International Journal of Science and Research Archive

eISSN: 2582-8185

Cross Ref DOI: 10.30574/ijsra

Journal homepage: https://ijsra.net/

(Review Article)

Technical Review: Key concepts of health database management for public health workforce development in resource-limited settings

GV Fant *

Jodhpur School of Public Health (JSPH), Rajasthan, Jodhpur, India.

International Journal of Science and Research Archive, 2023, 09(02), 222–230

Publication history: Received on 28 May 2023; revised on 08 July 2023; accepted on 11 July 2023

Article DOI: https://doi.org/10.30574/ijsra.2023.9.2.0526

Abstract

The academic and professional technology communities recognize that data science and big data management are essential methods for handling and using data in modern society for decision-making that ultimately leads to societal improvements. These insights are applicable to the world-wide practice of public health. The purpose of this technical review is to provide the public health decision-maker with an overview of key concepts of health database management to help guide additional individual or workforce development in order to support public health practice. This discussion is organized with the following subheading: Definition: Health Data Management; Data Preparation Actions; Basic SQL Commands in Microsoft Access; and Basic Data Management Procedures. Public health workforce development efforts that include discussion and skill-building activities dealing with issues and topics in health data management could contribute to developing the public health competencies of all public health professionals, including those in resource-limited settings, for the benefit of reporting disease outbreaks and other health outcomes of interest to local public health authorities.

Keywords: Health Data Management; Data Preparation Actions; Basic SQL Commands; Microsoft Access; Basic Data Management Procedures; Public Health Workforce

1 Introduction

The academic and professional technology communities recognize that data science and big data management are essential for handling and using data in modern society for decision-making that ultimately lead to societal improvements [1]. These insights are applicable to the world-wide practice of public health. For example, the World Health Organization (WHO) Coronavirus COVID-19 Dashboard likely utilizes various data science and health data management techniques for handling its various data sources in order to visualize data (see dashboard: https://covid19.who.int/).

The ability of information systems to ingest different sources of data, including coded medical diagnoses and medical procedures performed on individuals in a community, requires public health and health information management professionals with essential database management skills [2]. However, these technical skills are not those needed by database administrators. Rather, they are the appropriate skills necessary to handle data for public health purposes. Some of these skills are the same that database administrators, data scientists, or data engineers require, along with an understanding of how the data will be used in public health practice, especially in resource-limited settings.

Fifty competency domains have been identified for the effective application of health informatics, including practical methods and abilities for using health database management systems [3]. Public health professionals with health database management skills need to understand, for instance, the metadata that accompany any database registry used

* Corresponding author: Gregory Fant

in connection with medical trials [4] and how to leverage this metadata in the process of creating actionable insights for the data stored in a database registry. Eventually, the study and practice of health informatics and health information technology adoption has been essential to real-time, population-based public health practice [5].

2 Public Health Professional Competency 1A4

The need to identify the areas of competency for public health professionals has led some to formally discuss and identify these competencies and link them to a set of essential public health services. Examining these efforts [6] with special attention to health information skills to support public health practice, several competencies support the utilization of data and information in order to help plan and deliver public health services, including Competency 1A4:

> Uses information technology in accessing, collecting, analyzing, using, maintaining, and disseminating data and information

This competency is important because it has been shown to align with essential public health service, such as being able to monitor the health status of a population in order to identify and investigate community health problems. The competency also allows public health professionals to inform community members on pressing health issues facing their populations.

The purpose of this technical review is to provide the public health decision-maker with an overview of the key concepts of health database management to help guide additional individual or workforce development in order to support public health practice. The explanation of these concepts may help public health professionals with continued skill development activities. This discussion is organized with the following subheading: Definition: Health Data Management; Data Preparation Actions; Basic SQL Commands in Microsoft Access; and Basic Data Management Procedures. The implications of these topics for global public health practice are briefly considered.

