IST-Africa

IST-Africa 2022 Conference Proceedings
Miriam Cunningham and Paul Cunningham (Eds)
IST-Africa Institute and IIMC, 2022
ISBN: 978-1-905824-68-7

Toward Improved Data Quality in Public Health: Analysis of Anomaly Detection Tools applied to HIV/AIDS Data in Africa

Folashikemi Maryam Asani OLANIYAN[1], Adebowale OWOSENI[2]
[1]*African Village (Tech Innovation), Adeniran Ogunsanya, Surulere, Lagos, Nigeria*
Email: folashikemiolaniyan@gmail.com
[2]*Centre for Computing and Social Responsibility, De Montfort University, Leicester, UK*
Tel: +44 116 250 4553, Email: adebowale.owoseni@dmu.ac.uk

Abstract: The study examined the data quality efficiency of the WHO Data Quality Review (DQR) toolkit and PyCaret anomaly detection algorithms. The tools were applied to the African HIV/AIDS data (2015-2021) extracted from a public data repository (data.pepfar.gov). The research outcome suggests that unsupervised anomaly detection algorithms could complement the efficiency of the WHO DQR toolkit and improve Data Quality Assessment (DQA). In particular, the study showed that anomaly detection algorithms through python programming provide a more straightforward and more reliable process for detecting data inconsistencies, incompleteness, and timeliness appears more accurate than the WHO tool. Consequently, the study contributed to ongoing debates on improving health data quality in low-income African countries.

Keywords: Data quality review, anomaly detection, data quality assessment, public health, low-income country

1 Introduction

Reliable data is critical to effective decision-making across all spectra of health systems and plays a prominent role in health policy development, implementation, governance, regulations, financing, and service delivery. Accurate and reliable data is crucial to global health practice [1]. Premier health institutions such as WHO, USAID, PEPFAR, UNAIDS and other local health organisations recognise that data quality constitutes a core medium to promote best practices in all areas of health delivery. Thus, they consistently seek to adopt an optimal strategy to facilitate data-driven quality health service delivery through empirical techniques [2]. However, recent data collection and review methodology is becoming increasingly complex and complicated due to the variety, volume, and velocity of data, coupled with frequent changes in underlining digital technologies that drive the act of making sense of data. This phenomenon makes the gaps in data quality obvious, especially in low-income countries with low investment in Health Information Systems (HIS) for data collection, storage, analysis, and reporting. Predictably, the quality of information derived from a dataset cannot be better than the quality of the dataset primarily. Therefore, as the need to improve global health management through defined policies and strategies intensifies, data collation and review accuracy becomes more pertinent. More attention should be focused on ascertaining the quality of data collection and collation.

National health authorities and partner organisations appear coordinated in their approach to data quality because they adopt the same Data Quality Review (DQR) toolkit to examine the quality of health facility data. DQR toolkit supports consistent, periodic,

independent, and objective assessments of facility-reported data. The main goal of the tool is to examine the quality of data generated by health facility-based Information systems; it incorporates guidelines and techniques that outline the basis for a shared understanding of data quality. Also, the toolkit aims to advance the institutionalisation of DQR at national level. The essence of the tool is to assess data completeness, consistency, and timeliness; in simple terms, the toolkit is expected to detect any anomaly in data. The toolkit adopts simple statistical techniques to achieve this overarching goal. However, technology advancement using artificial intelligence and machine learning techniques may present a better approach to DQR. In recent developments, supervised, semi-supervised and unsupervised machine learning models such as anomaly detection algorithms have been implemented to solve various problems in healthcare sectors, from monitoring real-time outlier occurrence in sensory data of Wireless Sensor Networks [3] to fraud detection cases using data derived from hospitals in Brazil [4]. Although global health organisations do not currently implement anomaly detection models in solving data quality issues, [5] opined that machine learning algorithms could support big data in healthcare by detecting data completeness and consistency; an example of the machine Learning library for anomaly detection is PyCaret Library, which runs the PyCaret's model [6].

Due to the complexity and dynamism of reporting health data at administrative levels and to sustain the accuracy of data analysis and subsequent critical decisions that will be drawn from the data, this paper comparatively considered the data quality efficiency of the WHO Data Quality Review (DQR) toolkit and PyCaret anomaly detection model. The tools were applied to the Nigerian HIV/AIDS data (2015-2021) extracted from a public data repository (data.pepfar.gov).

