Automated Data Quality Control System in Health and Demographic Surveillance System

Joseph Tlouyamma, Sello Mokwena
University of Limpopo, Private Bag X1106, Sovenga, 0727, Polokwane, 0700, South Africa.

Abstract

The automated data quality management system serves as a comprehensive solution developed to enhance the precision, dependability, and uniformity of data within data-driven organizations. Such systems play an important role in eliminating the shortcomings associated with manual data quality management, which is prevalent in health and demographic surveillance systems (HDSS). The ongoing difficulty of ensuring data quality through manual processing hinders the HDSS's capacity to optimize data quality effectively. To address this challenge, our study adopted design science methodologies to provide guidelines for the design and implementation of the automated data quality control system. The open source technologies (Pentaho data integration, R Studio, SQL, Windows task scheduler) were used to facilitate the automation and validation of the incoming and database resident data. The quality of data has vastly improved since the implementation of the proposed system. The findings suggest that the automated data quality control system exhibited superior performance compared to the manual methods, thereby minimizing errors and time-wasting efforts.

Keywords: *Automated data quality control, Application programming interface, Data quality framework, Health and demographic surveillance system, Data quality*

1. Introduction

Significant volumes of data are generated daily, presenting a formidable challenge pertaining to ensuring quality control. Data management operations are rendered inefficient and strenuous when data quality is executed manually. Exponential technological advancements have enabled groundbreaking developments in data automation processes [1], [2], [3], fundamentally transforming the way enterprises administer data quality. Automation substantially reduces the likelihood of errors, inconsistencies, and inaccuracies in datasets by mechanizing the conventional labor intensive procedures of data cleansing and validation [4], [5]. With the ongoing adoption of digital transformation by organizations, the incorporation of automation technology improves operational efficiency, but also creates opportunities for development and innovation in many sectors. Automation may also provide enormous benefits to Health and Demographic Surveillance Systems (HDSS) data which are mostly managed manually.

HDSS is a longitudinal data collection system that collects data on health, demographics, and social factors in well-defined areas. They track births, deaths, and migration, providing valuable research data. This data informs policy formulation, and improves Health Information Systems. By engaging communities and investing in technology, data scope, and local capacity, HDSS can become more powerful tool to improve the health of the entire population. Data generated through HDSS is subjected to human quality control, resulting in a multitude of complications. The time required to manually examine flagged data records shows a

Corresponding author: Joseph Tlouyamma *(joseph.tlouyamma@ul.ac.za)*
Received: *12 February 2024*; Revised: *25 April 2024*; Accepted: *9 May 2024;* Published: *20 May 2024*

substantial increase in proportion to the quantity of the data set [6], leading to longer data turnaround times. The use of an automated quality control system can effectively detect and eliminate inaccurate data records, eliminating the need for manual review [7].

This study proposes an automated data quality control system to eliminate erroneous and inefficient manual handling of HDSS data. The paper is organized as follows: Section 2 explores the literature to uncover the truth about the topic while Section 3 discusses methodological procedures used in conducting the study. The analysis of the results is presented in Section 4, while a discussion of these results is provided in Section 5. Section 6 concludes the paper.

2. Related Work

Methodologies for automated data management have been established in various fields [8], [9], [10], [11]. The implementation of these methods requires the maintenance of machine-readable data definition files, including code to establish goal states linked to data assets. When these data definition files are used, the intended target states can be achieved [12]. AutoControl and LiteGreen are tools that have been developed within the realm of virtualized data centers [13]. Their purpose is to enable automated adaptation to dynamic changes and the achievement of service-level objectives, while ensuring optimal use of resources, performance of applications, and minimal power consumption. The foundation of these tools lies in virtualization technology, control theory, and optimization methods.

Business intelligence technologies, such as SAS, have the potential to be used for HDSS data integration, hence streamlining the process [14]. Using these technologies can increase the precision and effectiveness of HDSS data cleansing and integration [13]. Data quality testing and analysis are made possible through the use of automated technologies, which also enable the identification and reporting of outlier variables and trends in datasets [15].

3. Research Methods and Design

3.1 Study context

This study was conducted in Dimamo, an HDSS established in 1995, located approximately 30km from Polokwane, the capital city of Limpopo province in South Africa. The HDSS collects data from more than 21,000 households with an estimated population of 100 000. Researchers use the collected data to monitor the burden of diseases, socioeconomic statuses, demographics, and other health related issues. Data quality controllers manually check each variable for completeness, accuracy, and validity. Despite the ability of this approach to detect and rectify the majority of data quality concerns, the process is laborious, inaccurate and ineffectual, hence compromising the quality of the data. A breach of data quality has the potential to result in substandard judgments, erroneous models, misleading conclusions, and in some cases, financial loss.

3.2 Research approach

Design science was used in the conceptualization and development of an automated data quality control system. The steps involved problem identification and definition (dealt with in the previous sections), planning and designing the solution, and implementation.

