G&I Genomics & Informatics

pISSN 1598-866X eISSN 2234-0742
Genomics Informm 2013;11(4):186-190
http://dx.doi.org/10.5808/GI.2013.11.4.186

REVIEW ARTICLE

Perspectives on Clinical Informatics: Integrating Large-Scale Clinical, Genomic, and Health Information for Clinical Care

In Young Choi[1], Tae-Min Kim[2], Myung Shin Kim[3], Seong K. Mun[1,4], Yeun-Jun Chung[5]*

[1]Department of Medical Informatics, The Catholic University of Korea College of Medicine, Seoul 137-701, Korea,
[2]Department of Medical Informatics, MRC, IRCGP, The Catholic University of Korea College of Medicine, Seoul 137-701, Korea,
[3]Departments of Clinical Laboratory, The Catholic University of Korea College of Medicine, Seoul 137-701, Korea,
[4]Open Source Electronic Health Record Agent (OSEHRA), Arlington Innovation Center, Virginia Tech, Arlington, VA 22203, USA
[5]Department of Microbiology, Department of Medical Informatics, Integrated Research Center for Genome Polymorphism (IRCGP), MRC, The Catholic University of Korea College of Medicine, Seoul 137-701, Korea

The advances in electronic medical records (EMRs) and bioinformatics (BI) represent two significant trends in healthcare. The widespread adoption of EMR systems and the completion of the Human Genome Project developed the technologies for data acquisition, analysis, and visualization in two different domains. The massive amount of data from both clinical and biology domains is expected to provide personalized, preventive, and predictive healthcare services in the near future. The integrated use of EMR and BI data needs to consider four key informatics areas: data modeling, analytics, standardization, and privacy. Bioclinical data warehouses integrating heterogeneous patient-related clinical or omics data should be considered. The representative standardization effort by the Clinical Bioinformatics Ontology (CBO) aims to provide uniquely identified concepts to include molecular pathology terminologies. Since individual genome data are easily used to predict current and future health status, different safeguards to ensure confidentiality should be considered. In this paper, we focused on the informatics aspects of integrating the EMR community and BI community by identifying opportunities, challenges, and approaches to provide the best possible care service for our patients and the population.

Keywords: clinical data warehouse, database, electronic health records, medical informatics

Introduction

The development of electronic medical record (EMR) systems began as a means to document clinical activities for in-patients and out-patients [1]. They have evolved as the primary front-line patient care clinical tool for medical professionals. The completion of the Human Genome Project opened the era of research in genomics and proteomics. Genome research provides keys to understanding the mechanisms of disease. In clinical informatics, the widespread adoption of the EMR system has generated large amounts of heterogeneous clinical data—some structured and others unstructured. In addition, the explosive health-related contents from online communities, mobile applications, and electronic personal health records increased the availability of non-traditional data on individual activities and life style [2]. In genetics, since the completion of the Human Genome Project in 2003, the acquisition, analysis, and presentation of whole-genomic data has become faster, cheaper, and more reliable day by day [3]. Such dramatic technological advances affect the development of new prevention, diagnosis, and treatment patterns for routine clinical care. The massive amount of heterogeneous data from two different domains is expected to provide personalized, preventive, and predictive healthcare services in the near future [4].

Integrated use of EMR and bioinformatics is beginning to influence the changes in the research paradigm—that is, rapid introduction of new concepts into the point of care. Dr. Want used clinical bioinformatics (CBI) with the definition

Received October 18, 2013; Revised November 18, 2013; Accepted November 20, 2013

***Corresponding author:** Tel: +82-2-2258-7343, Fax: +82-2-537-0572, E-mail: yejun@catholic.ac.kr

of "the clinical application of bioinformatics-associated sciences and technologies to understand molecular mechanisms and potential therapies for human disease" [5]. CBI aims to deal with the challenge of integrating genomic and clinical data to accelerate the translation of knowledge into effective treatment plan development and personalized prescription. It is to assist clinicians in various ways, including new biomarker discovery, identification of genotype and phenotype correlations, and pharmacogenomics at the point of care.

