International Journal of Population Data Science

Journal Website: www.ijpds.org

Thirty-three myths and misconceptions about population data: from data capture and processing to linkage

Peter Christen[1,2,*] and Rainer Schnell[3]

Submission History

Submitted:	24/09/2022
Accepted:	09/12/2022
Published:	31/01/2023

[1]School of Computing, The Australian National University, Canberra, ACT 2600, Australia
[2]Scottish Centre for Administrative Data Research (SCADR), University of Edinburgh. UK
[3]Methodology Research Group, University Duisburg-Essen, Germany

Abstract

Databases covering all individuals of a population are increasingly used for research and decision-making. The massive size of such databases is often mistaken as a guarantee for valid inferences. However, population data have characteristics that make them challenging to use. Various assumptions on population coverage and data quality are commonly made, including how such data were captured and what types of processing have been applied to them. Furthermore, the full potential of population data can often only be unlocked when such data are linked to other databases. Record linkage often implies subtle technical problems, which are easily missed. We discuss a diverse range of myths and misconceptions relevant for anybody capturing, processing, linking, or analysing population data. Remarkably, many of these myths and misconceptions are due to the social nature of data collections and are therefore missed by purely technical accounts of data processing. Many are also not well documented in scientific publications. We conclude with a set of recommendations for using population data.

Keywords
data quality; record linkage; data linkage; personal data; administrative data; data editing; data errors

*Corresponding Author:
Email Address: peter.christen@anu.edu.au (Peter Christen)

https://doi.org/10.23889/ijpds.v8i1.2115

Introduction

Many domains of science increasingly use large administrative or operational databases that cover whole populations to replace – or at least enrich – traditional data collection methods such as surveys or experiments [1–3]. This kind of data are now seen as a crucial strategic resource for research [4–6]. Governments and businesses have also recognised the value that large population databases can provide to improve decision-making [7–9].

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 [10].

However, often neither data quality issues (how population data were captured and processed) nor the techniques used to link population data, are clear to decision makers and researchers used to smaller data sets. Although there is much work on general data quality, very little specific to population data is published or taught in data science courses.

Therefore, the kind of problems we consider in this paper are usually underestimated by non-specialists, leading to inflated expectations. Such over-expectations might cause costly mismanagement in areas such as public health or in government decision-making. Furthermore, failing population data projects, such as census operations or health surveillance, might even result in the loss of trust in governments and science by the public [8, 11]. In the context of research, myths and misconceptions[1] about population data can lead to wrong outcomes of research studies that can result in conclusions with severe negative impact [12, 13].

We focus on personal data about individuals covering (nearly) whole populations. Following a recent definition of *Population Data Science* [14], we define population data as *data about people at the level of a population*. The focus on populations is important, as it refers to the scale and complexity of the data being considered. These make (manual) processing and assessment of data quality that are normally conducted on the much smaller data sets used in traditional medical studies or social science surveys challenging.

Personal data include personally identifiable information (PII) [15], such as the names, addresses, or dates of birth of people. Most administrative and operational data collected by governments and businesses can also be categorised as personal data, including electronic health records, (online) shopping baskets, and people's educational, social security, financial, and taxation data [16].

A crucial aspect of population data is that they are not primarily collected for research, but rather for operational or administrative purposes [5, 10, 14, 17]. As a result, researchers have much less control over such data and how they are processed, and only limited ways to learn about the data's provenance, making it unclear if such data are fit for the purpose of a specific research study [18]. Quoting Brown et al. [19], "in science, three things matter: the data, the methods used to collect the data (which give them their probative value), and the logic connecting the data and methods to conclusions". When both the data and their collection are outside the control of a researcher, conducting proper science can become challenging.

Most of the discussion about the use of population data for research has been about privacy and confidentiality [15, 20, 21]. Much less consideration has been given to how data quality and assumptions about such data can influence the outcomes of a research study. For administrative data, decades of experience have shown that many unexpected patterns (such as unusual combinations of attribute values) in large databases are due to data errors instead of anything of actual interest [10]. The same can be said about population covering databases of personal data.

Not much has been written about the characteristics of personal data and how they differ from other types of data, such as scientific or financial data. During the writing of this article, we conducted extensive literature research to find scientific publications, government reports, or technical white papers that describe experiences or challenges when dealing with population databases. The number of identified publications was scarce, indicating that many challenges encountered and lessons learnt are not being shared, even though these would be of high value to anybody who works with population data.

Many of the misconceptions we discuss below are therefore drawn from our several decades of experience working with large real-world population databases and in collaborative projects with both private and public sector organisations across diverse disciplines (including health, finance, and national statistics).

We do not advocate to abandon the use of population data for research or decision making. Rather, with this work we aim to improve the handling of population databases. By presenting common misconceptions about population data, we hope to show how researchers could identify and avoid the resulting traps.

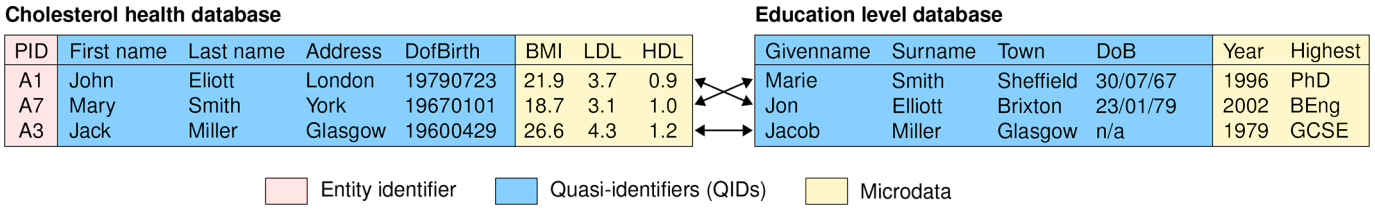
Characteristics and uses of population data

Population data are observational data that most commonly occur in administrative or operational databases as held by government or business organisations [10]. As we illustrate in Figure 1, in their most general form each *entity* (person) in a population is represented by one or more *records* (like rows in a spreadsheet) in such a database, each consisting of a set of *attributes* (or *variables*).

The attributes that represent entities in population data can be categorised into *identifiers* and *microdata*. The first category can either be an *entity identifier* (such as patient identifiers) that is (supposed to be) unique for each person in a population [22]. Alternatively they can be a group of attributes that, when enough are combined, become unique for each individual. Such attributes are known as *quasi-identifiers* (QIDs), and they include names, addresses, date and place

[1] According to Merriam Webster (https://www.merriam-webster.com), a myth is a "popular belief or tradition that has grown up around something or someone", while a misconception is "a wrong or inaccurate idea or conception". For brevity, throughout the paper we will only use misconception.

Figure 1: Two small example population databases, where only one contains unique entity identifiers (the PID attribute), but both contain quasi-identifiers (QIDs) and microdata



The first database consists of eight attributes and the second of six, where both have three records. The QIDs have been used to link records across these two databases. Note the variations and missing values in QIDs.

of birth, and so on [15]. Many of the misconceptions about population data are about entity identifiers and QIDs, and how they are captured, processed, and used to link population databases to transform them into a form suitable for a research study.

The second component of personal data are known as *microdata* [15] (or payload data), and they include the data of interest for research, such as the medical, educational, financial, or location details of individuals. Much of these microdata are highly sensitive if they are connected to QID values because combined they can reveal private personal details about an individual. Research into data anonymisation and disclosure control [23, 24] is addressing how sensitive microdata can be made available to researchers in anonymised form.

