Medical Data Compatibility Problems in the Tasks of Information Systems Integration. Conceptualization.

Ozar Mintser,[a] Larysa Babintseva,[a] Stanislav Mokhnachov,[a] Olga Sukhanova,[a] Pavlo Hanynets,[a] Alexandr Sarcanich[a]

[a]Shupyk National Healthcare University of Ukraine, Kyiv, Ukraine

DOI: 10.25929/695sar10

ABSTRACT

Integration of medical data is a critical component of the functioning of modern healthcare systems and a primary task of personalized medicine. Aggregating data from disparate sources, such as electronic medical records and medical devices, allows service providers to obtain a complete picture of patients' health status and optimize workflows.

It is noted that the strategy of integration is closely linked with the logic of medical data compatibility. The problems of integration also reflect in the tasks of creating Portfolios for physicians and pharmacists. Research objective. Summary of biomedical data compatibility issues. It is emphasized that data compatibility depends on the consistency of standards applied in programs. The quality of data also requires special attention. This directly affects the quality of the decisions made. Although data interoperability is one of the primary requirements of information and communication systems (ICS), it is often overlooked. As a result, data exchange is not performed, significantly limiting the flow of information. The problem of large dimensionality is also serious. It is evident that such multidimensional and labor-intensive computational processes are a primary task for modern algorithms and models of artificial intelligence (AI) and machine learning (ML).

It is highlighted that one of the ways to solve the problems of standardization and integration of large-scale medical data are metadata, which are also useful for improving statistical analysis, probabilistic models, and ML models. Conclusions. 1. Precision Medicine (4P Medicine) was introduced as a new paradigm approach to healthcare with a more predictable, preventive, personalized, and participatory manner. Precision medicine is closely related to data-intensive approaches, as well as to ML and AI. 2. Integration of data for placement in structures useful for precision medicine is only possible after several previous stages of their processing, namely: data (metadata) collection, processing, obtaining 'clean' data, data compression. 3. To realize the prospects of precision medicine, approaches to computational learning must evolve with the help of well-chosen and well-integrated digital data ecosystems.

KEYWORDS

Integration of medical data, interoperability of medical data, Portfolio of doctors and pharmacists, information and communication systems, 4P Medicine, interoperable data standards, metadata

1. Introduction

The primary goal of precision medicine is the aggregation and integration of extensive arrays of diverse data into analytical structures that enable the development of individualized, context-dependent diagnostic and therapeutic approaches. In this regard, artificial intelligence (AI) and machine learning (ML) approaches can be used to construct analytical models for complex diseases and be utilized in predicting personalized health outcomes. [1]

Computational approaches in medicine are characterized by large datasets that combine both structured and unstructured formats. Clinical and biomedical data offer a wide selection of shapes and formats, sizes, complexity levels, and are often poorly annotated and unstructured, creating heterogeneity in many situations. For instance, various types of variables (different encoding), different datasets (electronic health records [EHRs] from different hospitals), inconsistent distributions or scaling, diverse modalities of data (continuous signals, intervals, categories, etc.), and different formats (various standards of medical reporting). Each of these issues poses a challenge for effective modeling using AI and ML.

A range of problems also arises in analyzing medical data in modern personalized medicine settings, depending on time, available computational power, and bioethical constraints. One of the primary challenges in deriving knowledge from electronic medical records is that they represent extremely heterogeneous data sources with complex arrays of quantitative, qualitative, and transactional data.

Different types of data include ICD (International Classification of Diseases) codes, biochemical and laboratory analyses, clinical (textual) notes, historical archives of medical interventions, treatment methods. These data sources are often collected by dozens of people, separately for each case, making the data from Electronic Health Records (EHR) quite challenging to analyze. EHRs were not designed as a resource for automated learning; hence the data structures representing information for preservation were not considered during their usage. As EHRs are primarily adapted for clinical and hospital logistics, modeling such data and training AI algorithms based on them often encounter problems related to structural heterogeneity or issues of adaptation through existing strategies or by redesigning these medical records. [2,3] The problem of dimensionality is also significant. The dimensionality problem is particularly evident in the case of genomic and transcriptomic analysis, where the number of genes or transcripts is about tens of thousands, while the number of samples seldom exceeds a few hundred or a few thousand.

2. Research objective

Summarize the problems of compatibility of biomedical data. Our work is aimed at presenting general options for approaches to overcoming these problems by applying machine learning and artificial intelligence to analyze biomedical and health-related data in the context of precision medicine.The methodological basis of the study is the conceptualization of contemporary problems of ensuring the compatibility of biomedical data.

