A Minimum Metadataset for Data Lakes Supporting Healthcare Research

(Discussion paper)

Davide Piantella[*], Pierluigi Reali, Priyansh Kumar and Letizia Tanca[†]

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

Abstract

While data lakes have emerged as a solution for storing vast amounts of heterogeneous and often unstructured data, responding to the growing need for flexible data storage, integration, and analytics in different domains, the digital transformation of healthcare processes has led to an exponential increase in various types of health records, necessitating efficient data management solutions and making this domain an ideal arena for experimenting data lake efficacy. In data lakes, effective metadata extraction and management are crucial for describing raw data, establishing connections, and ensuring interoperability among datasets ingested into the lake. To address this, we propose a minimum set of metadata tailored for clinical research, which includes relevant information common to significant branches of healthcare. Our metadataset not only streamlines data ingestion processes but also enhances the accessibility and usability of healthcare datasets for research purposes. By standardizing the collected metadata within the clinical research domain, we also facilitate data integration, analysis, and exploration, facilitating comprehensive data description and management within the data lake environment.

Keywords

medatata, healthcare, data lakes, interoperability

1. Introduction

Responding to the pressing demand for flexible and easily-accessible data analytics [1], an emerging trend involves data lakes as repositories for vast amounts of data and documents in the big-data context [2]. Notably, data lakes operate without a predefined schema, enabling the ingestion of raw data in various formats (including relational data, images, text, data streams, and logs) without the need for prior preprocessing [3]. This adaptability empowers users and organizations to seamlessly store and access their data, facilitating data analytics, data-driven applications, and machine learning tasks.

In the field of medicine, the transition to digital healthcare processes and services has led to an exponential increase in medical data. Within hospitals, daily operations generate a multitude of (often unstructured) digital documents, including medical images, nursing notes, discharge

SEBD 2024: 32nd Symposium on Advanced Database Systems, June 23-26, 2024, Villasimius, Sardinia, Italy

[*] Corresponding author.

[†] PNRR - M4 C2, Invest 1.3 - D.D. 1551.11-10-2022, PE00000004). CUP MICS D43C22003120001.

davide.piantella@polimi.it (D. Piantella); pierluigi.reali@polimi.it (P. Reali); priyansh.kumar@mail.polimi.it (P. Kumar); letizia.tanca@polimi.it (L. Tanca)
0000-0003-1542-0326 (D. Piantella); 0000-0003-3041-4004 (P. Reali); 0000-0003-2607-3171 (L. Tanca)

letters, and laboratory results. Moreover, advancements in medical devices, applications, and monitoring technologies have digitized patient data, resulting in the collection, analysis, and storage of vast amounts of heterogeneous information. In fact, it seems that, by 2025, the annual growth rate of healthcare data will surpass that of generic data, reaching 36% compared to circa 27% [4]. These challenges make the realm of medicine the ideal one for experimenting the effectiveness of the use of Metadata.

With the increasing availability of Electronic Health Records facilitating real-world-evidence clinical trials [5], a significant application of healthcare data management is medical research. In this context, the ability to collect and analyze data from heterogeneous sources is crucial [6], and, given the diverse formats of healthcare data and its sheer volume, a data lake is a very interesting solution. Since the datasets ingested by a data lake are extremely heterogeneous, accessing and manipulating the stored raw data can be very expensive in terms of computational and time complexity, therefore effective metadata extraction and management, establishing connections among the ingested datasets [7], are essential for describing raw data. In fact, metadata provide valuable information regarding the data without the need to directly analyze the datasets.

To achieve this, we propose a minimum set of metadata (i.e., a *minimum metadataset*) for the context of clinical research, which encloses the relevant information common to the main branches of healthcare. Data feeders can then specify additional metadata that further describe the datasets.

2. Methodology and Related Work

We consider metadata models and tools specifically tailored for the healthcare context.

Our primary objective is to construct a minimum metadata model that not only offers essential information relevant to associated healthcare data but also creates a distinctive framework that facilitates the sharing of clinical data coming from diverse formats and sources. This improves interoperability, which, in turn, supports seamless data exchange and collaboration across different healthcare organizations. Our proposed minimum metadata model serves as a foundation that can be further enhanced and specialized for each specific scope of use.

We now briefly describe the existing clinical metadata models and management tools we analyzed, which contributed to the design of our minimum metadataset.

