New Zealand

Statistical disclosure risk in data repositories

Shaun Roberts and Barry Milne

COMPASS
RESEARCH CENTRE
FACULTY OF ARTS
THE UNIVERSITY OF AUCKLAND
Whare Wānanga o Tāmaki Makaurau

COMPASS Colloquium
Statistics New Zealand, Wellington
7 August 2019

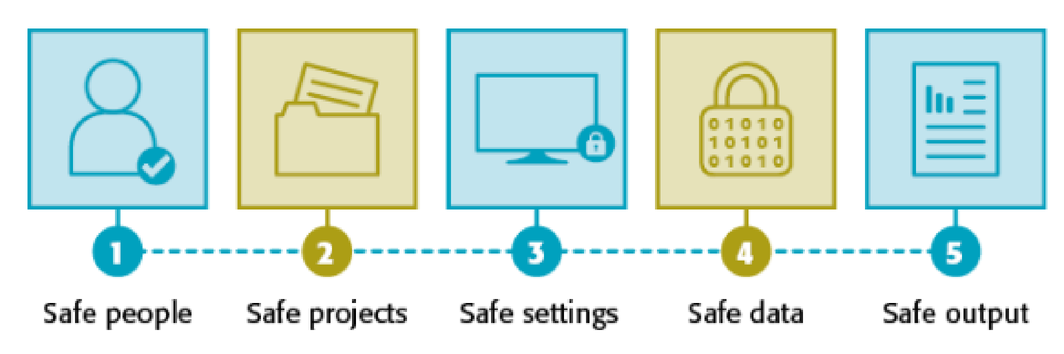
The University of Auckland

Context – Disclosure Risk

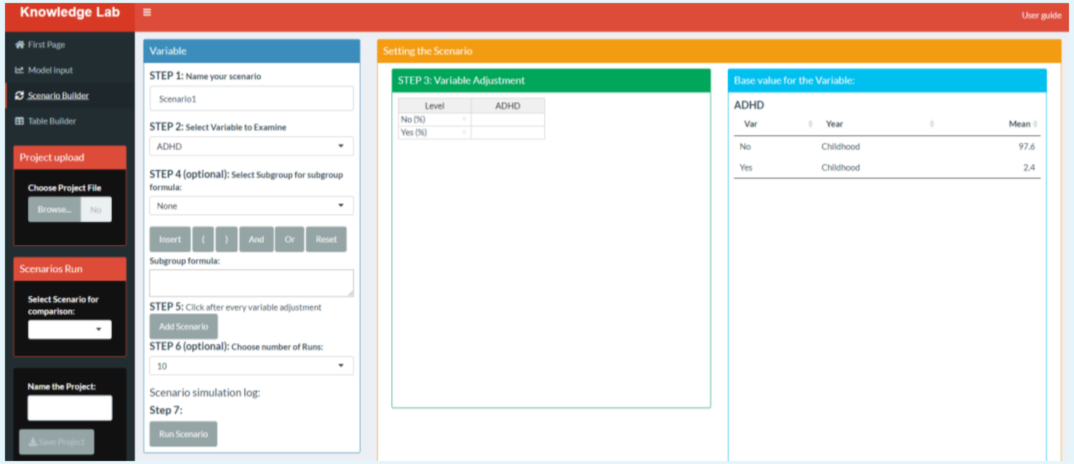
- Statistical offices take Disclosure Risk seriously

  “Risk of identifying a unit in a data file”

  - StatsNZ Microdata



  - Basefile for micro-simulation



Variable

STEP 1: Name your scenario
Scenario1

STEP 2: Select Variable to Examine
ADHD

STEP 4 (optional): Select Subgroup for subgroup formula:
None

Subgroup formula:

STEP 5: Click after every variable adjustment

STEP 6 (optional): Choose number of Runs:
10

Scenario simulation log:

Step 7:

Select Scenario for comparison:

Name the Project:

Setting the Scenario

STEP 3: Variable Adjustment

Level	ADHD
No (%)	
Yes (%)	

Base value for the Variable:

ADHD

Var	Year	Mean
No	Childhood	97.6
Yes	Childhood	2.4

Context – Disclosure Risk

- Identity disclosure
  - Identifying that it is YOU in the datafile. This is bad because it may allow:
- Attribute disclosure
  - Identifying things about YOU that wouldn't otherwise be known
    - May – or may not – be sensitive (a judgement call)

- Direct identifiers
  - Names, addresses, phone numbers
- Quasi-identifiers
  - Anything else that could be used to identify YOU, especially when used in combination
  - Date of birth, gender, location