It is commonly recognised that isolated population databases are of limited use when trying to investigate and solve today's complex challenges [14, 18], such as how a pandemic spreads through a population. Therefore, many projects that are based on population data employ *data* or *record linkage* [22, 25] to link all records that correspond to the same person across two or more databases. Linked records at the level of individuals, rather than aggregated data, are generally required to allow the development of accurate predictive models.

We illustrate the general pathway of population data in Figure 2, and discuss each stage (how data are captured, processed, and linked) and the corresponding misconceptions below. While many of these misconceptions seem obvious, they are often not taken into consideration when population data are used in research studies or for decision making.

Many data issues are due to humans being involved in the processes that generate population data, including 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. Research managers, and policy and decision makers, often also assume any kind of question can be answered with highly accurate and unbiased results when analysing population databases [18].

While the literature on general data quality is broad [26, 27], given the wide-spread use of personal data at the level of populations it is surprising that only little published work seems to discuss data quality aspects specific to personal data [15, 28, 29]. One reason is due to the perceived sensitivity of this kind of data: Population data are generally covered by privacy regulations, such as the European General Data Protection Regulation (GDPR) or the US Health Insurance Portability and Accountability Act (HIPAA) [15], and the processes and methods employed are often covered by confidentiality agreements. Furthermore, detailed data aspects are generally not included in scientific publications where the focus is on presenting the results obtained in a research study rather than the steps taken to obtain these results.

Misconceptions about population data

Following from Figure 2, we categorise the identified misconceptions due to how population data are captured, how such data are processed, and how they are linked. We do not discuss any misconceptions related to the analysis of population data – how to prevent pitfalls in statistical data analysis and machine learning has been discussed extensively elsewhere [30, 31].

Misconceptions due to data capturing

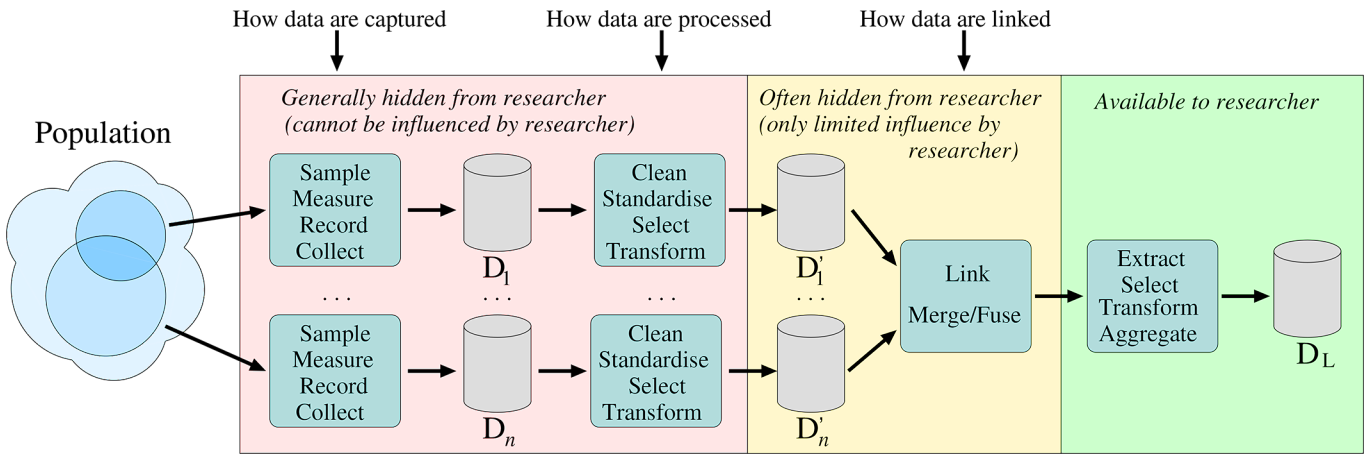
Misconceptions under this category can occur due to how information about people is captured. We refer to data capture as any processes and methods that convert information from a source into electronic format. This involves decisions about selecting or sampling individuals from an actual real population to be included into a population database, how to measure the characteristics of these people, and the methods employed to actually collect and record this information.

Many data capturing methods and processes involve humans who can make mistakes or behave in unexpected ways, or equipment that can malfunction or be misconfigured. Data capturing methods include manual data entry, optical character recognition, automatic speech recognition, and sensor readings (like biometrics from fingerprint readers and smart watches, or location traces from smart phones). While each of these methods can introduce specific data quality issues (such as keyboard typing or scanning errors), there are some common misconceptions about capturing population data.

(1) A population database contains all individuals in a population

Even databases that are supposed to cover whole populations, such as taxation or census databases, very likely have subpopulations that are under-represented or absent (for example children or residents without a fixed address).

Figure 2: The general pathway of population data from their source to a researcher's computer



We assume $\mathbf{D}_1$ to $\mathbf{D}_n$ (with $n \geq 2$) are the source databases that likely cover different subsets of a real population. These databases are generally processed independently by their respective data owners into the databases $\mathbf{D}'_1$ to $\mathbf{D}'_n$, and then linked and processed further (often by different organisations) into the linked database $\mathbf{D}_L$ to make it fit for the purpose of a research study

Individuals who do not have a national health identifier number, like tourists or international students, and therefore are not eligible for government health services might be missing from population health databases. There are also always individuals who refuse to participate in government services for personal reasons, which can be influenced by their ethnicity, religion, or political beliefs.

The digital divide [32], the division of people who have access to digital services and media versus those who do not, likely also results in biased population databases. Younger and more affluent individuals are more likely included in databases collected using digital services compared to older people, migrants, or those with a lower socio-economic status.

Organisations might also refuse to provide their data for commercial reasons or due to confidentiality concerns. Not all patients will be included in a health database if, for example, certain private hospitals refuse to participate in a research study. This will likely introduce bias [33], since patients with higher income will more likely go to exclusive private hospitals.

The assumption that all individuals in a population are represented in a database leads to the illusion that it will be possible to identify any subpopulation of interest and explore this group of people, even if it is very small [10].

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

The reasons for records about individuals to be included (or not) into a database are crucial to understanding the population covered in that database. Some population databases can be based on mandatory inclusion of individuals (think of government taxation or residency databases) while others are based on voluntary or self-selected inclusion (think of health databases such as registries where patients can opt-out when asked to provide their medical details).

The definitions and rules used to extract records about individuals into a population database might not be known to those who are processing and linking the database, and even less likely to the researchers who will be analysing it [18]. Furthermore, these rules might differ between organisations or change over time, or they might contain minor inconsistencies or mistakes such as ill-defined age or date ranges. For example, COVID-19 cases can be included into a database based on the date of symptom onset, collection of samples, or diagnosis [34], where each of these will result in different numbers of records being added to a database each day.

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

Many population databases contain different pieces of information for different sets of records. This sparseness is because large databases are commonly generated by compiling different individual databases each only covering a part of a population, or by collecting records over time where changes in regulations and data capturing methods and processes can lead to different attributes (both QIDs and microdata) being collected. The resulting sparse patterns may prevent many statistical analyses of the data if not all relevant information has been captured.

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

Individuals might have left the population of interest for a research study because the criteria for inclusion in a database are no longer given. For example, a person may have died (although this might be of interest in itself) or have left the geographical area of a study. Because many organisations are not notified of an individual's death or relocation until sometime after the event (in some cases never), at any given time a population database will likely contain records that are outside the scope of interest. Including records about these people can affect research studies as well as operational systems.

