REVIEW

A systematic review of the quality and timeliness of public health data

Wilfred Bonney*

Viskas Informatics®, Halifax, Nova Scotia, Canada

Received: March 28, 2023 **Accepted:** May 14, 2023 **Online Published:** May 21, 2023
DOI: 10.5430/jha.v12n1p16 **URL:** https://doi.org/10.5430/jha.v12n1p16

ABSTRACT

The quality and timeliness of public health data is a topic of prime concern in this information age. Many epidemiologists, health scientists and researchers in the public health domain have consistently emphasized on the importance of the need for the right timely data for the right decision-making at the right time. In other words, there is an urgent need to ensure that the right data reaches the right people at the right time. However, this urgent need appears to be misleading and not achievable in the current public health practices and workflow processes. The workflow processes in the current healthcare environments enable data collection to be delayed and only to be captured when the events have already occurred. In this paper, a systematic review of relevant scientific literature was used to not only explore the complexity and uniqueness of public health data, but also explain why improving the quality and timeliness of public health data is a challenging endeavor for many epidemiologists, health scientists and researchers. Recommendations for streamlining the public health workflow processes to support the generation of high-quality and timely public health data were also discussed in the paper.

Key Words: Public health, Public health data, Data quality, Public health informatics, Public health research

1. INTRODUCTION

The coronavirus disease 2019 (COVID-19) pandemic caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and the declaration of Monkeypox as a global public health emergency on July 23, 2022 by the World Health Organization have highlighted the importance of accessibility to quality and timely public health data.[1–8] These public health emergencies have not only exposed weaknesses and data gaps in public health surveillance and reporting systems,[1,4,9] but also led to a situation whereby many public health agencies are scrambling for real-time data to justify mandated actions and thus, placing major strains on public health information systems and infrastructures. More importantly, many federal, state, tribal, local, territorial, and international public health agencies are systematically collecting, generating, aggregating, interpreting, and reporting data from diverse data sources to justify purported actions that are often susceptible to errors and inconsistencies. These data are reported in a manner that are not timely and believable.

Furthermore, the heterogeneous data sources of public health data bring abundant data types and complex data structures, thereby, increasing the difficulty of data integration and semantic interoperability.[9–12] It is in this regard that Acosta et al.[13] asserted that "precision, granularity, and timeliness of public health data continue to challenge the ability to make responsive decisions" (p.12). This perspective is shared by Kadakia et al.,[14] who argued that barriers to timely data

*Correspondence: Wilfred Bonney, Ph.D.; ORCiD ID: https://orcid.org/0000-0002-7946-4419; Email: wbonney@dal.ca; Address: Viskas Informatics®, Halifax, Nova Scotia, Canada.

collection and health information exchange hindered many health departments throughout COVID-19 (p.385). Similarly, de Bienassis et al.[2] found that many countries, at the onset of the COVID-19 pandemic, lacked rudimentary and timely data for effective public health decision-making. Good-quality public health data drive public health decision-making, priority setting, strategic deployment of resources, and evaluation of public health actions.[15,16]

The objective of this paper was to use a systematic review of relevant scientific literature to not only explore the complexity and uniqueness of public health data, but also explain why improving the quality and timeliness of public health data is a challenging endeavor for many epidemiologists, health scientists and researchers. The first part of the paper gives an overview of the complexity and uniqueness of public health data. In the second part, the focus is on improving the quality and timeliness of public health data. The third part focuses on streamlining the public health workflow processes to support the generation of high-quality and timely public health data.

2. METHODS

A systematic review of published research work from 2007 to 2022 was used to not only explore the complexity and uniqueness of public health data, but also explain why improving the quality and timeliness of public health data is a challenging endeavor for many epidemiologists, health scientists and researchers. The methodology followed the PRISMA (Preferred Reporting Items for Systematic reviews and Meta-Analysis) guidelines[17,18] and involved a systematic methods of collating and synthesizing findings from peered reviewed publications, found and accessed with the help of EBSCOhost, ProQuest and PubMed databases. Additional publications were retrieved using citation searching and data sources such as ACM digital libraries, Google Scholar, IEEE Xplore, and ScienceDirect. The search strategy consisted of a series of targeted search terms with different combination of keywords and/or phrases including: (a) public health; (b) public health data; (c) health data; (d) data quality; (e) quality of public health data; (f) public health AND data quality; (g) quality of health data; (h) data quality of public health; (i) case definitions; (j) timeliness of health data; (k) timeliness of public health data; (l) health data AND timeliness; (m) public health data AND timeliness; and (n) health data AND/OR public health data AND/OR data quality.

