"Big Data" and the Electronic Health Record

M. K. Ross*, Wei Wei*, L. Ohno-Machado

Division of Biomedical Informatics, University of California, San Diego, USA

Summary

Objectives: Implementation of Electronic Health Record (EHR) systems continues to expand. The massive number of patient encounters results in high amounts of stored data. Transforming clinical data into knowledge to improve patient care has been the goal of biomedical informatics professionals for many decades, and this work is now increasingly recognized outside our field. In reviewing the literature for the past three years, we focus on "big data" in the context of EHR systems and we report on some examples of how secondary use of data has been put into practice.

Methods: We searched PubMed database for articles from January 1, 2011 to November 1, 2013. We initiated the search with keywords related to "big data" and EHR. We identified relevant articles and additional keywords from the retrieved articles were added. Based on the new keywords, more articles were retrieved and we manually narrowed down the set utilizing predefined inclusion and exclusion criteria.

Results: Our final review includes articles categorized into the themes of data mining (pharmacovigilance, phenotyping, natural language processing), data application and integration (clinical decision support, personal monitoring, social media), and privacy and security.

Conclusion: The increasing adoption of EHR systems worldwide makes it possible to capture large amounts of clinical data. There is an increasing number of articles addressing the theme of "big data", and the concepts associated with these articles vary. The next step is to transform healthcare big data into actionable knowledge.

Keywords

Electronic health records, data mining, natural language processing, privacy, security, quality improvement

Yearb Med Inform 2014:97-104
http://dx.doi.org/10.15265/IY-2014-0003
Published online August 15, 2014

Introduction

The term "big data" encompasses concepts in existence for decades, and its definition is evolving. The term seems to have been first derived from an IT strategic consulting group's approach to manage data volume, velocity, and variety [1]. In a recent review exploring the definition of "big data," Ward and Barker amalgamate concepts of size, complexity, and technology to define "big data" as "*the storage and analysis of large and/or complex data sets using a series of techniques including…machine learning"* [2]. The term applies to many fields including marketing, astronomy, search engines, cellular data, social media, politics, and healthcare [3].

The electronic health record (EHR) itself could be considered "big data" and hence extend to the manipulation and application of data stored in EHRs. Incentives from the Health Information Technology (HITECH) Act of 2009 in the United States have, in part, led to an adoption rate approaching 80 percent of certified EHRs in acute care hospitals [4]. Electronic health record adoption rates have also increased worldwide [5-10]. It has been suggested that, in the United States alone, there will soon be one billion patient visits documented per year in EHR systems [11]. In addition to the patient data housed in the EHR, the amount of additional data available about medical conditions, underlying genetics, medications, and treatment approaches is high. Yet human cognition to learn, understand, and process the data is finite. Thus, computer-based methods to organize, interpret, and recognize patterns from these data are needed [12].

While EHR adoption for healthcare is reassuring, it is important that data continue to have secondary use for quality improvement and research that helps improve patient care and potentially limit healthcare costs [13]. Over the years, EHR data have been used with the intent to improve care [14-16], increase patient engagement [17, 18], perform quality improvement [16], build shared models and standardization across institutions [19, 20], create new knowledge [14, 16], conduct research in a "real-world" settings instead of in controlled trials [14, 21], enable public health surveillance and intervention [16,19], and facilitate personalized care and decision-making [15, 16]. The ultimate goal is to create a continually learning healthcare infrastructure with real-time knowledge production [16, 22-26] and create an ecosystem that is predictive, preventive, personalized, and participatory [27].

Utilizing the EHR system to answer healthcare questions differs from the traditional research approach of collecting data after a question is asked. Although EHRs have been in existence for many years (and so has the idea of secondary use of the data), the process is currently not streamlined and many challenges exist. Main challenges include limitations of processing ability[15], interoperability and lack of standardization [3, 18, 19, 28], accuracy and completeness of records [29], cost [30], security and privacy concerns [21], and inability to extract the needed information [31]. In regard to information completeness, Weiskopf *et al.* estimate that, if stringent definitions are

* These authors contributed equally to this manuscript

utilized, less than a quarter of all records are considered complete. These authors encourage the scientific community to raise awareness about this issue and call for researchers to define completeness by four criteria: documentation, breadth, density, and predictive ability [29]. Steps toward improved information extraction and analysis in the USA include the formation of alliances between companies and healthcare institutions [18, 30, 32].

While work in this field has existed for decades, for this *year-in-review* article we focus on the most recent published approaches to secondary use of data in EHR systems and aim to summarize current trends, projects, approaches and challenges.

Methods

Our literature search is limited to "big data" in the context of EHRs included in PubMed from January 1, 2011 to November 1, 2013. Big data research is conducted in multiple well-established disciplines such as statistics, computer science, and biology. Terms that are widely used in big data research articles seem to be derived from existing terms or borrowed from other fields. A good proportion of these terms are not yet included as Medical Subject Heading (MeSH) terms. To capture the latest trends, we searched for the new terms even if they were not MeSH terms. As a result, we identified some articles that would not have otherwise been detected, but the tradeoff was that we may have failed to identify relevant articles that use the most traditional terms. We developed a key word determination strategy to cover as many terms as possible, in which new keywords were iteratively extracted from the articles that were retrieved with predetermined keywords.

