22

A Matter of Trust: Data Quality and Information Integrity

Sarah B. Macfarlane and Carla AbouZahr

1 Introduction

At the time of the 2016 Australian census, the Statistics Society of Australia criticized the Australian Bureau of Statistics for making changes to the census protocol. The Bureau had decided to maintain respondents' identifiers for four years (up from 18 months) so that they could link data with previous censuses and other data sources. The Bureau argued that the increased time period would allow them to build more comprehensive datasets and produce key indicators for government decision-making. The Statistics Society was concerned that the Bureau had not properly consulted the public about the change and warned that people might withhold their names and compromise subsequent data analyses. The president of the Statistics Society warned that the controversies 'may impact upon the quality of the data collected and may be raising unnecessary fears in the community' [1]. Following preliminary data analyses, an independent panel concluded that privacy concerns had impacted response with 1 per cent of respondents giving no name or a fake name and 3 per cent giving their age instead of their date of birth [2].

S. B. Macfarlane (✉)
Department of Epidemiology and Biostatistics, School of Medicine, and Institute for Global Health Sciences, University of California San Francisco, San Francisco, CA, USA
e-mail: sarah.macfarlane@ucsf.edu

C. AbouZahr
CAZ Consulting Sarl, Bloomberg Data for Health Initiative, Geneva, Switzerland

S. B. Macfarlane, C. AbouZahr (eds.), *The Palgrave Handbook of Global Health Data Methods for Policy and Practice*, https://doi.org/10.1057/978-1-137-54984-6_22

Census data provide fundamental information about the demographic structure of society, but their integrity depends on the quality of the data their enumerators collect and process. Like other forms of data collection, censuses rely on the willingness of respondents to cooperate which, in turn, depends on their trust in the value of the data and how their data will be handled. Similarly, users of data, and information based on the data—whatever its source—need to distinguish data that are reliable and trustworthy from those that are inaccurate or misleading.

Epidemiologists describe comprehensive approaches for assessing survey quality by examining sampling and non-sampling error (see Chap. 8). The concept of *total survey error* provides a framework for describing errors that can occur during the design and conduct of a survey and how they can affect population estimates based on the data [3]. Holders of official data and statistics use similar approaches to assess data quality and information across several types of data source, including surveys, but they add dimensions that address their *fitness for use*. The Canadian Institute for Health Information (CIHI), for example, which hosts most Canadian public health data, developed a comprehensive *data quality framework* to guide data producers throughout data collection, management and documentation from a range of data sources [4]. In 2016, CIHI published a revised and renamed *information quality framework* which 'provides an overarching structure for all of CIHI's quality management practices related to capturing and processing data and transforming it into information products' [5].

We examine a common approach to assessing data and information quality across three critical data sources for health, that is, censuses, registries (with civil registration as our example) and population surveys, although the approach applies to most data sources. We provide guidance about preventing, detecting, addressing and documenting errors that can impact the quality of data and information, and we explore ways the products of data collection can be shared, combined, linked and triangulated to multiply information.

2 Producing and Assessing Quality Data and Information

The purpose of data collection—whether for a one-off survey or to maintain an ongoing register—is to describe *target* characteristics of a *target* population consisting of population units. A census aims to count an entire country's population and describe its socio-demographic characteristics at a point in time. Civil registration intends to count all births and deaths in a defined geographic area as they occur and describe causes of death. A household survey might, for example, aim to describe the demographic structure and the prevalence of health conditions among adults in a country at a point in time.

Having established clear objectives, the programme team focusses on planning, collecting and processing data, and disseminating information based on the data. This process is never perfect. The team monitors and documents the process so that it can understand the extent to which the resulting observations actually represent the intended target characteristics of the intended target population.

The *dataset* is central to transforming data into information. During planning, the programme team aims to ensure that the dataset is adequate to provide the intended information. During data collection, the team aims to obtain data for the dataset as planned. During data management, the team aims to clean and organize the dataset so that it can be analysed. While analysing the dataset, the team aims to provide the required information. The team must defend the trustworthiness of the dataset when it disseminates information based on the data or when the team shares the dataset for secondary analysis or linkage with other datasets. Much hinges on the quality of the dataset.

In its simplest form, a dataset consists of a matrix of rows and columns. Each row (or record) represents an observed population unit, for example a newborn infant, a death or a household. Each column represents characteristic, for example birthweight, cause of death or type of dwelling. Each cell represents the value of an observed (or edited) characteristic for an observed unit (Table 22.1). Users of the dataset need to know: (1) whether the records in the rows represent the units in the target population, that is whether there are any missing, duplicated or redundant units that could bias the results; (2) whether the programme team has specified the measurements in each column correctly, that is, whether the team has used the correct instruments to measure or specify the characteristics it intends to describe; and (3) whether the cell values are correct, that is whether the enumerators have measured and edited them correctly and if there are any missing values that could bias results. Biemer developed the matrix like the one in Table 22.1 as a *total error framework* to summarize major errors that can occur for any data source, not just for surveys [6]. The extent of errors in the dataset depends on how well the programme team plans and implements data collection and processes the data, as we describe in Sects. 3 and 4.

Table 22.1 Representation of total error in a dataset

Records of units in the dataset	Measurements of characteristics					Rows: Do the units represent the target population?
	Column errors: Have the measurements been specified correctly?					
	1	2	....			Row errors: are there any missing, duplicated or redundant records?
1						
2						
....						
Columns and cells: Do the observed values represent the target characteristics?	Cells errors: Are there any errors in cell values or any missing values?					Overall: How well does the dataset represent the target characteristics of the target population?

Adapted from Biemer [6]

3 Planning to Produce the Highest Quality Data and Information

Protocols or standard operating procedures (SOP) address why producers intend to collect data and the strategy they will employ to do so. They may describe, for example, how a civil registration and vital statistics (CRVS) system will record data on vital events or how a survey will meet its objectives. All protocols/SOP detail the context of the work, who needs the data/information and why, and how the data will be collected, managed and processed to achieve the project or programme objectives. Protocols/SOP also address ethical considerations and data security, include a budget and timeline, and describe roles and responsibilities, and how the programme team will manage and supervise data collection and processing. We focus on how the protocol/SOP attempts to ensure the quality of the dataset.

