EHR Big Data Deep Phenotyping

Contribution of the IMIA Genomic Medicine Working Group

L. J. Frey[1], L. Lenert[2], G. Lopez-Campos[3]

[1] Chair IMIA Genomic Medicine WG, Biomedical Informatics Center, Medical University of South Carolina, Charleston, SC, USA

[2] Biomedical Informatics Center, Medical University of South Carolina, Charleston, SC, USA

[3] Vice-Chair IMIA Genomic Medicine WG, Health and Biomedical Informatics Centre, The University of Melbourne, Parkville, Victoria, Australia

Summary

Objectives: Given the quickening speed of discovery of variant disease drivers from combined patient genotype and phenotype data, the objective is to provide methodology using big data technology to support the definition of deep phenotypes in medical records.

Methods: As the vast stores of genomic information increase with next generation sequencing, the importance of deep phenotyping increases. The growth of genomic data and adoption of Electronic Health Records (EHR) in medicine provides a unique opportunity to integrate phenotype and genotype data into medical records. The method by which collections of clinical findings and other health related data are leveraged to form meaningful phenotypes is an active area of research. Longitudinal data stored in EHRs provide a wealth of information that can be used to construct phenotypes of patients. We focus on a practical problem around data integration for deep phenotype identification within EHR data. The use of big data approaches are described that enable scalable markup of EHR events that can be used for semantic and temporal similarity analysis to support the identification of phenotype and genotype relationships.

Conclusions: Stead and colleagues' 2005 concept of using light standards to increase the productivity of software systems by riding on the wave of hardware/processing power is described as a harbinger for designing future healthcare systems. The big data solution, using flexible markup, provides a route to improved utilization of processing power for organizing patient records in genotype and phenotype research.

Keywords

Deep phenotype, ontology, big data, genome, electronic health record

Yearb Med Inform 2014:206-11
http://dx.doi.org/10.15265/IY-2014-0006
Published online August 15, 2014

Introduction

Enormous volumes of electronic data are collected on patients across the world at an ever-increasing rate. Traditional clinical workflows are already overwhelmed with the problem of too much data and too little time [1]. The idea of finding relevant knowledge for the patient at hand from the millions of clinical experiences accumulated in Electronic Health Records (EHR) is a daunting task--how do we search for such knowledge and how do we know the associations observed are valid? Starren and colleagues argue that including next generation sequencing within EHRs will further overwhelm workflows in clinical practice [2] and poses a host of challenges [3].

How can informaticists create the next generation of information systems that address both the present and future challenges of data scope and diversity? Stead et al. [4], in a seminal paper, describe the problem of building large health information systems as a tension between investment in software-based and hardware/processor-power-based solutions. While most developers and informaticists believe that a solution lies in better software capable of handling the new complexities of medical care, the upshot of the Stead and colleagues article is that investment in more complex software systems may not be the best approach to solving the growing complexity of clinical data. They argue that software development improves linearly, with about five percent gains yearly in efficiency and capabilities to address complexities. In contrast, hardware processing power has improved exponentially (Moore's Law), doubling every 18 months. Simpler software systems that rely on hardware processing power to address complexity have an inherent advantage.

To shift healthcare to a model that relies more on processing power than on software complexity, Stead et al. propose the use of an internet like approach to health information system development using light prescriptive standards (e.g., URL, HTTP and HTML) that support flexible markup coupled with massive indexing of simple data stores. Heralding the rise of "big data" architectures, they propose combining "light" methods with reference standards based on ontologies and vocabularies for flexible markup of EHR data. The tagging of clinical information at different levels of specificity is modeled on the standards approaches that enabled the rapid expansion of the Internet. Reference standards can evolve over time, which is important in domains where knowledge is rapidly evolving such as genomic data in medicine. To date, the approach remains theoretical, but the success of big data architectures in other industries suggests that it is both feasible and advantageous.