3 Definition: Health Data Management

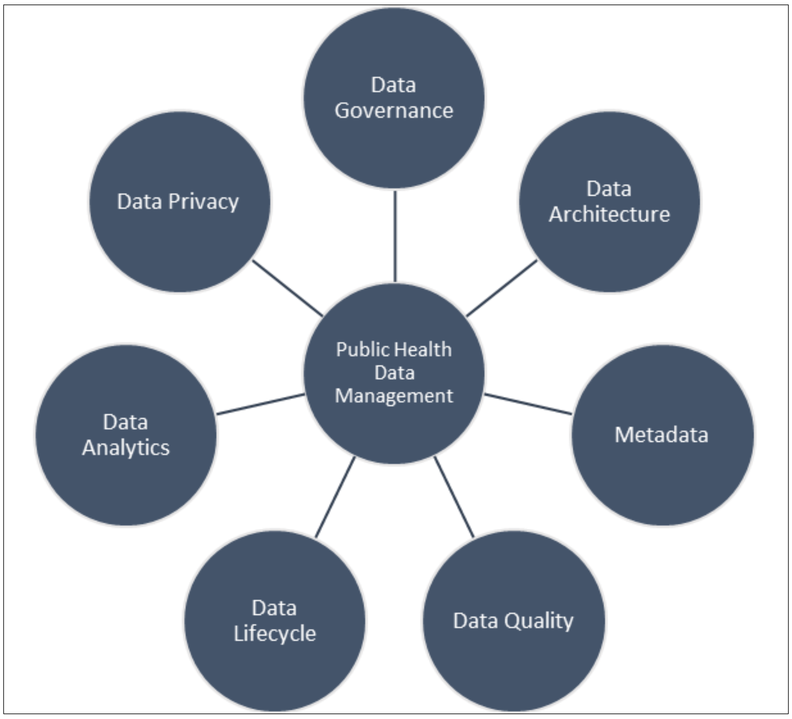
It is important for decision-makers to understand the main concepts of health data management. A functional definition of the concept is adapted from the Encyclopedia of Public Health [7]:

> Health data management comprises all activities related to the management of health data as a valuable resource. It encompasses acquiring, entering, processing, coding, outputting, retrieving, and storing data gathered from different areas of public health and health care. This concept embraces the validation and control of data according to legal or professional requirements.

This definition acknowledges that health data are valuable resources for public health decision-making and for the public health organization as a whole. Health data management is a tacit activity of a public health information system where health data are managed. In this section, we briefly describe some important concepts of health data management.

3.1 Principle Components of Public Health Data Management Framework

Health data are powerful tools for helping leaders to understand the health needs of a given population, the utilization of organizational resources to address those needs, and to permit the evaluation of efforts that guide further public health action [8]. Public health decision-makers need to have confidence that the data are accurate, consistent, and comprehensive. Accurate health data are correct, valid, and free from error. Consistent health data are reported reliably and in the same format over time. Comprehensive health data means that the data have been collected for the task at hand [9]. To ensure these characteristics, a data framework is needed for most organizations (Figure 1).



Adapted from: https://www.trellance.com/data-management-framework-7-essential-components/ (accessed: 27 December 2022)

Figure 1 Principal Components of Public Health Data Management Framework

Data relied upon for public health decision-making depend on following a good data management framework. A data infrastructure that has been constructed on sound principles is likely to yield reliable data for public health decision-making. A public health data management framework, adapted from Trellance [10], has seven components:

3.2 Principle Components of Public Health Data Management Framework

- **Data Governance** which provides the overarching support to data management through stewardship, policies, processes, standards, and adherence to leading practices.
- **Data Architecture** which provides the infrastructure for the storage, integration, and use of data throughout the organization.
- **Metadata** which allows you to use data more efficiently by providing critical information about data attributes.
- **Data Quality** which provides the structure necessary to have data that fulfills the needs of the public health organization.
- **Data Lifecycle** which follows the data throughout the public health organization, providing integrity from the initial introduction into the organization through the final deletion from or storage within the public health organization.
- **Data Analytics** which applies statistical and visualization techniques that lead to valuable insights that can help the public health organization implement better public health actions.
- **Data Privacy** which supports the needs of the public health organization to share data both internally and externally using agreed upon practices and electronic health information exchange protocols.

When the public health data are considered and managed using the above framework, the utilization of these data in a public health information system is more authoritative. A multi-disciplinary team is also needed for the various aspects of this data framework to be successful. Public health decision-makers can be confident that the data collected are accurate, consistent, and comprehensive. Then the decision-makers can move forward with requesting health data analytics and the interpretation of data in order to address the needs of the population and/or the organization.

