Integration of genomics data and electronic health records toward personalized medicine: A targeted review

Ali Najafi¹, Neda Emami², Taha Samad-Soltani²*²

¹Molecular Biology Research Center, Systems Biology and Poisonings Institute, Baqiyatallah University of Medical Sciences, Tehran, Iran
²Department of Health Information Technology, Faculty of Management and Medical Informatics, Tabriz University of Medical Sciences, Tabriz, Iran

**Article Info**

**Article type:** Review

**Article History:**
Received: 2021-04-26
Accepted: 2021-08-17
Published: 2021-08-22

**Corresponding author:**
Taha Samad-Soltani
Department of Health Information Technology, Faculty of Management and Medical Informatics, Tabriz University of Medical Sciences, Tabriz, Iran

Email: samadsoltani@tbzmed.ac.ir

**Keywords:**
Genetics
Personalized Precision
Electronic Record Pharmacogenomics

Introduction: Integration of rapidly expanding high-throughput omics technologies and electronic health record (EHR) has created an unprecedented advantage in terms of acquiring routine healthcare data to accelerate genetic discovery. In this regard, EHR can also provide several important advantages to omics research if the integration challenges are well handled. The main purpose of the present study was to review available and published knowledge in the related literature and then to classify and discuss stakeholders’ requirements in this domain.

Material and Methods: At first, a broad electronic search of all available literature in English was conducted on the topic through a search in the databases of Medline, Web of Science, Institute of Electrical and Electronics Engineers (IEEE), Scopus, and Cochrane. Then, stakeholders’ requirements were tabulated, and finally, a word cloud was generated and analyzed to achieve functional and non-functional cases.

Results: A total of 81 articles were included in the given analysis. Integration requirements also consisted of nine functional cases including a uniform approach to the interpretation of genetic tests, standardized terminologies and ontologies, structured data entry as much as possible, an integrated online patient portal, multiple data source handling, machine-readable storing and reporting, research-oriented requirements, pharmacogenomics decision support capabilities, and phenotyping algorithms and knowledge base. Besides, there were three non-functional cases comprised of interoperability of multiple systems, ethical, legal, security factor, and big data computations.

Conclusion: The main challenges in this way could also have semantic and technical themes. Therefore, system developers could guarantee the success of systems by overcoming the given challenges.

Cite this paper as:

INTRODUCTION

Large-scale investments in fundamental sciences have resulted in major advances in clinical medicine. Researchers have also discovered hundreds of genes and constantly examined their structures, functions, and behaviors. In this respect, personalized or precision medicine (PM) has been recognized as a wide and rapidly expanding area of healthcare in which clinical, genomic and other omics characteristics of individuals are being discussed. Accordingly, a combination of these data sources has been labeled as multi-omics. Healthcare embracing PM similarly provides an integrated and evidence-based approach for continuous delivering and persistent care in an individual manner. In this emerging area, genomic medicine is being used to achieve a molecular understanding of a disease in terms of the development of prevention and drug strategies at the early stages of the disease and even before its occurrence [1, 2].

PM hopes to provide new tools for promoting accuracy, prevention, and effectiveness of healthcare. This promising vision is relying on the usefulness of applications derived from modern knowledge in biological systems as well as its combination with
information technology (IT) power. Therefore, revision and revolution in various sectors of health systems within this quick move can be inevitable. For example, innovations and their related costs provided by policy-makers, adequate evidence, and comparative research together with their standards that are necessary to formulate solutions concerning the necessity of establishing PM based on clinical evidence, data standards within clinical research, patient safety, and public health, consumer care tools as well as health-related IT [3, 4].

Accordingly, the integration of rapidly expanding high-throughput genotyping technologies and EHR has established an unprecedented advantage in terms of acquiring routine healthcare data to accelerate genetic discovery. Healthcare systems will also form such visions by developing integrated biobanks of EHR [4]. Moving towards EHRs has even led to the accessibility of medical information. However, the low quality and granularity of the recorded data have limited their usage in the domains of omics research and evidence-based medicine (EBM). The given challenge equally varies between healthcare institutes, and different infrastructures of EHR have brought about different outcomes. Hence, a national and pervasive model can be assume as the best choice for incorporating multi-omics data into the EHR. In this respect, personalized or precision medical interventions require more genome sequencing, generating big data and analyzing them, as well as linking outcomes to EHR structure [5]. As the challenges of generating and storing multi-omics data are gradually met, the challenges of management, integration, interoperability, analysis and interpretation of latent knowledge are on an increase more and more. Thus, there is a requirement to create a bridge to the integration gap [6].