3. RESULTS

3.1 Systematic review

Seven hundred and seventy-nine abstracts and articles were screened, and forty-five of them met the inclusion criteria and were reviewed in full. Articles and/or studies were included in the systematic review if they reported not only on the complexity and uniqueness of public health data, but also discussed the quality and timeliness of public health data. The inclusion criteria also required that the retrieved articles were: (a) published in English language; (b) published in the date range between January 1, 2007 and December 31, 2022; and (c) electronically available in full-text.

The methodology excluded articles that were not available electronically in full-text format. Retrieved electronic full-text articles not published in English language were also excluded from the systematic review. Findings from the reviewed articles were quoted, paraphrased, synthesized, and grouped under four broad themes: Defining Public Health Data; Complexity and Uniqueness of Public Health Data; Improving the Quality of Public Health Data; and Improving the Timeliness of Public Health Data. The PRISMA 2020 flow diagram[17,18] for the systematic review is shown in Figure 1.

3.2 Defining public health data

Data play a significant role in public health practice because of the heterogeneous and voluminous data being exchanged across the continuum of different healthcare providers in the healthcare environments. More specifically, public health is inherently a data-driven and data-intensive domain.[12,19–22] The current and emerging public health emergencies require timely collection of high-quality public health data for designing appropriate public health interventions to improve population health outcomes.[20]

The definition of public health data varies among researchers and requires basic understanding of the meaning of the two terms: Public Health and Data. While the term Public Health is defined as a "branch of medicine concerned with the prevention and control of disease and disability, and the promotion of physical and mental health of the population on the international, national, state, or municipal level",[23] the term Data is defined as the "representations of facts, concepts, or instructions in a manner suitable for communication, interpretation, or processing by humans or by automated means"[24] (p.21). Drawing upon those two definitions, public health data can simply be defined as the representations of facts, concepts and/or instructions related to disease prevention and health promotion.

The facts and concepts represented in public health are mostly clinical and administrative data, derived from heterogeneous data sources such as Electronic Medical Records, Electronic Health Records, Personal Health Records, Census Data, Genomic Data, Hospital Utilization, Insurance

Claims and Social Media Data.[21] These fragmented and heterogeneous data sources of public health data are complex and difficult to model in statistical environments to inform decision-making in a timely manner. Improving the quality and timeliness of public health data is, therefore, a necessity that cannot be ignored.

It is the duty and best interest of public health agencies to collect, transform, analyze, and report public health data in a timely manner to not only support any mandated actions and recommendations, but also enable analyses of health priorities and trends. According to Lee and Gostin,[25] public health agencies collect, store, and use public health data to identify and control public health threats and improve public health services (p.82). Even though advancement in artificial intelligence promises to provide real-time decision-making, it is still in its infancy and not viable when dealing with public health data that are not readily available and subject to workflow processes that are hard to ascertain and predict.

