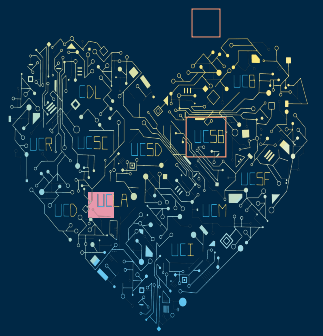
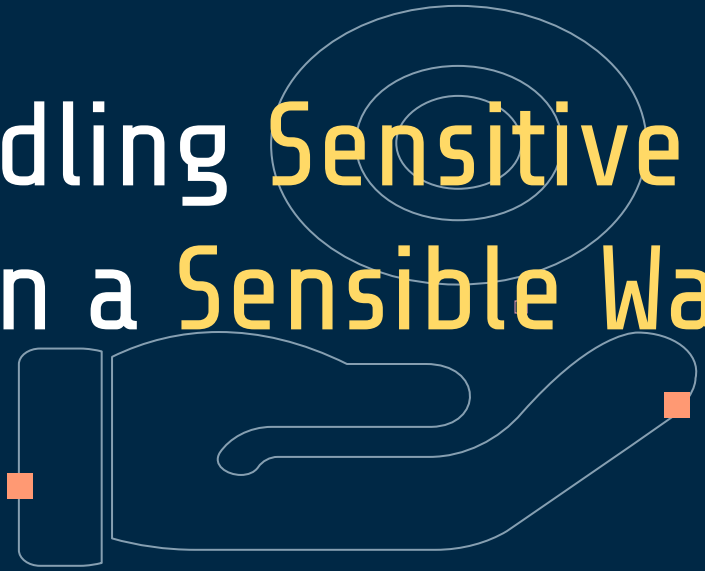
UC 💗 Data Week 2022

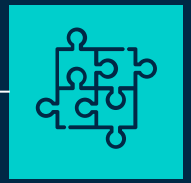


Handling Sensitive Data in a Sensible Way



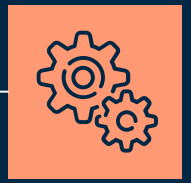
Renata Curty
Social Sciences Research Facilitator
UCSB Library – Research Data Services
Feb. 17 2022

AGENDA FOR THIS SESSION



01

What is Sensitive Data? Why should we care?



02

Existing Protocols & Recommendations

03

Ethical & Practical Considerations

SENSITIVE DATA
Definition

Data that must be protected against unwanted disclosure and which access should be safeguarded.

Protection of sensitive data may be required for legal or ethical reasons, for issues pertaining to personal privacy, or for proprietary considerations.

SENSITIVE DATA

Whose protection?

Any human subject data that can potentially disclose people's identity and damage individual or collective reputations, rights, or best interests.

It also includes data, which, if disclosed without precaution, may infringe upon ethical agreements and threaten the ownership, representation, and existence of vulnerable communities, protected lands and species.

SENSITIVE DATA

Beyond humans subjects!

GBIF Global Biodiversity Information Facility

4 levels of sensitivity according to **biological significance** and **threat** from exploitation

Key recommendation:
Generalize the spatial locality or geographic coordinates.

Chapman AD (2020) Current Best Practices for Generalizing Sensitive Species Occurrence Data. Copenhagen: GBIF Secretariat. https://doi.org/10.15468/doc-5jp4-5g10.

Endangered Species

Amazon "Pink" Dolphin

Protected Research Sites

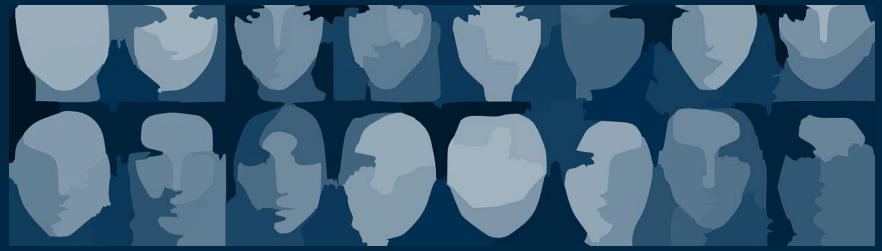
Pacunam
FUNDACIÓN PATRIMONIO CULTURAL Y NATURAL MAYA

Threatened or Commercially Exploited Plants

Prehistoric Wollemi Pine

HUMAN SUBJECTS DATA

Personal identifiable information (PII)



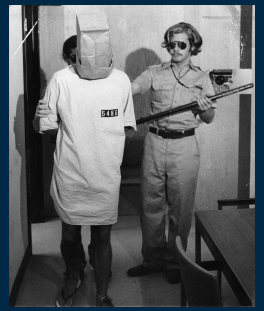
Why it is so important to protect participants' rights and privacy?

NOT THAT LONG AGO...

Infamous Experiments

- Participants were unwary exposed to physical and psychological risks
- Exploitation of vulnerable groups and economically or educationally disadvantaged persons

Milgram's Obedience to Authority Shock (1961)



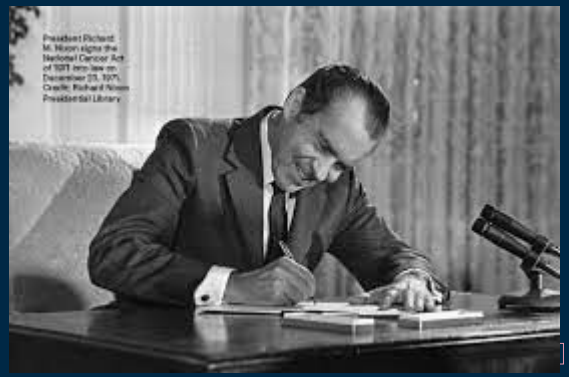
Stanford Prison (1971)