Context – Disclosure Risk

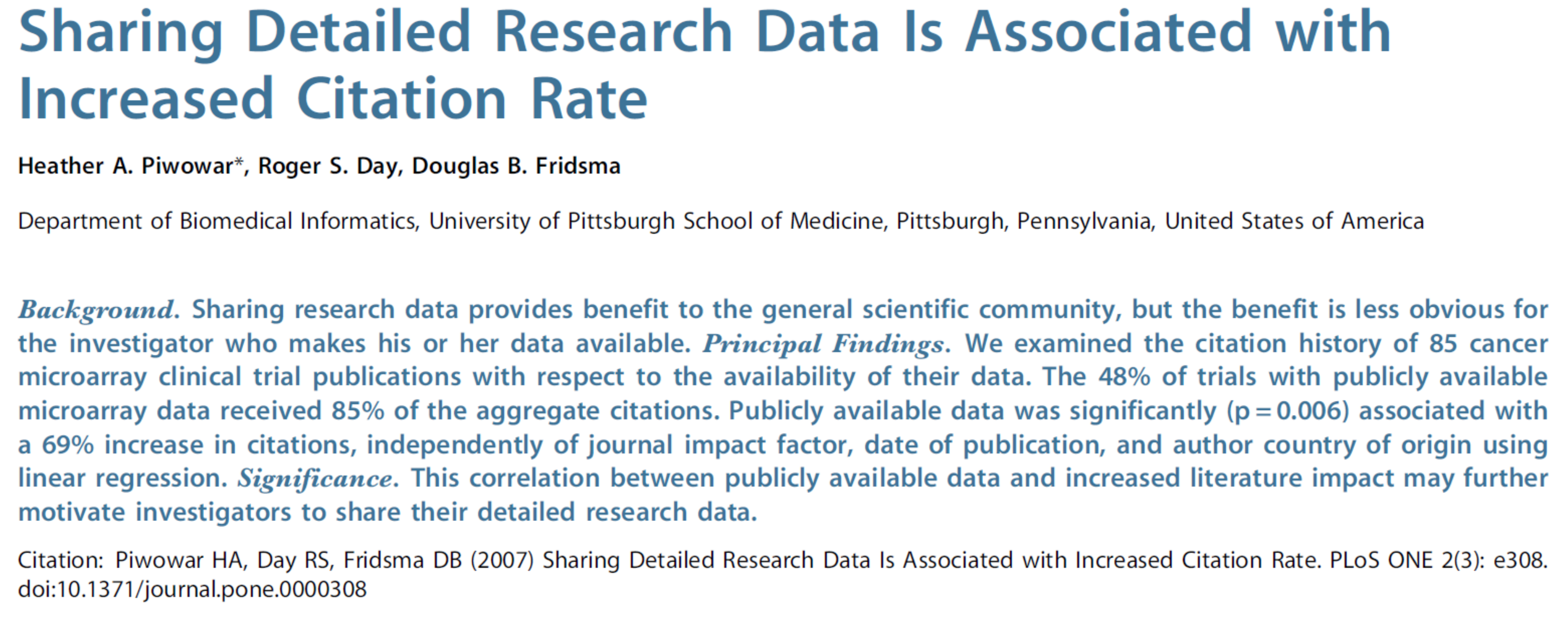
- Anonymous **does not** mean non-identifiable
- Re-identification often possible
  - With some extra information
  - Stratification by many factors
- A real issue in a small place like New Zealand
  - 60–70,000 people of the same age
  - So… ~100 people who share the same date of birth and are male and ~100 people who share the same date of birth and are female
    - Stratifying by other factors may produce n=1 (or close to)
- Failure to recognise this can lead to disaster…

Disclosure disaster

- Massachusetts Group Insurance Commission released anonymised health data on state employees (1997)
  - To enable research to improve healthcare
  - Mass Governor assured public the privacy was protected by removal of identifiers
- MIT CompSci grad student, Latanya Sweeney
  - Accessed the health data
  - Accessed electoral roll ($20) for Cambridge, Mass (where Governor lived), incl name, address, ZIP code, birthdate and sex of every voter
  - Six in Cambridge shared Gov’s birthdate; only three were men and only one lived in his ZIP code
  - She mailed all of the Gov’s health records to him…

Which got me thinking…

- Researchers are being asked to deposit data files in publically available repositories, for re-analysis by other researchers
  - In the interest of ‘open science’



Sharing Detailed Research Data Is Associated with Increased Citation Rate

Heather A. Piwowar*, Roger S. Day, Douglas B. Fridsma

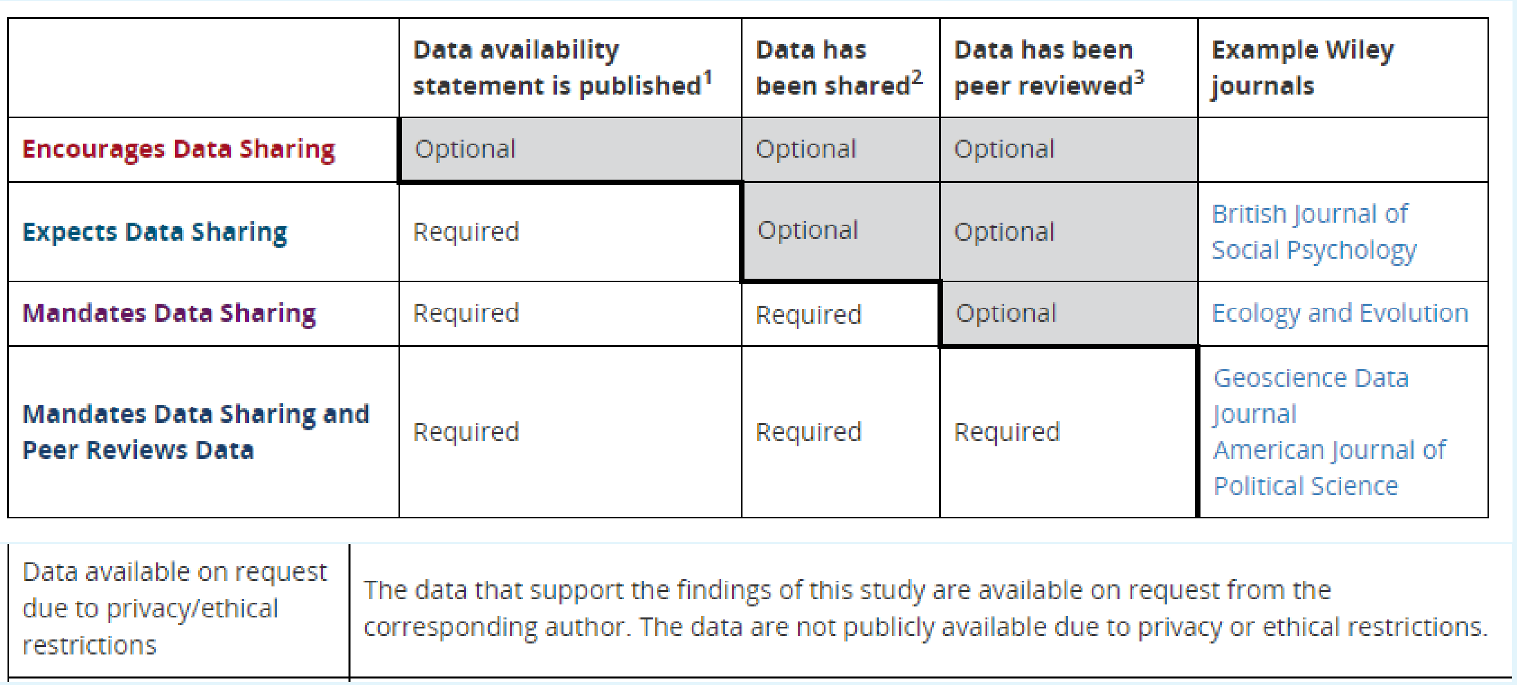
Department of Biomedical Informatics, University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania, United States of America

Background. Sharing research data provides benefit to the general scientific community, but the benefit is less obvious for the investigator who makes his or her data available. ***Principal Findings.*** We examined the citation history of 85 cancer microarray clinical trial publications with respect to the availability of their data. The 48% of trials with publicly available microarray data received 85% of the aggregate citations. Publicly available data was significantly ($p = 0.006$) associated with a 69% increase in citations, independently of journal impact factor, date of publication, and author country of origin using linear regression. ***Significance.*** This correlation between publicly available data and increased literature impact may further motivate investigators to share their detailed research data.