In Sect. 3.1, we examine the path by which data reach the rows of the dataset to understand the extent to which the records represent the target population. In Sect. 3.2, we examine the path by which the data reach the columns and cells of the dataset to understand the extent to which the observations represent the target characteristics. In Section 4, we examine how data processing can address some of these errors and can introduces cell and row errors. We follow Groves' survey lifecycle approach to quality [7] adapted by Zhang to include registries (Fig. 22.1) [8].

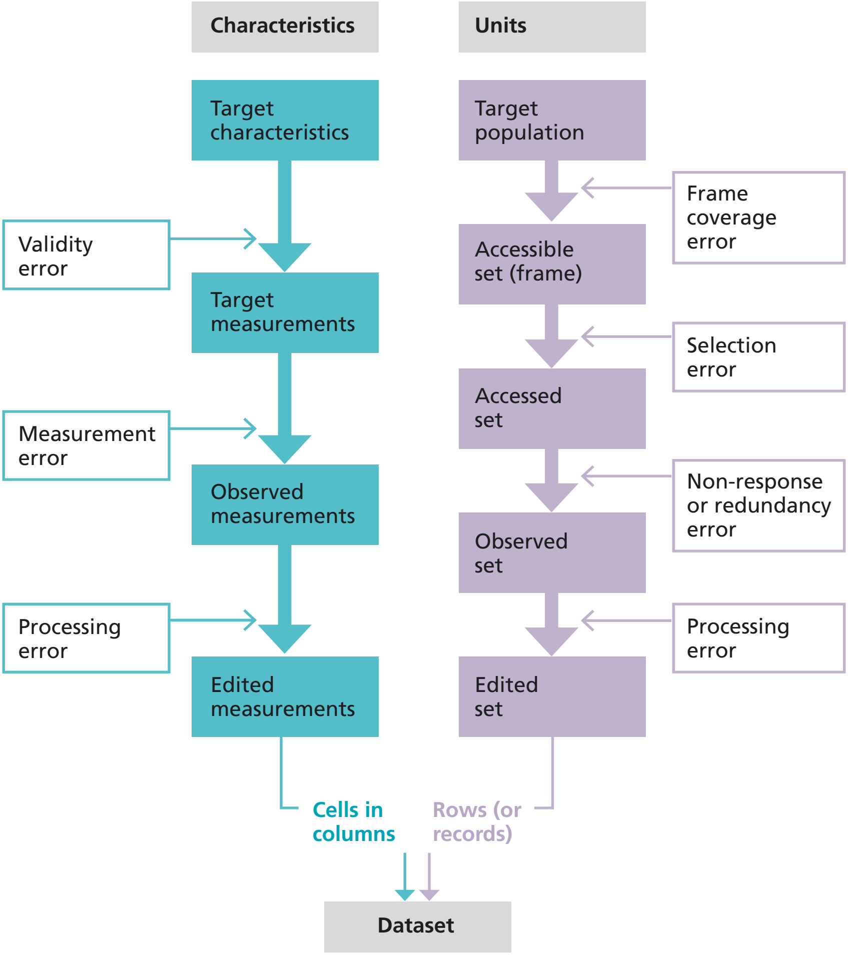

Fig. 22.1 Life cycle of data from planning to dataset from a quality perspective. Adapted from Groves [7] and Zhang [8]

3.1 Observing Units That Represent the Target Population

The protocol/SOP describes procedures either to recruit all units in the target population (for censuses and civil registration) or to recruit a sample of units that represents the target population (for surveys). There are three steps in this process. First the planning team attempts to establish a list or frame of all units in the target population, for example a list of households in a village, of census enumeration areas, or of hospitals reporting births and deaths to the

CRVS system. Secondly, the team accesses units from the frame, for example by selecting a sample for a survey, or using the complete list for censuses, or registering all births and deaths reported by the hospitals. Finally, enumerators attempt to observe the accessed units by interviewing and recording data about them. Errors occurring during each of these steps affect how completely the rows in the resulting dataset (observed units) represent the target population.

Establishing a Frame of Units in the Target Population: Preventing Coverage Error

Investigators attempt to draw up a list of all units in the target population or to describe a process to identify them. We use the term *frame* to describe this list or process. The frame is seldom complete. The frame may omit some units belonging to the target population (*under-coverage*) or contain redundant units that do not belong to the population or are duplicated (*over-coverage*). Thus the frame contains a list of accessible population units which are not necessarily the same as the units in the target population. The difference between the units in the list and the units in the target population results from coverage errors.

Under-coverage in a survey is when an investigator prepares a frame of villages in an area but omits some that have not yet been mapped, thus excluding those villages from being sampled. Over-coverage occurs when the list duplicates some villages giving them a higher chance of being selected. Under-coverage occurs in a census frame if organizers omit dwellings or enumeration areas from the frame. Under-coverage of births and deaths is a major problem for CRVS in low- and middle-income countries where registration systems are not available to the whole population, for example, ethnic minorities may be excluded from registration (see Chap. 7).

In developing the protocol/SOP, the planning team needs to assess the quality of alternative frames regarding the extent of coverage errors they might introduce and choose the frame that minimizes anticipated coverage error. Since the census frame, which attempts to locate all dwellings in a geographic area, also serves as a frame for CRVS and sample surveys, it is essential that the census team maintains it between censuses. Once the team has selected a frame, it should document the possibility of different types of coverage errors and assess the likely impact of these errors on the team's ability to describe the target population. If the team documents potential coverage errors, data analysts can take them into account during processing.

