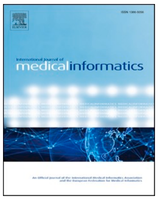
Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.elsevier.com/locate/ijmedinf



Effective data quality management for electronic medical record data using SMART DATA

Seunghee Lee[a], Gyun-Ho Roh[b], Jong-Yeup Kim[a,c], Young Ho Lee[e], Hyekyung Woo[d,*], Suehyun Lee[e,*]

[a] Healthcare Data Science Center, Konyang University Hospital, Daejeon, 35365, Republic of Korea
[b] Biomedical Research Institute, Seoul National University Hospital, Seoul, Republic of Korea
[c] Department of Biomedical Informatics, College of Medicine, Konyang University, Daejeon, 35365, Republic of Korea
[d] Department of Health Administration, Kongju National University, Kongju, 32588, Republic of Korea
[e] Department of Computer Engineering, Gachon University, Seongnam, Republic of Korea

ARTICLE INFO

Keywords:
Data quality management system
Medical data
Electronic medical record
SMART DATA

ABSTRACT

Objectives: In the medical field, we face many challenges, including the high cost of data collection and processing, difficult standards issues, and complex preprocessing techniques. It is necessary to establish an objective and systematic data quality management system that ensures data reliability, mitigates risks caused by incorrect data, reduces data management costs, and increases data utilization. We introduce the concept of SMART data in a data quality management system and conducted a case study using real-world data on colorectal cancer.
Methods: We defined the data quality management system from three aspects (Construction - Operation - Utilization) based on the life cycle of medical data. Based on this, we proposed the "SMART DATA" concept and tested it on colorectal cancer data, which is actual real-world data.
Results: We define "SMART DATA" as systematized, high-quality data collected based on the life cycle of data construction, operation, and utilization through quality control activities for medical data. In this study, we selected a scenario using data on colorectal cancer patients from a single medical institution provided by the Clinical Oncology Network (CONNECT). As SMART DATA, we curated 1,724 learning data and 27 Clinically Critical Set (CCS) data for colorectal cancer prediction. These datasets contributed to the development and fine-tuning of the colorectal cancer prediction model, and it was determined that CCS cases had unique characteristics and patterns that warranted additional clinical review and consideration in the context of colorectal cancer prediction.
Conclusions: In this study, we conducted primary research to develop a medical data quality management system. This will standardize medical data extraction and quality control methods and increase the utilization of medical data. Ultimately, we aim to provide an opportunity to develop a medical data quality management methodology and contribute to the establishment of a medical data quality management system.

1. Introduction

Medical data can enhance the ability of healthcare professionals to engage in evidence-based knowledge management and support decision making to improve medical outcomes. Electronic medical records (EMRs) facilitate the access to and aggregation of clinical data; thus, interest in conducting research using data collected during medical care has increased [1,2]. The secondary use of data collected from EMRs also provides a window into patient-centered treatment, new medical research, and discoveries [3,4]. As the paradigm of medical research shifts to a data-driven approach, the use of medical data has been emphasized [5,6]. The dependence on data to make accurate clinical decisions is increasing, and the use of data is a common challenge in the medical field [7].

The amount of medical data is huge, but the types, forms, and attributes vary widely depending on each institution and disease, so if each institution generates secondary data with only simple technical methods as it does now, there are many limitations in the usability

* Corresponding authors.
E-mail addresses: hkwoo@kongju.ac.kr (H. Woo), leesh@gachon.ac.kr (S. Lee).

https://doi.org/10.1016/j.ijmedinf.2023.105262
Received 15 August 2023; Received in revised form 3 October 2023; Accepted 11 October 2023
Available online 12 October 2023

aspect of the data. Past medical data was difficult to build and utilize high quality data due to various obstacles such as ambiguous data definition, difficult data deployment design, complex management, special analysis tools, and environmental management [8,9]. The quality control of medical data has also traditionally been based on human experience and basic statistical methods.