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

A common assumption is that population databases are generated and maintained as person-level databases, with

one record per person in a population. However, it is not uncommon for a population database to contain duplicate records referring to the same person due to errors or variations in QID values. While data entry validation may prevent exact duplicates, fuzzy or approximate duplicates [22] might be missed by (automatic) checks. In real-world settings, the same person can therefore be registered multiple times at different institutions and their duplicate records are not being identified as referring to the same individual.

Some duplicates are very difficult to find, for example women who change their last name and address when they get married and therefore only their first name, gender, and place and date of birth stay the same. The flip side is that several people with highly similar personal details (similar QID values), such as twins who only have different first names, might not be recognised as two individuals but instead as duplicates.

Duplicate records are possible even if entity identifiers (such as social security numbers or patient identifiers) are available that should prevent multiple records by the same individual from being added to a population database. Due to human behaviour and errors such identifiers are not always provided or entered correctly. It might therefore be beneficial to apply deduplication (also known as internal linkage) [22] to a population database to identify any possible duplicate records that refer to the same person, and then handle such duplicates appropriately.

An example are people in a social security database who should have one record but were registered multiple times because they changed their names or addresses and might have forgotten their previous registration (and their unique identifier number), or they might be interested in multiple registrations to obtain social benefits more than once. Duplicate records have even been identified in domains where high data quality is crucial, such as in voter registration databases [35].

(6) Records in a population database always refer to real people

Real-world databases can contain records of people who never existed [36]. These can be records added to train data entry personnel or test software systems. Often these records are not removed from a database. While potentially easy to identify by humans ('Tony Test' living in 'Testville'), these records are difficult to detect by data cleaning algorithms because they were designed to have the characteristics of real people and often contain high amounts of variations and errors.

In population databases collected from social media platforms, complete records might furthermore correspond to fake users, and increasingly even AI (artificial intelligence) bots that generate human-like content [17, 18]. In one reported case of research fraud, records have been created in order to boost the size of a database being analysed in order to make research findings more convincing [37].

(7) Errors in personal data are not intentional

There are social, cultural, as well as personal reasons why individuals would decide to provide incorrect personal details. These include fear of surveillance by governments, trying to prevent unsolicited advertisements from businesses, or simply the desire to keep sensitive personal data private. Fear of data breaches and how personal data are being (mis)used by organisations are clearly influencing the reluctance of individuals to provide their details unless deemed necessary [8, 38]. In domains such as policing and criminal justice, faked QID values such as name aliases occur commonly as criminals try to hide their actual identities [39].

Incorrectly provided data might only be modified slightly from a correct value (such as a date of birth a few days off), be changed completely (a different occupation given), not be provided at all (no value entered if an input field is not mandatory), or be made up (such as a telephone number in the form of '1234 5678').

The decision to provide incorrect or withhold personal details is dependent upon the context in which this information is being collected. Generally, it is less likely for an individual to provide incorrect values (for non-crucial QIDs) on an official government form or when opening a bank account compared to when ordering a book in an online store. However, the opposite might be true in cases where an individual does not trust the institution that is collecting their data [40].

(8) Certain personal details do not change over time

While some personal details, such as names and addresses, are known to change over time for many individuals, it is often assumed that others are fixed at birth. These include ethnic and gender identification, as well as place and country of birth. In many population databases, ethnic identification is self-reported, where the available categories depend upon how a society values different subpopulations. For example, the US Bureau of the Census has repeatedly changed the details of their policy regarding ethnic groups[2]. Socially influenced attributes might be changed by individuals over time, a recent example being the Black Lives Matter movement which made many individuals become more proud to be of colour. It can therefore be problematic to use these values in the context of, for example, longitudinal data analysis or record linkage [15].

It is even possible for values fixed at birth to change in the real world. An example is the Eastern German city of Chemnitz, which from 1953 until 1990 was named Karl-Marx-Stadt. Individuals born during that period have a country of birth (German Democratic Republic) and a place of birth that both do not exist anymore.

(9) Personal name variations are incorrect

People's names are a key component of QIDs in many population databases. Unlike with most general words, for many personal names there are multiple spelling variations (such as 'Gail', 'Gayle', and 'Gale'), and all of them are correct [22]. When data are entered, for example over the telephone, differently sounding name variations might be recorded for the same individual due to mispronunciation or misunderstanding.

Furthermore, there are many cultural aspects of names, including different name orders and structures, ambiguous transliterations from non-roman into the roman alphabet, or name changes over time for religious reasons, to name a few. Name variations are a known problem when names are compared between records when linking databases [41].

[2] https://www.census.gov/about/our-research/race-ethnicity.html.

Working with names can therefore be a challenging undertaking that requires expertise in the cultural and ethnic aspects of names [42].

(10) Coding systems do not change over time

Categorical QID values and microdata are often coded using systems such as the International Standard Classification of Occupations (ISCO)[3] or the International Classification of Diseases (ICD)[4], the latter currently in its eleventh revision. It is commonly assumed that such codes are fixed over time and unique in that a certain item, such as an occupation or disease, is only assigned one code, and that this assignment does not change. However, many coding systems are revised over time with new codes being added, outdated and unused codes being removed, and whole groups of codes being recoded (including codes being swapped). A database might therefore contain codes which are no longer valid. Furthermore, at a given time, different revisions of coding systems might occur in a population database, for example if the transition to a new version of a system is not conducted at the same time by all organisations that contribute to that database.

An example are the codes of the Australian Pharmaceutical Benefit Scheme (PBS) [43], where the antidepressant Venlafaxine had the code N06AE06 until 1995, when it was given the code N06AA22, which was then changed to N06AX16 in 1999. Using such codes to group or categorise records can therefore lead to wrong results of a research study if records have been collected over time.

(11) Data definitions are unambiguous

Many population databases contain information that is based on definitions such as how to categorise records or create categorical QID or microdata values. As with coding systems, data definitions can change over time, and they can also be interpreted differently. A recent example are the definitions for death or hospitalisations due to COVID-19 infections, where different US states used various definitions that resulted in databases that could not be used for comparative analysis [12].

Unless metadata (see misconception 25 below) are available that clearly describe such definitions and their changes, it can be difficult to identify the effects of any changed definitions because any such change might have subtle effects on the characteristics of only some individuals in the population of interest for a research study.

(12) Temporal data aspects do not matter

Given the dynamic nature of personal details, the time and date when population data are captured and stored in a database can be crucial because differences in data lag can lead to inconsistent data that are not suitable for research studies [1, 12]. If it takes different amounts of time for different organisations to capture data about the same events then clearly these data are not comparable, resulting in misreporting for example of the numbers of daily COVID-19 deaths [34], vaccination rates [1, 44], or education levels of migrants [45].

Daily, weekly, monthly, or seasonal aspects can influence data measurements, as can events such as public holidays and religious festivities which likely only affect certain subpopulations. For example, daily reporting of new COVID-19 infections might be limited on weekends and the beginning of a week due to less testing and delayed laboratory diagnosis on weekends. Similar delays will happen during and after public holidays.

Data corrections are not uncommon, especially in applications where there is an urgent need to provide initial data as quickly as possible, for example to better understand a global pandemic [1, 34]. Later updates and corrections of data might not be considered, leading to wrong conclusions of research studies.