Accessing Units from the Frame: Preventing Selection Error

The programme team selects units from the frame for surveys and attempts to include all units for censuses and CRVS but either way this process introduces selection errors. For surveys that use random sampling, investigators select a sample from a *sampling* frame. This introduces *sampling error* which analysts use during data analysis to estimate indicators based on the data. To be able to describe the sampling error, the protocol must specify a procedure that ensures that every unit in the population has a known probability of being selected. Investigators' choice of sample size will affect the sampling error; the larger the sample size the smaller the sampling error. Chapter 8 describes the basics of sampling for surveys, or the reader can refer to an epidemiology textbook [7].

Selection errors can also occur when accessing units for censuses and registries. Since these errors are not usually planned as they are for random sampling, they can bias the dataset. In a census, for example, enumerators may exclude a street or village mistakenly or intentionally, although the SOP usually elaborate procedures to prevent this from happening. Alternatively, census forms may get lost in the mail and not reach the intended households. Similarly, in CRVS, certain hospitals may not systematically notify the civil registration system of the occurrence of a vital event. Such errors are difficult to take into account in the analysis since they are not random. This is also true for surveys that do not use random sampling, for example, when investigators select units *conveniently, consecutively or* interview *volunteers.* Investigators drawing up a protocol/SOP should seek to minimize unintended selection errors or, if intended, describe how they could impact their conclusions.

Observing Accessed Units: Preventing Non-response and Redundancy Error

Well-trained enumerators intend to observe all selected units but this is not always possible. *Unit non-response* occurs when selected units do not participate in the survey or census as planned, for example, survey enumerators are unable to interview heads of households, no one is present in the household, or the potential respondent declines to participate. Non-response is an increasing problem for household surveys (see Chap. 8). To reduce non-response, investigators should plan to visit households when participants are likely to be there and explain the purpose and benefits of the study very carefully.

Although in most settings householders are legally obliged to participate in censuses, they may omit people who should be included in the household on that day. In countries where civil registration doesn't function well, many people do not present themselves at the civil registry office to register births and deaths even though the system exists, in which case registration is described as incomplete. Chapter 7 describes ways in which civil registration systems can enhance registration, for example, through legislation, improved service provision, and by promoting the value of birth and death registration. Over-response is also possible, but less common, for all these data sources. In a census, the same person may be listed in two dwellings, for example, a child whose parents live in different dwellings and both include the child on their respective forms. Births and deaths can be counted twice when hospitals and the relatives report the same birth/death. Relatives may report twice if they are unsure the event has been registered, forget they have already registered the event, or if they register at both place of occurrence and place of usual residence.

3.2 Collecting Valid, Accurate and Complete Measurements: Preventing Validity and Measurement Errors

The protocol/SOP describes the way in which an enumerator, registrar or respondent will make and record each observation or measurement aiming to ensure that all are valid, accurate and complete. The best way to avoid these errors is to adopt or modify standardized questionnaires and measurement tools that have been tried and tested. This will also help to ensure that measurements are comparable to those in other datasets collecting similar data.

The protocol/SOP itself controls the *validity* of each measurement, that is, the extent to which the technique the enumerator will use to measure a characteristic actually measures what it is intended to measure. For example, if the team wishes to measure height and weight but instructs enumerators to measure participants with their shoes on, they will not measure the actual height or weight of each individual. Similarly, if the programme team wants to measure age but does not specify that enumerators should collect information on exact date of birth, the resulting data will be a poor approximation for exact age.

Measurement errors occur when enumerators do not take measurements as instructed or a when a respondent does not answer questions correctly. This can result in incorrect or missing values (*item non-response*). Programmes can reduce measurement errors by only observing characteristics that are essential to their objectives thus keeping the questionnaire/interview short. Protocols/SOP should detail thorough training and guidance for enumerators, and plan to pilot questionnaires and procedures to test instruments, and resolve ambiguities in questions and definitions. Enumerators can reduce missing values by carefully explaining the purpose of data collection to potential respondents.

If enumerators or registrars record information within a short period of the occurrence of an event or activity this reduces problems of data recall errors and can increase the accuracy of the data collected. Household surveys, for example, ask about recent health-related events or health-related behaviour. The time frame will depend on the nature of the event/behaviour but may extend from the previous month (tobacco use) to the previous three years (use of antenatal care). CRVS systems have a legal basis that makes registration compulsory for all people living in a defined area and aim to register births and deaths as they occur, usually within a maximum of 30 days, to maximize accuracy and minimize missing information items (see Chap. 7). If relatives report late, for example when a birth certificate is needed for entrance to secondary school, they may not give the correct date of occurrence.

Inaccuracies can also stem from failure to apply uniform standards when recording information. Even though, in principle, trained physicians attending hospital deaths determine cause of death using international standards, cause-of-death data are often problematic due to excessive use of so called *garbage codes,* that is, ill-defined or vague and unspecific causes of death (see Chap. 7). WHO estimates that the percentage of garbage codes is below 10 per cent where CRVS functions well and deaths are routinely medically certified by trained physicians, but over 30 per cent in settings where CRVS functions less well and physicians are not so well trained in how to complete the medical certificate of death, such as in Azerbaijan, Bahrain, Egypt, Georgia, Oman, Saudi Arabia, Sri Lanka and Thailand [9].

4 Checking, Cleaning and Processing the Dataset and Preventing Processing Errors

Whether data managers continually update data for civil registration or they collect data for a one-off study, they must check for deviations from the protocol/SOP and search for errors that occur during data collection. When enumerators record data manually on questionnaires, their supervisors may check for specific types of error, and data managers look for further errors as and after they transcribe the data into a database. When enumerators collect data on electronic devices such as smartphones or tablets for direct transmission to the database, integrated software can recognize and query errors at the moment of data capture.