EHR can also provide several important advantages to omics research, including cost savings, high accessibility of clinical big data, and the ability to process data in the form of time series as the most important ones. The ability to reuse this data can also be provided at a low cost to researchers by applying EHR integrated with multi-omics data. Additionally, it is possible to carry out extensive meta-analysis studies. The findings in this line have shown that EHR-based approaches save about 82% of the costs of each sample. In addition, some studies have focused on reducing the design and implementation time of research compared with that of traditional methods [7-9].

The other advantage of the integration of EHR with molecular information is to provide an analytical platform for a large amount of data created and to support this platform for establishing large cohorts aimed at subsequent analysis, which will generate a new perspective. In this regard, a study had also mentioned that the average number of samples in biobanks had reached 460,000 records and it was growing dramatically [10]. Another advantage of an integrated infrastructure between multi-omics data and EHR is associating with the accessibility of trends in an optional time series, representing a valuable investment in genetic research. Moreover, genotypes and phenotypes can be discussed during naturally occurring time intervals such as disease and drug side effects and their progress, survival rates, as well as responses to treatments. Therefore, low-cost information update in EHR leads to high accuracy in omics computational algorithms [11-13].

Considering the importance of integration of multi-omics data into EHR, the present study was conducted to propose an applied solution by reviewing available and published knowledge in the related literature and presenting object-oriented modeling and prototyping in the form of a standard conceptual model for effective integration.

The main objectives of the present study included:

- Extracting and organizing available knowledge about genomic-enabled challenges and requirements of EHR, using an umbrella review of stakeholders’ viewpoints; and,
- Modeling and discussing stakeholders’ functional and non-functional requirements.

**MATERIAL AND METHODS**

The present study fulfilled two major phases to achieve the pre-determined objectives below:

First, a wide electronic search of all published items in English was conducted until 30 July 2020 using five databases: PubMed (the National Center for Biotechnology Information (NCBI)), Web of Science, Scopus, Institute of Electrical and Electronics Engineers (IEEE), and Cochrane. The searches included online books, published and in-press articles, as well as conference papers to ensure the inclusion of as many articles as possible. Then, the titles, abstracts, keywords, and the bibliographies of all the selected studies were subsequently searched for potentially relevant articles.

Then, the full texts of the reviews or perspectives were checked with focus on mentioned challenges and implications of EHR and omics data integration in the articles. The inclusion criteria were (a) articles in English, (b) reviews, (c) systematic reviews (d), reviews with data/narrative synthesis, (e) meta-analyses, (f) perspectives, (g) reports or formulations in reputed journals (h), brief or comprehensive communications, as well as (h) letters and commentaries publishes until 30 July 2020. The studies were excluded if not accessible through university networks on databases. A combination of
the following terms was used to retrieve related papers.

(Personalized OR Precision) AND Medicine AND (EHR OR Electronic Medical Record OR EMR)

It should be noted that all the selected articles were included because all of the mentioned implications and challenges were stakeholders’ requirements addressed in the present model. According to software development methodologies, one of the major sources of stakeholders’ requirements was their comments, interviews, reports, and perspectives. Therefore, all the articles were included for requirement analysis. Distribution of included studies over time as shown in Fig 1.

Fig 1: Distribution of included studies over time

Additionally, some text processing methods such as word cloud generator, remove stop words and tokenization were employed to visualize a more beneficial outcome of the review phases. Finally, the extracted evidence was discussed to support the requirements.

