Data Disclosure Risk

Carla Graebner
Keshav Mukunda
SFU Library

Disclosure risk and you!

Intro

Think about these:

1. Data with serious disclosure risk are easy to recognize.
2. Only data collected explicitly for research have disclosure issues.
3. Disclosure of non-sensitive information is acceptable.

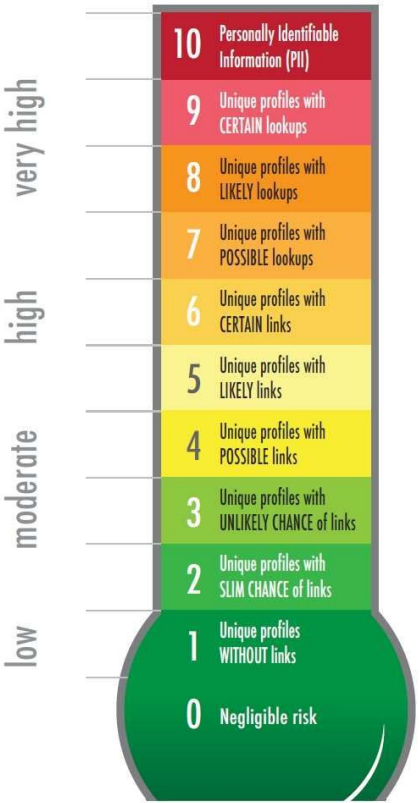
Data disclosure

Disclosure is the unauthorized release of information about an individual or an organization. Disclosive data could lead to the identification of specific individuals or organizations in a study.

Disclosure risk is the risk that a dataset could reveal information about individual participants in a study. This is rare with research data, but is growing as researchers collect more and more detailed data for studies.

Whether disclosure is intended or accidental doesn't matter – it's still harmful to individuals, hurts research, and could be violating laws.

Re-identification risk and harm



Low	0	No Harm
	1	Little Harm
	2	Humiliation
	3	Reputation Damage
Moderate	4	Financial Loss
	5	Health Threat
	6	Legal Jeopardy
High	7	Prison
	8	Physical Injury/Impairment
	9	Disfigurement
	10	Death

Re-identification case: Netflix

The *Netflix Prize* was an open competition to create the best algorithm for predicting users' film ratings based **only** on previous ratings.

Two researchers created a re-identification methodology to identify individual subscribers' records in the Prize dataset, and "using the Internet Movie Database as the source of background knowledge, [they] successfully identified the Netflix records of known users, uncovering their apparent political preferences and other potentially sensitive information."

Narayanan, A., and Shmatikov, V., *Robust De-anonymization of Large Sparse Datasets* 2008 IEEE Symposium on Security and Privacy, Oakland, CA, 2008, pp. 111-125. doi: 10.1109/SP.2008.33

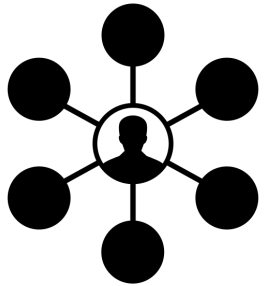
Types of identifying information

Direct identifiers are values that *explicitly* point to particular individuals. Often these are collected for survey administration. Any value that serves as an explicit name can be a direct identifier, for example:

- first and last names
- personal health numbers
- license numbers
- vehicle identifiers
- SIN
- institutional IDs
- IP addresses
- fingerprints
- e-mail addresses
- account numbers
- phone numbers
- voiceprints

id by Gregor Cresnar from the Noun Project

Types of identifying information



Quasi-identifiers (or **indirect identifiers)** are values that *could* be used in combination with one another to identify an individual in a dataset. For example:

race or ethnicity	age	postal code
size of household	annual income	height
place of birth	weight	marital status
medical treatments*	gender*	linguistic group

What increases disclosure risk?★

1. **Exact values** in the dataset. Exact age, income, etc.

2. **Small cell size**. Combinations of variables could produce unique or nearly unique individuals.

3. **Small or special populations**, or **small geographic areas**. For example, Statistics Canada uses a population size > 70,000 for reporting health regions, and the US HIPAA Privacy Rule has a value > 20,000 to release geographic information.

★ This is not a complete list!

What increases disclosure risk?★

4. **Longitudinal data** typically have a lot of information in them, for analytic reasons. Individual records can be connected across datasets.