3.2.1 Planning and design of the solution

Contemporary automation control design requires that solutions be planned and designed in an automated context. It involves the establishment of a virtual environment and the integration of automation control algorithms with physical equipment [16]. Figure 1 shows an automated data quality control workflow implemented in a Hyper-V-based server environment.

The environment allowed field data collectors and their supervisors to sync data through virtual private networks (VPN) to the PostgreSQL database. Data can be autonomously exported using the R application programming interface (R API) in preparation for quality control. Subsequently, the Pentaho Data Integration (PDI) job was configured to autonomously extract the data from the repository and load it into MS SQL server database.

Structured query language (SQL) scripts were designed to verify incoming data to validate its quality. In a manner that is devoid of human interaction, each variable undergoes a comprehensive, integrity, and accuracy verification process. The SQL script included rules to protect against data quality violations.

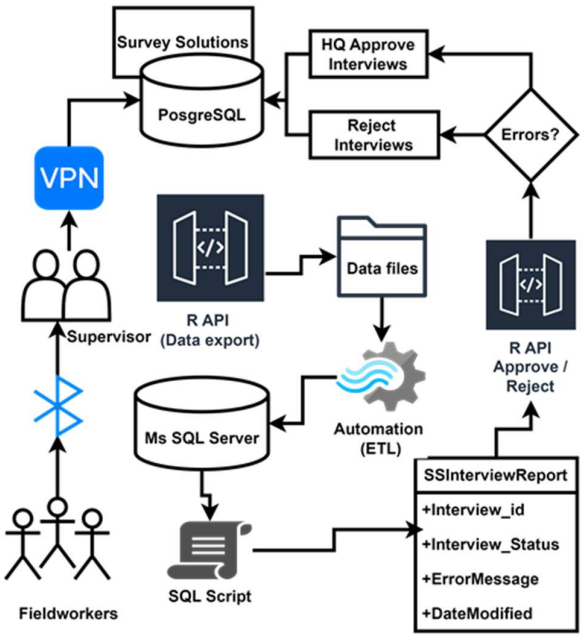

Figure 1. A workflow of an automated data quality control system.

The data quality concerns are identified and, based on the data's status, the R API either rejects or approves the data. In the event of rejection, the data record would be returned to the field data collector for error correction; otherwise, it would be permanently saved in the database. The aforementioned procedures were coordinated and automated to effectively apply validation rules to quality control data in the system. Automated data quality control solutions provide a multitude of advantages. These systems have the ability to identify issues in the data, resulting in a substantial reduction in error rates [17]. They use quality control procedures and accuracy thresholds to ensure the delivery of high-quality data [18]. Automating data quality control procedures can guarantee the accuracy, dependability, and timeliness of the data [19].

3.2.2 Data quality control process

The data was validated at three different levels to ensure optimal quality. First level involved validation algorithms or C# macros and skip logic patterns at the application (application used for data collection) level. At this level, the checks were made against missing values, incorrectly captured values, invalid date formats, and data compliance with previously collected data. The latter was ensured by preloading data at the application level. These measures ensured the acquisition of comprehensive and precise data.

At the second level, the data from previous level was fed into SQL script to perform further validation or data quality checks. In addition to data accuracy and completeness, this level verified consistency and validity of the incoming data to the historic data in the database. For example, a record for male participant associated with pregnancy or linked to a child as the mother must be considered invalid. Conversely, an inconsistent error message is generated when, for instance, a member who was formerly classified as male is presently identified as female.

Third level involved the implementation of stored procedures at the database level. At this level, data irregularities that evaded detection in preceding levels were identified and then rejected for rectification. This level validates data against data entry errors, logical errors, and spatial errors.

These levels of data quality assurance were automated through the process described in section 3.2.1.

3.2.3 Implementation and Validation

Automation was achieved through the integration and deployment of various tools such as the R API, PDI, Bash-programmed files, and the Windows task scheduler. The system was configured in a Windows 2016 server environment with Windows task scheduler used to autonomously launch R API and PDI. The automation was coordinated to ensure a sequential execution of tasks and guarantee the quality of an end product. The task scheduler executes a file with extension .bat in silent mode. This file was programmed using the Bash programming language, and its code snippet is shown below. The code can be used to launch any type of file in a Windows environment.

```
if not DEFINED IS_MINIMIZED set IS_MINIMIZED=1 && start "BO darbi" /min "%~f0" %* && exit
 "F:\data-integration\kitchen.bat" /file:F:\Data-App-ETL\ApD0001-Master.kjb /level:basic
exit /b
REM EXECUTE TASK KILL TO CLOSE CMD PROGRAM.-
wmic Path win32_process Where "Caption Like 'cmd%.exe'" Call Terminate
```

The command prompt (CMD) is launched in a minimized mode to execute kitchen.bat, which was used to execute PDI job (ApD0001-Master.kjb). In the current context, the code was designed to autonomously launch the PDI job and the R API to process, integrate, and quality control data. Automation was achieved through the use of the Windows task scheduler to configure and schedule the execution of the PDI and R API. The entire system was scheduled to run autonomously at 12:00 midnight for data quality control. Running data quality management tasks at midnight offers a number of advantages.

