



Medical Big Data Storage in Precision Medicine: A Systematic Review

Mostafa Langarizadeh (PhD)¹, Mehdi Hajebrahimi (PhD Candidate)^{1*}

¹Department of Health Information Management, School of Health Management and Information Sciences, Iran University of Medical Sciences, Tehran, Iran

ABSTRACT

Background: The characteristics of medical data in Precision Medicine (PM), the challenges related to their storage and retrieval, and the effective facilities to address these challenges are importantly considered in implementing PM. For this purpose, a secured and scalable infrastructure for various data integration and storage is needed.

Objective: This study aimed to determine the characteristics of PM data and recognize the challenges and solutions related to appropriate infrastructure for data storage and its related issues.

Material and Methods: In this systematic study, coherent research was conducted on Web of Science, Scopus, PubMed, Embase, and Google Scholar from 2015 to 2023. A total of 16 articles were selected and evaluated based on the inclusion and exclusion criteria and the central search theme of the study.

Results: A total of 1,961 studies were identified from designated databases, 16 articles met the eligibility criteria and were classified into five main sections: PM data and its major characteristics based on the volume, variety and velocity (3Vs) of medical big data, data quality issues, appropriate infrastructure for PM data storage, cloud computing and PM infrastructure, and security and privacy. The variety of PM data is categorized into four major categories.

Conclusion: A suitable infrastructure for precision medicine should be capable of integrating and storing heterogeneous data from diverse departments and sources. By leveraging big data management experiences from other industries and aligning their characteristics with those in precision medicine, it is possible to facilitate the implementation of precision medicine while avoiding duplication.

Keywords

Information Storage and Retrieval; Precision Medicine; Medical Informatics; Data Accuracy; Public Health Infrastructure

Introduction

The National Institute of Health (NIH) defines precision medicine (PM) as “an emerging approach for disease treatment and prevention that takes into account individual variability in genes, environment, and lifestyle for each person” [1]. In simple terms, this approach involves physicians using information from a patient’s genes, environment, and lifestyle in conjunction with conventional medical knowledge to make diagnostic and treatment decisions. It provides physicians with a better understanding of the underlying mechanisms of diseases and therefore increases the potential to predict patients’ future health conditions. Consequently, it is more effective in selecting

*Corresponding author: Mehdi Hajebrahimi
Department of Health Information Management, School of Health Management and Information Sciences, Iran University of Medical Sciences, Tehran, Iran
E-mail: m.hajebrahimi@gmail.com

Received: 20 February 2024
Accepted: 9 March 2024

optimal strategies for preventing, diagnosing, and treating a patient's medical issues [2].

Working with diverse and heterogeneous data as well as data sources in precision medicine has led to differences in collecting, storing, and retrieving information and changes in workload, compared to common information systems. We face a wide range of big data in healthcare, which have not been experienced before [3]. Handling these large volumes of information is crucial for PM. Researchers are actively seeking a scalable infrastructure that can effectively integrate and store various data types. This infrastructure is vital for the successful implementation of the PM [4].

The collection, storage, and retrieval of information in health information systems are performed at three levels: data, information, and knowledge. Health information systems generate substantial volumes of data on a daily basis, which are then stored in integrated or distributed databases. These systems encompass decision support systems and information dashboards, both in clinical and administrative fields that play a pivotal role in converting these data into actionable knowledge [5].

Researchers, aware of the challenges and opportunities posed by this situation, are actively addressing the vast amounts of data by developing methods and technological tools to effectively manage and harness them. Additionally, it's essential to consider big data policies when handling PM data.

This paper, explored the fundamental traits of big data, specifically volume, variety, and velocity (often referred to as the 3Vs), which are widely recognized by researchers. The aim was to understand the manifestation of these characteristics in the context of precision medicine. The study considered the appropriate storage infrastructure and related issues based on these three major characteristics of big data (3Vs). Subsequently, recent storage techniques and technologies, data quality, cloud technology facilities, data storage, standards of data sharing, and finally, data security and privacy

were considered and discussed from the medical informatics perspective.

Material and Methods

The Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) method was used for literature review in this study [7].

Data sources

Data were collected by searching four major databases: Web of Science (WOS), PubMed, Scopus, and Embase, as well as the Google Scholar search engine (which was used to retrieve relevant grey literature) (see Figure 1).

The search spanned from 2015 to 2023, a period chosen due to a significant increase in publications related to precision medicine from 2015 [8].

The following keywords were applied to the title/abstracts and subject fields of the databases (using Mesh and EMtree terms). The final literature search was conducted in July 2023. The language filter was set to English.

Search strategy included two groups of keywords:

- 1) keywords identifying "precision medicine" and related terms and synonyms
- 2) keywords identifying "data storage and retrieval" and related terms and synonyms

These two groups of keywords were merged by using Boolean operators. Additionally, truncation symbols, phrase searching, and other search techniques were used to enhance the comprehensiveness and specificity of the search.

Eligibility criteria

The inclusion and exclusion criteria are provided in Table 1.

Study Selection

Two authors screened the studies in the initial phase based on their titles and abstracts. Subsequently, the eligible studies were selected, and the authors independently screened

the full texts of the studies according to the inclusion and exclusion criteria. In case of any discrepancies, the authors collaboratively reviewed the cases to resolve disagreements and reach a consensus.

Data Extraction

Based on the study objectives, data extraction was performed, and all data were migrated to a Microsoft Excel worksheet featuring

specific data categories: “name of the author,” “publication year,” “title of the study,” “volume characteristics,” “variety characteristics,” “velocity characteristics,” “standards,” and “new theme.” The authors independently conducted data selection for each source and documented the findings in the spreadsheet. In cases of discrepancies in the collected data, the authors collaboratively worked together to resolve the differences and reach a consensus.

