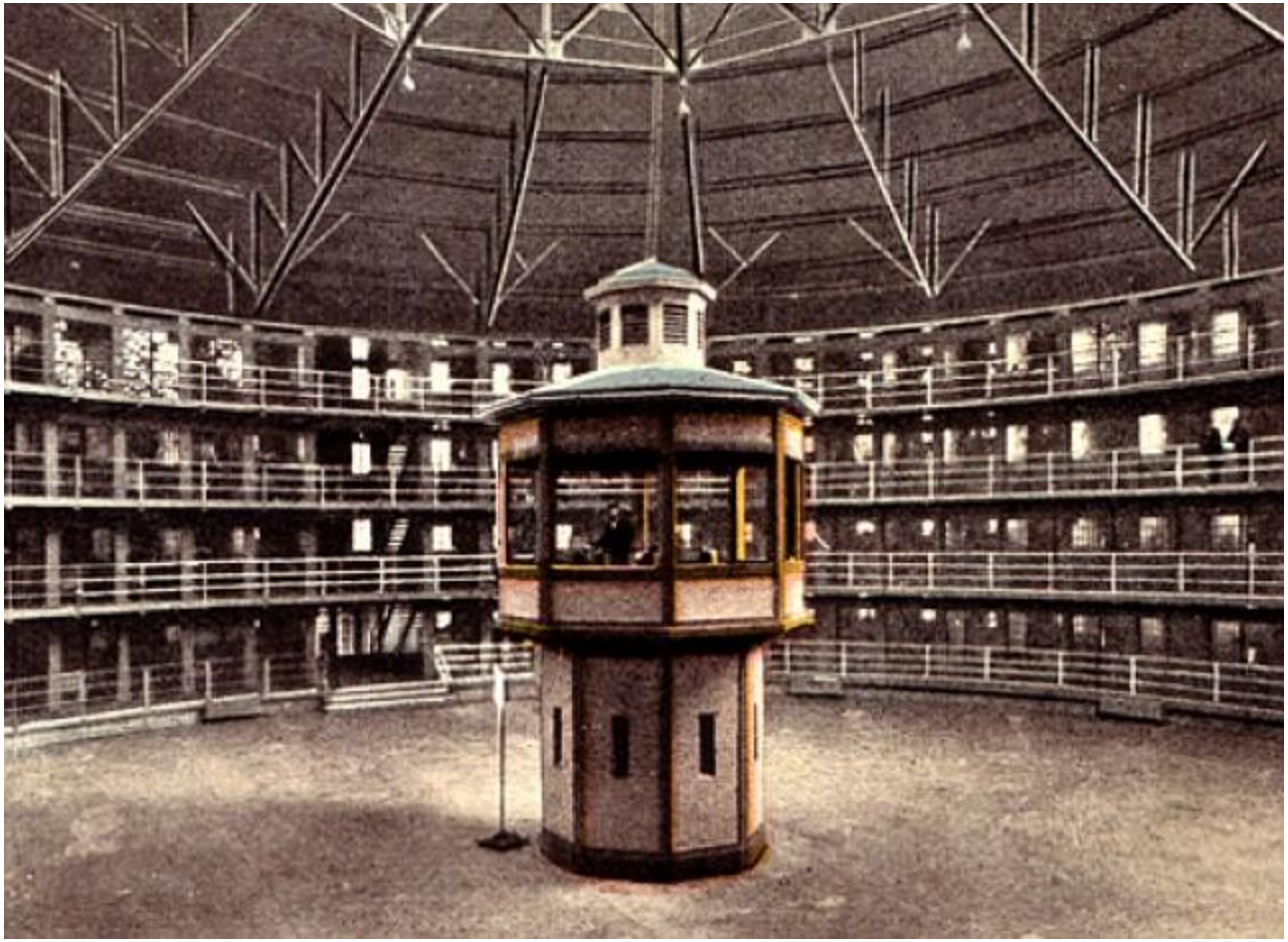
- Linked research administrative/ big data for public benefit research
- Where consent is disproportionately difficult to attain

Administrative Data Research Network

Producing regulated spaces of research

- Production of a ‘regulated space’ - Space in which the data exists, is:
- (firstly) only accessible to individuals who are felt to be very unlikely, due to character, professional position and training, to attempt to gain this information
- (secondly) allows an ‘unequal gaze’ to exist, where there is the constant *possibility* of observation