• Reduced impact on users: One way to reduce the impact on users is to carry out data quality management procedures during off-peak hours. Due to the general trend of reduced user activity and interaction with the system, midnight is characterized by reduced traffic. This ensures that the system maintains full functionality throughout standard business hours.

• Increased system availability: While performing autonomous data quality control activities, substantial system resources may be used. It is possible to decrease the probability of service interruptions or slowdowns that may impact users during key business hours by organizing these tasks for midnight, when the system is expected to encounter reduced demand.

• Optimized resource use*:* Optimization of system resource consumption is achieved by performing data quality management operations during off-peak hours. Midnight is characterized by a reduction in resource demands and conflicting activities, which facilitates the execution of data management tasks with greater efficiency and expedites their completion.

• Improved data accuracy: A contribution to enhanced data precision may be made by autonomously performing data quality control operations when the system is inactive. The execution of data quality management operations can occur uninterrupted, hence reducing the likelihood of conflicts or inconsistencies that may arise from concurrent user activity.

3.2.4 Error validation and detection

The proposed automated system was configured to receive data from the field and validate it to meet acceptable quality standard.

Table 1. Example of errors generated by automated data quality system.

ERROR CATEGORY	TYPE OF ERROR
Logical errors	Unable to save BSU Document Id 274365 Interview key 17-02-68-71. An error occurred while updating the entries. See the inner exception for details. ERROR P60: This would make this pregnancy "p11002" to ID1203, ending 03 Jan 2021, end within less than 140 days of the end of another! The transaction ended in the trigger. The batch has been aborted.
Data entry error	Unable to save BSR Document Id 281301 Interview key 48-66-93-74. Invalid Location start date ##N/A##
Logical errors	Unable to save BSR Document Id 281879 Interview key 61-23-74-25. Member [member name] cannot be its own mother
Data entry error	Unable to save BSR Document Id 282156 Interview key 03-42-66-67. 'Other' in not a valid membership start for member [member name] - create an in-migration for this member
Data entry error	Error processing BSR Document Id 282200 Interview key 77-73-56-23 VPDate 20291127 not within bounds 2006/04/24 and 2020/02/24
Logical errors	Unable to save BSR Document Id 418034 Interview key 19-03-03-02. Member [member name] cannot be in a conjugal relationship with herself
Logical errors	Unable to save BSU Document Id 424199 Interview key 36-25-64-14. More than one pregnancy outcome linked to the same baby "[member name] - Female - 33" is involved
Data entry error	Unable to save Call Centre Document Id 448744 Interview key 40-43-60-38. Individual null for member #7, 'Is new member?' probably not answered
Data entry error	Unable to save BSU Document Id 455946 Interview key 54-21-25-71. No socio-economic data for household '3b319138-69f1-e811-b803-00505696b21a'
Spatial error	Unable to save BSR Document Id 617016 Interview key 36-85-27-01. No GPS coordinates for location

The data quality was validated from over 900 variables. These variable were classified as texts, dates, spatial and numeric. These variables were validated against data entry errors (typographical errors and missing data), logical errors (inconsistencies in birth dates, age at death, or implausible values for certain variables), and spatial errors (incorrect GPS coordinates). The algorithms deployed by automated system accurately identified and classified errors, allowing clean data into the database. If any data anomaly is detected, then the detailed error thrown. Table 1 provides some examples of errors produced during data validation by automated data quality system.

3.2.5 System evaluation

The automated data quality control system was tested and deployed in a real-world setting over a period of one year from June 2021 to 2022. The quantitative approach was used to evaluate the efficiency of the proposed automated system. The evaluation was based on the system's ability to produce a data of good quality. The metrics considered for the evaluation were accuracy, completeness, timeliness, and validity.

The utilization of these metrics for assessing data quality is not only righteous but also indispensable in a HDSS, where complete, accurate, and timely data is vital for successful decision-making and public health interventions. Ensuring the dependability of information, accuracy is emphasized as the most critical component of data quality [20]. While assessed less often, timeliness is a crucial factor in guaranteeing that decision-making data is current and pertinent. Data completeness is a parameter that is extensively evaluated across numerous research; it signifies the degree to which every essential data element is included in the records [22], [23]. The combined influence of these three metrics enhances the overall data quality in HDSS, hence empowering policymakers and researchers to formulate well-informed judgments grounded in dependable, up-to-date, and all-encompassing information.