4 Data Preparation Actions

In order for health data to be accurate, consistent, and comprehensive, it must be prepared for use. Data are generally messy at the time of collection and many actions are needed to make the data ready for use. These data preparation actions include many steps [11].

4.1 Description of data preparation steps

- **Raw Data.** Extracting raw data from a database or spreadsheet. These data were obtained with no additional cleaning or error correction having been performed. Therefore, it retains errors, missing values, omissions, and other inconsistencies.
- **File Formats.** Data Scanning: Recognize the file format found in the dataset. Common file formats include comma separated values (.csv), tab separated data (.tab), Microsoft Excel Open XML spreadsheet (.xlsx), JavaScript Object Notation (.json), text file (.txt) and other types. This will affect data import procedures.
- **General Inspection.** Data Scanning: Visually inspect the data file in the computing environment. Be aware of the values in the rows and columns as well as the completeness of data cells. Consult a data dictionary that describes the data columns and critically access the data file for agreement with the data dictionary. Notice the data types for each column and confirm which column is the target variable.
- **Combine Datasets.** Determine whether the various data files will need to be stacked or joined to form a more complete data file.
- **Reshape Datasets.** Sometimes data files contain transactional data for individuals in the row, and the result is multiple rows of data for the same person because each transaction is recorded on a row resulting in a long datafile. The more common approach for data analysis is to utilize a datafile where each row contains data by column name for each person. It may be necessary to reshape the data file from transaction-based/long file to a record-based/wide file.
- **Descriptive Statistics.** Data Scanning: Perform descriptive statistics on the reshaped data file to get a better idea of the characteristics of each data column as well as the entire data file.
- **Data Visualization.** Data Scanning: In additional to descriptive statistics, the use of data visualization techniques helps to better understand the data file. Data visualization may be useful in data transformation, identifying data errors, and recoding data columns or variables.
- **Data Cleanup.** After scanning the data and reshaping it, important actions must be taken to correct errors in or clean-up a data file prior to data analysis. It may also be the time to construct a new clean data file for data analysis.
- **Data Leakage.** If data science methods will be used, then it is necessary to partition the clean data file (creating a training dataset and, separately, a test dataset) that will be used for analysis. Data leakage occurs when training data is inadvertently included in the test data.
- **Missing Data.** Instances of missing data will have to be addressed.

The completion of these data preparation steps within a relational database management system (RDBMS) before the data analysis is conducted is an important part of the public health data management framework. The prior steps contribute to assuring Data Quality in the framework. Additionally, the steps also support the Data Architecture, Metadata, and Data Lifecycle components of the public health data management framework (Figure 1).

5 Basic SQL Commands in Microsoft Access

A discussion of health data management for the public health decision-maker should include a description of the basic commands used in a RDBMS and also a comparison of basic data management operations performed by selected software programs. As commonly understood, RDBMS is a collection of data tables, linked by at least one primary key to identify a unique record. Each data table contains columns of data that align with the variables of interest from the operational question of interest. Each data table also contains rows of data, with data for each column, and for a unique record, or observational unit. These data columns and rows are often manipulated in a RDBMS using a database programming language known as SQL.

The basic SQL commands used in a relational database are described below [12]. These commands are organized by database operation, the most common of which are known by the acronym CRUD which stands for Create, Read, Update, and Remove.

5.1 Database Operation: Creating a database

5.1.1 Create table statement

Syntax

```
CREATE TABLE Name
(Column Name Data type Field Size, [NULL | NOT NULL]
[optional constraints]);
```

5.1.1.1 Description of statement

This is the basic syntax used to create a database table. The CREATE TABLE statement defines the table name, column names, data types and field size.

5.2 Database Operation: Reading (or Selecting) records from a database

5.2.1 SELECT statement

Syntax

```
SELECT ColumnName(s)
FROM TableName(s);
```

Description of statement

This is the basic syntax to retrieve a record from the database table. The SELECT statement specifies the data column(s) that are needed while the FROM statement identifies the data table containing the data column.