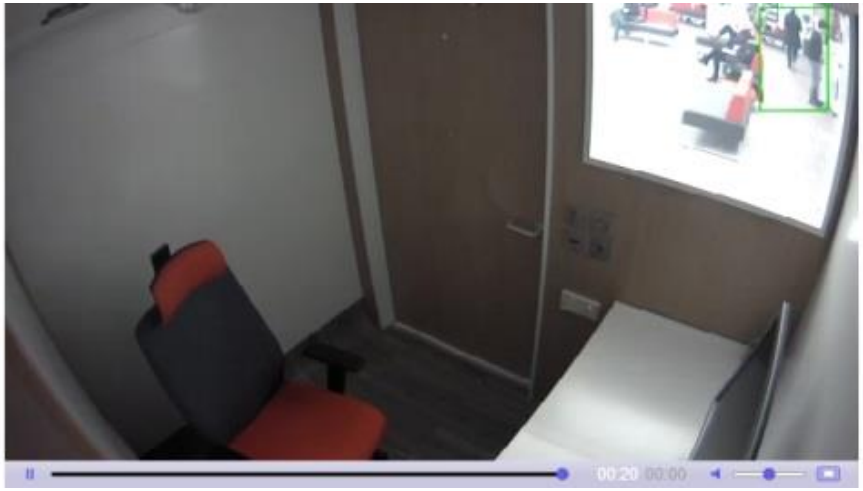
‘unequal gaze’ - Panoptican



Safe settings, shaping behaviour and producing regulated spaces

- Controlled access (ie it is not in a public space) to the space housing the work space.
- Methods for controlling who has access to or is in the work space.
- Methods for monitoring behaviour in the workspace – the ‘gaze’ of the research centre.
- That the remotely executed work is only visible to those in the work space.
- Strong environmental ‘cues’ that encourage safe behaviour.
- A built structure that needs to resist a deliberate but not a determined malicious intruder (eg an intruder using considerable force or tools).

CCTV camera



View from camera

CCTV camera monitors and records activity remotely. Alerts are sent on motion activation.

Conclusions

- Focusing on only the data - absolute anonymity is probably impossible
- Focusing on empirically plausible threats
- Allows - functional anonymisation
- A workable and robust response?

ELSEVIER

Computer Law & Security Review

Volume 34, Issue 2, April 2018, Pages 204-221

Functional anonymisation: Personal data and the data environment

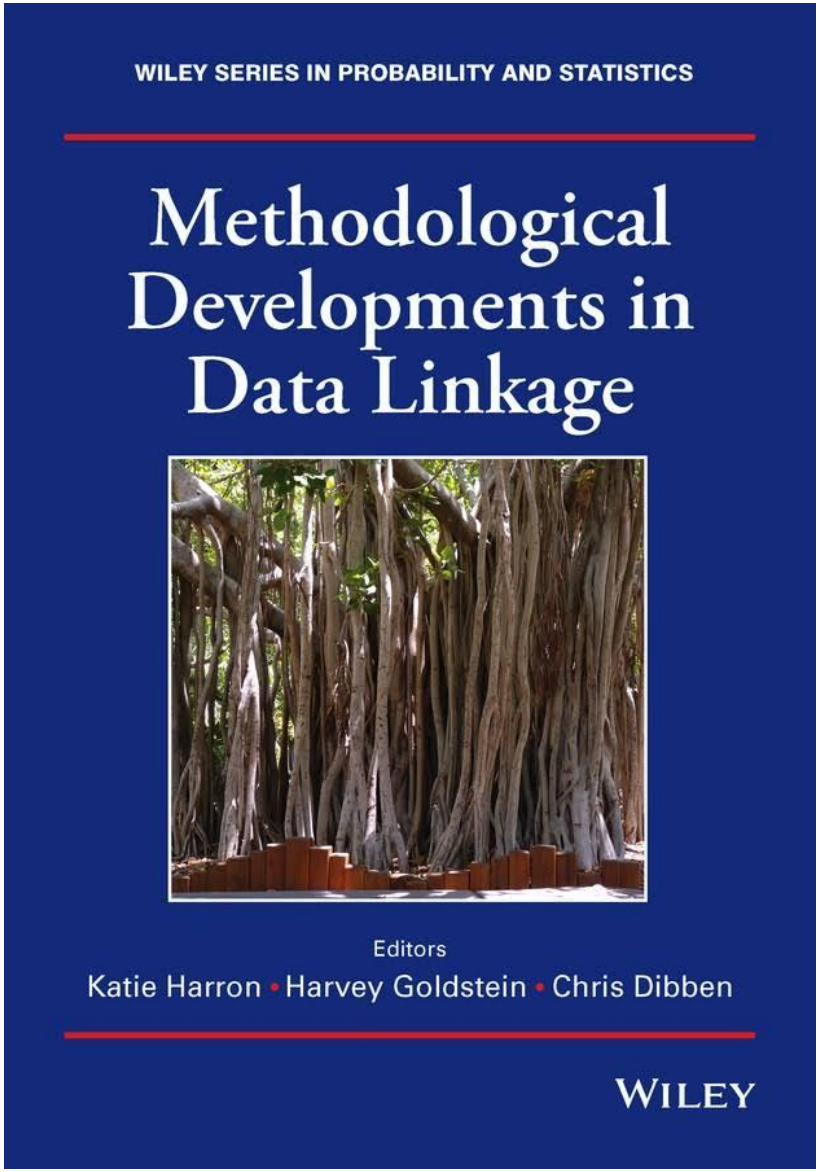
Mark Elliot [a], Kieron O'Hara [b], Charles Raab [c], Christine M. O'Keefe [d], Elaine Mackey [a], Chris Dibben [c], Heather Gowans [e], Kingsley Purdam [a], Karen McCullagh [f]