Citation: Piwowar HA, Day RS, Fridsma DB (2007) Sharing Detailed Research Data Is Associated with Increased Citation Rate. PLoS ONE 2(3): e308. doi:10.1371/journal.pone.0000308

Encouraged by journals…

- Wiley
  - “When data is **FAIR** (Findable, Accessible, Interoperable, Reusable), the process becomes more efficient as you can access and analyze each others’ findings and reuse it to inform new findings.”

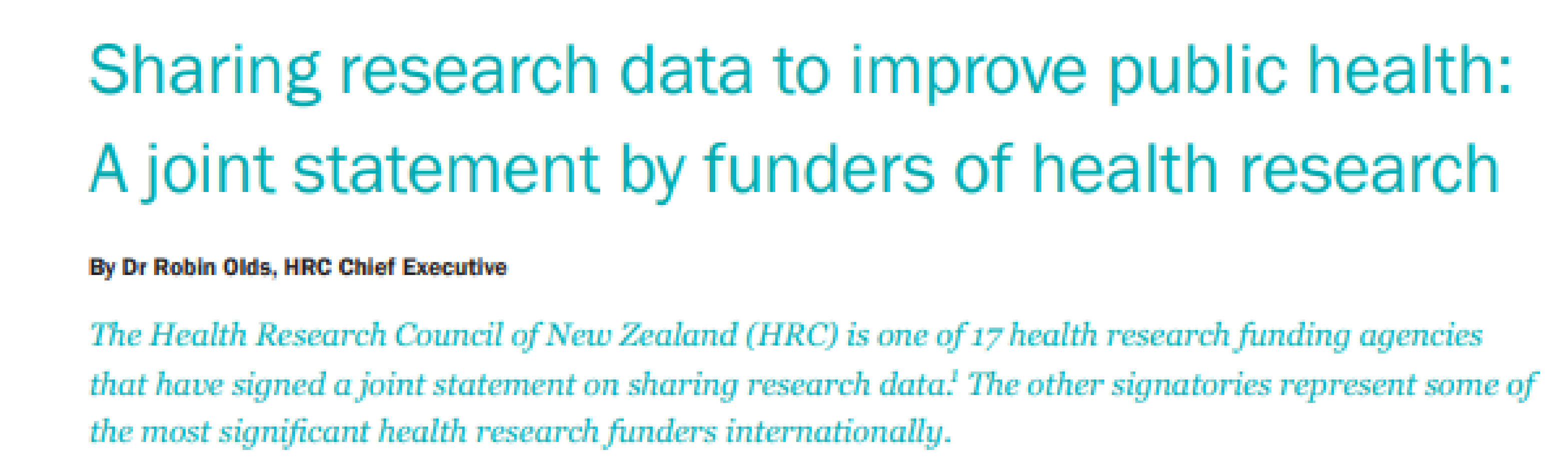


	Data availability statement is published[1]	Data has been shared[2]	Data has been peer reviewed[3]	Example Wiley journals
Encourages Data Sharing	Optional	Optional	Optional	
Expects Data Sharing	Required	Optional	Optional	British Journal of Social Psychology
Mandates Data Sharing	Required	Required	Optional	Ecology and Evolution
Mandates Data Sharing and Peer Reviews Data	Required	Required	Required	Geoscience Data Journal American Journal of Political Science

Data available on request due to privacy/ethical restrictions	The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

Recommended by funders…

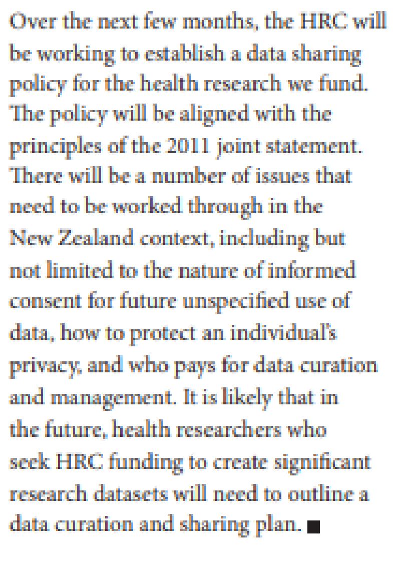
COMPASS RESEARCH CENTRE
FACULTY OF ARTS
THE UNIVERSITY OF AUCKLAND
Whare Wānanga o Tāmaki Makaurau



Sharing research data to improve public health: A joint statement by funders of health research

By Dr Robin Olds, HRC Chief Executive

The Health Research Council of New Zealand (HRC) is one of 17 health research funding agencies that have signed a joint statement on sharing research data.[1] The other signatories represent some of the most significant health research funders internationally.



Over the next few months, the HRC will be working to establish a data sharing policy for the health research we fund. The policy will be aligned with the principles of the 2011 joint statement. There will be a number of issues that need to be worked through in the New Zealand context, including but not limited to the nature of informed consent for future unspecified use of data, how to protect an individual's privacy, and who pays for data curation and management. It is likely that in the future, health researchers who seek HRC funding to create significant research datasets will need to outline a data curation and sharing plan. ■

Medical Research Council UK
The MRC expects valuable data arising from MRC-funded research to be made available to the scientific community with as few restrictions as possible so as to maximise the value of the data for research and for eventual patient and public benefit. Such data must be shared in a timely and responsible manner.

What about disclosure risk?

- By ‘doing the right thing’, are researchers inadvertently risking identity (and attribute) disclosure for those on whom data have been collected?

RQ1: What is the disclosure risk of datasets in publically available data archives?

- Is enough being done to ensure researchers ‘do the right thing’ safely?

RQ2: Do data archives have policies on disclosure risk, and do they highlight risk of disclosure to researchers?

