Data Ethics

Data Analysis for Journalism and Political Communication
(Spring 2026)

Prof. Bell

Tuskegee Syphilis Study

The Belmont Report (1978)

- Summarizes ethical principles and guidelines for conducting research with human subjects

The Belmont Report (1978)

- Summarizes ethical principles and guidelines for conducting research with human subjects
- Codified into law as the “Common Rule”, which covers all federally-funded research

The Belmont Report (1978)

Respect for Persons

Individuals should be treated as autonomous agents, and persons with diminished autonomy are entitled to protection.

The Belmont Report (1978)

Respect for Persons

Individuals should be treated as autonomous agents, and persons with diminished autonomy are entitled to protection.

Beneficence

(1) Do not harm and (2) maximize possible benefits and minimize possible harms.

The Belmont Report (1978)

Respect for Persons

Individuals should be treated as autonomous agents, and persons with diminished autonomy are entitled to protection.

Beneficence

(1) Do not harm and (2) maximize possible benefits and minimize possible harms.

Justice

Groups who bear the burden of research should also be the beneficiaries of that research.

Other Considerations

- Informed Consent
  - Subjects must affirmatively agree to participate in research

Other Considerations

- Informed Consent
    - Subjects must affirmatively agree to participate in research
    - Participants must have all information necessary to make an informed decision

Other Considerations

- Informed Consent
  - Subjects must affirmatively agree to participate in research
  - Participants must have all information necessary to make an informed decision
  - What about studies that require deception? Studies that compensate participants?

Other Considerations

- Informed Consent
  - Subjects must affirmatively agree to participate in research
  - Participants must have all information necessary to make an informed decision
  - What about studies that require deception? Studies that compensate participants?
- Weighing benefits and risks of research
  - Is there a way to gain the information sought in the study that minimizes risk?

Other Considerations

- Informed Consent
  - Subjects must affirmatively agree to participate in research
  - Participants must have all information necessary to make an informed decision
  - What about studies that require deception? Studies that compensate participants?
- Weighing benefits and risks of research
  - Is there a way to gain the information sought in the study that minimizes risk?
  - If risks cannot be reduced, how great is the potential benefit to society and to the subject?

Other Considerations

- Informed Consent
  - Subjects must affirmatively agree to participate in research
  - Participants must have all information necessary to make an informed decision
  - What about studies that require deception? Studies that compensate participants?
- Weighing benefits and risks of research
  - Is there a way to gain the information sought in the study that minimizes risk?
  - If risks cannot be reduced, how great is the potential benefit to society and to the subject?
  - Examples: Stanford Prison Experiment, Milgram Experiment

Other Considerations

- Informed Consent
  - Subjects must affirmatively agree to participate in research
  - Participants must have all information necessary to make an informed decision
  - What about studies that require deception? Studies that compensate participants?
- Weighing benefits and risks of research
  - Is there a way to gain the information sought in the study that minimizes risk?
  - If risks cannot be reduced, how great is the potential benefit to society and to the subject?
  - Examples: Stanford Prison Experiment, Milgram Experiment
vs. randomized drug trials

Expanding Principles for Equity (Urban Institute)

- Seek and include communities' interests in research design
- Seek out and incorporate communities' interpretation of the data
- Return data and research results to community members in a form they can use
- Be aware of how sensitive topics can affect people and communities
- Minimize the amount of personally identifiable information (PII) collected
- Avoid undue burden
- Share data to reduce the burden of duplicate data collection

Example: Project Nightingale

In 2019, the *Wall Street Journal* reported that the nation's second-largest health care provider, Ascension, provided Google with access to tens of millions of health records to develop AI tools that "make health records more useful, more accessible, and more searchable."

Example: Project Nightingale

In 2019, the *Wall Street Journal* reported that the nation's second-largest health care provider, Ascension, provided Google with access to tens of millions of health records to develop AI tools that "make health records more useful, more accessible, and more searchable."

- There was no informed consent from patients

Example: Project Nightingale

In 2019, the *Wall Street Journal* reported that the nation's second-largest health care provider, Ascension, provided Google with access to tens of millions of health records to develop AI tools that "make health records more useful, more accessible, and more searchable."

- There was no informed consent from patients
- There is potential harm from the release of health records

Example: Project Nightingale

In 2019, the *Wall Street Journal* reported that the nation's second-largest health care provider, Ascension, provided Google with access to tens of millions of health records to develop AI tools that "make health records more useful, more accessible, and more searchable."