Recently, several studies have explored medical quality evaluation models that integrate traditional evaluation methods and machine learning-based technologies using EMR data [10–13]. These data can potentially be utilized for accurate and effective clinical support after data quality evaluation-based noise processing, and outliers can reflect specific medical quality information [14]. As the value of EMR data increases, so does the need for quality assessment and quality management to ensure that the data is of high quality beyond the usability of the data.

Data quality management generally refers to the development and implementation of activities that apply quality management techniques to data to ensure that the data meets the requirements in a specific context. In order to manage data quality, you must first understand the basic structure of a specific situation, as well as data quality evaluation technology, management approach, and optimization technology according to the data life cycle from various aspects. Therefore, these developments must go beyond the existing technologies used in databases, clinical research data warehouses, and initiatives to establish new methodological and framework-based technologies optimized to support end-to-end quality management within the data lifecycle [15].

In this study, we introduce the concept of SMART data in a data quality management system and conducted a case study using real-world data on colorectal cancer. The purpose of this study is summarized as follows.

• **Problem or Issue:** The use of EMR data has many limitations due to various factors such as poor data quality, lack of domain knowledge, and structure unsuitable for research. Data quality issues need to be addressed throughout the data lifecycle, from data creation to use.

• **What is Already Known:** In the past, it was difficult to build and use high-quality medical data due to various obstacles such as ambiguous data definition, difficult data deployment design, complex management, special analysis tools, and environmental management. In addition, quality control of medical data has traditionally relied on human experience and basic statistical methods. Recently, several studies have explored medical quality assessment models that integrate traditional assessment methods with machine learning-based techniques using EMR data.

• **What this Paper Adds:** For more efficient quality control of medical data, we propose the following: We introduce SMART DATA as high quality data collected and managed according to the data lifecycle from data creation to use. This concept encapsulates the basic steps of a data quality management system and has been implemented using a real colorectal cancer case study. We expect to see a more robust system through collaboration with data and clinical experts.

2. Medical data quality management

Data quality management has many definitions, but generally involves the development and implementation of activities that apply quality management techniques to data to ensure that the data meets the specific needs of an organization in a given context [4,16]. Data quality is very important as it is considered to be at the core of all organizational activities. Poor data quality leads to inaccurate reporting, which leads to poor decisions and economic losses. Effective data governance requires harmonizing data from diverse sources, creating and monitoring data use policies, and eliminating inconsistencies and inaccuracies that can negatively impact the accuracy and compliance of data analysis. A data quality policy is an essential part of data governance and involves developing and establishing a defined and agreed set of rules and standards to manage all data within an organization.

Our ultimate goal is to develop a quality management system that is organically connected and has a feedback structure by understanding the data lifecycle and governance principles from creation to use of medical data.

2.1. SMART DATA

We define “SMART DATA” as systematized, high-quality data collected based on the life cycle of data construction, operation, and utilization through quality control activities for medical data (Fig. 1).

2.1.1. Level 1: Data quality guiding indicators for standardized data extraction

Quality issues in medical data include lack of standardization, missing or incomplete data, and incompatible data types and element phenotypes. One cause of data quality issues is inconsistent terminology [17–19]. Evaluating consistency by comparing extracted data with a dictionary of data models is essential for data quality control.

In this work, we present a series of examples by domain in a distributed network data validation study of electronic health data. Medication information typically includes the drug dispensed or prescribed, number of days supplied, number of units, and date dispensed or prescribed; standard coding terms include Rxnorm or NDC. As for procedure or diagnosis information, the codes include HCPCS, ICD-9-CM, and SNOMED-CT [18]. A series of item mapping-based standardization studies should be conducted during the data construction phase. Standardized methods for collecting patients’ electronic health data include the Centers for Disease Control and Prevention cancer registry data [20] and observational medical outcomes partnership common data model (OMOP CDM) [21].

The methods for data extraction from medical contents data can be divided into rule- and machine learning-based approaches. Rule-based natural language processing can be labor-intensive, requiring considerable manual work, and is not easy to reuse. Nevertheless, rule-based approaches provide reliable results when extracting complex and structured templates [22]. By contrast, deep learning model results for extracting and normalizing biomarker status from unstructured EMR data are reflected in the data construction stage [23].

