4.4 Collection and management of good quality data

Authors

Fernando Gouvea-Reis, Public Health England, London, United Kingdom.

Marcelo Farah Dell'Aringa, CRIMEDIM–Research Center in Emergency and Disaster Medicine, Università del Piemonte Orientale, Novara, Italy.

Virginia Murray, Public Health England, London, United Kingdom.

4.4.1 Learning objectives

To understand key aspects of data collection for research in health emergency and disaster risk management (Health EDRM), including:

1. Different sources and methods for data collection, along with their advantages and limitations.
2. Challenges involved in collecting data in disaster settings, and how these might be overcome.
3. The importance of data quality, data storage and data sharing.

4.4.2 Introduction

The timely collection of good quality data on key aspects relevant to disaster risk management, including emergency response is critical to Health EDRM research, as research outcomes are dependent on data quality and outputs. High quality research and data are invaluable to enable:

- Planners and responders to implement Health EDRM for effective and efficient action in the areas where their work is most needed.
- Policymakers to influence evidence informed best policy and practice in Health EDRM.

Good quality research requires data that are relevant to the research question and objectives, which may include demography, morbidity, mortality, infrastructure, different health factors, environmental characteristics, and so on. Such data are needed to manage disaster risk so that future disasters can be avoided or their impact minimized. It also supports the planning, management, and evaluation of post-disaster interventions. Poor quality data will lead to poor quality research and, potentially, to misinformed policies. Therefore, it is key to ensure the collection of high-quality data during any study.

This chapter discusses important aspects that should be considered before, during and after the process of data collection in order to ensure that good quality data are used and available in disaster research. It explores the planning and preparation processes, different methods for data collection, as well as the challenges that a researcher may face when studying disasters and tools that might help them to address these challenges. Finally, it will discuss how to ensure good quality data are stored and made accessible to others so that it can bring additional benefits.

4.4.3 Preparation

Successful data collection demands careful preparation. It is important to formulate a clear and specific research question or hypothesis to be tested, and then to plan what specific data and what collection strategy will provide adequate and sufficient information to answer that question or allow the hypothesis to be accepted or rejected. Although it can be tempting to adapt the data collection or methods of an ongoing study to collect additional data to test another hypothesis, without proper reflection and planning, this can result in the presence of confounding factors in the collected data, leading to biased results. Alternatively, it can also compromise the statistical power of the results. Having a clear research question and aim at the start of the planning process can help to avoid such issues (Chapter 3.5).

It is also important to have a clear, written protocol before data collection begins, and this may be needed when seeking ethics approval (Chapter 6.4). This includes the research question, aim and objectives, definitions of exposure, outcome, and other terms, the expected sample size, the methods to be used, how participants will be recruited and how the data will be curated and analysed after collection. Furthermore, agreement on clear hazard definitions is key to allow comparability on data collected from different sources. Conducting a literature or scoping review before you write your protocol is an important tool to understand how other researchers studying disasters and disaster risk management have collected data on similar contexts, or how they answered similar questions (Chapters 2.6 and 3.6). This can help in the understanding of what strategies work best, as well as with anticipating the main challenges as encountered by others, so that the researcher is prepared to deal with these should they appear during their study.

4.4.4 Data collection methods

Once a research question and the variables of interest have been defined, the next step is to determine how these parameters will be measured. Depending on factors such as the study design, funding, time and human resources available, the researcher may decide between collecting new data or studying data that have been previously collected by others. These different approaches are also known as primary and secondary data collection methods.

4.4.5 Primary data

Primary data are data collected for the first time and for the purpose of a specific study. The researchers conducting that study decide where, how and when the data will be collected to specifically address their research question. However, this approach can be expensive and time consuming, and may demand technical resources. Methods of primary data collection can be broadly divided into two approaches:

Quantitative methods are used for numerical data. They include analysis of the data using descriptive and comparative statistical techniques (see Chapters 4.2 and 4.5) to answer specific questions about, for example, how commonly something occurs, or differences between groups. In Health EDRM, this approach can be used to estimate morbidity and mortality. It can also be used in the construction of more complex models to estimate, for example, the economic impact of a flooding in an affected area (Chapters 4.6 and 4.7). Data collection methods in quantitative research can involve surveys (Chapter 3.1), the measurement of outcomes in experiments or observational studies (Section 2 and Chapter 4.1), and the use of routinely collected data from different monitoring systems (Chapter 2.4). It usually requires large sample sizes and appropriate sampling of the participants from whom the data will be collected, in order to ensure the desired generalizability of the results.