Tuskegee syphilis study (1932-1972)

NATIONAL RESEARCH ACT

Series of congressional hearings on human-subjects research

Signed into law in 1974, creating the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research



THE BIG SHIFT (1979)

Biomedical and behavioral human subjects research

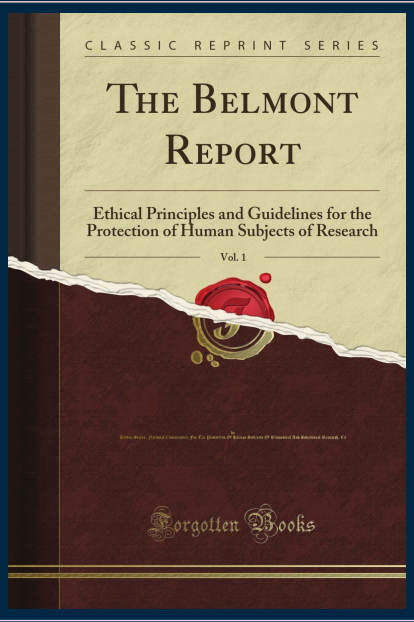
Practice vs. Research

Basic Principles:

- Respect for Persons
- Beneficence
- Justice

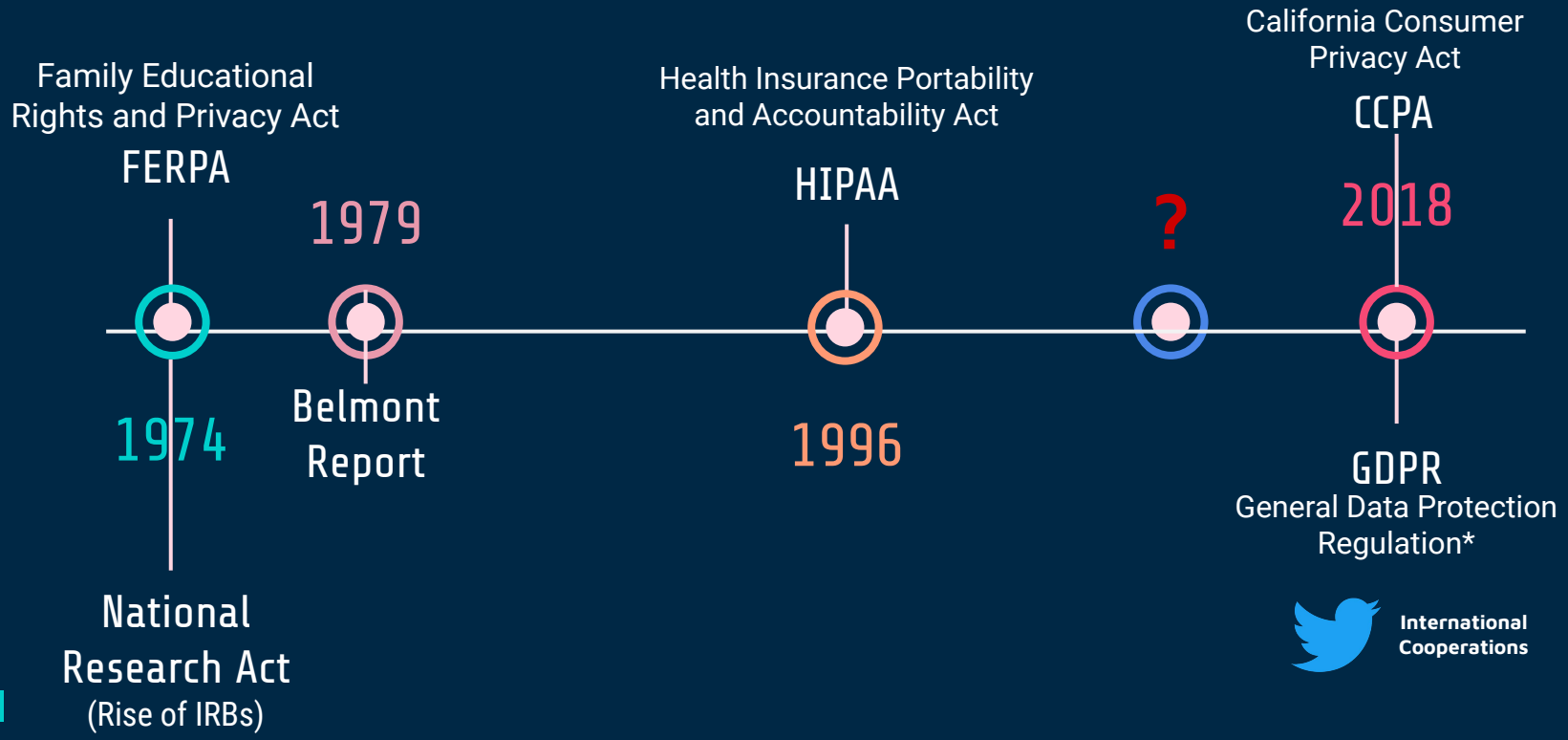
Applications:

- Informed Consent
- Assessment of Risks and Benefits
- Selection of Subjects



https://www.hhs.gov/ohrp/sites/default/files/the-belmont-report-508c_FINAL.pdf

SOME HALLMARKS



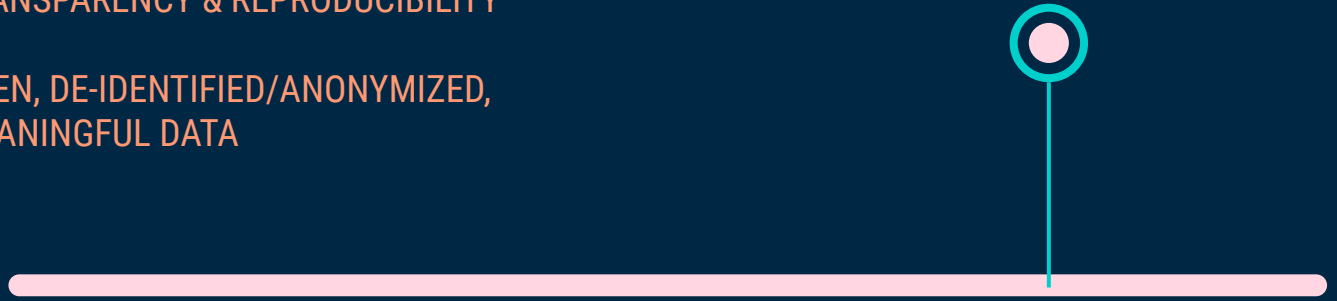
A TENSION

TRANSPARENCY & REPRODUCIBILITY

OPEN, DE-IDENTIFIED/ANONYMIZED, MEANINGFUL DATA

NSF Data Management Plans & Sharing Mandates

2011



- Back then, little practical guidance on how and where data should be shared
- Confidentiality and anonymity key for publishing or sharing data relating to individuals, as means to minimize risks to subjects privacy
- Previous consent for publication of appropriately anonymized raw data from participants

Informed Consent Language (Exempt Research)

We are asking you to take part in a research study being done by [*list researcher's name*] at the University of California. Being in this study is optional.

If you choose to be in the study, you will complete a survey. This survey will help us learn more about [*briefly describe the purpose of the research*]. [*Optional: If unclear, explain why subjects are being asked to participate and/or how they were selected.*] The survey will take about [*XX minutes or hours*] to complete.

You can skip questions that you do not want to answer or stop the survey at any time. The survey is anonymous, and no one will be able to link your answers back to you. Results will be only presented in aggregated form.

Questions? Please contact [*researcher's name*] at [*contact info*]. If you have questions or concerns about your rights as a research participant, you can call the Institutional Review Board at [phone number]

Adapted from https://irb.ucsf.edu/

POINTS TO CONSIDER

Even if there is low/minimal risks to participants on exempted research:

- Data sharing and reuse?
- Are we making promises we can keep?
- Are we considering the risk of re-identification or data breach?

HUMAN SUBJECTS PRIVACY (NOT BINARY, IT'S A SPECTRUM!)

Types of Identifiable Data

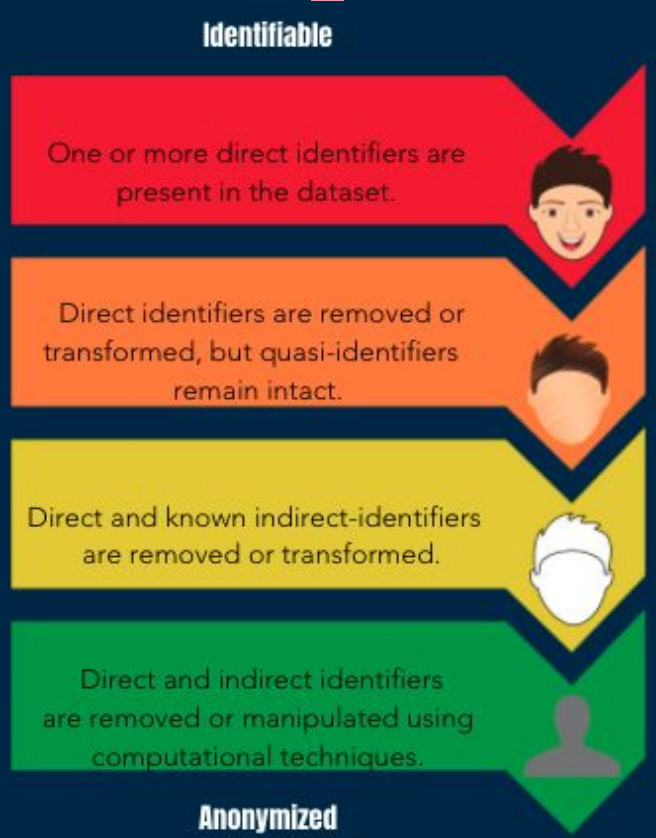
Direct identifiers
Unique to individuals
Examples:

- Name
- Email
- SSN
- IP address
- Phone number
- Full-face images
- Medical record number

Indirect-identifiers
Attributes that combined can disclose one's identity
Examples:

- Race or ethnicity
- Age
- Gender
- Zipcode
- Political opinion
- Religious orientation
- Affiliation/profession

Risk of Re-identification



IT IS A SPECTRUM

WITHIN A LIFECYCLE

1. Data collection (e.g., protocols and instruments)
2. Storage/Backup (e.g., authentication, safety measurements)
3. Cleaning/Processing
4. Reporting
5. Sharing/Archiving (e.g., repository, auxiliary documentation)
6. Access & Reuse (e.g., restriction levels, license agreements, attribution)

NOT as anonymous as you think!

