CASE STUDY e-ISSN: 2349-0659 p-ISSN; 2350-0964

Public Health Data Science: Main Concepts and Case Study for Public Health Professionals in India

G. V. Fant[1,2]*, A. Purohit[3]

Abstract

There is growing use of big data and public health data sources in many areas of public health need. Data science is an interdisciplinary field that may have great utility in public health and epidemiology in the search for solutions to many public health needs or problems. Real-world, big data that are continually being collected permit the use of data science methods to address public health problems. This review describes essential concepts in data science from a public health perspective. A case study that utilizes data science thinking is offered to illustrate the application of public health data science for public health professionals in India and the South-East Asia Region.

Keywords: Cross-industry standard process for data mining, Data modeling, Data science, Domains of public health data science, KNIME analytics platform, Public health

Asian Pac. J. Health Sci., (2020); DOI: 10.21276/apjhs.2020.7.4.18

Introduction

Globally, there is enthusiasm and appreciation for the use of big data and public health data sources in health services, public health, environmental health, behavioral and social health, and many areas of health need.[1] This led to the need to describe how modern analytic methods can be used to address issues pertaining to deriving meaningful insights from data for public health and epidemiologic action.[2-4] Machine learning includes a family of methods that can be used to help address the data needs for public health action. The methods of machine learning are part of the professional knowledge, methods, and applications of data science.[5-7] Therefore, it seems essential to describe data science and its application in public health.

In this review, the main concepts of data science will be described, and such the key principles of supervised and unsupervised learning will also be discussed. This will be followed by a description of the domains relevant to public health data science. Finally, a possible application and case study of public health data science will be presented using a data science open-source software platform to address a public health problem.

Concepts in Data Science for Public Health

Data Science

Data science is not the application of scripting computer programming to a dataset. Rather, it is an interdisciplinary field of scientific methods, processes, and systems to extract knowledge or insights from messy, big data in various forms, either structured or unstructured data, and is similar to data mining.[5,6] The amount of public health data collected globally in any year requires methods that utilize advanced professional knowledge and methods from the field of data science that can be used with labeled data (or structured data) and unlabeled data (or unstructured data) to address public health or epidemiologic problems affecting the population.

Data science is a discipline that uses quantitative methods, including statistics, mathematics, technology, domain knowledge, and professional curiosity, to develop algorithms designed to discover patterns, predict outcomes, and find optimal solutions to complex problems of a specific domain. This discipline employs techniques and theories from mathematics, statistics, information science, computer science, data lakes, data mining, data warehousing, databases, data visualization, artificial intelligence, and big data, to name just a few. A sound foundation in applied statistics, including graphing data, seems to be a prerequisite to embarking upon the study and use of data science methods for populations.

For this review, we offer a functional definition of public health data science. This definition seems necessary as we consider the application of data science for public health problems: Public health data science is the application of data science knowledge and methods along with traditional analytic methods of public health (public health epidemiology and biostatistics) to address public health issues and matters of concern.

Data Science: Data Science Steps in Problem – Solving in Public Health

When attempting to address a public health problem using a big data source, a problem-solving framework is useful. The

[1]Visiting Professor, Public Health (Epidemiology and Biostatistics), Jodhpur School of Public Health, Rajasthan, India

[2]Executive Director, Society for Epidemiology, Jodhpur School of Public Health, Rajasthan, India

[3]Founder and President/CEO, Jodhpur School of Public Health, Jodhpur, Rajasthan, India

Corresponding Author: Dr. G. V. Fant, Jodhpur School of Public Health, Plot No 131, 2nd Polo Ground, Paota, Jodhpur, Pin 342006, Rajasthan. E-mail: gregory.fant@jsph.in

How to cite this article: Fant GV, Purohit A. Public Health Data Science: Main Concepts and Case study for Public Health Professionals in India. Asian Pac. J. Health Sci., 2020; 7(4):70-76

Source of support: Nil

Conflicts of interest: None

Received: 30/08/2020 **Revised:** 30/09/2020 **Accepted:** 28/10/2020

framework adopted by those using data science methods is a step-wise process adapted from the data mining field. This process is known as the Cross-industry Standard Process for Data Mining (CRISP-DM).[6,7]