3. Discussion

Since the late 20th century, the field of medicine and healthcare has employed an approach to the proper verification of existing clinical and biomedical research, known as evidence-based medicine (EBM). It aims at the comprehensive utilization of confirmed accumulated scientific and clinical data for the development of health-related measures and policies. With the emergence of large, qualitatively selected data arrays obtained from large-scale clinical trials (large-scale data analytics, LSDA) and powerful methods of such data analysis as well as the development of mechanisms for converting them into useful information, the ideals of EBM have been incorporated into Precision Medicine. [5] Later, biological databases were added to such data, which include individual EHR data as well as social information (social determinants of health). [6,7]

Advancements in artificial intelligence play a central role in the development of such integrated structures. Data integration indeed is a highly complex task. The ability to perform complex queries, construct heterogeneous models, and develop hierarchically nested data search operations in multiple databases is the primary goal of data integration strategies, useful for artificial intelligence and machine learning models in precision medicine.

The necessity of exchanging clinical research data to ensure the reproducibility of results, plan subsequent research stages, perform quantitative comparisons of diagnostic or treatment methods' effectiveness, accelerate the reporting of results, and ensure continuous medical education becomes evident. Optimal utilization of shared data is associated with the need for their standardization, which becomes a central task in both medical research and personalized clinical practice.

In resolving issues of standardization and integration of large-scale medical data (LSDA), metadata have become a central element in modern solutions, which can help to ensure efficient exchange, analysis, and usage of information in healthcare. For this reason, the aspiration to create high-quality, well-formatted, and standardized metadata has become highly relevant. [8]

Metadata also prove useful in enhancing statistical analysis, probabilistic models, and ML models. The use of metadata can improve query optimization through resampling and initial loading, standardization of data sets, and as an auxiliary source for multi-dimensional Bayesian analysis, and analysis of datasets with different dynamic ranges. [9–12]

Integration of multiple data sources and metadata further requires the design, development, and implementation of analysis algorithms capable of processing heterogeneous data in conditions of noise accumulation, false correlations, and random endogeneity, while maintaining a balance between statistical accuracy, computational efficiency, and interpretability. Addressing such issues may require new models for the implementation of metadata reporting standards.

Standardization of the method of presentation and acquisition of metadata in biomedical and clinical settings is critically important for the development of comprehensive machine learning approaches that fully utilize these unified data structures. [13]

The current development of complex and quite effective AI algorithms and the accompanying proliferation of large-scale data sources in biomedical conditions have heightened expectations regarding the many potential benefits that can be derived from the combination of "good" methods and "good" data ("clean" data). However, to make these large volumes of data useful for creating high-quality artificial intelligence models, it is necessary to address not only the dimensionality challenge. There's a need to create a system of "clean" data that can amalgamate diverse sources, technologies, infrastructures, and processes for the collection, storage, processing, analysis, and utilization of medical data in a specific organization or in the industry as a whole. Such systems are called digital data ecosystems.

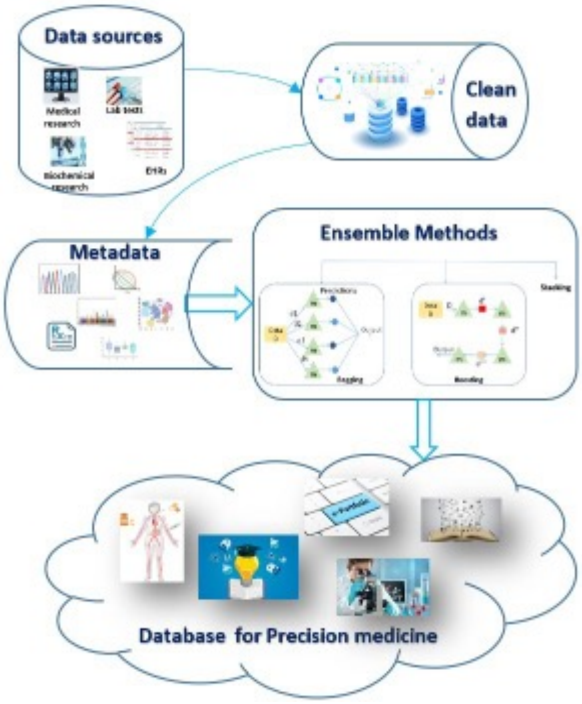

Figure 1. Digital data ecosystems.

