International Journal of Community Medicine and Public Health
Owhonda G et al. Int J Community Med Public Health. 2022 Jun;9(6):2437-2441
http://www.ijcmph.com
pISSN 2394-6032 | eISSN 2394-6040

DOI: https://dx.doi.org/10.18203/2394-6040.ijcmph20221517

Original Research Article

The data analysis and validation engine: an application of artificial intelligence in the improvement of COVID-19 data management

Golden Owhonda[1], Anwuri Luke[2]*, Japheth Russell Inyele[1], Chidinma Eze-Emiri[1]

[1]Department of Public Health & Disease Control, Rivers State Ministry of Health, Port Harcourt, Rivers State, Nigeria
[2]Department of Community Medicine, College of Medicine, Rivers State University, Nkpolu-Oroworukwo, Port Harcourt, Rivers State, Nigeria

Received: 22 April 2022
Accepted: 10 May 2022

***Correspondence:**
Dr. Anwuri Luke,
E-mail: ndimekz2010@gmail.com

ABSTRACT

Background: The proper management of healthcare data is fundamental to the health system processes; artificial intelligence has proven its value in these processes. Artificial intelligence can simplify the management of information, improve data security, and automate data flow. It is also useful in the analysis and interpretation of big data. Hence, it has the possibility of screening and diagnosing diseases, categorizing disease severity, detecting therapeutic agents, and forecasting outbreak spots.
Methods: A data analysis and validation engine was developed to perform data quality control checks, classify addresses, and generate epidemiology numbers using the index and parse command on the command-line interface of DAVE.
Results: DAVE correctly formatted data and created a local copy of the datastore and the index. It also returned previous EPID numbers to each entry and assigned a new EPID number to missed entries. DAVE imported the entries into the data template of the existing data management tool and generated a sample manifest that is then sent to the Laboratory. The data flow from the point of collection to storage and reporting was assessed as 100% accurate without errors and in real-time; there was also the ability to roll back if any error occurred.
Conclusions: DAVE is a semi-autonomous system that operates with minimal human intervention; it is automatically faster as it leverages computing power to parse, store, and retrieve data while practically eliminating the need for manual data quality assessment. The DAVE functionality can be extended to incorporate additional features like forecasting outbreaks of emerging/re-emerging diseases, categorizing the severity of diseases and analysis of data in our setting.

Keywords: Data analysis and validation engine, Artificial intelligence, COVID-19, Data management

INTRODUCTION

Over the years, surveillance in the field of epidemiology has evolved from manual to digital methods through the use of traditional computer programming to artificial intelligence (AI) on improving the collection, collation, storage, interpretation, and presentation of data.[1,2] The practice of medicine is gradually changing with the advent of digitalized data acquisition, machine learning (ML), and computing infrastructure.[3,4] Though the AI-based tools are stilled being evaluated, these digital data management techniques have become the order of the day, as they are used to determine the etiopathogenesis, screening/ diagnosis, categorizing of disease severity, detecting therapeutic agents, vaccine production, and forecast of outbreak spots.[5-7] In recent times, AI-driven techniques have increasingly been used

in weather prediction and climate analysis to improve their capacity to predict, prevent, and detect the trends in outbreak prevalence, location, and future global health risks.[8,9] Additionally, AI and its applications are expanding into areas previously believed to be spearheaded by human experts, and proven to execute key healthcare tasks than human expertise.[10-12] Hence, utilization of AIs' will promote the health and well-being of individuals irrespective of their socioeconomic status as it works towards achieving the United Nations Sustainable Development Goals (SDG-3) by the year 2030.[10,11,13] To improve health programs in many resource-poor nations, the open data kit (ODK), open-source mobile software has been used by healthcare workers to rapidly collect, collate, and support the manipulation of intricate and large volumes of data types (text, location, images, audio, video, barcodes).[14-16] The ODK is the most vastly utilized, as it reinforces the function of many other electronic data collection (EDC) tools such as; Ona, Kobo Toolbox, Survey CTO, COMM Care, Enketo, and triangulates with Redcap, DHIS2, and R through application program interfaces (APIs).[14,17] ODK permits vast deployment and evaluation of the diagnostic test, therapeutics, and access to the compassionate use of safe and efficacious vaccines to curb the spread of this dangerous viral haemorrhagic disease.[14] Additionally, many data mining techniques have been applied by researchers with real-time datasets using numerous forms of processes, not limited to machine learning classifiers.[6,18,19]