Biomedical informatics is a more popular term. Biomedical informatics is defined as an emerging, multi-disciplinary field, and it is the integration of the computational methods and diverse technologies used in life science research, such as genomics, proteomics, systems biology, computer sciences, and healthcare applications, such as electronic health records (EHRs) [6]. The adoption of EMRs enables one to conduct comprehensive phenotypic-genotypic association studies using the genotypes obtained from whole genome sequencing of a given cohort in combination with the phenome data of the same population, as available in the EMR database, as shown in Fig. 1.

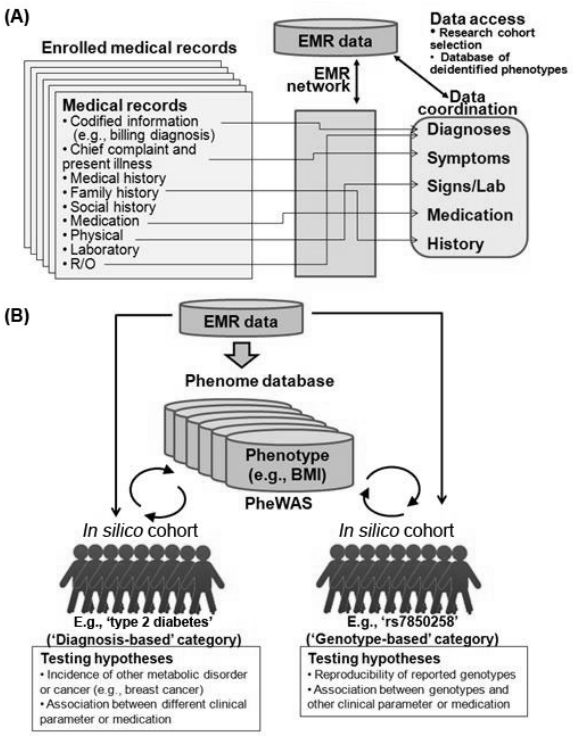

Fig. 1. Electronic medical record (EMR) databases and PheWAS. (A) Dataflow shows how EMRs of enrolled patients are curated and incorporated into a database. (B) *In silico* cohort is generated from an EMR database with respect to various categories (e.g., billing diagnoses or disease-related genotypes). The cohort is tested for the association with various phenotypes, as available in EMR databases (PheWAS). R/O, rule out; BMI, body mass index.

The major challenge in enabling such convergent research is to provide easy storage, user-friendly visualization, speedy analysis, knowledge generation, and presentation of clinically relevant information at the point of care. Relevant information should be extracted and linked with medical records in a clinically applicable manner. The convergence of discovery research for clinical implementation can only be accomplished through stringent data management, analysis, interpretation, and quantification in a multidisciplinary research environment. In addition, since genetic information can be easily identifiable, ethical issues, such as informed consent and stewardship over this database, also should be considered as the data grow. The communities of EMRs and bioinformatics (BI) have different histories. While the EMR community focused on clinical activities and clinical workflows, the BI community originated from the biological research community, which included physicists, computer scientists, statisticians, and clinical researchers. What are the optimal ways to integrate the tremendous advances in BI into routine clinical work? In this paper, we will primarily focus on the informatics aspects of this large question by identifying opportunities, challenges, and approaches for the ultimate goal of providing the best care possible for our patients and the population.

Recent Projects Using Bio-enabled EHR

A number of research projects using large-scale health record datasets have been conducted in various communities around the globe. For example, the NSF BIGDATA program solicitation (http://www.nsf.gov), which is partially funded by the National Institutes of Health (NIH), includes large-scale data collection and analysis. This program aims to advance the core scientific and technological means of managing, analyzing, visualizing, and extracting useful information from large, diverse, distributed, and heterogeneous datasets.

INBIOMEDvision (http://www.inbiomedvision.eu) is a two-year initiative funded by the European Commission 7th Framework Program of Information and Communication Technologies (ICT), with the aim of bridging the communities of bioinformatics and medical informatics. It is a coordination support action (FP7) that has the aim of promoting biomedical informatics by means of permanent monitoring of the scientific state-of-the-art and existing activities in the theme, execution of prospective analyses on the emerging challenges and opportunities, and dissemination of knowledge in the field [7].