The metrics were modelled by considering database D_k as consisting of data values $x_1, x_2, x_3 \dots \dots \dots \dots x_n$. The dirty data (D_i) may be represented as $D_i \in \{ x_1, x_2, x_3 \dots \dots \dots \dots x_n\}$ and clean data, denoted by C_j for $C_j \in \{ x_1, x_2, x_3 \dots \dots \dots \dots x_n\}$ where $i, j, k = 1,2,3, \dots \dots \dots n \ \ \forall \, n \in Z+$. This implies that $D_i \cup C_i = D_k$.

The accuracy or completeness can be computed as

$$Q_x = \frac{\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} m_i}{\sum_{i=1}^{n} x_i} = 1 - \frac{\sum_{i=1}^{n} m_i}{\sum_{i=1}^{n} x_i} \qquad (1)$$

Where m_i represents missing or inaccurate values and $0 \leq \frac{\sum_{i=1}^{n} m_i}{\sum_{i=1}^{n} x_i} \leq 1$

In order to optimize the output Q_x , it is necessary to minimize the value of $\frac{\sum_{i=1}^{n} m_i}{\sum_{i=1}^{n} x_i}$

Next, we seek to compute data timeliness, which demonstrates the proximity of shared data to real-time that must be maximized. Therefore, the data must be current as a reference to the current moment [24]. To assess this, we model data timeliness (T_x) as $S_i \leq C_i \leq E_i$. C_i is the current date value bounded between the start observation date (S_i) and end observation date (E_i). T_x can then be computed as

$$T_x = \frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} C_i + \sum_{i=1}^{n} E_i} - \frac{\sum_{i=1}^{n} E_i}{\sum_{i=1}^{n} C_i + \sum_{i=1}^{n} E_i} \qquad (2)$$

Since C_i and E_i are incoherent subsets of x_i then, it implies that $\sum_{i=1}^{n} C_i + \sum_{i=1}^{n} E_i = \sum_{i=1}^{n} x_i$. So equation (2) can be simplified as

$$T_x = 1 - \frac{\sum_{i=1}^{n} E_i}{\sum_{i=1}^{n} C_i + \sum_{i=1}^{n} E_i} = 1 - \frac{\sum_{i=1}^{n} E_i}{\sum_{i=1}^{n} x_i} \qquad (3)$$

Obtaining the highest proportion of timely data is contingent upon the minimization of $\sum_{i=1}^{n} E_i$, which has an impact on the ultimate output of T_x .

Lastly, we compute the validity of all the date variables in the system. Data validity involves checking if a data item complies with its definition's syntax (format, type, range). The objective is to determine the proportion of data points whose values fall within the permitted range. The validity of data V_d can be defined by

$$V_d = l_b \leq x_i \leq u_b \qquad (4)$$

Where l_b is the lowest expected value and u_b the highest expected value with x_i being the actual date value stored in the database.

This study employs a comprehensive data quality evaluation process that includes automated data quality checks, external validations, and ongoing monitoring. This integration guarantees that the gathered data is reliable and of superior quality, thus facilitating meaningful health and demographic analyses.

4. Analysis of Results

This study implemented an automated data quality management system to improve HDSS data quality and eliminate the shortcomings associated with manual data quality administration Analysis of the results provides an insight into the appropriateness of each approach in enhancing data quality. Manual and automated data quality procedures were performed in the database hosted on a Windows server environment. The statistics for manual operations were generated prior to the implementation of the automated system. Comparative analysis was performed to determine the best methods for controlling the quality of the data, especially with a growing amount of data.

In a manual system, quality controllers examine data to identify potential quality concerns and recommend appropriate remedial actions if necessary. On the other hand, automated system uses a set of pre-defined rules to quality control the data and attempt to resolve the issues, if any. Both systems were evaluated on their ability to produce data of good quality. Finding the optimal balance between thoroughness and speed is crucial, as the data validation time can affect data quality in several ways.

A comparison of the time required by manual and automated quality control procedures to validate the quality of data is presented in Figure 2. Verifying the quality of data manually is faster, especially when there are fewer records; however, it becomes more time-consuming as the data size increases. This is particularly true because quality controllers lack the capacity to process large amounts of data in the shortest space of time possible. Manually verifying a substantial volume of data compromise data quality as a result of mistakes, time limits, and human fallibility.

Although automated quality control exhibits lower time spans for data quality verification, the prevailing tendencies are comparable to those reported in human data quality verification. That is, an increase in data size results in a corresponding increase in data quality validation time intervals. Errors in the data may be introduced when a substantial volume of data is manually validated. For example, Figure 3 presents empirical evidence that demonstrates an exponential increase in error rates as the amount of the data grows. For every megabits (MB) of data that has gone through the quality control process, the errors were collected, aggregated and analyzed. A significant proportion of data errors go undiscovered during manual data quality checks and compound as the volume of data increases. Automated quality control, on the contrary, started with slightly higher error rates notwithstanding the lower amount of data. This was attributable to some issues in the validation algorithms. Updating the algorithm resulted in a drop of error rates, however, picks up again as the amount of data increases. While the error rate dropped after the algorithm update, it marginally increased as the amount of data increased. Error rates were much lower for automated system compared to manual data quality validation.

