Frontiers in Health Informatics
2021; 10: 86

www.ijmi.ir
Open Access

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

Ali Najafi[1], Neda Emami[2], Taha Samad-Soltani[2]*

[1]Molecular Biology Research Center, Systems Biology and Poisonings Institute, Baqiyatallah University of Medical Sciences, Tehran, Iran
[2]Department 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

A B S T R A C T

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:
Najafi A, Emami N, Samad-Soltani T. Integration of genomics data and electronic health records toward personalized medicine: A targeted review. Front Health Inform. 2021; 10: 86. DOI: 10.30699/fhi.v10i1.299

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

Copyright© 2021, Published by Frontiers in Health Informatics. This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International (CC BY) License (http://creativecommons.org/).

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.

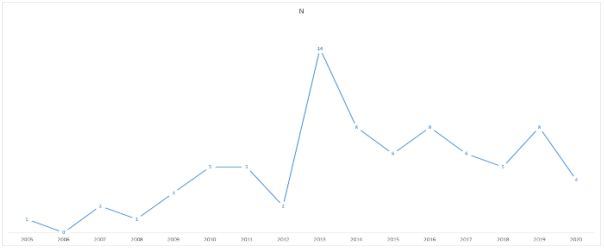

Fig 1: Distribution of included studies over time

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.

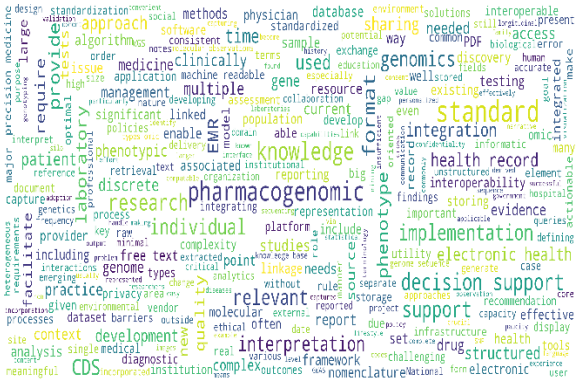

Fig 2: A word cloud generated from table of challenges and requirements after text preprocessing

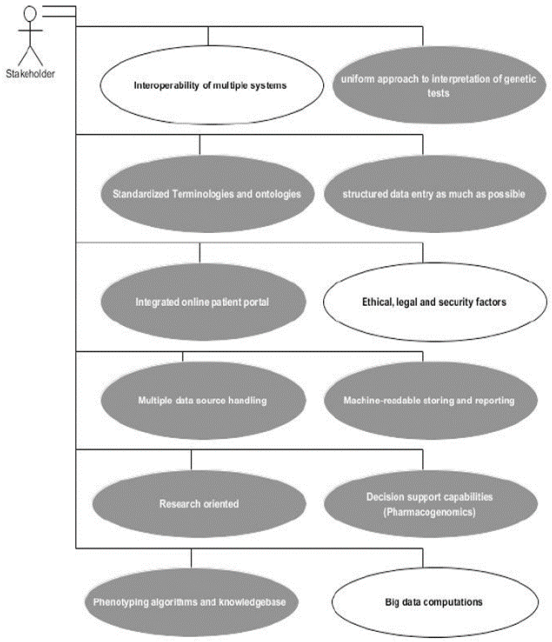

Fig 3: Functional (gray) and non-functional (white) requirements

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 continuous news and information updated in a simple and easy-to-use manner. Although many practitioners lack the requisite genomic knowledge to provide adequate counseling, genomic portals have an up-to-date knowledge body and can provide patient empowerment educational programs, online educational videos to engage the public with better perceptions, mobile-based accessibility to enhance communications, and online courses to improve literacy and genomic training between nurses [22]. Some quality attributes (non-functional requirements) should be correspondingly taken into account in the design of genomics patient portals such as patient privacy, timely accessibility, portal security, data confidentiality, consistency, usability, and reliability [22, 23]. Applying tools such as pictograms [29] and decision rules to share genomic information to patients and the public can make it possible to integrate clinical DSSs in online portals by generating a notification to patients of their genetic test results with human-friendly interpretations [23, 30]. In addition, laboratory technicians can rapidly distribute genetic results to patients in their profiles and implement surveillance strategies at the heart of the portal [30]. As a result, the present solution should be satisfied by quality attributes as well as integrating with clinical DSSs in warnings, alarms, and notifications in the preferred form of a mobile application.