The world health organization (WHO) and researchers are working diligently to develop more efficient and robust diagnostic techniques using advanced technologies to analyze data as well as; prevent, detect, and forecast outbreak spots. It may also be able to predict therapeutic drugs and vaccines in the circumstance.[20-24] For instance, CRISPR-Cas is a broadly used nucleic acid detection system designed using machine learning techniques to precisely target and cleave to the genome of SARS-CoV-2 for therapeutic purposes.[23-26] Moreover, data analyzed from these processes can enable collaborative preparedness in fighting epidemics.[13,27] In 2015, Nigeria adopted the surveillance outbreak response management and analysis system (SORMAS) reporting tool as an open-source mobile and web application software for the management of epidemic-prone diseases.[28-30] The SORMAS platform is a multifaceted, bidirectional information exchange data management software used for prevention, detection, contact tracing, and surveillance of disease outbreaks, particularly in resource-poor settings; and was adopted at the onset of the COVID-19 pandemic.[28,31] Although it has its strengths, there are several limitations in our setting such as Complexity in data entry, resulting in a lower throughput; the inability to classify addresses, therefore the collected data is consequentially problematic for automated systems that are meant to parse and classify them by their Local Government areas.

This study aimed at identifying and solving the challenges of our data management tools; thereby improving COVID-19 data management in Rivers State.

METHODS

A data analysis and validation engine (DAVE) was developed. This engine performed data quality control checks and implemented an artificially intelligent system that influences machine learning to classify addresses. The system contrives a Naive Bayes classifier which is a probabilistic classifier based on Bayes' theorem with strong (naive) independent assumptions, coupled with kernel density estimation to achieve accuracy levels that surpass a human classifier. The system was trained on a dataset of more than 250,000 pre-classified addresses leveraging the computing power of a cloud server to generate a model that can be utilized locally to classify addresses with minimal hardware requirements. DAVE is a suite of applications that automates data processing and its components. The command-line interface of DAVE exposes two commands, the index, and the parse command. The index command kept track of assigned EPID numbers for each Local Government Area (LGA) with version history and allowed for instantaneous rollbacks if there were issues. The parse command does the actual data processing and quality control checks. The checks involved removal of non-American Standard Code for Information Interchange (ASCII) characters; trimming of excess whitespace; validation and conversion of dates to the YYYY-MM-DD format.

RESULTS

DAVE processing mechanism

Entries are pulled from the data collection tool and passed to DAVE as a file with the mime-type "application/vnd. openxmlformatsofficedocument.spread sheetml.sheet" or "application/vnd.ms-excel". DAVE checks if the file structure matches the data template returning a descriptive error message if the check fails. Next, Dave checks individual fields removing characters that are outside the ASCII encoding table and fixing data formats. Once this step is completed DAVE pulls from the server and creates a local copy of the datastore and the index so that the system can be rolled back to previous states if any critical error occurs. At this point, DAVE begins assigning EPID numbers to each entry. To do this, DAVE iterates through each entry, compares the current entry to already existing data using the Levenshtein function to account for typographical errors, and returns the already existing EPID number when a match is found but assigns a new EPID number when there is none.