(13) The meaning of data is always known

It is not uncommon for population databases to contain attributes that are not (well) documented. These can include codes without known meaning, irrelevant sequence numbers, or temporary values that have been added at some point in time for some specific purpose. If no documentation is available, database managers are generally reluctant to remove such attributes. As a result, spurious patterns might be detected if such attributes are included into a data analysis.

(14) Missing data have no meaning

Missing data are common in many databases [46]. They can lead to problems with data processing, linking, and analysis [15]. Missing data can occur at the level of missing records (no information is available about certain individuals in a population), missing attributes (some QIDs or microdata contain no data values for all records in a database), or missing QID or microdata values for individual records (specific missing attribute values) [18].

There are different categories of missing data [46, 47]. In some cases a missing value does not contain any valuable information, in others it can be the only correct value (children under a certain age should not have an occupation), or it can have multiple interpretations. A missing value for a question about religion in a census, for example, can mean an individual does not have a religion or they choose not to disclose it. Missing data can also occur in settings where resources are limited and therefore data entries had to be prioritised, such as in busy emergency departments [34].

Care must therefore be taken when considering missing data. Removing attributes or even records with missing values, or imputing missing values [41, 47], can result in errors and structural bias being introduced into a population database that can lead to incorrect outcomes of a research study.

(15) All records in a population database were captured using the same process

Since population databases are often collected over long periods of time and wide geographical areas, records are commonly generated or entered by a large number of staff, giving rise to different interpretations of data entry rules. For example, if an input field requires a mandatory value, humans will enter all kinds of unstandardised indicators for

[3] https://www.ilo.org/public/english/bureau/stat/isco/.

[4] https://www.who.int/standards/classifications/classification-of-diseases.

missingness, ranging from single symbols (like '–' or '.'), acronyms ('NA' or 'MD'), to texts explaining the missing data (like 'unknown'). If a population database is compiled from independent organisations these different interpretations of data entry rules will cause the need for standardisation before analysis. Manual data entry such as typing can furthermore lead to different error characteristics between data entry personnel [15], resulting in subsets of records in a population database with different data quality.

Data might be captured at different temporal and spatial resolution, such as postcodes or city names only versus detailed street addresses. As a result, the characteristics of both QID and microdata values can differ between subsets of records in a population database, making their comparison and analysis challenging [34].

(16) Attribute values are correct and valid

Any data values captured, either by some form of sensor or manually entered into a database, can be subject to errors coming from equipment malfunction, human data entry (typing mistake), or cognitive mistakes (such as confusion about the data required or difficulties recalling correct information), or even malicious intent [18, 27].

In the medical domain, manual typing errors, wrong interpretations of forms (think of handwritten prescriptions by doctors), entering values into the wrong input fields, or making mistakes interpreting instructions (when prescribing drugs) are commonly occurring mistakes, where rates ranging from 2 to 514 mistakes per 1,000 prescriptions have been reported [48].

While data validation tools can detect values outside the range or domain of what is valid (such as 31st of February) [27], without external validation it is generally not feasible to ascertain the correctness of any given value. If a patient is really 42 years old can only be validated if authoritative information (likely from an external database) about the patient's true age is available.

Furthermore, while individual QID values in a given record can each be valid, they might contradict each other. For example, a record with first name 'John' and gender 'f' likely contains one QID value that is incorrect. Many, but not all, such contradictions can be identified and corrected using appropriate edit constraints [41].

(17) Data values are in their correct attributes

Data entry personnel do not always enter values into the correct attribute. Many Asian and some Western names, for example, can be used interchangeably as first and last names, leading to misinterpretation. For example, 'Paul', 'Thomas', 'Chris', and 'Dennis' are all used as first and last names. The ordering of how first and last names are written can also depend on the culture and origin of an individual.

(18) Data validation rules produce correct data

To ensure data of high quality, many data management systems contain rules that need to be fulfilled when data are being captured. For example, registering a new patient in a hospital requires both a valid address and a valid date of birth. In some cases, such as in emergency admissions, not all of this information will be known. Due to such data validation rules, default values are often used. A common example is the 1st of January being used for individuals with unknown day and month of birth. While these are valid, if not handled properly such defaults can result in skewed data distributions that can adversely affect research studies. Data entry personnel might also have ad-hoc rules they apply in order to bypass data entry requirements and to ensure any entered records fulfil all data validation steps.

(19) All relevant data have been captured

Because the primary purpose of most population databases is not their use for research studies, not all relevant information that is of importance for a given study might be available for all records in a database, or it might only be available in subsets of records. This can, for example, be due to changes in data entry requirements over time, or because data might have been withheld by the owner due to confidentiality concerns or for commercial reasons, or data might only be provided in aggregated or anonymised form. If a statistical model is generated from such data, a probable causal variable might be missed, since it is not captured at all.

Data that are not available are known as *dark data* [46], data we do not know about but that could be of interest to a research study. As a result, certain required or desired information might be missing for a given research study, making a given population database less useful or requiring the use of alternative data for that study.

(20) Population data provide the same answers as survey data

Population data, as captured from administrative or operational databases, are about what people are and what they do [10]. This is unlike survey data where commonly questions about attitudes, beliefs, expectations, or intentions are asked with the aim to understand the behaviour of people. Factual information about people can provide different answers compared to questions about what they claim to do, while inferring people's beliefs from their behaviour might not be possible.

(21) Population data are always of value

Both private and public sector organisations increasingly make databases publicly available to facilitate their analysis by researchers. However, many of these databases either lack metadata or context for them to be of use, or they are aggregated or anonymised due to privacy and confidentiality concerns [24]. A main reason for this is because past experiences have shown that sensitive personal information about individuals can sometimes be re-identified even from supposedly anonymised databases [38, 49].

Population data without context are unlikely to be of use for research studies. A database of QID values (such as names and addresses) without any (or only limited) microdata is, by itself, of little value for research. Having the educational level of individuals in a database only becomes useful if this database can be linked with other data at the level of individuals. Furthermore, due to the dynamic

nature of people's lives, population data become out of date quickly and therefore need to be updated regularly. Without adequate metadata (see misconception 25 below), context, useful detailed microdata, and regular updates, many publicly available databases are of little value for research.

Misconceptions due to data processing

It is rare for population databases to be used for research without any processing being conducted. The organisation(s) that collect population data, and those that further aggregate, link (or otherwise integrate) such data, as well as the researchers who will analyse the data, all will likely apply some form of data processing [34].

Processing can include data cleaning and standardisation, parsing of free text values, transformation of values, numerical normalisation, recoding into categories, imputation of missing values, and data aggregation. The use of different database management systems and data analysis software can furthermore result in data being reformatted internally before being stored and later extracted for further processing and analysis. Each component of a data pipeline can result in both explicit (user applied) as well as implicit (internally to software) data processing being conducted, leading to various misconceptions.

(22) Data processing can be fully automated

Much of the processing of population data has to be conducted in an iterative fashion, where data exploration and profiling lead to a better understanding of a database which in turns helps to apply appropriate data processing techniques [27]. This process requires manual exploration, programming of data specific functionalities, domain expertise with regard to the provenance and content of a database, as well understanding of the final use of a database. Data processing is often the most time-consuming and resource intensive step of the overall data analytics pipeline, commonly requiring substantial domain as well as data expertise. In national statistical agencies, it has been reported that as much as 40% of resources are used on data processing [18].

Time and resource constraints might mean not all desired data processing can be accomplished. Manual editing and evaluation is also unlikely to be possible on large and complex population databases, and therefore compromises have to be made between data quality and timeliness for a database to be made available for research [18].