2 Objectives

In healthcare delivery, there are three Data Quality Dimensions (DQD): data completeness and timeliness, internal consistency of data, and external consistency of data. This study compared the effectiveness of WHO's DQR toolkit with PyCaret's anomaly detection model, along with data completeness and timeliness/consistency dimensions. Consequently, this paper made recommendations following a systematic presentation of the results to articulate the findings and their implications - merits and demerits in practice. The study evaluates the current methods for DQR by WHO, identifies areas requiring improvements, and makes recommendations.

3 Literature Review

In this section, we extend the motivation for this study and the ongoing argument on DQR. We explore literature that speaks to data quality in healthcare and from a low-income country viewpoint.

3.1 Data Quality in Healthcare

Poor data quality can cause health facilities to under or overstate the performance of program indicators [2]. Moreover, the variations in the quality of data from sub-Saharan Africa could impede the utilisation of these data to improve the health care system. Extant literature outlines the features of health-related data quality; these include accuracy, continuity, coherence, consistency, availability, clarity, accessibility, timeliness, credibility, interpretability, reliability, authorisation, usability [7]. [8] shows that frequently, dimensions are reported to include accuracy, timeliness, consistency, objectivity, transparency, reputation, and security. The components of data quality vary by author, and some of these attributes overlap based on the worldview of the proponent. However, the following qualities appear consistent across literature: availability, accuracy, timeliness,

completeness, and consistency. This study focused on data completeness and consistency, one of the critical dimensions of DQA.

3.2 Data Quality Assessment (DQA)

DQA is critical in understanding the confidence level in health facility data and discovering problematic areas in generating quality data [9]. It is a restorative procedure that allows organisations, policymakers and managers to determine the quality of data at any given time. In addition, DQA helps organisations develop and implement strategies to deal with data quality issues. Data quality assessment focuses on reported routine standard indicators through health facility information systems [2].

Data verification is a crucial component of DQA; it enables data from source points (registers and tally sheets) to be recounted quantitatively and compared to reported data through the Health Management Information System (HMIS). On the other hand, there is also an assessment of the report's availability, timeliness, and completeness, which measures how service delivery sites and intermediate aggregation sites collect, compile, and report data promptly. DQA ensures that information from source documents is accurately transmitted to the next level of reporting, and all reporting hierarchy levels are verified (from service delivery level to the National Level). Thereby, it exposes systematic errors in reported data to be recognised and estimates the extent of discrepancies resulting from over and under-reporting [2]. Without verification, it will be difficult to establish whether data from any organisation is accurate; importantly, data completeness and timeliness is pivotal to data verification and, in the bigger context, DAQ.

3.3 Data Completeness and Timeliness as Data Quality Dimensions (DQD)

Data completeness is the extent to which all essential steps in data collection, data cleaning, data entry, and data analysis have been executed thoroughly [10]. They further argued that data completeness ensures that no data are missing, no responses are incomplete and uncollected. Routine data quality assessment would examine the completeness of data at two levels: the completeness of reports submitted to the district level and the completeness of data elements in registers and forms. This means the number of reports submitted and the number of health facilities expected to report for the assessment period is paramount for evaluating completeness reporting trends [2]

Outcomes of previous studies on data completeness and timeliness of healthcare data have varied results; for instance, a study conducted by [11] on the National Assessment of Data Quality and Associated Systems-Level Factors in Malawi found a median data completeness score of 0.92 (0.79,1.00) and 0.99 (0.98,1.00) at District Hospitals and District Health Office (DHO), respectively. Thus, this indicates the high completeness of reports at the facility and District office. Another study [10] found that data completeness of age, parity, and haemoglobin level was 99.1%, 98.0%, and 85.8%, respectively, in the source register. Also, they found out that the mean Percentage for these variables from primary data sources from all districts was 94.3%. The common factor in this study underscores the importance of data quality and the need to consider the underlining mechanism for assessing data quality. Logically, anomaly detection appears to be an essential and integral component of the DQA. The standard and generally convenient methods for anomaly detection were packaged as the WHO DQR toolkit, and to some health data analysts, these appear like a black box.