2.1. Genosurf

Genosurf [8] is a metadata integration and search system designed to efficiently analyze genomics datasets from various sources in biological and clinical research settings. It leverages a Genomics Conceptual Model (GCM) and implements a multi-ontology semantic search system. The metadata repository includes millions of metadata entries from multiple datasets, focusing on significant genomics data. The system offers a web-based interface that allows users to perform targeted searches based on specific metadata attributes and values. In this way, users can inspect descriptions of matching datasets, explore the related metadata, and obtain the link to the original datasets. Moreover, Genosurf facilitates free-text searches and offers query preparation functionalities for further data processing.

2.2. PDXFinder

Patient-Derived tumor Xenograft (PDX) models are essential tools to study the effects of chemotherapy on tumors. PDXFinder [9] provides centralized access to an extensive collection of PDX models. It supports advanced search functionalities that enable users to filter and refine their searches based on specific criteria such as cancer type, molecular characteristics, and treatment history. As a result, researchers can access detailed information about each PDX model, including clinical annotations, molecular profiling data, histopathological features, and associated research publications.

2.3. HL7 - FHIR

The Health Level Seven International (HL7) Fast Healthcare Interoperability Resources (FHIR) [10] is widely acknowledged as a fundamental metadata model for achieving healthcare data interoperability, presenting a standardized approach to data representation and exchange. FHIR provides extensive information, encompassing various aspects of healthcare data, such as patient demographics, clinical observations, medications, and procedures. These resources are designed to be easily accessible using RESTful APIs[11], further enhancing its appeal and ease of implementation. The standardized nature of FHIR and its support for RESTful APIs can enable data exchange and sharing between diverse healthcare systems, regardless of their underlying technology and platforms.

2.4. Datacite

Datacite [12, 13] is an internationally recognized organization that provides persistent identifiers, known as DOIs (Digital Object Identifiers), for research data. Although not specifically a metadata model, Datacite significantly contributes to data discoverability, access, and reference. By assigning DOIs to research datasets, Datacite ensures their long-term accessibility and establishes a standardized approach for referencing and linking data. The metadata offered by Datacite includes essential details about the dataset, such as its title, authors, publisher, publication date, version, and any related resources. This metadata is usually presented in a standardized format, leveraging a metadata schema, defining specific data elements and their required or recommended attributes.

2.5. EOSC FAIR principles

The European Open Science Cloud (EOSC) initiative[1] aims to create a seamless and open research environment by providing access to research data, services, and infrastructures across Europe. EOSC recently published its guidelines and recommendations to promote the Findability, Accessibility, Interoperability, and Reusability (FAIR) of research data and services [14]. The EOSC FAIR principles emphasize the importance of making research data and related resources easily discoverable, accessible, and interoperable. By adhering, researchers and data providers

[1] https://digital-strategy.ec.europa.eu/en/policies/open-science-cloud. These principles align with the broader FAIR data movement, which seeks to maximize the value and impact of research data by ensuring its usability and long-term preservation.

adopt standardized metadata (categorized as mandatory, recommended, and optional), data formats, and interoperability standards. This enables efficient data discovery, access, integration, and reuse, promoting collaboration, knowledge sharing, and interdisciplinary research within the EOSC ecosystem.

2.6. Standardized terminologies and coding systems

Standardized terminologies and coding systems are vital for achieving healthcare data interoperability [15], providing a common vocabulary and coding structure, to ensure that clinical concepts and metadata are represented in a consistent and standardized manner across different healthcare systems and datasets. For example, SNOMED-CT [16] is a comprehensive clinical vocabulary widely used in healthcare. It allows for the precise and uniform encoding of clinical observations, diagnoses, procedures, and other medical concepts. Similarly, LOINC [17] is a standardized coding system specifically designed for clinical laboratory observations and results. It provides a unified representation of laboratory tests, measurements, and observations.

2.7. Clinical Document Architecture

The Clinical Document Architecture (CDA) [18], developed by HL7, serves as a minimum metadata model for exchanging clinical documents. CDA defines the structure and semantics of clinical records, enabling the standardized sharing of healthcare information. It enables interoperability across different healthcare organizations by providing a common framework for representing patient clinical summaries, discharge letters, progress notes, and other healthcare documents.

3. Minimum metadataset

We report in Table 1 our proposed minimum metadataset for healthcare. Following several research works [19, 20, 21], we decided to employ for our model three main categories: (i) *administrative* metadata, (ii) *data provenance* metadata, and (iii) *descriptive* metadata.