RESULTS

Characteristics of Articles and Related Challenges

The distribution of the selected articles over time was illustrated in Fig 2. A text-preprocessing phase was fulfilled to extract the most important keywords in terms of challenges and requirements. Stop words were also removed and general tokens such as “may, must, clinical, genomic, data, EHR, etc.” were added to stop word dictionary to ignore items with low information value. As shown in Fig 2, a word cloud with a maximum of 300 words was generated to achieve some frequent key challenges in the given articles. The open-source Word Cloud for Python documentation library was also used [14].

To summarize the given challenges and requirements, frequent key challenges and requirements were re-analyzed and converted into some clean stakeholders’ scenarios in two different categories for requirement engineering, functional (use cases) and non-functional (quality attributes) requirements. The unified modeling language (UML) (ver. 2) use case diagram was also employed to display the requirements [15]. Fig 3 shows the associated items.

Fig 2: A word cloud generated from table of challenges and requirements after text preprocessing

Fig 3: Functional (gray) and non-functional (white) requirements

As a general result of extracted challenges from the review, the present solution should take into account of interoperability, uniform reporting, standardization, structured data entry, integration with other information systems, confidentiality and privacy, multiple data source, machine-readable store and retrieval, research capabilities, decision support system (DSS), knowledge management, and big and high dimensional data management and computation as the main requirements.

Solutions, technical notes, and recommendations

All solutions, tasks, artifacts, and recommendations of the reviewed articles were extracted to propose a
uniform conceptual model for multi-omics integration into EHR. The requirements were also discussed by an explanation of current evidence and achievements.

In this respect, Shoenbill et al. [16] mentioned that standardized genetic terminologies and methods for data transfer as well as a standard structure and language were required to exchange information between clinical systems and transform knowledge in clinical DSSs. Examples of available standards included health level 7 (HL7) for messaging, genomic variation format (GVF) for genetic data annotation, logical observation identifiers names and codes (LOINC), human genome variation society (HGVS), human gene nomenclature committee’s (HGNC) terminology, and systematized nomenclature of medicine: clinical terms (SNOMED CT) [17].

Currently, the primary standards applied for genomics test management and exchange include LOINC, HGVS, HGNC, the database of single nucleotide polymorphisms (dbSNP), reference sequences NCBI (RefSeq), and the international system for human cytogenetic nomenclature (ISCN) that contain information about genetic test findings and risk factors [16]. Unfortunately, in 2017, Sitapati et al. [18] stated that these terminologies had not been adopted by a high portion of laboratories. Additionally, traditional terminologies of health information technology had not supported genetic concepts (such as diseases) well. Online mendelian inheritance in man (OMIM) had also listed many of these diseases but had not been integrated with EHR and a gap had remained between it and SNOMED CT. Moreover, some countries had applied the international classification of diseases (ICD) codes for registration of patient clinical conditions which was insufficient for genomics [19]. Today, a synergy exists between extensions based on extensible markup language (XML) especially HL7 and service-oriented architecture (SOA). HL7 defines message formats that can store various laboratory result forms and other health information, SOA or web services also offer a strong solution in which laboratory and electronic health systems can be connected [20].

In this respect, Deckard et al. [21] explained that LOINC could cover the scope of laboratory testing (e.g. microbiology), and a diverse spectrum of clinical measurements (e.g., vital signs or radiology reports) by more than 73,000 terms. It could similarly represent laboratory concepts such as basic attributes, answer lists, observation panels, and other details like help text, language variants, and units of measure in a powerful data model. They further focused on more than 1400 terms currently applied in reporting of genetic tests. LOINC could also employ HGNC’s terminology to name the gene(s) and HGVS’s terms to label the variation(s) in genetic tests [21]. Consequently, the researchers had to follow a combined approach based on discussed technologies and extensions to develop architecture and database of their models.

Hazin et al. [22] shed light on improving genomic literacy between patients, nurses, physicians, and laboratory technicians. They recommended patient educational programs and online educational services to help patients gain better perceptions. Furthermore, a combination of EHR systems with specialized training on interpreting genomic information was suggested in the core of activities. Additionally, a test template was recommended that could provide a uniform structure to standardized genetic testing, enable primary care, and reduce reporting misunderstandings. In addition, Kullo et al. [23] emphasized consistency and reliability of genotype reporting via applying EHR mechanisms. At this point, it was emphasized that DSS reporting needed a similar approach. HL7 also published an implementation guideline for reporting structured clinical genetic tests [23] whose benefits were as follow:

1) creating a common infrastructure to communicate patient-specific genetic data between stakeholders,

2) enabling both human and computer readability of genetic data to develop clinical DSSs and real-time computations, and

3) supporting healthcare practices and clinical trials together [24].