The CRISP-DM process has the following steps:

1. Problem statement – Articulation of the problem to be addressed using data science methods and techniques. This "problem statement" should be amenable to analysis using data science methods.
2. Subject matter understanding – The problem statement has its basis in a specific domain of knowledge. Professional domain knowledge is essential for the successful completion of a data science project.
3. Data understanding – Identifying the data source and verifying data quality.
4. Data preparation – Selecting data, cleaning data, preparing the data, and reformatting the data for data analysis and data modeling.
5. Data modeling – Selecting the data modeling technique, partitioning the data (training data; test data), building the data model, and assessing the data model.
6. Data model evaluation – Evaluating the results of the data model before deployment to determine how the model addresses the "problem statement" using test data.
7. Data model deployment – Using the data model on unseen, process data to generate a report or produce another dataset for analysis; monitor and maintain the process of using the data model on unseen, process data.

The use of data science knowledge and methods in public health is different from the use of public health research methods and statistics to address a public health problem. In the first instance, real world, existing data are used with data science methods to develop data model that is then used with real-world, unseen data of the type that was used to develop the data model to address a public health problem. In the second instance, data are collected using public health research methods and statistics to develop a proposed model to explain the phenomenon. This explanation is, then, generalized to the population in the hope that the proposed model behaves probabilistically in the population as it did with the collected data to address the public health problem.

Data Preparation

In practice, identifying the source of real-world data that will be used to address the "problem statement" and preparing the data for use in model development is known as Data Preparation or Data Wrangling. Keeping in mind that not all data are useful, the data preparation phase requires many technical skills, logic, intuition, and curiosity to be applied to understandings of various sources of data and preparing the data for data modeling. While various authors have highlighted certain technical skills and abilities necessary for data preparation,[5-9] the main actions include the following:

- Finding the data (raw data/initial data)
- Accessing the data for quality (raw data/initial data)
- Understanding the layout and patterns in the raw data/initial data
- Selecting data elements from the raw data/initial data for the project
- Preparing the data for a new data set that will be used for the project
- Creating a new dataset for the data science project that will address the "problem statement"
- Understanding the layout and patterns of this new dataset
- Partitioning the new dataset into a training dataset (80%) and testing dataset (20%).

Many of these same authors[5-9] explain that the process of getting the data, understanding it, cleaning it, and preparing it for further data visualization and data modeling as both difficult and time-consuming from a data science perspective. This may be linked to using real-world data that were not explicitly collected to address a public health problem. By contrast, from a statistics perspective involving public health research methods and sample size, data preparation is likely less involved because the data were specifically collected to address a public health problem.

Data Visualization

A fundamental part of data science includes data visualization.[6,10] Various types of business intelligence software, statistical software packages, open-source programming languages, and even spreadsheet applications can assist with data visualization (it is beyond the purpose of this review to show the various types of data visualization techniques). That said, it is important to realize that data visualization is a graphical way to examine data to identify patterns in it or to communicate a message.

The study and practice of epidemiology have been using data visualization methods for a very long time. Recall the important contributions of John Snow to the study and practice of epidemiology [Figure 1].

This figure includes the citation (Wikipedia; National Geographic):

> *"This map of London was created by John Snow in 1854. London was experiencing a deadly cholera epidemic, when Snow tracked the cases on this map. The cholera cases are highlighted in black. Using this map, Snow and other scientists were able to trace the cholera outbreak to a single infected water pump."*

The principles of data visualization, for the perspective of public health data science, are as simple and profound as John Snow and the "Cholera Map" from London in 1854: He told a story with data and a map.