Risk of Re-identification

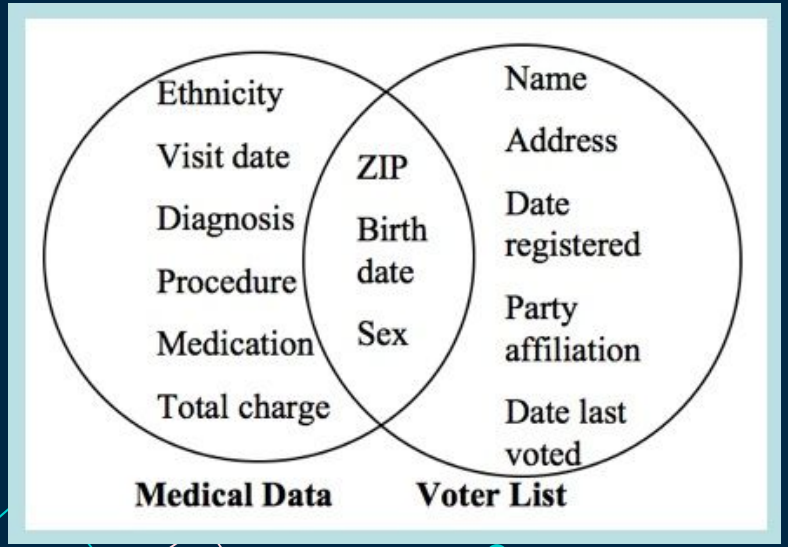
- 1990 census data ~ 90% of the U.S. population could be identified by just a ZIP code, date of birth, and gender, in combination with secondary health data sold by pharmacies and analytics companies.

Dr. Latanya Sweeney (K-anonymity)

NOT as anonymous as you think!

Risk of Re-identification

- (1997) re-identification experiment wherein she successfully identified then Massachusetts governor, William Weld, to his medical records using publicly accessible records.



K-Anonymity

RISK ASSESSMENT AND MITIGATION

"Hiding in the crowd guarantee"

At least *k* individuals in the dataset who share the set of attributes that might become identifying for each individual.

3 to 5 "matching cases" desired

ID	Age	Zipcode	Diagnosis
1	28	13053	Heart Disease
2	29	13068	Heart Disease
3	21	13068	Viral Infection
4	23	13053	Viral Infection
5	50	14853	Cancer
6	55	14853	Heart Disease
7	47	14850	Viral Infection
8	49	14850	Viral Infection
9	31	13053	Cancer
10	37	13053	Cancer
11	36	13222	Cancer
12	35	13068	Cancer

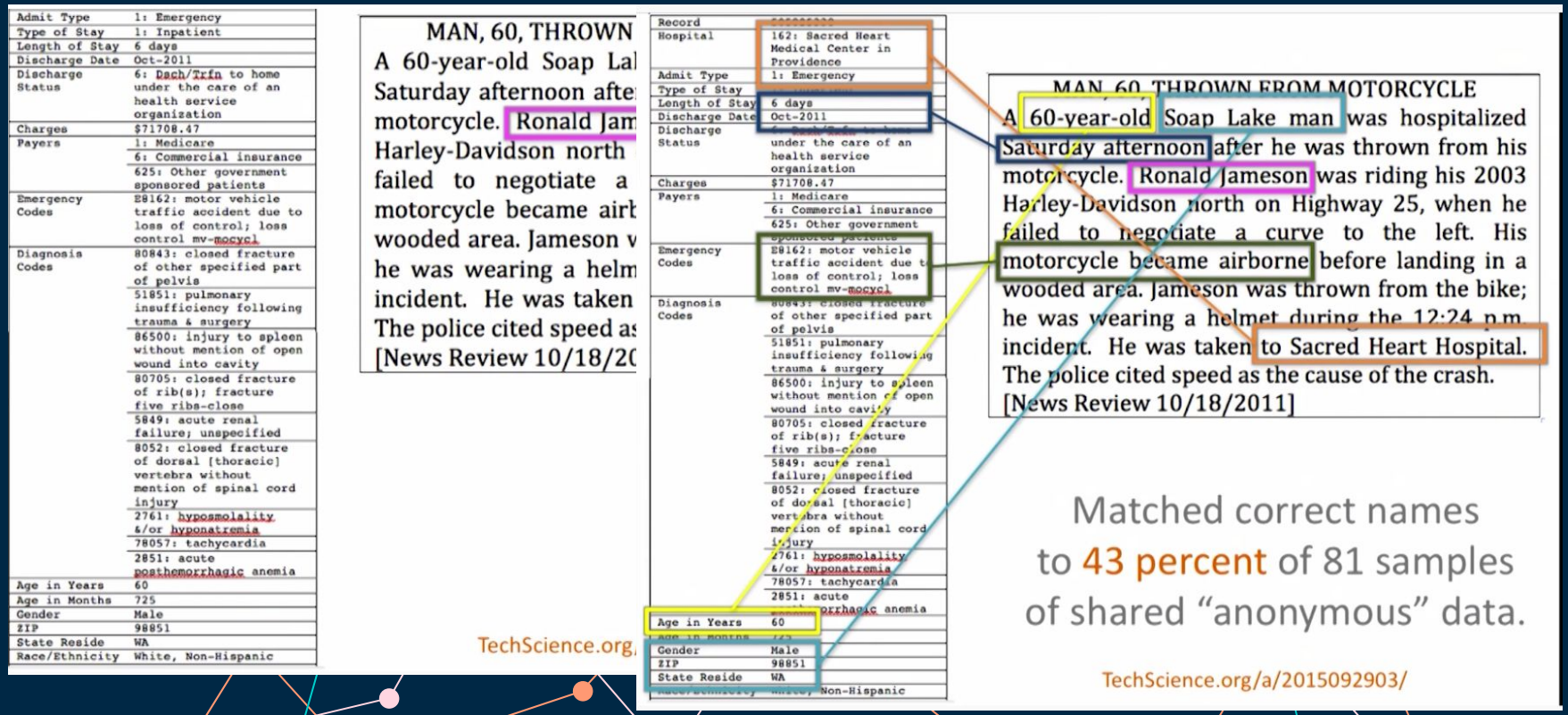
ID	Age	Zipcode	Diagnosis
1	[20-30]	130**	Heart Disease
2	[20-30]	130**	Heart Disease
3	[20-30]	130**	Viral Infection
4	[20-30]	130**	Viral Infection
5	[40-60]	148**	Cancer
6	[40-60]	148**	Heart Disease
7	[40-60]	148**	Viral Infection
8	[40-60]	148**	Viral Infection
9	[30-40]	13***	Cancer
10	[30-40]	13***	Cancer
11	[30-40]	13***	Cancer
12	[30-40]	13***	Cancer