3.4 Anomaly Detection as DAQ Tool

Anomaly detection detects unusual objects or occurrences in databases that are out of the ordinary [12]. Unsupervised anomaly detection focuses on unlabelled data by using only the intrusive information of the data to detect deviations from the majority. Anomaly

detection is a core problem in machine learning and data mining discussed in various practical applications, including network intrusion detection, fraud detection, and the life science and medical domains. Research on anomaly detection for DQA of facility-level public health datasets appears scarce. Limited literature exists on anomaly detection algorithms to improve data quality assessment in HIV service delivery by the WHO, PEPFAR or any other public health organisation.

In a study of discrete sequence health data, [13] proposed unsupervised anomaly detection models using LSTM neural network and Empirical Distribution Function (EDF) algorithms to automate fraud detection of healthcare management systems. [14]. Another study [15] suggests the Linear Support Vector Method (SVM) algorithm as a statistical modelling method for detecting abnormal wireless sensor networks (WSN) in healthcare. Also, [16] proposed using Convolutional Neural Networks (CNN) to detect structural health monitoring data anomalies. However, it will be beneficial to study how machine learning algorithms could enhance DQA from a global health regulatory perspective and what the outcome could mean to different layers of health facilities. Additionally, from a low-income country perspective with varied and limited access to emerging data management techniques. In essence, this study sought to compare the effectiveness of WHO's DQR toolkit with PyCaret's anomaly detection model along with data completeness and consistency dimensions.

4 Methodology

4.1 Data Collection and Analytic approach

This study used a publicly available HIV/AIDS dataset captured at health facility level data from the United States President's Emergency Plan for AIDS Relief (PEPFAR) program (United States Department of State, 2021). The data were retrieved from data.pepfar.gov in March 2021 and included data on several key program indicators from October 1, 2014, to December 31, 2021, in the excel workbook format (.xlsx) 1,099,274 records were retrieved. Using HIV/AIDS data, the study evaluates the data quality infrastructure of the WHO DQR Toolkit and PyCaret's anomaly detection algorithm.

We conducted the comparative analysis in two phases; the first stage assessed the WHO Quality Toolkit in Excel format from the Measure, Evaluation, and Reporting (MER) site. In the second phase of the analysis, we ran the HIV/AIDS data through PyCaret's anomaly detection model using the Python library – PyCaret via Jupyter Notebook. Subsequently, the efficacy and efficiency of the two toolkits were evaluated based on: ease of use, the accuracy of the results and replication factors in concluding on a better method for data quality assessment. The study employed descriptive evaluation and inferential statistics to provide the comparative outcomes of the findings.

Table 1: Data Analysis summary using WHO Took Kit and PyCaret Anomaly Detection

DQR Approach	**WHO Tool Kit**	**PyCaret Anomaly Detection Models**
Programming Language	Microsoft Excel Macros	Python
Platform	Microsoft Excel	Jupyter Notebook
Library	None (Excel Functions)	Scikit-Learn and PyCaret

Analysis Steps: ***DQD for Completeness***	Impute Assessment Information Define Data Quality Thresholds Impute values for metric analysis Review results	Import relevant libraries into Jupyter Notebook **Determine Data Completeness;** A simple mathematical operation (code) groups each Operating Unit and returns the Percentage of missing values in each Quarter. Any quarter reporting missing values greater < 25% (i.e., the reported service delivery is less than 75%) is reporting poorly. **Treat missing values:** The ***sklearn.impute.SimpleImputer*** is used as an imputation transformer for the missing values.
DQD For Consistency - Outlier Detection	Method imputes manually collected values and uses simple statistical techniques to return extreme outliers for districts with greater than 3 standard deviations from the mean value	PyCaret Anomaly Detection Library compares the following algorithms: Isolation Forest, k-Nearest Neighbour, and local outlier factor (LOF). Data is analysed, and the best model is determined based on the Accuracy Score generated.
DQD For Consistency – Timeliness	Imputes aggregated district-level data for the last three years to view a trend (line chart) of reporting trends and compares it with the value of the current year	A time-series anomaly detection model is advised to determine district level consistency of values over time.