Background

Although extant EHR systems could be overwhelmed by genomic scale data, combining data with omics resources and associating deep phenotypes with patient records will advance knowledge discovery of genetic disorders [5] and support the practice of personalized medicine [6]. To reach this

goal without overwhelming EHRs there are a number of challenges that must be overcome: phenotyping in the EHR, omics data representation and big data analytics in medicine. Researchers are actively addressing these challenges and accelerating advancement.

Phenotyping in the EHR

Phenotype is specified through the expressed characteristics of an organism that result from variation in its genotype interacting with an environment [7]. Deep phenotyping [6] extends phenotyping into clinical data with the variation of clinical concepts collected over patient-clinician encounters used to define phenotypic cohorts of patients. From the perspective of clinical records, a phenotype is a collection of clinical traits and measurements that describe fundamental attributes of a patient. The phenotype can be a single trait such as race or a collection of events that compose a cohort of patients that are of interest for a particular question being investigated. The use of constellations of events to define phenotype provides a way of specifying the criteria for a cohort of patients. The use of groups of events allows the definition of a phenotype to go beyond diagnosis coded with International Classification of Disease, version 9 codes (ICD9) and potentially assesses the accuracy of assigned codes [8].

A prerequisite to achieving precision medicine is the systematic study of phenotype abnormalities through deep phenotyping that identifies human deviations in morphology, physiology and behavior [6]. Through controlled experiments with precise phenotype definitions, phenotypes have been developed extensively in animal models [9]. In clinical settings, on the other hand, the data are noisy and collected for the purpose of delivering medical treatment at the point of care rather than phenotyping. Consequently, data stored in EHRs do not have the same consistency and precision of data collected for experiments.

Hripcsak and Albers discuss the challenges to phenotyping in the EHR such as incompleteness, inaccuracy, complexity and bias [8]. They propose studying the complexities of the EHR as a means of improving phenotype collection and improving EHR processes to support phenotype development. They also touch upon expanding the way in which phenotypes can be defined in terms of time series data analysis. They describe the process of phenotyping in the EHR as an iterative approach in which experts curate a set of cohort patients expressing a phenotype in order to create a training data set. Features are then extracted from the EHR, rules are constructed, and sensitivity along with specificity are measured until they reach an acceptable validated level for the training data. The rules are then applied to the full data set. The consistency of vocabularies and ontologies used to describe EHR data can be explored to determine the impact they have on the rules generated and complexity of extracting phenotype from medical data.

Prescriptive phenotype definition is not the only approach to the problem; it is also possible to define a phenotype by example based on medical events. The field is currently advancing techniques in ontology construction for developing phenotype and similarity measures to search and match phenotypes [10, 11]. Ontologies and vocabularies are central to phenotype definition within EHR data. The Unified Medical Language System (UMLS) [12] provides a repository of vocabularies with which to consistently markup medical records. Expanding potential applications of the repository, conceptual similarity has been used to relate it to other knowledge sources [13]. The UMLS contains vocabularies such as the Logical Observations Identifiers, Names, Codes (LOINC) used to encode lab tests [14], RxNorm [15] for normalized medication names, ICD9 [16] for diagnostic codes and Systematized Nomenclature of Medicine--Clinical Terms (SNOMED CT) with high concept coverage and explicit semantic relationships [17], which combined, become very relevant for the specification of phenotype.

Ontologies

Using a reference standard to harmonize phenotype research, the Human Phenotype Ontology project created a common terminology and ontological representation that can be used for consistently categorizing phenotypes in human disease [18]. It connects with the Online Mendelian Inheritance in Man (OMIM) resource and extends it through the use of a controlled vocabulary for consistent labeling [19]. Doelken et al. [10] describe the Human Phenotype Ontology project and its continuous build architecture. They also describe software that facilitates consistency management with resources such as OMIM.