One of the core requirements of genomics-EHR integration characterized the design of some pathways to clinical and research use of data [16]. In this regard, Scheuermann et al. [26] stated that secure and ongoing collection of gene-based and other molecular tests from EHRs and other health information systems (HISs) to the data warehouse and linkage with population-based registries could support researchers in making relationships among personal traits, interventions, and outcomes. It should be noted that the translation of molecular data into clinically actionable insights and research feeds is not possible by unstructured data formats as well as low-quality and non-normalized clinical ones [31]. Accordingly, computational methods can apply filters on documents and free texts and provide different levels of insights into disease-oriented studies and further research [32].

Ulman et al. [19] mentioned that data warehouses and translational application frameworks would benefit PM by providing toolboxes for researchers to select study cohorts and to combine phenotype-genotype data into their research processes and systems. Alterovitz et al. [33] also suggested an abstraction layer on top of file formats in the form of application protocol interfaces (API) reflecting the formats and data manipulation patterns. These APIs were able to contextualize genomics data with other systems for discovery and secondary research purposes. Besides, they recommended a combination of simple modular architecture research tools (SMART) on fast healthcare interoperability resources (FHIR) to enable standardized support of research.

Another framework for research-oriented genomics clinical DSSs was proposed by translating knowledge into clinical DSSs rules and involving a layer of standards such as HL7 and LOINC data elements as well as ontology and rule creators such as CPIC. This approach could standardize knowledge by providing a standard language between researchers and software developers [33].

There are some limitations of the EHR for research

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 socio-cultural 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 genomic 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

