Metadata as a Key Driver in Healthcare Data Analysis

(Discussion paper)

Chiara Criscuolo[†], Davide Piantella[*,†], Pierluigi Reali[†], Maria Gabriella Signorini and Letizia Tanca

Politecnico di Milano - Department of Electronics, Information, and Bioengineering
Via G. Ponzio 34/5, 20133 Milano, Italy

Abstract

In medicine, the digitization of healthcare processes and health services is generating an incredible amount of medical data. However, the huge data volume and variety of formats significantly impact the efficient sharing of data collected across different hospitals. This could compromise the quality of multicentric studies and hamper the potentiality of modern medical research through AI-based systems and machine learning analysis. In this context, being able to extract and manage good-quality metadata is paramount, since, especially when dealing with heterogeneous and unstructured datasets, metadata provides valuable ready-to-use information regarding the dataset without the need to directly analyze its content. Several data models exist that are specific for storing and conveniently organizing clinical metadata, such as the Observational Medical Outcomes Partnership Common Data Model (OMOP CDM), providing a flexible solution for multiple types of healthcare data. Furthermore, being compliant with the EU AI Act is a necessary requirement for medical AI-systems, thus metadata can also support ethical data science. The role of metadata in clinical contexts has been studied and analyzed in the Health Big Data project, whose goal is involving 51 Italian research hospitals (IRCCS) to maximize the interoperability of healthcare datasets and enhance clinical research. In this discussion paper, we describe how effective management of metadata in clinical datasets is crucial for ensuring data usability, harmonization, and ethics in AI-driven healthcare applications.

Keywords

Healthcare, Data science, Metadata, Big Data, Ethics

1. Introduction

The digitization of healthcare processes has led to an explosion of medical data, with the availability of large datasets allowing researchers to leverage federated architectures and applying innovative analytical techniques. However, this transformation is occurring in an environment of uncertainty and rapid change, where current decisions will shape the future of healthcare data management and analysis. In addition, the diversity of data formats and huge volume pose significant challenges for efficient data sharing across hospitals. These barriers can compromise the quality of multicentric studies and limit the potential of AI-driven medical research. Moreover, as automation in data collection and analysis increases—along with the capability of identifying large-scale patterns in biomedical data—it becomes crucial to question which systems govern these processes and how they are regulated.

In this context, metadata plays a key role in managing heterogeneous and unstructured datasets, offering valuable, ready-to-use information without requiring direct content analysis. In addition, ensuring compliance with regulations such as the European *AI Act* [1] is essential for promoting ethical and trustworthy AI-driven applications in healthcare. Addressing these challenges, the Health Big Data project, involving 51 Italian research hospitals, aims to enhance healthcare data interoperability and improve clinical research through effective metadata management.

SEBD 2025: 33rd Symposium on Advanced Database Systems, June 16-19, 2025, Ischia, Italy
*Corresponding author.
†These authors contributed equally.
chiara.criscuolo@polimi.it (C. Criscuolo); davide.piantella@polimi.it (D. Piantella); pierluigi.reali@polimi.it (P. Reali); mariagabriella.signorini@polimi.it (M. G. Signorini); letizia.tanca@polimi.it (L. Tanca)
0000-0002-1345-2482 (C. Criscuolo); 0000-0003-1542-0326 (D. Piantella); 0000-0003-3041-4004 (P. Reali); 0000-0002-9391-9846 (M. G. Signorini); 0000-0003-2607-3171 (L. Tanca)

1.1. Contributions and Paper Structure

In this paper, we provided three key contributions that highlight metadata as a fundamental driver in healthcare data analysis:

- *C1* We describe the role of metadata in managing complex, heterogeneous, and unstructured healthcare data.
- *C2* Considering a real-world use case, we apply a well-established metadata framework (i.e., the Observational Medical Outcome Partnership Common Data Model, OMOP CDM [2]) to enhance electrocardiogram (ECG) data collection and analysis by identifying relevant metadata and proposing a workflow for their automatic extraction from raw data.
- *C3* We illustrate how metadata supports ethical data science in the medical domain, ensuring that healthcare datasets and systems remain trustworthy and regulation-compliant.

The rest of this paper is organized as follows. Section 2 introduces some preliminary concepts such as Big Data, Electronic Health Records, and metadata classifications. Section 3 presents our case study: considering ECG data, we identify relevant metadata and present a workflow for automatically extracting them from raw datasets. Section 4 describes the role of metadata in advancing ethical data science in the healthcare domain. Section 5 concludes the paper and suggests novel directions for future research.