Moreover, supporting cohort studies were proposed by Kohane et al. [25]. The problem of phenotypic misclassification recorded in genomic research due to rarely codified data in narrative texts could also lead to insufficient computable data to drive phenotype-genotype studies. Hence, guidelines, templates, and uniform structures were applied in developing the architecture of the present model.

Much of the valuable data found in EHR are represented in free texts and are even unstructured such as radiology, laboratory, and genetic test results [23]. Moreover, genetic variants extracted from the sequence would be integrated into EHR as structured data and must be stored in a structured and machine-readable format to trigger clinical DSSs [26]. By extending the structured data entry, computerized physician order entry (CPOE) systems and clinical alerts can be thus designed through being integrated with clinical DSSs and pharmacogenomics [19].

Semantic machine processability (SMP) is also a concept that normalizes specific information in a common structure to realize machine readability. Within SMP, relevant concepts and relationships are explicitly revealed by computational methods such as neuro-linguistic programming (NLP) and semantic interoperability since unstructured and structured data in clinical data sources can have different representations at semantic levels consisting of codes mapped from various local, national, and international coding systems as well as a variety of
reference ranges, units, and timeline templates. It is also considered a daunting challenge to the efforts towards doing semantic explicit analysis. Therefore, without formal constraints and operational mechanisms, it would be complex to capture deep richness in a data source and enable an effective data usage [27]. Moreover, existing clinical informatics architectures are incapable to store, search, annotate, and share genomic sequence data across healthcare systems requiring IT developers to collaborate with clinical experts and scientists to redesign EHR based on recommended solutions [16]. HL7 could internationally release an implementation guide for designing structured genetic test reports [26]. Structured data entry is also directly correlated with standardized terminologies. Some of the popular examples of standards, which could be used as an architectural facilitator, were genomic variation format (GVF), HL7, HGNC, genetic data annotation, HGVS, and genetic data representation for clinical use (LOINC, SNOMED CT, RXNorm, and etc.). As mentioned by Shoenbill et al. [16], the main standards utilized for genetic laboratory tests include LOINC, HGVS, HGNC, RefSeq, dbsNP, HGNC, and ISCN. Unfortunately, coding systems are not the only challenge in existing EHR. Data may be thus hidden in the unstructured or free text up to 80% of the value in EHR [28]. It requires NLP to derive textual entities from these notes. However, in current functional requirements, it is supposed that a new electronic health/ medical/laboratory system will be designed with a focus on coding standards and structured data entry. Therefore, structured data entry as well as mapping these fields has been recommended to coded concepts based on popular coding systems.

To improve genomic literacy among patients and providers as well as pre/posttest counseling by trained medical geneticists, online portals have been proposed. Kullo et al. [23] highly recommended a web-based, user-friendly, and secure portal to empower patients. Our results showed that integrated online patient portals are an important feature in the implementation of personalized medicine. This portal can disseminate useful knowledge between patients by c

...
include complete data capture, data quality and validation, system knowledge, and heterogeneity among systems [34]. Specially, for genomics research several challenges are containing missing data, little information (e.g. poor documentation of family history, environmental and lifestyle factors), finding reliable genomics information, imprecision in trait ascertainment, lack of interoperability, bias, and confidentiality [35]. Most of these problems can be handled by computational techniques for example there are several strategies to overcome missing data problems [36, 37]. Designing checklists with more detailed information can be helpful for elimination little information problem. Confidentiality issues needed to be addressed to provide continued research and development by collaborating with third parties [38]. Therefore, there was a requirement for some preprocessing algorithms to be applied on confidentiality requirements as to quality attributes. As a result, in the present framework, a proper process was proposed about the selected architecture and functional requirements.