The use of ontological information enables the patient history or event streams to be organized with hierarchical data structures that allow flexible annotation and searching of the data. An organizational mechanism based on directed acyclic graph (DAG) representations can be used to encode the ontology. The use of DAGs is widespread in genomics (e.g., the Gene Ontology and Sequence Ontology).

A strategy for managing the dynamic domain of genotype and phenotype data is to combine highly indexed flexible ontological markup of data files with ontology distance measures of events in the patient population. Software systems store events from patients as clinical encounters with clinical content (much of it is free text) and metadata on context of collection. By marking up observations from each clinical encounter using a flexible and extensible format, a stream of events can be associated with each patient. The primary interest is to draw similar cohorts from streams of clinical events in the EHR.

Semantic and Temporal Similarity

There are multiple approaches to calculating similarity measures between patient cases such as path length between terms in an ontology or similarity of temporal sequences that occur in the patient's record. A distance measure can be computed between terms by searching for the shortest path connecting them. Köhler et al. [20] examine clinical diagnostics using semantic similarity searches on clinical feature that describe phenotypes. The tool uses the Human Phenotype Ontology to augment the searches. They validated the tool with sim-

ulated data and describe how the approach can be applied to assist in diagnostic workflow. Girdea et al. [11] present PhenoTips a web-based tool for documenting phenotype information that is used within clinical encounters. The open source software uses the Human Phenotype Ontology and connects with OMIM. PhenoTips has been deployed and used to collect anonymized patient phenotype information for three research projects in hospitals across Canada. The tool is specifically designed to fit within clinical workflows and has incorporated feedback from clinicians using the system.

Inherent in the concept of similarity are representations of patient-history. While the genome of a patient may be stable, the interpretation of the variants, the effects of disease and environment evolve over time. The evolution of a patient's health events related to a disease may follow the same or a similar trajectory to other patients. Leveraging the concept of similar trajectories, an approach to predicting an index patient's health events could use temporal alignment of the health records from comparator patients [21].

Systems such as Lifelines2, discussed in the visualization literature, apply techniques to identify and align patients. Specifically, these systems use an index case to find matching cases, those with exact matches of ordered events to the index case. The ordered events from the matched cases are aligned to the index case and used to predict future events for the index case [22]. When multiple cases are aligned to an index case, the range of outcomes from the matched cases provides a prediction guide for the index case.

The approach of finding patients that match an index patient assumes that the underlying illness and the course of prior events are similar [22]. This is a reasonable assumption, if one assumes that the clinical phenotype of a patient is given by both underlying medical conditions and the aligning medical events. Expanding this conceptual definition of phenotype, Wongsuphasawat [23] extends Lifelines2 to include differentiation between and filtering out of unimportant events, inclusion of demographic features, and modeling of the trajectory those patients took to reach the alignment point [24]. This further expands the conceptual definition of phenotype. Phenotype is not only the disease but also its response over time to native homeostatic mechanisms and to treatment.

Even if there is similarity in the underlying conditions (e.g., a myocardial infarction), the health events ordered across time (e.g., a myocardial infarction, followed by congestive heart failure and low blood pressure) may also serve an important role in defining phenotype. If a patient has had several prior heart attacks or had a heart transplant prior to the heart attack, knowledge of the path that patients followed prior to the aligning event may be critical in developing a phenotype that could be used for treatment planning. When using EHRs and as the number of events considered increases, the probability of finding cases with an exact match in the sequence becomes reduced. Lee and colleagues [25, 26] have explored relaxing the requirements for exact match by weighting the differences in comparisons of sequence events using dynamic programming methods, however, this approach faces challenges with exponential complexity in the number of patterns. Adding difficulty, similar events in a sequence must be considered as well. For example, a patient who has asthma that follows treatment with the beta blocker metoprolol is similar to one who has asthma that follows treatment with a beta blocker propranolol, but not necessary to one who has asthma that follows treatment with aspirin.