2. Preliminaries

For the scope of this paper, some preliminary concepts must be defined. We regard *Big Data* both as a technological and cultural phenomenon since *(i)* it necessitates considerable computational power and algorithmic precision to collect, analyze, link, and compare vast datasets [3], and *(ii)* it allows for uncovering patterns that inform decisions in various fields, including economics, social sciences, technology, and, more specifically, healthcare and medicine.

The "4 Vs" of Big Data – Velocity, Volume, Variety, and Veracity – represent the core characteristics of this phenomenon. Many researchers extended the traditional "Vs" with additional features [4, 5, 6, 7], such as: Complexity, Value, and Variability. Although not traditionally included in the definition of Big Data, these concepts emphasize even more the importance of effective data management for maximizing the intrinsic value of data. In healthcare, this factor is particularly critical, as the variability of healthcare data can impact the quality and reliability of analyses. In fact, Big Data can be viewed as a combination of structured (e.g., relational or tabular), semi-structured (e.g., XML or JSON), and unstructured data (e.g., natural language texts, raw images), or any combination thereof, collected by organizations [8].

In this context, the OMOP CDM offers powerful tools to standardize the structure and representation of data from different sources [2]. Its main components comprise a set of standardized *tables* (with associated predefined fields) and interconnected *vocabularies*. The tables store the clinical data and related metadata of interest; the vocabularies provide coded concepts to identify the described measures, diagnoses, procedures, etc.

Once datasets are collected, they typically require preparation to align with the specific research goals. This preparation includes cleaning and preprocessing the datasets, making them ready for analysis and suitable for machine learning or AI-based applications. These steps constitute the so-called *Data Science Pipeline* [9]. In healthcare research, this pipeline presents unique challenges, such as the variety of datasets and their formats, prompting additional steps for the integration or harmonization of different data sources. Moreover, in the pipeline, it is essential to not only consider data quality, but also to address ethical issues, including the identification of potential biases related to labels, representation, and sampling [10] and, possibly, their mitigation [11].

2.1. Managing Electronic Health Records (EHRs)

The increasing digitization of healthcare processes results in a huge amount of medical data, such as clinical images, laboratory results, and discharge letters. It has been estimated that by 2025 the annual

growth rate of healthcare data will reach 36%, significantly higher than the general data growth rate, estimated at 27% [12].

The term Electronic Health Records (EHRs) defines a comprehensive, cross-institutional, and longitudinal collection of healthcare data to encompass the entire clinical history of a patient [13]. EHRs store both structured (e.g., date of birth, diagnosis codes, and laboratory results) and unstructured (e.g., clinical notes, and medical images) data. While unstructured data are more challenging to manage and extract information from, they offer a richer description of patient conditions and valuable contextual information that structured formats struggle to capture (e.g., social history, anamnesis).

One possible solution to enhance data analysis capabilities when dealing with unstructured heterogeneous data is to leverage *metadata*, i.e., information describing the different characteristics of the data itself. For instance, metadata for natural language texts may specify the language used and the topics discussed. In the case of a medical image, metadata should include the scanned body region, type and configuration of the imaging device. Metadata can be associated with different levels of information, from a single data point (e.g., the patient's age in an ECG signal) to entire datasets (e.g., the age range and ethnicity of patients in a multicentric clinical trial) or even a whole data provider (e.g., its trustworthiness and other quality metrics).

The increasing availability of EHRs has enabled *(i)* real-world-evidence clinical trials [14], *(ii)* the use of deep learning algorithms for advancing healthcare data—particularly, medical images—analysis [15], and *(iii)* the generative-AI application for natural language processing tasks (i.e., reading comprehension, summarization, translation, and question answering) [16], as well as real-world risk forecasting and clinical research to study disease progression, simulate interventions, and support medical education [17]. Thus, the ability to collect, harmonize, and integrate data from multiple heterogeneous datasets becomes paramount [18]. A common approach for storing unstructured datasets in heterogeneous formats is the use of *data lakes*: schema-less data repositories capable of ingesting raw data without preprocessing [19]. However, to exploit the full potential of data lakes and EHRs, we must properly describe and catalog datasets in a structured and harmonized manner: metadata are crucial for outlining raw datasets and establishing meaningful connections among them [20].

2.2. Metadata classifications

Analyzing some of the most relevant general-purpose metadata classifications [21, 22, 23, 24], we can identify three main clusters of metadata categories: *(i) administrative*, supporting data governance, management and administration; *(ii) data provenance*, providing a detailed record of the dataset's lineage and evolution throughout its lifecycle; *(iii) descriptive*, allowing the users to understand the content, purpose and relevance of the datasets.