After data capture, data managers use standardized procedures to clean the dataset, that is, to detect and correct or remove incomplete or inaccurate records. They aim to ensure the dataset is consistent with other similar datasets and conforms with data quality standards. They screen the dataset looking for oddities such as data gaps or duplications, outliers, inconsistencies, and unexpected patterns and results. Data cleaning may be limited to removing typographical errors or involve harmonizing, standardizing or imputing values for erroneous observations. Most database and statistical software—for example, Epi Info™, SPSS and STATA—incorporate (sometimes dual) data capture and data cleaning tools. The data processing itself introduces errors which should be monitored and flagged when checking and cleaning the data.

4.1 Checking and Cleaning Data Records

The magnitude and distribution of response errors across population subgroups indicate the overall quality of the dataset. Examination of duplicated records against eligibility criteria will identify units that were wrongly included in the dataset. Missing records can be identified from the complete list of selected units that the team intended to observe. Redundant records (carefully checked) can be removed from the dataset, but missing records can introduce bias. In routine, ongoing data collection systems, duplicate records may occur due to the use of different spellings of names, unclear addresses, and absence of clear and unambiguous identifying characteristics. The use of a unique identification number in every registration record and associated certificate can help avoid this.

4.2 Checking and Cleaning Data Items

Ideally, the protocol/SOP, associated training and supervision of data enumerators, and automatic data capture procedures will keep data errors to a minimum. In practice, errors always occur and are often not apparent until the dataset is examined.

Typical measurement errors include: (1) values that are wrong, impossible, or missing, for example, incorrect dates or implausible coding; (2) values that fall outside the measurement range, for example, a haemoglobin count of 2 grams per decilitre or an adult height of 0.5 metres; (3) values that are inconsistent with other data items, for example, a child whose weight is impossibly low for its height, or a child of five years attending secondary school; and (4) measurements that are inconsistent between enumerators or coders over time, that is they consistently take the measurements or ask the questions differently from each other.

Age heaping commonly occurs in situations where respondents don't have birth certificates or when enumerators accept rounded ages instead of obtaining dates of birth. Heaping can occur for any measurement, for example, if weight recorded in kilograms to only one decimal point shows last digit preferences for zero and five, this would indicate that enumerators have measured weights poorly. Heaping is easily identified by looking at the distribution of the last digit (which should be evenly distributed). Heaping is best assessed during a pilot and rectified by giving enumerators further training or by using a different method of measurement. Demographers have developed indices for measuring age heaping and then accounting for them in their analyses (see Chap. 17).

Data managers must follow rules in dealing with missing values; and if they have corrected or imputed any values, they need to document what they have done. They have to decide what to do with problematic data. The options are to delete the data points, to correct them or to leave them unchanged. When a data point is biologically impossible—for example, a maternal death in a male—it should either be corrected or deleted. It is sometimes possible to recalculate data that have been poorly coded, for example, redistributing cause of death data that have been assigned to garbage codes [10]. Van den Broeck et al. provide advice about data cleaning, presenting it as a 'three-stage process, involving repeated cycles of screening, diagnosing, and editing of suspected data abnormalities' [11].

Box 22.1 illustrates how Demographic and Health Surveys (DHS) check for the quality of captured records and data items.

Box 22.1 Demographic and Health Surveys: Data Editing and Quality Assurance [12]

- Questionnaires are checked when they first arrive from the field, for the correct numbers of questionnaires and selection of eligible respondents. Responses that are open-ended (such as 'other' responses) or those that require coding (such as occupation) are also coded at this point.

- All questionnaires are checked after data entry to ensure that all that were expected were in fact entered. The numbers of questionnaires are also checked against the sample design.
- All questionnaires are entered twice and verified by comparing both data sets. All discrepancies are resolved.
- The entered data are checked for inconsistencies and where possible, they are resolved. Some missing data, such as dates of events, are imputed where possible.
- A set of quality control tables is generated on a regular basis. These tables indicate potential problems in the field. The tables include information on response rates, age displacement, and completeness of data. This information is then relayed to the field teams to help them improve the quality of data in the field.

4.3 Assessing the Overall Dataset and Making Adjustments During Analysis

Once the dataset has been cleaned, the data are available to produce basic tabulations and indicators. This is another opportunity to check information quality, for example, by checking the consistency of indicators with similar indicators based on datasets from previous years or based on other datasets. Analysis at this stage can also assess the possible impact on the findings of errors anticipated from the frame or discovered during data checking and to make adjustments to estimates, if that is possible. This is the time to return to the original question: how well does the dataset represent the target characteristics of the target population?

To check for bias caused by non-responders, analysts can compare any of their known characteristics with those of responders. This is difficult since non-responders by definition don't answer questions, but it may be possible to compare publically known demographic characteristics of the person, type of household or geographic area (obtained perhaps through the frame). The non-responder or a relative may have given a reason for their absence or refusal which can be helpful in understanding non-response. Documentation of the dataset must include a full description of response rates by important subgroups, such as those living in remote areas, persons without a fixed address, minorities, ethnic groups and so on. By definition, and by law, both the census and CRVS systems should cover all persons residing in the country or territory, irrespective of nationality. Special studies may be required to assess the extent to which such groups are excluded from the census or CRVS systems, whether for *de jure* or *de facto* reasons.

Item non-response can introduce bias for missing measurements. It is easier to assess the bias their absence might have introduced by comparing key information from their records with information from the records with a recorded value, for example, whether age was missing more often for units in rural than in urban areas.

When data collection includes all population units—as intended for censuses and CRVS systems—and depending on the assessed accuracy of the measurements, the indicators calculated from the data will be the *true* population values (as measured by the SOP) at that time or period. If there are gross measurement and coverage errors, the indicators could represent another population.

For sample surveys, the calculated indicators are estimates of the *true* population values. If the survey team has used probabilistic sampling, it can measure the uncertainty around the estimate usually expressed as a confidence interval. Again, gross measurement and coverage errors could affect the population that the survey describes. At this point, it is important to assess the uncertainty around the estimated indicators and assess if they are of an acceptable width to allow conclusions to be drawn. If investigators have over-sampled certain sub-groups, then they must weight estimates during processing to reflect the true proportions of each sub-group in the population.