Category	Attributes
Administrative	GUID, Creator, Owner, Rights, Terms of access
Data provenance	Publication year, Upload date, Acquisition method, Acquisition tools, Download URL, Checksum, Encryption algorithm, File version, Update/modification date, Update frequency
Descriptive	File description, File format, Min age, Max age, Ethnicity, Patient sex, Blood group, Primary site, Collection site, Disease names, Disease types, Disease variants

Table 1
Proposed minimum metadataset for healthcare

Administrative metadata refer to the administrative aspects of data management, facilitating effective data governance, management, and administration. They include the following metadata, regarding ownership, access authorizations, and policies:

- *GUID (Globally Unique Identifier)*: a unique identifier assigned to each dataset for identification and referencing purposes.
- *Creator*: the entity responsible for creating or generating the dataset.
- *Owner*: the entity that owns the dataset and holds responsibility for its management.
- *Rights*: the permissions or restrictions associated with accessing and using the dataset.
- *Terms of access*: the terms and conditions that govern the access and usage of the dataset.

Data provenance metadata serve as a detailed record of the data lifecycle. They offer valuable insights on reliability and quality by capturing information about collection methods, processing steps, and modifications:

- *Publication year*: the year in which the dataset was officially published or made available.
- *Upload date*: the date when the dataset was uploaded into the repository.
- *Acquisition method*: a description of the acquisition process employed to collect the data.
- *Acquisition tools*: SW and HW used to collect the data, with the related version details.
- *Download URL*: it refers to the specific web address that enables users to download the dataset to their local systems.
- *Checksum*: a hash value that acts as a verification mechanism for data integrity.
- *Encryption algorithm*: if, for privacy reasons, the dataset is encrypted, this reports the algorithm used to protect sensitive information.
- *File version*: an identifier or label that denotes the version or the revision of the dataset. It allows healthcare professionals, researchers, and stakeholders to track and manage different instances of the dataset, ensuring proper documentation and version control.
- *Update/modification date*: it stores the date when the dataset was last updated or modified. This provides valuable information about the currency and freshness of the data, allowing the users to ascertain the relevance and applicability of the dataset for their specific needs.
- *Update frequency*: it indicates the regularity or frequency at which the dataset is updated.

Descriptive metadata refer to the content and characteristics of a dataset, freeing the users from the need to examine the resource itself in detail. This category is essential for classifying and organizing datasets, enabling efficient search and retrieval, and facilitating decision-making about which resources better fit the needs of the users:

- *File description*: a brief description or summary of the dataset, providing an overview of its purpose, scope, and data content.
- *File format*: the specific file format in which the dataset is stored (e.g., CSV, XML, or DICOM).
- *Min age*: the minimum age of the patients represented in the dataset.
- *Max age*: the maximum age of the patients represented in the dataset.
- *Ethnicity*: the ethnic background of the patients represented in the dataset.
- *Patient sex*: the sex of the patients included in the dataset.
- *Blood group*: the blood type of the patients included in the dataset.
- *Primary site*: the primary anatomical site or organ associated with the data collected.
- *Collection site*: the location or institution where the data was collected or originated.
- *Disease names*: names of the diseases or medical conditions primarily represented in the dataset.

- *Disease types*: the classification or type of diseases or medical conditions.
- *Disease variants*: specific variants or subtypes of diseases or medical conditions.

4. Adequacy of the proposed model

In Table 2 we compare our minimum metadataset with the models described in Section 2, leveraging some of the main metadata commonly employed for both general-purpose [19, 20, 21, 22, 23] and healthcare-related [24, 25] domains. Moreover, we evaluated the quality of our model by studying how it addresses some of the most common challenges encountered in – but not limited to – clinical data science and integration, which are acknowledged as critical barriers also in the *National Institute of Health (NIH) strategic plan for data science research* [26].