SUPPRESSION + GENERALIZATION

Recommended videos:
https://www.youtube.com/watch?v=X3MKP_-FrWE
https://www.youtube.com/watch?v=_0Io6c1MPOY

You may assess it in SPSS, Stata and R (plyr)

ANOTHER CASE



Admit Type	1: Emergency
Type of Stay	1: Inpatient
Length of Stay	6 days
Discharge Date	Oct-2011
Discharge Status	6: Dsch/Trfn to home under the care of an health service organization
Charges	$71708.47
Payers	1: Medicare
	6: Commercial insurance
	625: Other government sponsored patients
Emergency Codes	E8162: motor vehicle traffic accident due to loss of control; loss control mv-mocycl
Diagnosis Codes	80843: closed fracture of other specified part of pelvis
	51851: pulmonary insufficiency following trauma & surgery
	86500: injury to spleen without mention of open wound into cavity
	80705: closed fracture of rib(s); fracture five ribs-close
	5849: acute renal failure; unspecified
	8052: closed fracture of dorsal [thoracic] vertebra without mention of spinal cord injury
	2761: hyposmolality &/or hyponatremia
	78057: tachycardia
	2851: acute posthemorrhagic anemia
Age in Years	60
Age in Months	725
Gender	Male
ZIP	98851
State Reside	WA
Race/Ethnicity	White, Non-Hispanic

MAN, 60, THROWN

A 60-year-old Soap Lak
Saturday afternoon afte
motorcycle. Ronald Jam
Harley-Davidson north
failed to negotiate a
motorcycle became airb
wooded area. Jameson v
he was wearing a helm
incident. He was taken
The police cited speed as
[News Review 10/18/20

TechScience.org/

Record	[redacted]
Hospital	162: Sacred Heart Medical Center in Providence
Admit Type	1: Emergency
Type of Stay	[redacted]
Length of Stay	6 days
Discharge Date	Oct-2011
Discharge Status	[redacted] under the care of an health service organization
Charges	$71708.47
Payers	1: Medicare
	6: Commercial insurance
	625: Other government [redacted]
Emergency Codes	E8162: motor vehicle traffic accident due t loss of control; loss control mv-mocycl
Diagnosis Codes	[redacted] of other specified part of pelvis
	51851: pulmonary insufficiency following trauma & surgery
	86500: injury to spleen without mention of open wound into cavity
	80705: closed fracture of rib(s); fracture five ribs-close
	5849: acute renal failure; unspecified
	8052: closed fracture of dorsal [thoracic] vertebra without mention of spinal cord injury
	2761: hyposmolality &/or hyponatremia
	78057: tachycardia
	2851: acute [redacted] orrhagic anemia
Age in Years	60
Age in Months	[redacted]
Gender	Male
ZIP	98851
State Reside	WA
[redacted]	[redacted] Non-Hispanic

MAN, 60, THROWN FROM MOTORCYCLE

A 60-year-old Soap Lake man was hospitalized Saturday afternoon after he was thrown from his motorcycle. Ronald Jameson was riding his 2003 Harley-Davidson north on Highway 25, when he failed to negotiate a curve to the left. His motorcycle became airborne before landing in a wooded area. Jameson was thrown from the bike; he was wearing a helmet during the 12:24 p.m. incident. He was taken to Sacred Heart Hospital. The police cited speed as the cause of the crash. [News Review 10/18/2011]

Matched correct names
to 43 percent of 81 samples
of shared "anonymous" data.

TechScience.org/a/2015092903/

https://www.youtube.com/watch?v=tivCK_fBBfo

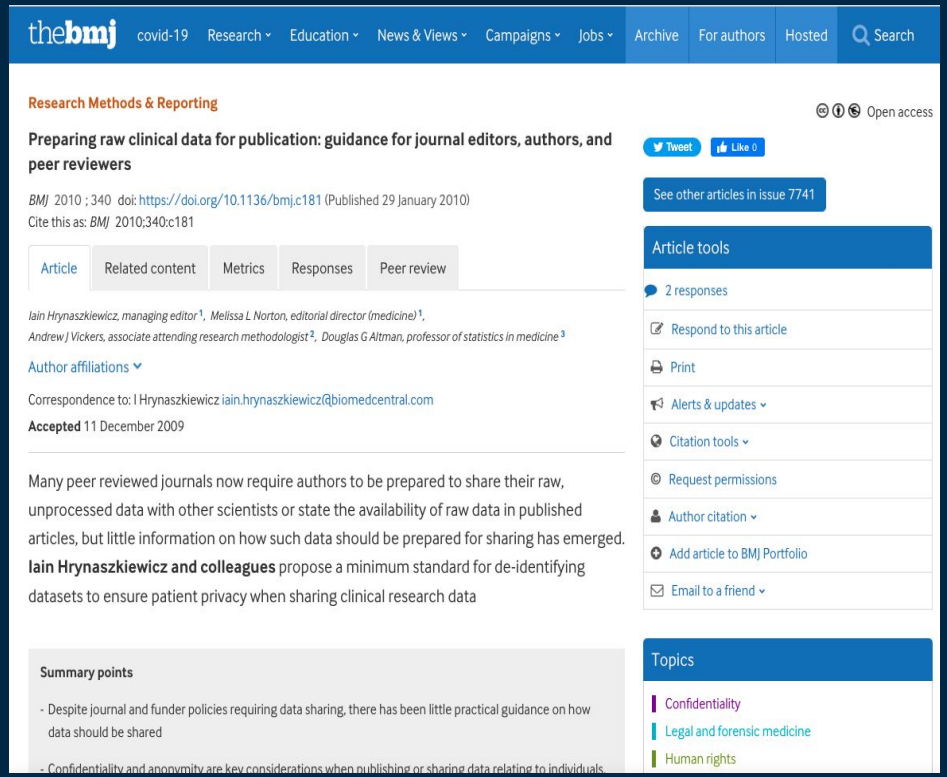
HIPAA (18 IDENTIFIERS)