(23) Data processing is always correct

There are often multiple methods available to process data, for example to normalise numerical values, impute missing data, or standardise free-format text [26, 41]. Converting 'dirty' data into 'clean' data can therefore result in incorrectly cleaned data [22]. Sometimes there is no single correct value for a given ambiguous input value. For example, within a street address, the abbreviation 'St' can stand for either 'Street' or 'Saint' (as used in a town name like 'Saint Marys').

Given data processing commonly involves human efforts, mistakes in the use and configuration of software can lead to incorrect data processing, as can bugs in or the use of different or outdated versions of software. The use of unsuitable tools for a given project (such as spreadsheet software instead of a proper database management system or statistical analysis software) can furthermore results in mistakes when data are being processed. It has been reported [50] that on the 2nd October 2020 a total of 15,841 positive COVID-19 cases (around 20%) in England were missed because when recording daily cases an old file format of the Microsoft Excel spreadsheet software was used which allowed a maximum of 65,536 rows. Software features such as auto-completion and automatic spelling correction can furthermore lead to the wrong correction of unusual but valid data values that are not available in a dictionary.

As low quality data (possibly due to mistakes in data processing) are identified over time, improvements can be made to data processing methods that result in improved data quality. While this generally means that data quality can improve over time, changes in the actual data (data trends) as well as in data capturing might also mean that data processing again becomes less efficient, leading to lower data quality. Examples include data cleaning rules that have been correct in the past but might generate wrong results, such as when postcode boundaries are changing, or coding systems are revised.

(24) Aggregated data are sufficient for research

Highly aggregated data, for example at the level of states, counties, or large geographical units, are hardly of use for scientific research intending causal statements. Although results based on aggregated data might seem to be interesting, the number of possible alternative explanations for the same set of facts based on aggregated data is usually so large that no definitive conclusions are possible.

A major problem here is the *ecological fallacy*, describing the mistake of an aggregate relationship implying the same relationship for individuals [51]. For example, if increased mortality rates are observed in regions where vaccination rates are high, the false conclusion would be that vaccinated people have a higher probability to die. But actually the reverse might be true: People observing other people dying might be more willing to get vaccinated.

How data are aggregated depends on how aggregation functions are defined and interpreted. Weekly counts can, for example, be summed Monday to Sunday or alternatively Sunday to Saturday. Data that are aggregated inconsistently, or at different levels of aggregation, will unlikely be of use for any research studies (or only after additional data processing has been conducted).

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

Metadata (also known as *data dictionaries*) are describing a database, how it has been created, populated, and its content captured and processed. Metadata include aspects such as the source, ownership, and provenance of a database, licensing and access limitations, costs, description of all attributes including their domains and any coding systems used, summary statistics and descriptions of data quality dimensions [15], as well as any data cleaning, imputation, editing, processing, transformation, aggregation, and linkage that was conducted on the source

database(s) to obtain a given population database. Relevant documentation should be provided, including who conducted any data processing using what software (and which version of it), and containing a revision history of that documentation. Metadata are crucial to understand the actual structure, content, and quality of a database at hand.

Unfortunately, metadata are often not available, or they are incomplete, out of date, they need to be purchased, or can only be obtained through time-consuming approval processes [5]. A lack of metadata can lead to misunderstandings during data processing, linking and analysis, wasted time, misreporting of results, or can make a population database altogether useless [12].

Misconceptions due to data linkage

Linking databases is generally based upon comparing the QID values of individuals, such as people's names, addresses, and other personal details (as illustrated in Figure 1), to find records that refer to the same person [25, 41]. These QID values, however, can contain errors, be missing, and they can change over time. This can lead to incorrect linkage results even when modern linkage methods are employed [52]. Linking databases can therefore be the source of various misconceptions about a linked data set. While all of the following misconceptions can occur when data from two sources are being linked (or even when duplicate records need to be identified within a single database [22]), in situations where records from multiple (more than two) sources have to be linked any of these misconceptions can become even more challenging and more difficult to deal with.

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

Due to data quality issues and the record linkage technique(s) employed [22], a linked data set likely contains wrong links (Type I errors, two records referring to two different individuals were linked wrongly) while some true links have been missed (Type II errors, two records referring to the same person were not linked) [53]. The performance of most record linkage techniques can be controlled through parameters [52], allowing a trade-off between these two types of errors. Many linked data sets with different error characteristics can therefore be generated by changing parameter settings, where each generated data set provides an approximation of the actual population it is supposed to represent.

(27) Population databases represent the conditions of people at the same time

Data updates on individuals often occur at different points in time, usually when an event such as a medical condition occurs, or a data error is detected during a triggered data transaction such as a payment. In the German Social Security database, for example, education is entered when a record for a given person is newly created, but is not regularly updated afterwards. Therefore, highly trained professionals might have a record stating a low educational level because they were pupils at the time of their first paid job. Similar issues have been identified in Sweden for the education data of migrants, where for different subpopulations their educational levels are updated at varying rates [45]. As a result, records that represent the same individual can have different values (in both QID and microdata attributes) across the databases being linked. The assumption that the QID values of all records in the databases being linked are up-to-date might therefore not be correct, and outdated information can lead to wrong linkage results [15].

Data corrections and updates can furthermore occur when incorrect historical data are being discovered and errors rectified [1]. Unless it is possible to re-conduct a linkage, which is unlikely for many research studies due to the efforts and costs involved in such a process, a linked data set might contain errors which have influenced the conclusions of the original study.

(28) A linked data set contains no duplicates

When linking databases, pairs or groups of records that refer to the same individual might not be linked correctly (missed true links, Type II errors). One reason for this to occur is if a wrong entity identifier has been assigned to an individual (by accident or on purpose), as has been reported even in voter registration databases [35]. If a linkage requires agreement on such unique identifier values, then two records with different values in the unique identifier will not be linked even if they have many highly similar (or even agreeing) QID values. Another reason is if crucial QID values of an individual have changed over time, such as both their name and address details, or are missing, resulting in two records that are not similar with each other [22]. Therefore, many linked data sets do contain more than one record for some individuals in a population.

(29) A linked data set is unbiased

Linkage errors generally do not occur at random [5], rather they depend upon the characteristics of the actual QID values of individuals, which can differ in diverse subpopulations. Examples include name structures of migrants that are different from the traditional Western standard of first, middle, and last name formats [54], or different rates of mobility (address changes) for young versus older people. As a result, there can be structural bias in a linked data set in subpopulations defined by ethnic or social categories, age, or gender (for example if women are more likely to change their names compared to men when they get married) [55].

Recent work has also shown that even small amounts of linkage error can result in large effects on false negative (Type II) error rates in research studies. This is especially the case with small sample sizes that can occur with the rare effects that are often sought to be identified via record linkage from large population databases [56]. If the aim of a study is to analyse certain (potentially small) subpopulations, or compare, for example, health aspects between subpopulations, then a careful assessment of the potential bias introduced via record linkage is of crucial importance [53].

(30) Attribute values in linked records are correct

Given a supposedly correct link, there might be contradicting attribute values in the corresponding records. Data fusion is the process of resolving such inconsistencies, where often a decision needs to be made which of many available fusion operations to apply [57]. Even if the links made between records are correct, how records are fused or merged can therefore introduce errors both into QID values as well as microdata.