Most census offices undertake a *post-enumeration survey* (PES) to assess the census population count. They conduct a sample census in a random sample of areas and observe the differences between this count and the census count. They then adjust the reported census count. For example, Statistics South Africa conducted a PES after its 2011 census. The uncorrected census population count was 42.51 million people, but the PES indicated that this figure omitted 6.29 million people, so the final count became 49.79 million people (indicating a net undercount of 14.6 per cent). They also used the PES findings to assess the content quality of key characteristics such as age and sex [13].

5 Additional Criteria for Assessing Data and Information Quality

All data quality frameworks focus on the extent to which the dataset reflects what it is intended to measure. These are issues that concern data producers, but users are also concerned about other quality dimensions of the dataset. The CIHI, for example, orients its information quality framework around *fitness for use*; seeking 'to ensure a level of quality relative to the intended use of the information.' CIHI considers data and information fit for use if they satisfy the needs of users ranging from health system planners through healthcare providers and researchers. The CIHI information quality framework [5] assesses and rates the quality of information using dimensions (Box 22.2) based on the United Nations Statistical Commission's [14].

Box 22.2 Dimensions by Which the Canadian Institute for Health Information Assesses Information Quality [5]

Relevance: Does the information meet users' current and potential needs?

Accuracy and reliability: Does the information correctly and consistently describe what it was designed to measure?

Comparability and coherence: Is the information consistent over time and across providers, and can it be easily combined with other sources?

Timeliness and punctuality: Is the information current and released on schedule?

Accessibility and clarity: Is the information and its supporting documentation easily accessed and clearly presented in a way that can be understood?

Mahapatra et al. provide an assessment framework for vital statistics from civil registration systems, demonstrating how the dimensions in Box 22.2 can be used to assess vital statistics and cause-of-death statistics [15]. Statistics South Africa illustrates their use in its report of 2015 death notifications [16].

An additional essential trust dimension is security, that is, protection of data or information from unauthorized access or editing. All institutions handling personal data must ensure that they protect and de-identify, where necessary, all personal data and that they publish and monitor their procedures for maintaining data confidentiality. The CIHI complements its information quality framework with a *Privacy and Security Risk Management Framework* to 'ensure CIHI protects the privacy of Canadians and maintains the confidentiality, security and integrity of their personal health information throughout the life cycle' [17].

Guidelines are available for maintaining key data sources and for assessing data quality and information integrity, including for census [18]; CRVS [19, 20]; household surveys [7]; routine health information systems [21–23]; surveillance of communicable [24, 25] and non-communicable diseases [26]; and for research studies [27, 28].

6 Documenting the Products of Data Collection

The major products of data collection are the dataset and any reports based on analysis of the data. Thorough documentation allows others to understand and assess quality and further analyse the data using more sophisticated techniques.

6.1 Documenting the Dataset

Whether the data are for the sole use of an investigating team or to be made publically available for others to analyse, datasets must be well-documented with data organized and stored in an accessible format—with clear description, or metadata (see Chap. 23). Standardized and consistent metadata standards are essential for data sharing. The Organization of Economic Co-operation and Development (OECD) Health Statistics publication, for example, links to a comprehensive metadata dictionary that covers data definitions, sources and methods for all the indicators [29]. The United Nations maintains a metadata dictionary for Sustainable Development Indicators [30].

The CIHI suggests that data managers document a metadata repository under the following headings: (1) description of the dataset with detailed background information about the context in which the data were collected; (2) criteria for selecting the units of observation; (3) methods of data collection and capture; (4) data processing procedures including description of data editing; (5) any data analysis and dissemination already undertaken; (6) details of data storage; and (7) all relevant documentation dealing with data quality [4].

Although post-collection data errors occur in all datasets, data managers rarely describe their data cleaning processes, especially for routine data collection activities [31]. To enhance the users' trust, data managers should specify how they have cleaned the data to address problems such as miscoding and follow rules in dealing with missing values; and if they have imputed any values, they need to document what they have done. This helps reassure data users of the integrity of the data and absence of manipulation.

Data producers may not publish data reports because they don't want to reveal poor data quality. We consider this to be a mistake. Data producers are more likely to gain the trust of data users if they are transparent about data limitations. And nothing is more conducive to improving data quality that making information available and throwing the light of day on the dark corners of a dataset. For example, the 2015 Statistics South Africa report of mortality and causes of death is explicit about data quality limitations [16].

6.2 Documenting Information Based on the Dataset

Most reports based on these data sources present information quite simply as trends in indicators disaggregated by population sub-groups, time and space. Census reports publish the actual breakdown of the counted population by age, sex, enumeration area and provide further tables depending on the census questionnaire (see Chap. 6). Civil registration reports provide estimates of birth rates and death rates by age, sex, and cause, broken down by socio-demographic and geographic areas (see Chap. 7). National household surveys publish detailed cross-tabulations and estimates for specific population sub-groups. Reports should include tables showing non-response and missing values and summarize and assess the likely impact of any errors introduced during data collection and cleaning.

Tables provide the most detailed information, but diagrams can illustrate distributions of indicators between key groups, across time and by geographic area. Most people find simple visual presentations such as charts and maps easier to understand than large tables or long lists of numbers. Other chapters in this handbook illustrate line graphs (Chaps. 6, 7, 9,17), population pyramids (horizontal histograms) (Chap. 6), bar charts (Chap. 11), maps (Chaps. 12, 15, 20) and results of predictive modelling (Chaps. 19, 20, 21). However, there are many ways that visuals provide misleading information, whether deliberately or, as is more often the case, unintentionally [32].

The basic principles for interpreting both tables and diagrams are to ascertain: (1) the number of units on which the table/diagram is based; (2) whether there are any missing values and how they are distributed among sub-groups; (3) how percentages were calculated (using the total number of units, or numbers in sub-groups in their denominators); (4) the range or standard deviation of indicators expressed as averages (for example, average blood pressure); and (5) for diagrams, check the scales of each axis and whether there is any break in the axis that could misrepresent findings. Reference materials are available to guide the presentation of demographic and epidemiological information [33].