SAFE HARBOR METHOD

- Name
- Address (all geographic subdivisions smaller than state)
- All elements (except years) of dates (e.g., birthdate, admission date, discharge date, date of death)
- Telephone numbers
- Fax number
- Email address
- Social Security Number
- Medical record number
- Health plan beneficiary number
- Account number
- Certificate or license number
- Vehicle identifiers
- Device identifiers and serial numbers
- Web URL
- Internet Protocol (IP) Address
- Finger or voice print
- Photographic image
- Any other characteristic that could uniquely identify the individuals

WAIT, THERE ARE 10 MORE...

Aggregated from policy documents and research guidance from major UK and US funding agencies, governmental health departments and statutes, and three internationally recognised publication ethics resources for editors of biomedical journals



thebmj covid-19 Research Education News & Views Campaigns Jobs Archive For authors Hosted Search

Research Methods & Reporting

Open access

Preparing raw clinical data for publication: guidance for journal editors, authors, and peer reviewers

BMJ 2010 ; 340 doi: https://doi.org/10.1136/bmj.c181 (Published 29 January 2010)
Cite this as: *BMJ* 2010;340:c181

Tweet Like 0

See other articles in issue 7741

Article Related content Metrics Responses Peer review

Iain Hrynaszkiewicz, managing editor [1], Melissa L Norton, editorial director (medicine) [1], Andrew J Vickers, associate attending research methodologist [2], Douglas G Altman, professor of statistics in medicine [3]

Author affiliations

Correspondence to: I Hrynaszkiewicz iain.hrynaszkiewicz@biomedcentral.com

Accepted 11 December 2009

Many peer reviewed journals now require authors to be prepared to share their raw, unprocessed data with other scientists or state the availability of raw data in published articles, but little information on how such data should be prepared for sharing has emerged. **Iain Hrynaszkiewicz and colleagues** propose a minimum standard for de-identifying datasets to ensure patient privacy when sharing clinical research data

Summary points

- Despite journal and funder policies requiring data sharing, there has been little practical guidance on how data should be shared
- Confidentiality and anonymity are key considerations when publishing or sharing data relating to individuals

Article tools

2 responses
Respond to this article
Print
Alerts & updates
Citation tools
Request permissions
Author citation
Add article to BMJ Portfolio
Email to a friend

Topics

Confidentiality
Legal and forensic medicine
Human rights

https://doi.org/10.1136/bmj.c181

GOLDEN RULE

STRICTLY NECESSARY

IDENTIFIERS ONLY

Some De-identification Techniques

Redaction/Suppression

Removal of identifiers that can put subjects' identities and privacy at risk.

Swapping/Shuffling

Data for one or more variables are switched with another record, so that the data does not know whether the real data values correspond to certain records.

Pseudonymization/ Tokenization/Hashing

Meaningful piece of data are turned into a into an alias or a random string of characters which serves as reference to the original data, but cannot be used to guess those values.

Noising/Disturbing

Slightly modification of the original dataset by applying techniques that round numbers and add random variation.

EXAMPLE Redaction/Suppression

FIRST NAME	LAST NAME	EMAIL	COURSE	GRADE
Mathew	Keaton	mpine@gmail.com	HIS200	7.0
Patricia	Mason	pattymas@gmail.com	ANT300	8.5

FIRST NAME	COURSE	GRADE
Mathew	HIS200	7.0
Patricia	ANT300	8.5

EXAMPLE Pseudonymization/Tokenization/Hashing

FIRST NAME	LAST NAME	EMAIL	COURSE	GRADE
Mathew	Keaton	mpine@gmail.com	HIS200	7.0
Patricia	Mason	pattymas@gmail.com	ANT300	8.5

NAME	EMAIL	COURSE	GRADE
S00001	6c429dwf	HIS200	7.0
S00009	8f156lmp	ANT300	8.5

EXAMPLE Swapping/Shuffling

FIRST NAME	LAST NAME	EMAIL	COURSE	GRADE
Mathew	Keaton	mpine@gmail.com	HIS200	7.0
Patricia	Mason	pattymas@gmail.com	ANT300	8.5

NAME	EMAIL	COURSE	GRADE
S00001	6c429dwf	ANT300	8.5
S00009	8f156lmp	HIS200	7.0

EXAMPLE Noising/Disturbing

FIRST NAME	LAST NAME	EMAIL	COURSE	GRADE
Mathew	Keaton	mpine@gmail.com	HIS200	7.0
Patricia	Mason	pattymas@gmail.com	ANT300	8.5

NAME	EMAIL	COURSE	GRADE
S00001	6c429dwf	HIS200	7.5
St00009	8f156lmp	ANT300	8.0

- "Corrupting" the dataset in a controlled manner. Proceed with caution. Only for large datasets.