Additionally, healthcare data management requires specialized metadata categories to ensure accurate description, privacy protection, and interoperability. We now present two healthcare-specific metadata models that align with the general-purpose classifications while addressing domain-specific requirements. Pierson et al. [25] introduce a classification designed for handling medical data, identifying six primary metadata categories:

- *Patient-related:* these metadata include simple information regarding the patient (e.g., sex, age).
- *Image-related:* medical images are often characterized by dimensions, voxel size, and encoding.
- *Acquisition-related:* decisions taken during the acquisition process significantly impact the resulting medical images. For this reason, we store information such as the acquisition device, the set of parameters for the acquisition process, and the acquisition date.
- *Hospital-related:* this category includes the department responsible for the acquisition and hospital information in general.
- *Medical record:* medical history is fundamental for interpreting medical exam results, which are often compared with previous exams of the same patient.
- *Security-related:* sensitive information must be kept private. Therefore, we must store information regarding authorization and encryption.

Similarly, Badawy et al. [26] report the following classification of healthcare metadata, endorsed by a consortium of 33 domain experts:

- *Person-related:* this category includes relevant information regarding the subjects of the study (i.e., the patients), such as age, sex, medical history, and concomitant medications. When applicable, it also includes information regarding care providers (e.g., clinicians or relatives) to avoid biases.
- *Observation-related:* this category includes data collected during a study or analyzed in post-study evaluations, integrating information from both digital health technologies and human participants. Examples are devices used for acquisition, software names and versions used for the analysis, and sensor precision.
- *Context of collection:* they include details of the clinical study, conduct, eligibility criteria, processes, and procedures.
- *Time-related:* they provide temporal information regarding the data collected and reported in the datasets, such as start time, end time, time precision, time format, and time zone.

3. Case study: metadata for ECG datasets

As a case study, we focus on electrocardiograms (ECGs), biosignals that record the heart's electrical activity over time. We identify a set of relevant metadata to describe this specific type of dataset and propose a workflow to automatically extract these metadata from raw ECG data.

3.1. Minimum set of metadata

When we have multiple data providers (e.g., in a multicentric study), a best practice consists in identifying a minimum set of metadata that each source must attach to the dataset before sharing it with other partners. This will ensure that essential metadata are present for each dataset, leaving the possibility for data feeders to specify additional metadata that further describes their datasets.

To properly represent ECG data, we extend the healthcare minimum metadataset illustrated in [27] with additional metadata, specifically tailored to describe an ECG, shown in bold font in Table 1.

Category	Attributes
Administrative	GUID, Creator, Owner, Rights, Terms of access
Data provenance	Publication year, Upload date, Acquisition method, Acquisition tools, **Number of leads, Sampling frequency, Bandwidth**, Download URL, Checksum, Encryption algorithm, File version, Update/modification date, Update frequency
Descriptive	File description, File format, Age, Ethnicity, Sex, Blood group, Disease name, **Heart Rate Variability indices, Arrhythmias**

Table 1
Minimum set of metadata for ECG data

Specifically, the *Number of leads* defines the number of ECG channels, each capturing the electrical activity along a different direction and enriching the information content of the signal. The *Sampling frequency* describes the temporal resolution of the ECG, while the *Bandwidth* informs about the lowest and highest frequencies represented in the signal; these metadata are fundamental because certain analyses necessitate high-resolution ECGs [28]. *Heart Rate Variability* (HRV) *indices* are time-domain, frequency-domain, and non-linear metrics derived from ECG signals through appropriate processing [29]; together with automatically detected *Arrhythmias*, they can be queried to identify signals presenting specific characteristics of interest for a clinical study.

Furthermore, as concerns ECG signals, the *Acquisition method* is used to differentiate 2-minute or shorter recordings (e.g., diagnostic ECGs) from 24-hour or longer ones (Holter ECGs), providing essential context to interpret the extracted HRV indices correctly [30].

3.2. Extraction and representation of ECG-specific metadata

Common ECG data formats (e.g., EDF [31], BIDS [32], CSV + header file or JSON) typically store, in addition to raw signals, also several provenance and descriptive metadata in a structured manner, facilitating their identification and retrieval. However, *HRV indices* and *Arrhythmias* must be derived from the ECG traces through specific processing. To enhance interpretability and semantic interoperability with datasets from other sources, we insert these features into OMOP CDM structures and map them to unique concepts, leveraging OMOP vocabularies. This is achieved through a data processing and standardization pipeline we developed [33] that takes raw ECGs as input, extracts HRV features and the detected arrhythmias, and structures the output according to the OMOP CDM. Suitable concepts to describe the metadata of interest in the OMOP Vocabulary are identified through the Athena web interface[1]. In the following, we summarize the main steps of this pipeline.