Clinical data require special attention owing to the sensitivity of personal information. For anonymization, the appropriateness of representative processing methods, such as pseudonymization, totalization, data deletion, data categorization, and data masking, should be reviewed and managed by encrypting and de-identifying personal information using a monitoring system to prevent identification. The current domestic medical data collection system is based on creating registries according to the rules set by each hospital and loading them using ETL(Extract, Transform and Load) tools. Additionally, actual users can only view the metadata of the data requested from hospitals. To check the data, users must directly access a closed network with strict accessibility restrictions.

Health and medical Information Managers check the accuracy of information in the internal databases of medical institutions and ensure that patient information is up to date. In particular, their role differs from that of medical recorders as they also maintain the accuracy of information by linking it to externally provided data. Such medical information exchange further strengthens and expands the capacity to manage medical records and medical information within medical institutions. We thus recommend cooperation among **Health and medical Information Manager**s during the construction stage (Fig. 2).

2.1.2. Level 2: Data quality verification through semantic information extraction

The detailed techniques for general data quality evaluation include outlier detection, text matching, text clustering, name error detection, quality error pattern classification, and data quality scoring. This section focuses on an outlier detection method for evaluating the quality of

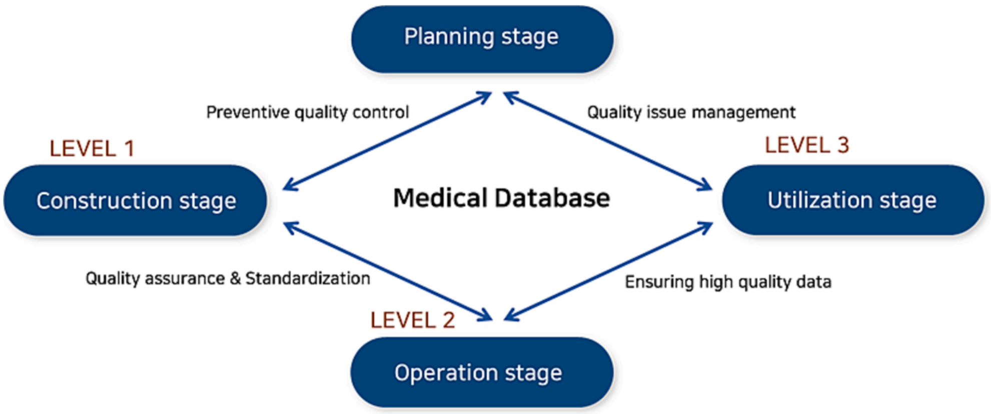


Fig. 1. SMART DATA quality management system.

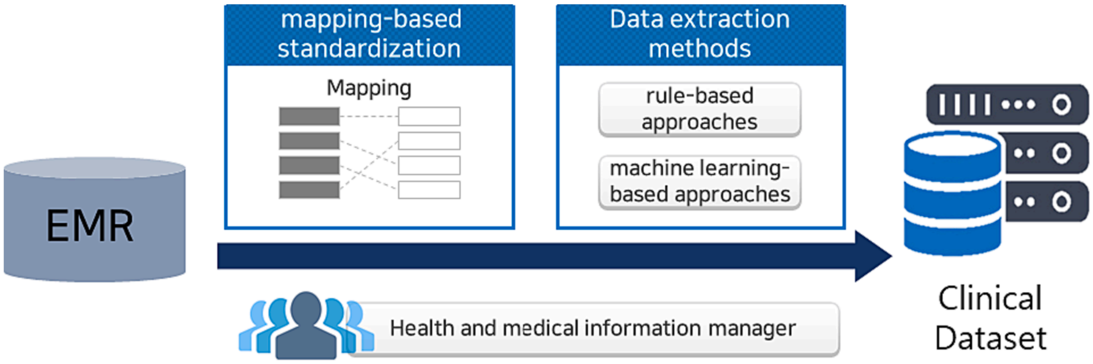

mapping-based standardization
Mapping
Data extraction methods
rule-based approaches
machine learning-based approaches
EMR
Health and medical information manager
Clinical Dataset

Fig. 2. The role of health and medical information managers.

clinical data.