MORE WAYS TO MITIGATE RE-IDENTIFICATION

(QUANTITATIVE DATA TRANSFORMATIONS)

If a variable might act as an indirect identifier you may also consider:

- Aggregation reducing the precision of the variable or the detail of its characteristics (e.g., remove last 4 digits of a zip code, then turn into county name)
- Top-coding restricting the upper range of a variable. (e.g., income categories, the top might be noted as 200,000 and above). By leaving the top category as identified only on the low end, it would be impossible for a user to identify the very few people person in the study sample who makes 300,000 per year.
- Collapsing and/or combining variables merging data recorded in two or more variables into a single category. This is particularly useful if the initial data collection created several categories with very few subjects in each of them.
- Bracketing/Categorization the process of transforming continuous variables into categorical variables by reporting a variable range rather than its specific value (e.g., number of years in the institution, range of years)

OTHER STRATEGIES
(QUANTITATIVE DATA)

- Use weighted data to generate output
- Avoid submitting tables with small cell sizes (i.e., cells with fewer than 5 respondents)
- Restrict cross-tabular analysis to two or three dimensions
- Be cautious when using small subgroups or small areas
- Avoid listings of cases with outliers

QUALITATIVE DATA?

Whose narratives?

P001, 29
Dentist

P045 32
K-12 Teacher

P019, 61
Contractor

P023, 54
Marketing Analyst

P005, 41
IT Manager

P036, 45
Business Owner

A METANARRATIVE BASED ON CONNECTIONS FOUND IN INDIVIDUAL NARRATIVES

QUALITATIVE DATA DE-ID

De-identification or anonymization may distort or otherwise affect the value of the narratives and personal experiences shared by human subjects.

Importance to find a balance between keeping your participants' information confidential and unnecessarily reducing the analytic value of the data by removing too much information. If you are having difficulties striking that balance,

Techniques that may be applied to qualitative data include:

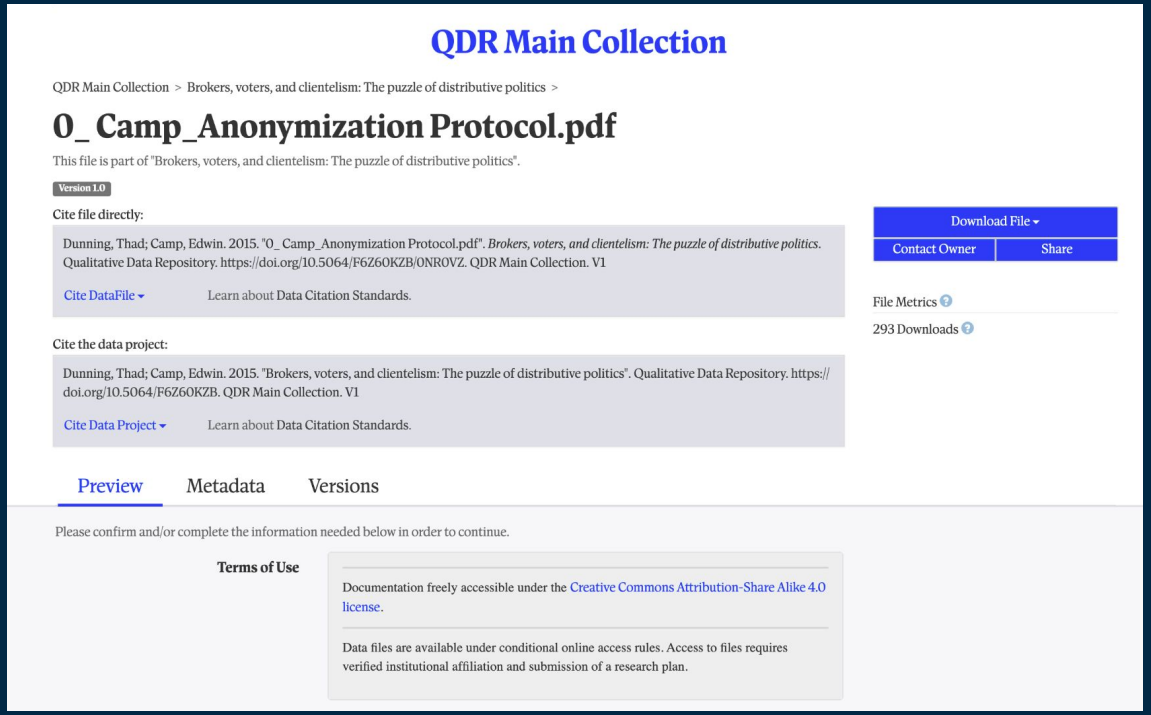
- Using pseudonyms in place of actual names
- Employing abstract systems of coding responses
- Removing elements or whole blocks of sensitive text

QUALITATIVE RESEARCH DE-ID

- Engage in de-identification soon after your interactions with your human participants, marking up elements that require redaction during transcription and/or analysis.
- Clearly and consistently indicate any changes you make to the original file, e.g., by placing square brackets around passages that have been changed.
- Give preference to pseudonyms, or aggregate nouns (e.g., refer to the state in which an individual lives rather than the town or county) or categories (e.g., "… was born in [1975-1980]" instead of 1977) over redacting.
- Check the document properties of files, which may contain identifiers such as original file names identifying interview respondents.
- Keep a list of de-identification rules, both for yourself, or for your team should you collaborate. This list serves as important documentation when you share your data. This document is separate from the key that links de-identified entries to the actual individuals or entities interviewed, which should not be included when you share your data.