We initialized the search with "big data AND Electronic Health Records" and "data mining AND Electronic Health Records". The term "big data" is not a MeSH term; it is embodied by the MeSH terms of "automatic data processing", "electronic data processing", and "data mining". However, this was important to identify recent articles that were directly related to "big data". The retrieved articles were reviewed and selected manually; additional search terms were determined from the key word section, title, abstract and the main context of selected articles. Table 1 shows the queries we used. The category names were arbitrarily assigned. Inclusion and exclusion criteria were determined with the purpose of identifying articles that focused on secondary use of EHR data.

Inclusion criteria: articles highlighting the secondary use of EHR data to improve healthcare and clinical research; addressing patient data privacy and de-identification, EHR data sharing and access problems; developing and applying new data analysis methods and data visualization techniques to EHR data; natural language processing (NLP), information extraction and information retrieval methods and applications to EHR data; translational research involving at least 1000 patients; new information technologies for healthcare quality improvement.

Exclusion criteria: articles on the implementation of EHR systems including techniques and evaluation, bioinformatics articles without significant clinical emphasis (genomics, genome-wide association studies), public policy, user-interface design, and medical imaging. The latter would be relevant but was excluded due to time and space constraints.

Results

From our initial query terms we retrieved 227 distinct papers. After manual review, we used 84 papers for the results section. Article categories are shown in Figure 1.

Table 1 Query terms derived from collected articles. Statistical model names such as linear regression are not displayed in the category of data analysis. Terms with asterisks are not MeSH terms. We keep non MeSH terms in the keyword list because they appear frequently in big data related articles in biomedical informatics. "Electronic medical records" covers more articles than "electronic health records". "Big data" is a core concept in this field. "Cloud computing" is developed from distributed computing and attempts to host EHR systems in a cloud have been reported. Predictive model is a widely used term in modeling. "Data visualization" is critical for interpreting study outcomes and this technique helps clinicians accept new methods. "Monitoring" is a broad concept, including drug monitoring, patient monitoring, etc., we use "monitoring" to cover all relevant topics. "Access control" is a new but popular topic in patient data privacy research. "Privacy mechanism" covers techniques to protect privacy such as differential privacy. "De-identification" in biomedical informatics usually means removing sensitive personal information to protect privacy following the HIPAA Privacy Rule. Disease "phenotyping" based on clinical characteristics in the EHR is a newer approach in working toward personalized medicine.

Category	Query terms	Category	Query terms
EHR	Electronic Health Records Electronic Medical Records* Personal Health Records	Privacy and Security	Access Control* Confidentiality Consent Patient Data Privacy Data Sharing Privacy Privacy Mechanism* Computer Security
Data Analysis	Artificial Intelligence Big Data* Cloud Computing* Data Mining Data Interpretation Computer Simulation Predictive Model* Statistical Model Visualization	NLP	Controlled Vocabulary De-Identification* Information Extraction Information Storage and Retrieval Knowledge Representation Natural Language Processing Text Mining
Health Information Technology	Clinical Decision Support System Monitoring* Quality Improvement Social Media	Translational Informatics	Surveillance Pharmacovigilance Phenotyping*

Data Mining

Predictive analytics and automated systems to assist with Knowledge Discovery in Databases (KDD) [16, 33] are necessary to build a learning healthcare system. In the past three years, traditional data analysis methods and tools were widely used on EHR data and some examples include disproportionality analysis [34], support vector machine [35] and conditional random fields [36]. New approaches were also introduced. For example, Hrovat *et al.* combined association rule mining, which was designed for mining large transaction datasets, with model-based recursive partitioning to predict temporal trends (e.g. behavioral patterns) for subgroups of patients based on discharge summaries [37]. Sun *et al.* used a sparse regression model to combine EHR data and expert knowledge to identify risks related to adverse conditions [38]. Topics that were popular, possibly due to the availability of large data sets and funding were pharmacovigilance, EHR phenotyping, and NLP applications to extract complex concepts and relations from clinical documents and the literature.