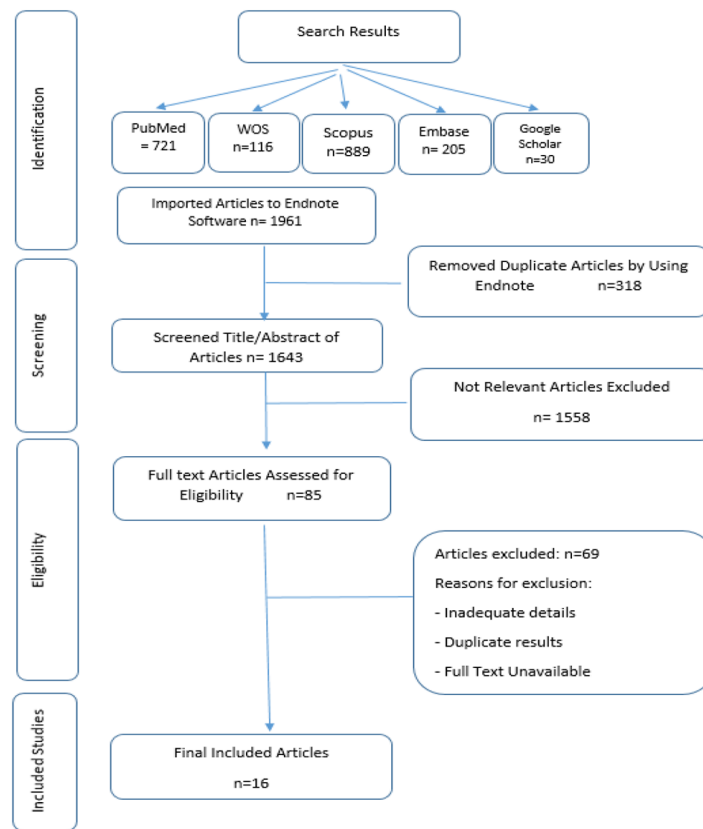


Figure 1: PRISMA flow diagram of data collection based on central search theme

Table 1: Inclusion and exclusion criteria

Inclusion criteria	Exclusion criteria
- Published from 2015 to 2023	- Articles whose full text can not be accessed
- Study types: several types of studies (RCTs, review articles, original articles, paper proceedings ...)	- Not relevant full text articles
- The study must published in English language	

RCTs: Randomized Control Trials

The extracted data from each source were categorized and analyzed by using inductive thematic analysis, aligning with the key concepts which emerged from the dataset.

Results

The initial studies (1961 studies), found by the strach strategy, were exported to Endnote software 9. Then, 318 duplicate studies were identified and removed. The rest of the studies were screened on the basis of their titles and abstracts, and 1558 irrelevant articles were excluded, leaving a total of 85 articles for further examination.

The remaining studies underwent a thorough review of their full texts to determine their eligibility. Ultimately, 16 articles met the eligibility criteria and were included in the study (see Table 2). A visual representation of the complete article selection process is presented in Figure 1.

The articles were mainly cited for general reference on the topic even though their primary content did not directly align with Precision Medicine. Furthermore, we made reference to several websites and blogs that offered introductory-level content, including PM global initiatives or specific-level websites, such as those affiliated with standard organizations or cloud computing service corporations.

One of the primary findings of the study was the classification of the “variety” of PM data (Figure 2).

Findings derived from various studies reviewed in accordance with the research objectives, are presented in Figure 3. The results will be discussed in detail in the following section.

Discussion

PM Data and Its Major Characteristics in the Context of Big Data’s 3Vs

In general, the challenges encountered in data storage and retrieval within precision

medicine reflect those faced in handling big data; consequently, the solutions used to manage big data can be extensively used in precision medicine.

Volume: Precision medicine on the shoulders of giant data

The volume of medical and biological science information and publications has been increasing exponentially since 1879 when the publication of the largest bibliographic index, i.e. Index Medicus, began in the fields of life science and biomedical science information. In the 1960s, the term “information explosion” was commonly used due to the challenges encountered in collecting and storing information [25]. The constant use of this term indicates that the speed of information production has always exceeded the advancement of technologies pertinent to their storage and management.

After the advent of the Internet, there has been a significant increase in health information systems and computer transactions in hospitals, organizations, and health centers. This, as well as the expansion of e-health, tele monitoring, mHealth, and IOT has resulted in a large volume of data flooding the field of health and treatment. Therefore, researchers have recently started to use another term: Data Flood [26].

To this huge volume of data, add genomic data, which can compete with the data generated from big data sources, such as Twitter, YouTube and astronomical data [27]. In addition, data from platforms, such as Twitter, YouTube, Instagram, and Facebook can also be considered as a subset of precision medicine data and can be processed to extract information about individuals’ lifestyle and environment [28].

Variety

With various data from multiple sources, heterogeneity is considered a natural property of PM data, and data variety is the most challenging characteristics in data management [29].

Table 2: Descriptive summary of included articles based on central theme of study

Authors	Year	Description(core subject)	Reference Type
Karacic Zanetti et al. [9]	2023	Ethical, privacy and security issues in data storage in Health wallets	Article
Kaul et al. [10]	2017	Challenges related to the limited analytic capabilities of existing computational and storage infrastructure of genomic data	Article
Kumar et al. [11]	2020	Proposing a hybrid approach-WBTC (Word Based Compression Technique) based on statistical and substitution model for genome data compression	Article
Lynch et al. [12]	2023	Qualitatively explore the Australian public's views and preferences for storing and sharing genomic data	Article
Mansouri et al. [13]	2017	cloud-based data stores	Article
Nepal et al. [14]	2017	Presenting and describing a trusted storage cloud for scientific workflows, called TruXy.	Article
Peng et al. [15]	2019	Using automated storage system for biobanks in Wuhan University, China	Article
Saranya et al. [16]	2022	Reviewing major strategies, challenges, and current developments in information management, storage space, and information retrieval	Book chapter
Vatian et al. [17]	2019	Proposing a system for storing the creative associations of the doctor	Article
Agrawal et al. [18]	2016	Big data storage and management	Article
Anžel et al. [19]	2021	A survey to rank different storage properties adapted for visualization and reporting different storage devices over time while ranking them by their properties	Article
Doricchi et al. [20]	2022	A review on approaches to DNA data storage	Article
Yang et al. [21]	2021	Overview of current memory systems and advances in data storage technology	Article
Gupta et al. [22]	2023	Artificial intelligence in healthcare data management for precision medicine	Book
Kalra et al. [23]	2019	Introducing EMIF platform that provide integrated framework for the large-scale health and life sciences data in precision medicine	Article
Symvoulidis et al. [24]	2021	Introducing an EHR cloud-based system that utilizes an object storage architecture to store healthcare data	Article

EMIF: European Medical Informatics Framework, EHR: Electronic Health Record

Many researchers define variety as different forms of data in a dataset, including structured, semi-structured, and unstructured data [6, 29, 30]; however, it is too complicated than this definitions. Abawayj presents taxonomy of big data variety into four classifications:

structure diversity, source diversity, content diversity, and processing diversity [31]. After reviewing previous studies, we categorized PM data variety into four major classes.