5.3 Database Operations: Retrieving and Grouping records from a database

5.3.1 WHERE statement

Syntax

```
SELECT ColumnName(s)
FROM TableName(s)
WHERE [search condition];
```

Description of statement

This basic syntax is used to filter records retrieved from a database table. The SELECT statement specifies the data column(s) that are needed while the FROM statement identifies the data table containing the data column. The WHERE statement specifies a specific condition for the retrieval of the record(s).

5.3.2 GROUP BY statement

Syntax

```
SELECT Column Name(v1), COUNT (ColumnName) AS TotalItems
FROM TableName(s)
GROUP BY [ColumnName (v1) for unique record];
```

Description of statement

This basic syntax is used to group records that have been retrieved from a database table. The SELECT statement specifies the data column(s) that are needed, and COUNT (columnname) is used to count the rows in a column excluding NULL values. The FROM statement identifies the data table containing the data column, and the GROUP BY statement is used with the aggregate function COUNT to combine groups of records into a single record.

5.3.3 DISTINCT keyword

Syntax

```
SELECT DISTINCT ColumnName
FROM TableName;
```

Description of statement

This is the basic syntax to retrieve a record from the database table. The SELECT statement along, with DISTINCT keyword is used to show values in a specific column. The FROM statement identifies the data table containing the data column.

5.3.4 Database Operations: Updating database records

Syntax

```
UPDATE TableName
SET ColumnName=Value
```

WHERE [search condition];

Description of statement

This is the basic syntax to update records in a database table. The UPDATE statement identifies a specific database table that will be updated with more current data. The SET keyword specifies both the data column that will be updated and the data value that will be inserted into the data column. The WHERE statement specifies a specific condition for the retrieval of the record(s).

5.3.5 Database Operations: Deleting a database record

DELETE Statement

Syntax
DELETE FROM TableName
WHERE [search condition(s)];

Description of statement

This is the basic syntax to remove records from the database table. The DELETE FROM statement identifies the database table that will be changed and the FROM statement identifies the specific record(s) that will be removed.

6 Basic Data Management Procedures

Public health decision-makers may find it useful to be knowledgeable about the basic data management tasks that are needed to prepare data for use. There are several components of the data management process [13]:

- **Get and define data.** Identify pertinent source data, import the data into an appropriate software package where data management will be performed, and, using a data dictionary, define the key elements of the imported data file in a way that will be read by the data management software package.
- **Combine data from various sources.** Import, store, and combine various database files in the data management software package.
- **Clean the data.** Clean and organize the database files that will be used in data analysis, then look for anomalies and begin the correction process, where possible.
- **Aggregate, select, sort, and weight cases.** Continue the data cleaning process by organizing the data and determining whether or not the entire database file will be used in subsequent analysis or if a sample of the data are needed. Aggregate, select, sort, store, and weight cases in the database file that will be used, as needed.
- **Transform data.** Continue with the data cleaning process and recode or transform data columns, as needed. Address issues of missing data. Create new data variables that are ratios of data variables and are needed for later analysis.
- **Restructure data for analysis.** Restructure database files as needed for the expected analysis methods.
- **Export data and results.** Export the final database file that will be used in the analysis in the file format required by another application (e.g., IBM SPSS Statistics, NCSS, KNIME, Orange, IBM SPSS Modeler, etc.).

Software is available to perform many data management tasks and this software is capable of performing the required tasks. The choice of which software to use depends, in part, on what was taught to the individual who will perform the data management tasks, as well as the features of the software that are needed in the data preparation efforts. The following table (Table 1) compares three common software packages that can perform basic data management tasks for public health information systems.

Table 1 Comparison of Basic Data Management Tasks Using Software for Public Health Information Systems [14]

	SPSS Menu	**SPSS Syntax (using ODBC)**	**NCSS Menu**	**SAS Programming**

Load Data	File>Open>Data (select database file)	File>New>Syntax
		GET DATA /TYPE=ODBC /CONNECT= ‘(insert the path to the database)’ /SQL = ‘SELECT * FROM CombinedTable’.
		EXECUTE
			File>Open (select database file)	DATA DISKIN;
				INFILE 'filename.DAT';
				INPUT v1 v2 v3;
				RUN;