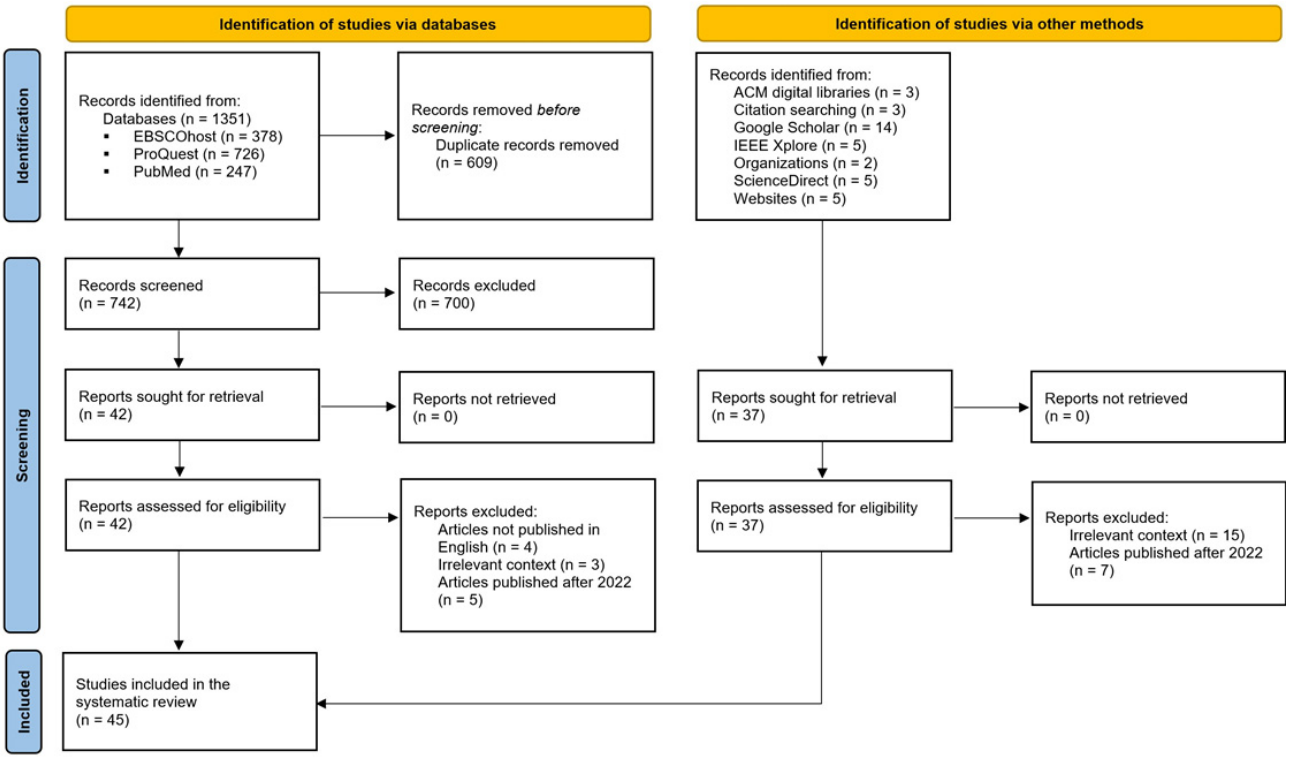

Figure 1. The PRISMA 2020 flow diagram for the systematic review

In their systematic review work on the barriers to data sharing in public health, van Panhuis et al.[22] defined public health data as "data that were primarily collected by public health agencies for routine purposes such as disease surveillance or program monitoring without primary intention of research" (p.2). Specifically, van Panhuis et al.[22] characterized public health data as routinely collected data that are not research-ready. Acknowledging that public health data "could be quantitative, qualitative, imaging, or genomic output" (p.2), the U.S. Centers for Disease Control and Prevention (CDC)[26] defined public health data as "digitally recorded factual material commonly accepted in the scientific community as a basis for public health findings, conclusions, and implementation" (p.1). These definitions and properties of public health data epitomize their complexity and uniqueness in public health practice.

3.3 Complexity and uniqueness of public health data

The single source of truth for public health data is often debated among many epidemiologists, health scientists and researchers. In her article about, "The making of public health data: paradigms, politics, and policy", Nancy Krieger[27] postulated that public health data "do not simply exist: the variables included or excluded from any given data set reflect the choices of individuals and institutions with the power to make these decisions" (p.412). This assertion by Krieger[27] is still valid today. Unfortunately, not much have changed in the public health domain since Krieger's article was published over 30 years ago. Public health data continue to be held up by institutions and public health agencies with power to analyze and interpret those datasets. As a case in point, there is still no available public health data to confirm and/or substantiate the origin of the COVID-19 pandemic.

Likewise, Sundararaman and Ramanathan[21] asserted that public health data "are generated from public health practice, with data sources being population-based and institution-based" (p.65). Whereas the population-based data "are collected through census, civil registrations, and population surveys"; the institution-based data "are obtained from individual health records and administrative records of health institutions"[21](p.65). In reality, the majority of public health data resides in different independent or siloed public health information systems either in tangible or intangible form.[21] Hence, the accessibility of public health data from one provider to another at the right time remains a major challenge in public health practice.[11] Sundararaman and Ramanathan[21] consequently noted that high-quality data in individual siloed systems does not necessarily add value without longitudinal view of the entities in consideration.