1) Variety in data types [30]: (e.g. omics, signals, images, audios, videos, texts, transac-

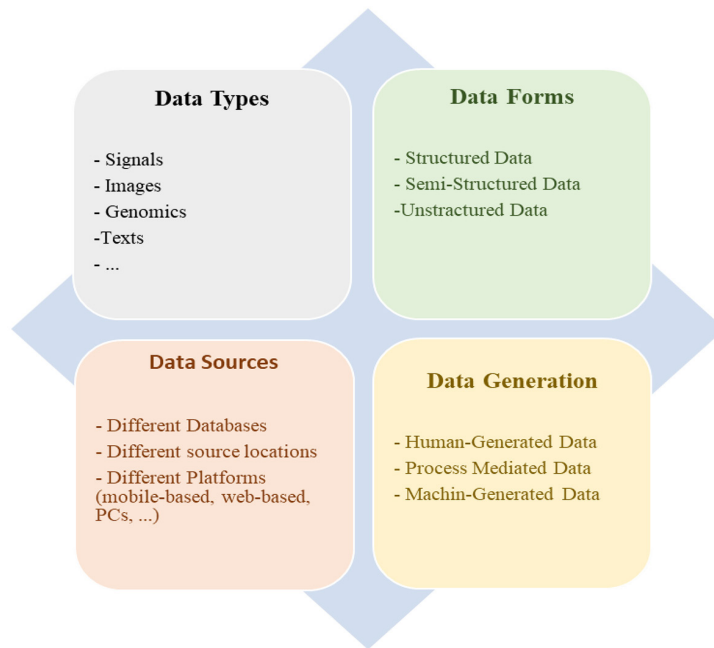


Figure 2: Variety of precision medicine data in four major categories.

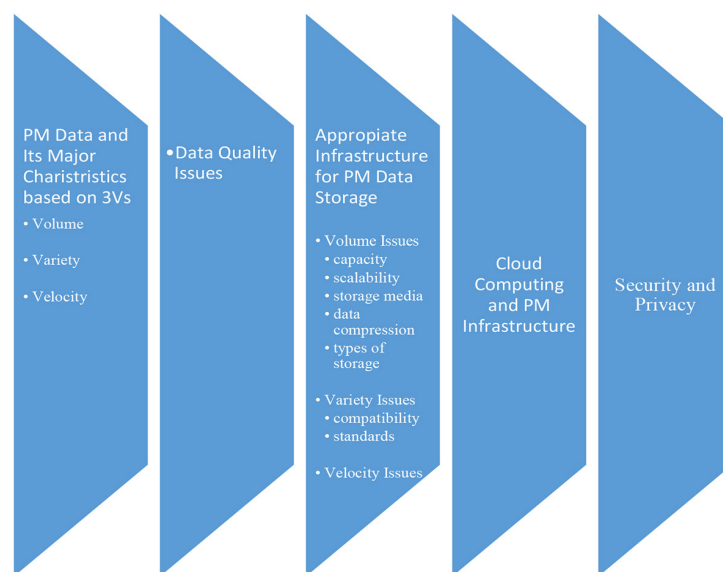


Figure 3: A graphical representation of results

tional data, ...)

2) Variety in forms of data [6, 29, 30]: structured, semi-structured, and unstructured data

3) Variety and heterogeneity in data sources [29, 32]: different databases which may be also in different locations (e.g. Genomic databases, mobile-based monitoring applications data, biobanks, wearable biosensors data, ...)

4) Variety in types of data generation: human generated, process mediated, and machine generated [31, 33]:

a. Human-generated data are created as a result of human- computer interactions [31]. These data might be generated by healthcare personnel or by patients, e.g. Emails, Genomic data, physician interpretation notes on Rx images, CPOE intervention orders or even social media posts, and so on.

b. Process- mediated data: “This is the information that concerns some business events of interest, like the purchase of a camera in an e-commerce site or the sign-up of clients in a system” [33].

c. Machine-generated data: generally refers to data automatically generated by devices, (31) such as wearable sensors and biosensors, video surveillance, IOT data, and so on (see Figure 2).

Velocity:

Velocity grants us access to continuous and real-time data. It also enables predicting a patient’s condition at any given moment. Hulsen believes that although not utilizing a large volume of valuable retrospective data is a waste of resources, medical data are being generated at such a speed that their control and exploitation can be more valuable [34]. This is especially crucial due to the ongoing influx of health-related data from various sources. Innovations like wearable devices, sensors, and continuous monitoring tools provide real-time updates on a patient’s health status. The constant flow of these data allows for immediate insights, contributing to timely decision-making, personalized treatment strategies, and improved patient outcomes [29, 32].

Transforming big data into useful knowledge for real time use at the patient’s bedside, through high speed analysis and processing techniques, can turn the challenge of big data into its most useful feature. Therefore, the physician will have the most up-to-date analyses at the patient’s bedside and can make the fastest and the best decisions about the patient.

Efficient management of data velocity involves handling and processing large volumes of data. It requires a robust infrastructure, advanced analytics, and decision support systems capable of swiftly processing, analyzing, and interpreting data to derive meaningful insights.

Data quality: which data should be stored?

When considering data storage, the primary concern revolves around determining the specific data to be stored. Given the considerable volume, costs, and distinct objectives, it becomes evident that it is neither feasible nor necessary to store every piece of generated data. Numerous data sources generate vast quantities of data on a daily basis; however, a significant portion of these data lacks relevance, while many relevant datasets lack the necessary quality for utilization. This concern is particularly pertinent to healthcare data as compared to industrial big data [34].

Data quality (DQ) refers to how well a dataset aligns with a user’s specific requirements [35]. The users’ needs can vary significantly; therefore, the perception of data quality may differ based on individual viewpoints and distinct needs. Indeed, data quality is a multifaceted concept as what may be considered high-quality data for one business could be considered irrelevant for another. The key lies in understanding the user’s specific needs. By aligning data collection objectives with these needs and carefully selecting the volume and type of big data for analysis, we can accurately assess the data quality. Consequently, different data quality elements may carry varying

degrees of importance depending on the context [30].

In general, the most important and common data quality elements in healthcare studies, include accuracy, accessibility, timeliness, credibility, consistency, integrity, completeness, comprehensiveness, uniqueness, coherence, precision, security, relevance, readability, accessibility, and usefulness [30, 35-37].

Furthermore, two often overlooked characteristics of Data Quality (DQ) elements that are crucial for usability and cost implications are authorization (pertaining to the organization's data access permissions, whether unrestricted or restricted) and structure (the expense related to converting unstructured data to structured data) [30]. These attributes can significantly impact user objectives and requirements indicating that data usability hinges on more than just traditional quality elements. Certainly, meaningless, noisy, and non-valuable data cannot help us make informed medical decisions.

Some big data extensions such as validity, value and especially veracity emphasize the importance of data quality although no specific definition for these terms has yet been provided [6].