5 Results and Discussion

This section summarises the results of analysing the PEPFAR PMTCT data using the WHO Data Quality Assessment Toolkit and Anomaly Detection Models. Figures 1, 2 and 3 provide a graphical representation of the anomaly detection outcomes.

5.1 DQD for Completeness

The WHO Data Quality Dimensions defined by the MER report identifies data as incomplete if the Percentage of reporting at district levels is below 75%. The analysis using the WHO DQA toolkit gives a detailed result of the data completeness; however, the integrity of the results can be questioned because the district values are first aggregated to facility levels and then manually imputed into the Excel Workbook. The imputation process is slow and tedious.

In comparison, the python programming code imports the raw data without the necessity for aggregation or manual imputation. The process is agile, and the results are reliable. However, knowledge of programming is required. This suggests the python code provides a more reliable model for detecting data incompleteness at all reporting levels.

5.2 DQD for Consistency – Outlier Detection

The WHO DQA Tool does not provide an automated method of detecting outliers beyond the facility level detection. The analysis results detected extreme outliers using simple

statistical techniques for deviations from the mean value at greater than three standard deviations. The results of the WHO DQA toolkit require further investigation as the raw data is aggregated before imputation, and the result is for only 1000 inputted facility data (see *Figure 1*).

DOMAIN 2: INTERNAL CONSISTENCY OF REPORTED DATA

Indicator 2a: Identification of Outliers

Indicator 2a1: Extreme Outliers (>3 SD from the mean)		2021		
	National score	Districts with >= 1 extreme outliers relative to the mean in the year of values		
Program Area and Indicator	%	No.	%	Name
Prevention - PMTCT_STAT	10.0%	12	1.2%	Benguela, Huambo, Gaborone District, Kweneng East District, Serowe District, DS Bujumbura sud, DS Gitega, DS Ngozi, Ayos
Total (all indicators combined)				

Interpretation of results - Indicator 2a1:
- **Overall, of the 97 districts, there is good consistency in the data reported**

Figure 1: Outlier detection results using the WHO DQA tool

The Pycaret model bins the rows with anomalous data helping subject matter experts identify inconsistency at the facility level in the data. Consequently, compared to the WHO tool, the steps involved in identifying facilities reporting anomalies is faster and more scalable. The unsupervised learning approach makes for identifying anomalous grouping facilities=1 and non-anomalous facilities=0. The model evaluates the accuracy of the model.

The isolation forest model provided the most accurate model for detecting the facility records with anomalies. A comparison of the efficacy of the algorithms (iforest, KNN and LOF) using uMap, t-SNE plots shows the isolation forest model bins the anomalies more closely, proving a higher accuracy at predicting facilities reporting inconsistently.

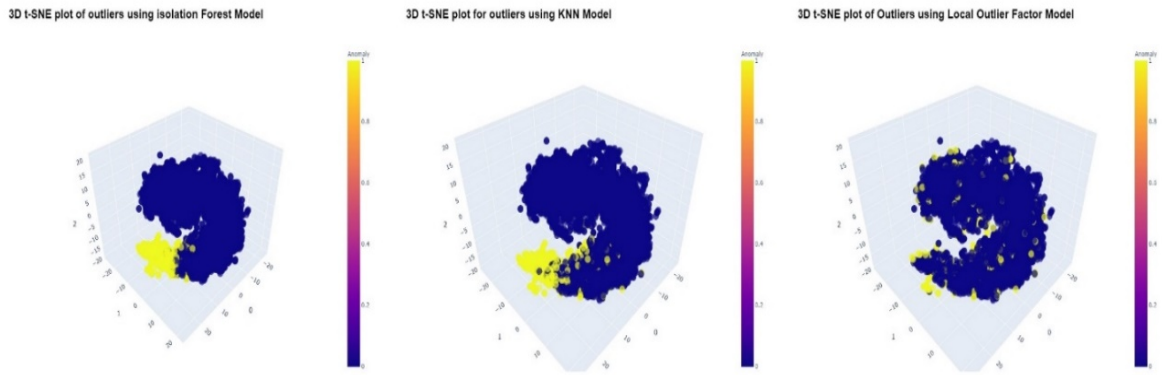

Figure 2: uMap plot showing model accuracy of the isolation forest model compared to KNN and LOF.