1. Ginsburg, GS, Willard HF. Genomic and personalized medicine: foundations and applications. Transl Res. 2009; 154(6): 277-87. PMID: 19931193 DOI: 10.1016/j.trsl.2009.09.005 [PubMed]
2. Jain KK. Non-genomic aspects of personalized Medicine. In: Jain KK. Textbook of personalized medicine. Springer; 2021.
3. Downing GJ. Key aspects of health system change on the path to personalized medicine. Transl Res. 2009; 154(6): 272-6. PMID: 19931192 DOI: 10.1016/j.trsl.2009.09.003 [PubMed]
4. Wei W-Q, Denny JC. Extracting research-quality phenotypes from electronic health records to support precision medicine. Genome Med. 2015; 7(1): 41. PMID: 25937834 DOI: 10.1186/s13073-015-0166-y [PubMed]
5. Joyner MJ, Paneth N. Seven questions for personalized medicine. JAMA. 2015; 314(10): 999-1000. PMID: 26098474 DOI: 10.1001/jama.2015.7725 [PubMed]
6. Bourgey M, Dali R, Eveleigh R, Chen KC, Letourneau L, Fillon J, et al. GenPipes: An open-source framework for distributed and scalable genomic analyses. GigaScience. 2019; 8(6): giz037. PMID: 31185495 DOI: 10.1093/gigascience/giz037 [PubMed]
7. Monda KL, Chen GK, Taylor KC, Palmer C, Edwards TL, Lange LA, et al. A meta-analysis identifies new loci associated with body mass index in individuals of African ancestry. Nat Genet. 2013; 45(6): 690-6. PMID: 23583978 DOI: 10.1038/ng.2608 [PubMed]
8. Howey R, Shin S-Y, Relton C, Smith GD, Cordell HJ. Bayesian network analysis complements Mendelian randomization approaches for exploratory analysis of causal relationships in complex data. PLoS Genet. 2020; 16(3): e1008198. PMID: 32119656 DOI: 10.1371/journal.pgen.1008198 [PubMed]
9. Postmus I, Trompet S, Deshmukh HA, Barnes MR, Li X, Warren HR, et al. Pharmacogenetic meta-analysis of genome-wide association studies of LDL cholesterol response to statins. Nat Commun. 2014; 5: 5068. PMID: 25350695 DOI: 10.1038/ncomms6068 [PubMed]
10. Henderson GE, Cadigan RJ, Edwards TP, Conlon I, Nelson AG, Evans JP, et al. Characterizing biobank organizations in the U.S.: Results from a national survey. Genome Med. 2013; 5(1): 3. PMID: 23351549 DOI: 10.1186/gm407 [PubMed]
11. Delaney JT, Ramirez AH, Bowton E, Pulley JM, Basford MA, Schildcrout JS, et al. Predicting clopidogrel response using DNA samples linked to an electronic health record. Clin Pharmacol Ther. 2012; 91(2): 257-63. PMID: 22190063 DOI: 10.1038/clpt.2011.221 [PubMed]
12. Monnin P, Legrand J, Husson G, Ringot P, Tchechmedjiev A, Jonquet C, et al. PGxO and PGxLOD: A reconciliation of pharmacogenomic knowledge of various provenances, enabling further comparison. BMC Bioinformatics. 2019; 20(Suppl 4): 139. PMID: 30999867 DOI: 10.1186/s12859-019-2693-9 [PubMed]
13. Salem J-E, Shoemaker MB, Bastarache L, Shaffer CM, Glazer AM, Kroncke B, et al. Association of thyroid function genetic predictors with atrial fibrillation: A phenome-wide association study and inverse-variance weighted average meta-analysis. JAMA Cardiol. 2019; 4(2): 136-43. PMID: 30673079 DOI: 10.1001/jamacardio.2018.4615 [PubMed]
14. Mueller A. Wordcloud for python documentation [Internet]. 2018 [cited: 15 Mar 2021]. Available from: http://amueller.github.io/word_cloud/index.html
15. Arlow J, Neustadt I. UML 2 and the unified process: Practical object-oriented analysis and design. Pearson Education; 2005.
16. Shoenbill K, Fost N, Tachinardi U, Mendonca EA. Genetic data and electronic health records: a discussion of ethical, logistical and technological considerations. J Am Med Inform Assoc. 2014; 21(1): 171-80. PMID: 23771953 DOI: 10.1136/amiajnl-2013-001694 [PubMed]
17. Moore B, Rynearson S, Cunningham F, Ritchie G, Eilbeck K. Using GVF for clinical annotation of personal genomes. AIMM; 2012.
18. Sitapati A, Kim H, Berkovich B, Marmor R, Singh S, El-Kareh R, et al. Integrated precision medicine: The role of electronic health records in delivering personalized treatment. Wiley Interdiscip Rev Syst Biol Med. 2017; 9(3): 1378. PMID: 28207198 DOI: 10.1002/wsbm.1378 [PubMed]
19. Ullman-Cullere MH, Mathew JP. Emerging landscape of genomics in the electronic health record for personalized medicine. Hum Mutat. 2011; 32(5): 512-6. PMID: 21309042 DOI: 10.1002/humu.21456 [PubMed]

20. Marsolo K, Spooner SA. Clinical genomics in the world of the electronic health record. Genet Med. 2013; 15(10): 786-91. PMID: 23846403 DOI: 10.1038/gim.2013.88 [PubMed]

21. Deckard J, McDonald CJ, Vreeman DJ. Supporting interoperability of genetic data with LOINC. J Am Med Inform Assoc. 2015; 22(3): 621-7. PMID: 25656513 DOI: 10.1093/jamia/ocu012 [PubMed]

22. Hazin R, Brothers KB, Malin BA, Koenig BA, Sanderson SC, Rothstein MA, et al. Ethical, legal, and social implications of incorporating genomic information into electronic health records. Genet Med. 2013; 15(10): 810-6. PMID: 24030434 DOI: 10.1038/gim.2013.117 [PubMed]

23. Kullo IJ, Jarvik GP, Manolio TA, Williams MS, Roden DM. Leveraging the electronic health record to implement genomic medicine. Genet Med. 2013; 15(4): 270-1. PMID: 23018749 DOI: 10.1038/gim.2012.131 [PubMed]

24. Crump JK, Fiol GD, Williams MS, Freimuth RR. Prototype of a standards-based EHR and genetic test reporting tool coupled with HL7-compliant infobuttons. AMIA Jt Summits Transl Sci Proc. 2018; 2017: 330-9. PMID: 29888091[PubMed]