	Our model	GENOSURF [8]	PDXFINDER [9]	CDA/FHIR [18, 10]	DATACITE [12, 13]	EOSC [14]
Domain	Healthcare (in general)	Genomic	Cancer	Health documents	General purpose	General purpose
Terms&Conditions	✓	✗	∼[⋆]	✗	✓	✓
File details	✓	✗	✗	✗	✓	✓
Content description	✓	✗	✗	✗	✓	✓
File format and structure	✓	✓	✓	✓	✓	✓
Provenance	✓	∼[*]	✓	✗	✗	∼[†]
Publication date	✓	✗	✗	✓	✓	✓
Reference to vocabularies	✗	✗	✗	✓	✗	✗
Access rights	✓	✗	✗	✗	✗	✓
Integrity information	✓	∼[+]	✗	✗	✗	✓
Encryption information	✓	✗	✗	✗	✗	✗
Patient: blood group	✓	✗	✗	✓	✗	✗
Patient: age	✓	✓	∼[‡]	✓	✗	✗
Patient: gender	✓	✓	✗	✓	✗	✗
Patient: ethnicity	✓	✓	✗	✗	✗	✗
Observation: disease	✓	✓	✓	✓	✗	✗
Observation: collection site	✓	✓	✓	✓	✗	✗

[⋆] License type only.
[*] Techniques only.
[†] Software used only.
[+] Claimed to be stored, although not displayed.
[‡] Predefined ranges only.

Table 2
Comparison of our minimum metadataset for healthcare with other models

4.1. Lack of standard structure and policies

The lack of consistent data standards and formats across different healthcare systems poses a significant challenge in achieving effective interoperability. Healthcare organizations often employ diverse coding schemes, data structures, and terminologies, leading to inconsistencies and incompatibilities when exchanging health information. This inconsistency may result in errors and misinterpretations and possibly make data integration more complex.

Proposed Solution The minimum metadataset we suggest in Table 1 provides a standardized model for organizing and describing essential information about healthcare data for research purposes. By adopting this model, healthcare organizations can establish a common structure for data representation, promoting consistency and compatibility in data exchange. To ensure high interoperability, this solution is in line with state-of-the-art standards and frameworks, such as HL7 FHIR [10] and DICOM (Digital Imaging and Communications in Medicine)[27], as well as the others mentioned in Section 2. Moreover, leveraging the proposed minimum metadata model, healthcare systems could map their local data elements and terminologies to the standardized model, facilitating accurate interpretation and integration of health information. Finally, the standardized metadata elements can also help in designing a data catalog for a data lake that can easily accommodate different types of healthcare data.

4.2. Privacy concerns

Healthcare data is inherently sensitive and requires robust protection to maintain confidentiality and ensure the secure and reliable exchange of health information. The already stringent privacy regulations implemented in the US (HIPAA [28]) and Europe (GDPR [29]) must comply with other privacy regulations specific to each country or region.

Proposed Solution Our model prioritizes data protection and privacy by excluding sensitive information such as patient names, dates of birth, and unique identifiers that could potentially disclose patient or clinician identities. The model minimizes the risk of privacy breaches by carefully selecting and including only non-identifying attributes. This approach aligns with privacy and security policies, safeguarding the confidentiality of healthcare data and promoting a secure environment for data exchange and interoperability.

4.3. Incomplete and inaccurate data

The quality and usefulness of healthcare datasets can be compromised by inconsistent data capture and incomplete or inaccurate data entry practices. These issues significantly impact the integrity and reliability of the exchanged data. Inconsistent data capture refers to variations in how data is collected and recorded across different healthcare systems or organizations. This can be due to discrepancies in terminology, coding systems, and acquisition processes, making it challenging to compare and integrate information accurately. Incomplete or inaccurate data entry practices further compound these challenges by introducing errors or missing information into the exchanged data.

Proposed Solution The model incorporates essential attributes describing data provenance, ensuring both the elicitation of details regarding the acquisition process and the tracking and management of different versions of datasets over time. These attributes allow healthcare professionals and researchers to clearly understand the acquisition methods and identify the most up-to-date version of a dataset, reducing the risk of utilizing outdated or incomplete data. By clearly indicating the dataset version, our model promotes data integrity and ensures that users work with the most accurate and complete information.

4.4. Data bias

Data bias is a significant concern in healthcare research [30, 31], as it can lead to unequal treatment, inaccurate research findings, and disparities in patient outcomes. Bias can arise from several aspects, e.g., the demographics of the population sampled, the methods used to collect and analyze data, and other intrinsic biases. For example, if a dataset primarily includes information from individuals of a certain age or ethnicity, the findings and conclusions drawn from that data may not be applicable or representative of the broader population. Similarly, biases can occur when selecting variables to be measured, leading to incomplete or skewed representations of health conditions. Addressing data bias is crucial to ensure fair and reliable research insights.