The definition of an outlier may vary depending on the specific search objective; however, it generally refers to an observation or a set of observations that deviate significantly from the rest and do not align with the overall pattern or trends observed in the data [24].

The detection and handling of outliers are crucial because they can affect the representative values within the data distribution. Enhanced data reliability and accuracy can be ensured by identifying and addressing outliers. Differences may exist depending on the distribution of the data.

The most representative distribution- and density-based approaches are tested as outlier detection methods applied to medical data [25,26]. Distribution-based approach designates a specific type of statistical distribution and finds outliers with a low probability distribution. There are z-score, IQR, Grubb's test, iForest, etc. Density-based approach compares the density of local outlier elements and considers data points with large differences in density as candidates for outliers. There are LOF(Local outlier factor), k-NN, etc. Various methodologies can be employed to detect outliers, such as the FindFPOF technique, which utilizes a neural network-based approach, and the local search algorithm, which considers differences in data entropy. These approaches are effective in identifying and handling outliers in data analysis. In addition, research that defines and explores outliers using various distance concepts is worthwhile.

Evaluating the quality of an entire dataset is difficult because general outlier detection techniques aim to identify outliers [27]. However, because of the characteristics of medical data, outliers can be redefined as data with heterogeneous characteristics that reflect specific medical attributes. Thus, data with general characteristics that exclude specificity can be refined and utilized.

We screened the data detected as outliers in this manner, named them as a Critical Clinical Set (CCS), and defined them as data requiring additional clinician review. We expect to explore and refine rare traits using CCS (Fig. 3). However, we agree that collaborative research with relevant experts is essential to determine thresholds for CCS deployment.

In particular, we should pay attention to understanding data through various distance concepts. It should be accompanied by a data review procedure through the concept of semantic distance, reflecting clinical knowledge.

2.1.3. Level 3: Data quality verification based on real clinical scenarios

In this study, data on colorectal cancer patients from a clinical oncology network (CONNECT) medical institution (Konyang University Hospital) were used to select a colorectal cancer prediction model research topic with a very high CRC incidence as a scenario. The CONNECT platform is hosted by the National Cancer Center and collects standardized data from 10 medical institutions across the country.

A standard metadata system was established through word, term, and domain definitions. The quality verification of the data extraction results included a variable distribution check, outlier check, cross-check between related variables, semantic error check, variable reclassification, and filling up of missing data. Correction and additional normalization procedures were conducted for values that deviated from the extraction logic. Metadata were created by excluding sensitive information that could identify a specific individual, and each patient's join key was managed by assigning an anonymous number based on the patient number. A dashboard was created, and quality was secured by identifying the degree of conformity of the variables to understand the overall data status.

This study used k-means clustering, local outlier factor (LOF), and isolation forest (iForest) to search for CCS. In the k-Means Clustering

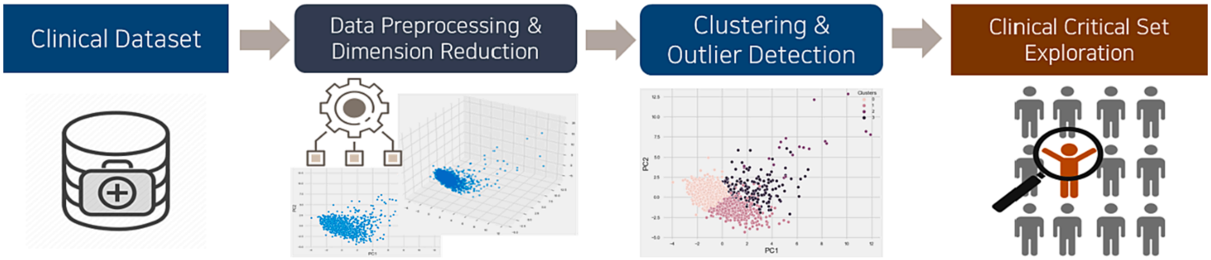

Clinical Dataset
Data Preprocessing & Dimension Reduction
Clustering & Outlier Detection
Clinical Critical Set Exploration
PC1
PC2