Veracity has an opposite relationship with the 3Vs. As we face higher volume, variety, and velocity in data generation, we are likely to face worse or undefined veracity.

Veracity highlights three critical characteristics of data: uncertainty, incompleteness, and inconsistency [29]. Among these, uncertainty permeates various aspects of medicine. Even when working with high-quality data, there remains no absolute certainty in the final decision [38]. Thus, if the quality of data in precision medicine is poor, the ultimate outcome could indeed be disastrous. Deleting low-quality, useless, and non-purposeful data has another advantage: it reduces the final data volume, leading to cost savings in the processes of data storage, analysis, and retrieving. However, identifying useless data is not always

straightforward. Hasty deletion of data without a careful plan could deprive us of the inherent benefits of big data. Surprisingly, seemingly unrelated factors often reveal unexpected and valuable insights when analyzed from seemingly non-purposeful data. This serendipity is a crucial and unique advantage of big data analysis. Therefore, exercising intelligence in discerning noisy data from high-quality data becomes vital for successful outcomes.

Data quality can be improved continuously from data generation to analysis and storage. Froukhi et al. presented a seven-step chain called the Big Data Value Chain (BDVC), in which different elements of data quality can reach a higher quality level at each step [39]. The chain includes: data generation, data acquisition, data processing, data storage, data analysis, data visualization, and data exposition. Through this chain, it is possible to transform raw data into high-quality data, high-quality data into knowledge, and knowledge into insights.

Data generation is the first step. Structured and standardized generation of data can greatly reduce the workload in the next steps. In general, the standardization of data sources and data (which are considered important challenges) can help increase the quality of data [40]. Data quality in this phase largely depends on multiple sources from which it originates. These sources are most effective when they are accurate, timely, structured, comprehensive, error-free, and unbiased [41].

The most important step in transforming raw data into high-quality data is the data preprocessing step where various techniques are used in data cleaning, data transformation, data integration, and data reduction. This will enhance the quality of data in terms of completeness, consistency, uniqueness, and validity, and finally turn transform them into valuable data [42].

In all seven steps of BDVC, researchers employ various techniques to improve data quality elements. The improvement and use

of these techniques, such as machine learning [43], data mining [44], data profiling [45], sentiment analysis [36], and credibility analysis [37], can bring us closer to the goals of precision medicine.

Appropriate infrastructure for PM data storage

Infrastructure is one of the most critical issues that should be given priority in the implementation of precision medicine. It is necessary to predict an infrastructure that can accommodate the required space for storing, retrieving, integrating, analyzing, and sharing data in a secure and private environment while ensuring high speed. This infrastructure should be flexible enough to meet both predicted and unpredicted future needs. In order to store data for precision medicine, hardware and software facilities should be provided at least in accordance with the three main characteristics of big data: (volume, variety, and velocity).

Volume

When considering an infrastructure appropriate to the volume of precision medicine data, it is important to take the following cases into consideration.

Capacity

It indicates the space required for storing and processing a given volume of data. In addition, it refers to the necessary space for data archiving and data backup [18].

Scalability

The required infrastructure should be expandable to accommodate population growth, increased data sources, and the integration of new types of data [46].

Storage media

When it comes to data storage or the implementation of specific storage-related goals or policies, it's essential to choose an appropriate storage medium. The most desirable characteristics of storage media include smaller size, faster storage and retrieval, lower energy consumption, greater durability, the ability to

compress, and store more data in less space. Anzel et al. Identified the basic properties of storage media as follows: 1) accessibility, 2) capacity, 3) lifespan, 4) mutability (which defines the functions of a device: write, read, or both), 5) typology (such as optical, magnetic, molecular, etc.), 6) energy use, and 7) data density [19].

Indeed, our journey through data storage technologies has been remarkable. It began with drum memories in 1932, followed by HDDs in 1956, compact discs (CDs) in 1981, and now it encompasses cutting-edge innovations, such as synthetic deoxyribonucleic acid (synthetic DNA) and synthetic metabolomes. These advancements have significantly enhanced our capabilities for long-term storage, increased capacity, improved data density, and more [19, 20]. However, despite this improvement, the constant pace of data generation continues to surpass our progress. The pure volume of data being produced challenges even our most sophisticated technologies. While synthetic DNA and metabolomes hold great promise, they still face barriers—chiefly their high cost and time-consuming processes [20].

Data compression

This is performed to reduce the size of data for storage or transmission [47]. The advancement of technologies has led to the creation of higher-quality data and, thus, a higher volume. For example, advances in medical imaging have provided physicians with highly detailed details, and this trend extends to other types of data, such as audio data and ECGs. Simultaneously, there have been advances in data compression, including methods, techniques, and algorithms which are based on data quality (lossless), data types (video, audio, text, image), application (wireless sensor networks, medical imaging, etc.), and coding schemes (e.g., Huffman coding, and dictionary-based coding) [47]. A vast volume of genomic data is generated, presenting a significant challenge for efficient storage and

transmission in PM [10]. Therefore, in order to tackle this issue, a hybrid lossless genome compression algorithm called WBTC (Word Based Compression Technique) has been developed. Built on statistical and substitution models, WBTC aims to effectively compress vast genomic datasets [11]. Compression technologies can help reduce the required space and cost by minimizing data size.

Types of storage

The use of appropriate type of storage can enhance availability, durability, access speed, scalability, and cost-effectiveness of stored data.

There are three types or architectures for data storage on media: file storage, block storage and object-based storage. Each of which has its own characteristics and is used for storage on different platforms. In file-based storage, data are stored hierarchically, with data stored in files, files stored in folders, and folders stored in directories.

To access the desired data, metadata is used to show the path of the folder, file, and data. In block storage, data are broken into blocks, each of which is stored with its own ID for retrieval. Each of these two methods has its own advantages and disadvantages. In file-based storage, almost everything can be stored, and the retrieval of complex data is easily possible. However, managing large volumes of data or expanding capacity, can be complicated. Despite high performance and high speed in data storage and retrieval, block storage is limited in handling unstructured data, scalability issues and the addition of metadata capabilities [48-50]. In object-based storage, each stored object contains data, metadata, and an ID for accessing data. It is also used for storing large amounts of data, such as genomics and data analytics [51]. Object-based storage is known for customizability, flexibility, and scalability all of which make it suitable for handling unstructured and large data sets. Furthermore, it ensures high availability for stored data at a lower cost in cloud-based infrastructures

[24]. Considering the data characteristics of personalized medicine, which often involve unstructured large volumes of data requiring scalable infrastructure, object-based storage is more appropriate compared to the other two alternatives [18, 24, 51].