The Electronic Medical Records and Genomics (eMERGE) Network is a national consortium organized by National Human Genome Research Institute (NHGRI) to develop,

disseminate, and apply approaches to research. It combines DNA biorepositories with EMR systems for large-scale, high-throughput genetic research with the ultimate goal of returning genomic testing results to patients in a clinical care setting. The network is currently exploring more than a dozen phenotypes (with 13 additional electronic algorithms having already been published). Various models of returning clinical results have been implemented or planned for pilot at sites across the network. Themes of bioinformatics, genomic medicine, privacy, and community engagement are of particular relevance to eMERGE [8].

In addition to nationwide collaboration projects, there is a research project using large-scale EHRs. Hanauer *et al*. [9] used large-scale, longitudinal EHR data to conduct research on associations of medical diagnoses and to explore patterns of specific disease progression. Lin *et al*. [10] proposed a symptom-disease treatment (SDT) association rule, mining a comprehensive EHR of approximately 2.1 million records from a major hospital. Based on selected International Classification of Disease (ICD-9) codes, they were able to identify clinically relevant and accurate SDT associations from patient records in seven distinct diseases, ranging from cancers to chronic and infectious diseases.

Challenges for Genome-Enabled EMR

Research combining biology and clinical information must address "the storage, retrieval, analysis, and dissemination of molecular information in a clinical setting," as suggested by the America Medical Informatics Association (AMIA)-initiated genomics working group [11]. The AMIA proposed specific areas as follows: 1) development of a database structure to unify clinical and genomic data, 2) connecting biology information with patient health records, 3) development of a genome-enabled EMR system, 4) linking clinical trial and drug discovery information, 5) supporting the development benchmark clinical/molecular datasets, 6) developing clinical decision-support tools utilizing molecular information, and 7) visualizing and modeling the molecular basis of disease.

The major challenge is how to integrate the heterogeneous data into one database system. Should there be a single database or should one consider a federated model? Furthermore, one should consider various cases by various users, which would determine the overall system architecture. It is not possible to move large amounts of genomic data. Then, how and where should the high-intensity computation be managed? The expected raw sequencing data for one person is approximately 4 terabytes. The integrated database can have a potential impact on the prevention, diagnosis, and treatment of disease. To make this desire come true, it is important to connect genomics data with clinical information. Genetic test results are already used to assess the risk of breast cancer patients, determine the potential adverse drug reactions on individual patient metabolism, and identify treatment plans for cancers. Genetic test results have suggested a diagnosis for patients with neuropathy, inflammatory bowel disease, and Proteus syndrome and have guided therapeutic care for patients with arterial calcifications, movement disorders, and Miller syndrome [12-17]. The number of applications of genomics in diagnosing diseases and guiding treatment procedures in the clinic will continue to increase.

Key Informatics Issues to Consider

The integrated use of EMR and BI data needs to consider four key informatics areas: data modeling, analytics, standardization, and privacy.

Data modeling and data warehouse

Data modeling and warehouse are two key concepts within proper systems architecture for medicine and bioinformatics. The first generation of clinical data warehouses (CDWs) is mostly stored in commercial relational database management systems and collects structured contents, and the healthcare community has developed a large clinical database, which is called the CDW [18]. The extraction, transformation, and load is essential for converting and integrating distributed healthcare data. Ad hoc reporting tools using online analytical processing technology are used to gain intuitive and simple analysis results in clinical informatics.

Bioclinical data warehouses that integrate heterogeneous patient-related clinical or omics data should be considered. Applications analyzing bioclinical data warehouses will include genetic epidemiology and evaluation of decision-support systems before production systems.

In less than a decade, the Human Genome Project has been established to generate a large amount of biological data to practice diagnostics, prognostics, and therapeutics [19]. The central products using standardized data model are GenBank [20], SWISS-PROT [21], Exon-Intron [22], and IMGT [23].

However, standardized data models for integrated bioclinical data warehouses have not been developed yet. This area will be important to allow researchers to rapidly spread information throughout the world and inspire thousands of research projects.

The experience in designing storages for digital radiological imaging, also known as picture archiving and communication system (PACS), may provide some guidance in

bioinformatics data warehouses. In PACS, one has multiple image storages that are not necessarily part of EMR; however, the diagnostic reports are part of the EMR. The PACS images are then made available to physicians as needed.

Analytics

The analytics technology that is commonly used for these systems is mainly traditional data mining techniques developed in the 1980s. The popular term of analysis technology in the business and computer science communities in the 1990s was business intelligence. Recently, technologies that require advanced and unique data storage, management, analysis, and visualization became important in applications that are very large and complex databases.