Fig. 3. Detection flow of clinical critical sets from clinical dataset.

approach, we divided the dataset into four clusters, considering the inherent structure and distribution of the data. Post clustering, we measured the Euclidean distance between each data point and its corresponding cluster center to discern outliers. The outlier threshold was determined statistically, set as the mean distance plus a constant multiple of the standard deviation within each cluster. Specifically, the threshold was formulated as:

$$\text{Threshold} = \text{Mean Distance} + k \times \text{Standard Deviation}$$

where k is a predetermined constant, representing the level of sensitivity in detecting outliers. Points residing at a distance greater than this threshold from their respective cluster centers were classified as outliers, providing a clear and objective criterion for outlier identification. The LOF method was applied to detect anomalies based on the density of the dataset. The number of neighbors was set to 40, reflecting the density and distribution of data points in the feature space. The contamination parameter was set at 0.01, implying that approximately 1 % of the data points were considered outliers. This value was carefully chosen after extensive experimentation, ensuring a balanced approach to outlier detection by considering the local density variations in different regions of the data space.

Lastly, the iForest method was employed, with the contamination parameter also set to 0.01, designating 1 % of the data points as outliers. This parameter was meticulously optimized to maintain a balance between recognizing true outliers and reducing the false positive rate, considering the unique characteristics and distribution of the dataset. Outlier candidate data for k-means clustering (19), iForest (18), and LOF (18) were extracted and reviewed from 3,000 colorectal cancer datasets (Fig. 4); After comparing the CCS and non-CCS groups, we observed that the CCS group exhibited significantly higher averages in clinical attributes. ALT levels were notably elevated at 199.96 (standard deviation: 272.72) compared to 22.56 (standard deviation: 24.08) in the non-CCS group. AST values averaged 53.8, whereas the non-CCS group had an average of 30.1. CA-19-9 levels reached a peak of 6147.6 in the CCS group, compared to 300.6 in the non-CCS group, and CEA levels were at 4.4, which was double the non-CCS group's average of 2.2. These disparities were further emphasized by highly significant t-test p-values, approaching zero across all the features. As expected, reflecting the CCS data set was advantageous for developing a survival prediction model for colorectal cancer patients (support vector machine score of 0.86 vs. 0.90, random forest score of 0.87 vs. 0.90, gradient boosting score of 0.87 vs. 0.91).

3. Discussion

In this study, we introduced the concept of SMART DATA in three stages within a quality management system based on the medical data life cycle, and considered key points and issues at each stage.

As in other fields, the quality of medical data must be considered throughout the entire data lifecycle, and all stages are organically interrelated. (1) Data construction: During data collection and storage, the accuracy, completeness, consistency, and validity of data must be evaluated. Issues, such as input errors, omissions, duplications, and inconsistencies, can occur during data collection, affecting the data quality. Therefore, data quality issues must be identified and addressed during data building. (2) Data operation: During regular updates and maintenance, the consistency, accuracy, and timeliness of data must be maintained. In doing so, a data quality monitoring and improvement plan must be established to identify and resolve data quality issues. (3) Data utilization: During data analysis and decision making, the validity, consistency, completeness, and reliability of data are crucial. Therefore, a data quality diagnosis must be conducted during analysis to identify and fix data quality issues. Furthermore, data security and personal information protection issues must be considered during the data utilization stage.

Medical data must be understandable from the perspective of the data producer, taking into account domain expertise such as medical databases and systems. Data operations must also develop clinical evidence-based data quality assessment models and quantitative indicators, as well as data quality management and improvement plans. And to assess usefulness from the perspective of data users, there must be evidence of clinical reliability. Therefore, for the purposes of medical data quality control, data that have undergone a clinical evidence-based data curation process are recognized as SMART DATA.