The technologies available to public health epidemiologists (public health data scientists) do not displace the fundamentals of what John Snow illustrated for us: Every story has a beginning, middle, and an end. Public health and epidemiologic data help us tell a story and to see the patterns in data that lead us to public health problem-solving and public health action.[5-6,11]

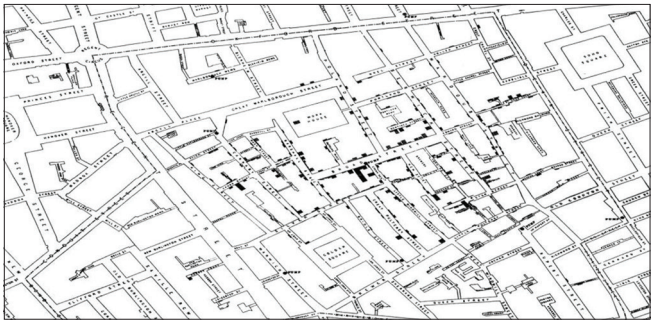

Figure 1: 1854 Map used by John Snow and the Broad Street Pump Handle (Wikipedia)

Data Modeling

In data science, data modeling has its foundations in applied statistics and data analytics. In data science, professional knowledge and methods seem expanded to include labeled quantitative data, unlabeled data, and text data (Image data are usually connected to artificial intelligence and not a part of this review).

Remember, from the data preparation phase, a data set was prepared to address the "problem statement." Moreover, the dataset is very large. These data may be divided into two parts (or the data were partitioned): Training set (80% of the total dataset) and test set (20% of the total dataset) depending on the data modeling-data analysis needs. Using the training dataset, statistical and analytic methods permit the extraction of meaningful insights and solve problems. We will compare the results from the training dataset with the testing dataset.

There are three broad families of data modeling methods: Supervised learning; unsupervised learning; and associations.[5-10,12] The reader who is skilled in epidemiology and biostatistics (possibly learned in a Master of Public Health MPH degree program) can think of the various statistical methods from these subjects to be assigned to supervised learning in the study and practice of data science.

In Table 1, the families of data modeling methods are briefly categorized.[5,7,12] This categorization may enable the reader to better learn how these methods behave in a software package of her/his choosing. It does not matter what software package is used to learn these methods provided that the software package is able to perform the analysis.

Machine learning (ML) includes techniques such as Bayesian methods; neural networks; natural language processing; and decision trees. ML is about data pattern recognition by the computer without explicit programming to perform a specific task. In this way, the computer teaches itself to employ these techniques on big data to identify patterns in it.[12,13] The data modeling phase is also concerned with issues of model accuracy, such as comparing results of the modeling between the training and test datasets.

Table 1: Categorization of common data modeling families used in data science

Data modeling family	*Brief description*
Supervised learning	Labeled data to predict a target field, using one or more predictors Examples: • Linear regression • Logistic regression • Decision trees
Unsupervised learning	Unlabeled data to group or cluster records based on more or more fields Examples: • K-Means • Anomaly • Natural language processing (used for text data)
Association	To describe relationships between categories Examples: • Apriori • Sequence • Association rules

Data Model Evaluation

In the evaluation phase, the data modeling efforts are examined in terms of the original "problem statement."[6,7] Domain knowledge is needed to evaluate the data modeling efforts for whether or not the data model addresses the "problem statement" and the utility of the result. Accuracy of the data model (from the data modeling phase) is a different matter from the evaluation of the data model for its ability to address the "problem statement."

Data Model Deployment and Decision-making

In the deployment phase, we are taking our prior model and using it in a real-world setting. Unseen, real-world data are now presented to the data model.[6,7] The outcome or deliverable of the deployment phase could be as simple as generating a report of results, using the model to create a dataset for further analysis, deploying the model to address a real-world problem in the domain knowledge area, monitoring the use of the data model in a real-world application, etc. In the context of the real-world setting and the original "problem statement," we are now making decisions based on the real-world data and the data model that were constructed.

Domains of Data Science Used to Address the Public Health and Epidemiologic Conditions Facing a Population

Upon reflection, three main skill domains – programming, statistics, and data mining, along with public health and epidemiology – likely comprise public health data science [Figure 2]. Other areas of expertise may also be needed such as typical public health research and statistical methods, public health database management skills, and an appreciation for computer algorithms. A tacit, necessary skill set for the public health epidemiologist who intends to practice in the field of public health data science is the need to upskill personal technology capabilities (e.g., refine statistical software skills, ETL – extract, transform, load – skills, relational database skills, familiarity with an "open-source" programming language, such as the current version of Python, and data mining software). The study and practice of public health data science need all of these skill sets applied to messy big data to help address public health and epidemiological problems impacting a population.