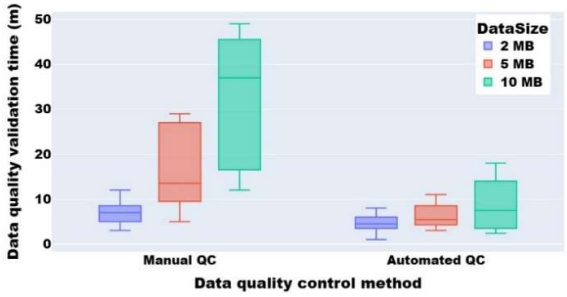

Figure 2. The amount of time spent in quality control of the data.

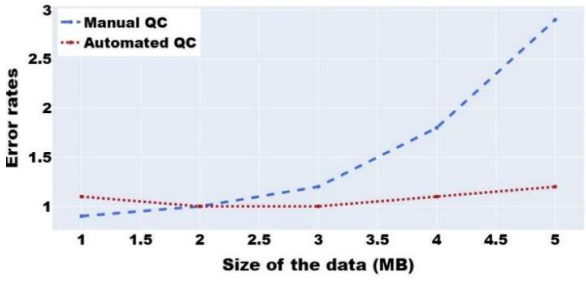

Figure 3. The effect of the amount of data on the data quality.

With lower error rates, automated data quality control has a potential to improve the data quality with the possibility of reducing human's intervention to minimal. In assessing the effectiveness of automated data quality management system in quality improvement, metrics such as data completeness, timeliness, validity and accuracy were considered. The quality of the data was assessed both prior to and subsequent to the system's implementation in order to gauge the system's efficacy in regulating data quality. Figure 4 shows a data quality dashboard that provides an overview on the level of data quality in the database. The dashboard was built using dash-plotly, a web-based implementation of Python. Prior to the deployment of the system, the level of data completeness was at 78%. It then increased by 3% and 6% to 81% and 87% respectively, following several months of automated system usage. A 14% enhancement in data completeness was attained, resulting in an overall data quality of 92%.

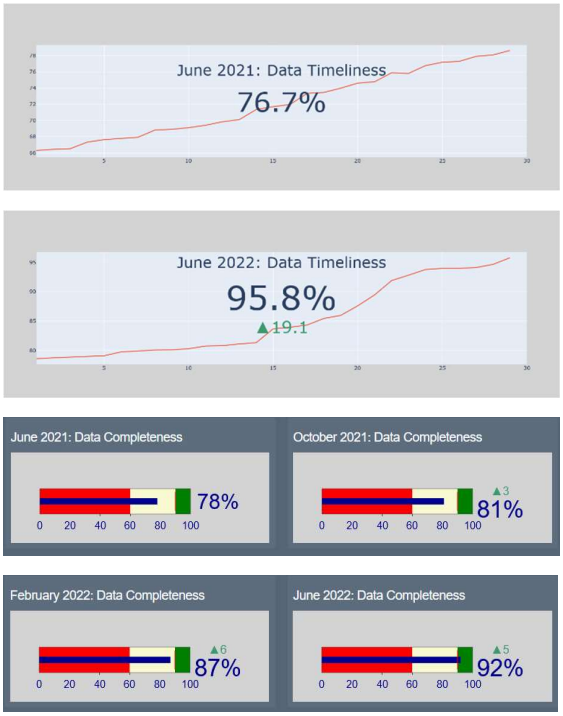

Figure 4. Measurement and reporting on data completeness and timeliness.

On the other hand, data timeliness improved by 19.1%, culminating in a quality level of 95.8%. Data quality with respect to these two metrics improved by 14% and 19.1% between June 2021 and June 2022.

Figure 5 presents other crucial data quality metrics that are used to evaluate the efficacy of the proposed system. Data validity was measured between two points and was reported to have increased by 12%, making quality levels to reach 96.7%. In order to improve data quality, data accuracy was also taken into account. Data accuracy refers to the degree to which the data accurately reflects real world things. The degree to which data represent real-world entities has a bearing on its usability and validity especially for decision making and research purposes. It, therefore, remains crucial to assess with the aim of improving the accuracy of data.

It is evident from Figure 5 that the baseline data quality (data accuracy) before the implementation of an automated data quality control system was 70.2%. A 3% decrease in the level of data accuracy was observed four months after the system was deployed. This was attributable to some faulty data quality validation rules, which allowed some inaccurately captured data items to pass through to the database. Correct adjustment of the rules resulted in 11.6% and 19.1% improvement in the level of data accuracy. Generally, the data accuracy was improved by 28.1% since the implementation of an automated data quality control system.