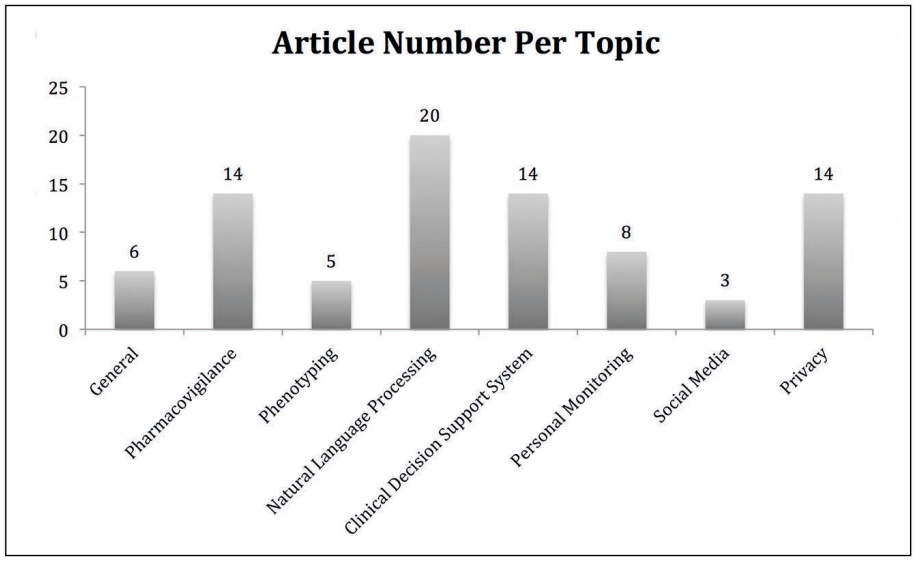

Fig. 1 Articles grouped by topics discussed in the result section

Pharmacovigilance

Adverse drug events are estimated to occur in 30% or more of hospital stays and cost billions of dollars annually [34, 39]. Although medications are tested in a controlled environment through a formal approval process, the methods are not infallible and medications may need to be removed from market, as controlled trials may not represent the general population that ends up actually exposed to the medication [40]. Therefore, post-marketing surveillance is important but traditionally achieved through manual reports by healthcare professionals, patients, and manufacturers. This process is not proactive or expeditious [41]. One study noted that for every hour spent in the development of their semi-automated approach, an estimated 20 hours were saved in manual review [42]. Applying predictive analytics and decision support to the EHR to improve post-marketing surveillance is an important aspect of achieving a learning healthcare system.

Chazard *et al.* used over 100,000 hospital EHRs to show adverse drug event detection in anticoagulant therapy and hyperkalemia, using decision trees and association rules [43]. Ji and colleagues applied a fuzzy-logic-based model to over 16,000 patient records in the VA Health system to evaluate adverse drug events related to specific medications of interest in the drug classes of cholesterol-lowering and angiotensin converting enzyme inhibitors [44]. LePendu *et al.* were able to recreate the significantly increased risk for myocardial infarction in patients with rheumatoid arthritis taking rofecoxib (Vioxx) by extracting clinical data from over one million patient reports using ontologies to map data into standardized clinical concepts [38]. The authors noted that this significant risk was not detected when using ICD-9 codes alone. Other approaches to identify adverse drug events (ADEs) from the clinical record were also successful, including a variety of approaches that used machine learning, keyword features and pattern matching to identify sentences that indicate side effects associated with drug pairings [45], combined molecular structure of medications with clinical data in the EHR to predict other drug interactions [46], processed and normalized concepts found in narrative text for data harmonization [47], identified ADEs related to more than one medication [48], and utilized algorithms to calculate odds ratio of abnormal labs in drug-exposed patients compared to non drug-exposed patients [49]. Harpaz *et al.'s* group demonstrated that clinical information from EHRs can augment signal detection from adverse event reporting databases [50]. Finally, Oliveira *et al.* evaluated a model that combined drug-event data from over 30 million European patients' EHRs and Health Information Exchanges (HIEs) with the literature and used protein and biologic pathway data to identify new ADEs [51].

Phenotyping

Genetic diseases can have different clinical presentations in individuals despite seemingly identical mutations in the DNA. The association of genotypes and phenotypes contains more information for the understanding of disease progression than genotypes or phenotypes alone [20, 52]. However, an important challenge to determine these associations is related to obtaining phenotypic information because of poor standardization across sites. Recent initiatives in this area include the Electronic Medical Records and Genomics (eMERGE) Network, which maps phenotype information from EHRs to standard vocabularies and associates these phenotypes with geno-

types and phenome-wide association scans (PheWAS), to map clinical data to single nucleotide polymorphisms (SNPs) [53, 54]. A recent PheWAS study by Pendergrass *et al.* performed an analysis of over 70,000 participants in the Population Architecture using Genomics and Epidemiology (PAGE) Network. The authors assessed associations between SNPs and phenotype variables including lab values and biomarkers and found 33 potentially new associations [52].

Identification of specific phenotypes within a broad disease category has often utilized data collected in clinical trials. By utilizing EHRs, specific phenotypes within disease categories have been identified in conditions such as diabetes [55] and neuropsychiatric disorders [56, 57]. The identified phenotypes can be assessed to determine their association with individual risk factors for disease, efficacy of different treatment options, and outcomes.

Natural Language Processing

Natural language processing (NLP) is a core technique employed in biomedical informatics. A large amount of EHR data, such as discharge summaries, is unstructured; therefore, employing NLP techniques to extract and structure this narrative text information is an important step in many EHR data secondary use studies. Natural language processing can also help address some aspects of incomplete data by augmenting the amount of computable data. Several recent NLP applications focused on information extraction and identification of ADEs from EHR data.

Information extraction (IE) is a traditional area in which NLP is used to identify and classify name entities (e.g. concepts and assertions) and relations from narrative text. The focus is shifting from simple name entity recognition (e.g. HIPAA defined fields for data de-identification), and from a single source such as a radiology report to more complex concepts and relations from multiple sources [58-60]. The complex concepts include medical problems [61], medications [62], tests and treatments [61], assertion status [61, 63, 64], disease phenotypes [64, 65]. Since these concepts are not easily extractable by rules, machine-learning-based methods and hybrid methods are widely adopted [61, 63, 66].