25. Kohane IS. Using electronic health records to drive discovery in disease genomics. Nat Rev Genet. 2011; 12(6): 417-28. PMID: 21587298 DOI: 10.1038/nrg2999 [PubMed]

26. Scheuermann RH, Milgrom H. Personalized care, comparative effectiveness research and the electronic health record. Curr Opin Allergy Clin Immunol. 2010; 10(3): 168-70. PMID: 20431366 DOI: 10.1097/ACI.0b013e328338c232 [PubMed]

27. Shabo Shvo A. Meaningful use of pharmacogenomics in health records: semantics should be made explicit. Pharmacogenomics. 2010; 11(1): 81-7. PMID: 20017674 DOI: 10.2217/pgs.09.161 [PubMed]

28. Kho AN, Rasmussen LV, Connolly JJ, Peissig PL, Starren J, Hakonarson H, et al. Practical challenges in integrating genomic data into the electronic health record. Genet Med. 2013; 15(10): 772-8. PMID: 24071798 DOI: 10.1038/gim.2013.131 [PubMed]

29. Barros IM, Alcântara TS, Mesquita AR, Santos ACO, Paixão FP, Lyra JrDP. The use of pictograms in the health care: A literature review. Res Social Adm Pharm. 2014; 10(5): 704-19. PMID: 24332470 DOI: 10.1016/j.sapharm.2013.11.002 [PubMed]

30. Peterson JF, Bowton E, Field JR, Beller M, Mitchell J, Schildcrout J, et al. Electronic health record design and implementation for pharmacogenomics: A local perspective. Genet Med. 2013; 15(10): 833-41. PMID: 24009000 DOI: 10.1038/gim.2013.109 [PubMed]

31. Sheldon J, Ou W. The real informatics challenges of personalized medicine: Not just about the number of central processing units. Per Med. 2013; 10(7): 639-45. PMID: 29768753 DOI: 10.2217/pme.13.16 [PubMed]

32. Sethi P, Theodos K. Translational bioinformatics and healthcare informatics: computational and ethical challenges. Perspect Health Inf Manag. 2009; 6(Fall): 1h. PMID: 20169020 PMCID: PMC2804463 [PubMed]

33. Alterovitz G, Warner J, Zhang P, Chen Y, Ullman-Cullere M, Kreda D, et al. SMART on FHIR genomics: Facilitating standardized clinico-genomic apps. J Am Med Inform Assoc. 2015; 22(6): 1173-8. PMID: 26198304 DOI: 10.1093/jamia/ocv045 [PubMed]

34. Kim E, Rubinstein SM, Nead KT, Wojcieszynski AP, Gabriel PE, Warner JL. The evolving use of electronic health records (EHR) for research. Semin Radiat Oncol. 2019; 29(4): 354-61. PMID: 31472738 DOI: 10.1016/j.semradonc.2019.05.010 [PubMed]

35. Safarova MS, Kullo IJ. Using the electronic health record for genomics research. Curr Opin Lipidol. 2020; 31(2): 85-93. PMID: 32073412 DOI: 10.1097/MOL.0000000000000662 [PubMed]

36. Wells BJ, Chagin KM, Nowacki AS, Kattan MW. Strategies for handling missing data in electronic health record derived data. EGEMS (Wash DC). 2013; 1(3): 1035. PMID: 25848578 DOI: 10.13063/2327-9214.1035 [PubMed]

37. Marwala T. Computational intelligence for missing data imputation, estimation and management: Knowledge optimization techniques. Information Science Reference; 2009.

38. Hemingway H, Asselbergs FW, Danesh J, Dobson R, Maniadakis N, Maggioni A, et al. Big data from electronic health records for early and late translational cardiovascular research: Challenges and potential. Eur Heart J. 2018; 39(16): 1481-95. PMID: 29370377 DOI: 10.1093/eurheartj/ehx487 [PubMed]

39. Gottesman O, Kuivaniemi H, Tromp G, Faucett WA, Li R, Manolio TA, et al. The electronic medical records and genomics (eMERGE) network: Past, present, and future. Genet Med. 2013; 15(10): 761-71. PMID: 23743551 DOI: 10.1038/gim.2013.72 [PubMed]