An ECG signal is made up of characteristic waveforms representing different phases of the cardiac cycle, as shown in Figure 1. Both the time distance (specifically, its variability over time) between such waveforms and their morphology can provide critical information on patient health. After traditional preprocessing, established algorithms (e.g., Pan-Tompkins [34]) are applied to detect the position of the R peaks in the ECG traces, obtain the RR-interval time series, and calculate typical HRV indices, including the RR-interval mean, standard deviation, and root mean square of successive differences (RMSSD) [29]. In addition, waveform classification models are employed to automatically identify the occurrence of arrhythmias of particular interest for clinical research, including atrioventricular blocks, bundle branch blockades, atrial fibrillation, bradycardia, and tachycardia.

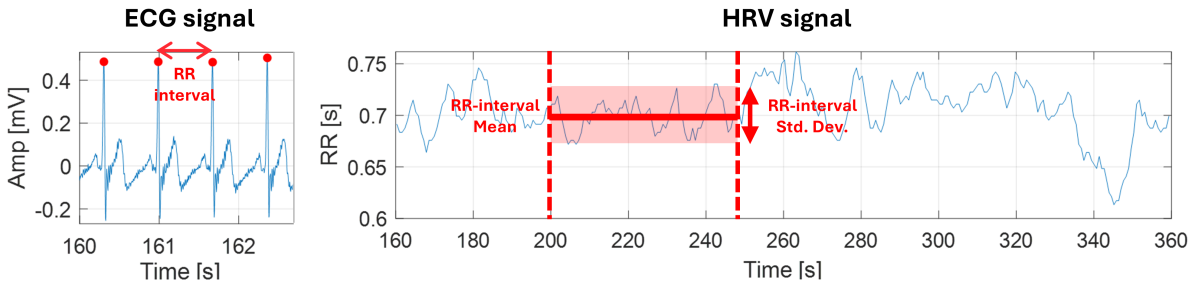

Figure 1: An example of ECG and derived HRV signal and features.

Open-source tools, such as the *Neurokit2* toolbox [35] and specialized Deep Neural Networks (DNN) [36], can be leveraged for this purpose, with the additional benefit of favoring tracking of software version, code, and characteristics of the applied feature extraction methods.

Once feature extraction is completed, HRV indices and automated diagnoses are mapped into a suitable OMOP CDM structure. Given the patient-centric approach of the CDM, the ***Person*** table is the first one to be populated, which can conveniently accommodate the *Descriptive* metadata related to patient demographics (i.e., Age, Ethnicity, Sex) by means of dedicated fields[2]. Then, a ***Procedure_occurrence*** table is initialized that stores the basic properties of the collected ECG signals. For example, a specific field (`procedure_type_concept_id`) stores the previously described *Acquisition method* metadata, allowing for differentiating between diagnostic and Holter recordings. The other ECG-tailored *Data provenance* metadata, namely *Number of leads*, *Sampling frequency*, and *Bandwidth*, can be allocated in distinct instances of the ***Observation*** table, each one mapped to a specific OMOP Vocabulary concept ensuring unambiguous representation (e.g., *Sampling frequency* maps to "Digital Sampling Rate", OMOP ID: 37533243).

The ***Observation*** table is also used to store the automatically detected arrhythmias, which are represented through the `observation_concept_id` "ECG automated diagnosis" (OMOP ID: 35810893), defining the specific metadata reported in that instance, and the `value_as_concept_id` field populated

[1] https://athena.ohdsi.org/search-terms/start
[2] https://ohdsi.github.io/CommonDataModel/cdm54.html