Data Screening	Analyze>Descriptive Statistics>Explore	Easiest method: Utilize the menu window, then before “OK” select “Paste.” This will “Paste” the SPSS Syntax into a separate window that can be “saved” and modified using SPSS Programming Syntax rules and applied to any SPSS dataset when correctly “opened.”	Analysis>Descriptive Statistics>Data Screening	PROC Sort DATA=sample;
				By v1;
				RUN;
				PROC Print Data=sample LABEL;
				By v1;
				ID ids;
				VAR v2 v3
				FORMAT v2 v3 3.0;
				RUN;

Describe Data	Analyze>Descriptive Statistics> Descriptives	Easiest method: Utilize the menu window, then before “OK” select “Paste.” This will “Paste” the SPSS Syntax into a separate window that can be “saved” and modified using SPSS Programming Syntax rules and applied to any SPSS dataset when correctly “opened.”	Analyze>Descriptive Statistics>Descriptive Statistics	PROC Means data=filename;
				VAR v1 v2 v3 v4;
				RUN;

Recode Data	Transform>Recode>Into Different Variable Transform>Recode> Into Same Variable	Easiest method: Utilize the menu window, then before “OK” select “Paste.” This will “Paste” the SPSS Syntax into a separate window that can be “saved” and modified using SPSS Programming Syntax rules and applied to any SPSS dataset when correctly “opened.”	Data>Transformation>Recode Transformation	SAS does not have a recode command, so a series of if-then/else commands must be used.

Data Visualization	Graphs>Gallery> Scatter	Easiest method: Utilize the menu window, then before “OK” select “Paste.” This will “Paste” the SPSS Syntax into a separate window that can be “saved” and modified using SPSS Programming Syntax rules and applied to any SPSS dataset when correctly “opened.”	Graphics>(make selection)	Title “height and weight;
				
				PROC sgplot data=(filename);
				scatter x=height
				y=weight;
				RUN;
				title;

Save Data	File>Save or Save As (select format)	Easiest method: Utilize the menu window, then before “OK” select “Paste.” This will “Paste” the SPSS Syntax into a separate window that can be “saved” and modified using SPSS Programming Syntax rules and applied to any SPSS dataset when correctly “opened.”	File>Save or Save As (select format)	libname U “U:\” ;
				
				data u.class;
				set sashelp.class;
				
				RUN;

6.1 Implications

The World Health Organization (WHO) Coronavirus COVID-19 Dashboard [ibid] relies on accurate, good-quality health data for data analysis, reporting, and visualization for all countries. By using proper data handling methods (including those discussed here), good-quality data will be available for health data analysis, health planning, and health data visualization efforts in order to aid public health decision-making. Health data management, data preparation, the role of Microsoft Access for SQL database operations, and the basic data management procedures are not unique to the public health arena but rather take-on added significance when considered in the process of using health data for protecting the health status of a population and sharing essential health data with other public health professionals, especially during local outbreaks or pandemics. Public health workforce development efforts that include discussion and skill-building activities dealing with issues and topics in health data management could contribute to developing the public health competencies of all public health professionals, including those in resource-limited settings, for the benefit of reporting disease outbreaks and other health outcomes of public health interest to local public health authorities. Then, public health authorities could share good health data with national, regional, and global decision-makers for the benefit of global public health.

7 Conclusion

Public health practice is based, in large part, on the availability and interpretation of data. These data need to be of high quality and require careful collection, preparation, analysis, interpretation and storage. This technical review provides an overview of key concepts of health database management. The framework for the consideration of public health data management (Figure 1) reminds all members of the public health workforce of the critical components that lead to data and insights upon which public health practice can be based. This technical review may provide a user-friendly overview to topics that can be expanded upon in a hands-on training environment for public health professionals who are interested in pursuing and leading activities in health data management for public health action.

Compliance with ethical standards

Acknowledgments

I appreciate the ongoing support provided to me by the Jodhpur School of Public Health, Jodhpur, India, where I have been a pro-bono/volunteer faculty member since 2013. I am grateful for the opportunity to help and work with other public health professionals in India and the WHO South-East Asia Region.

Disclosure of conflict of interest

There are no financial conflicts of interest related to this technical review.