Relation identification is a promising field in IE, although there are still many obstacles. Two frequently studied topics are co-reference identification and temporal relation extraction. Co-reference resolution aims to recognize two mentions that refer to the same entity in a sentence or across sentences [67]. Current co-reference resolution studies use rule-based [68, 69], machine learning [35, 70, 71] and hybrid systems [72] to identify noun phrases, including person, pronoun, and concepts such as medical tests. The performance of these systems varies depending on the data quality [73].

Temporal relation extraction is a research focus in IE because it is useful in identification of complications, patient outcome predictions, and ADE detection. Warner *et al.* introduced a method to identify hospital-acquired complications using temporal phenomes (i.e. a selected set of phenotypes) [74]. Temporal status was also utilized in the abstraction of Emergency Department (ED) CT imaging reports [75].

Data Application and Integration

EHRs have been utilized to retrospectively assess treatment effectiveness in real-world settings, quality of care and cost (76]. Prospectively, clinical decision support systems (CDSSs) use information in the EHR system and specific algorithms to guide health care providers' decision-making. Decision support tools have existed for many years and continue to evolve. Personal monitoring devices and social media may eventually be integrated into EHRs and CDSSs to enhance predictions [3]. Challenges include privacy concerns, low adoption of data standards, and poor interoperability [16, 18].

Clinical Decision Support

Clinicians have human limitations in the ability to multi-task, reason, and comprehend information; therefore, CDSSs can play an important role in big data processing [12]. CDSSs are not new: for decades there have been high expectations that CDSSs will improve clinical documentation, increase guideline adherence, help predict outcomes, and assist clinicians in making diagnoses and preventing errors [77]. Current approaches employ rule-based systems, heuristics, fuzzy logic, artificial neural networks, Bayesian networks, and other machine-learning techniques [12]. The field continues to develop and some healthcare systems are partnering with industry to build CDSSs [78]. Recent work includes an NLP approach to assess adherence to treatment protocols and guidelines [79, 80], automated medication dosing reminders in the operating room [81], screening for disease [82, 83], prediction of hospital readmission [84, 85], creation of a life-expectancy index for hospitalized elderly patients [86], determination of early indicators of patient deterioration [87], and guided urinary tract infection treatment [88]. Clinical decision support is not fully automated and issues such as human error in algorithm design can lead to underperformance [89]. Other limitations have been lack of adoption, system accuracy, integration with workflow, and lack of natural language processing tools [12]. Eventually, the hope is for a CDSS that includes learning algorithms to make recommendations based on previous successful treatments [3]. Some prototype systems are in development [3, 90-92].

Personal Monitoring and Social Media

Mobile health technology (mHealth), i.e., applying mobile communication technology to healthcare and patient wellness, has become increasingly popular. The potential is enormous as there are an estimated 3.2 billion unique mobile users worldwide with over 30,000 available healthcare apps (not all regulated) [93]. Interventions have been targeted directly to patients, such as apps for smoking cessation and weight loss [94, 95]. Real-time aggregated data from individual patients in web sites such as the Quantified Self can address work productivity, posture, blood glucose levels, and can be leveraged for health outcomes [6, 96]. These data could be integrated into the Personalized Health Record or EHRs, and would contribute to building an automated system to identify at-risk populations and send automated health messages to patients.

mHealth uses sensors, global positioning satellite receivers, and accelerometers that

continuously monitor data [97]. Examples of recent real-time mobile monitoring of patients include projects that estimated air pollution exposure [98] and assessed physiological responses to changes in position [98, 99]. Once risk factors are determined for various conditions, the hope is to integrate these factors into the EHR to assist physicians in identifying at-risk patients and build predictive models [100].

Besides analysis of personal information, surveillance of population databases can be utilized to alert physicians of potential at-risk patients, behaviors and outcomes [77]. One example is the New York City Department of Health and Mental Hygiene's (NYCDH) project in which an infrastructure was built so that the healthcare department could gather information on specific neighborhoods and alert physicians to conditions such as obesity. Another example is the University of Wisconsin's UW eHealth-PHINEX program which developed a framework to map asthma and diabetes data from the EHR to socioeconomic information found in a public health data exchange to help delineate community patterns of disease [101]. In addition, Lin *et al.* linked cancer registry data and discharge summary data to examine treatment outcomes and disparities, building on previous work in this field [102].

Web-based information is another potential source for patient information that can be linked to EHR data. Resources that are increasingly utilized include social networks, web site visit history, blogs, forums, user-generated ratings of items, and evaluation of links to previously viewed web sites. Patterns can be used for tailoring patient education and for recruiting for clinical trials [103]. Vickey *et al.* processed over two million fitness-related Twitter posts to demonstrate that tracking social media markers can provide insight to personal health behavior [104]. In addition, Myslin *et al.* analyzed Twitter posts to understand user sentiment of newer tobacco products versus older delivery methods, which can help guide public health intervention approaches [105]. Once privacy and interoperability issues are addressed, this information could be potentially linked or integrated into the EHR to assist caregivers in providing personalized care to their patients.

Privacy and Security

Some authors report that EHRs improve healthcare quality and efficiency and further the relationship between patients and healthcare providers [106, 107]. However, privacy concerns must be addressed before employing EHR data for clinical research, and other secondary use purposes [108].

