Avoiding the Peak of Inflated Expectations: Common Misconceptions in Population Data

Peter Christen[1] and **Rainer Schnell**[2]

[1] The Australian National University, Australia
peter.christen@anu.du.au

[2] University Duisburg-Essen, Germany
rainer.schnell@uni-due.de

German RLC

This work was supported by Deutsche Forschungsgemeinschaft Grant 407023611 and Universities Australia

The long version: A draft paper

- Peter Christen and Rainer Schnell (2021 / 22): Common Misconceptions about Population Data arXiv:2112.10912 https://doi.org/10.48550/arXiv.2112.10912

- Comments are welcome!

Population data

- A shift in many domains of science towards the use of large and complex databases that cover (nearly) whole populations.
- These replace – or at least enrich – traditional data collection methods (such as surveys or experiments).
- Following McGrail et al. (IJPDS, 2018), we define population data as *“data about people at the level of a population”.*
- Ideally, a population database should contain one record per entity (person), and all elements (fields / attributes / variables) of relevance for a study for all entities.

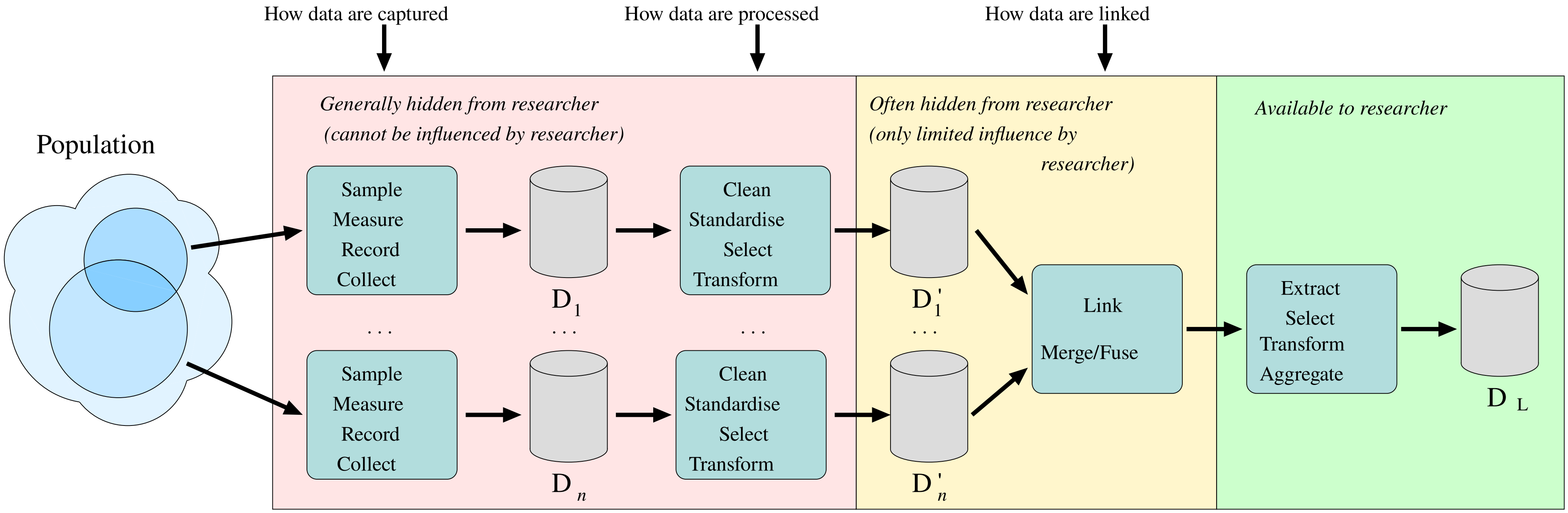
Rapid adoption and inflated expectations

- Due to the perceived advantages of population data, the number of projects adopting existing databases for research and planning is increasing.
- The use of buzzwords like Big data, AI, and machine learning, in the context of population data seems to suggest for non-technical users and decision makers that any kind of question can be answered when analysing population databases.
- Neither data quality issues (how population data was captured and processed), nor the techniques used to link population data, are clear to decision makers and researchers who are used to smaller data sets.

Data quality and misconceptions of population data

- There is much work on general data quality, but only little specific to population data.
- Therefore, the kind of problems we consider here are usually underestimated by non-specialists, leading to inflated expectations.
- Such over-expectations might cause costly mis-management in areas such as public health or in government decision making.
- Failing population data projects might even result in the loss of trust in governments and science by the public.
- Hence, misconceptions on population data must be avoided.

The pathway of population data


How data are captured
How data are processed
How data are linked
Population
Generally hidden from researcher (cannot be influenced by researcher)
Often hidden from researcher (only limited influence by researcher)
Available to researcher
Sample Measure Record Collect
D 1
Clean Standardise Select Transform
. . .
Sample Measure Record Collect
D n
Clean Standardise Select Transform
D' 1
D' n
Link Merge/Fuse
Extract Select Transform Aggregate
D L

- Many issues are due to humans being involved in the processes that generate population data.
- The mistakes and choices people make, changing requirements, novel computing and data entry systems, limited resources and time, as well as decision making influenced by political or for economical reasons.

Types of misconceptions

- In our arXiv paper we identified 32 misconceptions across the three stages – here we give one example each.
- **Data capturing**: People and organisations can behave mischievous or fraudulent.
  - The `Jedi’ religion in national censuses.
  - Results for at-home COVID tests (a positive result can mean loss of income).
  - Financial incentives for COVID test numbers done by labs.
- **Data processing**: Handling of missing data is done in different ways by different organisations and / or data entry personnel (and often outside the control of a researcher).
  - What is entered into a name field / attribute can be a blank, or an indicator for missing information is used, such as “John Doe”, “Baby”, “n/a”, etc.
  - These can be difficult to detect.
- **Data linkage**: Temporal aspects when records were recorded or updated are rarely considered (unlikely the same date for all records)
  - For example, education data in the German Social Security database is most commonly added when a record was generated (a person’s first job), but rarely updated.