Approaches to address this complexity may require application of tools such as Ayers et al.'s [27] Sequential Pattern Mining Using Bitmaps (SPAM) algorithm [28] or artificial intelligence methods for temporal abstraction, such as those applied in Shahar's and colleagues Knave II application [29]. Nonetheless, deep phenotyping has to include the course a patient has followed to reach a particular point in time, as this ultimately reflects concepts of disease progression, complications of illness, and response (or non response) to treatment.

Extending analysis approaches with time series information from the EHR enables the recognition of trending patterns that enhance phenotypic description. By constructing deep phenotypes from clinical notes and other medical findings, a more precise description of a patient's health and treatment options can be developed and similarities between patients can be identified, especially in relationship to genomic data and molecular drivers of the phenotypes.

Omics Data Representation

In the last two decades medicine has witnessed a revolution in the development and use of different molecular biology and "-omics" technologies and methodologies. Traditionally used as powerful tools towards a better understating of the mechanisms associated with disease, they are transitioning to critical tools for achieving improved healthcare by means of precision medicine. In the last five years this trend has been reinforced by the developments and evolution of sequencing methods that have substantially reduced the costs of accessing these technologies, and thus, facilitate their adoption for clinical applications [30]. These new methodologies have brought with them an explosion in the volumes of data generated and pose a challenge for their management and interpretation.

Almost simultaneously, there has been a movement towards the widespread use and adoption of EHRs in the clinical setting. Consequently, EHR developers have to incorporate new forms and data types generated in the genomics and molecular fields, manage them effectively and present the assay results in a meaningful way to users. The speed in the advances and changes in genomics represents additional challenges for EHRs in terms of variability and quantity of data generated.

This increasing complexity caused by the evolution of genomic technologies is exemplified in the changes associated with differences in the management of laboratory results. Initially laboratories were focused on a single gene analysis or gene panels. This was followed by the management of millions of variants identified in genotyping experiments based on genome wide association studies (GWAS) and more recently the management of whole exome sequencing (WES) and whole genome sequencing (WGS). In 2003, when the Human Genome Project was finished, it meant the successful

accomplishment of a multibillion dollar international project that required several years to complete, now new advanced next generation sequencing techniques have reduced the costs and data turnaround in a manner that have made gathering individual genomes for clinical purposes a reality. The investment in WES technology combined with patient data has resulted in over 100 Mendelian disease variants being identified in the past three years [5].

The big volumes of data generated with the latest genomic approaches add an important challenge for biomedical informatics and EHRs not just because of their size itself but also because of the processing, analysis requirements and methodologies that must be applied in order to present the data in a useful and meaningful manner for clinical users. These processes need to generate key metadata to be included with the genomic data to aid in their interpretation. The metadata should include aspects such as the laboratory techniques, bioinformatics, tools, databases and pipelines used to generate those data.

Masys et al. [3] described some of the challenges associated with the inclusion of genomic data in EHRs. They identified seven challenges: Separation of raw data and interpreted data; Annotation of data generation and processing; Requirement of lossless compression methods to reduce data footprint; Presentation of the clinically actionable data; Use of human and machine readable formats to facilitate the design and implementation of decision support methods; Anticipation changes in genomic variation; and finally Design of systems supporting clinical care and research.

Although the term "omics" covers a multitude of different approaches (e.g., proteomics, transcriptomics, metabolomics, microbiomics), most of the current efforts to incorporate these data in the EHRs have been focused on the integration of genomics information. Even this has been generally limited to the inclusion of a single genome per individual whereas it is possible to find different genomes within an individual in different situations such as transplant recipients, chimerism or cancer. Therefore the possibilities of having to integrate multiple genomes into the EHRs are real and should be considered in the design of future systems. Additionally in the last couple of decades the landscape of gene expression analysis has been dominated by microarray technology but the reduction in the cost of sequencing technologies is leading towards the replacement of microarrays by RNA-Seq (RNA sequencing) analysis bringing an additional source of large volumes of sequencing data that should be managed.