Sophisticated software makes it easy to produce charts, maps and innovative visualizations but a balance must be struck between design and function; complicated visualizations can fail to communicate [34]. Infographics are increasingly used to convey information that tells a story using easy to understand visuals and minimal text. Major challenges are how to present probabilities and uncertainty, particularly when users have different levels of statistical literacy. Spiegelhalter et al. offer some sound advice on ways of visualizing uncertainty that are relevant to charts and figures in general [35]. Chapter 20 of this handbook shows how predictive maps incorporate uncertainty.

Uncertainty accompanies all data collection activities but cannot always be measured. However, sample surveys are designed to describe uncertainty. For example, the Nepal DHS 2016 estimated 95 per cent confidence intervals for neonatal mortality to be from 16.5 to 26.4 deaths per 1,000 live births in urban areas—that is, there is a high probability that this range contains the true neonatal mortality rate for the target population in urban areas. The study also estimated 95 per cent confidence intervals for rural areas to be from 26.0 to 39.7 deaths per 1,000 live births. There is very little overlap between the two ranges, suggesting that neonatal mortality is higher in rural areas [36].

All reports or papers should provide the context for the data collection. A conflict of interest statement can enhance users' trust in data, whether derived from routine data collection or from special studies and surveys. These statements can be particularly sensitive in areas where the interests of public health and of private sector businesses intersect. For example, in 2017, the World Health Assembly called for a consultation to bring together representatives from health, industry, NGOs, governments and civil society to examine ways of 'addressing and managing conflicts of interest in the planning and delivery of nutrition programmes at country level.' [37]

7 Sharing and Combining the Products of Data Collection

We have described processes for checking and documenting the quality of data products before their release. Not all users have the technical knowledge to critically review these products, but they can work with their technicians to satisfy their conclusions about the quality dimensions described in Box 22.2. Users can then combine and triangulate these data and information with those from other products, of which there are many!

Opening and Linking Datasets The Open Government movement encourages governments to make data publically available either as datasets or as indicators (see Chap. 23). Websites collate data from specific types of data across the world, for example, census data are available in one place [38] as they are for DHSs [39]**.** Researchers undertake more sophisticated analyses of individual or combinations of published datasets and so multiply available information. Fabic et al. reviewed 1,117 peer-reviewed papers based on DHSs published between 1985 and 2010 and noted a progressive increase in the number published each year with a widening range of topics. They also noted an increase in publications that analysed multiple surveys either across time or across countries (representing 34 per cent of all the publications) [40].

Triangulating Data and Information Ministries need to compare indicators across different sources both to validate specific indicators but also to gain more insight into a specific issue. Rutherford et al. describe a triangulation approach which they have used to understand the dynamics of HIV transmission and to measure the impact of public health programmes. They define *public health triangulation* as 'the process of reviewing and interpreting existing data and trends in those data from multiple data sources that bear on different facets of a broad public health question in order to identify factors that underlie the observed data and to assist with public health decision-making and actions.' [41] Rutherford et al. describe steps from framing questions, through identifying, gathering, and reviewing data and interpreting and using the results to inform public health action. Qualitative and quantitative data include primary and secondary information from censuses, surveillance, public health programmes and results of local research studies. They emphasize the need to check for data quality in the way we have described, and they warn about the dangers of analysing trend data without an underlying model (*ecological fallacy*), dredging data without a hypothesis and ensuring the reproducibility of results.

Estimating SDG Indicators Since 1980, WHO and the United Nations Children's Fund have used triangulation to annually review data to estimate national immunization coverage. They do this by reviewing government reports and survey findings from published and unpublished literature, and in consultation with local experts and programme managers. They publish the estimates after feedback from national authorities [42]. In Chap. 21, Mathers et al. describe more complex statistical methods of bringing together disparate data to come up with indicator estimates, and in Fig. 21.1, they demonstrate the many sources used to estimate child mortality estimates in Nigeria from 1964 to 2017. Murray uses the term *systematic review* to describe the process of reviewing and using all available quantitative data and attempting to reconcile differences between data sources [43].

Combining Products with Big Data With burgeoning amounts of data now in the public domain, the discipline of data science has emerged. Data science is the process of finding, developing and communicating actionable information that stems from multiple sources, including from social media. For instance, data science might bring together information from household surveys, routine health information systems, censuses and non-traditional sources like Facebook activity, Google searches, tweets and mobile phone data with the purpose of ascertaining people's health-seeking activities (see Chap. 16) modelling disease outbreaks (see Chap. 10) or predicting the effects of health interventions. Data scientists use *big data* which is defined by its *variety*, *velocity* and *volume*. Compared to the data we have described which are collected by *design*, big data are *found*, and so the same rules of analysis and interpretation don't necessarily apply [44].

We have suggested rigorous ways of checking designed data but what about the quality of found big data? Biemer [6] and subsequently Japec et al. [44] explore evolving Table 22.1 to become a Big Data Total Error Framework. Although big data do not arrive in matrix form, data scientists reduce them to such for analysis. To illustrate some data quality issues, we expand on one of Japec et al.'s examples. Suppose a data scientist decides to describe opinions of a population in a geographic area by harvesting Google searches made from *URLs* based in that location during the past week. Row errors would result from people being excluded because they didn't have Internet access (undercoverage), people conducting many searches (redundance) or searches being conducted by a robot not a person (ineligible). Column errors of misspecification could include inappropriate classification of phrases used in the Google search. Cell errors are similar to those in any other form of data collection, for example, misclassification or miscoding, and wrong content when people meant something different to what was recorded, and missing values such as people not including certain terms in their search.

8 Conclusion

We have described criteria for maintaining and assessing the quality of data and information but even adherence to the highest quality standards does not necessarily engender the user's trust.