At a minimum, public health data are complex and come with the *5Vs of Big Data*: Volume (i.e., tremendous volume of the data); Velocity (i.e., data being formed at an increasing speed); Variety (i.e., different kinds of data types and formats); Value (i.e., low-value density); and Veracity (i.e., provenance and trustworthiness of the data source).[6, 10, 28–30] Burgun et al.[31] also observed that public health data are mainly made of heterogeneous and low-throughput data. These low-throughput data sources are commonly referred to as data lakes.

A data lake is "a central storage repository that holds big data from many sources in a raw, granular format".[32] These raw and granular data could be structured, semi-structured, unstructured and data as-is. For the purpose of making any meaningful evidence-based and justifiable conclusion to support clinical and public health decision-making, the collected data stream must not have ambiguity in its collection, aggregation and analysis, but rather support certain predefined standardized coding that is interpretable to the data analyst and the public at large. Where there are missingness of data fields/columns, the challenge for the generalizability of the data becomes a huge bottleneck for researchers, thereby inviting further subjective interpretations of the data.[33–35] For example, data fields such as ethnicity and race are persistently unavailable in many public health surveillance systems and are inconsistently reported by many jurisdictions within the United States.[9, 14, 15]

In order to achieve perfect generalization or sound inferences from missing data fields, statisticians and epidemiologists tend to use missing value imputation to support their predictive models.[33–35] The question, however, is whether it is appropriate, or ethical, for researchers to use such methodology in public health research.[36] As Box's aphorism[37] correctly stated: "All models are wrong but some are useful" (p.202). This aphorism might explain why many infectious disease models and predictions have gone wrong and led to multiple debates on the reliability and trustworthy of public health data.

3.4 Improving the quality of public health data

The quality of public health data is far from perfect due to the *5Vs of Big Data*.[21] The *5Vs of Big Data* make it difficult to judge the quality of public health data within a reasonable amount of time.[10] More importantly, the quality and timeliness of public health data are critical for good decision-making on public health interventions, priorities, and strategic allocation of resources.[38] High-quality data are the prerequisite for guaranteeing their added value in public health research.[10, 31] High-quality public health data are also "the prerequisite for better information, better decision-making and better population health" outcomes[21] (p.65).

In a study conducted to capture the dimensions of data quality that are important to data consumers (i.e., those who use data), Wang and Strong[39] asserted that the requirement for improving data quality relies on the perception or viewpoint of data consumers. Different data consumers appear to have different interpretations of what data quality really means.[10, 39, 40] Perhaps, the most succinct definition of data quality that fits the workflows of public health researchers could be attributed to the work of Liaw et al.[42] Citing the work of Wang,[43] Liaw et al.[42] defined data quality in the context of fit-for-purpose or fit-for-use. Wang and Strong[39] also defined data quality as "data that are fit for use by data consumers" (p.7). These definitions characterize data quality as a desired quality state that can be assessed and improved over time.

A major issue with conducting public health research on routinely collected and/or generated public health data is that they are often not fit-for-purpose or research-ready.[22, 40] Public health data are only research-ready if they are free of defects and possess desired quality features.[40] In other words, public health data are only as useful as their desired quality.[40, 44] Therefore, the need to ensure that public health data, generated and/or extracted for research purposes, are of high quality is of great importance. Epidemiologists, health scientists and researchers depend heavily on high-quality public health data to support proactive public health actions, accurate public health disease forecasting, and analyses of health priorities and trends in a timely manner.[13]

The quality of public health data increases with decreasing quantity of data.[40] Population health outcomes will be imperfect and distorted if the underlying quality of the public health data is poor and ineffective.[40] Assessing and im-

proving the quality of public health data will guarantee their meaningfulness and usefulness in public health research and other secondary uses.[40] CDC[26] also emphasized on the urgent need to evaluate all phases of public health data for quality before making them accessible to public health researchers and the general public. Essentially, the quality of public health data gains competitive advantages in public health research.[40]

3.5 Improving the timeliness of public health data

Drawing upon the work of Wang and Strong,[39] the timeliness of public health data could be defined as the extent to which the age of the public health data is appropriate for the task at hand. In most cases, if the data analyst performing the tasks at hand is not aware of the public health context on which the data were collected, then there is a tendency for the time spent on pre-processing and interpreting the data to be longer than expected. More importantly, the intended uses of routinely collected public health data is not often the same as its secondary uses in research. This in turn affects the lifecycle of the data usefulness, thus causing the data pre-processing to iterate until a reliable and interpretable data are obtained.