This contributes to ensuring the reliability of medical data quality assessment while minimizing the time and human cost of data validation. Research on smart data is driven by the need for a process to understand how data are produced in hospitals, determine applied quality

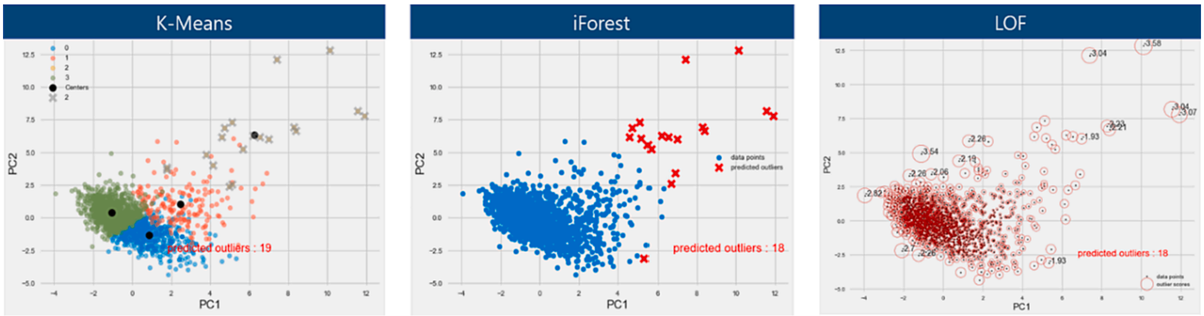


Fig. 4. CCS extraction results for colorectal cancer.

controls, and understand how de-identification leads to an analytical environment. We anticipate that medical data intelligence will improve through continued research and quality control.

Rather than a database, the institution's quality management foundation and capabilities should be evaluated [28]. Setting clear targets and planning for quality control procedures makes it possible to establish the usability and consistency of data. The most important thing in quality management is managing interoperability, which allows the integration and use of multiple data [29]. Quality management for integrated data should be implemented from the planning stage to standardize quality management and enable scalable connections between data.

For feasible data quality management, it should also be considered to present simplified data guidelines so that anyone can apply them. Simplified data guidelines should be provided for each disease, including cancer, dementia, metabolic diseases, and natural products, and quality control should be performed through guidelines.

For example, a complex disease such as cancer requires planning and presentation of a cancer disease management strategy at the national level, with quality control of data through a guideline-based cancer registry [30]. In addition, guidelines are being proposed for medical devices, new drugs, and biologic data, which are exponentially growing and difficult to manage and explain, and there is a need for guidelines that cover all diseases [31].

Medical data quality management must be coordinated with lifecycle governance aspects. Most quality control focuses on finding outliers in the data [32]. However, the types of data are very diverse. There is a need to develop more diverse data quality control methods so that quality control for various data can be actively performed [33]. You also need to think about meaning beyond what you're looking for in outliers. We wanted to consider a detailed view of outliers through the CCS concept.

In this study, we curated 1,724 learning data and 27 CCS data for predictive modeling of colorectal cancer, following a step-by-step process to transform raw data into "smart" data. A deep dive into the CCS concept, we have gained a more nuanced understanding of outliers. Moreover, by highlighting the clear disparities between the CCS and non-CCS groups across multiple features, our analysis underscores the significance of the CCS detection mechanism. This suggests that the CCS group represents patients with pronounced or specialized clinical attributes. Our observations, if carefully incorporated into clinical opinion, may improve our understanding of the various factors associated with the development of colorectal cancer and contribute to the risk assessment of individual patients and the development of individualized treatment strategies. We expect that this CCS detection will improve the development of learning models and provide the ability to proactively select and filter outliers in the dataset by continuously monitoring updated datasets that share CCS characteristics.