Recommendations

- It should be made to clear to researchers and decision makers that population data cannot be generated fast, free of costs, and without errors.
- The limits of databases collected by humans on human behaviour should be element of scientific education in all domains expected to use such data.
- Researchers should aim to get involved in the production of data planned to be used for their research.
- At least detailed information how the data was captured, processed, and linked is required (Metadata).
- Publish data issues and lessons learnt after a population data project has been implemented.

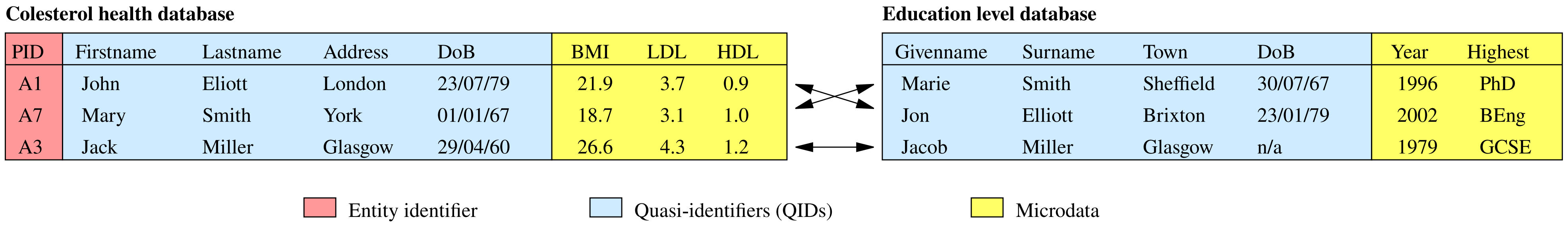
Common Misconceptions in Population Data

- For more please see our arXiv paper: https://doi.org/10.48550/arXiv.2112.10912

Thanks

peter.christen@anu.edu.au
rainer.schnell@uni-due.de

Characteristics of population data



Colesterol health database

PID	Firstname	Lastname	Address	DoB	BMI	LDL	HDL
A1	John	Eliott	London	23/07/79	21.9	3.7	0.9
A7	Mary	Smith	York	01/01/67	18.7	3.1	1.0
A3	Jack	Miller	Glasgow	29/04/60	26.6	4.3	1.2

Education level database

Givenname	Surname	Town	DoB	Year	Highest
Marie	Smith	Sheffield	30/07/67	1996	PhD
Jon	Elliott	Brixton	23/01/79	2002	BEng
Jacob	Miller	Glasgow	n/a	1979	GCSE

🟥 Entity identifier 🟦 Quasi-identifiers (QIDs) 🟨 Microdata

- Generally, each entity (person) in a population database is represented by one or more records (rows).
- Each record is made of attributes (fields) that can be categorised into identifiers (unique or quasi-identifiers, QIDs) and microdata (payload data).
- QIDs are generally not used in research studies, however they are often important for linking databases.
- QID values, such as names and addresses, can change over time, contain errors and variations, or be missing.

Misconceptions due to data capturing (1)

(1) A population database contains all individuals in a population *(tourists in national health databases)*

(2) The population covered in a database is well defined

(3) Population databases contain complete information for all records in the database

(4) All records in a population database are within the scope of interest *(dead people)*

(5) Each individual in a population is represented by a single record in a database

(6) Records in a population database always refer to real people *(test records like "Tony Test living in Testville")*

Misconceptions due to data capturing (2)

(7) Errors in personal data are not intentional *(my phone number is +44 756 1234 5678)*

(8) Certain personal details do not change over time.

(9) Personal name variations are incorrect *(Gail versus Gayle)*

(10) Coding systems do not change over time *(ICD-10 / -11)*

(11) Data definitions are unambiguous *(COVID onset date based on symptoms, test collection, or diagnosis)*

(12) Temporal data aspects do not matter

(13) The meaning of data is always known

(14) Missing data have no meaning *(employment for children)*

Misconceptions due to data capturing (3)

(15) All records in a population database were captured using the same process *(multiple data entry personnel)*

(16) Attribute values are correct and valid

(17) Data values are in their correct attributes *(first and last names “Paul” and “Thomas”)*

(18) Data validation rules produce correct data *(1 January)*

(19) All relevant data have been captured

(20) Population data provide the same answers as survey data *(what people are and what they do, versus what they say they are and do)*

(21) Population data are always of value

Misconceptions due to data processing

(22) Data processing can be fully automated *(still an iterative semi-automated process with much involvement of manual decision making based on domain expertise)*

(23) Data processing is always correct *(sometimes no single ‘correct’ solution, furthermore mistakes can be made in using or configuring software)*

(24) Aggregated data are sufficient for research *(ecological fallacy, describing the mistake of an aggregate relationship implying the same relationship for individuals)*

(25) Metadata are correct, complete, and up-to-date

Misconceptions due to data linkage

(26) A linked data set corresponds to an actual population

(27) Population databases represent the conditions of people at the same time *(unlikely all source records were collected or updated on the same date)*

(28) A linked data set contains no duplicates

(29) A linked data set is unbiased *(different linkage rates for different sub-populations)*

(30) Attribute values in linked records are correct

(31) Linkage error rates are independent of database size

(32) Modern record linkage techniques can handle databases of any sizes