Methods

1. Identify data archives to investigate
2. Identify disclosure risk method(s)
3. Calculate disclosure risk for (sample of) data sets in the data archives (RQ1)
4. Investigate policies of data archives regarding disclosure risk (RQ2)

Data archives

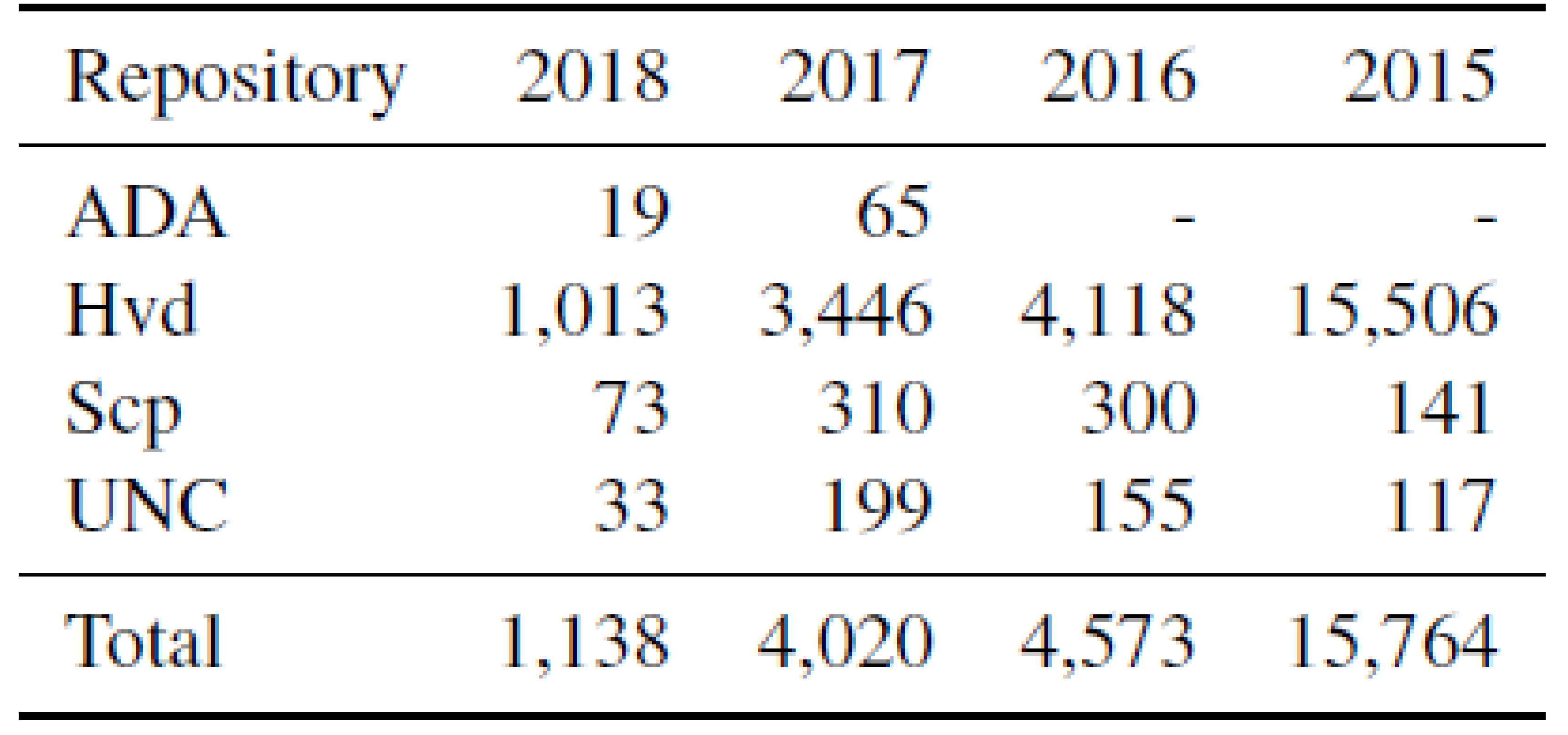
- Sought to identify repositories that were:
  - Open access
  - Contain microdata (not tabular data) about people
  - Cover a broad range of research disciplines
- Many ruled out that were not ‘general discipline’
  - e.g. BMC, National Archives of Criminal Justice, Association of Religious Data Archives

Data archives

- ‘Dataverse Project’ met requirements
  - 32 repositories. We chose to look at the three largest English-language repositories
    - The Harvard Dataverse - general repository for research data which caters to all disciplines of researchers
    - Scholars Portal Dataverse - used by a variety of Canadian universities, colleges, and polytechnics, mostly from Ontario
    - The University of North Carolina (UNC) Dataverse - caters to a general audience of researchers mostly in social science and medical fields; many users from UNC and surrounding universities
  - Additionally we chose to investigate
    - The ADA Dataverse as a ‘local’ dataverse which NZ Social Science Data Service (maintained by COMPASS) is now using

Data archives

- One year of uploads (2017 – most recently completed year)



Repository	2018	2017	2016	2015
ADA	19	65	-	-
Hvd	1,013	3,446	4,118	15,506
Scp	73	310	300	141
UNC	33	199	155	117
Total	1,138	4,020	4,573	15,764

Data archives

- Sample up to
  - n=25 data ‘collections’ tagged as ‘replication’ – linked to a published paper/report
  - n=25 remaining data ‘collections’
- Inclusions
  - Data about humans
  - Metadata sufficient to calculate disclosure risk
- Exclusions
  - Publically available data
  - Access request denied, or external request form required
  - Not microdata

Data archives

	Replication	Other
ADA	0	27
Harvard	82	25
Scholars Portal	1	16
UNC	30	17

Disclosure Risk methods

- File-level disclosure risk
  - Overall mean disclosure risk across records in a file
- Record-level disclosure risk
  - Disclosure risk for every record in a file

- Need method(s) that rely on information from the file itself and minimal external information

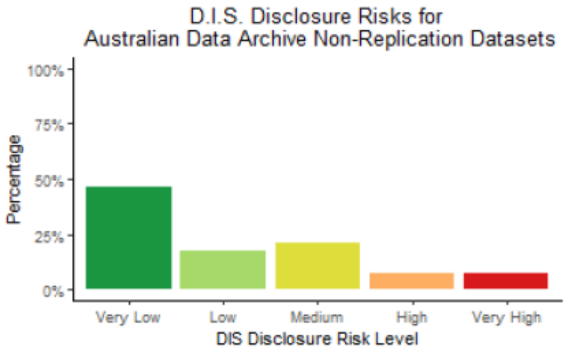
Disclosure Risk methods

- File level
  - **Data Intrusions Simulation (DIS)** appeared to be the most intuitive and implementable method
    - File-wide probability that unique matches (on a set of quasi-identifiers) are correct matches
    - A function of number of sample uniques/pairs and sampling fraction, so only the datafile itself and some estimate of the sampling fraction is needed (a bit of effort…).
- Record level
  - Special Uniques Detection Algorithm (SUDA) scores each record based on the number and size their unique patterns
  - **DIS-SUDA** calibrates the scores to the overall file-level (DIS) score
    - Just need file itself and the DIS estimate from above.