Case Study: Public Health Data Science for Public Health Professionals in India

Background

In a blog-post at PATH.ORG, Neeraj Jain, PATH country director for India since 2016, reflected on the ways in which public health in India changed over the last decade.[14] Jain discussed seven public health issues:

1. Decreasing trend in communicable disease
2. Focus on prevention
3. Reduced neonatal mortality rates
4. Addressing antimicrobial resistance
5. Improved nutrition
6. Use of digital health and artificial intelligence for social impact
7. Stronger government accountability.

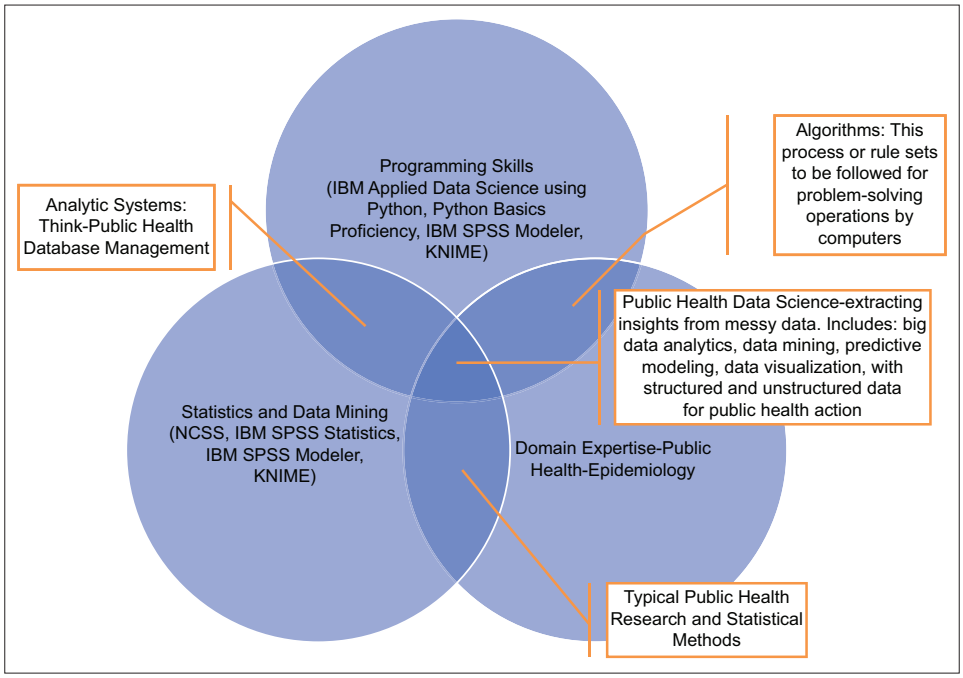

Figure 2: Domains comprising public health data science

Problem Statement

The question that comes to mind is a simple one: Do we really see the improved nutritional status in India using open, public-use data for the past decade? We would expect to see an annual improvement in nutritional status in India from the prior decade.

Public Health Context

From a public health perspective, the nutritional status of a population is a major concern.[15,16] The study and practice of epidemiology remind the public health professional that population-level, trend data on a public health condition is an indicator of the pattern of the health condition in the population and is influenced by many factors, including governmental action and public health interventions.[17-19] Public health knowledge development using big data includes:

- Looking for data patterns (including data visualization)
- Studying prevalence and incidence in a population
- Predictive analytics: Predictive modeling and simulation.

Malnutrition in India and the Indian Subcontinent is among the highest in the world. The measurement of malnutrition in a population (as in some type of annual, prevalence value for a population) would provide public health leaders, planners, and epidemiologists with a sense of the nutritional status in a country.