Cai and Zhu[10] defined timeliness as the "time delay from data generation and acquisition to utilization" (p.5). If the time delay is very significant, then the usefulness of the interpreted data would be outdated and had no added value for use in public health emergency. Specifically, the timeliness of public health data "limits the availability of actionable public health information as the traditional route for the data moves from patient self-report to a physician, through diagnostic confirmations, and then from a physician or laboratory facility to a public health authority"[19] (p.8).

The different public health practices and workflow processes also contribute to the timely release, availability, and accessibility of public health data. The data wrangling process (i.e., importing, restructuring, cleaning, de-identifying, and organizing data for analysis) is often tedious, non-linear, and iterative.[45] Often, the data collection and data wrangling processes create significant time lags measured in years,[13,40] which inherently affects the timeliness and interpretability of the analyzed public health data. In a public health emergency, making key decisions with public health data with significant time lags can negatively impact public health.[13]

Many data analysts, epidemiologists, and statisticians prefer to import public health data into software applications (i.e., Python, SAS, STATA, and R statistical environment) before cleaning the data. Informaticians, on the other hand, prefer to standardize and clean the public health data first before importing them into analytical software applications for faster analyses. While the former approach is error-prone and time consuming, the latter approach reduces the time required to analyze and interpret the public health data. These two different workflow approaches significantly impact the timeliness of data analytics pre-processing and interpretations of the public health data.

4. Discussion

The standard operating protocol for handling and processing public health data, in an event of a pandemic, conflicts with the inherent need to make public health data available and accessible in a timely manner. In most public health settings, when a new infectious disease is identified, a new case definition is proposed that set the scene for which minimum core datasets should be collected for the identified infectious disease.[46] In the current public health practices, until the case definition is set up, there is no established way to initiate data collection of individuals who have been infected with the identified disease. These case definitions, unfortunately, differ widely among different disease domains.[46] However, certain demographic and social determinants of health data would still remain the same for the infected individuals.

Findings from the review of the scientific literature support the fact that public health data are complex and unique in public health practice. Improving the quality and timeliness of public health data remains a big challenge for public health researchers, and would ultimately require implementing the following recommendations:

- Move towards the development of a generic case definition that will support proactive data collection for all emerging disease pathogens.
- Streamline the public health workflow processes to support the generation of high-quality and timely public health data.
- Develop integrated public health surveillance and reporting systems that seek to proactively collect data only on patients' symptoms versus patients' demographics and symptoms.
- Avoid the use of multiple siloed systems to collect and store public health data.
- Develop enabling public health information systems that are not constraining and overwhelming to the workflow processes of epidemiologists, health scientists and public health researchers.
- Develop public health surveillance systems with built-in logic of populations' demographic data to enable researchers in the medical record/data linkage domain to be able to rapidly link datasets during pandemic time to evaluate and analyze data for decision makers in a timely manner.
- Enhance computer processing technology to support viable data sources of public health data.[10]

• Utilize data science and informatics knowledge to support and improve how public health captures more timely relevant and actionable data.[13,45,47]

This systematic review is limited to the number of electronic articles retrieved and reviewed in full for the paper in the published date range. It is possible that including paper-based and electronic articles that were not published in English language would have enhanced the themes and quality of the paper outcome. However, extensive review of those foreign articles would have been difficult to synthesize with language limitations and comprehensions. It is, therefore, possible that the paper is biased towards those electronic articles written in the English language, and may have missed valuable insights and perspectives from articles that were not available electronically.

5. Conclusions

This paper has demonstrated the use of a systematic review of scientific literature to describe the complexity and uniqueness of public health data. It has also explained the reason why improving the quality and timeliness of public health data to deliver right timely data to decision makers at the right time is a challenging endeavor for many epidemiologists, health scientists and public health researchers. Recommendations for streamlining the public health workflow processes to support the generation of high-quality and timely public health data were also discussed in the paper.