5. **Sampling frame**. Knowing the larger frame from which a representative sample was created, the actual sample could be recreated.

6. **Linkages**. A dataset may have links to other external data sources (other studies, administrative data, social media). This external information could be used to statistically match records to disclose identities.

★ This is not a complete list!

How would you assess the risk?

Opinion poll on an upcoming election with respondent gender, age, and postal code.

Gauge by Dirk-Pieter van Walsum from the Noun Project

How would you assess the risk?

Survey of elementary school children and their mode of transportation to school.

Gauge by Dirk-Pieter van Walsum from the Noun Project

How would you assess the risk?

Survey of hospital discharges with dates of intake and discharge.

Gauge by Dirk-Pieter van Walsum from the Noun Project

Disclosure risk assessment process

1. Gather background information about the study, the methods, and the dataset. Are subjects vulnerable, could disclosure lead to a high level of harm? Was confidentiality promised? What consent was obtained?
2. Look for direct identifiers. Are ID variables randomized, or in a sequence based on geography or name?
3. Look for indirect identifiers. Geographic variables often influence the risk in all other variables. Any identifiers with low observation counts?
4. Check for any linkages with previous studies or outside information.

Classify variables by risk: example

A dataset has come to you for disclosure risk assessment!

(see the one-page handout for more details)

Classify each variable in the dataset as:

a. re-identification
b. sensitive
c. no risk
d. indeterminate

Are there any combinations of variables that increase disclosure risk?

Masking vs de-identification

Masking. Data fields are hidden or redacted, which greatly minimizes the risk of identifying an individual. This is typically applied to direct identifiers. Masked variables have no secondary research value.

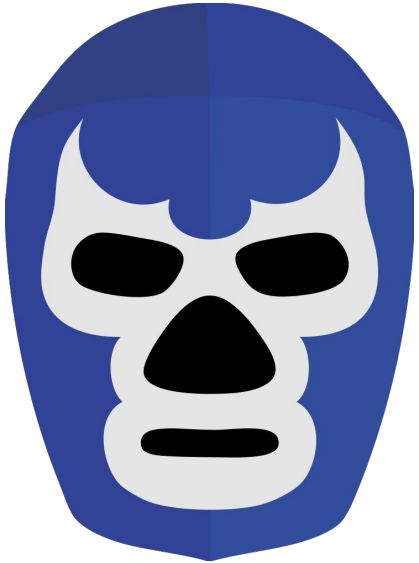
De-identification. Changes are made to the individual data values but they are not completely hidden. This is usually applied to demographic and socioeconomic fields. De-identified variables still have research analytic value.

Masking techniques

1. Create **pseudonyms** for data values using a secure method that is either irreversible (e.g., *hashing*) or difficult to reverse (e.g., *encryption*).

2. **Suppress** an entire field in the dataset, typically all direct identifiers.

3. Replace original variables with **randomized** ones. For numerical variables, random noise can be added such that the distribution of values remains the same as the original.



Approaches to de-identification

1. Use **lists** of data fields that need to be dealt with. This is a simple approach favoured by research and government organizations, but it doesn't give any assurance of low risk.

2. Discipline-based **heuristics** and rules of thumb have evolved over the years, but these are usually not based on specific risk metrics or evidence.

3. A more systematic **risk-based methodology** would be consistent with both current standards and best practices from regulators. The assumption here is that the risk of re-identification can be *estimated*.

Estimate risk of re-identification

Based on reasonable values to the likelihood of these risks:

1. Someone could **deliberately** attempt re-identification.
   Conservative estimate: maybe 10% of people with access to the data?

2. Someone might **inadvertently** recognize a record in the dataset.
   Depends on the geography of the dataset; how small is the group?

3. There is a data **breach**.
   Likelihood is based on previous evidence of reportable breaches by the relevant organization.

De-identification techniques

By **generalization**, to reduce the precision of a field.
e.g., *date of birth* becomes *month and year*, or *year*, or *decade*.

By **suppression**, replacing a value in the dataset with a NULL value.
e.g., a 50-year old mother may be unique in a birth registry, so her age value can be suppressed.

By **subsampling**, where only a random sample of the dataset is released.
e.g. only a 30% sample.

Terminology

Equivalence class: All records that have the same values on each of a selection of indirect identifiers.

Equivalence class size: The number of records in an equivalence class. These can change during de-identification.

k-anonymity: The size of each equivalence class in the dataset must be at least *k*.