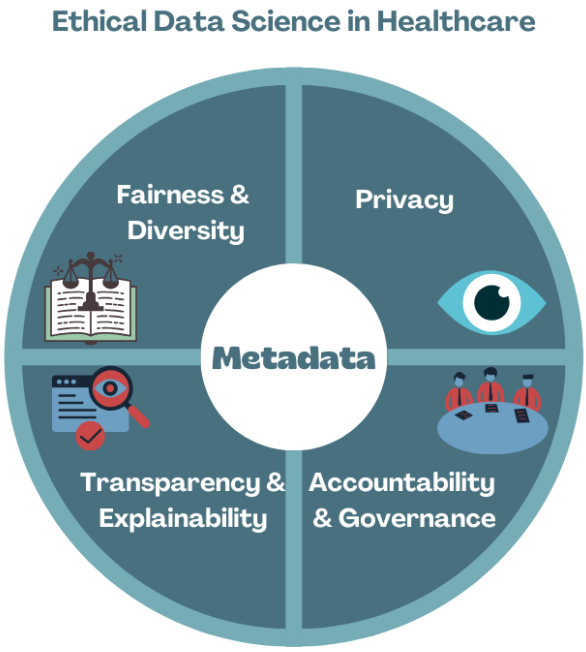

Figure 2: Ethical Data Science: the four key areas

with the arrhythmia (if any) identified in the ECG (e.g., "ECG: atrial fibrillation", OMOP ID: 4064452). With a similar strategy, the calculated HRV indices are allocated in separate instances of the ***Measurement*** table, which is preferred for storing numerical values. All the instances of the ***Observation*** and ***Measurement*** tables are connected to the previous ***Procedure_occurrence*** table, so the association between the mapped metadata and the ECG of origin is clearly defined.

Finally, the remaining patient-related *Descriptive* metadata, i.e., *Blood group* and *Disease name*, can be represented as records of the ***Condition_occurrence*** table, which is directly linked to the Person table and primarily stores patient diagnoses (also associated with a specific period of time), as well as generally immutable facts (e.g., blood group). Certain patient characteristics included in the designed OMOP CDM structure, such as age, sex, and ethnicity, demand careful consideration from an ethical perspective.

4. Metadata supporting Ethical Data Science

Healthcare data, such as clinical trials, are well-suited for analysis through mining methods and machine learning techniques. However, the European *AI Act* [1] document classifies the application of Data Science to healthcare data for medical diagnosis as a high-risk application due to its significant impact on patients' lives. Consequently, such systems must necessarily undergo rigorous compliance checks and adhere to the principles outlined in the document. Among the principles listed in the *AI Act*, we first report the three pillars required to achieve trustworthy AI systems[3]:

- **Lawful**: systems must respect all the applicable laws and regulations, and in the medical domain should be also compliant with domain-specific rules.
- **Ethical**: systems should incorporate ethical principles and values.
- **Robust**: systems must be designed and used to prevent any unintentional harm from both technical and social perspectives.

Ethical considerations are fundamental for ensuring that a system is trustworthy and, thus, applicable in real-world contexts. *Ethical Data Science* is a broad field that focuses on minimizing harm, ensuring

[3]https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

moral rights, and evaluating practices involved in the generation, collection, analysis, and dissemination of data that could potentially affect people and society adversely. By identifying specific risks and challenges, Figure 2 categorizes ethical issues and concerns in the following areas: fairness and diversity, privacy, transparency and explainability, accountability and governance. We believe that metadata can play a crucial role in enhancing many of these ethical aspects.

Fairness and Diversity. In the context of healthcare and Machine Learning (ML) tasks, fairness refers to *the equitable treatment of individuals or groups by ML models, irrespective of sensitive or protected characteristics.* To identify and measure fairness, analysts require access to protected characteristics, such as sex, ethnicity, or age, which should not serve as discriminatory factors in predictions. As previously shown, these attributes can be recorded as metadata within the dataset. Additionally, metadata can capture the distribution of these characteristics by reporting the percentage of occurrences for each value. This enables *diversity* analysis, helping to identify representation inequities that could lead to systemic biases in data-driven decisions. Such biases may result in greater prediction errors for underrepresented groups or minorities, ultimately exacerbating disparities in healthcare outcomes.

Privacy. Privacy must be guaranteed through the entire AI system lifecycle, ensuring that personal information is protected from unauthorized users, and that data usage is traceable at every stage of the data science pipeline (i.e., monitor data flow, track changes in the data transformation/processing, identify who can access or modify the data). This is particularly critical in the healthcare sector, where *informed consent* plays a central role in clinical trials. Informed consent ensures that participants are fully aware of the experiment, voluntarily agree to participate, and retain the right to withdraw their consent at any time. Metadata can significantly enhance privacy in AI systems by indicating whether a dataset contains personally identifiable information and by tracing the history of privacy-preserving transformations applied to the records [37, 38]. An additional metadata usage in this area could be a log tracking consent status, ensuring researchers only use data from patients who have given informed consent.

Transparency and Explainability. Transparency and explainability are essential for fostering trust in AI systems. When users understand how algorithmic decisions are made, they are more likely to trust and accept the outcomes. Additionally, regulations such as GDPR and the AI Act emphasize the need for transparency in AI applications. Transparency and explainability ensure that the appropriate information reaches the relevant stakeholders [39]. These principles are particularly relevant in medical diagnosis. For instance, if an AI system is used to screen patients at high risk for cancer, a medical researcher needs to understand the factors contributing to the diagnosis. Similarly, when an AI system predicts a particular medical condition, it is crucial that its data sources, analytical processes, and decision-making logic are well-documented and accessible. Metadata can play a key role in enhancing transparency and explainability by providing detailed records of AI system functionalities, data sources, and decision rationales. Metadata could contain model interpretability metrics, ensuring that predictions are explainable to medical experts. Furthermore, metadata can be tailored to different audiences, ensuring that explanations are adapted to various levels of expertise and backgrounds, thereby making AI systems more accessible and interpretable.