Several privacy protection mechanisms have been proposed to enable data sharing with reduced information loss [109-111]. In some patient data privacy studies, each patient record is represented in a table. For each patient, the table has the value that corresponds to attributes such as name, date of birth, diagnosis, etc. The privacy mechanisms we discuss here are based on this representation. The outdated *k*-anonymity method either generalized or suppressed attributes to guarantee each row was identical to at least *k* other rows [109]. The *l*-diversity algorithm was designed to solve limitations of *k*-anonymity [112]. In this approach, equally outdated, a quasi-identifier equivalence class, which contained rows of records sharing identical values of non-sensitive attributes, showed diverse values in each sensitive attribute. An updated *k*-anonymity method developed for EHR data adopts ideas from *l*-diversity to keep sensitive attributes diverse [113]. A new framework for protecting privacy has been more utilized lately. *Differential privacy* algorithms can provide strong privacy guarantees but there are still concerns about the resulting data utility when these algorithms add much noise to data [111]. One example of the application of differential privacy is SHARE, a system designed to aggregate statistics on data found in health information systems [114]. A recent method introduces wavelet transformations into differential privacy to improve data utility by adding noise after transformations; this can be potentially useful for publishing aggregated clinical information [115].

In practice, in recent years, an internationally discussed topic in EHR privacy is access control. With a belief that encouraging patients to visit their own medical records will improve healthcare quality, a privacy framework named "Points to Consider"(P2C) was designed to guide patients to access EHR data under control policies and assist in developing EHR query tools [116]. To efficiently create access control policies, methods based on social network analysis were used to identify stable interaction pairs and groups from EHR log data and provide access policy suggestions [117]. In addition to role-based access control for EHRs, audit systems to detect unauthorized and suspicious accesses have already been successfully implemented, such as the machine-learning model to detect suspicious accesses to EHR data [118].

Privacy is an important consideration in collaborative research. "De-identification" (i.e., removal of certain identifiers) of EHR data can facilitate certain types of data sharing for research. Ferrández *et al.* developed an automated text de-identification system for Veterans Health Administration (VHA) clinical documents, using a hybrid approach of rule-based and machine learning methods to improve upon current techniques [119]. Deleger *et al.*'s work shows that NLP-based de-identification tools perform at levels comparable to human annotators [59], but those levels are unfortunately not yet ideal.

In order to process the vast amount of biomedical data available, researchers and institutions need HIPAA-compliant computational environments to host confidential EHR data. This is time-consuming and expensive to set up. One solution is cloud computing, in which users can "lease" HIPAA-compliant computer hardware and software over the Internet and remain adherent to privacy rules [120]. Some public cloud-based EHR systems are on the market. A recent access control model designed for cloud-based EHR systems grants users different levels of permission using hierarchical key management [121]. HIPAA-compliant private clouds have also been developed to host clinical and translational research data; this type of cloud also has the potential of hosting EHR systems [122, 123].

Limitations

Due to the nature of a yearbook review, the scope and depth of this article is limited. We focused on recent publications knowing that we are unable to detail all historical aspects

of the field. Some terms like "big data" and others chosen as keywords may not be widely accepted and are not identified by the National Library of Medicine as MeSH terms. Nevertheless, in order to capture the most recent developments, we felt these non-MeSH keywords were important. Finally, we structured our review on article topics, but the topic selection was subjective. This resulted in three major sections of *data mining*, *data application and integration*, and *privacy and security*. However, there are overlaps among these sections and some other potentially relevant categories may not have been represented.

Conclusion

The increasing adoption of EHR systems worldwide makes it possible to capture massive amounts of clinical data. The next step is to truly transform these big healthcare data into knowledge. New data mining and natural language processing techniques are key components of analytics for EHR data. Critical for future progress are security and privacy mechanisms that facilitate EHR and other healthcare data sharing. Access control methods and security measures allow EHR systems to protect sensitive patient information. The development and application of big data analysis methods on EHRs may help create a continually learning EHR ecosystem. In the future, it may be possible to combine data from the EHR with other sources such as social media, environmental information, and gene sequencing data. Additionally, with the globalization of biomedical research and healthcare, it will be important to develop means to harmonize and compute with big data originating from different countries in a way that respects national and international policy and legislation as well as patient preferences.

Acknowledgements

The authors were funded in part by NIH grants T15LM011271 (MKR), U54HL108460 (WW, LOM), UL1TR0001000 (LOM), and D43TW007015 (LOM). We thank Lisa Naidoo and Mary Wickline for assistance with article retrieval.