- There was no informed consent from patients
- There is potential harm from the release of health records
- Google claims that the data was sufficiently protected and staff properly trained

Example: Project Nightingale

In 2019, the *Wall Street Journal* reported that the nation's second-largest health care provider, Ascension, provided Google with access to tens of millions of health records to develop AI tools that "make health records more useful, more accessible, and more searchable."

- There was no informed consent from patients
- There is potential harm from the release of health records
- Google claims that the data was sufficiently protected and staff properly trained
- Improving health care is a general societal good

Example: Project Nightingale

In 2019, the *Wall Street Journal* reported that the nation’s second-largest health care provider, Ascension, provided Google with access to tens of millions of health records to develop AI tools that “make health records more useful, more accessible, and more searchable.”

- There was no informed consent from patients
- There is potential harm from the release of health records
- Google claims that the data was sufficiently protected and staff properly trained
- Improving health care is a general societal good
- Because there is no “opt-in,” the data is not biased in favor of certain demographic groups

Example: Project Nightingale

In 2019, the *Wall Street Journal* reported that the nation’s second-largest health care provider, Ascension, provided Google with access to tens of millions of health records to develop AI tools that “make health records more useful, more accessible, and more searchable.”

- There was no informed consent from patients
- There is potential harm from the release of health records
- Google claims that the data was sufficiently protected and staff properly trained
- Improving health care is a general societal good
- Because there is no “opt-in,” the data is not biased in favor of certain demographic groups
- There are financial benefits that accrue to the companies, not the research subjects

Case Studies

1. Home DNA Testing
2. Crisis Text Line
3. Diversity in Faces (DiF) dataset

Case Studies

1. What are the relevant ethical principles and practices?
2. What concerns are there about violations of ethical principles?
3. How could the research have been conducted more ethically?

Time left:

08min 00sec

Controls:

⏮ ◁ ▷ ⏭

De-identification

- Researchers often promise ***anonymity*** or ***confidentiality*** to participants in order to reduce the risks of participating in research

De-identification

- Researchers often promise ***anonymity*** or ***confidentiality*** to participants in order to reduce the risks of participating in research
- There are two types of identifiers that must be removed before sharing this data:

De-identification

- Researchers often promise ***anonymity*** or ***confidentiality*** to participants in order to reduce the risks of participating in research
- There are two types of identifiers that must be removed before sharing this data:
  1. **Direct identifiers**: information that would be sufficient on its own to disclose an identity, such as names, addresses, and phone numbers.

De-identification

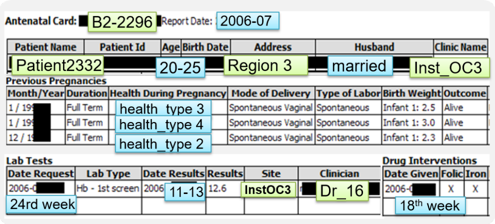
- Researchers often promise ***anonymity*** or ***confidentiality*** to participants in order to reduce the risks of participating in research
- There are two types of identifiers that must be removed before sharing this data:
  1. **Direct identifiers**: information that would be sufficient on its own to disclose an identity, such as names, addresses, and phone numbers.
  2. **Indirect identifiers**: information that *in combination* would be sufficient to disclose an identity

De-identification

- Researchers often promise ***anonymity*** or ***confidentiality*** to participants in order to reduce the risks of participating in research
- There are two types of identifiers that must be removed before sharing this data:
  1. **Direct identifiers**: information that would be sufficient on its own to disclose an identity, such as names, addresses, and phone numbers.
  2. **Indirect identifiers**: information that *in combination* would be sufficient to disclose an identity
- De-identification is the process of removing direct and indirect identifiers

DE-IDENTIFICATION

1. Remove direct identifiers



Antenatal Card: [redacted] B2-2296 [redacted] Report Date: 2006-07

Patient Name	Patient Id	Age	Birth Date	Address	Husband	Clinic Name
Patient2332	[redacted]	20-25	[redacted]	Region 3	[redacted] married [redacted]	Inst_OC3

Previous Pregnancies

Month/Year	Duration	Health During Pregnancy	Mode of Delivery	Type of Labor	Birth Weight	Outcome
1 / 19[redacted]	Full Term	health_type 3	Spontaneous Vaginal	Spontaneous	Infant 1: 2.5	Alive
1 / 19[redacted]	Full Term	health_type 4	Spontaneous Vaginal	Spontaneous	Infant 1: 3.0	Alive
12 / 1[redacted]	Full Term	health_type 2	Spontaneous Vaginal	Spontaneous	Infant 1: 2.3	Alive

