Data management

Øystein Ariansen Haaland
Oslo, October 2019

Health Registries for Research
Norway

NordForsk

Error search

Large data sets will almost always contain errors

What kind of errors have you encountered or heard about?

Take 2 minutes to talk with your neighbor

Error search

- IDs
- Missing data
- Contradictory information
- Impossible values

IDs

Each individual should have a unique ID

Sort data according ID

Identify and delete duplicate records (rows)

IDs

Each individual should have a unique ID

Sort data according ID

Identify and delete duplicate records (rows)

What happens if we do not?

Take 2 minutes to talk with your neighbor

IDs

Each individual should have a unique ID

Identify records with missing ID

IDs

Each individual should have a unique ID

Identify records with missing ID

How should they be handled?

IDs

Each individual should have a unique ID

Identify records with missing ID

Sort by other variables to see if there are duplicates

ID	Mat. age	GA at birth	Weight at birth	Pat. age	Birth year	Length at birth	Sex
113	29	37	3400	33	2007	53	m
.	29	37	3400	33	2007	53	m

IDs

Each individual should have a unique ID

Identify records with missing ID

One register
- Keep records with missing ID that are not duplicates

Several registers
- Impossible to merge without ID
- Delete records with missing ID

Missing data

Summarize missing values for each variable

Missing data

Summarize missing values for each variable

How do you handle missing data?

1-2 minutes to discuss with neighbor

Missing data

Summarize missing values for each variable

Delete records

- Risky if number of individuals decreases by a lot
- Not AS risky if number of individuals is still very large
- Not AS risky if observations are missing completely at random (MCAR)
    - Corresponds to rolling a die

Missing data

Summarize missing values for each variable

Model missingness (Height is 1.95m → Probably male)

- Multiple imputation
    - Replace missing observations by predictions based on other variables
    - Generate several imputed datasets
    - Perform analysis on all sets and combine at the end
    - Requires missing at random (MAR) or MCAR

Missing data

Summarize missing values for each variable

Model missingness (Height is 1.95m → Probably male)

- Maximum likelihood
    - Assume distribution of missingness
    - Build distribution directly into model
    - Not as common as multiple imputation
    - No general software (that I am aware of)

Missing data

Summarize missing values for each variable

Keep missing as its own value

Smoking status for pregnant women:
“Yes” and “Missing” have similar characteristics on other variables

Contradictory information

Similar variables may contain contradictory information

How can two different variables contain contradictory information?

Take 2 minutes with your neighbor

Contradictory information

Similar variables may contain contradictory information

Cleft lip with or without cleft palate: Yes

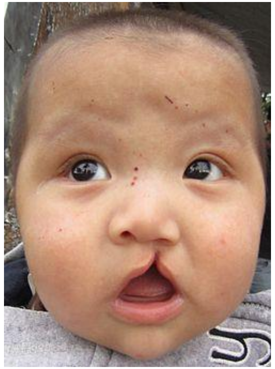

Photo: Wikipedia

Contradictory information

Similar variables may contain contradictory information

Cleft lip with or without cleft palate: Yes

Cleft palate only: Yes

Photo: Wikipedia

Contradictory information

Similar variables may contain contradictory information

Cleft lip with or without cleft palate: Yes
Cleft palate only: Yes

Year of death: 2008 (Cause of death registry)

Contradictory information

Similar variables may contain contradictory information

Cleft lip with or without cleft palate: Yes
Cleft palate only: Yes

Year of death: 2008 (Cause of death registry)
Alive in 2018: Yes (Birth registry)

Contradictory information

Similar variables may contain contradictory information

Cleft lip with or without cleft palate: Yes
Cleft palate only: Yes

Year of death: 2008 (Cause of death registry)
Alive in 2018: Yes (Birth registry)

No fixed solution. Use logic or experience.

Impossible values

Sometimes variables have impossible values

How can this happen?

Take 2 minutes and discuss with your neighbor

Impossible values

Sometimes variables have impossible values
- Year of birth: 1880
- Birth weight: 8400g

Detect by plotting variables
- Histograms

Try to adjust (1880 → 1980)

Consider as missing
- Often one particular value even means “missing” (typically -1, 0 or 9)?

Poor quality (variables)

Variables with a lot of missing observations

How would you handle such variables?

Take 2 minutes and discuss with your neighbor

Poor quality (variables)

Variables with a lot of missing observations

- Possible to use other variables?

    - Mother’s education instead of father’s education

Poor quality (variables)

Variables with a lot of missing observations

- Possible to use other variables?
    - Mother’s education instead of father’s education
- Possible to create composite variables?
    - Parent’s highest education

Poor quality (variables)

Variables with a lot of missing observations

- Possible to use other variables?
    - Mother’s education instead of father’s education
- Possible to create composite variables?
    - Parent’s highest education
- Reason for missingness?
    - Education at age 25, but many parents are too young

Poor quality (variables)

Variables with a lot of missing observations

- Use techniques described earlier
    - Model missingness
    - Use as own category
- Does “missing” just mean “no”?
    - Merging data sets
    - Rare conditions

Poor quality (individuals)

Individuals with a lot of missing observations

- Not very common in registries
- Reason for missingness?
    - Foreigner
    - Too young
    - Other reasons?
- Not a problem if there are few such individuals
    - Unless the reason for the missingness is related you your research question!
    - Education in immigrants vs. risk of cancer

Summary

The possibilities for errors in registry research is almost endless

No fixed solutions

Use your head

Generally do not delete records just because they have missing data