Accountability and Governance. Can an AI system be held accountable for its actions? Accountability refers to the responsibility of individuals and organizations to ensure that their data processes and algorithms operate ethically and fairly, preventing harm and addressing ethical lapses when they arise. Through effective governance, accountability should be maintained at every stage of the AI system lifecycle or data science pipeline. However, assigning responsibility to specific actors within an AI system is inherently challenging. In the healthcare sector, ethical review boards typically oversee data science practices to ensure responsible management and adherence to ethical standards. Metadata can also play a crucial role in strengthening governance. Metadata can document the version history of an ML model, tracking changes in training data, algorithm updates, and human interventions. By recording assessments of algorithms, data, and design processes, metadata facilitates auditability and enhances oversight. Additionally, it can document potential redress strategies for addressing issues in data science methodologies, further supporting responsible AI development.

To summarize, the metadata contributions to ethical data science in healthcare on the four areas are:

- **Identification of bias and supporting fairness analysis**: *(i)* in bias detection metadata can record demographic attributes (e.g., Age, Ethnicity, Sex) to assess and mitigate bias in AI models; *(ii)* in data distribution insights they can store statistical summaries of dataset diversity to prevent representation inequities; *(iii)* metadata can be used to report fairness metrics (i.e. fairness-related performance indicators).
- **Tracking privacy**: *(i)* metadata can identify personal, anonymized, or sensitive health data to ensure proper handling; *(ii)* regarding consent management, metadata can log patient consent status and usage permissions, ensuring compliance with regulations; *(iii)* metadata can record privacy-preserving actions (e.g., encryption, de-identification) applied to data.
- **Documenting transparency**: *(i)* metadata can document how an AI system arrives at a specific diagnosis or prediction; *(ii)* regarding model interpretability, metadata can store explanations of model behavior and reasoning for end-users; *(iii)* metadata can adapt explanations based on the user's expertise (e.g., doctors vs. patients).
- **Enabling accountability**: *(i)* in audit trails, metadata can capture version histories, documenting changes in data, models, and decisions; *(ii)* metadata can help compliance by marking datasets and models that meet ethical and legal standards; *(iii)* metadata can log incorrect predictions and associated corrective actions to improve error and redress tracking.

5. Conclusions and Future Work

In this work, we analyzed the role of metadata as a key driver in healthcare data management and analysis. We demonstrated how metadata facilitates the handling of complex, heterogeneous, and unstructured healthcare data by providing a precise description of metadata in clinical contexts, including a dedicated metadata set for ECG data. Additionally, we applied a well-established metadata framework (OMOP CDM) to enhance real-world use cases, specifically focusing on ECG data collection and analysis. Finally, we explored the role of metadata in ethical data science within the medical domain, particularly in bias detection and fairness analysis, privacy tracking, transparency augmentation, and accountability enablement. Future work consists in:

- Studying how to reduce representation ambiguities in metadata values. This could be achieved by enriching the OMOP CDM vocabulary with a multi-language compendium of medical ontologies such as Unified Medical Language System [40].
- Validate our ECG metadata representation on a real-world multicentric study by applying it to healthcare datasets from multiple sources and assessing its impact on data interoperability and AI-driven analysis.
- Expand the state-of-the-art metadata frameworks to encompass additional aspects of ethical data science and validate the approach in real-world scenarios.

Acknowledgments

This work has been supported by the Health Big Data project, funded by the Italian Ministry of Economy and Finance and coordinated by the Ministry of Health. We also thank Davide Martinenghi for his support and advice.

Declaration on Generative AI

During the preparation of this work, the authors used GPT-4 for: grammar and spelling check. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] The European Parliament, Artificial Intelligence Act, Official Journal of the European Union, 2024. URL: https://artificialintelligenceact.eu/the-act/, last accessed on January 30, 2025.

[2] J. M. Overhage, P. B. Ryan, C. G. Reich, A. G. Hartzema, P. E. Stang, Validation of a common data model for active safety surveillance research, Journal of the American Medical Informatics Association 19 (2012) 54–60. doi:10.1136/amiajnl-2011-000376.

[3] D. Boyd, K. Crawford, Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon, Information, communication & society 15 (2012) 662–679.