The complexity of integrating genomic data into EHRs and the clinical workflow become a rationale for intermediate solutions, specifically, the development of ancillary systems that incrementally evolve increasing functionality [2]. Ancillary technologies provide a route to gradually incorporate big data infrastructure into EHR systems.

To achieve a successful integration of genomic data into the EHRs it is necessary to adapt and develop available standards to ensure efficient data and information exchange between the laboratories where the data are generated, the EHR and their users and in some cases as well with the possible ancillary repositories where the genomic data are stored [2]. Standards for genomic variants such as the genome variation format (GVF) [31] can be used to store data along with some existing terminologies and ontologies, such as LOINC, SNOMED-CT, that have been adapted for these new data. The approach of combining consistent markup with efficient genomic storage is a key aspect to ensure the successful use of genomic data in the EHRs.

Despite the many hurdles, there are numerous examples where genomic data have been successfully included in the EHRs for both research and clinical purposes such as those from the electronic Medical Records and Genomics (eMERGE) consortium [32]. Newton et al. [33] provide a comprehensive description of phenotyping processes in the eMERGE network. They touch upon the complexity of the task and methodologies for achieving it through the use of machine learning and data reduction methods.

Another major challenge from a technology perspective is integrating the different types of omics data with phenotype identification. Ontologies can be used to both integrate data from diverse sources through unified semantics and to provide relationships for computational analysis such as semantic similarity. Similarity metrics can be calculated through ontologies and other algorithms to model the degree of content similarity to identify phenotypes. This explosion of data requires adoption of new technology such as big data approaches for managing and analyzing it.

Big Data Analytics in Medicine

The term, big data, describes a collection of data that pose challenges to traditional data processing approaches (e.g., relational databases). The challenges are derived from the following characterization: volume (denoting size), variety (indicating heterogeneity), veracity (representing accuracy) and velocity (designating processing speed). Combinations of these four characteristics can result in a big data problem that cannot be scaled using traditional databases and analysis systems. Medical data over time combined with genomic data becomes a big data problem due primarily to volume and variety and becomes a velocity problem with the use of real-time data.

New methods for big data parallel processing have been developed using a functional program paradigm. With these new big data approaches, the algorithms are brought to where the data are instead of shipping the data to different cores in a computing environment. The methods are based on Google's file system [34] and BigTable [35], a sparse, distributed, persistent multidimensional sorted map that increases performance by taking a large number of records and parallelizing their processing over many machines. An ecosystem of open source tools (e.g., Hadoop, MapReduce, Spark, HBase, Accumulo, Mahout, CouchDB and MongoDB) have been implemented and applied to big data problems such as Facebook's real-time messaging environment [36]. The use of NoSQL database solutions combined with parallelized MapReduce jobs applied to medical data has the potential to change the way deep phenotypes are constructed. The discovery of deep phenotypes can be expanded and scaled through the use of big data methodologies to include patterns of time series of events from the EHR.

Another strength of the big data Hadoop system is the ability to commoditize the use of hardware for scaling data storage and computation. New nodes can be added to scale with storage needs. Because the knowledge around genomic variant and omics data will change significantly over time, computationally powerful and dynamic systems are needed that can re-analyze data as new knowledge is created. A value added component of big data systems is the ability to store and process variant information and utilize it when its relevance is identified. Given the growth of genomic data in clinical systems, such as the VA's Million Veterans Project, the ability to incrementally scale the storage and analysis platform is highly desirable.

