# Anonymisation: a collection of thoughts

Becca Wilson, Newcastle University

Additional slides:
Prof Harvey Goldstein, University of Bristol
Prof Paul Burton, Newcastle University

Newcastle University

# Risk vs. Anonymisation vs. Usefulness

*"Given the development of increasingly powerful data sharing, matching and mining techniques – and a backdrop of strong political and commercial pressure to make more data available - it can seem inevitable that re-identification risk will increase exponentially."*

The Anonymisation Decision-Making Framework

*"Data can either be useful or perfectly anonymous, but never both"*

Paul Ohm

Newcastle University

# The Problem

- Problem: Release of large (pseudonymised) datasets for analysis potentially allows ‘statistical attack’ via searching for records satisfying certain constraints (e.g. age, location, medication)
- Common Solution: degrade data values to make it ‘unlikely’ that an attacker could correctly identify individuals. e.g. k-anonymity is met:

  *the information for each person contained in the release cannot be distinguished from at least k-1 individuals whose information also appear in the release*

# Statistical Disclosure Methods

Non-perturbative methods – reduce information content

- Delete sensitive / identifiable variables
- Remove ‘cells’ with small counts if data in tabular form, preserving margins
- Group/bin categories or categorise continuous variables of disclosive variables such as postcode, age….
- Sub-sample

Newcastle University

# Statistical Disclosure Methods

Perturbative methods – alters the data to increase uncertainty of identification

- Add random ‘noise’ to increase uncertainty around correct identification (this includes random misclassification for categorical variables)
- micro-aggregation of similar cases (effectively reduces variation)
- Create ‘synthetic’ data values while preserving data structure

Newcastle University

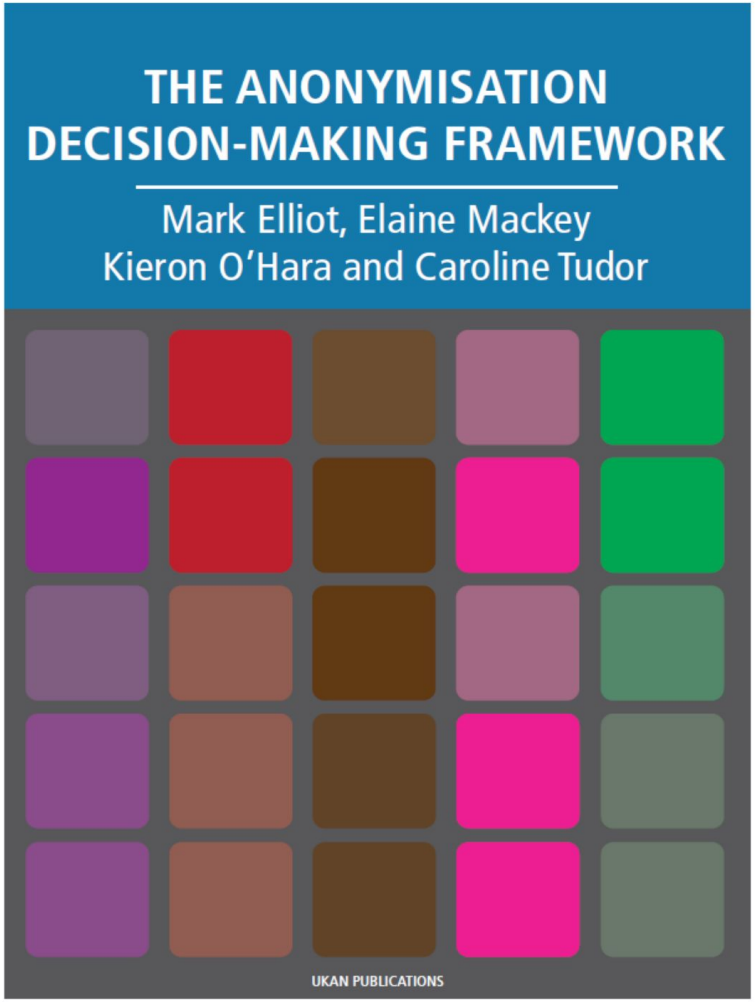
# Effects on statistical analysis: concerns