Identification of Data and Data Preparation

Global public health data of high quality are difficult to find. To address this, the first place consulted was the open-data repository of the World Bank, entitled "Databank." A dataset entitled "Prevalence of undernourishment (% of population)" was found.[20] From 2000 to 2018, the prevalence of undernourishment (% of population) decreased worldwide from a high of 13.7% to a low of 8.6%. The World Bank report (in the notes accompanying the data):

"Population below minimum level of dietary energy consumption (also referred to as prevalence of undernourishment) shows the percentage of the population whose food intake is insufficient to meet dietary energy requirements continuously. Good nutrition is the cornerstone for survival, health and development. Well-nourished children perform better in school, grow into healthy adults and in turn give their children a better start in life. Well-nourished women face fewer risks during pregnancy and childbirth, and their children set off on firmer developmental paths, both physically and mentally."

The dataset contained 263 rows of data for countries and regions of the world, with 62 columns of data in text, numbers, or blank/missing. The numerical data did not contain the same number of decimals. This was a messy dataset.

The main data preparation issue was how to extract only the data for India and only the columns of data that contained pertinent data values. KNIME Analytics Platform[9,21] was the open-source, data science software package used to prepare the data. "Visual programming" was used to prepare the data, and the annotated nodes highlight the programming operations that occurred in the workflow on the Open KNIME Workbench [Figure 3].

KNIME analytics platform was downloaded to a personal computer. The Open KNIME Workbench is understood from left to right. The dataset was downloaded from the World Bank open-data website onto a personal computer using an Excel Reader node. A Row Filter node was attached to extract only the data for India, the country of interest. Next, a Column Filter node was used to select only the columns of data for India that were needed to examine annual undernourishment in the population. The data were wide and needed to be transformed into a vertical array using a transpose node. Finally, the data were output in two different ways: (1) A line plot and (2) a CSV dataset for further analysis.

Data Modeling

The line plot produced by KNIME [Figure 3] had the general shape of the data reports found at the World Bank website. The CSV dataset was, indeed, exported to a specific folder on the personal computer for further analysis.

The CSV dataset was then opened using MS-Excel [Figure 4] to produce a line graph [Figure 5]. It was comforting to notice that the CSV dataset looked the same in both KNIME(see Node for "line plot" in the KNIME Workbench) and MS-Excel. This Excel-based line graph was easier to use.

Data Model Evaluation

The line graph in Figure 5 indeed shows a decreasing trend in undernutrition in India from 2001 to 2018. Steep and steady declines in undernutrition were seen in the population data reported for India between 2004 and 2018. This decrease in undernutrition in the population of India suggests an improvement in nutritional status for the country. This is consistent with the blog-post from PATH.ORG that was cited at the beginning of this application case study.