40. Denny JC. Surveying recent themes in translational bioinformatics: Big data in EHRs, omics for drugs, and personal genomics. Yearb Med Inform. 2014; 9(1): 199-205. PMID: 25123743 DOI: 10.15265/IY-2014-0015 [PubMed]

41. Levy KD, Decker BS, Carpenter JS, Flockhart DA, Dexter PR, Desta Z, et al. Prerequisites to implementing a pharmacogenomics program in a large health-care system. Clin Pharmacol Ther. 2014; 96(3): 307-9. PMID: 24807457 DOI: 10.1038/clpt.2014.101 [PubMed]

42. Roden DM, Denny JC. Integrating electronic health record genotype and phenotype datasets to transform patient care. Clin Pharmacol Ther. 2016; 99(3): 298-305. PMID: 26667791 DOI: 10.1002/cpt.321 [PubMed]

43. Koumaditis K, Themistocleous M, Rupino Da Cunha P. SOA implementation critical success factors in healthcare. Journal of Enterprise Information Management. 2013; 26(4): 343-62.

44. Downing GJ, Boyle SN, Brinner KM, Osheroff JA. Information management to enable personalized medicine: Stakeholder roles in building clinical decision support. BMC Med Inform Decis Mak. 2009; 9(1): 44. PMID: 19814826 DOI: 10.1186/1472-6947-

9-44 [PubMed]

45. Kuehn BM. Pilot programs seek to integrate genomic data into practice. JAMA. 2017; 318(5): 410-2. PMID: 28700793 DOI: 10.1001/jama.2017.7181 [PubMed]

46. Bender D, Sartipi K. HL7 FHIR: An agile and restful approach to healthcare information exchange. International Symposium on Computer-Based Medical Systems. IEEE; 2013.

47. Janjua NK, Hussain M, Afzal M, Ahmad HF. Digital health care ecosystem: SOA compliant HL7 based health care information interchange. International Conference on Digital Ecosystems and Technologies. IEEE; 2009.

48. Alonso SG, de la Torre Díez I, Rodrigues JJPC, Hamrioui S, López-Coronado M. A systematic review of techniques and sources of big data in the healthcare sector. J Med Syst. 2017; 41(11): 183. PMID: 29032458 DOI: 10.1007/s10916-017-0832-2 [PubMed]

49. Wiewiórka MS, Messina A, Pacholewska A, Maffioletti S, Gawrysiak P, Okoniewski MJ. SparkSeq: Fast, scalable and cloud-ready tool for the interactive genomic data analysis with nucleotide precision. Bioinformatics. 2014; 30(18): 2652-3. PMID: 24845651 DOI: 10.1093/bioinformatics/btu343 [PubMed]

50. Chung W-C, Chen C-C, Ho J-M, Lin C-Y, Hsu W-L, Wang Y-C, et al. CloudDOE: A user-friendly tool for deploying Hadoop clouds and analyzing high-throughput sequencing data with MapReduce. PLoS One. 2014; 9(6): e98146. PMID: 24897343 DOI: 10.1371/journal.pone.0098146 [PubMed]

51. Reynolds SM, Miller M, Lee P, Leinonen K, Paquette SM, Rodebaugh Z, et al. The ISB cancer genomics cloud: A flexible cloud-based platform for cancer genomics research. Cancer Res. 2017; 77(21): e7-10. PMID: 29092928 DOI: 10.1158/0008-5472.CAN-17-0617 [PubMed]

52. Botsis T, Hartvigsen G, Chen F, Weng C. Secondary use of EHR: Data quality issues and informatics opportunities. Summit Transl Bioinform. 2010; 2010: 1-5. PMID: 21347133 PMCID: PMC3041534 [PubMed]

53. Lan K, Wang D-T, Fong S, Liu L-S, Wong KKL, Dey N. A survey of data mining and deep learning in bioinformatics. J Med Syst. 2018; 42(8): 139. PMID: 29956014 DOI: 10.1007/s10916-018-1003-9 [PubMed]

54. Angulo C, Crespo P, Maldonado JA, Moner D, Pérez D, Abad I, et al. Non-invasive lightweight integration engine for building EHR from autonomous distributed systems. Int J Med Inform. 2007; 76 (Suppl 3): S417-24. PMID: 17600763 DOI: 10.1016/j.ijmedinf.2007.05.002 [PubMed]