The OECD has developed survey modules for countries to assess the public's trust in official statistics [45]. They suggest that trust in official statistics depends on trust in both the statistical products and the institution producing them. The OECD uses criteria similar to those in Box 22.2 to assess data products but they suggest that trust in the institution producing them depends on: the extent to which they are or are perceived to protect data confidentiality and operate impartially without political interference; produce statistics openly and transparently; and maintain an honest relationship with the public and other key stakeholders. The latter includes the institution disseminating information about how and why it collects data, holding regular consultations, listening to criticism, correcting erroneous data and publically addressing misleading media reports.

External factors also influence the public's trust in data and these include the political environment. As we write, the United States is planning to introduce a question about citizenship into its 2020 census and to report these data by census block—which could be an apartment building. People are threatening to *#LeaveItBlank.* Despite all the checks we have described, and the privacy and security commitments that may be in place, the quality of data and information depend on the population's trust in the institution collecting the data, and in the government that finances and plans to use the results.

Key Messages

- All datasets are prone to errors that arise during data collection, design, implementation, compilation and analysis.
- Methods are available to prevent and manage errors so that users can be confident in the integrity of the information.
- Data producers should provide detailed metadata for each dataset and document the methods they used to maximize data quality.
- Data producers and analysts are responsible for building and maintaining trust in statistics as a global public good.

References

1. Karp P. Census controversy shows ABS 'needs to do better', says Statistical Society. 2016 [cited 2018 8th April]. Available from: https://www.theguardian.com/australia-news/2016/aug/09/census-controversy-shows-abs-needs-to-do-better-says-statistical-society
2. SBS News. Census 2016: Fake names given amid privacy fears. 2017 [cited 2018 8th April]. Available from: https://www.sbs.com.au/news/census-2016-fake-names-given-amid-privacy-fears
3. Groves RM, Lyberg L. Total survey error past, present, and future. Public Opinion Quarterly. 2010 Jan 1;74(5):849–79. Available from: http://dx.doi.org/10.1093/poq/nfq065
4. Canadian Institute for Health Information. The CIHI Data Quality Framework. Ottawa, Canada: CIHI. 2009 [cited 2018 5th November]. Available from: https://www.cihi.ca/en/data_quality_framework_2009_en.pdf
5. Canadian Institute for Health Information. CIHI's Information Quality Framework. Ottawa, Canada: CIHI. 2017 [cited 2018 5th November]. Available from: https://www.cihi.ca/sites/default/files/document/iqf-summary-july-26-2017-en-web_0.pdf

6. Biemer P. Total Error: Adapting the paradigm for big data. 2014 [cited 2018 8th April]. Available from: https://www.niss.org/sites/default/files/biemer_ITSEW2014_Presentation.pdf
7. Groves RM, Fowler Jr. FJ, Couper Mick P, Lepkowski JM, Singer E, Tourangeau R. Survey methodology. Second ed. Hoboken, USA: John Wiley & Sons, Inc.; 2009.
8. Zhang Chun. Topics of statistical theory for register-based statistics and data integration. Statistica Neerlandica. 2012;66(1):41–63. Available from: http://dx.doi.org/10.1111/stan.508
9. World Health Organization, World health statistics 2018, page 57. [cited 2018 5th November]. Available from http://www.who.int/gho/publications/world_health_statistics/2018/en/
10. World Health Organization. WHO methods and data sources for global burden of disease estimates 2000–2015. 2017 [cited 2018 8th April]. Available from: http://www.who.int/healthinfo/global_burden_disease/GlobalDALYmethods_2000_2015.pdf
11. Van den Broeck J, Argeseanu Cunningham S, Eeckels R, Herbst K. Data cleaning: Detecting, diagnosing, and editing data abnormalities. PLoS Medicine. 2005 Sep 6;2(10):e267. Available from: http://dx.doi.org/10.1371/journal.pmed.0020267
12. The DHS Program. Data processing. [cited 2018 8th April]. Available from: https://dhsprogram.com/data/data-processing.cfm
13. Statistics South Africa. Census 2011 Post-enumeration survey results and methodology. 2012 [cited 2018 8th April]. Available from: http://www.statssa.gov.za/census/census_2011/census_products/Census_2011_PES.pdf
14. United Nations Statistical Commission. Guidelines for the template for a generic National Quality Assurance Framework (NQAF). 2012 [cited 2018 8th April]. Available from: https://unstats.un.org/unsd/statcom/doc12/BG-NQAF.pdf
15. Mahapatra P, Shibuya K, Lopez AD, Coullare F, Notzon FC, Rao C, et al. Who counts? Civil registration systems and vital statistics: successes and missed opportunities. Lancet. 2007 Nov;370(9599):1653–63. Available from: http://dx.doi.org/10.1016/s0140-6736(07)61308-7
16. Statistics South Africa. P0309.3. Mortality and causes of death in South Africa, 2016: Findings from death notification. 2018 [cited 2018 8th April]. Available from: http://www.statssa.gov.za/?page_id=1854&PPN=P0309.3&SCH=7308
17. Canadian Institute for Health Information. Privacy and security risk management framework. 2015 [cited 2018 8th April]. Available from: https://www.cihi.ca/sites/default/files/psrm_framework_enweb_0.pdf
18. United Nations, Department of Economic and Social Affairs, Statistics Division. Handbook on the management of population and housing censuses. Second Revision. New York, USA: United Nations. 2016 [Cited 2018 5th November]. Available from: https://unstats.un.org/unsd/demographic-social/Standards-and-Methods/files/Principles_and_Recommendations/Population-and-Housing-Censuses/Series_M67Rev2-E.pdf
19. World Health Organization. Improving the quality and use of birth, death and cause-of-death information: guidance for a standards-based review of country

practices. 2010 [cited 2018 8th April]. Available from: http://whqlibdoc.who.int/publications/2010/9789241547970_eng.pdf