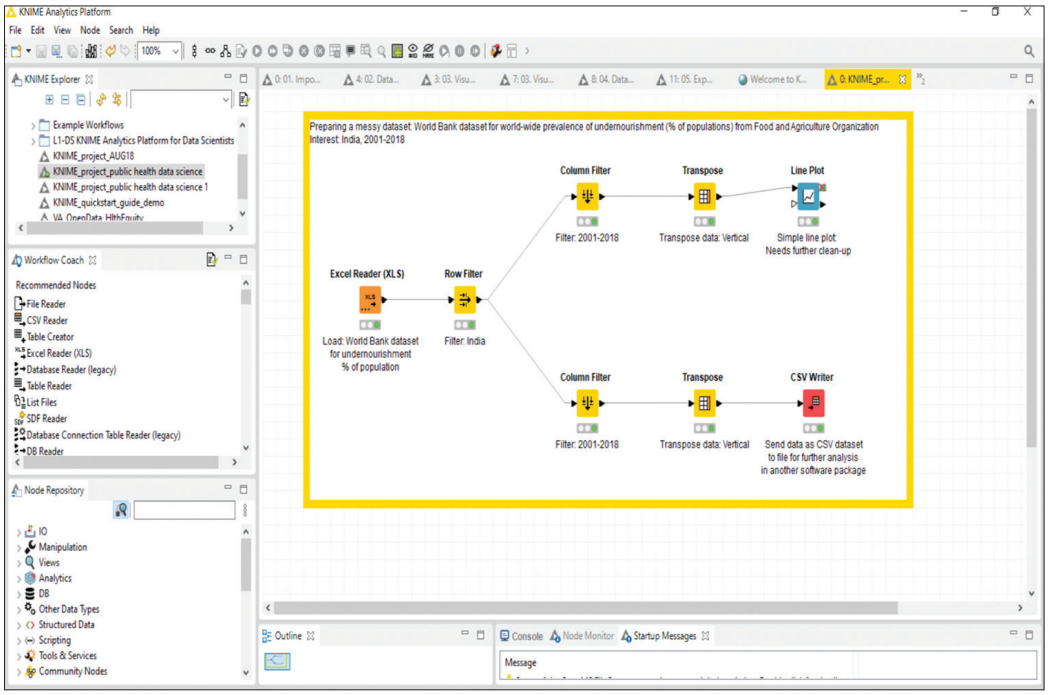

Figure 3: Data preparation using KNIME analytics platform

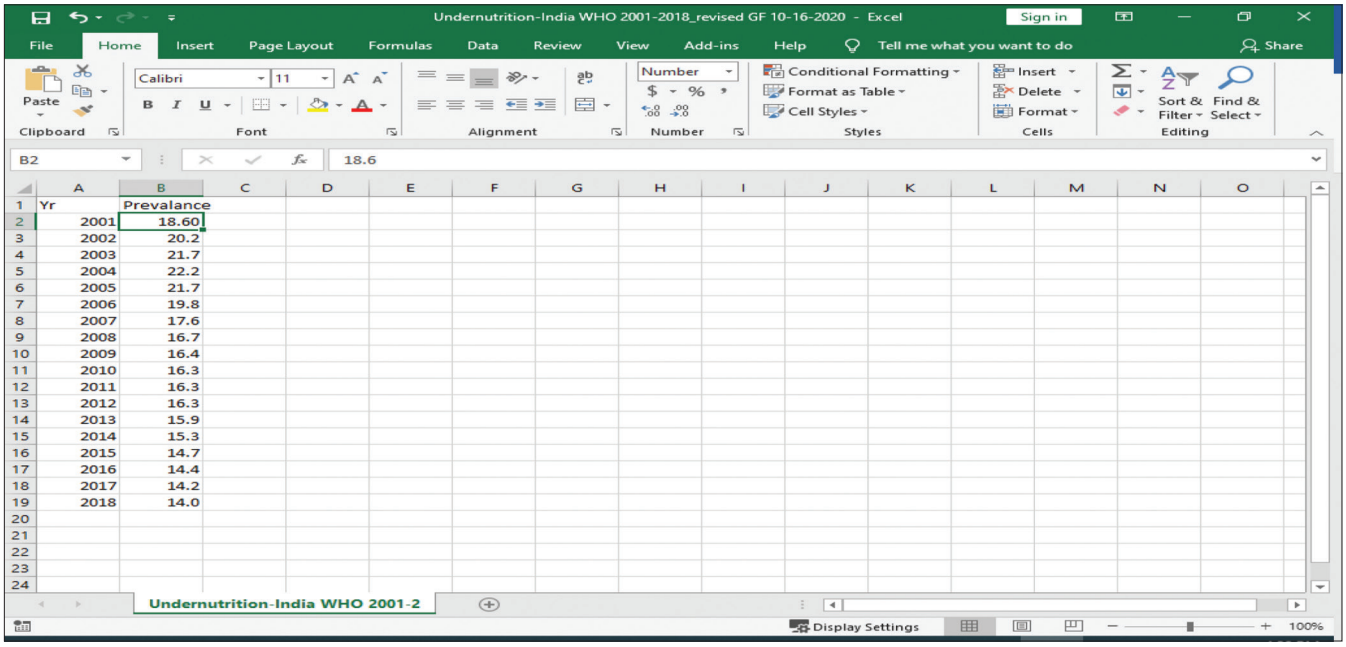

Figure 4: CVS file "opened" in MS-excel

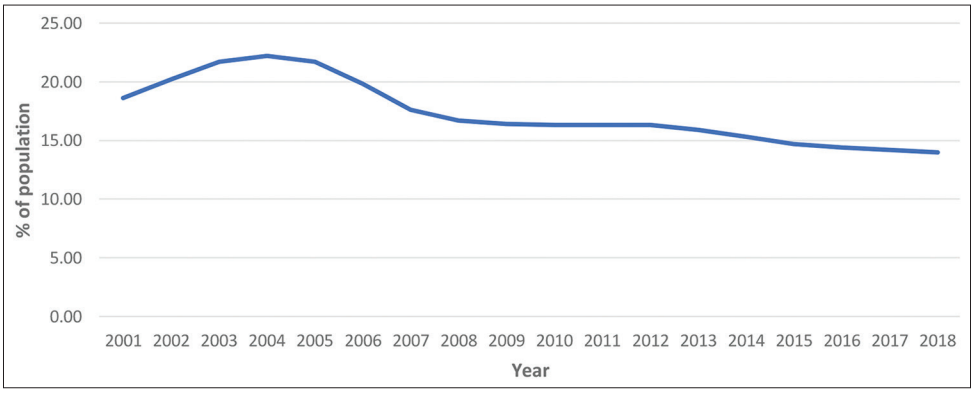

Figure 5: India: Prevalence of undernourishment (% of population), 2001–2018; data on undernourishment are from the Food and Agriculture Organization of the United Nations and measure food deprivation based on average food available for human consumption per person, the level of inequality in access to food, and the minimum calories required for an average person

Data Model Deployment

While the findings from the data modeling phase, the real issue is whether this pattern in the data model will be seen in the population of India when data from 2019 to 2020 are added. Public health decision-makers and public health planners probably would subnational and demographic data to better understand the patterns of undernutrition in various parts of India (e.g., Rajasthan) and in various subpopulation groups in India. For example, issues related to combating malnutrition could play in the management of diabetes (including diabetic retinopathy) at a population-level in India.[22] Furthermore, the child population of India along with the prevalence of child hunger in the country may call for further examination of the prevalence of undernutrition to guide public health policy and action with more recent data used in the data model.[23] These considerations may occupy this deployment phase.

Implications of Case Study for Public Health Professionals in India

This brief case study showed the basic steps of the data science used to address a public health issue. Please notice the messy data and how it was prepared for data modeling in the data preparation phase with the help of the KNIME Analytics Platform. This "visual programming" is not possible using traditional statistical software packages.