Variety

To reduce infrastructure problems related to big data variety, the following issues should be considered.

Compatibility

Appropriate infrastructure for precision medicine must be capable of integrating, storing, analyzing, and sharing all types of data from multiple sources and platforms. This requires compatibility with all existing information systems, data sources, and devices (such as PCs, mobile phones, biosensors, etc.), enabling them to interoperate and allowing the system to receive, store, and analyze all types of inputs and produce appropriate outputs. Compatibility with existing systems is a key factor in the adoption of PM [52].

Standardization and use of standards in the precision medicine ecosystem can contribute to its compatibility with different information systems [53].

Standards

Various standards are used in different aspects of precision medicine data management, including data generation, integration, storage, retrieval, sharing, and exchange between different information systems. The most important issue in PM is to ensure that both the data and their sources use a common language across different processes to reduce data heterogeneity. This common language is known as standards which enable the exchange and communication between different care systems, increase data quality, and reduce their preparation time [54]. Standardization is one of the important challenges in integrating and storing data in precision medicine [55].

The diverse landscape of data and data sources has given rise to a multitude of standards. Choosing the right standards for the care

ecosystem poses both challenges and complexities. However, within this complexity lies an opportunity. While some researchers advocate for creating entirely new standards for precision medicine, the existing standards already exhibit remarkable diversity. By expanding or subtly adjusting these existing standards, we can effectively meet the specific needs of precision medicine. Thus, navigating this intricate terrain offers both challenges and opportunities for innovation [46].

Consequently, the standards used in precision medicine can be divided into three categories:

A- General medical informatics standards. There are standards that can be used in all types of systems and for all purposes. These standards are prerequisites for establishing a common language and interoperability across various computer systems in healthcare. For example, nomenclature standards such as UMLS or SNOMED CT, which facilitate the integration of biomedical and health language in computer systems.

Another example is DICOM, which is a standard for the storage, retrieval, and exchange of medical images [56]. Moreover, FAIR standard defines the basic data characteristics to generate standard data for computer systems [57].

B- Existing standards in medical informatics, which should be expanded to better support the needs of precision medicine, such as security and privacy standards and IHE profiles to address interoperability challenges [58].

C- The standards that have been developed or are currently being developed for use in precision medicine to address the lack of required standards in this field. For example, Advancing Standards for Precision Medicine (ASPM) project which is being implemented by ONC in partnership with NIH and as part of the Precision Medicine Initiative (PMI) project. ASPM project, started in 2018, aims to facilitate the sharing,

aggregating and synthesizing of health data in the areas of m-health, wearable sensors, and social determinants of health (SDOH) data [59].

Velocity

Perhaps the best outcome would be that data can be stored, processed, and utilized with the same speed and efficiency at which they are produced. The use of real time data at the point of care enables the physicians to make the fastest decisions in emergencies; therefore, strong tools for rapid data storage and processing is required. There are three strong technologies which have the capability to store and process massive amounts of data rapidly and can be used for managing PM big data infrastructures: 1- Hadoop clusters with thousands of nodes, that are characterized by low cost, low latency, high flexibility, and scalability, 2- NOSQL technology, which works based on the concept of distributed databases and allows for the storage of unstructured data across multiple nodes, 3- Massively Parallel Processing (MPP) which enables thousands of processors to work on different parts of a program [60, 61].

Cloud computing tools as a solution package for PM

For an organization, it is crucial to have a suitable infrastructure which aligns with the above-mentioned characteristics. This infrastructure must be capable of providing space for integrating and storing significant amount of data and their analysis, securely sharing data and knowledge, enabling high-speed access from anywhere and at any time. However, such an infrastructure requires large data centers, complex networking, and large investment in hardware, software, and human resources. Therefore, few organizations can afford such an infrastructure due to its continuous support and maintenance costs.

Cloud computing based infrastructures are considered as an effective solution to deal with the problems of data acquisition,

integration, storage, distribution, and processing in precision medicine [13].

Nowadays, many healthcare organizations use this technology to achieve a balanced analysis and improve data consistency in their data integration processes [3]. The use of this technology and its related technologies can be significantly effective in saving resources and reducing costs [62]. The significance of cloud computing in the realization of precision medicine is such that it can be stated that the synergy between technology and medicine is crucial for success in this domain.

Major providers of public cloud services, such as MS Azure and Amazon AWS provide various services like Infrastructure as a Service (IaaS) and Storage as a Service (StaaS) to implement large-scale projects which involve big data in their data centers with thousands of servers [18]. Some companies, like Intel, have made significant investments in providing cloud-based services in healthcare sector to offer a variety of services. Advances in scalable cloud computing can help remove limitations in various PM processes. Additionally, cloud-based technologies, such as fog computing and edge computing can reduce costs, increase data security, enhance data processing speed, and facilitate the dynamic, real-time provision of services [63, 64]. Infrastructure as Code methodology (a cloud-based methodology) is an example of cloud-based infrastructure that has been introduced by Frey et al. to accelerate the implementation of precision medicine. It serves as a tool for storing, integrating, and managing various medical data [3].

Security and privacy issues

Ensuring data access, security, and privacy, along with seamless interoperability of systems, remains a critical challenge in data management. Researchers continually seek solutions to overcome these barriers, particularly when dealing with genomic data- a domain that is both powerful and vulnerable to misuse. The sensitivity surrounding genomic

information emphasizes the need for robust safeguards [65-67]. Diagnostic genomic sequencing generates vast quantities of data that besides its primary purpose, holds potential for various secondary applications, such as research and shaping future healthcare for the data contributor. These opportunities might necessitate sharing data with third parties, which could intensify concerns among individuals [12].

Individuals are the primary data owners and can withhold their data from the system. This may impede the progress of precision medicine. Disclosure of such data could potentially lead to threats related to medical records, unemployment, or an increase in insurance premiums [68]. Significant progress has been made in the realms of security and privacy regarding data access and interactions between systems and data sharing [3, 67]. Researchers hope to achieve this goal by adapting strategies tailored to different types of data and increasing public participation. The four basic solutions proposed for this purpose include cryptography, use of blockchain technology, access control and security analysis, and the use of protocols and standards related to network security [65].

Two solutions have also been suggested to ensure the privacy of personal information: anonymization and pseudonymization [68]. Among these, the use of blockchain technology, apart from security, can also be considered in data management. Blockchain can reduce our concern about the integration of big data. Furthermore, safeguarding the privacy of individuals' identity and using tools related to the Internet of Things provide secure access to data and result in the realization of patient-centered medicine [69, 70]. Moreover, blockchain has the potential to gain people's trust by resolving data ownership issues and promoting transparency in precision medicine [70]. Managing data sharing, establishing a secure ecosystem for medical data to conduct collaborative research, and addressing

interoperability issues in EHR systems for e-health are additional blockchain capabilities requiring increased focus [69, 71].