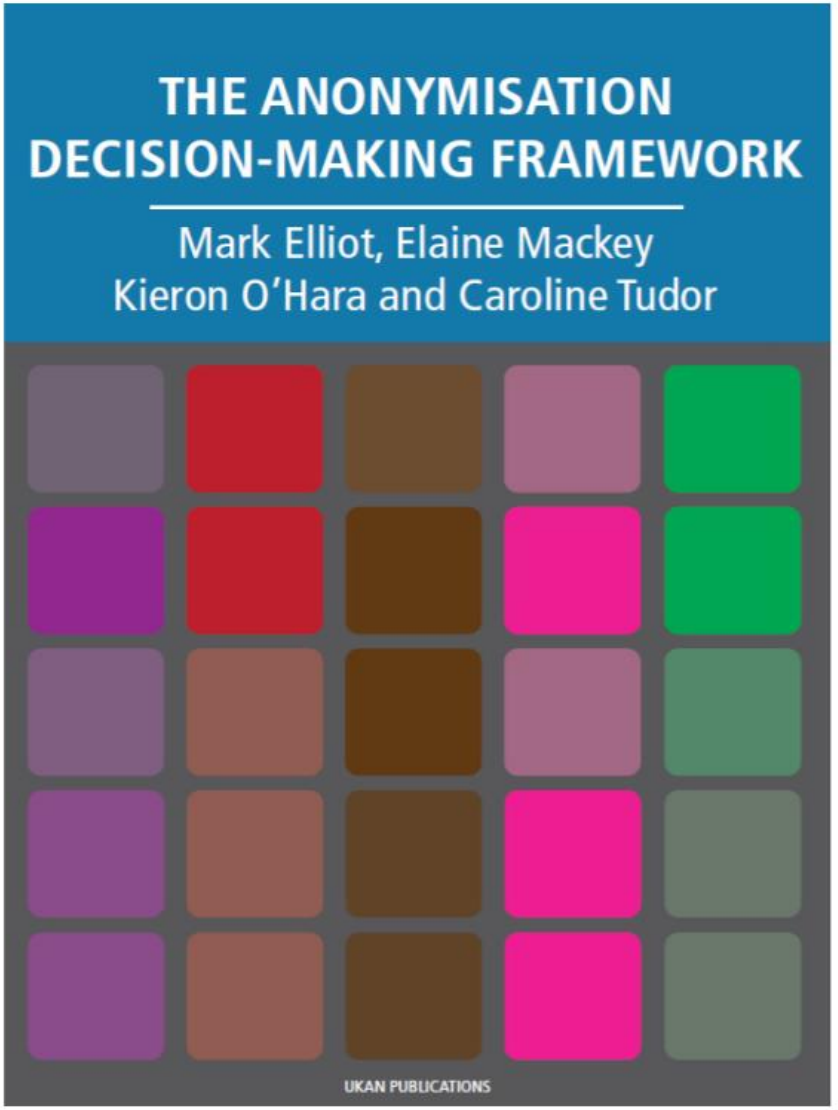
Show more

https://doi.org/10.1016/j.clsr.2018.02.001

Abstract

Anonymisation of personal data has a long history stemming from the expansion of the types of data products routinely provided by National Statistical Institutes. Variants on anonymisation have received serious criticism reinforced by much-publicised apparent failures. We argue that both the operators of such schemes and their critics have become confused by being overly focused on the