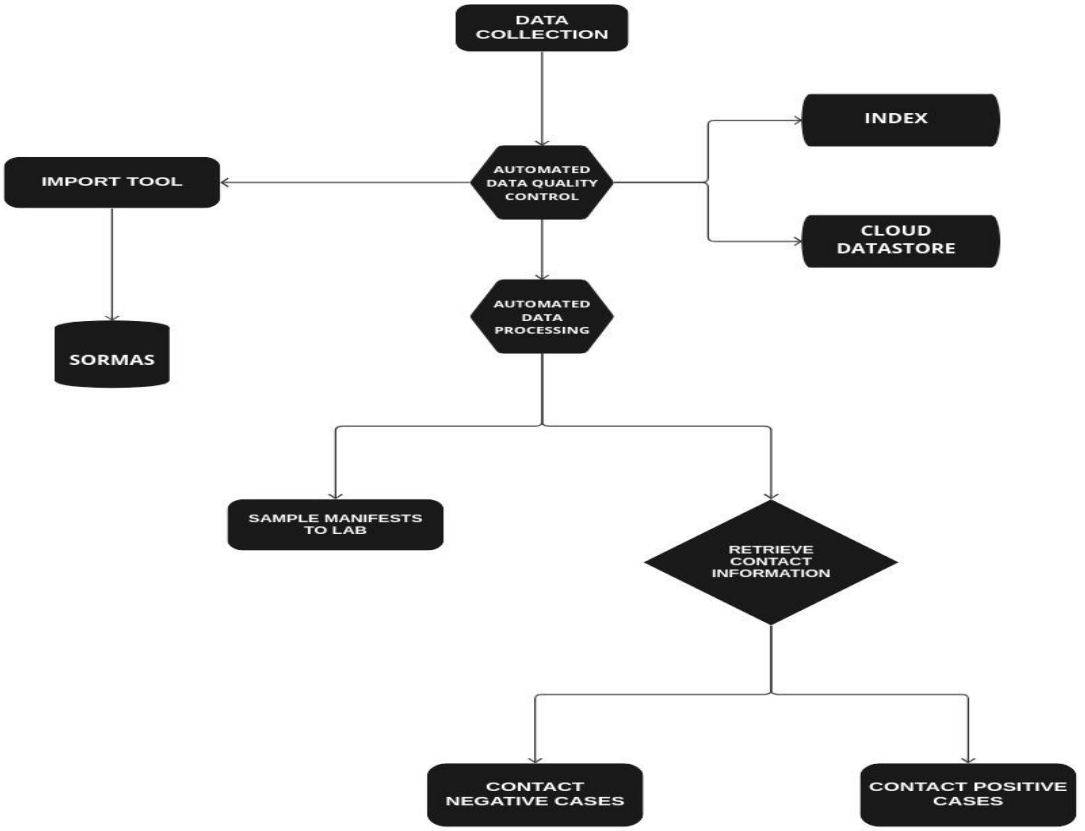

Figure 1: Data analysis and validation engine data processing system.

EPID number assignment

To assign EPID numbers, DAVE uses a pre-trained model based on the Naive Bayes classification algorithm to catalogue each address data entered into the LGAs. Once the classification is done, the system checks the index for the last entry for that LGA and increments it. This process is repeated until all entries have been assigned EPID numbers, then new data is uploaded back to the server and the local copies are updated. The entries with assigned EPID numbers get written to a manifest file in the output directory and sent to the laboratory. Dave then writes this data into files that can be imported into the data template of our management tools and generate a sample manifest that is then sent to the laboratory. Finally, the engine updates the local copy of the data and uploads it to the server.

The import component

The Import component of DAVE is an in-browser script that simulates input from a human interface device. It accounts for network conditions and handles dialogues and possible errors messages while it runs the upload. The import tool cuts the runtime to only a fraction of what would be required if the system was being operated manually.

The data retrieval component

To retrieve the contact information of individuals from the datastore, DAVE takes a list of names that is extracted from the result sheet gotten back from the Laboratory, runs the same data quality checks and applies the same fixes as it does with entries from the data collection tool. The data retrieval of DAVE used the levenshtein distance to compare entries, identify matches and pull their contact information. This data is then written to a file with the mime-type "text/csv" to avoid duplication while assigning EPID numbers.

DISCUSSION

The DAVE system is a semi-autonomous system that operates with minimal human intervention, requiring a reduction in the personnel management needed to operate a fully functional data unit. Dave is automatically faster necessitating a reduction in data processing time as it

leverages computing power to parse, store, and retrieve data while practically eliminating the need for manual data quality assessment. It also can process roughly about 10,000 entries in five minutes and upload about 1500 entries in an average of 15-minutes compared to approximately five entries with existing management tools. The proof of concept which involves data flow from the point of collection to storage and reporting is assessed as 100% accurate without errors and in real-time; there is also the ability to roll back if any error occurs. To automatically parse and classify data, regular expressions (RegEx) are used; the approach involves compiling a database of popular addresses and comparing entries against them for similarities, updating frequently and storing multiple references to the same physical location. A weakness of this approach includes a severe runtime overhead that grows exponentially with the increase in the number of entries that need to be parsed. The DAVE system circumvents these weaknesses. Although the DAVE system is efficient it requires high-end computational infrastructure to train the model. This serves as a limitation to its implementation. Also, the quality of the output to an extent is still dependent on the quality of data entered.[33,34]

CONCLUSION

The DAVE was effective in the implementation of AI in COVID-19 data management in Rivers State. Artificial intelligence and computer science at large are useful when properly applied in the practice of medicine and not only restricted to addressing the challenges encountered in COVID-19 data management. Although the development of the data analysis and validation engine (DAVE) application is complex, it only requires basic computer knowledge to operate which makes it an effective tool, by improving data quality, simplifying the retrieval of data, and eliminating the common bottlenecks associated with the existing data management systems in developing countries.

Recommendations

The data analysis and validation engine (DAVE) functionality can be extended to incorporate additional features like forecasting outbreaks of emerging/re-emerging diseases, categorizing the severity of diseases, predicting diagnostic measures for diseases, identifying possible therapeutic agents for diseases, and assessing the effectiveness of potential vaccines.