This method protects against re-identification in the case where general background information is known about a specific individual in the dataset. Here there is a 1-in-k chance that the specific individual can be identified.

Gender	Age	Postal Code	Language	Disease
Female	27	R2C 0V6	Mandarin	Heart-related
Female	34	T5A 1P8	Mandarin	Pneumonia
Female	28	R2C 2G4	Cree	Heart-related
Male	24	G1B 1R8	Arabic	Pneumonia
Female	35	T5A 0X4	Spanish	Cancer
Male	23	G1B 2M3	Korean	Heart-related
Male	29	G1B 0P7	Mandarin	No condition
Female	29	R2C 1P2	Spanish	Cancer
Male	27	G1B 2E7	Spanish	Viral infection
Female	36	T5A 2J2	Cree	Viral infection

Gender	Age	Postal Code	Language	Disease
Female	27	R2C 0V6	NULL	Heart-related
Female	34	T5A 1P8	NULL	Pneumonia
Female	28	R2C 2G4	NULL	Heart-related
Male	24	G1B 1R8	NULL	Pneumonia
Female	35	T5A 0X4	NULL	Cancer
Male	23	G1B 2M3	NULL	Heart-related
Male	29	G1B 0P7	NULL	No condition
Female	29	R2C 1P2	NULL	Cancer
Male	27	G1B 2E7	NULL	Viral infection
Female	36	T5A 2J2	NULL	Viral infection

Gender	Age	Postal Code	Language	Disease
Female	27	R2C	NULL	Heart-related
Female	34	T5A	NULL	Pneumonia
Female	28	R2C	NULL	Heart-related
Male	24	G1B	NULL	Pneumonia
Female	35	T5A	NULL	Cancer
Male	23	G1B	NULL	Heart-related
Male	29	G1B	NULL	No condition
Female	29	R2C	NULL	Cancer
Male	27	G1B	NULL	Viral infection
Female	36	T5A	NULL	Viral infection

Gender	Age	Postal Code	Language	Disease
Female	[20, 30)	R2C	NULL	Heart-related
Female	[30, 40)	T5A	NULL	Pneumonia
Female	[20, 30)	R2C	NULL	Heart-related
Male	[20, 30)	G1B	NULL	Pneumonia
Female	[30, 40)	T5A	NULL	Cancer
Male	[20, 30)	G1B	NULL	Heart-related
Male	[20, 30)	G1B	NULL	No condition
Female	[20, 30)	R2C	NULL	Cancer
Male	[20, 30)	G1B	NULL	Viral infection
Female	[30, 40)	T5A	NULL	Viral infection

Gender	Age	Postal Code	Language	Disease
Female	[20, 30)	R2C	NULL	Heart-related
Female	[30, 40)	T5A	NULL	Pneumonia
Female	[20, 30)	R2C	NULL	Heart-related
Male	[20, 30)	G1B	NULL	Pneumonia
Female	[30, 40)	T5A	NULL	Cancer
Male	[20, 30)	G1B	NULL	Heart-related
Male	[20, 30)	G1B	NULL	No condition
Female	[20, 30)	R2C	NULL	Cancer
Male	[20, 30)	G1B	NULL	Viral infection
Female	[30, 40)	T5A	NULL	Viral infection

Anonymization tools ★

Does anyone have experience with these or other tools?

Amnesia -- data anonymization tool
https://amnesia.openaire.eu/

sdcMicro (an R package)
https://cran.r-project.org/package=sdcMicro

Privacy Analytics
https://privacy-analytics.com/

★ This is not a recommendation!

True or false? (or, “it depends”)

1. Data with serious disclosure risk are easy to recognize.
   False. *Inferential disclosure risk is often subtle.*

2. Data without direct identifiers such as names, addresses, or SINs may still have serious disclosure risk.
   True. *Inferential disclosure is usually the issue.*

3. Two variables may be enough for inferential disclosure.
   True. *e.g., birth dates or income, combined with geography.*

True or false? (or, “it depends”)

4. Geographic variables are always recognizable in data.
   False. *Geographic data is often imbedded in sampling clusters.*

5. Only data collected explicitly for research have disclosure issues.
   False. *Administrative data are usually highly disclosive.*

6. Disclosure of non-sensitive information is acceptable.
   Depends. Usually false, but depends on the terms of collection.