Considered as an important requirement, support of pharmacogenomics before drug prescription or detecting high-risk individuals based on genomics data was mentioned as a core component of novel and mature genomics HISs [39]. Therefore, EHR could plays a critical informational role for adoption of pharmacogenomics as a part of routine medical care [40]. To order purposeful pharmacogenomics tests, the test outcomes were required to be easy to human understanding; as well, they needed to include clinically validated pharmacogenomics guidelines. Pharmacogenomics testing must be also integrated into EHRs and pharmacy-ordering systems especially by supporting international terminologies and coding systems such as world health organization international classification of diseases (WHO ICD). Moreover, these results should be machine-readable in order to design and develop pharmacogenomics clinical DSSs as well as alerting and recommender systems. Thus, pharmacogenomics guidelines are need to be converted into rules and algorithms; and then embedded into clinical DSS knowledge base to create clinician notifications and reports. Preferably, it has been fed through health economic models and a wide range of data with costs associated with adverse events related to selected medications in an evolutionary pharmacogenomics ecosystem [41]. CPIC has also defined and shared the best guides and practices for pharmacogenomics knowledge management and clinical DSSs [30]. However, the meaningful use of pharmacogenomics inpatient administration depends on standard mark-ups for gene expression and genetic variations [30]. This requirement can be fulfilled by structured data entry as well as formatted and coded data elements, which were discussed earlier. Ullman-Cullere et al. [19] also provided some evidence that structured pharmacogenomics data and clinical DSS could be embedded within CPOE.

Genotype-phenotype structures and algorithms were the next requirements of mature genomics-integrated EHRs. In this respect, electronic phenotyping in being performed robustly based on EHR. This procedure has been also followed by genome-wide association studies (GWAS). In this respect, Gottesman et al. [39] developed and published a code library of electronic phenotyping algorithms on health records in which phenotype knowledge base (PheKB) and formal phenotyping language were other achievements of their study. Accordingly, they suggested creating a repository of the most important pharmacogenetic variants to activate and support future genotype-phenotype research such as identification of novel genotype-phenotype associations. Furthermore, Levy et al. [41] argued that raw genotype information did not include phenotypic interpretation. Furthermore, they categorized the results to develop a translation layer via assigning an identical phenotype category. The genotype-phenotype algorithms were then derived from multiple data sources and they needed to be developed and validated. Information on sociocultural determinants of health status, patient-reported acquired data, and bio-bank derived data including omics data were among the major data sources of these algorithms [42].

An un-interoperable information system or repository was not helpful in meaningful data exchange with other systems such as government agencies, healthcare providers, and various types of stakeholders like pharmaceutical corporations. In a wrong relationship, incorrect understanding could also mix up with interpretations. Information also requires a high degree of interoperability for sending and receiving genomics [19, 27]. Healthcare institutions with locally constructed information systems are thus suitable for implementation of pharmacogenomics by providing an appropriate architecture for facilitating interoperability. In this regard, HL7 proposed a model to exchange genetic tests with the contribution of EHR developers. Another implementation of interoperability in genomics-oriented information systems was developed by Illumina VeraCode® Absorption, Distribution, Metabolism, and Excretion (ADME) core panel for PREDICT project. They suggested that genetic variant data could be converted into a portable document format (PDF) or a plain text. In addition, a query in the form of an automated script to filter database views was embedded periodically into a central web service available to all sub-systems and components [41]. As an essential software architecture, SOA can thus play a critical role in the success of healthcare systems due to their particular nature [43]. Therefore, SOA and web services offer a great potential to overcome interoperability
challenges. Using a translation layer to assign a coded phenotype category and to generate some valuable text strings to genetic test results also represented another solution proposed in predict [41]. At a higher level, knowledge sharing requires appropriate standards for interoperable data representations. Messaging frameworks adapted with HIT terminologies are also a critical task. Unfortunately, traditional terminologies do not support genetic diseases as expected. For example, a mapping between SNOMED CT and OMIM terminologies remains as a semantic gap [19]. These standards enable information in each of the separated systems to be related to each other and also generate new knowledge [44]. In 2017, Kuehn [45] argued that putting a whole genome in EHRs was not effectively operational. He also adopted a standard language for genomic information that had been developed on FHIR. HL7 FHIR standard was released to achieve healthcare system interoperability. FHIR was then developed based on HL7 v2 and v3. As a resource-oriented standard; patients, devices, and documents can be defined in the present model as resources. Simplicity has been also embedded into the core of FHIR. Therefore, it has attracted a lot of attention from health informatics communities [46]. Moreover, HL7 is an SOA compliant standard designed based on XML [47]. So, considering the development and advancement of SOA as well as increased popularity and application of FHIR, it seems that there are clear horizons for full interoperability of EHR and genetics data. It is worth noting that maintaining privacy and confidentiality in the interoperable genomics-integrated EHR is an essential work task [22].