WILEY SERIES IN PROBABILITY AND STATISTICS
Methodological Developments in Data Linkage
Editors
Katie Harron • Harvey Goldstein • Chris Dibben
WILEY


THE ANONYMISATION DECISION-MAKING FRAMEWORK
Mark Elliot, Elaine Mackey
Kieron O'Hara and Caroline Tudor
UKAN PUBLICATIONS

Acknowledgements

MARK ELLIOT,[a] KIERON O'HARA,[b] CHARLES RAAB,[c] CHRISTINE M. O'KEEFE,[d] ELAINE MACKEY,[a] CHRIS DIBBEN,[c] HEATHER GOWANS,[e] KINGSLEY PURDAM,[a] KAREN MCCULLAGH[f]

[a]University of Manchester
[b]University of Southampton
[c]University of Edinburgh
[d]CSIRO, Canberra
[e]University of Oxford
[f]University of East Anglia

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

Economic and Social Research Council
Shaping Society

Dave the mind reader

People with autism 'die younger', warns charity

By Dominic Howell
BBC News

18 March 2016 | Health

Top Stories

EU leaders put migrant deal to Turkey
EU leaders hold talks with Turkey's prime minister in an attempt to reach a deal over the migrant crisis.
1 hour ago

No 'concessions' over disability cuts
2 minutes ago

Ben Nevis gains a metre thanks to GPS
18 March 2016

Features

BJPsych The British Journal of Psychiatry (2016)
208, 232–238. doi: 10.1192/bjp.bp.114.160192

Premature mortality in autism spectrum disorder

Tatja Hirvikoski, Ellenor Mittendorfer-Rutz, Marcus Boman, Henrik Larsson, Paul Lichtenstein and Sven Bölte

Background
Mortality has been suggested to be increased in autism spectrum disorder (ASD).

Aims
To examine both all-cause and cause-specific mortality in ASD, as well as investigate moderating role of gender and intellectual ability.

Method
Odds ratios (ORs) were calculated for a population-based cohort of ASD probands ($n=27\,122$, diagnosed between 1987 and 2009) compared with gender-, age- and county of residence-matched controls ($n=2\,672\,185$).

Results
During the observed period, 24 358 (0.91%) individuals in the general population died, whereas the corresponding figure for individuals with ASD was 706 (2.60%; OR = 2.56; 95% CI 2.38–2.76). Cause-specific analyses showed elevated mortality in ASD for almost all analysed diagnostic categories. Mortality and patterns for cause-specific mortality were partly moderated by gender and general intellectual ability.

Conclusions
Premature mortality was markedly increased in ASD owing to a multitude of medical conditions.

Declaration of interest
None.

Table 3 Risk for all-cause mortality for the entire autism spectrum disorder (ASD) group, as well as separately for females and males, and low-functioning ASD and high-functioning ASD groups

	Controls Number of deaths (%)	ASD OR (95% CI) Number of deaths (%)	Low-functioning ASD OR (95% CI) Number of deaths (%)	High-functioning ASD OR (95% CI) Number of deaths (%)
Total	24 358 (0.91)	2.56 (2.38–2.76) 706 (2.60)	5.78** (4.94–6.75) 169 (2.71)	2.18 (2.00–2.38) 537 (2.57)
Females	11 693 (1.39)	2.24 (1.99–2.51) 296 (3.51)	8.52 (6.55–11.08) 61 (3.00)	1.88 (1.65–2.14) 235 (3.67)
Males	12 665 (0.69)	2.87* (2.60–3.16) 410 (2.19)	4.88 (4.02–5.93) 108 (2.57)	2.49 (2.22–2.80) 302 (2.08)

ASD, autism spectrum disorder; OR, odds ratio; CI, confidence interval.
*Partial likelihood ratio test for interaction effect ASD × gender, $P=0.001$.
**Partial likelihood ratio test for model selection (low-functioning ASD/high-functioning ASD), $P<0.001$.