Discussion

Data integration is a key component to building huge data systems filled with biological, genomic, clinical, phenotype and other health related data. Data integration involves combining or linking data from multiple sources to enable data sharing, expanded data sets, secondary analysis/reuse of data and broadening multidisciplinary collaborations. In Seoane et al. [37] review of data integration in genomic medicine, they observed that data integration approaches of cross-linking, data warehouses and federation are suitable for particular applications, but are not general solutions. The problem is a plethora of small heterogeneous data sets that resist integration through the complexity of variety. The cases of EHR data, omics data and deep phenotyping involve the big data variety problem. Hadoop data stores offer a new approach to reduce and manage the complexity of high variety data.

Although difficult, the variety problem in big data can be addressed through the use of BigTable paradigm because the data can be stored in a raw format and transformed at the time it is needed with as much precision as the raw format encodes. Rather than the data warehouse paradigm that needs to harmonize to a canonical representation, the new *big data methods for integration have the ability to store data without normalizing it in a relational data model*. The reference standard approach supports the management of the flexible markup of patient records along with using ontologies to organize and search those records [4]. The phenotype and genotype of the patient can then be maintained through the denormalized markup language developed for knowledge discovery in EHR data. The NoSQL solution does not preclude the development of standardized representations; it just does not make heavy standardization and normalization a prerequisite to integrating high variety data into the system.

Using deep phenotype information at the point of care introduces the need for real-time analysis to meet the requirements of point of care services. To present potential deep phenotypes to the clinician that incorporates time series data from the EHR, the analysis system needs to have parallelizable components that can break the task into independent chunks. This enables parallel algorithms to speed processing time and to deliver results with acceptable response times.

Given the existence of scalable big data stores and analysis capabilities, deep phenotyping analysis can be applied to time series and trending information in medical records. Clinical data are loaded into the Hadoop cluster and analyzed along with other data within the cluster. Having all the data within the Hadoop cluster allows phenotype and genotype data to be linked and analyzed on a common platform. The analysis approach involves writing the appropriate MapReduce program to assess similarity of patients in deep phenotypic cohorts. The Mahout MapReduce code base has been developed for machine learning using Hadoop. It implements learning algorithms such as nearest neighbor and handles running the mappers and reducers across a cluster and outputs the resulting classification model. Mahout provides a solution within the Hadoop framework to parallelize analysis and increase performances to potentially address speed requirements at point of care.

Conclusions

The scientific advances in the types of data available and the diversity of algorithms for phenotyping analysis of clinical data put software development further behind. Traditional relational data warehouses are well suited for facts aggregation over dimensions that can be preconfigured for fast query answering. This is useful for identifying patterns indexed by dimension tables, but is difficult to apply to time series information, which is core to understanding or predicting a patient's health trajectory. A patient's health information over time is critical for deep phenotyping. Specifically, it is necessary to understand a patient's response to treatment where clinical measurement trends with the delivery of a therapeutic intervention. New approaches are needed to deal with the scale of clinical data and the rapidly expanding diversity of algorithms. These approaches, heralded in Stead et al. [4], focus on simple data models augmented by extensive, flexible indexing driven by raw computing power.

Many of the discussed challenges are tightly intertwined and are critical for achieving precision medicine. Omics is an extremely dynamic field and our knowledge about the effects associated with the different variants is continuously evolving and being updated. Analysis and interpretation of omics data are supported by the existing knowledge and predictions available at the moment of the analysis. Interpretation of some variants may change over time and it is necessary to keep open the possibility of reinterpreting the data using the advances in knowledge and interpretation of the human variants as well as applying improved analytical tools and pipelines. Advances require gaining access to new and extended datasets that inform knowledge discovery on health and diseases across different populations. For this reason it is important that omics data are accessible for research purposes under the appropriate ethical approval. The relevance of this sharing process for research comes because in many aspects and despite the noise and difficulties to mine and extract information from the EHRs, they represent the best data annotation source, that when combined with omics data, can advance the practice of precision medicine.