References

1. Laney D. 3D Data Management: Controlling Data Volume, Velocity, and Variety. META Group; 2001.
2. Ward JS, Barker A. Undefined By Data: A Survey of Big Data Definitions. arXiv:1309.5821; 2013.
3. Murdoch TB, Detsky AS. The inevitable application of big data to health care. Jama. 2013;309(13):1351-2.
4. Index for Excerpts from the American Recovery and Reinvestment Act of 2009. Health Information Technology (HITECH) Act 2009. p. 112-64.
5. Charles DK J, Patel V, Furukawa M. Adoption of Electronic Health Record Systems among U.S. Non-federal Acute Care Hospitals: 2008-2012; 2013. http://www.healthit.gov/sites/default/files/oncdatabrief9final.pdf. Accessibility verified April 20, 2014.
6. Shah NH. Translational bioinformatics embraces big data. Yearb Med Inform 2012;7(1):130-4.
7. Heinze O, Birkle M, Koster L, Bergh B. Architecture of a consent management suite and integration into IHE-based Regional Health Information Networks. BMC Med Inform Decis Mak 2011;11:58.
8. Tejero A, de la Torre I. Advances and current state of the security and privacy in electronic health records: survey from a social perspective. J Med Syst 2012;36(5):3019-27.
9. Mense A, Hoheiser-Pfortner F, Schmid M, Wahl H. Concepts for a standard based cross-organisational information security management system in the context of a nationwide EHR. Stud Health Technol Inform 2013;192:548-52.
10. Faxvaag A, Johansen TS, Heimly V, Melby L, Grimsmo A. Healthcare professionals' experiences with EHR-system access control mechanisms. Stud Health Technol Inform 2011;169:601-5.
11. Hripcsak G, Albers DJ. Next-generation phenotyping of electronic health records. J Am Med Inform Assoc 2013;20(1):117-21.
12. Wagholikar KB, Sundararajan V, Deshpande AW. Modeling paradigms for medical diagnostic decision support: a survey and future directions. J Med Syst 2012;36(5):3029-49.
13. Rowley J. The wisdom hierarchy: representations of the DIKW hierarchy. J Inf Sci 2007;33(2):163-80.
14. Sullivan P, Goldmann D. The promise of comparative effectiveness research. JAMA 2011;305(4):400-1.
15. Fernandes L, O'Connor M, Weaver V. Big data, bigger outcomes: Healthcare is embracing the big data movement, hoping to revolutionize HIM by distilling vast collection of data for specific analysis. J AHIMA 2012;83(10):38-43; quiz 44.
16. Harper EM. The economic value of health care data. Nurs Adm Q 2013;37(2):105-8.
17. Colpas P. Integration, analytics key to next-generation EMRs. Industry experts discuss the year ahead in EMRs/EHRs. Health Manag Technol 2013;34(1):6-8, 10-1.
18. Weinstock M. 2013 most wired. Hosp Health Netw 2013;87(7):26-37.
19. Hoffman S, Podgurski A. Big bad data: law, public health, and biomedical databases. J Law Med Ethics 2013;41 Suppl 1:56-60.
20. Chute CG, Ullman-Cullere M, Wood GM, Lin SM, He M, Pathak J. Some experiences and opportunities for big data in translational research. Genet Med 2013;15(10):802-9.
21. Ackerman MJ. Big data. J Med Pract Manage 2012;28(2):153-4.
22. Liyanage H, Liaw ST, de Lusignan S. Accelerating the development of an information ecosystem in health care, by stimulating the growth of safe intermediate processing of health information (IPHI). Inform Prim Care 2012;20(2):81-6.
23. Bonney S. HIM's role in managing big data: Turning data collected by an EHR into information. J AHIMA 2013;84(9):62-4.
24. Leventhal R. Trend: big data. Big data analytics: from volume to value. Healthc Inform 2013;30(2):12, 4.
25. Glaser J, Overhage JM. The role of healthcare IT: becoming a learning organization. Healthc Financ Manage 2013;67(2):56-62, 4.
26. Institute of Medicine. Best Care at Lower Cost: The Path to Continuously Learning Health Care in America; 2012.
27. Hood L, Flores M. A personal view on systems medicine and the emergence of proactive P4 medicine: predictive, preventive, personalized and participatory. N Biotechnol 2012;29(6):613-24.
28. Phan JH, Quo CF, Cheng C, Wang MD. Multiscale integration of -omic, imaging, and clinical data in biomedical informatics. IEEE Rev Biomed Eng 2012;5:74-87.
29. Weiskopf NG, Hripcsak G, Swaminathan S, Weng C. Defining and measuring completeness of electronic health records for secondary use. J Biomed Inform 2013;46(5):830-6.
30. Conn J. Pairing up. Early adopters of big data seek advisers, partners. Mod Healthc 2013;43(24):8-9.
31. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. Nat Rev Genet. 2012;13(6):395-405.
32. Barr P. Analyze this: health systems get help with big data. Hosp Health Netw 2013;87(6):16.
33. Bellazzi R, Zupan B. Predictive data mining in clinical medicine: current issues and guidelines. Int J Med Inform 2008;77(2):81-97.
34. Lependu P, Iyer SV, Fairon C, Shah NH. Annotation Analysis for Testing Drug Safety Signals using Unstructured Clinical Notes. J Biomed Semantics 2012;3 Suppl 1:S5.
35. Rink B, Roberts K, Harabagiu SM. A supervised framework for resolving coreference in clinical records. J Am Med Inform Assoc. 2012;19(5):875-82.
36. Patrick JD, Nguyen DH, Wang Y, Li M. A knowledge discovery and reuse pipeline for information extraction in clinical notes. J Am Med Inform Assoc. 2011;18(5):574-9.
37. Hrovat G, Stiglic G, Kokol P, Ojstersek M. Contrasting temporal trend discovery for large healthcare databases. Comput Methods Programs Biomed 2014;113(1):251-7.
38. Sun J, Hu J, Luo D, Markatou M, Wang F, Edabollahi S, et al. Combining knowledge and data driven insights for identifying risk factors using electronic health records. AMIA Annu Symp Proc 2012;2012:901-10.
39. Bates DW, Spell N, Cullen DJ, Burdick E, Laird N,