The generation and preservation of high-quality and timely public health data are the prerequisite for conducting high-quality public health research. The current public health workflow processes do not support proactive data collection, as critical data continue to lag in public health surveillance and reporting systems. There is, however, a general belief that improving the quality and timeliness of public health data together with increased sharing, linking and reuse of public health data would improve efficiencies in healthcare delivery, generate new knowledge and evidence-based public health practice, and ultimately drive scientific innovation in both the private and public health sector.[22,26,31,48,49]

Conflicts of Interest Disclosure

The author declares no conflicts of interest.

References

[1] Axelrath S. Challenges Encountered in the Public Health Data Collection of COVID-19 Cases Among People Experiencing Homelessness. JAMA Netw Open. 2022; 5(8): e2229703. PMid: 35980641. https://doi.org/10.1001/jamanetworkopen.2022.29703

[2] de Bienassis K, Fujisawa R, Hashiguchi TCO, et al. Health data and governance developments in relation to COVID-19: How OECD countries are adjusting health data systems for the new normal. 2022 [cited 2022 Oct 13]. Available from: https://www.oecd-ilibrary.org/social-issues-migration-health/health-data-and-governance-developments-in-relation-to-covid-19_aec7c409-en

[3] Dyda A, Purcell M, Curtis S, et al. Differential privacy for public health data: An innovative tool to optimize information sharing while protecting data confidentiality. Patterns. 2021; 2(12): 100366. PMid: 34909703. https://doi.org/10.1016/j.patter.2021.100366

[4] Huyser KR, Horse AJY, Kuhlemeier AA, Huyser MR. COVID-19 Pandemic and Indigenous Representation in Public Health Data. Am J Public Health. 2021; 111(S3): S208-14. PMid: 34709868. https://doi.org/10.2105/AJPH.2021.306415

[5] Nuzzo JB, Borio LL, Gostin LO. The WHO Declaration of Monkeypox as a Global Public Health Emergency. JAMA. 2022; 328(7): 615. PMid: 35895041. https://doi.org/10.1001/jama.2022.12513

[6] Peddireddy AS, Xie D, Patil P, et al. From 5Vs to 6Cs: Operationalizing Epidemic Data Management with COVID-19 Surveillance [Internet]. In: 2020 IEEE International Conference on Big Data (Big Data). Atlanta, GA, USA: IEEE; 2020 [cited 2022 Oct 13]. 1380-7 p. https://doi.org/10.1109/BigData50022.2020.9378435

[7] Soualmia LF, Hollis KF, Mougin F, et al. Health Data, Information, and Knowledge Sharing for Addressing the COVID-19. Yearb Med Inform. 2021; 30(01): 004-7. PMid: 34479377. https://doi.org/10.1055/s-0041-1726541

[8] World Health Organization. Naming the coronavirus disease (COVID-19) and the virus that causes it. 2020 [cited 2022 Sep 22]. Available from: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/technical-guidance/naming-the-coronavirus-disease-(covid-2019)-and-the-virus-that-causes-it

[9] Martin LT, Nelson C, Yeung D, et al. The Issues of Interoperability and Data Connectedness for Public Health. Big Data. 2022; 10(S1): S19-24. https://doi.org/10.1089/big.2022.0207 PMid:36070509 PMCid:PMC9508439

[10] Cai L, Zhu Y. The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. CODATA. 2015; 14(2): 1-10. https://doi.org/10.5334/dsj-2015-002

[11] Chelladurai U, Pandian S. A novel blockchain based electronic health record automation system for healthcare. J Ambient Intell Human Comput. 2022; 13(1): 693-703. https://doi.org/10.1007/s12652-021-03163-3

[12] Shapiro JS, Mostashari F, Hripcsak G, et al. Using Health Information Exchange to Improve Public Health. Am J Public Health. 2011; 101(4): 616-23. PMid: 21330598. https://doi.org/10.2105/AJPH.2008.158980

[13] Acosta JD, Chandra A, Yeung D, et al. What Data Should Be Included in a Modern Public Health Data System. Big Data. 2022; 10(S1): S9-14. PMid: 36070507. https://doi.org/10.1089/big.2022.0205

[14] Kadakia KT, Howell MD, DeSalvo KB. Modernizing Public Health Data Systems: Lessons from the Health Information Technology for