Lab Tests

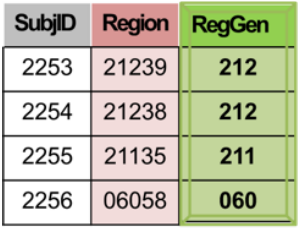
Date Request	Lab Type	Date Results	Results	Site	Clinician
2006-0[redacted] 24rd week	Hb - 1st screen	2006 11-13	12.6	InstOC3	[redacted] Dr_16 [redacted]

Drug Interventions

Date Given	Folic	Iron
2006-0[redacted] 18th week	X	X

De-identification

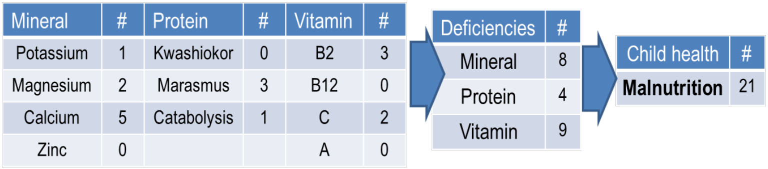
1. Remove direct identifiers
2. Aggregate or reduce the precision of a variable
   - Generalize the meaning of categories



SubjID	Region	RegGen
2253	21239	**212**
2254	21238	**212**
2255	21135	**211**
2256	06058	**060**

De-identification

1. Remove direct identifiers
2. Aggregate or reduce the precision of a variable
   - Generalize the meaning of categories
   - Collapse categories



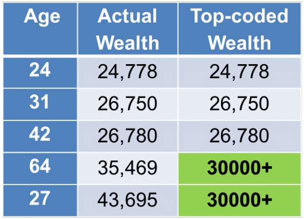
Mineral	#	Protein	#	Vitamin	#
Potassium	1	Kwashiokor	0	B2	3
Magnesium	2	Marasmus	3	B12	0
Calcium	5	Catabolysis	1	C	2
Zinc	0			A	0

Deficiencies	#
Mineral	8
Protein	4
Vitamin	9

Child health	#
Malnutrition	21

De-identification

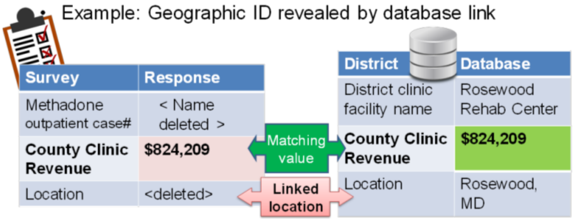
1. Remove direct identifiers
2. Aggregate or reduce the precision of a variable
   - Generalize the meaning of categories
   - Collapse categories
   - Restrict the upper or lower ranges



Age	Actual Wealth	Top-coded Wealth
24	24,778	24,778
31	26,750	26,750
42	26,780	26,780
64	35,469	**30000+**
27	43,695	**30000+**

De-identification

1. Remove direct identifiers
2. Aggregate or reduce the precision of a variable
    - Generalize the meaning of categories
    - Collapse categories
    - Restrict the upper or lower ranges
3. Anonymize keys that link to other datasets



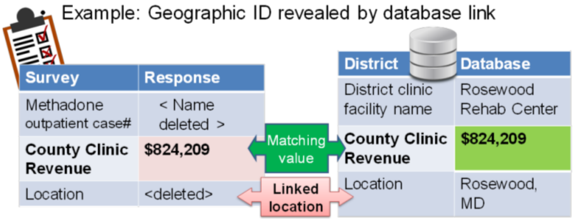
Example: Geographic ID revealed by database link

Survey	Response
Methadone outpatient case#	< Name deleted >
County Clinic Revenue	**$824,209**
Location	<deleted>

District	Database
District clinic facility name	Rosewood Rehab Center
County Clinic Revenue	**$824,209**
Location	Rosewood, MD

De-identification

1. Remove direct identifiers
2. Aggregate or reduce the precision of a variable
   - Generalize the meaning of categories
   - Collapse categories
   - Restrict the upper or lower ranges
3. Anonymize keys that link to other datasets
4. Maintain a master log of all replacements, aggregations, or removals and keep it in a secure location separate from the de-identified data files



Example: Geographic ID revealed by database link

Survey	Response
Methadone outpatient case#	< Name deleted >
County Clinic Revenue	**$824,209**
Location	<deleted>

District	Database
District clinic facility name	Rosewood Rehab Center
County Clinic Revenue	**$824,209**
Location	Rosewood, MD

De-identification Exercise

Time left:

05min 00sec

Controls:

⏮ ◁ ▷ ⏭