Petersen LA, et al. The costs of adverse drug events in hospitalized patients. Adverse Drug Events Prevention Study Group. JAMA 1997;277(4):307-11.
40. Warrer P, Hansen EH, Juhl-Jensen L, Aagaard L. Using text-mining techniques in electronic patient records to identify ADRs from medicine use. Br J Clin Pharmacol 2012;73(5):674-84.
41. Coloma PM, Trifiro G, Patadia V, Sturkenboom M. Postmarketing safety surveillance : where does signal detection using electronic healthcare records fit into the big picture? Drug Saf 2013;36(3):183-97.
42. Haerian K, Varn D, Vaidya S, Ena L, Chase HS, Friedman C. Detection of pharmacovigilance-related adverse events using electronic health records and automated methods. Clin Pharmacol Ther.2012;92(2):228-34.
43. Chazard E, Ficheur G, Bernonville S, Luyckx M, Beuscart R. Data mining to generate adverse drug events detection rules. IEEE Trans Inf Technol Biomed 2011;15(6):823-30.
44. Ji Y, Ying H, Dews P, Mansour A, Tran J, Miller RE, et al. A potential causal association mining algorithm for screening adverse drug reactions in postmarketing surveillance. IEEE Trans Inf Technol Biomed 2011;15(3):428-37.
45. Sohn S, Kocher JP, Chute CG, Savova GK. Drug side effect extraction from clinical narratives of psychiatry and psychology patients. J Am Med Inform Assoc 2011;18 Suppl 1:i144-9.
46. Vilar S, Harpaz R, Santana L, Uriarte E, Friedman C. Enhancing adverse drug event detection in electronic health records using molecular structure similarity: application to pancreatitis. PLoS One 2012;7(7):e41471.
47. Eriksson R, Jensen PB, Frankild S, Jensen LJ, Brunak S. Dictionary construction and identification of possible adverse drug events in Danish clinical narrative text. J Am Med Inform Assoc 2013;20(5):947-53.
48. Iyer SV, Harpaz R, Lependu P, Bauer-Mehren A, Shah NH. Mining clinical text for signals of adverse drug-drug interactions. J Am Med Inform Assoc 2013.
49. Yoon D, Park MY, Choi NK, Park BJ, Kim JH, Park RW. Detection of adverse drug reaction signals using an electronic health records database: Comparison of the Laboratory Extreme Abnormality Ratio (CLEAR) algorithm. Clin Pharmacol Ther 2012;91(3):467-74.
50. Harpaz R, Vilar S, Dumouchel W, Salmasian H, Haerian K, Shah NH, et al. Combing signals from spontaneous reports and electronic health records for detection of adverse drug reactions. J Am Med Inform Assoc 2013;20(3):413-9.
51. Oliveira JL, Lopes P, Nunes T, Campos D, Boyer S, Ahlberg E, et al. The EU-ADR Web Platform: delivering advanced pharmacovigilance tools. Pharmacoepidemiol Drug Saf 2013;22(5):459-67.
52. Pendergrass SA, Brown-Gentry K, Dudek S, Frase A, Torstenson ES, Goodloe R, et al. Phenome-wide association study (PheWAS) for detection of pleiotropy within the Population Architecture using Genomics and Epidemiology (PAGE) Network. PLoS Genet 2013;9(1):e1003087.
53. Denny JC, Ritchie MD, Basford MA, Pulley JM, Bastarache L, Brown-Gentry K, et al. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. Bioinformatics 2010;26(9):1205-10.
54. Pathak J, Wang J, Kashyap S, Basford M, Li R, Masys DR, et al. Mapping clinical phenotype data elements to standardized metadata repositories and controlled terminologies: the eMERGE Network experience. J Am Med Inform Assoc 2011;18(4):376-86.
55. Richesson RL, Rusincovitch SA, Wixted D, Batch BC, Feinglos MN, Miranda ML, et al. A comparison of phenotype definitions for diabetes mellitus. J Am Med Inform Assoc 2013.
56. Roque FS, Jensen PB, Schmock H, Dalgaard M, Andreatta M, Hansen T, et al. Using electronic patient records to discover disease correlations and stratify patient cohorts. PLoS Comput Biol 2011;7(8):e1002141.
57. Lyalina S, Percha B, Lependu P, Iyer SV, Altman RB, Shah NH. Identifying phenotypic signatures of neuropsychiatric disorders from electronic medical records. J Am Med Inform Assoc 2013.
58. Ferrandez O, South BR, Shen S, Friedlin FJ, Samore MH, Meystre SM. Evaluating current automatic de-identification methods with Veteran's health administration clinical documents. BMC Med Res Methodol 2012;12:109.
59. Deleger L, Molnar K, Savova G, Xia F, Lingren T, Li Q, et al. Large-scale evaluation of automated clinical note de-identification and its impact on information extraction. J Am Med Inform Assoc 2013;20(1):84-94.
60. Pantazos K, Lauesen S, Lippert S. De-identifying an EHR database - anonymity, correctness and readability of the medical record. Stud Health Technol Inform 2011;169:862-6.