The public health epidemiologist can imagine using KNIME (or another similar software package) to connect concurrently to a specific table of a relational database, a CSV dataset, an Excel data file, and, perhaps, a proprietary dataset using Python on the Open KNIME Workbench with specific "visual programming" nodes: Specific columns of data from these various data sources linked to each other and used to construct a new dataset. This new dataset would be used in data modeling and a data model exported from KNIME into any format for use with unseen, real-world data to address a public health problem.

Summary

In this review, the main concepts of data science were described for use with public health and epidemiologic conditions facing a population. We described the domains of data science used to address the public health and epidemiologic conditions facing a population and showed the intersection where the needs of public health and domain expertise combine with statistics and data mining, along with programming skills, to define a new area of knowledge and practice: Public health data science. Because the field of public health data science will likely help public health decision-makers, public health planners, and public health epidemiologists identify needs and where public health action can help improve the public health status of nations, a case study was presented to illustrate the use of data science concepts in public health.

Acknowledgments

GVF grateful to KNIME Analytics Platform professional staff for helping him learn the use of this software package for public health.

Biographical Statement

Dr. GV Fant is a public health epidemiologist and visiting faculty member at JSPH 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. He earned two master's degree – in the health sciences/public health and public administration and, later, a graduate certificate in health sciences (specialization: Epidemiology). Dr. Fant completed a year-long Executive Management Fellowship in data science (emphasis: Healthcare, public health, and managerial epidemiology). 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 and as an International Practitioner of the Faculty of Public Health of the Royal Colleges of Physicians of the United Kingdom (IPFPH-UK) in 2017.