By treating EHR data as clinical event streams, a number of new big data methods can be developed and adapted from the technology sector. The content of these streams can be processed in combination with strat-

egies for conceptual markup of events and matching of event streams, to rapidly retrieve and identify phenotypes. Specifically, big data solutions can use tagged data coupled with ontologies to identify phenotypes. The growth of clinically relevant deep phenotyping in this genomic medicine era depends on the application of flexible and evolving approaches to nosology, that is in turn, enabled by a move to new, computational-intensive big data architectures.

Acknowledgments

The authors LJF and LL were supported by the grant 1R01GM108346-01:BIGDATA: Mid-Scale: DA: Techniques to Integrate Disparate Data: Clinical Personalized Pragmatic Predictions of Outcomes (Clinical3PO; Co-PIs: Lenert & Frey) awarded by the National Institute of General Medical Sciences: NIGMS.

References

1. Cases M, Fulong LI, Albanell J, Altman RB, Bellazzi R, Boyer S, et al. Improving data and knowledge management to better integrate health care and research. J Intern Med 2013:321-8.
2. Starren J, Williams MS, Bottinger EP. Crossing the omic chasm: a time for omic ancillary systems. JAMA 2013 Mar 27;309(12):1237-8.
3. Masys DR, Jarvik GP, Abernethy NF, Anderson NR, Papanicolaou GJ, Paltoo DN, et al. Technical desiderata for the integration of genomic data into Electronic Health Records. J Biomed Inform 2012 Jun;45(3):419-22.
4. Stead WW, Kelly BJ, Kolodner RM. Achievable Steps Toward Building a National Health Information Infrastructure in the United States. J Am Med Inform Assoc 2005;12(2):113-21.
5. Rabbani B, Mahdieh N, Hosomichi K, Nakaoka H, Inoue I. Next-generation sequencing: impact of exome sequencing in characterizing Mendelian disorders. J Hum Genet 2012 July:621-32.
6. Robinson, PN. Deep Phenotyping for Precision Medicine. Hum Mutat 2012;33(5):777-80.
7. Dawkins R. The Extended Phenotype. Oxford: Oxford University Press; 1989.
8. Hripcsak G, Albers DJ. Next-generation phenotyping of electronic health records. J Am Med Inform Assoc 2013;20(1):117-21.
9. Darvasi A. Experimental strategies for the genetic dissection of complex traits in animal models. Nat Genet 1998;18:19-24.
10. Doelken SC, Mungall CJ, Bauer,S, Firth HV, Bailleul-Forestier I, Black GCM., et al. The Human Phenotype Ontology project : linking molecular biology and disease through phenotype data. Nucleic Acids Res 2013:1-9.
11. Girdea M, Dumitriu S, Fiume M, Bowdin S, Boycott KM, Chitayat D, et al. PhenoTips: Patient Phenotyping Software for Clinical and Research Use. Hum Mutat 2013;34(8):1057-65.
12. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res 32 (Database issue). 2004:267-70.
13. Bodenreider O, Burgun A. Aligning knowledge sources in the UMLS: methods, quantitative results, and applications. Medinfo Proc 2004:327–31.
14. Wilson PS, Scichilone RA. LOINC as a data standard: how LOINC can be used in electronic environments. J AHMIA 2011;82(7): 44-47.
15. Nelson SJ, Zeng K, Kilbourne J, Powell T, Moore R. Normalized names for clinical drugs: RxNorm at 6 years. J Am Med Inform Assoc 2011; 8:441-8.
16. International Classification of Diseases version 9. [15 December 2013]; http://www.icd9data.com/2006/Volume1/
17. Melton GB, Parsons S, Morrison FP, Rothschild AS, Markatou M, Hripcsak G. Inter-patient distance metrics using SNOMED CT defining relationships. J Biomed Inform 2006;39:697-705.