[4] H. Liu, A. Gegov, M. Cocea, Unified framework for control of machine learning tasks towards effective and efficient processing of big data, Data science and big data: An environment of computational intelligence (2017) 123–140.

[5] A. De Mauro, M. Greco, M. Grimaldi, What is big data? a consensual definition and a review of key research topics, AIP Conference Proceedings 1644 (2015) 97–104.

[6] S. Suthaharan, Big data classification: Problems and challenges in network intrusion prediction with machine learning, ACM SIGMETRICS Performance Evaluation Review 41 (2014) 70–73.

[7] A. Katal, M. Wazid, R. H. Goudar, Big data: issues, challenges, tools and good practices, in: 2013 Sixth international conference on contemporary computing (IC3), IEEE, 2013, pp. 404–409.

[8] W. Pedrycz, S.-M. Chen, et al., Information granularity, big data, and computational intelligence, Springer, 2015.

[9] W. Wang, J. Gao, M. Zhang, S. Wang, G. Chen, T. K. Ng, B. C. Ooi, J. Shao, M. Reyad, Rafiki: Machine learning as an analytics service system, Proceedings of the VLDB Endowment 12 (2018).

[10] C. Criscuolo, T. Dolci, M. Salnitri, Towards assessing data bias in clinical trials, in: VLDB Workshop on Data Management and Analytics for Medicine and Healthcare, Springer, 2022, pp. 57–74.

[11] C. Criscuolo, T. Dolci, M. Salnitri, Mitigating unfairness in machine learning: A taxonomy and an evaluation pipeline (2024).

[12] D. R.-J. G.-J. Rydning, J. Reinsel, J. Gantz, The digitization of the world from edge to core, Framingham: International Data Corporation 16 (2018) 1–28.

[13] B. Shickel, P. J. Tighe, A. Bihorac, P. Rashidi, Deep ehr: a survey of recent advances in deep learning techniques for electronic health record (ehr) analysis, IEEE journal of biomedical and health informatics 22 (2017) 1589–1604.

[14] U. FDA, Framework for fda's real-world evidence program, Silver Spring, MD: US Department of Health and Human Services Food and Drug Administration (2018).

[15] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, C. I. Sánchez, A survey on deep learning in medical image analysis, Medical image analysis 42 (2017) 60–88.

[16] A. J. Thirunavukarasu, D. S. J. Ting, K. Elangovan, L. Gutierrez, T. F. Tan, D. S. W. Ting, Large language models in medicine, Nature medicine 29 (2023) 1930–1940.

[17] Z. Kraljevic, D. Bean, A. Shek, R. Bendayan, H. Hemingway, J. A. Yeung, A. Deng, A. Baston, J. Ross, E. Idowu, et al., Foresight–generative pretrained transformer (gpt) for modelling of patient timelines using ehrs, arXiv preprint arXiv:2212.08072 (2022).

[18] H. Kondylakis, L. Koumakis, M. Tsiknakis, K. Marias, Implementing a data management infrastructure for big healthcare data, in: 2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI), IEEE, 2018, pp. 361–364.

[19] N. Miloslavskaya, A. Tolstoy, Big data, fast data and data lake concepts, Procedia Computer Science 88 (2016) 300–305.

[20] F. Ravat, Y. Zhao, Metadata management for data lakes, in: New Trends in Databases and Information Systems, ADBIS 2019 Short Papers, Workshops BBIGAP, QAUCA, SemBDM, SIMPDA, M2P, MADEISD, and Doctoral Consortium, Bled, Slovenia, September 8-11, 2019, Proceedings, volume 1064 of *Communications in Computer and Information Science*, Springer, 2019, pp. 37–44.

[21] C. Lagoze, C. A. Lynch, R. Daniel Jr, The Warwick Framework: A Container Architecture for Aggregating Sets ofMetadata, Technical Report, Cornell University, 1996.