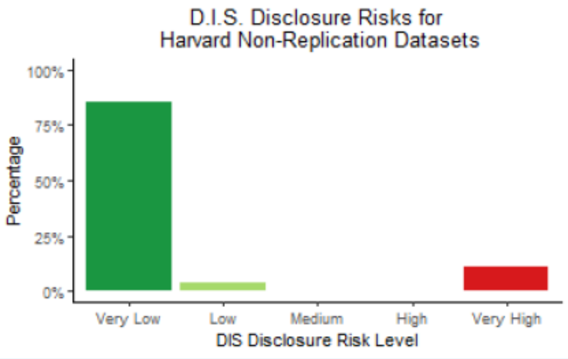
Quasi-identifiers

- Assessed uniqueness on the following quasi-identifiers
  - Age/DOB
  - Gender
  - Ethnicity/Country of birth
  - Employment status
  - Education
  - Marital status
  - Region
- Ignored missings

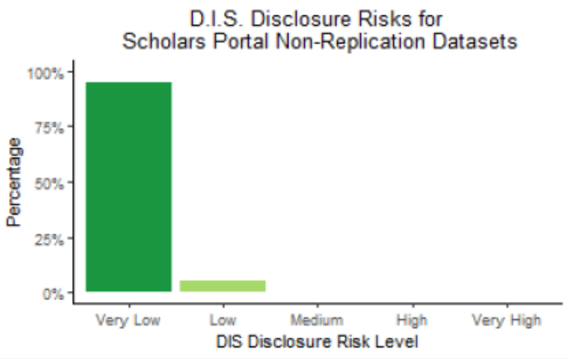
Results
RQ1: File-level


D.I.S. Disclosure Risks for
Australian Data Archive Non-Replication Datasets

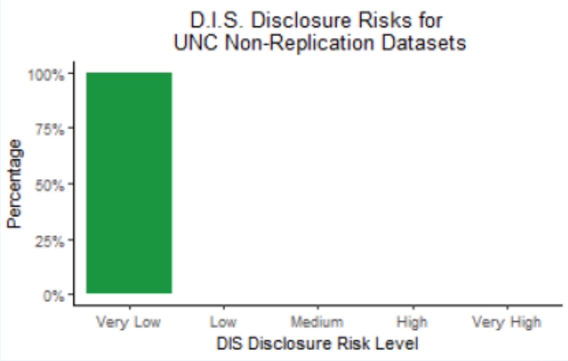
n=27


D.I.S. Disclosure Risks for
Harvard Non-Replication Datasets

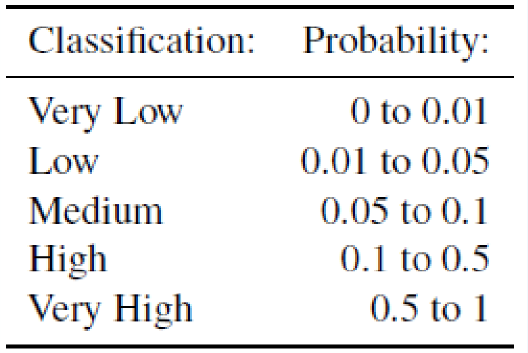
n=25


D.I.S. Disclosure Risks for
Scholars Portal Non-Replication Datasets

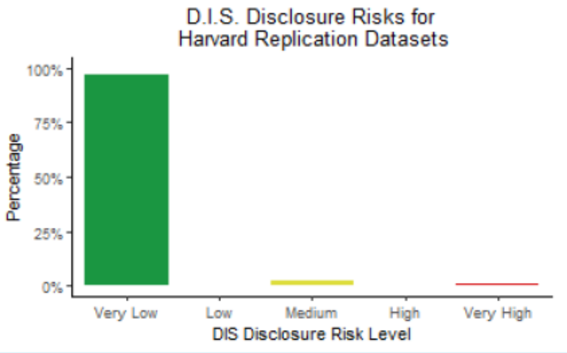
n=16


D.I.S. Disclosure Risks for
UNC Non-Replication Datasets

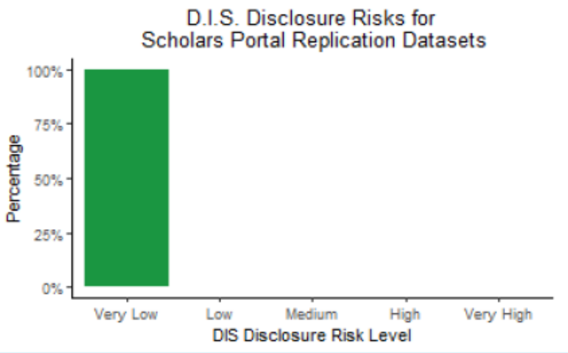
n=17



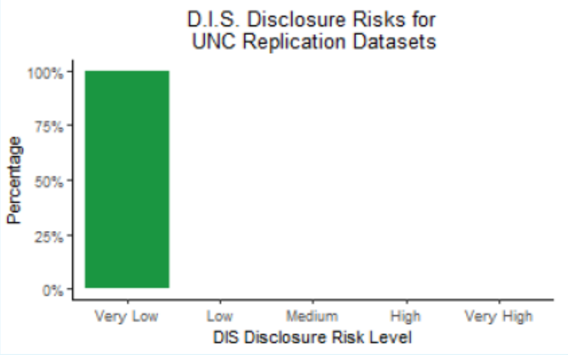
Classification:	Probability:
Very Low	0 to 0.01
Low	0.01 to 0.05
Medium	0.05 to 0.1
High	0.1 to 0.5
Very High	0.5 to 1