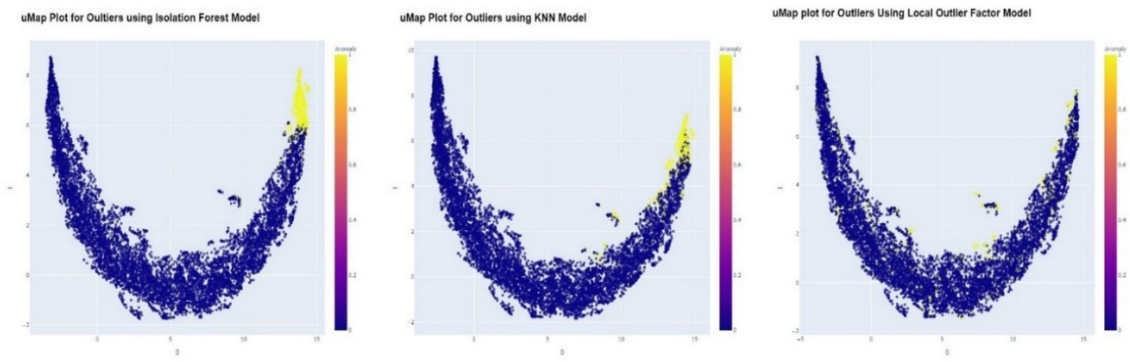

Figure 3 tSNE plot showing model accuracy of isolation forest model compared to KNN and LOF

5.3 DQD for Consistency – Time

The result of the research analysis has discovered four significant findings with the comparison of the WHO DQA toolkit and the PyCaret Models for improving data quality in public health:

1. The suggested model provided a better suited and more accurate result for reporting completeness over the WHO DQA toolkit.
2. The unsupervised anomaly detection model required data preprocessing to deal with missing values. Therefore, a statistical imputation algorithm was applied to input the mean of the reported values at each facility level in the missing rows.
3. The suggested model automates detecting outliers, and the isolation forest algorithm gave the most accurate model for detecting outliers in the data.
4. The time-series anomaly detection method requires timestamps for each facility recording service delivery to be implemented. The WHO tool provides an excellent at-a-glance view of the consistency of reporting over the last three years.

The WHO DQA Tool does not provide an automated method of detecting outliers beyond the facility level detection. The analysis results detected extreme outliers using simple statistical techniques.

6 Implications of Findings

The study suggests that the current method of ensuring data quality in public health by WHO requires manual aggregation and imputation of data which is time-consuming and unrealistic for the large volume of data generated by healthcare organisations. Although the WHO tool currently provides immediate support to health administrators, the accuracy of the results from the tool appears less reliable at facility levels of reporting than the anomaly detection algorithm. This is due to the WHO tool's heavy reliance on manual data imputation. The anomaly detection model does not require data manipulation before use. Instead, a data pre-processing step determines the treatment of missing values.

The anomaly detection technique is replicable and more accurate since the process is scalable. Consideration should be given to anomaly detection algorithms in assessing data quality at the facility level. One of the limitations of this recommendation is the digital skills gap among health workers at the local level. The workers appear more used to Microsoft Excel, but the use of WHO could continue to impede the reliability of health data analysis and the quality of decisions made from such research. Making this change may not be easy, but relevant health organisations should prioritise this improved data quality as a significant deliverable.

7. Conclusion

The study set out to compare the effectiveness of WHO's DQR toolkit with PyCaret's anomaly detection model, along with data completeness and timeliness/consistency dimensions. Using HIV/AIDS data, the study evaluates the data quality infrastructure of the WHO Toolkit and anomaly detection algorithm. The outcome of the comparative analysis indicates that the anomaly detection model appears more reliable in assessing health data quality.

This study focused on only one indicator from the PEPFAR program area and evaluated only the data quality dimensions of completeness and consistency. From the fundamental knowledge of the capabilities of the anomaly detection tool, having an insight on multivariate use of the algorithm in detecting thickness across other indicators will be an excellent way to implement other ML solutions to problems of poor data quality in public health – and this area could be considered in future research. Additionally, other datasets from public health organisations like UNAIDS can be analysed to expand the scope of research on how anomaly detection models could improve health data quality.