Standardization

The lack of standardized vocabularies for clinical informatics has hampered the development of automated clinical decision support systems. Finding the laboratory term representing the same meaning of serum sodium for different database systems is troublesome in multi-center data integration. The National Library of Medicine has developed the Unified Medical Language System (UMLS) to enable these different vocabularies to be interoperable by developing a vocabulary at least at a basic level [24]. The same problems are known in bioinformatics. DNA sequences have different names and are joined in some databases only with varying levels of confidence. Codification of molecular diagnostic or cytogenetic results using existing medical vocabularies will have difficulties due to a lack of sufficient terms for molecular findings. For example, the Systematic Nomenclature of Medicine (SNOMED) has minimal codes related to the description of molecular diagnostic findings. The Logical Observation Identifiers Names and Codes (LOINC) vocabulary has recently added a significant number of molecular pathology terms; however, it lacks the rich context-defining relationships provided by ontology [24]. The Clinical Bioinformatics Ontology (CBO) addresses this gap by providing uniquely identified concepts related to clinically significant molecular findings. The CBO consists of nearly 7,000 concepts, each of which is associated with a global unique identifier, and is associated with more than 15,500 relationships [25-27]. As the efforts to integrate two different domains are increased, vocabulary issues will be essential.

Privacy

Individual genomic data are easily identifiable and can be used to predict current and future health status. Thus, extracting knowledge from large health data employs a significant risk of privacy information breach; thus, researchers need to consider Health Insurance Portability and Accountability Act (HIPAA) and Institutional Review Board (IRB) requirements for building a privacy-preserving and trustworthy database infrastructure and conducting research [28]. When personal genetics data can be incorporated into EMR systems, different safeguards to ensure confidentiality will be required. A de-identified bio-data warehouse combining traditional clinical and genomic information will be essential to conduct translational research.

Era of Open Global Collaboration

One major difference between the EMR community and BI community is the degree of multidisciplinary open collaboration. In EMRs, multidisciplinary collaborative efforts have been limited. Most of the EMR systems in use today are based on proprietary software that hampers data exchange with other systems. In the BI community, global collaboration, aided by wide use of open source software and development methodologies, has been a key success factor. The BI community has been global in scope from the beginning, and information sharing and free exchange of software tools have ushered in the era of open source software in health informatics. There is a great need for the EMR community to adopt more open source software methodologies that would allow rapid global collaboration (http://www.osehra.org).

Conclusion

Genomic technologies hold the potential to improve the diagnosis and treatment of inherited and complex diseases —including cancer—and facilitate the move towards personalized predictive medicine. The higher throughput and rapidly falling costs of next-generation sequencing have resulted in voluminous genomic data and downstream computational challenges. Thus, the shift from this powerful discovery research to clinical implementation can only be accomplished with careful integration with EMRs, a frontline patient care tool. The most prominent reason to integrate clinical information and biology information under the same system is to provide opportunities for bi-directional exchange of data, technology, and knowledge between two disciplines with different histories and cultures. Additionally, open global cooperation will provide opportunities to make rapid progress in understanding, treating, and preventing human diseases.

Acknowledgments

This study was supported by a grant from the Ministry for Health, Welfare and Family Affairs (A120175).