D.I.S. Disclosure Risks for
Harvard Replication Datasets

n=82


D.I.S. Disclosure Risks for
Scholars Portal Replication Datasets

n=1


D.I.S. Disclosure Risks for
UNC Replication Datasets

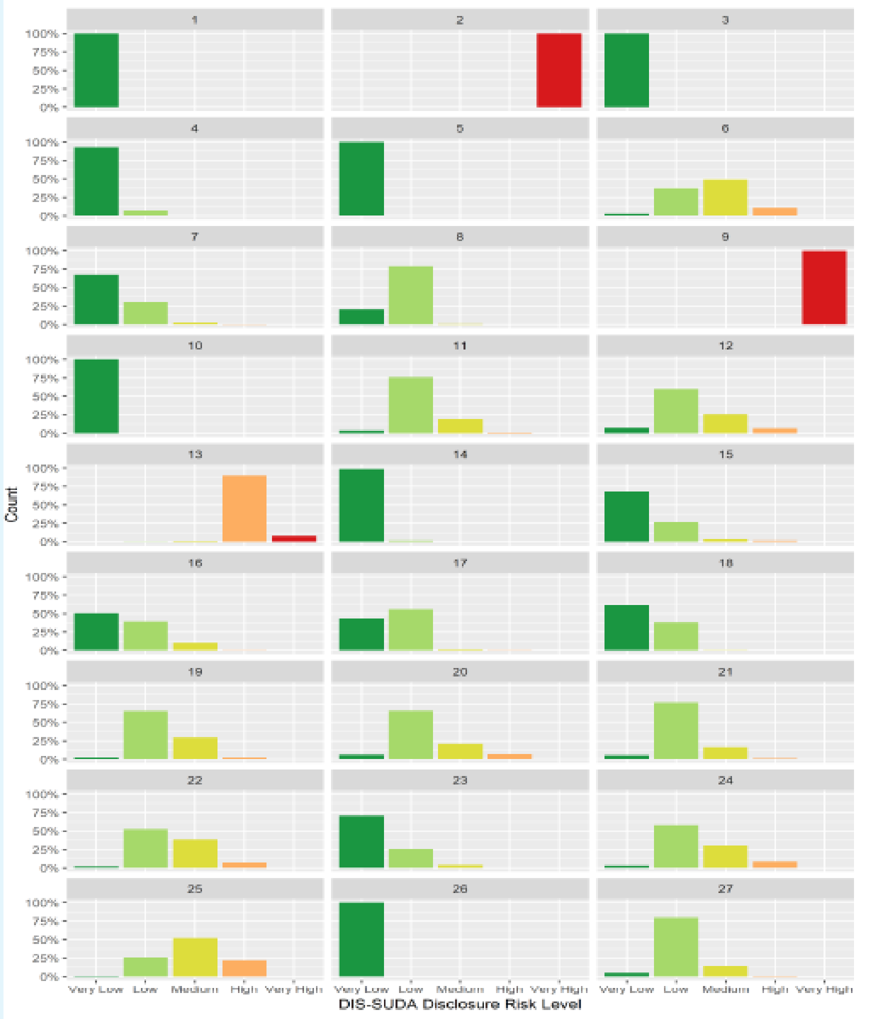
n=30

Results
RQ1: Record-level

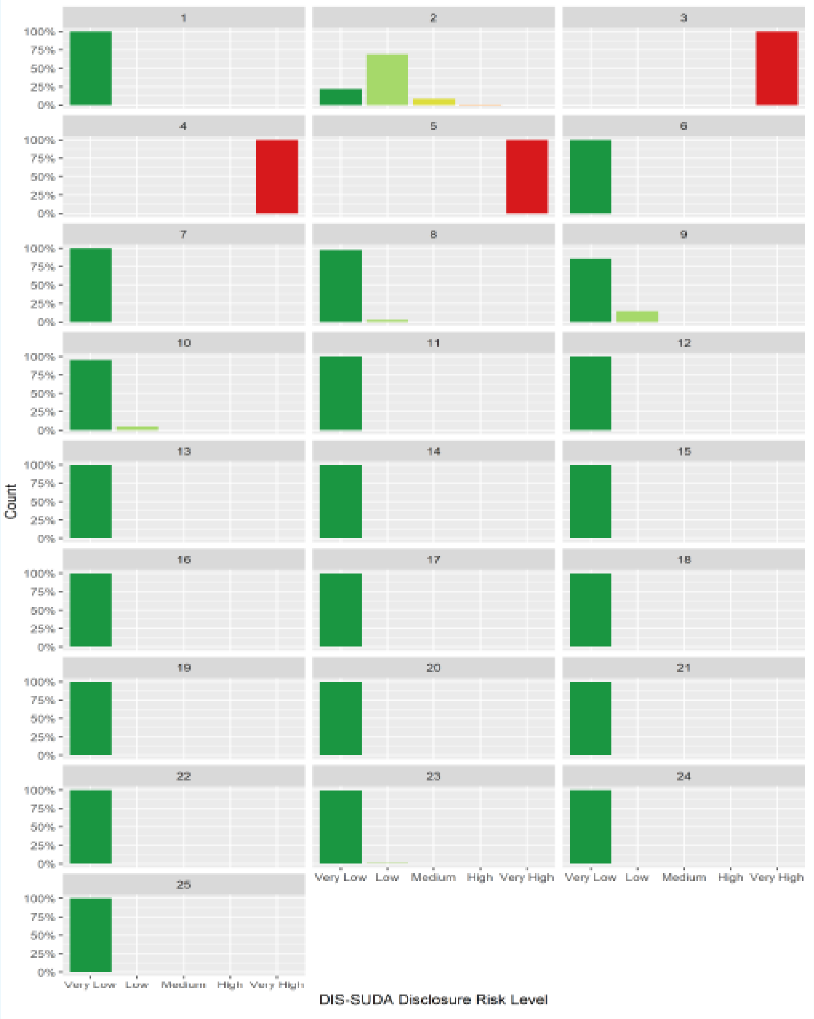
COMPASS RESEARCH CENTRE

FACULTY OF ARTS
THE UNIVERSITY OF AUCKLAND
Whare Wānanga o Tāmaki Makaurau

ADA



Harvard – non-replication



Comment
RQ1: Disclosure risk

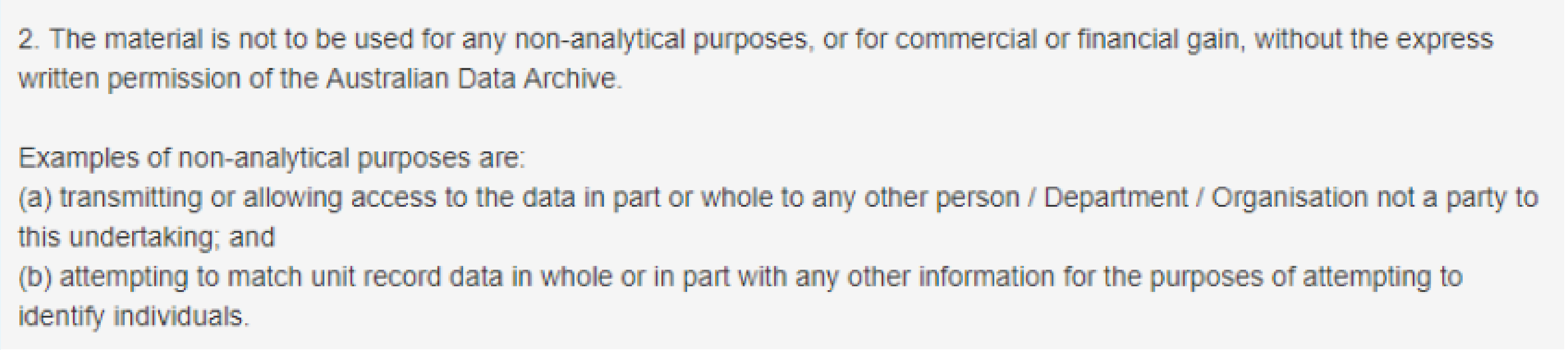
- Most files had very low disclosure risk.
- There were six, however, with very high disclosure risk
  - 3 ADA Dataverse files, which contained postcode – a very granular variable that meant that there were many sample uniques
  - 3 Harvard Dataverse files that contained full names and mobile numbers (yikes!)
- Note, access to most ADA files was by application (e.g., you have to state your purpose) which they assess on a case-by-case basis
  - ADA made aware we are assessing disclosure risk; researchers responsible for data agreed to release the files on that basis