Dr. A Purohit is a global public health specialist and HIV-AIDS policy expert. He has a demonstrated history of success, working internationally with government administrations. Dr. Purohit is the founder and current President/CEO of the Jodhpur School of Public Health in Jodhpur, India. He is skilled in Global Health Issues, Healthcare, Business Development, and Global Management.

References

1. Vayena E, Dzenowagis J, Brownstein JS, Sheikh A. Policy implications of big data in the health sector. Bull World Health Organ 2018;96:66-8.
2. Wiens J, Shenoy ES. Machine learning for healthcare: On the verge of a major shift in healthcare epidemiology. Clin Infect Dis 2018;66:149-53.

3. Roth JA, Battegay M, Juchler F, Vogt JE, Widmer AF. Introduction to machine learning in digital healthcare epidemiology. Infect Control Hosp Epidemiol 2018;39:1457-62.
4. Bi Q, Goodman KE, Kaminsky J, Lessler J. What is machine learning? A primer for the epidemiologist. Am J Epidemiol 2019;188:2222-39.
5. Grus J. Data Science from Scratch. Boston, United States: O'Reilly Media Inc.; 2019.
6. Jain VK. Data Science and Analytics. New Delhi, India: Khanna Book Publishing; 2019.
7. Salcedo J, McCormick K. IBM SPSS Modeler Essentials: Effective Techniques for Building Powerful Data Mining and Predictive Analytics Solutions. Birmingham, United Kingdom: Packt Publishing; 2017.
8. Zinoviev D, Dvorak K. Data Science Essentials in Python. Raleigh, United States: Pragmatic Programmers; 2016.
9. Bakos G. KNIME Essentials. Birmingham, United Kingdom: Packt Publishing; 2013.
10. McCormick K, Salcedo J. SPSS Statistics for Data Analysis and Visualization. Indianapolis, United States: Wiley; 2017.
11. Flowers J, Johnson K. In: Regmi K, Gee I, editors. Public Health Intelligence. Switzerland: Springer International Publishing; 2016.
12. IBM Training. Introduction to Machine Learning Models Using IBM SPSS Modeler (V18.2) SPVC. Armonk, United States: IBM Corporation; 2019.
13. Panesar A. Machine Learning and AI for Healthcare. New York, United States: Apress; 2019.
14. Jain N. 7 Ways Public Health in India has Changed over the Last Decade; 2018. Available from: https://www.path.org/articles/7-ways-public-health-india-has-changed. [Last accessed on 2020 Oct 16].
15. Markle WH, Fisher M, Smego R. Understanding Global Health. New York, United States: McGraw Hill-Lange; 2007.
16. Merson MH, Black RH, Mills AJ. Global Health: Diseases, Programs, Systems, and Policies. 3rd ed. Burlington, United States: Jones & Bartlett Learning; 2012.
17. Bonita R, Beaglehole R, Kjellström T. Basic Epidemiology. 2nd ed. Geneva, Switzerland: World Health Organization; 2006.
18. Friis RH, Sellers TA. Epidemiology for Public Health Practice. 4th ed. Boston, United States: Jones & Bartlett Learning; 2009.
19. Fos PJ, Fine DJ, Zuniga MA. Managerial Epidemiology for Health Care Organizations. 3rd ed. Hoboken, United States: Jossey-Bass; 2018.
20. World Bank. Available from: https://www.databank.worldbank.org/home.aspx.
21. KNIME Analytics Platform. Available from: https://www.knime.com.
22. Paswan SK, Verma P, Raj A, Azmi L, Shrivastava S, Rao CV. Role of nutrition in the management of diabetes mellitus. Asian Pac J Health Sci 2016;3:1-6.
23. Mitra R. Action on Ground Needed to Tackle Malnutrition: Experts on India's Poor Showing at Global Hunger Index. Chennai: The New Indian Express; 2020.