- Cell removal: may over - coarsen data and in particular remove interesting interaction effects
- Grouping: like above may smooth over complex relationships
- Addition of random noise will lead to incorrect standard errors and also biased coefficients in generalised linear models unless properly adjusted for
- Synthetic data may lead to severely biased coefficients if analysis models do not include variables used in the synthesis

*Perfectly anonymous data is perfectly useless data*

# The Anonymisation Decision-Making Framework

http://bit.ly/UKAnonD



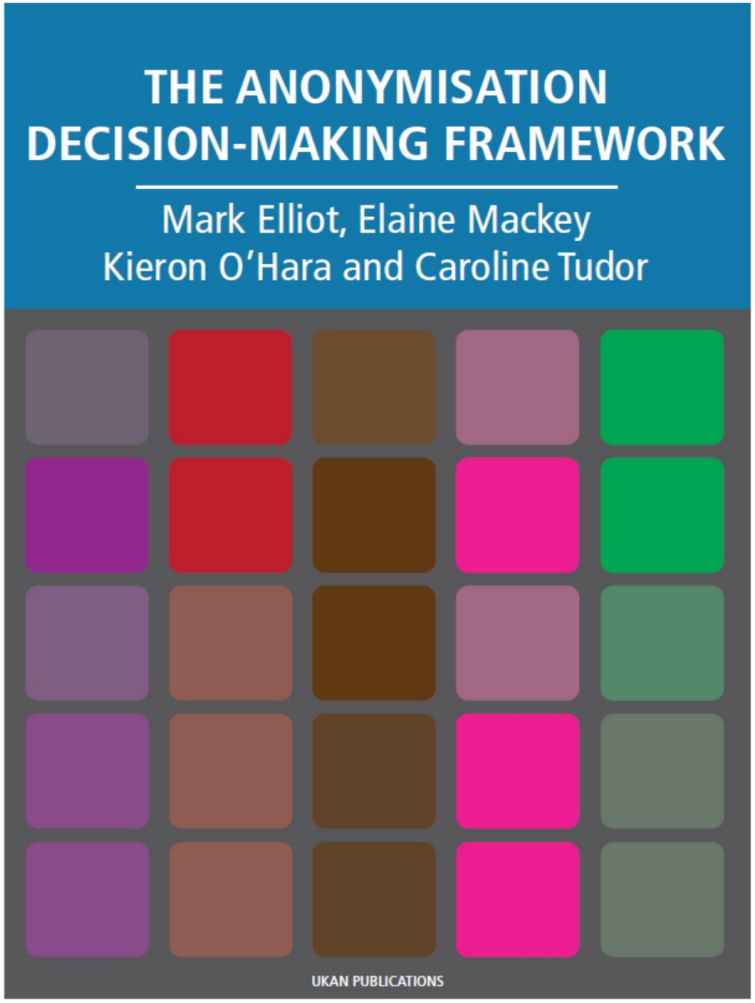

Practical guide to anonymisation for those with data they need to anonymise with confidence, typically in order to share it

Provides operational advice

*"whilst being less technical and forbidding than the statistics & computer science literature"*

# The Anonymisation Decision-Making Framework

http://bit.ly/UKAnonD




*"Our view has always been that anonymisation is a heavily context-dependent process and only by considering the data and its environment as a total system (which we call the data situation)"*