References

[1] M Helfert and Mouzhi G., "Big Data Quality - Towards an Explanation Model in a Smart City Context," in ICIQ 2016, p. 2:1-2:8. 2016

[2] World Health Organization. Assessing the National Health Information System: An Assessment Tool Version 4.00. World Health Organization: Geneva, Switzerland. 2008

[3] S.A, Haque; M Rahman and S.M Aziz. 'Sensor Anomaly Detection in Wireless Sensor Networks for Healthcare', Sensors, 15, pp. 8764-8786. Available at: https://doi.org/10.3390/s150408764. (Accessed: 12 April 2021) 2015

[4] Carvalho, L.F., Teixeira, C., Dias, E.C., Meira, W. and Carvalho, O., 2015. A simple and effective method for anomaly detection in healthcare. In Proceedings of the SIAM International Conference on Data Mining Workshop (Vol. 2015, pp. 16-24).

[5] S. Juddoo and C. George, "A Qualitative Assessment of Machine Learning Support for Detecting Data Completeness and Accuracy Issues to Improve Data Analytics in Big Data for the Healthcare Industry," *2020 3rd International Conference on Emerging Trends in Electrical, Electronic and Communications Engineering (ELECOM)*, , pp. 58-66, DOI: 10.1109/ELECOM49001.2020.9297009. 2020

[6] S. M Kasongo, and Y.Sun,. Performance Analysis of Intrusion Detection Systems Using a Feature Selection Method on the UNSW-NB15 Dataset. Journal of Big Data, 7(1). https://doi.org/10.1186/s40537-020-00379-6. 2020

[7] Cai, L. and Zhu, Y., The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. Data Science Journal, 14, p.2. DOI: http://doi.org/10.5334/dsj-2015-002. 2015

[8] Duda, S.N. Identifying, Investigating, and Classifying Data Errors: an Analysis of Clinical Research Data Quality from an Observational HIV Research Network in Latin America and the Caribbean. PhD thesis. Vanderbilt University. Available at: https://etd.library.vanderbilt.edu/etd-03252011-145804 (Accessed: 20 March 2021) 2011

[9] AbouZahr, C. 'Assessing and monitoring the performance of health information systems: Metrics and models'. Health Information Systems Knowledge Hub Working Paper 29. Herston, Australia: Health Information Systems Knowledge Hub, University of Queensland. Available at: Available from: www.uq.edu.au/hishub/. (Accessed: 24 April 2021). 2013

[10] Amoakoh-Coleman, M., Kayode, G.A., Brown-Davies, C., Agyepong, I.A., Grobbee, D.E., Klipstein-Grobusch, K. and Ansah, E.K. Completeness and accuracy of data transfer of routine maternal health services data in the greater Accra region. BMC research notes, 8(1), pp.1-9. 2015.

[11] O'Hagan, R., Marx, M.A., Finnegan, K.E., Naphini, P., Ng'ambi, K., Laija, K., Wilson, E., Park, L., Wachepa, S., Smith, J. and Gombwa, L.,. National assessment of data quality and associated systems-level factors in Malawi. Global Health: Science and Practice, 5(3), pp.367-381. 2017.

[12] Goldstein M, Uchida S. A Comparative Evaluation of Unsupervised Anomaly Detection Algorithms for Multivariate Data. PLoS ONE 11(4): e0152173. https://doi.org/10.1371/journal.pone.0152173. 2016

[13] Snorovikhina, Victoria and Alexey Zaytsev. "Unsupervised anomaly detection for discrete sequence healthcare data." *AIST* (2020).

[14] C. Li, L. Guo, H. Gao and Y. Li, "Similarity-Measured Isolation Forest: Anomaly Detection Method for Machine Monitoring Data," in *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1-12, 2021, Art no. 3512512, DOI: 10.1109/TIM.2021.3062684.

[15] Haque, Shah A., Mustafizur Rahman, and Syed M. Aziz. "Sensor Anomaly Detection in Wireless Sensor Networks for Healthcare" *Sensors* 15, no. 4: 8764-8786. https://doi.org/10.3390/s150408764. 2015

[16] Bao Y, Tang Z, Li H, Zhang Y. Computer vision and deep learning-based data anomaly detection method for structural health monitoring. Structural Health Monitoring.;18(2):401-421. doi:10.1177/1475921718757405. 2019