- Both Dataverses will be alerted to the high disclosure risk files

Comment
RQ1: Disclosure risk

- If anything, our disclosure risk estimates are underestimates
  - We had to estimate sampling fraction, and for that we had to estimate population size
  - If the population wasn't obvious, we defaulted to the 'adult population of the country'
  - If this population estimate was an overestimate then our sampling fractions (and disclosure risk estimates) will be underestimated
- Still… most files either had extremely low or extremely high disclosure risk – correctly estimating sampling fractions won't shift the 'extremely low' disclosure risks much

Results
RQ2: Policies on disclosure risk

- ADA allows data sets with high disclosure risk, but have protections around release
  - Application forms for most data sets
  - Terms and conditions expressly forbid re-identification:

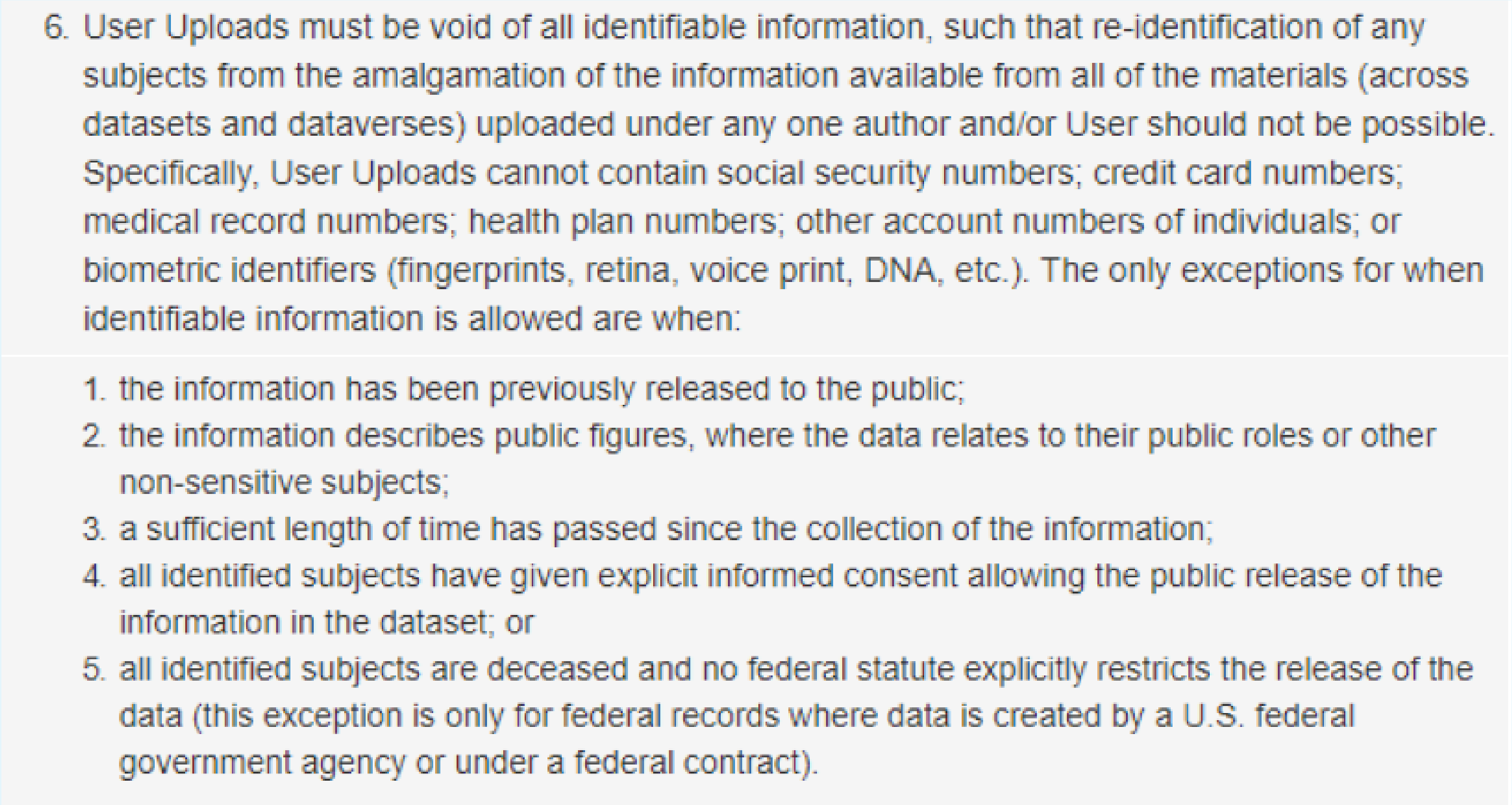


2. The material is not to be used for any non-analytical purposes, or for commercial or financial gain, without the express written permission of the Australian Data Archive.