Proposed Solution Addressing data bias is a complex task that requires a multifaceted approach. While our proposal focuses on a minimum metadata model, this does not solve the problem completely. Therefore, scientists, researchers, and medical professionals must employ various methodologies to tackle this issue comprehensively [32, 33]. We recognize the significance of including attributes such as ethnicity, sex, and collection site, which can help researchers and professionals analyze the demographic and geographical scope of the datasets, thus assessing potential biases and accounting for them in their analyses. By combining the strengths of the minimum metadata model, which addresses the identification of possible data bias through attribute inclusion, with other approaches [34, 35, 36], researchers can work towards mitigating and minimizing data bias, ultimately enhancing the quality and fairness of their research outcomes.

4.5. Data discovery

The rapid generation of healthcare data brings the difficulty of finding relevant datasets for specific research or clinical purposes, in terms of required variables, population demographics, or specific clinical parameters. This issue is further compounded by the lack of standardized data formats, inconsistent data labeling, and varying data storage practices across different healthcare systems and organizations.

Proposed Solution Researchers can leverage the proposed metadata model to establish a standardized framework for organizing and describing essential information about healthcare datasets. This includes attributes such as data structure, variables, demographics, diseases, and anatomical sites. Data consumers can then utilize this standardized metadata to efficiently search and filter through the vast amount of available datasets.

5. Conclusions and future works

Metadata can be used to support the storage, retrieval and analysis of complex datasets without the need to directly accessing raw data.In this paper we demonstratedthis possibility using the example of clinical metadata, which shows the essential information that data feeders should attach to each dataset before ingesting it into a data lake. We have shopwn as well that the use of metadata enhances data findability across multiple datasets, helping researchers acquire suitable data for their studies. Future extensions of this work will include bounding the values of the metadata fields to specific vocabularies to reduce representation ambiguities. In the clinical domain, an attractive solution could be exploiting the Unified Medical Language System (UMLS) [37], a controlled compendium of medical vocabularies including, among others, SNOMED-CT [16] and LOINC [17]. The multi-language support of UMLS could certainly facilitate the adoption and usage of our minimum metadataset by clinicians and researchers.

Acknowledgments

This work was carried out within the MICS (Made in Italy - Circular and Sustainable) Extended Partnership and received funding from Next-Generation EU (Italian PNRR - M4 C2, Invest 1.3 - D.D. 1551.11-10-2022, PE00000004). CUP MICS D43C22003120001.

References

[1] H. Fang, Managing data lakes in big data era: What's a data lake and why has it became popular in data management ecosystem, in: 2015 IEEE International Conference on Cyber Technology in Automation, Control, and Intelligent Systems (CYBER), IEEE, 2015, pp. 820–824.

[2] D. Piantella, A research on data lakes and their integration challenges, in: Proceedings of the 30th Italian Symposium on Advanced Database Systems, SEBD, volume 3194 of *CEUR Workshop Proceedings*, 2022, pp. 616–621.

[3] R. Hai, C. Koutras, C. Quix, M. Jarke, Data lakes: A survey of functions and systems, IEEE Transactions on Knowledge and Data Engineering (2023).

[4] 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).

[5] 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).

[6] 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.

[7] 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 23, Springer, 2019, pp. 37–44.

[8] A. Canakoglu, A. Bernasconi, A. Colombo, M. Masseroli, S. Ceri, GenoSurf: metadata driven semantic search system for integrated genomic datasets, Database 2019 (2019) 132.

[9] N. Conte, J. C. Mason, C. Halmagyi, S. Neuhauser, A. Mosaku, G. Yordanova, A. Chatzipli, D. A. Begley, D. M. Krupke, H. Parkinson, T. F. Meehan, C. C. Bult, PDX Finder: A portal for patient-derived tumor xenograft model discovery, Nucleic acids research 47 (2019) D1073–D1079.

[10] R. H. Dolin, L. Alschuler, Approaching semantic interoperability in health level seven, Journal of the American Medical Informatics Association 18 (2011) 99–103.

[11] A. Ehsan, M. A. M. Abuhaliqa, C. Catal, D. Mishra, RESTful API testing methodologies: Rationale, challenges, and solution directions, Applied Sciences 12 (2022) 4369.

[12] J. Brase, DataCite: a global registration agency for research data, in: 2009 fourth international conference on cooperation and promotion of information resources in science and technology, IEEE, 2009, pp. 257–261.

[13] P. Scott, R. Worden, Semantic mapping to simplify deployment of HL7 v3 clinical document architecture, Journal of biomedical informatics 45 (2012) 697–702.