18. Robinson PN, Mundlos S. The Human Phenotype Ontology. Clin Genet 2010;77:525-34.
19. McKusick VA. Mendelian Inheritance in Man. A Catalog of Human Genes and Genetic Disorders. Baltimore: Johns Hopkins University Press; 1998 (12th edition).
20. Köhler S, Schulz MH, Krawitz P, Bauer S, Dölken S, Ott CE, et al. Clinical Diagnostics in Human Genetics with Semantic Similarity Searches in Ontologies. Am J Hum Genet 2009;85:457-64.
21. Hajihashemi Z, Popescu M. An Early Illness Recognition Framework Using a Temporal Smith Waterman Algorithm and NLP. Proc AMIA Fall Symp 2013:549-57.
22. Wang TD, Plaisant C, et al. Aligning Temporal Data by Sentinel Events: Discovering Patterns in Electronic Health Records. CHI 2008 Proceedings: Health and Wellness, Florence, Italy; 2008.
23. Wongsuphasawat K, Gomez JAG, et al. LifeFlow: Visualizing an Overview of Event Sequences. CHI 2011, Vancouver, BC, Canada; 2011.
24. Wongsuphasawat,K, Gotz DH. Outflow: Visualizing Patient Flow by Symptoms and Outcome. IBM; 2011. p. 1-4.
25. Lee WN, Das AK. Local Alignment Tool for Clinical History: Temporal Semantic Search of Clinical Databases. AMIA Annu Symp Proc 2010:437–41.
26. Lee WN, Bridewell W, Das AK. Alignment and Clustering of Breast Cancer Patients by Longitudinal Treatment History. AMIA Annu Symp Proc 2011;2011:760-7.
27. Ayres J, Gehrke J, Yiu T, Flannick J. Sequential Pattern Mining using a Bitmap Representation. In: SIGKDD'02, Edmonton, Canada; 2002 July.
28. Papapetrou P, Kollios G, Sclaroff S, Gunopulos D. Mining frequent arrangements of temporal intervals. Knowl Inf Syst 2009;21(2):133–71.
29. Shahar Y, Goren-Bar D, Boaz D, Tahan G. Distributed, intelligent, interactive visualization and exploration of time-oriented clinical data. Artif Intell Med 2006;38(2):115–35.
30. 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 Jan;42(1):30-5.
31. Reese MG, Moore B, Batchelor C, Salas F, Cunningham F, Marth GT, et al. A standard variation file format for human genome sequences. Genome Biol 2010;11(8):R88.
32. McCarty CA, Chisholm RL, Chute CG, Kullo IJ, Jarvik GP, Larson EB, et al. The eMERGE Network : A consortium of biorepositories linked to electronic medical records data for conducting genomic studies. BMC Med Genomics 2011;4(1):13.
33. Newton KM, Peissig PL, Kho AN, Bielinski SJ, Berg RL, Choudhary V, et al. Validation of electronic medical record-based phenotyping algorithms : results and lessons learned from the eMERGE network. Medical Informatics 2013;20:e147-e154.
34. Ghemawat S, Gobioff H, et al. The Google File System. SOSP'03, Bolton Landing, New York, USA; 2003.
35. Chang F, Dean J, et al. Bigtable: A Distributed Storage System for Structured Data. OSDI '06: 7th USENIX Symposium on Operating Systems Design and Implementation; 2006. p. 205-18.
36. Borthakur D, Sarma JS, et al. Apache Hadoop Goes Realtime at Facebook. SIGMOD '11, Athens, Greece. 2011.
37. Seoane JA, Dorado J, Pazos A. Data Integration in Genomic Medicine: Trends and Applications. IMIA Yearbook 2012: Personal Health Informatics 2012:117-25.

Correspondence to:
Lewis J Frey
Chair IMIA Genomic Medicine WG
Biomedical Informatics Center
Public Health Sciences, Associate Professor
Hollings Cancer Center, Research Member
Medical University of South Carolina
135 Cannon Street, Suite 405K, MUSC 200
Charleston, SC 29425. USA
Tel: +1 843 792 4216
Fax: +1 843 792 5587
E-mail: Frey@musc.edu