ACKNOWLEDGEMENTS

Authors would like to thank all volunteers/Adhoc staff who collected the data for the study.

Funding: No funding sources
Conflict of interest: None declared
Ethical approval: The study was approved by the Institutional Ethics Committee

REFERENCES

1. Chen J, See KC. Artificial intelligence for COVID-19: Rapid Review. J Med Internet Res. 2020;22(10):12.
2. Horgan D, Hackett J, Westphalen CB, Kalra D, Richer E, Romao M, et al. Digitalization and COVID-19: The Perfect Storm. Biomed Hub. 2020;5(3):1-23.
3. Nguyen T, Larrivée N, Lee A, Bilaniuk O, Durand R. Use of artificial intelligence in dentistry: current clinical trends and research advances. J Can Dent Assoc. 2021;87(17): 8.
4. Ahmed Z, Mohamed K, Zeeshan S, Dong X. Artificial Intelligence with Multi-Functional Machine Learning Platform Development for Better Healthcare and Precision Medicine. Database J Biol Databases Curation. 2020;2020(10):35.
5. Randhawa GS, Soltysiak MPM, El Roz H, de Souza CPE, Hill KA, et al. Machine Learning Using Intrinsic Genomic Signatures for Rapid Classification of Novel Pathogens: COVID-19 Case Study. PLoS ONE. 2020;15(4):24.
6. Albahri AS, Hamid RA, Alwan JK, Al-qays ZT, Zaidan AA, Zaidan BB, et al. Role of biological Data Mining and Machine Learning Techniques in Detecting and Diagnosing the Novel Coronavirus (COVID-19): A Systematic Review. J Med Syst. 2020; 44(122):11.
7. Srinivasa Rao ASR, Vazquez JA. Identification of COVID-19 can be Quicker Through Artificial Intelligence Framework Using a Mobile Phone-Based Survey When Cities and Towns are Under Quarantine. Infect Control Hosp Epidemiol. 2020;41(7):826-30.
8. Bochenek B, Ustrnul Z. Machine Learning in weather prediction and climate analyses: applications and perspectives. Atmosphere J. 2022;13(180):16.
9. Pley C, Evans M, Lowe R, Montgomery H, Yacoub S. Digital and Technological Innovation in Vector-Borne Disease Surveillance to Predict, Detect, and Control Climate-Driven Outbreaks. Lancet Planet Health. 2021;5(10):739-45.
10. Owoyemi A, Owoyemi J, Osiyemi A, Boyd A. Artificial Intelligence for Healthcare in Africa. Front Digit Health. 2020;2(6):5.
11. Morgenstern JD, Rosella LC, Daley MJ, Goel V, Schünemann HJ, Piggott T. "AI's Gonna Have an Impact on Everything in Society, So It Has to Have an Impact on Public Health": A Fundamental Qualitative Descriptive Study of the Implications of Artificial Intelligence for Public Health. BMC Public Health. 2021;21(40):14.
12. Davenport T, Kalakota R. The potential for Artificial Intelligence in Healthcare. Future Health J. 2019; 6(2):94.
13. Vinuesa R, Azizpour H, Leite I, Balaam M, Dignum V, Domisch S, et al. The Role of Artificial Intelligence in Achieving the Sustainable