A great number of studies explored the integration of cloud computing and blockchain technologies. Experts believe that these two technologies can effectively address storage, sharing and security issues in precision medicine by covering each other's shortcomings. Cloud computing has been extensively used across various industries through multiple diverse of cloud services. However, the integration of blockchain within these models requires the re-engineering of cloud data centers, presenting researchers with many approaches and challenges [72, 73].

Despite the implementation of various technologies to enhance the privacy of individuals' data, obtaining explicit consent from the data donor could significantly mitigate potential legal and ethical challenges associated with sharing data with third parties [12].

Conclusion

Precision medicine, relies on data from new technologies and requires specialized facilities, technology, and related expenses. Establishing a secure infrastructure capable of integrating and storing heterogeneous data from multiple departments and sources, and subsequently preparing it for processing and analysis, necessitates fundamental alterations in the architecture of the existing systems and infrastructures.

Specialized human resources, hardware and software facilities, technology, budget, and support from policymakers, as well as patients who provide details of their lifestyle and genetic data to the healthcare system will play a significant role in this ecosystem.

Standardization, interoperability among incompatible systems, security, patient satisfaction, data ownership, medical ethics, data quality, unstructured data issues, and scalable storage infrastructure are the key challenges hindering the adoption of PM. Further research

is needed to tackle these challenges.

Training skilled personnel in precision medicine through collaboration among experts in medicine, medical informatics, bioinformatics, epidemiology, and computer science has the potential to revolutionize the future.

Despite the persistence of numerous challenges, the rapid advancements in information and communication technologies offer hope in addressing the existing obstacles and unveiling new horizons in disease prevention and treatment for humanity.

Acknowledgment

This study was supported by Iran university of medical sciences (IUMS); Grant No. 99-1-37-16894.

Authors' Contribution

M. Hajebrahimi was responsible for the conceptualization, methodology, writing of the original draft, and paper appraisal. M. Langarizadeh contributed through supervision, methodology, and editing. All the authors read, modified, and approved the final version of the manuscript.

Ethical Approval

This study is approved by ethical committee of Iran University of Medical Sciences (IR.IUMS.REC.1399.357).

Conflict of Interest

None

References

1. National Institute of Health (NIH). What is precision medicine? 2022. Available from: <https://medlineplus.gov/genetics/understanding/precisionmedicine/definition/>.
2. National Institute of Health (NIH). What are some potential benefits of precision medicine and the Precision Medicine Initiative? 2022. Available from: <https://medlineplus.gov/genetics/understanding/precisionmedicine/potentialbenefits/#:~:text=Wider%20ability%20of%20doctors%20to,by%20which%20various%20diseases%20occur>.
3. Frey LJ. Data integration strategies for predic-

- tive analytics in precision medicine. *Per Med.* 2018;**15**(6):543-51. doi: 10.2217/pme-2018-0035. PubMed PMID: 30387695. PubMed PMCID: PMC6277956.
4. Liu X, Luo X, Jiang C, Zhao H. Difficulties and challenges in the development of precision medicine. *Clin Genet.* 2019;**95**(5):569-74. doi: 10.1111/cge.13511. PubMed PMID: 30653655.
 5. Healthcare Information and Management Systems Society (HIMSS). Health informatics. 2023. Available from: <https://www.himss.org/resources/health-informatics>.
 6. Emmanuel I, Stanier C. Defining big data. *Proceedings of the International Conference on Big Data and Advanced Wireless Technologies*; 2016. p. 1-6.
 7. Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Int J Surg.* 2010;**8**(5):336-41. doi: 10.1016/j.ijsu.2010.02.007. PubMed PMID: 20171303.
 8. Frey LJ, Bernstam EV, Denny JC. Precision medicine informatics. *J Am Med Inform Assoc.* 2016;**23**(4):668-70. doi: 10.1093/jamia/ocw053. PubMed PMID: 27274018. PubMed PMCID: PMC4926751.
 9. Karacic Zanetti J, Nunes R. To wallet or not to wallet: the debate over digital health information storage. *Computers.* 2023;**12**(6):114. doi: 10.3390/computers12060114.
 10. Kaul G, Shah ZA, Abouelhoda M. A high performance storage appliance for genomic data. *Bioinformatics and Biomedical Engineering. Lecture Notes in Computer Science*, vol 10209. Granada, Spain: Springer, Cham; 2017. p. 480-8.
 11. kumar S, Agarwal S, Ranvijay. WBTC: a new approach for efficient storage of genomic data. *Int J Info Technol.* 2020;**12**(3):915-21. doi: 10.1007/s41870-020-00472-2.
 12. Lynch F, Meng Y, Best S, Goranitis I, Savulescu J, Gyngell C, Vears DF. Australian public perspectives on genomic data storage and sharing: Benefits, concerns and access preferences. *Eur J Med Genet.* 2023;**66**(1):104676. doi: 10.1016/j.ejmg.2022.104676. PubMed PMID: 36473622.
 13. Mansouri Y, Toosi AN, Buyya R. Data storage management in cloud environments: Taxonomy, survey, and future directions. *ACM Comput Surv.* 2017;**50**(6):1-51. doi: 10.1145/3136623.
 14. Nepal S, Sinnott RO, Friedrich C, Wise C, Chen S, et al. TruXy: Trusted storage cloud for scientific workflows. *IEEE Trans Cloud Comput.* 2015;**5**(3):428-42. doi: 10.1109/TCC.2015.2489638.
 15. Peng H, Qian K, Xiao Y, Zhang S, Chen H, Zhou Z, et al. Application of automated storage system in the operation of biobanking. *Biopreserv Biobank.* 2019;**17**(3):A60-A1.
 16. Saranya SM, Tamilselvi K, Mohanapriya S. Harnessing big data and artificial intelligence for data acquisition, storage, and retrieval of healthcare informatics in precision medicine. In: *Healthcare 40: Health Informatics and Precision Data Management*; 2022. p. 51-75.
 17. Vatian A, Ratnikova A, Gruntov A, Osipov S, Shalyto A, Gusarova N. Using associative links for storing personalized medical information. In: *Multi Conference on Computer Science and Information Systems, MCCSIS 2019-Proceedings of the International Conference on e-Health*; 2019. p. 211-5.
 18. Agrawal R, Nyamful C. Challenges of big data storage and management. *Glob J Info Technol.* 2016;**6**(1):1-10. doi: 10.18844/gjit.v6i1.383.
 19. Anžel A, Heider D, Hattab G. The visual story of data storage: From storage properties to user interfaces. *Comput Struct Biotechnol J.* 2021;**19**:4904-18. doi: 10.1016/j.csbj.2021.08.031. PubMed PMID: 34527195. PubMed PMCID: PMC8430386.
 20. Doricchi A, Platnich CM, Gimpel A, Horn F, Earle M, Lanzavecchia G, et al. Emerging Approaches to DNA Data Storage: Challenges and Prospects. *ACS Nano.* 2022;**16**(11):17552-71. doi: 10.1021/acsnano.2c06748. PubMed PMID: 36256971. PubMed PMCID: PMC9706676.
 21. Yang JQ, Zhou Y, Han ST. Functional applications of future data storage devices. *Adv Electron Mater.* 2021;**7**(5):2001181. doi: 10.1002/aelm.202001181.
 22. Gupta NS, Kumar P. Perspective of artificial intelligence in healthcare data management: A journey towards precision medicine. *Comput Biol Med.* 2023;**162**:107051. doi: 10.1016/j.compbiomed.2023.107051. PubMed PMID: 37271113.
 23. Kalra D. The importance of real-world data to precision medicine. *Per Med.* 2019;**16**(2):79-82. doi: 10.2217/pme-2018-0120. PubMed PMID: 30724116.
 24. Symvoulidis C, Kiourtis A, Mavrogiorgou A, Kyriazis D. Healthcare Provision in the Cloud: An EHR Object Store-based Cloud Used for Emergency. *J Healthinf.* 2021;**1**:435-42.
 25. Green JC. The Information Explosion-- Real or Imaginary? *Science.* 1964;**144**(3619):646-8. doi: 10.1126/science.144.3619.646. PubMed PMID: 17806987.
 26. Jeurgens C. Threats of the data-flood: An accountability perspective in the era of ubiquitous computing. In: *Archives in Liquid Times*. Amsterdam: Stichting Archiefpublicaties; 2017. p. 196-210.

27. Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, Efron MJ, et al. Big Data: Astronomical or Genomical? *PLoS Biol.* 2015;**13**(7):e1002195. doi: 10.1371/journal.pbio.1002195. PubMed PMID: 26151137. PubMed PMCID: PMC4494865.
28. Prospero M, Min JS, Bian J, Modave F. Big data hurdles in precision medicine and precision public health. *BMC Med Inform Decis Mak.* 2018;**18**(1):139. doi: 10.1186/s12911-018-0719-2. PubMed PMID: 30594159. PubMed PMCID: PMC6311005.
29. Hariri RH, Fredericks EM, Bowers KM. Uncertainty in big data analytics: survey, opportunities, and challenges. *J Big data.* 2019;**6**(1):1-6. doi: 10.1186/s40537-019-0206-3.
30. Cai L, Zhu Y. The challenges of data quality and data quality assessment in the big data era. *Data Sci J.* 2015;**14**:1-10. doi: 10.5334/DSJ-2015-002.
31. Abawajy J. Comprehensive analysis of big data variety landscape. *Int J Parallel, Emergent Distrib Syst.* 2015;**30**(1):5-14. doi: 10.1080/17445760.2014.925548.
32. Blazquez D, Domenech J. Big Data sources and methods for social and economic analyses. *Technol Forecast Soc Change.* 2018;**130**:99-113. doi: 10.1016/j.techfore.2017.07.027.
33. Arolfo F, Vaisman A, editors. Data Quality in a Big Data Context. In: *Advances in Databases and Information Systems*; Cham: Springer International Publishing; 2018.
34. Hulsen T, Jamuar SS, Moody AR, Karnes JH, Varga O, Hedensted S, et al. From Big Data to Precision Medicine. *Front Med (Lausanne).* 2019;**6**:34. doi: 10.3389/fmed.2019.00034. PubMed PMID: 30881956. PubMed PMCID: PMC6405506.
35. Ehsani-Moghaddam B, Martin K, Queenan JA. Data quality in healthcare: A report of practical experience with the Canadian Primary Care Sentinel Surveillance Network data. *Health Inf Manag.* 2021;**50**(1-2):88-92. doi: 10.1177/1833358319887743. PubMed PMID: 31805788.
36. Alaoui IE, Gahi Y. The Impact of Big Data Quality on Sentiment Analysis Approaches. *Procedia Comput Sci.* 2019;**160**:803-10. doi: 10.1016/j.procs.2019.11.007.
37. Zan S, Zhang X. Medical data quality assessment model based on credibility analysis. In: *IEEE 4th Information Technology and Mechatronics Engineering Conference (ITOEC)*; Chongqing, China: IEEE; 2018. p. 940-4.
38. Hatch S. Uncertainty in medicine. *BMJ.* 2017;**357**:j2180. doi: 10.1136/bmj.j2180. PubMed PMID: 28495912.
39. Faroukhi AZ, El Alaoui I, Gahi Y, Amine A. Big data monetization throughout Big Data Value Chain: a comprehensive review. *Journal of Big Data.* 2020;**7**:1-22. doi: 10.1186/s40537-019-0281-5.
40. Chaney K, Caban TZ, Rogers CC, Denny JC, White J. Health IT Advances Precision Medicine. *ONC: Health IT Buzz*; 2023.
41. Allen M. Successful Precision Medicine Hinges on High Quality Data. 2023. Available from: <https://verinovum.com/2022/05/precision-medicine-hinges-on-high-quality-data/#section3>.
42. Lincy SB, Kumar NS. An enhanced pre-processing model for big data processing: A quality framework. In *International Conference on Innovations in Green Energy and Healthcare Technologies (IGEHT)*; Coimbatore, India: IEEE; 2017. p. 1-7.
43. Mylavarapu G, Thomas JP, Viswanathan KA. An automated big data accuracy assessment tool. 4th International Conference on Big Data Analytics (ICBDA); Suzhou, China: IEEE; 2019. p. 193-7.
44. Taleb I, Serhani MA, Dssouli R. Big data quality assessment model for unstructured data. In *International Conference on Innovations in Information Technology (IIT)*; Al Ain, United Arab Emirates: IEEE; 2018. p. 69-74.
45. Capiello C, Samá W, Vitali M. Quality awareness for a successful big data exploitation. In *Proceedings of the 22nd International Database Engineering & Applications Symposium*; Villa San Giovanni, Italy: Association for Computing Machinery; 2018. p. 37-44.
46. McPadden J, Durant TJ, Bunch DR, Coppi A, Price N, Rodgeron K, et al. Health Care and Precision Medicine Research: Analysis of a Scalable Data Science Platform. *J Med Internet Res.* 2019;**21**(4):e13043. doi: 10.2196/13043. PubMed PMID: 30964441. PubMed PMCID: PMC6477571.
47. Jayasankar U, Thirumal V, Ponnurangam D. A survey on data compression techniques: From the perspective of data quality, coding schemes, data type and applications. *J King Saud Univ - Comput Inf Sci.* 2021;**33**(2):119-40. doi: 10.1016/J.JKSUCI.2018.05.006.
48. Dey E. All about Object Storage, File Storage, and Block Storage. 2021. Available from: <https://www.arvancloud.ir/blog/en/object-storage-vs-file-storage-vs-block-storage/>.
49. Sheldon R. Understanding object storage vs. block storage for the cloud. 2023. Available from: <https://www.techtarget.com/searchstorage/tip/Understanding-object-storage-vs-block-storage-for-the-cloud>.
50. Staimer M. Object vs. block vs. file storage: Which is best for cloud apps? 2023. Available from: <https://www.techtarget.com/searchstorage/tip/Object-storage-vs-file-storage-for-cloud-applications>.