[14] O. Corcho, M. Eriksson, K. Kurowski, M. Ojsteršek, C. Choirat, M. Van de Sanden, F. Coppens, EOSC interoperability framework, Report from the EOSC Executive Board Working Groups FAIR and Architecture, 2021.

[15] O. Bodenreider, R. Cornet, D. J. Vreeman, Recent developments in clinical terminologies SNOMED-CT, LOINC, and RxNorm, Yearbook of medical informatics 27 (2018) 129–139.

[16] K. Donnelly, SNOMED-CT: The advanced terminology and coding system for ehealth, Studies in health technology and informatics 121 (2006) 279.

[17] C. J. McDonald, S. M. Huff, J. G. Suico, G. Hill, D. Leavelle, R. Aller, A. Forrey, K. Mercer, G. DeMoor, J. Hook, W. Williams, J. Case, P. Maloney, LOINC, a universal standard for identifying laboratory observations: a 5-year update, Clinical chemistry 49 (2003) 624–633.

[18] R. H. Dolin, L. Alschuler, S. Boyer, C. Beebe, F. M. Behlen, P. V. Biron, A. Shabo, HL7 clinical document architecture, release 2, Journal of the American Medical Informatics Association 13 (2006) 30–39.

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

[20] A. J. Gilliland, Setting the stage, Introduction to metadata 2 (2008) 7.

[21] U.S. National Archives, Metadata in electronic records management, https://records-express.blogs.archives.gov/2016/11/21/metadata-in-electronic-records-management/, 2016. Online; accessed April-2024.

[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] 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.

[24] J. Pierson, L. Seitz, H. Duque, J. Montagnat, Metadata for efficient, secure and extensible access to data in a medical grid, in: Proc. 15th International Workshop on Database and Expert Systems Applications, 2004., IEEE Computer Society, 2004, pp. 562–566.

[25] R. Badawy, F. Hameed, L. Bataille, M. A. Little, K. Claes, S. Saria, J. M. Cedarbaum, D. Stephenson, J. Neville, W. Maetzler, A. J. Espay, B. R. Bloem, T. Simuni, D. R. Karlin, Metadata concepts for advancing the use of digital health technologies in clinical research, Digital biomarkers 3 (2020) 116–132.

[26] U.S. National Institutes of Health, NIH strategic plan for data science, https://datascience.nih.gov/nih-strategic-plan-data-science, 2018. Online; accessed April-2024.

[27] M. Mustra, K. Delac, M. Grgic, Overview of the DICOM standard, in: 2008 50th International Symposium ELMAR, volume 1, IEEE, 2008, pp. 39–44.

[28] I. G. Cohen, M. M. Mello, HIPAA and protecting health information in the 21st century, Jama 320 (2018) 231–232.

[29] C. J. Hoofnagle, B. Van Der Sloot, F. Z. Borgesius, The European Union general data protection regulation: what it is and what it means, Information & Communications Technology Law 28 (2019) 65–98.

[30] I. G. Cohen, R. Amarasingham, A. Shah, B. Xie, B. Lo, The legal and ethical concerns that arise from using complex predictive analytics in health care, Health affairs 33 (2014)

1139–1147.

[31] A. Rajkomar, M. Hardt, M. D. Howell, G. Corrado, M. H. Chin, Ensuring fairness in machine learning to advance health equity, Annals of internal medicine 169 (2018) 866–872.

[32] N. Norori, Q. Hu, F. M. Aellen, F. D. Faraci, A. Tzovara, Addressing bias in big data and AI for health care: A call for open science, Patterns 2 (2021).

[33] 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.

[34] J. R. Marcelin, D. S. Siraj, R. Victor, S. Kotadia, Y. A. Maldonado, The impact of unconscious bias in healthcare: how to recognize and mitigate it, The Journal of infectious diseases 220 (2019) S62–S73.

[35] C. FitzGerald, S. Hurst, Implicit bias in healthcare professionals: a systematic review, BMC medical ethics 18 (2017) 1–18.

[36] J. Odgaard-Jensen, G. E. Vist, A. Timmer, R. Kunz, E. A. Akl, H. Schünemann, M. Briel, A. J. Nordmann, S. Pregno, A. D. Oxman, Randomisation to protect against selection bias in healthcare trials, Cochrane database of systematic reviews (2011).

[37] O. Bodenreider, The unified medical language system (UMLS): integrating biomedical terminology, Nucleic acids research 32 (2004) D267–D270.