For example, assume three records that refer to the same person have been linked correctly, where each record contains a different salary value. Should the average, median, minimum, maximum, or the most recent of these three salary values be used for the fused record of this individual? How data fusion is conducted needs to be discussed with the researchers who will be analysing a linked data set because depending upon the fusion operation applied substantially different outcomes will potentially be obtained.

(31) Linkage error rates are independent of database size

The QID values used to link records can be shared by multiple individuals, potentially thousands in the case of city and town names or with popular first and last names. Therefore, when larger databases are being linked, the number of record pairs with the same QID values likely increases, resulting in more highly similar pairs. Making correct classification becomes increasingly challenging as there are more possibly matching pairs. Generalising linkage quality results obtained on small data sets in published studies to much larger population sized real-world databases can therefore be dangerous.

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

Many researchers, especially in the computer science and statistical domains, who develop record linkage techniques do not have access to large real-world databases due to the sensitive nature of population data. As a result, novel linkage techniques are often evaluated on small public benchmark data sets or on synthetically generated data [15]. While error rates for linkages obtained on such data sets can provide evidence of the superiority of a novel technique over existing methods, assuming that this new technique will produce comparable high quality linkage results on larger real-world databases is not guaranteed.

(33) Linkage techniques and their settings are easily transferrable

If a linkage method together with its parameter settings (for example how blocking is conducted, how values are compared, and how a classification threshold is set) has been successful deployed in a given linkage project, this does not mean that the same method and settings will provide comparable high linkage quality results on a different linkage project. For each linkage project, different methods and corresponding parameter settings will need to be established. Furthermore, the same holds even when linking large disparate population databases, where for different subpopulations different optimal parameter settings (such as classification thresholds) will need to be identified. Finally, repeated linkages over time, for example a yearly update, may also require different parameter settings.

Conclusions and recommendations

Due to misconceptions such as the ones we have discussed, the much hyped promise of big data requires some careful considerations when personal data at the level of populations are used for research studies or decision making. Given population data are increasingly used in many domains of science, as illustrated in Figure 2, researchers will potentially have less and less control over the quality of the data they are using for their studies and any processing done on these data [19]. They likely will also have only limited information about the provenance and other metadata that is needed to fully understand the characteristics and quality of their data. Because population data are commonly sourced from organisations other than where they are being analysed [10, 18], these limitations are inherent to this kind of data.

There are no (simple) technical solutions to detect and correct many of the misconceptions we have discussed. What is required is heightened awareness by anybody working with population data. While our list of misconceptions is unlikely to be exhaustive, our aim was to show that there is a broad range of issues that can lead to misconceptions. The following recommendations might help to recognise and overcome such potential misconceptions.

- If possible, data scientists and researchers should aim to get involved in the capturing, processing, and linking of any data they plan to use for their research. This involves discussions with database owners about what data to collect in what format, how to ensure high quality of these data, and that adequate metadata are collected. It also means proper planning and designing of the information systems that are required for data capture, processing and linkage, and their adequate support as well as updates over time. It is vital to have the involvement of data scientists and data policy managers in these processes.

- If at all possible, data scientists and IT personnel who are processing and linking population data need to work in close collaboration with the researchers who will conduct the actual analysis of these data. Both technical and strategic aspects of a project that involves population data should ideally be discussed with the analysts, data scientists, database managers, developers, project managers, as well as the owners of the population databases being used, processed, and linked. Forming multi-disciplinary teams with members skilled in data science, statistics, domain expertise, as well as 'business' aspects of research [5], is crucial for successful projects that rely upon population data. Interaction between data and domain experts might mean that a project based on population data becomes an iterative endeavour where data might have to be recaptured, reprocessed, and relinked until they are suitable for a research study.

- Cross-disciplinary training should be aimed at improving complementary skills [5]. Having data scientists who also have domain specific expertise will be highly valuable in any project involving population data. Equally crucial is for any researcher, no matter what their domain, to understand how modern data processing, record linkage, and data analytics methods work, and how these methods might introduce bias and errors into the data they are using for their research studies. Training in data exploration and data cleaning methods as well as data quality issues should be part of any degree that deals with data, including statistics, quantitative social science, computer science, and public health.

- While extensive methodologies about how to deal with uncertainties, bias, and data quality in surveys have been developed [58, 59], there is a lack of corresponding rigorous methods that can be employed on large population databases. The Big data paradox [60], the illusion that large databases automatically mean valid results, requires new statistical techniques to be developed. While certain data quality issues can be identified (and potentially corrected) automatically [27], novel data exploration methods are needed to identify more subtle data issues where traditional methods are inadequate.

- A crucial aspect is to have detailed metadata about a population database available, including how the database was captured, and any processing and linkage applied to it. All relevant data definitions need to be described, and information about all sources and types of uncertainties need to be collected [10]. Detailed data profiling and exploration should be conducted by researchers before a population database is being analysed so any unexpected characteristics in their data can be identified.

- Existing guidelines and checklists, such as RECORD [61] and GUILD [62], should be employed and adapted to other research domains. Frameworks such as the Big Data Total Error method [18] can be adapted for population data to better characterise errors in such data. Furthermore, data management principles such as FAIR (Findable, Accessible, Interoperable, Reusable) [63] should be adhered to, although in some situations the sensitive nature of personal data [15] might limit or prevent such principles from being applied. In such situations, at least metadata and any software used in a study should be made public in an open research repository. Following these principles, guidelines, and checklists will allow data scientists and researchers to highlight to data custodians that having access to metadata would be highly beneficial for their work.

- The lack of publications that describe practical challenges when dealing with population data can result in the misconceptions we have discussed here. We therefore encourage increased publication of data issues and the sharing of experiences with the scientific community about lessons learnt, as well as best practice approaches being implemented when dealing with population data.

We have discussed some aspects in modern scientific processes that are rarely considered when population data are being used for research studies or decision making. Since good data management is a key aspect of good science [19, 63], it is vital for anybody who uses population data to be aware of underlying assumptions concerning this kind of data. We hope the misconceptions and recommendations given here will help to identify and prevent misleading conclusions and poor real-world decisions, making population data the new oil of the big data era.

Acknowledgements

We like to thank S. Bender, S. Redlich, J. Reinhold, and C. Nanayakkara for their critical and helpful comments, A. Plöger for help with producing the figures, and S. Weiand for technical support. P. Christen likes to acknowledge the support of the University of Leipzig and ScaDS.AI, Germany, where parts of this work was conducted while he was funded by the Leibniz Visiting Professorship. P. Christen also gratefully acknowledges the support of the UK Economic and Social Research Council (ESRC), ES/W010321/1. The work by R. Schnell was supported by the Deutsche Forschungsgemeinschaft Grant 407023611. We finally like to thank the two anonymous reviewers whose comments have helped to improve our work.

Ethics statement

No ethics approval was required for this study because no actual data was involved.

Statement on conflicts of interest

The authors have no conflicts of interest.

References

1. Valerie C Bradley, Shiro Kuriwaki, Michael Isakov, Dino Sejdinovic, Xiao-Li Meng, and Seth Flaxman. Unrepresentative big surveys significantly overestimated US vaccine uptake. *Nature* 600, 695–700 (2021). https://doi.org/10.1038/s41586-021-04198-4

2. Liran Einav and Jonathan Levin. Economics in the age of Big data. *Science*, 346(6210), 2014. https://doi.org/10.1126/science.1243089

3. Ian Foster, Rayid Ghani, Ron S Jarmin, Frauke Kreuter, and Julia Lane, editors. *Big Data and Social Science*. CRC Press, Boca Raton, 2017. https://doi.org/10.1201/9781315368238

4. Roxanne Connelly, Christopher J Playford, Vernon Gayle, and Chris Dibben. The role of administrative data in the Big data revolution in social science research.