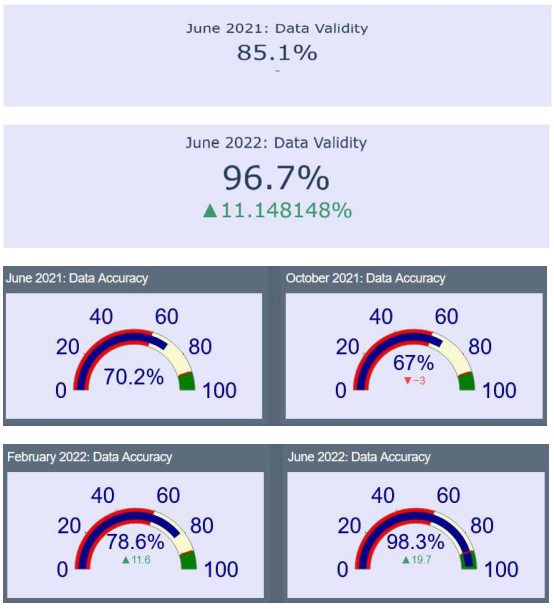

Figure 5. Measurement and reporting on data validity and accuracy.

In general, the increasing trend can be observed and that shows that the implementation of the automated data quality control system had a positive effect on the data quality. With the current trajectory, continued use of the system is likely to improve the quality of data to optimal levels. The quality of the data improved without user intervention, suggesting improved efficiency and optimized resource utilization.

5. Discussions

This study designed and deployed an automated data quality control system and assessed its effectiveness using metrics such as data accuracy, completeness, timeliness, and validity. The findings demonstrated that the overall quality of the data has improved since the implementation of the proposed system.

The evaluation of system efficiency through data quality evaluation is critical due to its substantial influence on the development, verification, and testing of decision-making systems [25]. By conducting a data quality evaluation within a designated system, potential flaws or deficiencies in the used technology can be detected and subsequently rectified in order to enhance the system's overall effectiveness [26]. A subjective and objective measure of data quality may be obtained through the use of complete data quality evaluation metrics [27], which take into account several aspects and accommodate the varying standards of data quality across users. Furthermore, these metrics have the potential to indicate the level of contentment with the data and aid in assessing the system's efficacy. However, assessing the effectiveness of a system presents a number of obstacles. One difficulty is the requirement to take into account contextual elements that affect performance. When assessing validity, it is crucial to consider the particularities of the operating environment of the system [28]. Improper modelling and measurement may also introduce a possible distortion to the system's operation, which is an additional obstacle. Inaccurate measurements have the potential to hamper innovation and disrupt the productivity of the system [29]. These challenges may ultimately affect the quality of the underlying data. This phenomenon occurs specifically due to the fact that the data generated by the system is influenced by its own quality [25]. When automated, the system has the potential to not only improve efficiency, but also accurately execute a set of rules to eliminate issues that affect the quality of the data. Quality checks on the data can be performed automatically, detecting and annotating data that have defects for further repair [30]. The use of completely automated data quality enhancement procedures can result in increased productivity and process simplification. This enables enterprises to optimize the value of their data, improve precision and dependability, and minimize mistakes [31].

In this study, practical approaches were proposed to detect inaccuracies in real time, resulting in more precise and reliable data. Automated alerts on data quality provide timely rectifications, therefore, averting the gradual spread of data anomalies. The improvement of overall efficiency involved the implementation of automated data quality checks, which suggested that human resources may be reallocated to more complicated tasks that are otherwise difficult to automate. Despite the potential benefits of the proposed system in improving the quality of data and efficiency, there exist some limitations. Few data quality metrics were selected to evaluate the efficiency of the automated data quality system. This constrains the ability to evaluate the effectiveness of the system in a more comprehensive framework. Measuring the validity of the data was based only on date variables, which may mean that other variables are not considered. Measuring the validity of free texts using traditional methods remains a challenge, which has resulted in other studies applying machine learning or artificial intelligence. In machine learning, data quality measurement is crucial because it dictates which datasets are utilized, maintained, and updated, especially when dealing with large-scale, high-dimensional data [32]. Although machine learning has the potential to enhance performance in data quality management, there are still several obstacles that must be resolved, including feature selection, model construction, and validation [33]. Our future work will focus on addressing these challenges to enable optimization of data quality in HDSS spheres. This is due to the machine learning's ability of to learn and adapt to intricate situations.

6. Conclusion

This study has shed light on the automation of data quality management systems, offering significant perspectives on data quality improvement. The results of this research provide insights into data quality management and have significant ramifications for data-driven organizations. By implementing automated data quality management, data-driven organizations can improve the efficiency and quality of their data. Undoubtedly, the topic on automated data quality management continues to be a dynamic and ever-growing field of research. Sustained investigation is necessary in order to enhance the comprehension and effectively tackle rising challenges in data quality management. In the quest for understanding, endeavor to reveal more layers of automation and make a positive contribution to the progression of data quality management is necessary.