Qualitative methods, which are discussed in detail in Chapter 4.12, are most often used to study research questions about how and why phenomena occur, and use observed and recorded non-numerical data, such as words and images, to understand meaning. The collection of such data is usually performed through in-depth interviews, focus groups, key-informant interviews, and observations. Because statistical methods are not used for qualitative research, there is no predetermined sample size. A qualitative approach is particularly useful when the objective is to understand underlying reasons, opinions and motivations during exploratory research, or to develop a theory. For example, it can be employed during a study that aims to understand the drivers of behaviour change related to the implementation of safe burial practices during an Ebola outbreak. It is also useful in the development of hypothesis to be tested in later quantitative studies.

4.4.6 Secondary data

Secondary data comprises data already collected or produced by others. Common sources of secondary data are government databases and publications, books, scientific papers, media channels and routine data. Routine data are data collected in a periodic, systematic manner by the government or other organizations (Chapter 2.4) and include:

- **Demographic data**, describing variables such as age, sex, ethnicity, migration patterns, marital status, and so on.
- **Health event data**, describing health variables that affect individuals or populations, including births, deaths, and population interaction with the health sector at different levels.
- **Circumstantial data**, describing factors associated with the social determinants of health, including data on education, employment, housing and environmental data.

- **National reference data**, which covers data that has not been issued purely for health purposes, but when integrated and combined with other variables can be useful in the understanding of different health issues.

Using secondary data means the researchers do not have full control over data quality, making it more difficult to ensure that the dataset they use is complete, unbiased, time accurate, and reliable. Table 4.4.1 highlights important key points on data quality that must be considered when using secondary data.

Table 4.4.1 Important considerations for the use of routine data *(1)*

Accuracy: to what extent is the dataset accurate? What are the potential biases?
Precision: Have appropriate measures of uncertainty been included (such as 95% confidence intervals)?
Completeness: how much of the data is missing?
Timeliness: were the data collected in a period that is relevant to the study?
Coverage: is the whole population of interest covered? If not, how does this impact the study?
Accessibility: who has access to the data, and how is this access controlled?
Confidentiality: have individual-level data been anonymized?
Original purpose of collection: can the data be used for a different purpose to the one for which it was collected? Who collected the data and how?
Analysis: have the data been standardized and presented in a comparable way?

4.4.7 Dealing with challenges in disaster data collection

Researchers can anticipate facing different challenges during data collection. Some examples are:

- limited access to certain areas due to infrastructural collapse (destruction of roads and other transportation systems, for example).
- Persistence of the hazard that originated the disaster, which might pose a risk for the research team (radiation after nuclear incidents, for example).
- emergence of infectious diseases outbreaks due to damaged or poorly functioning water and sanitation infrastructure, which can become a threat to the local community and researchers (cholera epidemics after floods, for example).
- political barriers (local authorities attempt to minimize or change disaster-related statistics, such as mortality estimates, or refuse access to the planned research site, for example).
- language barriers, when the researchers do not speak the local language, leading to the possibility of bias in the use of translators.

Case Study 4.4.1 illustrates how researchers in the field can face some of these barriers. The early consideration of the challenges that are most likely to be encountered can help choosing the most appropriate data collection strategy.

Case Study 4.4.1
Challenges in disaster data collection after the 2004 Indian Ocean Earthquake and Tsunami *(2)*

The 2004 earthquake and tsunami that occurred in the Indian Ocean affected 12 countries and left almost 230 000 people dead and approximately 1.7 million people displaced *(3)*. In the post-disaster environment, different groups conducted research aiming to understand how the event affected factors such as the health status of the local communities and their health needs. These groups faced various challenges in data collection.