References

1. Lee J, Kuo YF, Goodwin JS. The effect of electronic medical record adoption on outcomes in US hospitals. *BMC Health Serv* Res 2013;13:39.
2. Laakko T, Leppänen J, Lähteenmäki J, Nummiaho A. Mobile health and wellness application framework. *Methods Inf Med* 2008;47:217-222.
3. Hood L, Galas D. The digital code of DNA. *Nature* 2003; 421:444-448.
4. Weston AD, Hood L. Systems biology, proteomics, and the future of health care: toward predictive, preventative, and personalized medicine. *J Proteome Res* 2004;3:179-196.
5. Wang X, Liotta L. Clinical bioinformatics: a new emerging science. *J Clin Bioinforma* 2011;1:1.
6. Bernstam EV, Smith JW, Johnson TR. What is biomedical informatics? *J Biomed Inform* 2010;43:104-110.
7. Sanz F, Brunak S, Lopez-Alonso V. INBIOMEDvision: promoting and monitoring biomedical informatics in Europe. In: The 23rd International Conference of the European Federation for Medical Informatics Conference, 2011 Aug 28-31, Oslo.
8. The eMERGE Network. Nashville: eMERGE Network, 2013. Accessed 2013 Nov 1. Available from: http://emerge.mc.vanderbilt.edu/.
9. Hanauer DA, Zheng K, Ramakrishnan N, Keller BJ. Opportunities and challenges in association and episode discovery from electronic health records. *IEEE Intell Syst* 2011;26:83-87.
10. Lin YK, Brown RA, Yang HJ, Li SH, Lu HM, Chen H. Data mining large-scale electronic health records for clinical support. *IEEE Intell Syst* 2011;26:87-90.
11. American Medical Informatics Association. Bethesda: AMIA, 2013. Accessed 2013 Nov 26. Available from: http://www.amia.org/mbrcenter/wg/gen/.
12. Bainbridge MN, Wiszniewski W, Murdock DR, Friedman J, Gonzaga-Jauregui C, Newsham I, *et al*. Whole-genome sequencing for optimized patient management. *Sci Transl Med* 2011;3:87re83.
13. Lindhurst MJ, Sapp JC, Teer JK, Johnston JJ, Finn EM, Peters K, *et al*. A mosaic activating mutation in *AKT1* associated with the Proteus syndrome. *N Engl J Med* 2011;365:611-619.
14. Lupski JR, Reid JG, Gonzaga-Jauregui C, Rio Deiros D, Chen DC, Nazareth L, *et al*. Whole-genome sequencing in a patient with Charcot-Marie-Tooth neuropathy. *N Engl J Med* 2010; 362:1181-1191.
15. Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, Dent KM, *et al*. Exome sequencing identifies the cause of a mendelian disorder. *Nat Genet* 2010;42:30-35.
16. St Hilaire C, Ziegler SG, Markello TC, Brusco A, Groden C, Gill F, *et al*. NT5E mutations and arterial calcifications. *N Engl J Med* 2011;364:432-442.
17. Worthey EA, Mayer AN, Syverson GD, Helbling D, Bonacci BB, Decker B, *et al*. Making a definitive diagnosis: successful clinical application of whole exome sequencing in a child with intractable inflammatory bowel disease. *Genet Med* 2011;13: 255-262.
18. Chen H, Chiang RH, Storey VC. Business intelligence and analytics: from big data to big impact. *MIS Q* 2012;36:1165-1188.
19. Collins FS, Patrinos A, Jordan E, Chakravarti A, Gesteland R, Walters L. New goals for the U.S. Human Genome Project: 1998-2003. *Science* 1998;282:682-689.
20. Benson DA, Boguski MS, Lipman DJ, Ostell J, Ouellette BF, Rapp BA, *et al*. GenBank. *Nucleic Acids Res* 1999;27:12-17.
21. Bairoch A, Apweiler R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res* 2000;28:45-48.
22. Saxonov S, Daizadeh I, Fedorov A, Gilbert W. EID: the Exon-Intron Database-an exhaustive database of protein-coding intron-containing genes. *Nucleic Acids Res* 2000;28:185-190.
23. Ruiz M, Giudicelli V, Ginestoux C, Stoehr P, Robinson J, Bodmer J, *et al*. IMGT, the international ImMunoGeneTics database. *Nucleic Acids Res* 2000;28:219-221.
24. Lindberg D, Humpreys B. The Unified Medical Language System (UMLS) and computer-based patient records. In: *Aspects of the Computer-Based Patient Record* (Ball MJ, Collen MF, eds.). New York: Springer-Verlag, 1992. pp. 16-75.
25. Hoffman MA. The genome-enabled electronic medical record. *J Biomed Inform* 2007;40:44-46.
26. Hoffman M, Arnoldi C, Chuang I. The clinical bioinformatics ontology: a curated semantic network utilizing RefSeq information. *Pac Symp Biocomput* 2005:139-150.
27. Kohane IS. Bioinformatics and clinical informatics: the imperative to collaborate. *J Am Med Inform Assoc* 2000;7:512-516.
28. Gelfand A. Privacy and biomedical research: building a trust infrastructure: an exploration of data-driven and process-driven approaches to data privacy. *Biomed Comput Rev* 2011/2012 winter:23-28.