Newcastle University

# The Data Situation: Access

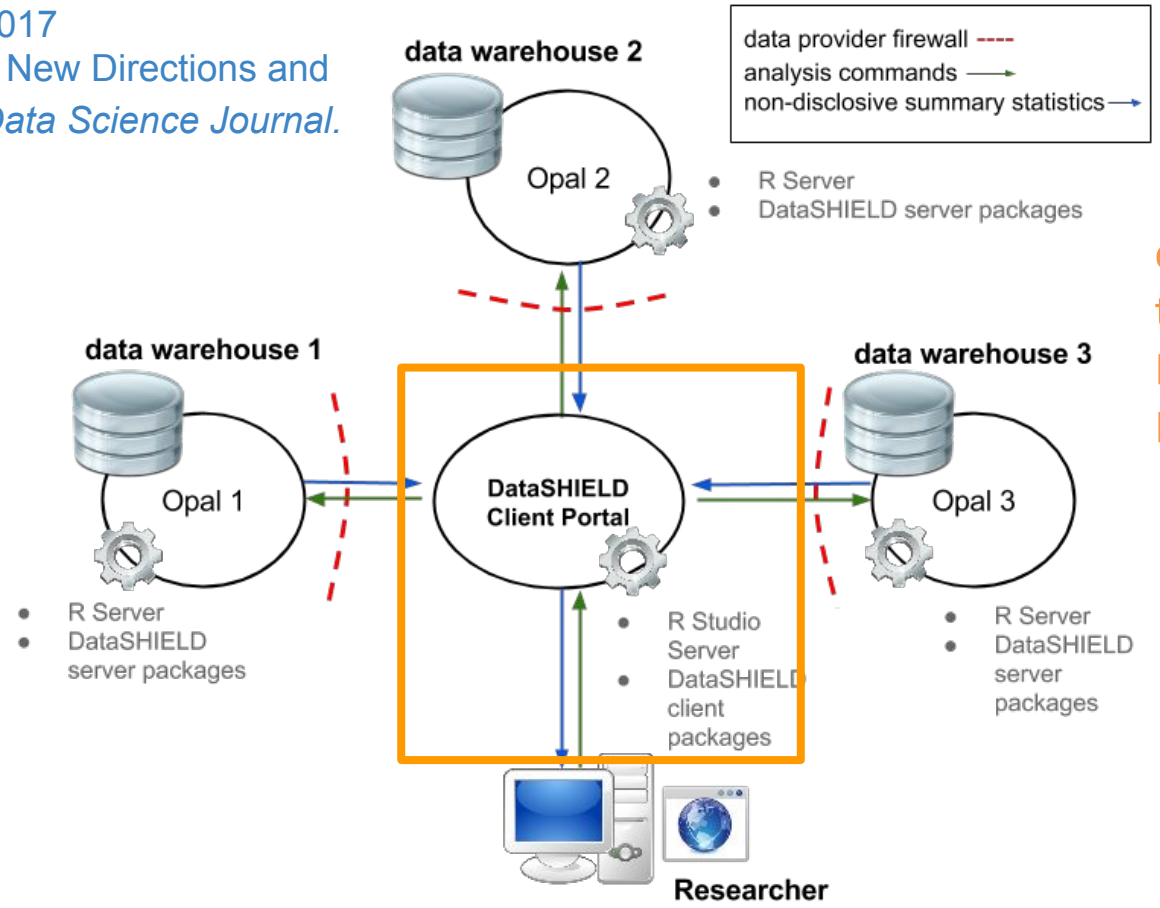
- safe room: analyse one study at a time, can not remove data, outputs checked
- safe haven: typically data anonymised/pseudonymised, what if you need to work on identifiable information e.g location?
- trusted third party: typically data anonymised/pseudonymised, require data sharing agreements, overcome governance restrictions - will they allow deposit of their data?
- Federated analysis?

# Example Data Situation: DataSHIELD

- DataSHIELD (www.datashield.ac.uk) is distributed approach that allows privacy-protected analysis of sensitive individual-level data from one study, and the co-analysis of several
- Data remains at the data owner, analysis taken to the data
- Built-in disclosure controls (computational and statistical approaches)
  - No requirement to manually check outputs of analysis

# Example Data Situation: DataSHIELD

Wilson et al., 2017
DataSHIELD – New Directions and Dimensions. *Data Science Journal.*




client commands sent to Opal via standard REST protocol over HTTPS

No requirement for anonymisation / pseudonymisation of data - but it is recommended

Newcastle University

# Summary

- Anonymisation / pseudonymisation doesn't completely remove the risk of re-identification
- Can destroy the usefulness of the data, degrading the information for the user
- In order to utilise the most appropriate methods and balance the usefulness consideration must be given to the nature of the data and the contextual environment

# The Data Situation