Economic and Clinical Health (HITECH) Act. JAMA. 2021; 326(5): 385. PMid: 34342612. https://doi.org/10.1001/jama.2021.12000

[15] Bauer UE, Plescia M. Addressing Disparities in the Health of American Indian and Alaska Native People: The Importance of Improved Public Health Data. Am J Public Health. 2014; 104(S3): S255-7. PMid: 24754654. https://doi.org/10.2105/AJPH.2013.301602

[16] Choi BCK. The Past, Present, and Future of Public Health Surveillance. Scientifica. 2012; 2012: 1-26. PMid: 24278752. https://doi.org/10.6064/2012/875253

[17] Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021; n71. PMid: 33782057. https://doi.org/10.1136/bmj.n71

[18] Page MJ, Moher D, Bossuyt PM, et al. PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. BMJ. 2021; n160. PMid: 33781993. https://doi.org/10.1136/bmj.n160

[19] Kass-Hout TA, Alhinnawi H. Social media in public health. Br Med Bull. 2013; 108(1): 5-24. PMid: 24103335. https://doi.org/10.1093/bmb/ldt028

[20] Papagari Sangareddy SR, Aspevig J. New Means of Data Collection and Accessibility. In: Magnuson JA, Dixon BE, editors. Public Health Informatics and Information Systems. Cham: Springer International Publishing; 2020 [cited 2022 Oct 13]. 289-305 p. Available from: https://doi.org/10.1007/978-3-030-41215-9_17

[21] Sundararaman A, Ramanathan SV. Open Research Issues and Emerging Research Directions in Data Quality for Public Health. In: 21st International Conference on Information Quality (ICIQ 2016). Ciudad Real, Spain; 2016. 65-74 p.

[22] van Panhuis WG, Paul P, Emerson C, et al. A systematic review of barriers to data sharing in public health. BMC Public Health. 2014; 14(1): 1144. PMid: 25377061. https://doi.org/10.1186/1471-2458-14-1144

[23] MeSH Browser. Public Health: MeSH Descriptor Data 2022. 2022 [cited 2022 Sep 22]. Available from: https://meshb.nlm.nih.gov/record/ui?ui=D011634

[24] Raymond SA, Gawrylewski HM, Ganter J, et al. CDISC clinical research glossary. Appl Clin Trials. 2007; 16: 12-52.

[25] Lee LM, Gostin LO. Ethical Collection, Storage, and Use of Public Health Data: A Proposal for a National Privacy Protection. JAMA. 2009; 302(1): 82-4. PMid: 19567443. https://doi.org/10.1001/jama.2009.958

[26] U.S. Centers for Disease Control and Prevention (CDC). Policy on Public Health Research and Nonresearch Data Management and Access. 2016 [cited 2022 Sep 30]. Available from: https://www.cdc.gov/maso/policy/policy385.pdf

[27] Krieger N. The Making of Public Health Data: Paradigms, Politics, and Policy. J Public Health Policy. 1992; 13(4): 412. PMid: 1287038. https://doi.org/10.2307/3342531

[28] Ishwarappa, Anuradha J. A Brief Introduction on Big Data 5Vs Characteristics and Hadoop Technology. Procedia Comput Sci. 2015; 48: 319-24. https://doi.org/10.1016/j.procs.2015.04.188

[29] Katal A, Wazid M, Goudar RH. Big data: Issues, challenges, tools and Good practices. In: 2013 Sixth International Conference on Contemporary Computing (IC3). Noida, India: IEEE; 2013 [cited 2022 Oct 13]. 404-9 p. Available from: https://doi.org/10.1109/IC3.2013.6612229

[30] Millham R, Agbehadji IE, Frimpong SO. The Paradigm of Fog Computing with Bio-inspired Search Methods and the "5Vs" of Big Data. In: Fong SJ, Millham RC, editors. Bio-inspired Algorithms for Data Streaming and Visualization, Big Data Management, and Fog Computing. Singapore: Springer Singapore; 2021 [cited 2022 Oct 13]. 145-67 p. https://doi.org/10.1007/978-981-15-6695-0_8

[31] Burgun A, Bernal-Delgado E, Kuchinke W, et al. Health Data for Public Health: Towards New Ways of Combining Data Sources to Support Research Efforts in Europe. Yearb Med Inform. 2017; 26(01): 235-40. PMid: 29063571. https://doi.org/10.15265/IY-2017-034

[32] Talend. What is a Data Lake? -[cited 2022 Sep 22]. Available from: https://www.talend.com/resources/what-is-data-lake/

[33] Bertsimas D, Pawlowski C, Zhuo YD. From predictive methods to missing data imputation: an optimization approach. J Mach Learn Res. 2017; 18(1): 7133-71.