For example, a study was conducted to determine the public health impact of the tsunami on the population of three communities in Aceh Jaya District, Indonesia. However, all health facilities in the three communities were destroyed during the tsunami, and the only health professionals to survive the disaster were two midwives. As a result, much of the data had to be obtained from secondary sources, such as reports from local authorities, and the results of the study were thus susceptible to recall, reporting and misclassification biases *(4)*. Another study found that poor health record keeping in facilities prior to the tsunami limited the comparative effectiveness of the health data collected after the tsunami. This led to issues in determining which health-related issues were the result of the disaster and which reflected pre-existing problems *(5)*.

In another study, the French Army medical service carried out an epidemiological survey to estimate health indicators in children during the weeks following the tsunami in Meulaboh. They reported issues with communication and translation during interviews, where sometimes it was difficult to communicate directly with the children or their parents, leading to errors of interpretation. Furthermore, the researchers also faced barriers related to the transportation of the data collection teams among the disaster settings *(6)*.

There are different approaches that can support researchers in gathering good quality data and overcoming the challenges involved in data collection for disaster research. The use of routine data, for example, is a useful tool in contexts where time and resources are constrained (Chapter 2.4). For example, using secondary, routine data can rapidly provide the necessary information to compare before and after disaster scenarios, demonstrate change in demand for specific healthcare services, and to evaluate its impact on local health systems, as demonstrated by Case Study 4.4.2.

4.4

Case Study 4.4.2
An ecologic study to evaluate the impact of the 2011 Rio de Janeiro landslides in the utilization of public mental health services *(7)*

Many areas of the south and south-eastern regions of Brazil are hit frequently by heavy rains during the summer months. These regions have some of the places with the highest population density in the country and many people living in disaster-prone areas. This leads to important vulnerabilities and thus many communities are under extensive disaster risk of landslides and floods. The 2011 landslides in the mountainous region of Rio de Janeiro State were the largest disaster by immediate death count in recent Brazilian history, with a report counting 845 immediate deaths, mostly by mud burial. Moreover, around 30 000 people were left homeless in 11 different municipalities and there was important damage to agricultural and industrial activities.

An ecologic study was performed using routine data from DATASUS (Departamento de Informática do SUS - Informatics Department of the Brazilian Public Health System in free translation). DATASUS comprises a wide range of open access data, and allows researchers to gather and analyse datasets regarding health outcomes, the incidence of diseases and on the utilization of the health services in different levels.

The study analysed data from the affected region of Rio de Janeiro state two years before and after the event and comparing it with unaffected regions of the state. The analysis of the data suggested a sustained increase in the search for mental health services by the affected population after the landslides, which was not found in the other regions of the state.

The use of routine data can also be helpful in the construction of models to leverage disaster risk reduction strategies. Case Study 4.4.3 presents an example where this approach was used to better prevent and respond to infectious diseases outbreaks.

Case Study 4.4.3
The combination of cholera outbreak data and satellite environmental information to estimate cholera risk *(8)*

Cholera is an infectious disease caused by the ingestion of contaminated water or food with the bacteria *Vibrio cholerae*. Water-related diarrheal diseases like cholera are estimated to kill approximately 1.5 million people every year. They are the second leading cause of death in children under five years old. The impact of cholera is higher in settings with poor availability of clean water, as well as places susceptible to floods and with heavy rainy seasons.

Scientists combined in an algorithm data related to the time and location of previous cholera outbreaks in sub-Saharan Africa with different satellite datasets, including precipitation, air temperature, and land surface temperature. The algorithm was tested in five cholera epidemic regions of Sub-Saharan Africa (Mozambique, Central African Republic, Cameroon, South Sudan, and Rwanda), and was able to identify and predict regions most at risk for an outbreak at least four weeks in advance *(8)*.

In Yemen, this model has been used to predict where and when the next increase in cases of cholera will happen. When risk areas are identified, local partners can work in managing disaster risk by directing emergency resources to the most critical areas, improving infrastructure where needed, chlorinating water and running educational and vaccination campaigns *(9–10)*.