20. United Nations, Department of Economic and Social Affairs, Statistics Division. Principles and recommendations for a vital statistics system. Revision 3. 2014 [cited 2018 8th April]. Available from: https://unstats.un.org/unsd/demographic/standmeth/principles/M19Rev3en.pdf
21. World Health Organization. What is the Health Information and Intelligence Platform (HIIP). 2014 [cited 2018 9th March]. Available from: http://hiip.wpro.who.int/portal/default.aspx
22. Aqil A LT, Moussa T, Barry A. Prism tools user guide. 2012 [cited 2018 8th April]. Available from: https://www.measureevaluation.org/resources/publications/ms-12-51
23. Health Metrics Network. Framework and standards for country health information systems. Second ed. [cited 2018 8th April]. Available from: http://www.who.int/healthinfo/country_monitoring_evaluation/who-hmn-framework-standards-chi.pdf
24. World Health Organization. World health statistics 2006. Geneva, Switzerland: World Health Organization. 2006 [cited 2018 5th November]. Available from: http://www.who.int/whosis/whostat2006/en/
25. Centers for Disease Control and Prevention. Updated guidelines for evaluating public health surveillance systems. MMWR Recommendations and Reports. 2001;50(13):1–35. [cited 2018 5th November]. Available from: https://www.cdc.gov/mmwr/preview/mmwrhtml/rr5013a1.htm
26. World Health Organization. The WHO STEPwise approach to surveillance of noncommunicable diseases (STEPS). 2003 [cited 2018 8th April]. Available from: http://www.who.int/ncd_surveillance/en/steps_framework_dec03.pdf
27. World health Organization. Recommended format for a research protocol. [cited 2018 8th April]. Available from: http://www.who.int/rpc/research_ethics/format_rp/en/
28. Equator Network. Enhancing the quality and transparency of health research. [cited 2018 8th April]. Available from: http://www.equator-network.org/
29. Organization of Economic Co-operation and Development. OECD health statistics 2017. Definitions, sources and methods. 2017 [cited 2018 8th April]. Available from: http://www.oecd.org/els/health-systems/Table-of-Content-Metadata-OECD-Health-Statistics-2017.pdf
30. United Nations, Department of Economic and Social Affairs, Statistics Division. SDG indicators metadata repository. [cited 2018 8th April]. Available from: https://unstats.un.org/sdgs/metadata/?Text=&Goal=17&Target=
31. Leek JT, Peng RD. Statistics: P values are just the tip of the iceberg. Nature. 2015 Apr 28;520(7549):612–612. Available from: http://dx.doi.org/10.1038/520612a
32. Hosseinpoor A, AbouZahr C. Graphical presentation of relative measures of association (correspondence). Lancet. 2010 Apr;375(9722):1254. Available from: http://dx.doi.org/10.1016/s0140-6736(10)60541-7
33. Duquia RP, Bastos JL, Bonamigo RR, González-Chica DA, Martínez-Mesa J. Presenting data in tables and charts. Anais Brasileiros de Dermatologia. 2014

Apr;89(2):280–5. Available from: http://dx.doi.org/10.1590/abd1806-4841.20143388

34. United Nations Economic Commission for Europe. Making data meaningful. Part 2. A guide to presenting statistics. 2009 [cited 2018 8th April]. Available from: http://www.unece.org/stats/documents/writing/MDM_Part2_English.pdf
35. Spiegelhalter D, Pearson M, Short I. Visualizing uncertainty about the future. Science. 2011 Sep 8;333(6048):1393–400. Available from: http://dx.doi.org/10.1126/science.1191181
36. Nepal Ministry of Health, New ERA, ICF. Nepal Demographic and Health Survey 2016. Kathmandu, Nepal: Ministry of Health, Nepal; 2017 [cited 2018 5th November]. Available from: https://www.dhsprogram.com/pubs/pdf/fr336/fr336.pdf
37. World Health Organization. Addressing and managing conflicts of interest in the planning and delivery of nutrition programmes at country level. Report of a technical consultation convened in Geneva, Switzerland, on 8–9 October 2015. [Cited 2018 5th November]. Available at http://www.who.int/nutrition/publications/COI-report/en/
38. IPUMS CPS. [cited 2018 10th April]. Available from: https://cps.ipums.org/cps/
39. DHS Program. Available datasets, [cited 2018 8th April]. Available from: https://dhsprogram.com/data/available-datasets.cfm
40. Fabic M, Yoon Joung C, Bird S. A systematic review of Demographic and Health Surveys: data availability and utilization for research. Bulletin of the World Health Organization. 2012 Aug 1;90(8):604–12. Available from: http://dx.doi.org/10.2471/blt.11.095513
41. Rutherford GW, McFarland W, Spindler H, White K, Patel SV A, Berle-Grasse J, et al. Public health triangulation: approach and application to synthesizing data to understand national and local HIV epidemics. BioMed Central Public Health. 2010 Jul 29;10(1). Available from: http://dx.doi.org/10.1186/1471-2458-10-447
42. Burton A, Monasch R, Lautenbach B, Gacic-Dobo M, Neill M, Karimov R, et al. WHO and UNICEF estimates of national infant immunization coverage: methods and processes. Bulletin of the World Health Organization. 2009 Jul 1;87(7):535–41. Available from: http://dx.doi.org/10.2471/blt.08.053819
43. Murray C. Towards good practice for health statistics: Lessons from the Millennium Development Goal health indicators. Lancet. 2007;369:862–73.
44. Japec L, Kreuter F, Berg M, Biemer P, Decker P, Lampe C, et al. Big data in survey research. Public Opinion Quarterly. 2015;79(4):839–80. Available from: http://dx.doi.org/10.1093/poq/nfv039
45. Organization of Economic Co-operation and Development. Measuring trust in official statistics: Cognitive testing. Report to the OECD of the Electronic Working Group on Measuring Trust in Official Statistics. 2011 [cited 2018 10th April]. Available from: https://www.oecd.org/sdd/50027008.pdf