[22] R. Gabriel, T. Hoppe, A. Pastwa, Classification of metadata categories in data warehousing - A generic approach, in: Sustainable IT Collaboration Around the Globe. 16th Americas Conference on Information Systems, AMCIS 2010, Lima, Peru, August 12-15, 2010, Association for Information Systems, 2010, p. 133.
[23] A. J. Gilliland, Setting the stage, Introduction to metadata 2 (2008) 1–19.
[24] J. Greenberg, A quantitative categorical analysis of metadata elements in image-applicable metadata schemas, Journal of the American Society for Information Science and Technology 52 (2001) 917–924.
[25] J. Pierson, L. Seitz, H. Duque, J. Montagnat, Metadata for efficient, secure and extensible access to data in a medical grid, in: 15th International Workshop on Database and Expert Systems Applications (DEXA 2004), with CD-ROM, 30 August - 3 September 2004, Zaragoza, Spain, IEEE Computer Society, 2004, pp. 562–566.
[26] R. Badawy, F. Hameed, L. Bataille, M. A. Little, K. Claes, S. Saria, J. M. Cedarbaum, D. Stephenson, J. Neville, W. Maetzler, et al., Metadata concepts for advancing the use of digital health technologies in clinical research, Digital biomarkers 3 (2020) 116–132.
[27] D. Piantella, P. Reali, P. Kumar, L. Tanca, A minimum metadataset for data lakes supporting healthcare research, in: Proceedings of the 32nd Symposium of Advanced Database Systems, 2024, volume 3741 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2024, pp. 681–691.
[28] L. G. Tereshchenko, M. E. Josephson, Frequency content and characteristics of ventricular conduction, Journal of Electrocardiology 48 (2015) 933–937. doi:10.1016/j.jelectrocard.2015.08.034.
[29] F. Shaffer, J. P. Ginsberg, An overview of heart rate variability metrics and norms, Frontiers in Public Health 5 (2017) 1–17. doi:10.3389/fpubh.2017.00258.
[30] M. Malik, J. T. Bigger, A. J. Camm, R. E. Kleiger, A. Malliani, A. J. Moss, P. J. Schwartz, Heart rate variability: Standards of measurement, physiological interpretation, and clinical use, European Heart Journal 17 (1996) 354–381. doi:10.1093/oxfordjournals.eurheartj.a014868.
[31] B. Kemp, J. Olivan, European data format 'plus' (edf+), an edf alike standard format for the exchange of physiological data, Clinical Neurophysiology 114 (2003) 1755–1761. doi:10.1016/S1388-2457(03)00123-8.
[32] K. J. Gorgolewski, T. Auer, V. D. Calhoun, R. C. Craddock, S. Das, E. P. Duff, G. Flandin, S. S. Ghosh, T. Glatard, Y. O. Halchenko, D. A. Handwerker, M. Hanke, D. Keator, X. Li, Z. Michael, C. Maumet, B. N. Nichols, T. E. Nichols, J. Pellman, J.-B. Poline, A. Rokem, G. Schaefer, V. Sochat, W. Triplett, J. A. Turner, G. Varoquaux, R. A. Poldrack, The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments, Scientific Data 3 (2016) 160044. doi:10.1038/sdata.2016.44.
[33] P. Reali, A. Carotenuto, D. Piantella, L. Tanca, P. Plebani, M. G. Signorini, Development of data ingestion pipelines for the federated use of biomedical data in research: The health big data project, in: 2024 IEEE 22nd Mediterranean Electrotechnical Conference (MELECON), IEEE, 2024, pp. 678–683. doi:10.1109/MELECON56669.2024.10608617.
[34] J. Pan, J. W. Tompkins, A real-time qrs detection algorithm, IEEE Transaction on Biomedical Engineering 32 (1985) 230–236. doi:10.1109/TBME.1985.325532.
[35] D. Makowski, T. Pham, Z. J. Lau, J. C. Brammer, F. Lespinasse, H. Pham, C. Schölzel, S. H. A. Chen, Neurokit2: A python toolbox for neurophysiological signal processing, Behavior Research Methods 53 (2021) 1689–1696. doi:10.3758/s13428-020-01516-y.
[36] A. H. Ribeiro, M. H. Ribeiro, G. M. M. Paixão, D. M. Oliveira, P. R. Gomes, J. A. Canazart, M. P. S. Ferreira, C. R. Andersson, P. W. Macfarlane, W. Meira, T. B. Schön, A. L. P. Ribeiro, Automatic diagnosis of the 12-lead ecg using a deep neural network, Nature Communications 11 (2020) 1760. doi:10.1038/s41467-020-15432-4.
[37] A. Pika, M. T. Wynn, S. Budiono, A. H. Ter Hofstede, W. M. van der Aalst, H. A. Reijers, Privacy-preserving process mining in healthcare, International journal of environmental research and public health 17 (2020) 1612.
[38] S. A. Sohail, F. A. Bukhsh, M. van Keulen, Multilevel privacy assurance evaluation of healthcare

metadata, Applied Sciences 11 (2021) 10686.
[39] R. Mariani, F. Rossi, R. Cucchiara, M. Pavone, B. Simkin, A. Koene, J. Papenbrock, Trustworthy ai—part 1, Computer 56 (2023) 14–18.
[40] O. Bodenreider, The unified medical language system (umls): integrating biomedical terminology, Nucleic acids research 32 (2004) D267–D270.