[34] Lin WC, Tsai CF. Missing value imputation: a review and analysis of the literature (2006-2017). Artif Intell Rev. 2020; 53(2): 1487-509. https://doi.org/10.1007/s10462-019-09709-4

[35] Madley-Dowd P, Hughes R, Tilling K, et al. The proportion of missing data should not be used to guide decisions on multiple imputation. J Clin Epidemiol. 2019; 110: 63-73. PMid: 30878639. https://doi.org/10.1016/j.jclinepi.2019.02.016

[36] Bonney W. Is it appropriate, or ethical, to use health data collected for the purpose of direct patient care to develop computerized predictive decision support tools? Stud Health Technol Inform. 2009; 143: 115-21. PMid: 19380924. https://doi.org/10.3233/978-1-58603-979-0-115

[37] Box GEP. Robustness in the Strategy of Scientific Model Building. In: Launer RL, Wilkinson GN, editors. Robustness in Statistics. Elsevier; 1979 [cited 2023 Apr 8]. 201-36 p. https://doi.org/10.1016/B978-0-12-438150-6.50018-2

[38] World Health Organization. 2016 Annual Report Communicable Diseases Cluster. 2017 [cited 2022 Sep 22]. Available from: https://apps.who.int/iris/bitstream/handle/10665/259634/9789290233930-eng.pdf?sequence=1&isAllowed=y

[39] Wang RY, Strong DM. Beyond Accuracy: What Data Quality Means to Data Consumers. J Manag Info Syst. 1996; 12(4): 5-33. https://doi.org/10.1080/07421222.1996.11518099

[40] Bonney W, Scobbie D, Nind T, et al. Profiling Clinical Datasets for Data Quality Assessment and Improvement [Internet]. In: BCS Health Informatics Scotland 2014 Conference. 2014 [cited 2022 Oct 13]. 1-8 p. Available from: https://scienceopen.com/document?vid=89e51c55-2e78-480b-93ca-4ccca7e91099

[41] Karr AF, Sanil AP, Banks DL. Data quality: A statistical perspective. Stat Methodol. 2006; 3(2): 137-73. https://doi.org/10.1016/j.stamet.2005.08.005

[42] Liaw ST, Rahimi A, Ray P, et al. Towards an ontology for data quality in integrated chronic disease management: A realist review of the literature. Int J Med Inform. 2013; 82(1): 10-24. PMid: 23122633. https://doi.org/10.1016/j.ijmedinf.2012.10.001

[43] Wang RY. A product perspective on total data quality management. Commun ACM. 1998; 41(2): 58-65. https://doi.org/10.1145/269012.269022

[44] Redman TC. Data quality: the field guide. Boston, MA: Digital Press; 2001.

[45] Goldsmith J, Sun Y, Fried LP, et al. The Emergence and Future of Public Health Data Science. Public Health Rev. 2021; 42: 1604023. PMid: 34692178. https://doi.org/10.3389/phrs.2021.1604023

[46] U.S. Centers for Disease Control and Prevention (CDC). Case definitions for infectious conditions under public health surveillance. 1997 [cited 2022 Sep 22]. Available from: https://wonder.cdc.gov/wonder/Prevguid/m0047449/m0047449.asp

[47] Asaro PV, Land GH, Hales JW. Making Public Health Data Available to Community-Level Decision Makers-Goals, Issues, and a Case Report: J Public Health Manag Pract. 2001; 7(5): 58-63. PMid: 11680032. https://doi.org/10.1097/00124784-200107050-00009

[48] Cheung S. Disambiguating the benefits and risks from public health data in the digital economy. Big Data Soc. 2020; 7(1): 1-15. https://doi.org/10.1177/2053951720933924

[49] Kostkova P. Disease surveillance data sharing for public health: the next ethical frontiers. Life Sci Soc Policy. 2018; 14(1): 16. PMid: 29971516. https://doi.org/10.1186/s40504-018-0078-x