To build a complete picture related to the hazard or disaster of interest, information from several data sources are likely to be needed. It is also important to note that, in different countries and contexts, the data of interest may be collected and curated by different organizations, which can include the Ministry of Health, National Statistics Offices, or even be fragmented through different levels of regional and local health departments (Chapter 2.4). This can result in extra time and resources needed to collect and standardise data provided by different sources.

However, in settings where local data collection for relevant parameters is poor or absent, the use of secondary data might be constrained. Depending on the availability of time and resources, you might choose to perform the primary data collection yourself using protocols with relevant ethical consent (Chapters 3.4 and 6.4). If this is not suitable to your context, the development of models can also be considered as an alternative strategy to fill the information gaps (Chapter 4.6). This can be an important opportunity to raise awareness among local governments, universities and independent organizations about the importance of initiating and maintaining good routine data collection and how this might help them prevent and respond to disasters.

4.4.8 Different approaches in data collection

There are a growing number of useful tools to support disaster research, and big data can be leveraged to provide important information in a variety of contexts. Big data includes data such as satellite imagery, images and videos from unmanned aerial vehicles (UAVs), sensor web and Internet of Things (IoT), airborne and terrestrial Light Detection and Ranging (LiDAR), simulation, crowdsourced information, social media, and mobile global positioning system (GPS) and Call Data Records (CDR) *(11).*

For example, the management of disaster risk can be supported through images and videos captured by satellites or UAVs to develop hazards maps and risk assessments. Similarly, the assessment of post-disaster damage through change detection, for instance, provides enhanced situational awareness, supporting and guiding action from rescue teams. It may also be possible to use crowdsourcing to gather these types of data (Chapter 5.2).

4.4.9 Data storage and data sharing

When the data has been collected and cleaned, the next step is to store it securely for current and future analysis, and to consider how it might be shared so that others can also benefit from it.

According to the type of research study, it is possible that data will be collected from multiple sources. Therefore, the design of a curation system should account for such differences and allow standardization. This can be achieved by a computerized database with clear rules for data entry. This involves facilitating the user role by requiring only the needed information to be added. For example, for discretionary variables, the adoption of drop down lists to be selected by the user instead of empty spaces for free text can help reducing entry errors and ensure standardization. Similarly, the implementation of rules such as limiting the valid range for variable fields and flagging errors if information is not adequately entered in a core field exemplify how the adoption of simple, good practices, help the achievement of a complete and accurate dataset *(12)*.

It is also important to consider that the usefulness of a dataset to others can be enhanced by providing data as disaggregated as possible, but while still safeguarding individual privacy. A simple example to understand this principle is when reporting on residents who have been affected by a local flood, a dataset which can be filtered according to sex, age, socioeconomical factors, health status and disability allows a much broader set of analysis to be made, such as developing hypothesis on the correlation of the outcome with possible risk factors. The more disaggregated a dataset can be to the individual level, the more invisible persons can be made visible. It can then be used as reliable evidence to inform policymaking, for example helping to direct resources to those affected who need it the most.

There is currently a widespread call across research for making data open and transparent, improving its usefulness so that others can also benefit from it. The ‘data revolution’ comprises the large increase in the volume and types of data that are currently collected by governments, private companies, NGOs, researchers and citizens. This is leading to an

unprecedented possibility of transforming such data into knowledge to not only manage disaster risk but also to better respond to disasters *(13)*. However, important data are often not released rapidly, or not shared at all, which compromises the potential re-usability of many datasets. The FAIR principles of data sharing were developed to assist in the production of good-quality data, with practical actions that can be adopted to increase findability, accessibility, interoperability and reusability of datasets *(14)*.

Examples of actions that can improve data quality and interoperability include the use of clear standards and definitions, as well as the use of data dictionaries to describe the variables and values present in a given dataset. A challenge faced by Health EDRM researchers is the great variety of hazards and the lack of agreed definitions on them. Different definitions for a given hazard hampers the comparability of results from different studies, for example. As a result, it is important to have clear case and hazards definitions when conducting research in emergencies and disasters, and to present data in a machine-readable format, so that it can be retrieved and processed by computers.