DOCUMENTING DE-ID

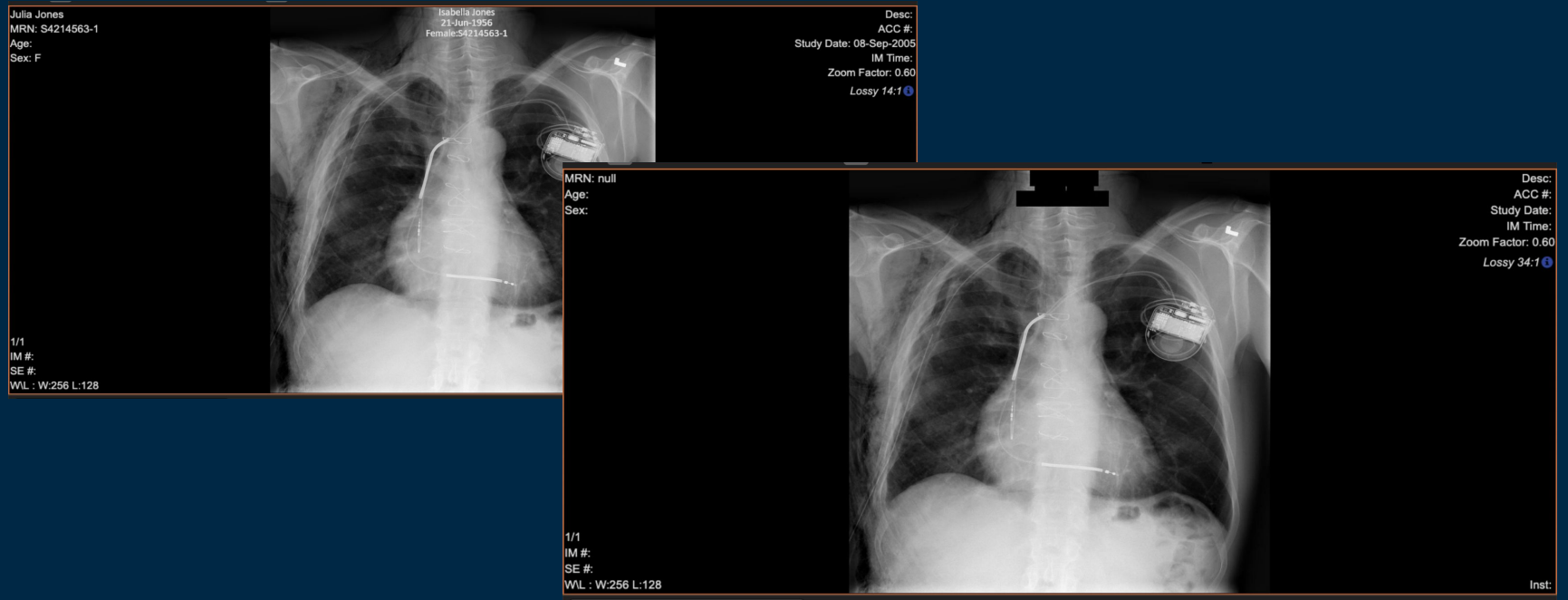
Guidelines on Making the Transcriptions Anonymous



Dunning, Thad; Camp, Edwin. 2015. "0_ Camp_Anonymization Protocol.pdf". *Brokers, voters, and clientelism: The puzzle of distributive politics.* Qualitative Data Repository. https://doi.org/10.5064/F6Z60KZB/0NR0VZ. QDR Main Collection. V1

IMAGES DE-ID

Digital Imaging and Communications in Medicine (DICOM)



https://cloud.google.com/architecture/de-identification-of-medical-images-through-the-cloud-healthcare-api

IMAGES DE-ID

Pixelation, how effective?

Timothy Gerstner, Doug DeCarlo, Marc Alexa, Adam Finkelstein, Yotam Gingold, Andrew Nealen
In Proceedings of the International Symposium on Non-Photorealistic Animation and Rendering (NPAR), Annecy, France, June 2012.

Multifactor models

Gross, Ralph, Latanya Sweeney, Jeffrey Cohn, Fernando de la Torre, and Simon Baker. "Face De-Identification." *Protecting Privacy in Video Surveillance.* Ed. Andrew Senior. Springer, 2009.

NO ONE SIZE FITS ALL

For some datasets, making it safely non-identifiable may require stripping away so much data that they may lose almost all analytic value.

REASONS NOT TO SHARE ARE ACCEPTABLE IN SOME CASES!

"Despite all the efforts to anonymize location data tied to specific individuals or devices, we anticipate that even highly aggregated location data about patterns of large groups of people can unintentionally reveal sensitive Information. Therefore, considering the potential risks for mobility data to be re-identified, the simulated generated using actual tracking data of state residents will not be publicly shared."

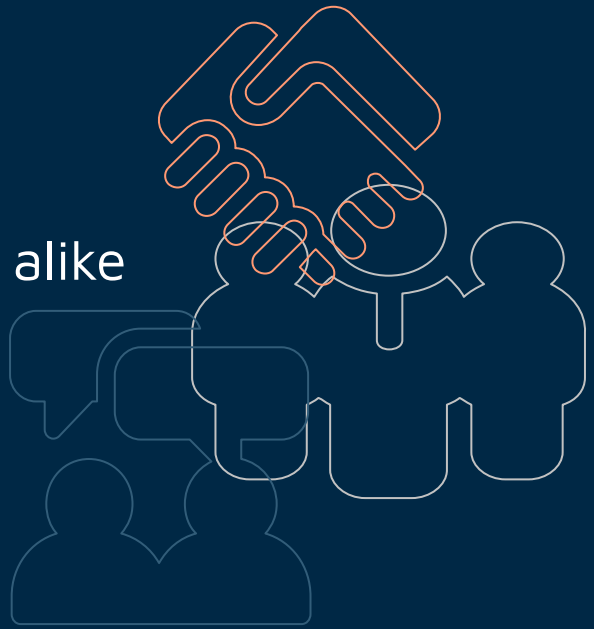
RESTRICTING ACCESS

as open as possible, as closed as necessary

- Licensing Agreements (ensure continued confidentiality compliance)
- Differential Privacy
- Data Enclave
- Metadata Only

NOT SURE WHERE TO START?

- IRB/Office of Research
- Program Officers
- (Data) librarians, curators, stewards and alike



QUESTIONS?
COMMENTS?

rcurty@ucsb.edu

THANKS