Social Science Research, 59(Supplement C):1–12, 2016. https://doi.org/10.1016/j.ssresearch.2016.04.015

5. Louisa Jorm. Routinely collected data as a strategic resource for research: priorities for methods and workforce. *Public Health Research Practice*, 25 (4), 2015. https://doi.org/10.17061/phrp2541540

6. Nathaniel D Porter, Ashton M Verdery, and S Michael Gaddis. Enhancing Big data in the social sciences with crowdsourcing: Data augmentation practices, techniques, and opportunities. *PloS one*, 15(6):e0233154, 2020. https://doi.org/10.1371/journal.pone.0233154

7. Susan Athey. Beyond prediction: Using Big data for policy problems. *Science*, 355(6324):483–485, 2017. https://doi.org/10.1126/science.aal4321

8. Jim Isaak and Mina J Hanna. User data privacy: Facebook, Cambridge Analytica, and privacy protection. *IEEE Computer*, 51(8):56–59, 2018. https://doi.org/10.1109/MC.2018.3191268

9. Foster Provost and Tom Fawcett. *Data Science for Business: What you need to know about Data Mining and Data-Analytic Thinking*. O'Reilly Media, Inc., 2013. https://learning.oreilly.com/library/view/data-science-for/9781449374273/.

10. David J Hand. Statistical challenges of administrative and transaction data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 181 (3):555–605, 2018. https://doi.org/10.1111/rssa.12315

11. Valerie Braithwaite. Beyond the bubble that is Robodebt: How governments that lose integrity threaten democracy. *Australian Journal of Social Issues*, 55(3):242–259, 2020. https://doi.org/10.1002/ajs4.122

12. Stephanie E Galaitsi, Jeffrey C Cegan, Kaitlin Volk, Matthew Joyner, Benjamin D Trump, and Igor Linkov. The challenges of data usage for the United States' COVID-19 response. *International Journal of Information Management*, 59:102352, 2021. https://doi.org/10.1016/j.ijinfomgt.2021.102352

13. Sarah Giest and Annemarie Samuels. 'For good measure': data gaps in a Big data world. *Policy Sciences*, 53(3):559–569, 2020. https://doi.org/10.1007/s11077-020-09384-1

14. Kimberlyn M McGrail, Kerina Jones, Ashley Akbari, Tellen D Bennett, Andy Boyd, et al. A position statement on population data science: The science of data about people. *International Journal of Population Data Science*, 3(1), 2018. https://doi.org/10.23889/ijpds.v3i1.415

15. Peter Christen, Thilina Ranbaduge, and Rainer Schnell. *Linking Sensitive Data*. Springer, Heidelberg, 2020. https://doi.org/10.1007/978-3-030-59706-1

16. Katie Harron, Chris Dibben, James Boyd, Anders Hjern, Mahmoud Azimaee, Mauricio L Barreto, and Harvey Goldstein. Challenges in administrative data linkage for research. *Big Data and Society*, 4(2):1–12, 2017. https://doi.org/10.1177/2053951717745678

17. Florian Keusch and Frauke Kreuter. Digital trace data: Modes of data collection, applications, and errors at a glance. In *Handbook of Computational Social Science, Vol 1*, pages 100–118. Taylor and Francis, 2021. https://doi.org/10.4324/9781003024583-8

18. Paul Biemer. Errors and inference. In: Ian Foster, Rayid Ghani, Ron S Jarmin, Frauke Kreuter, and Julia Lane, editors, *Big Data and Social Science*, chapter 10, pages 265–297. CRC Press, Boca Raton, 2017. https://doi.org/10.1201/9781315368238

19. Andrew W Brown, Kathryn A Kaiser, and David B Allison. Issues with data and analyses: Errors, underlying themes, and potential solutions. *Proceedings of the National Academy of Sciences*, 115(11):2563–2570, 2018. https://doi.org/10.1073/pnas.1708279115

20. Alessandro Acquisti, Laura Brandimarte, and George Loewenstein. Privacy and human behavior in the age of information. *Science*, 347(6221):509–514, 2015. https://doi.org/10.1126/science.aaa1465

21. Eric Horvitz and Deirdre Mulligan. Data, privacy, and the greater good. *Science*, 349(6245):253–255, 2015. https://doi.org/10.1126/science.aac4520

22. Peter Christen. *Data Matching*. Springer, Heidelberg, 2012. https://doi.org/10.1007/978-3-642-31164-2

23. George Duncan, Mark Elliot, and Juan-José Salazar-González. *Statistical Confidentiality: Principles and Practice*. Springer, New York, 2011. https://doi.org/10.1007/978-1-4419-7802-8

24. Mark Elliot, Elaine Mackey, and Kieron O'Hara. *The Anonymisation Decision-making Framework 2nd Edition: European Practitioners' Guide*. UKAN Manchester, 2020. https://msrbcel.files.wordpress.com/2020/11/adf-2nd-edition-1.pdf.

25. Katie Harron, Harvey Goldstein, and Chris Dibben. *Methodological Developments in Data Linkage*. John Wiley and Sons, 2015. https://doi.org/10.1002/9781119072454

26. Carlo Batini and Monica Scannapieco. *Data and Information Quality*. Springer, Heidelberg, 2016. https://doi.org/10.1007/978-3-319-24106-7

27. Won Kim, Byoung-Ju Choi, Eui-Kyeong Hong, Soo-Kyung Kim, and Doheon Lee. A taxonomy of dirty data. *Data Mining and Knowledge Discovery*, 7(1):81–99, 2003. https://doi.org/10.1023/A:1021564703268

28. Mark Smith, Lisa M Lix, Mahmoud Azimaee, Jennifer E Enns, Justine Orr, Say Hong, and Leslie L Roos. Assessing the quality of administrative data for research: a framework from the Manitoba Centre for Health Policy. *Journal of the American Medical Informatics Association*, 25(3):224–229, 2018. https://doi.org/10.1093/jamia/ocx078

29. Mihnea Tufiș and Ludovico Boratto. Toward a complete data valuation process. challenges of personal data. *ACM Journal of Data and Information Quality*, 13(4):1–7, 2021. https://doi.org/10.1145/344726

30. Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer, New York, 2 edition, 2009. https://doi.org/10.1007/978-0-387-84858-7

31. Patrick Riley. Three pitfalls to avoid in machine learning. *Nature*, 572 (7767):27–29, 2019. https://doi.org/10.1038/d41586-019-02307-y

32. Jan Van Dijk. *The Digital Divide*. John Wiley and Sons, Cambridge, UK, 2020. https://www.wiley.com/en-us/The+Digital+Divide-p-9781509534463.