Competing of Interest Statement

There are no competing interests declared by the authors of this study.

Data Availability Statement

The data used for this study is available and can be shared on request.

Author Contributions

The study was completed by the joint efforts of the authors, whose contributions are outlined below:

Joseph Tlouyamma: Conceptualization, methodology, analysis, data acquisition and writing. Sello Mokwena: Supervision, validation, writing, review and editing.

References

[1] S. K. Pradhan, H.-M. Heyn, and E. Knauss, 'Identifying and managing data quality requirements: a design science study in the field of automated driving', *Software Qual J*, May 2023, doi: 10.1007/s11219-023-09622-8.

[2] O. Ozonze, P. J. Scott, and A. A. Hopgood, 'Automating Electronic Health Record Data Quality Assessment', *J Med Syst*, vol. 47, no. 1, p. 23, Feb. 2023, doi: 10.1007/s10916-022-01892-2.

[3] S. Schelter, D. Lange, P. Schmidt, M. Celikel, F. Biessmann, and A. Grafberger, 'Automating large-scale data quality verification', *Proc. VLDB Endow.*, vol. 11, no. 12, pp. 1781–1794, Aug. 2018, doi: 10.14778/3229863.3229867.

[4] F. Curina, E. Abdo, and H. Mustapha, 'Reducing Human Error Through Automatic Kick Detection, Space Out and Dynamic Well Monitoring', in *Day 2 Wed, March 09, 2022*, Galveston, Texas, USA: SPE, Mar. 2022, p. D021S016R001. doi: 10.2118/208784-MS.

[5] K. J. Ruskin, A. C. Ruskin, and M. O'Connor, 'Automation failures and patient safety', *Current Opinion in Anaesthesiology*, vol. 33, no. 6, pp. 788–792, Dec. 2020, doi: 10.1097/ACO.0000000000000935.

[6] F. Daniel-Durandt and A. Rix, 'The Automation of Quality Control for Large Irradiance Datasets', in *2022 International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME)*, Maldives, Maldives: IEEE, Nov. 2022, pp. 1–6. doi: 10.1109/ICECCME55909.2022.9988452.

[7] J. Daniels, 'Automated training data quality process'. Google Patents, May 18, 2021.

[8] N. Balfe, S. Sharples, and J. R. Wilson, 'Impact of automation: Measurement of performance, workload and behaviour in a complex control environment', *Applied ergonomics*, vol. 47, pp. 52–64, 2015.

[9] K. C. Moffitt, A. M. Rozario, and M. A. Vasarhelyi, 'Robotic Process Automation for Auditing', *Journal of Emerging Technologies in Accounting*, vol. 15, no. 1, pp. 1–10, Jul. 2018, doi: 10.2308/jeta-10589.

[10] S. Madakam, R. M. Holmukhe, Bharati Vidyapeeth University, Pune, India, D. Kumar Jaiswal, National Institute of Industrial Engineering (NITIE), Mumbai, India, and FORE School of Management, New Delhi, India, 'The Future Digital Work Force: Robotic Process Automation (RPA)', *JISTEM*, vol. 16, pp. 1–17, Jan. 2019, doi: 10.4301/S1807-1775201916001.

[11] M. M. Adrita, A. Brem, D. O'Sullivan, E. Allen, and K. Bruton, 'Methodology for Data-Informed Process Improvement to Enable Automated Manufacturing in Current Manual Processes', *Applied Sciences*, vol. 11, no. 9, p. 3889, Apr. 2021, doi: 10.3390/app11093889.

[12] M. F. N. Twumasi, 'Standard operating procedures (SOPS) for health and demographic research data quality assurance: the case of VADU HDSS site', PhD Thesis, 2016.

[13] V. Dasari and S. Suresh, 'HDSS Data Cleaning and Integration Using SAS Business Intelligence Tool', *IBMRD's Journal of Management & Research*, pp. 427–436, 2013.

[14] M. Bushnell, 'Quality Assurance/Quality Control of Real-Time Oceanographic Data', in *OCEANS 2015 - MTS/IEEE Washington*, Washington, DC: IEEE, Oct. 2015, pp. 1–4. doi: 10.23919/OCEANS.2015.7404613.

[15] S. Van Den Berghe and K. Van Gaeveren, 'Data Quality Assessment and Improvement: A Vrije Universiteit Brussel Case Study', *Procedia Computer Science*, vol. 106, pp. 32–38, 2017, doi: 10.1016/j.procs.2017.03.006.

[16] V. M. Pomazan and S. R. Sintea, 'Creative Simulation Environment for Automation Control', *IOP Conf. Ser.: Mater. Sci. Eng.*, vol. 916, no. 1, p. 012087, Sep. 2020, doi: 10.1088/1757-899X/916/1/012087.