This study has several limitations. First, SMART DATA consists of steps with only the most simplified essential elements of a medical data quality management system, and each detailed step must be customized and elaborated in the future. Such effort will contribute to building a system for managing, maintaining, and repairing medical data outputs based on a platform infrastructure, including detailed tasks that may occur during the actual medical data life cycle (e.g., defining patterns of extracted items, storing pattern extraction results, supplementing verification results, and enhancing security). Second, we conducted a case study using data from a specific disease group at a single institution. We plan to apply this framework to the medical data of various diseases and institutions for verification and advancement. Third, we prioritized the inclusion of widely used methodologies. We expect this work to lay the foundation for a successful medical data quality assessment framework by reviewing various newly developed techniques for data verification, applying interactive verification algorithms that can minimize errors, and establishing criteria and procedures for error correction and improvement. Finally, although we attempted to reflect the opinions of practitioners in the medical field, the result was insufficient. Discovering and reflecting on more customization points by reflecting on the views of various practitioners may lead to a refined medical data quality assessment framework.

We anticipate that the discussion of how healthcare data quality management systems, with SMART DATA as the underlying concept, can be extended to larger datasets or adapted to different types of healthcare data beyond electronic health records will provide insight into the diversity of this diversity. In addition, we look forward to potential future updates to hospital systems, integration with emerging technologies (e.g., AI or blockchain), and evolution to respond to the changing healthcare data management environment. This allows us to provide a more holistic and practical perspective on SMART DATA and its applicability to medical data quality management systems.

4. Conclusion

In this study, SMART DATA is considered as an efficient method of a data life cycle-based medical data quality management system from data creation to utilization to improve medical data utilization. We have investigated major issues in data quality management through research on SMART DATA creation using EMR data. In the future, we expect to contribute to the establishment of a medical data quality management system by exploring medical data quality management methods through research to advance and systematize quality control procedures considering clinical field situations. Ultimately, this will create more opportunities for use by providing high-quality medical data.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We would like to thank Editage (www.editage.co.kr) for editing and reviewing this manuscript for English language.

Funding

This work was supported by the National Research Foundation of Korea (NRF) grant funded py the Korea government (No. RS-2023-00263735129820768701 01). The Korean Society for Medical Informatics supported this study.

References

[1] N.G. Weiskopf, C. Weng, Methods and dimensions of electronic health record data quality assessment: Enabling reuse for clinical research, J. Am. Med. Inform. Assoc. 20 (1) (2013) 144–151.
[2] S.T. Liaw, et al., Quality assessment of real-world data repositories across the data life cycle: A literature review, J. Am. Med. Inform. Assoc. 28 (7) (2021) 1591–1599.
[3] B.N.L. Guide, Capturing high quality electronic health records data to support performance improvement, Implementation Objective 2 (2013) 16.
[4] N.G. Weiskopf, et al., A data quality assessment guideline for electronic health record data reuse, EGEMS (wash DC) 5 (1) (2017) 14.
[5] M.-s. Choi, S. Lee, Current status and issues of data management plan in Korea, J. Korea Contents Association 20 (6) (2020) 220–229.
[6] I.H. Sarker, Machine learning: Algorithms, real-world applications and research directions, SN Computer Science 2 (3) (2021) 1–21.
[7] E. Tute, I. Scheffner, M. Marschollek, A method for interoperable knowledge-based data quality assessment, BMC Med. Inf. Decis. Making 21 (1) (2021) 93.
[8] N.G. Weiskopf, C. Weng, Methods and dimensions of electronic health record data quality assessment: Enabling reuse for clinical research, J. Am. Med. Inform. Assoc. 20 (1) (2013) 144–151.
[9] C. Xiao, E. Choi, J. Sun, Opportunities and challenges in developing deep learning models using electronic health records data: A systematic review, J. Am. Med. Inform. Assoc. 25 (10) (2018) 1419–1428.

[10] H. Chen, et al., Application of a four-dimensional framework to evaluate the quality of the HIV/AIDS data collection process in China, Int. J. Med. Inf. 145 (2021) 104306.

[11] M.M. Goez, et al., Joint pre-processing framework for two-dimensional gel electrophoresis images based on nonlinear filtering, background correction and normalization techniques, BMC Bioinf. 21 (1) (2020) 376.

[12] S. Henley-Smith, D. Boyle, K. Gray, Improving a secondary use health data warehouse: Proposing a multi-level data quality framework, EGEMS (wash DC) 7 (1) (2019) 38.