A data ecosystem creates a conducive environment for data management and utilization in making informed decisions, refining processes, and achieving strategic goals. It consists of various elements such as software, hardware, data, and people, fostering communication and collaboration at different stages.

In the context of healthcare, a digital data ecosystem can also denote a set of digital platforms and technologies used for interaction with doctors, patients, and other stakeholders based on prepared data of any nature and origin. The digital ecosystem should eliminate barriers in the diagnosis and treatment of patients and enable each participant to utilize the most advanced technologies and systems to meet their needs.

Conclusions.

1. Precision Medicine (4P Medicine) is introduced as a new paradigm approach to healthcare, characterized by more predictive, preventive, personalized, and participatory methods. Precision Medicine is closely linked with data-intensive approaches, as well as with ML (machine learning) and AI (artificial intelligence).

2. The integration of data for placement in structures beneficial for precision medicine is only possible after several preceding stages of data processing, namely: data (metadata) collection, processing, obtaining "clean" data, and data compression.

3. To realize the prospects of precision medicine, approaches to computational learning must evolve with the help of well-selected and well-integrated digital data ecosystems.

References and notes

[1] Martínez-García M, Hernández-Lemus E. Data Integration Challenges for Machine Learning in Precision Medicine. *Frontiers in Medicine*. 2021;8:784455. https://doi.org/10.3389/fmed.2021.784455.

[2] Choi E, Xu Z, Li Y, Dusenberry M, Flores G, Xue E, et al. Learning the graphical structure of electronic health records with graph convolutional transformer. In: *Proceedings of the AAAI Conference on Artificial Intelligence.* Vol. 34, New York, NY: 2020. p. 606–613.

[3] Unberath P, Prokosch HU, Gründner J, Erpenbeck M, Maier C, Christoph J. EHR-independent predictive decision support architecture based on OMOP. *Applied Clinical Informatics*. 2020;11:399–404. https://doi.org/10.1055/s-0040-1710393.

[4] Mirza B, Wang W, Wang J, Choi H, Chung NC, Ping P. Machine Learning and Integrative Analysis of Biomedical Big Data. *Genes (Basel)*. 2019;10(2):87. https://doi.org/10.3390/genes10020087.

[5] Chow N, Gallo L, Busse JW. Evidence-based medicine and precision medicine: Complementary approaches to clinical decision-making. *Precision Clinical Medicine*. 2018;1(2):60-64. https://doi.org/10.1093/pcmedi/pby009.

[6] Vazquez M, Valencia A. Patient Dossier: Healthcare queries over distributed resources. PLoS *Computational Biology.* 2019;15(10):e1007291. https://doi.org/10.1371/journal.pcbi.1007291.

[7] Shaibi GQ, Kullo IJ, Singh DP, Hernandez V, Sharp RR, Cuellar I, et al. Returning genomic results in a Federally Qualified Health Center: the intersection of precision medicine and social determinants of health. *Genetics in Medicine.* 2020;22(9):1552–1559. https://doi.org/10.1038/s41436-020- 0806-5.

[8] Özdemir V, Kolker E, Hotez PJ, et al. Ready to put metadata on the post-2015 development agenda? Linking data publications to responsible innovation and science diplomacy. *OMICS – A Journal of Integrative Biology.* 2014;18(1):1–9. https://doi.org/10.1089/omi.2013.0170.

[9] Sass S, Buettner F, Mueller NS, Theis FJ. A modular framework for gene set analysis integrating multilevel omics data. *Nucleic Acids Research*. 2013 Nov;41(21):9622–9633. https://doi.org/10.1093/nar/gkt752.

[10] Meng C, Kuster B, Culhane AC, Gholami AM. A multivariate approach to the integration of multi-omics datasets. *BMC Bioinformatics*. 2014 May 29;15:162. https://doi.org/10.1186/1471-2105-15-162.

[11] Fagan A, Culhane AC, Higgins DG. A multivariate analysis approach to the integration of proteomic and gene expression data. Proteomics. 2007 Jun;7(13):2162–2171. https://doi.org/10.1002/pmic.200600898.

[12] Fan J, Han F, Liu H. Challenges of Big Data Analysis. *National Science Review.* 2014 Jun;1(2):293–314. https://doi.org/10.1093/nsr/nwt032..

[13] Norgeot B, Quer G, Beaulieu-Jones BK, et al. Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. *Nature Medicine.* 2020 Sep;26(9):1320–1324. https://doi.org/10.1038/s41591-020-1041-y.