Development Goals. Nat Commun. 2020;11(233):10.
14. Marks M, Lal S, Brindle H, Gsell P-S, MacGregor M, Stott C, et al. Electronic Data Management for Vaccine Trials in Low Resource Settings: Upgrades, Scalability, and Impact of ODK. Front Public Health. 2021;9(665584):11.
15. Maduka O, Akpan G, Maleghemi S. Using Android and Open Data Kit Technology in Data Management for Research in Resource-Limited Settings in the Niger Delta Region of Nigeria: Cross-Sectional Household Survey. JMIR. 2017;5(11):8.
16. Bokonda PL, Ouazzani-Touhami, K, Souissi N. Open Data Kit: Mobile Data Collection Framework for Developing Countries. Int J Innov Technol Explor Eng. 2019;8(12):4749-54.
17. Nampa IW, Mudita IW, Riwu Kaho NPLB, Widinugraheni S, Lasarus Natonis R. The KoBoCollect for Research Data Collection and Management: An Experience in Researching the Socio-Economic Impact of Blood Disease in Banana. SOCA J Sos Ekon Pertan. 2020;14(3):545-56.
18. Abu-Dalbouh HM, Alateyah SA. Predictive Data Mining Rule-Based Classifiers Model for Novel Coronavirus (COVID-19) Infected Patients' Recovery in the Kingdom of Saudi Arabia. J Theor Appl Inf Technol. 2021;99(8):19.
19. AlMoammar A, AlHenaki L, Kurdi H. Selecting Accurate Classifier Models for a MERS-CoV Dataset. AISC. 2019;2018(868):1070-84.
20. National Centre for Disease Prevention and Control. National Interim Guidelines for Clinical Management of COVID-19. Abuja, Nigeria: National Centre for Disease Prevention and Control; 2020:44. Available at: https://covid19.ncdc.gov.ng. Accessed on 25 February 2022.
21. Managing Epidemics: Key Facts about Major Deadly Diseases. Geneva, Switzerland; 2018:257. Available at: https://apps.who.int/iris/handle/10665/272442. Accessed on 25 February 2022.
22. Biden JR. National Strategy for the COVID-19 Response and Pandemic Preparedness. Washington DC, United States of America; 2021:200. Available at: https://www.whitehouse.gov. Accessed on 17 March 2022.
23. Watters KE, Kirkpatrick J, Palmer MJ, Koblentz GD. The CRISPR Revolution and its Potential Impact on Global Health Security. Pathog Glob Health. 2019;115(2):80-92.
24. Malik YS, Sircar S, Bhat S, Ansari MI, Pande T, Kumar P, et al. How Artificial Intelligence May Help the Covid-19 pandemic: Pitfalls and Lessons for the Future. Rev Med Virol. 2020;2020(e2205):11.
25. Konwarh R. Can CRISPR/Cas Technology Be a Felicitous Stratagem Against the COVID-19 Fiasco? Prospects and Hitches. Front Mol Biosci. 2020;7 (557377):8.
26. Ding R, Long J, Yuan M, Jin Y, Yang H, Chen M, et al. CRISPR/Cas System: A Potential Technology for the Prevention and Control of COVID-19 and Emerging Infectious Diseases. Front Cell Infect Microbiol. 2021;11(639108):10.
27. Agrebi S, Larbi A. Use of artificial intelligence in infectious diseases. In: Artificial Intelligence in Precision Health. Artificial Intelligence in Precision Health; 2020;23:415-38.
28. Tom-Aba D, Silenou BC, Doerrbecker J, Fourie C, Leitner C, Wahnschaffe M, et al. The Surveillance Outbreak Response Management and Analysis System (SORMAS): Digital Health Global Goods Maturity Assessment. JMIR Public Health Surveill. 2020;6(2):9.
29. Silenou BC, Tom-Aba D, Adeoye O, Arinze CC, Oyiri F, Suleman AK, et al. Use of surveillance outbreak response management and analysis system for human monkeypox Outbreak, Nigeria, 2017-2019. Emerg Infect Dis. 2020;26(2):345-9.
30. Adeoye O, Tom-Aba D, Ameh C, Ojo O, Ilori E, Gidado S, et al. Implementing Surveillance and Outbreak Response Management and Analysis System (SORMAS) for Public Health in West Africa- Lessons Learnt and Future Direction. Int J Trop Dis Health. 2017;22(2):1-17.
31. Silenou BC, Nyirenda JLZ, Zaghloul A, Lange B, Doerrbecker J, Schenkel K, et al. Availability and Suitability of Digital Health Tools in Africa for Pandemic Control: Scoping Review and Cluster Analysis. JMIR Public Health Surveill. 2021;7(12):16.
32. Nguyen G, Dlugolinsky S, Bobák M, Tran V, López García Á, Heredia I, et al. Machine Learning and Deep Learning Frameworks and Libraries for Large-scale Data Mining: A Survey. Artif Intell Rev. 2019; 52(1): 77-124.
33. Alzubaidi L, Zhang J, Humaidi AJ, Al-Dujaili A, Duan Y, Al-Shamma O, et al. Review of Deep Learning: Concepts, CNN Architectures, Challenges, Applications, Future Directions. J Big Data. 2021; 8(53):74.
34. Ching T, Himmelstein DS, Beaulieu-Jones BK, Kalinin AA, Do BT, Way GP, et al. Opportunities and obstacles for deep learning in biology and medicine. J R Soc Interface. 2018;15(141):47.

Cite this article as: Owhonda G, Luke A, Inyele JR, Eze-Emiri C. The data analysis and validation engine: an application of artificial intelligence in the improvement of COVID-19 data management. Int J Community Med Public Health 2022;9:2437-41.