33. Richard Shaw, Katie Harron, Julia Pescarini, Elzo Pereira Pinto Junior, Mirjam Allik, Andressa Siroky, Desmond Campbell et al. Biases arising from linked administrative data for epidemiological research: a conceptual framework from registration to analyses. *European Journal of Epidemiology*, 1–10, 2022. https://doi.org/10.1007/s10654-022-00934-w

34. Rinette Badker, Kierste Miller, Chris Pardee, Ben Oppenheim, Nicole Stephenson, Benjamin Ash, Tanya Philippsen, Christopher Ngoon, Partrick Savage, Cathine Lam, et al. Challenges in reported COVID-19 data: best practices and recommendations for future epidemics. *BMJ Global Health*, 6 (5):e005542, 2021. https://doi.org/10.1136/bmjgh-2021-005542

35. Fabian Panse, André Düjon, Wolfram Wingerath, and Benjamin Wollmer. Generating realistic test datasets for duplicate detection at scale using historical voter data. In: *International Conference on Extending Database Technology*, 570–581, 2021. https://doi.org/10.5441/002/edbt.2021.67

36. Peter Christen, Ross W Gayler, Khoi-Nguyen Tran, Jeffrey Fisher, and Dinusha Vatsalan. Automatic discovery of abnormal values in large textual databases. *ACM Journal of Data and Information Quality*, 7(1-2):1–31, 2016. https://doi.org/10.1145/2889311

37. Servick, Kelly, and Martin Enserink. The pandemic's first major research scandal erupts. Science, 368(6495), 1041-1042, 2020. https://doi.org/10.1126/science.368.6495.1041

38. Eric Siegel. *Predictive Analytics: The Power to Predict who Will Click, Buy, Lie, Or Die*. John Wiley and Sons, 2013. https://learning.oreilly.com/library/view/predictive-analytics-the/9781118416853/?ar=.

39. Sarah Larney and Lucy Burns. Evaluating health outcomes of criminal justice populations using record linkage: the importance of aliases. *Evaluation Review*, 35(2):118–128, 2011. https://doi.org/10.1177/0193841X11401695

40. Finn Brunson and Helen Nissenbaum. *Obfuscation: A User's Guide for Privacy and Protest*. MIT Press, 2015. https://doi.org/10.7551/mitpress/9780262029735.001.0001

41. Thomas N Herzog, Fritz J Scheuren, and William E Winkler. *Data Quality and Record Linkage Techniques*. Springer Verlag, 2007. https://doi.org/10.1007/0-387-69505-2

42. Patrick McKenzie. Falsehoods programmers believe about names, https://www.kalzumeus.com/2010/06/17/falsehoods-programmers-believe-about-names/, 2010.

43. Leigh Mellish, Emily A Karanges, Melisa J Litchfield, Andrea L Schaffer, Bianca Blanch, Benjamin J Daniels, Alicia Segrave, and Sallie-Anne Pearon. The Australian Pharmaceutical Benefits Scheme data collection: a practical guide for researchers. *BMC Research Notes*, 8(1):1–13, 2015. https://doi.org/10.1186/s13104-015-1616-8

44. Alex Berry. Germany's vaccination rate could be higher than previously thought, https://p.dw.com/p/41oi7. Deutsche Welle, 7 Oct, 2021.

45. Samaneh Khaef. Registration of immigrants' educational attainment in Sweden: an analysis of sources and time to registration. *Genus*, 78(1):1–20, 2022. https://doi.org/10.1186/s41118-022-00159-5

46. David J Hand. *Dark Data: Why What You Don't Know Matters*. Princeton University Press, 2020. https://doi.org/10.2307/j.ctvmd85db

47. Roderick J Little and Donald B Rubin. *Statistical Analysis with Missing Data*. Wiley, Hoboken, 3 edition, 2020. https://doi.org/10.1002/9781119482260

48. Giampaolo P Velo and Pietro Minuz. Medication errors: Prescribing faults and prescription errors. *British Journal of Clinical Pharmacology*, 67(6): 624–628, 2009. https://doi.org/10.1111/j.1365-2125.2009.03425.x

49. Latanya Sweeney. *k*-anonymity: A model for protecting privacy. *International Journal of Uncertainty Fuzziness and Knowledge Based Systems*, 10 (5):557–570, 2002. https://doi.org/10.1142/S0218488502001648

50. Thiemo Fetzer and Thomas Graeber. Measuring the scientific effectiveness of contact tracing: Evidence from a natural experiment. *Proceedings of the National Academy of Sciences*, 118(33), 2021. https://doi.org/10.1073/pnas.2100814118

51. Glenn Firebaugh. Statistics of ecological fallacy. In: Neil J Smelser and Paul B Baltes, editors, *International Encyclopedia of the Social and Behavioral Sciences*, pages 4023–4026. Pergamon, Oxford, 2001. https://doi.org/10.1016/B978-0-08-097086-8.44017-1

52. Olivier Binette and Rebecca Steorts. (Almost) all of entity resolution. *Science Advances*, 8(12):eabi8021, 2022. https://doi.org/10.1126/sciadv.abi8021

53. James Doidge and Katie Harron. Reflections on modern methods: linkage error bias. *International Journal of Epidemiology*, 48(6):2050–2060, 2019. https://doi.org/10.1093/ije/dyz203

54. Louise Mc Grath-Lone, Nicolas Libuy, David Etoori, Ruth Blackburn, Ruth Gilbert, and Katie Harron. Ethnic bias in data linkage. *The Lancet Digital Health*, 3(6):e339, 2021. https://doi.org/10.1016/S2589-7500(21)00081-9

55. Megan A Bohensky, Damien Jolley, Vijaya Sundararajan, Sue Evans, David V Pilcher, Ian Scott, and Caroline A Brand. Data linkage: a powerful research tool with potential problems. *BMC Health Services Research*, 10(346):1–7, 2010. https://doi.org/10.1186/1472-6963-10-346

56. Sarah Tahamont, Zubin Jelveh, Aaron Chalfin, Shi Yan, and Benjamin Hansen. Dude, where's my treatment effect? Errors in administrative data linking and the destruction of statistical power in randomized experiments. *Journal of Quantitative Criminology*, 37(3):715–749, 2021. https://doi.org/10.1007/s10940-020-09461-x

57. Jens Bleiholder and Felix Naumann. Data fusion. *ACM Computing Surveys*, 41(1):1–41, 2008. https://doi.org/10.1145/1456650.1456651

58. Johnny Blair, Ronald F Czaja, and Edward A Blair. *Designing Surveys: A Guide to Decisions and Procedures*. Sage, Thousand Oaks, 3 edition, 2014. https://us.sagepub.com/en-us/nam/designing-surveys/book235701

59. Giles Reid, Felipa Zabala, and Anders Holmberg. Extending TSE to administrative data: A quality framework and case studies from Stats NZ. *Journal of Official Statistics*, 33(2), 2017. https://doi.org/10.1515/jos-2017-0023

60. Xiao-Li Meng. Statistical paradises and paradoxes in Big data (I): Law of large populations, Big data paradox, and the 2016 US presidential election. *The Annals of Applied Statistics*, 12(2):685–726, 2018. https://doi.org/10.1214/18-AOAS1161SF

61. Eric I Benchimol, Liam Smeeth, Astrid Guttmann, Katie Harron, et al. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statement. *PLOS Med*, 12(10):e1001885, 2015. https://doi.org/10.1371/journal.pmed.1001885

62. Ruth Gilbert, Rosemary Lafferty, Gareth Hagger-Johnson, Katie Harron, Li-Chun Zhang, Peter Smith, Chris Dibben, and Harvey Goldstein. GUILD: Guidance for information about linking data sets. *Journal of Public Health*, 40(1):191–198, 2017. https://doi.org/10.1093/pubmed/fdx037

63. Mark D Wilkinson, Michel Dumontier, IJsbrand J Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 3(1):1–9, 2016. https://doi.org/10.1038/sdata.2016.18