Authors information

Dr GV Fant is a public health epidemiologist, public health consultant, and visiting faculty member at JSPH, Rajasthan, India, having taught subjects in epidemiology, biostatistics, field epidemiology, and health database concepts to Indian MPH-degree students and public health professionals since 2013. He is, also, the Executive Director of the Society for Epidemiology at JSPH. Dr. Fant earned his doctorate (PhD) from University of Nebraska in 1997 and doctorate (PhD-h.c.) in health sciences and public service from Poornima University in 2021. Dr. Fant earned professional recognition as an Epidemiologist from the American College of Epidemiology (MACE) in 2002, the Society for Epidemiology at JSPH (MSEpi) in 2019, CIMP-Data Science in 2021, and CEHRS in 2022. Beginning in 1997, Dr. Fant has served as a U.S. civil servant and is an epidemiologist in Northern Virginia, USA.

References

[1] Paul PK, Hidalgo RS. Editorial: Data Science and Analytics-The emergence and issues in technology and academics. Intl J Applied Science and Engineering. 2022 10 (1):1-2

[2] Stanfill MH, Marc DT. Health Information Management: Implications of Artificial Intelligence on Healthcare Data and Information Management. Yearb Med Inform. 2019 Aug; 28(1):56-64. doi: 10.1055/s-0039-1677913. Epub 2019 Aug 16. PMID: 31419816; PMCID: PMC6697524.

[3] Jidkov L, Alexander M, Bark P, et al. Health informatics competencies in postgraduate medical education and training in the UK: a mixed methods study. BMJ Open 2019;9:e025460. doi:10.1136/bmjopen-2018-025460

[4] Stausberg J, Lobe M, Verplancke P, et al. Foundations of a Metadata Repository for Databases of Registers and Trials. In: Adlassnig KP (eds) Medical Informatics in a United and Healthy Europe. IOS Press; 2009. doi: 10.3233/978-1-60750-044-5-409

[5] Williams F, Oke A, Zachary I. Public health delivery in the information age: the role of informatics and technology. Perspect Public Health. 2019 September; 139(5): 236-254. doi: 10.1177/1757913918802308.

[6] The Council on Linkages between Academic and Public Health Practice. Crosswalk of the 2014 Core Competencies for Public Health Professionals and the Essential Public Health Services, October 2015. Available from: https://www.phf.org/resourcestools/Pages/Crosswalk_2014_Core_Competencies_and_Essential_Services.aspx

[7] Bocking W, Trojanus D. Health Data Management. In: Kirch W (ed) Encyclopedia of Public Health. Dordrecht: Springer, 2008.

[8] Christopher GG, Zimmerman EB, Chandra A, Martin LT. Charting a Course for an Equity-Centered Data System: Recommendations from the National Commission to Transform Public Health Data Systems. Princeton: Robert Wood Johnson Foundation; 2021.

[9] Davis N, Shiland B. Statistics and Data Analytics for Health Data Management. St Louis: Elsevier; 2017.

[10] Trellance.com. Data Management Framework – 7 Essential Components for Credit Unions [Internet] 2020 March [cited 2022 Dec 27] Available from: https://www.trellance.com/data-management-framework-7-essential-components/

[11] Hoyt M, Muenchen R. Data Preparation and Exploration – Applied to Healthcare Data. Informatics Education; 2020.

[12] Allison CL, Berkowitz NA. SQL for Microsoft Access, 2nd edition. Burlington: Wordware/Jones & Bartlett; 2008.

[13] Levesque R. SPSS Programming and Data Management – a guide for SPSS and SAS Users. Chicago: SPSS Inc; 2003.

[14] Sources used to construct Table 1: https://stats.oarc.ucla.edu/sas/modules/creating-and-recoding-variables-in-sas/
https://documentation.sas.com/doc/en/pgmsascdc/9.4_3.5/grstatproc/n19gxtzyuf79t3n16g5v26b73ckv.htm#n0iuhw0kbsunrqn1lep9q1c8mrv0
https://www.ssc.wisc.edu/~hemken/SASworkshops/SASWindows/saveSASdata.html; NCSS 2023 Statistical Software (2023). NCSS, LLC. Kaysville, Utah, USA, ncss.com/software/ncss