4.4.10 Conclusions

Overall, data collection in the context of disasters is a challenging task that demands careful preparation and planning. Different methods can be used to gather data, and the local context, time and resources available should be considered in selecting the most suitable approach for a specific study. Science-based policy making depends on high quality research, which in turn is dependent on high quality data. Therefore, it is important to ensure that data are collected, stored and shared at high standards. A careful preparation is essential to achieve this, including the construction of a research protocol containing a clear and specific research question, objectives, the strategy to be used during data collection and how the data will be curated and analysed at a later stage.

4.4.11 Key messages

- **A specific research question and a data collection strategy that will provide adequate and sufficient information to answer this with the available resources are important for high quality research.**
- **It is fundamental to acknowledge that despite good preparation, challenges may occur. Anticipating how to deal with them can help researchers to overcome future barriers.**
- **A careful plan on how the collected data will be stored and shared in the long term will ensure that others benefit from the study.**

4.4

4.4.12 Further reading

Fakhruddin B, Murray V, Gouvea-Reis F. Disaster loss data in monitoring the implementation of the Sendai Framework [Policy brief]; 2019. https://council.science/publications/disaster-loss-data-in-monitoring-the-implementation-of-the-sendai-framework (accessed 18 January 2020).

4.4.13 References

1. Goodyear M, Malhotra N. Collection of routine and ad hoc data. 2007. https://www.healthknowledge.org.uk/public-health-textbook/health-information/3a-populations/collection-routine-data (accessed 18 January 2020).
2. Morton M, Levy JL. Challenges in disaster data collection during recent disasters. Prehospital and Disaster Medicine; 2011: 26(3): 196-201.
3. Inderfurth AK, Fabrycky D, Cohen S. The 2004 Indian Ocean Tsunami: One Year Report. The Sigur Center Asia Papers. 2005.
4. Brennan RJ, Rimba K. Rapid health assessment in Aceh Jaya District, Indonesia, following the December 26 tsunami. Emergency Medicine Australasia; 2005: 17(4): 341-50.
5. Centers for Disease Control and Prevention. Assessment of health-related needs after tsunami and earthquake--three districts, Aceh Province, Indonesia, July-August 2005. MMWR: Morbidity and mortality weekly report; 2006: 55(4): 93-97.
6. Meynard JB, Nau A, Halbert E, Todesco A (2008). Health indicators in children from Meulaboh, Indonesia, following the tsunami of December 26, 2004. Military Medicine 173(9): 900-5.
7. Dell'Aringa M, Ranzani O, Bierens J, Murray V. Rio's Mountainous Region ("Região Serrana") 2011 Landslides: Impact on Public Mental Health System. PLoS Currents: Disasters; 2018: 10.
8. Khan R, Aldaach H, McDonald C, Alam M, Huq A, Gao Y, et al. Estimating cholera risk from an exploratory analysis of its association with satellite-derived land surface temperatures. International Journal of Remote Sensing; 2019: 40(13): 4898-909.
9. Department for International Development. World first as UK aid brings together experts to predict where cholera will strike next; 2018. https://www.gov.uk/government/news/world-first-as-uk-aid-brings-together-experts-to-predict-where-cholera-will-strike-next (accessed 18 January 2020).
10. National Aeronautics and Space Administration (NASA). NASA Investment in Cholera Forecasts Helps Save Lives in Yemen. 2018. https://www.nasa.gov/press-release/nasa-investment-in-cholera-forecasts-helps-save-lives-in-yemen (accessed 18 January 2020).
11. Yu M, Yang C, Li Y. Big data in natural disaster management: a review. Geosciences; 2018: 8(5): 165.

12. Busby A. Data Quality. In: Kreis IA, Busby A, Leonardi G, Meara J, Murray V, editors. Essentials of environmental epidemiology for health protection: a handbook for field professionals. Oxford University Press: Oxford; 2012.

13. A World that Counts – Mobilising the Data Revolution for Sustainable Development. New York: United Nations; 2014, Nov. http://www.undatarevolution.org/wp-content/uploads/2014/11/A-World-That-Counts.pdf (accessed 18 January 2020).

14. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data; 2016: 3: 160018.