[13] Z. Wang, et al., A rule-based data quality assessment system for electronic health record data, Appl Clin Inform 11 (4) (2020) 622–634.

[14] Y. Chen, Data quality assessment methodology for improved prognostics modeling, University of Cincinnati, 2012.

[15] I. Taleb, et al., Big data quality framework: A holistic approach to continuous quality management, Journal of Big Data 8 (1) (2021) 1–41.

[16] C. Cichy, S. Rass, An overview of data quality frameworks, IEEE Access 7 (2019) 24634–24648.

[17] S.-T. Liaw, et al., Towards an ontology for data quality in integrated chronic disease management: A realist review of the literature, Int. J. Med. Inf. 82 (1) (2013) 10–24.

[18] J. Brown, M. Kahn, S. Toh, Data quality assessment for comparative effectiveness research in distributed data networks, Med. Care 51.8 0(3) (2013) S22.

[19] Nicole Gray Weiskopf, Chunhua Weng, Methods and dimensions of electronic health record data quality assessment: Enabling reuse for clinical research, J. Am. Med. Inform. Assoc. 20 (1) (2013) 144–151.

[20] Centers for Disease Control and Prevention. Public Health Registry, Centers for Disease Control and Prevention, Office of Public Health Data, Surveillance, and Technology, November 8 2019, https://www.cdc.gov/datainteroperability/archive/specialized_registry.html.

[21] Franziska. Bathelt, The usage of OHDSI OMOP–a scoping review, Proceedings of the German Medical Data Sciences (GMDS) 95 (2021) 9520.

[22] A. Mykowiecka, M. Marciniak, A. Kupść, Rule-based information extraction from patients' clinical data, J. Biomed. Inform. 42 (5) (2009) 923–936.

[23] B. Zhao, Clinical data extraction and normalization of cyrillic electronic health records via deep-learning natural language processing, JCO Clinical Cancer Informatics 3 (2019) 1–9.

[24] V. Barnett T. Lewis Outliers in Statistical Data Vol. 3. No. 1 1994 Wiley New York.

[25] V.C. Pezoulas, et al., Medical data quality assessment: On the development of an automated framework for medical data curation, Comput. Biol. Med. 107 (2019) 270–283.

[26] Cheng Zhangyu Chengming Zou Jianwei Dong Outlier detection using isolation forest and local outlier factor Proceedings of the conference on research in adaptive and convergent systems. 2019.

[27] J. Lee, Y. Kim, K.M. Lee, A data quality evaluation method using the salient degree, Proceedings of KIIS Spring Conference 25 (1) (2015) 121–122.

[28] Agency, N.I.S. Public Data Quality Management Manual ver 2.0, National Information Society Agency: Daegu. 2018.

[29] Health, N.I.o. Final NIH policy for data management and sharing, N.G.F.O.o.T . Director, National Institutes of Health 2020 Bethesda, MD.

[30] Gupta, Sumitet, et al., Paediatric cancer stage in population-based cancer registries: the Toronto consensus principles and guidelines, The Lancet Oncology 17 (4) (2016) e163–e172, https://doi.org/10.1016/S1470-2045(15)00539-2.

[31] B. Lee S. Hwang P.G. Kim G. Ko K. Jang S. Kim J.H. Kim J. Jeon H. Kim J. Jung B.H. Yoon I. Byeon I. Jang W. Song J. Choi S.Y. Kim Introduction of the Korea BioData Station (K-BDS) for sharing biological data. Genomics Inform 21 1 2023 10.5808/gi.22073 e12 Epub 2023 Mar 31. PMID: 37037470; PMCID: PMC10085736.

[32] T. Kwon Y. Jeong D. Lee. Standardization and quality evaluation of health and medical big data, in Health Industry Brief, D.-c. Kwon, Editor, Korea Health Industry Development Institute: Osong. 2019.

[33] R.R. Colditz, et al., TiSeG: A flexible software tool for time-series generation of MODIS data utilizing the quality assessment science data set, IEEE Trans. Geosci. Remote Sens. 46 (10) (2008) 3296–3308.