Finally; big data computation, discussed in this review, were considered as the last non-functional requirement. Thus, EHRs need to simply achieve big data along with their generation. Although non-operational data do not require to be immediately available to practitioners, they must be stored for future data mining, knowledge discovery, and visualization [40, 48]. Big data have been posed by large genomic variations and clinical phenotypes in the pharmacogenomics scope. These data aggregated from multiple data sources after integration into a data warehouse using interoperability standards. Genomics research and innovation network (GRIN) have also been formed to create a broad database of annotated genomics and clinical data. The network adds transcriptome, proteome, and metabolome and intends to connect EHR and other systems. By the advancement of graphics processing units and deep neural networks, it is possible to mine a huge volume of big data in a shorter time. Also, cloud computing has made data manipulation easier. To handle the big data efficiently, Apache Spark and Hadoop have been introduced. These potential tools have led to some projects like the national institutes of health (NIH) genomics [49-51]. The results of all requirements in this regard are summarized in the way that genomics big data computation is the intersection point of other mentioned requirements to achieve an applied and appropriate integration. The ethical, legal, and security requirements had been also discussed by Hazin et al. in a detailed study [22]; therefore, they were not examined in this study.

DISCUSSION

In countries without a national EHR infrastructure, handling and combining data across multiple EHRs, information systems, genomics databases, and larger population studies are challenging. These data sources complement patient basic information. Likewise, some other type of data from informal sources such as patient groups in social networks, can play a positive role in genomics medicine [25]. Clinical and genomics data warehouses also aggregate data and facilitate retrospective analysis, data mining, knowledge discovery, and visualization. Standardized terminologies and comprehensive controlled vocabularies similarly play the main role in integrating heterogeneous data in a single clinical data warehouse. To reduce unavailable or incomputable health and genomics data within this hybrid data warehouse, data manipulation methods and technologies such as NLP and structured query language (SQL) are needed [52, 53]. In this respect, Angulo et al. described a lightweight message-oriented data integration engine that allowed concurrent connection to clinical information from various heterogeneous data sources. They applied XML-based technologies to generate messages and templates. The platform was an operational model which could be also used in genomics information systems [55]. Therefore, a data warehouse was proposed in the present study to handle multiple data sources for knowledge discovery and visualization based on findings and recommendations by Angulo et al. [54].

CONCLUSION

EHRs demonstrate potentials to enable pharmacogenomics and PM. The main challenges in this way include semantic and technical themes. To develop clinical DSSs and to extract modern knowledge from genomics data, a standardized format and machine-readable format was required. To integrate genotype-phenotype of a patient as well as the nature of multiple data sources in genomics, interoperability of related resources has been thus emphasized. To distribute extracted and visualized knowledge, online portals can facilitate the sharing process. As a whole puzzle, each requirement is dependent on others. Future studies can be directed towards introduced tools, methods, algorithms, as well as hardware and software in each requirement. A software methodology was also proposed to
develop various solutions and scenarios about the topics. This methodology was suggested as scientific and evidence-based requirement engineering by focusing on articles rather than people. Combination of current artifacts with other methods such as interviews can also lead to more accurate results.

**AUTHOR’S CONTRIBUTION**

All authors contributed to the literature review, design, data collection and analysis, drafting the manuscript, read and approved the final manuscript.

**CONFLICTS OF INTEREST**

The authors declare no conflicts of interest regarding the publication of this study.

**FINANCIAL DISCLOSURE**

No financial interests related to the material of this manuscript have been declared.

**REFERENCES**