51. Vergadia P. A map of storage options in Google Cloud. 2022. Available from: <https://cloud.google.com/blog/topics/developers-practitioners/map-storage-options-google-cloud>.
52. Park JH, Kim YB. Factors activating big data adoption by Korean firms. *J Comput Info Syst*. 2021;**61**(3):285-93. doi: 10.1080/08874417.2019.1631133.
53. Fukami Y. Open and Clarified Process of Compatibility Standards for Promoting Data Exchange. *Rev Socionetwork Strateg*. 2021;**15**(2):535-55. doi: 10.1007/s12626-021-00087-4. PubMed PMID: 35506053. PubMed PMCID: PMC8498773.
54. Wager KA, Lee FW, Glaser JP. Health care information systems: a practical approach for health care management. New York: John Wiley & Sons; 2021.
55. Afzal M, Islam SR, Hussain M, Lee S. Precision medicine informatics: principles, prospects, and challenges. *IEEE Access*. 2020;**8**:13593-612. doi: 10.1109/ACCESS.2020.2965955.
56. Shortliffe EH, Shortliffe EH, Cimino JJ, Cimino JJ. Biomedical informatics: computer applications in health care and biomedicine. New York: Springer; 2014.
57. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data*. 2016;**3**:160018. doi: 10.1038/sdata.2016.18. PubMed PMID: 26978244. PubMed PMCID: PMC4792175.
58. Lee KH, Urtnasan E, Hwang S, Lee HY, Lee JH, Koh SB, Youk H. Concept and Proof of the Life-log Bigdata Platform for Digital Healthcare and Precision Medicine on the Cloud. *Yonsei Med J*. 2022;**63**(Suppl):S84-92. doi: 10.3349/ymj.2022.63.S84. PubMed PMID: 35040609. PubMed PMCID: PMC8790588.
59. Office of the National Coordinator for Health Information Technology (ONC). Advancing Standards for Precision Medicine. 2023. Available from: <https://www.healthit.gov/topic/advancing-standards-precision-medicine>.
60. Sumit Gupta S. Real-Time Big Data Analytics. Birmingham: Packt; 2016. Available from: <https://www.perlego.com/book/4451/realtime-big-data-analytics-pdf>.
61. Edmondson J. Understanding big data infrastructure - a complete guide. 2022. Available from: <https://www.businesstechweekly.com/operational-efficiency/data-management/big-data-infrastructure>.
62. Dash S, Ahmad M, Iqbal T. Mobile cloud computing: a green perspective. Intelligent Systems, Lecture Notes in Networks and Systems; Singapore: Springer; 2021. p. 523-33.
63. Divaris K. Fundamentals of Precision Medicine. *Compend Contin Educ Dent*. 2017;**38**(8 Suppl):30-2. PubMed PMID: 29227115. PubMed PMCID: PMC5880533.
64. Oueida S, Kotb Y, Aloqaily M, Jararweh Y, Baker T. An Edge Computing Based Smart Healthcare Framework for Resource Management. *Sensors (Basel)*. 2018;**18**(12):4307. doi: 10.3390/s18124307. PubMed PMID: 30563267. PubMed PMCID: PMC6308405.
65. Blasimme A, Fadda M, Schneider M, Vayena E. Data Sharing For Precision Medicine: Policy Lessons And Future Directions. *Health Aff (Millwood)*. 2018;**37**(5):702-9. doi: 10.1377/hlthaff.2017.1558. PubMed PMID: 29733719.
66. Raza S, Hall A. Genomic medicine and data sharing. *Br Med Bull*. 2017;**123**(1):35-45. doi: 10.1093/bmb/ldx024. PubMed PMID: 28910995. PubMed PMCID: PMC5862236.
67. Raisaro JL, Troncoso-Pastoriza JR, El-Zein Y, Humbert M, Troncoso C, Fellay J, Hubaux JP. GenoShare: Supporting Privacy-Informed Decisions for Sharing Individual-Level Genetic Data. *Stud Health Technol Inform*. 2020;**270**:238-41. doi: 10.3233/SHTI200158. PubMed PMID: 32570382.
68. Thapa C, Camtepe S. Precision health data: Requirements, challenges and existing techniques for data security and privacy. *Comput Biol Med*. 2021;**129**:104130. doi: 10.1016/j.compbiomed.2020.104130. PubMed PMID: 33271399.
69. Shae Z, Tsai JJ. On the design of a blockchain platform for clinical trial and precision medicine. In 37th international conference on distributed computing systems (ICDCS); Atlanta, GA, USA: IEEE; 2017. p. 1972-80.
70. Hashim F, Harous S. Precision medicine block-chained: a review. In 2nd International Conference on Computation, Automation and Knowledge Management (ICCAKM); Dubai, United Arab Emirates: IEEE; 2021. p. 48-52.
71. Alonso SG, Arambarri J, López-Coronado M, de la Torre Díez I. Proposing New Blockchain Challenges in eHealth. *J Med Syst*. 2019;**43**(3):64. doi: 10.1007/s10916-019-1195-7. PubMed PMID: 30729329.
72. Pavithra S, Ramya S, Prathibha S. A survey on cloud security issues and blockchain. In 3rd International Conference on Computing and Communications Technologies (ICCCT); Chennai, India: IEEE; 2019. p. 136-40.
73. Gai K, Guo J, Zhu L, Yu S. Blockchain meets cloud computing: A survey. *J IEEE Commun Surv Tutor*. 2020;**22**(3):2009-30. doi: 10.1109/COMST.2020.2989392.