Examples of non-analytical purposes are:
(a) transmitting or allowing access to the data in part or whole to any other person / Department / Organisation not a party to this undertaking; and
(b) attempting to match unit record data in whole or in part with any other information for the purposes of attempting to identify individuals.

Results
RQ2: Policies on disclosure risk

- Harvard, Scholars Portal and UNC have clauses about exclusion of identifying information
  - Not well monitored (responsibility on researcher)
  - No clauses on re-identification risks of datafiles
  - Guidelines for use indicate that users should not attempt re-identification



6. User Uploads must be void of all identifiable information, such that re-identification of any subjects from the amalgamation of the information available from all of the materials (across datasets and dataverses) uploaded under any one author and/or User should not be possible. Specifically, User Uploads cannot contain social security numbers; credit card numbers; medical record numbers; health plan numbers; other account numbers of individuals; or biometric identifiers (fingerprints, retina, voice print, DNA, etc.). The only exceptions for when identifiable information is allowed are when:

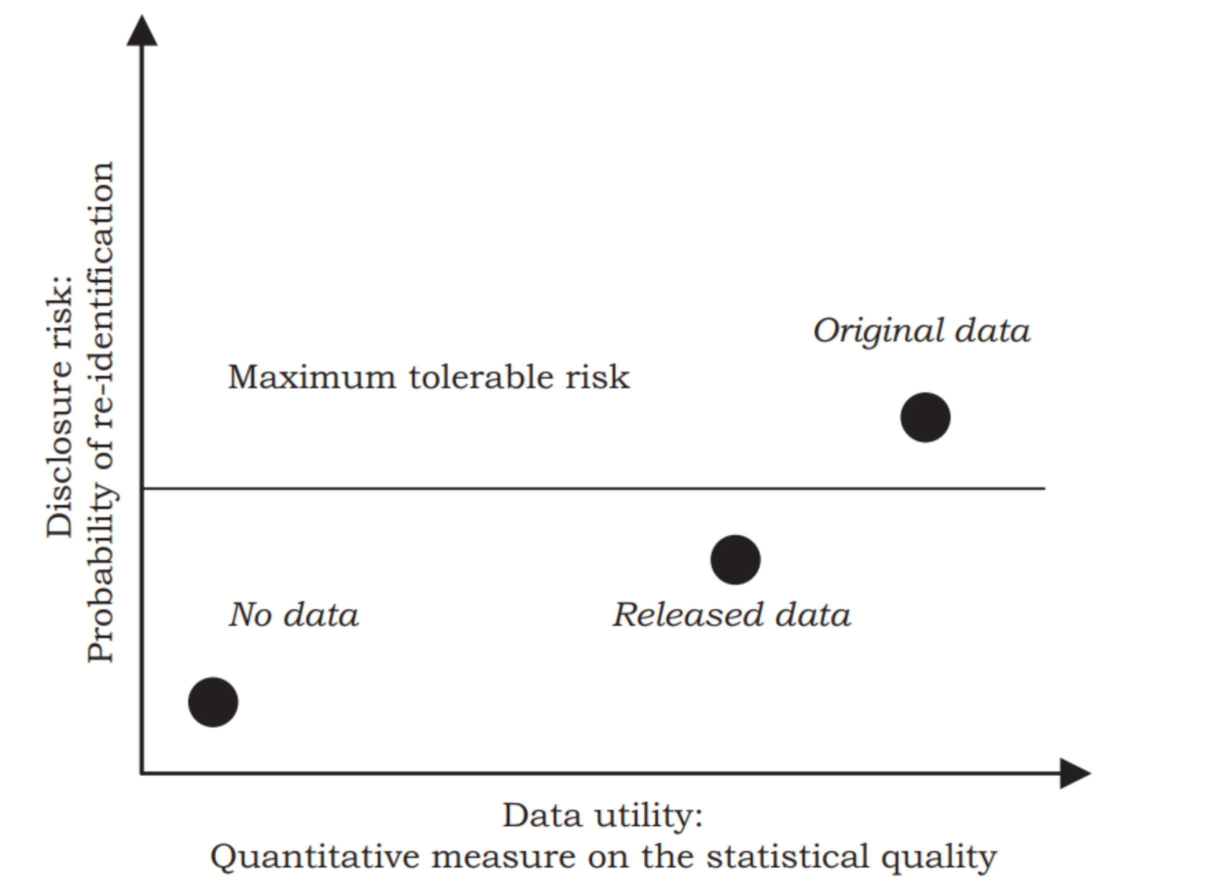
   1. the information has been previously released to the public;
   2. the information describes public figures, where the data relates to their public roles or other non-sensitive subjects;
   3. a sufficient length of time has passed since the collection of the information;
   4. all identified subjects have given explicit informed consent allowing the public release of the information in the dataset; or
   5. all identified subjects are deceased and no federal statute explicitly restricts the release of the data (this exception is only for federal records where data is created by a U.S. federal government agency or under a federal contract).

Conclusions

- Data sets with high disclosure risk can be found in data archives (including some data sets with direct identifiers)
  - Even though most data sets have extremely low risk, it is alarming to have any with high risk
- High disclosure is mostly caused by the inclusion of highly granular variables (e.g. postcode)
  - These should be removed (are they really needed for research??) or re-categorised into less granular variables
- More should be done by data archives, journals and funders to educate researchers about disclosure risk and how to protect against it

Conclusions

- It is always possible to keep data quality high while protecting disclosure of sensitive information



- THANKS!!
- Especially to Shaun for his excellent analytic work

- QUESTIONS??