[17] R. E. Billo, K. L. Needy, and T. A. Barbe, 'AUTOMATED DATA COLLECTION FOR AUTOMATED QUALITY CONTROL IN ASSEMBLY CELLS', in *Proceeding of Flexible Automation and Integrated Manufacturing 1996*, Atlanta, Georgia, USA: Begellhouse, 2023, pp. 131–138. doi: 10.1615/FAIM1996.140.

[18] G. Zibordi, D. D'Alimonte, and T. Kajiyama, 'Automated Quality Control of AERONET-OC LWN Data', *Journal of Atmospheric and Oceanic Technology*, vol. 39, no. 12, pp. 1961–1972, Dec. 2022, doi: 10.1175/JTECH-D-22-0029.1.

[19] M. A. Meira *et al.*, 'Quality control procedures for sub-hourly rainfall data: An investigation in different spatio-temporal scales in Brazil', *Journal of Hydrology*, vol. 613, p. 128358, Oct. 2022, doi: 10.1016/j.jhydrol.2022.128358.

[20] J. Mantas and others, 'Evaluation of Completeness, Comparability, Validity, and Timeliness in Cancer Registries: A Scoping Review', *Healthcare Transformation with Informatics and Artificial Intelligence*, vol. 305, p. 160, 2023.

[21] N. A. Valcik, M. Sabharwal, and T. J. Benavides, 'Obstacles for Public Organizations Using HRIS', in *Human Resources Information Systems: A Guide for Public Administrators*, Springer, 2023, pp. 129–152.

[22] M. Alwhaibi *et al.*, 'Measuring the quality and completeness of medication-related information derived from hospital electronic health records database', *Saudi Pharmaceutical Journal*, vol. 27, no. 4, pp. 502–506, 2019.

[23] S. L. Feder, 'Data quality in electronic health records research: quality domains and assessment methods', *Western journal of nursing research*, vol. 40, no. 5, pp. 753–766, 2018.

[24] M. S. Islam and P. Young, 'Modeling a Data Storage System (DSS) for Seamless Real-Time Information Support from Information Manufacturing System (IMS)', *DBKDA 2013*, p. 142, 2013.

[25] A.-Ş. Moloiu, G. Albeanu, H. Madsen, and F. Popenţiu-Vlădicescu, 'Data Quality Assessment for ML Decision-Making', in *Applications in Reliability and Statistical Computing*, H. Pham, Ed., in Springer Series in Reliability Engineering. Cham: Springer International Publishing, 2023, pp. 163–178. doi: 10.1007/978-3-031-21232-1_8.

[26] F. H. A. Beevi, S. Wagner, C. Pedersen, and S. Hallerstede, 'Data Quality Oriented Efficacy Evaluation Method for Ambient Assisted Living Technologies', in *Proceedings of the 10th EAI International Conference on Pervasive Computing Technologies for Healthcare*, Cancun, Mexico: ACM, 2016. doi: 10.4108/eai.16-5-2016.2263753.

[27] D. Xu, Z. Zhang, and J. Shi, 'A Data Quality Assessment and Control Method in Multiple Products Manufacturing Process', in *2022 5th International Conference on Data Science and Information Technology (DSIT)*, Shanghai, China: IEEE, Jul. 2022, pp. 1–5. doi: 10.1109/DSIT55514.2022.9943883.

[28] S. Guha, B. Cheng, and P. Francis, 'Challenges in measuring online advertising systems', in *Proceedings of the 10th ACM SIGCOMM conference on Internet measurement*, Melbourne Australia: ACM, Nov. 2010, pp. 81–87. doi: 10.1145/1879141.1879152.

[29] M. J. Overman *et al.*, 'Challenges of Efficacy Assessments in Pseudomyxoma Peritonea', *The Oncologist*, vol. 20, no. 3, pp. e3–e4, Mar. 2015, doi: 10.1634/theoncologist.2014-0415.

[30] C. Destici, K. Şemin, S. Koçak, and H. Özener, 'Monitoring and Data Quality Issues of Mining Activites Around BRTR, Turkiye', oral, other, May 2023. doi: 10.5194/egusphere-egu23-14132.

[31] R. Khare *et al.*, 'Predicting Causes of Data Quality Issues in a Clinical Data Research Network', *AMIA Jt Summits Transl Sci Proc*, vol. 2017, pp. 113–121, 2018.

[32] H. Cho and S. Lee, 'Data Quality Measures and Efficient Evaluation Algorithms for Large-Scale High-Dimensional Data', *Applied Sciences*, vol. 11, no. 2, p. 472, Jan. 2021, doi: 10.3390/app11020472.

[33] B. Aslam, A. Maqsoom, A. H. Cheema, F. Ullah, A. Alharbi, and M. Imran, 'Water Quality Management Using Hybrid Machine Learning and Data Mining Algorithms: An Indexing Approach', *IEEE Access*, vol. 10, pp. 119692–119705, 2022, doi: 10.1109/ACCESS.2022.3221430.