61. Jiang M, Chen Y, Liu M, Rosenbloom ST, Mani S, Denny JC, et al. A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries. J Am Med Inform Assoc 2011;18(5):601-6.
62. Xu H, Jiang M, Oetjens M, Bowton EA, Ramirez AH, Jeff JM, et al. Facilitating pharmacogenetic studies using electronic health records and natural-language processing: a case study of warfarin. J Am Med Inform Assoc 2011;18(4):387-91.
63. Clark C, Aberdeen J, Coarr M, Tresner-Kirsch D, Wellner B, Yeh A, et al. MITRE system for clinical assertion status classification. J Am Med Inform Assoc 2011;18(5):563-7.
64. Bejan CA, Xia F, Vanderwende L, Wurfel MM, Yetisgen-Yildiz M. Pneumonia identification using statistical feature selection. J Am Med Inform Assoc 2012;19(5):817-23.
65. Schuemie MJ, Sen E, t Jong GW, van Soest EM, Sturkenboom MC, Kors JA. Automating classification of free-text electronic health records for epidemiological studies. Pharmacoepidemiol Drug Saf 2012;21(6):651-8.
66. Minard AL, Ligozat AL, Ben Abacha A, Bernhard D, Cartoni B, Deleger L, et al. Hybrid methods for improving information access in clinical documents: concept, assertion, and relation identification. J Am Med Inform Assoc 2011;18(5):588-93.
67. Savova GK, Chapman WW, Zheng J, Crowley RS. Anaphoric relations in the clinical narrative: corpus creation. J Am Med Inform Assoc 2011;18(4):459-65.
68. Dai HJ, Chen CY, Wu CY, Lai PT, Tsai RT, Hsu WL. Coreference resolution of medical concepts in discharge summaries by exploiting contextual information. J Am Med Inform Assoc 2012;19(5):888-96.
69. Jindal P, Roth D. Using domain knowledge and domain-inspired discourse model for coreference resolution for clinical narratives. J Am Med Inform Assoc 2013;20(2):356-62.
70. Xu Y, Liu J, Wu J, Wang Y, Tu Z, Sun JT, et al. A classification approach to coreference in discharge summaries: 2011 i2b2 challenge. J Am Med Inform Assoc 2012;19(5):897-905.
71. Ware H, Mullett CJ, Jagannathan V, El-Rawas O. Machine learning-based coreference resolution of concepts in clinical documents. J Am Med Inform Assoc 2012;19(5):883-7.
72. Jonnalagadda SR, Li D, Sohn S, Wu ST, Wagholikar K, Torii M, et al. Coreference analysis in clinical notes: a multi-pass sieve with alternate anaphora resolution modules. J Am Med Inform Assoc 2012;19(5):867-74.
73. Uzuner O, Bodnari A, Shen S, Forbush T, Pestian J, South BR. Evaluating the state of the art in coreference resolution for electronic medical records. J Am Med Inform Assoc 2012;19(5):786-91.
74. Warner JL, Zollanvari A, Ding Q, Zhang P, Snyder GM, Alterovitz G. Temporal phenome analysis of a large electronic health record cohort enables identification of hospital-acquired complications. J Am Med Inform Assoc 2013.
75. Yadav K, Sarioglu E, Smith M, Choi HA. Automated outcome classification of emergency department computed tomography imaging reports. Acad Emerg Med 2013;20(8):848-54.
76. Kahn MG, Batson D, Schilling LM. Data model considerations for clinical effectiveness researchers. Med Care 2012;50 Suppl:S60-7.
77. Rothman B, Leonard JC, Vigoda MM. Future of electronic health records: implications for decision support. Mt Sinai J Med 2012;79(6):757-68.
78. Memorial Sloan-Kettering Center. Memorial Sloan-Kettering's Collaboration with IBM Watson Featured on CBS This Morning. 2013. Http://www.mskcc.org/blog/msk-s-collaboration-ibm-watson-featured-cbs-morning. Accessibility verified April 20, 2014.
79. Mishra NK, Son RY, Arnzen JJ. Towards automatic diabetes case detection and ABCS protocol compliance assessment. Clin Med Res 2012;10(3):106-21.
80. Klann JG, Anand V, Downs SM. Patient-tailored prioritization for a pediatric care decision support system through machine learning. J Am Med Inform Assoc 2013.
81. Nair BG, Newman SF, Peterson GN, Schwid HA. Automated electronic reminders to improve redosing of antibiotics during surgical cases: comparison of two approaches. Surg Infect (Larchmt) 2011;12(1):57-63.
82. Wagholikar KB, MacLaughlin KL, Henry MR, Greenes RA, Hankey RA, Liu H, et al. Clinical decision support with automated text processing for cervical cancer screening. J Am Med Inform Assoc 2012;19(5):833-9.
83. Murphy DR, Laxmisan A, Reis BA, Thomas EJ, Esquivel A, Forjuoh SN, et al. Electronic health record-based triggers to detect potential delays in cancer diagnosis. BMJ Qual Saf 2013.

84. Bradley EH, Yakusheva O, Horwitz LI, Sipsma H, Fletcher J. Identifying patients at increased risk for unplanned readmission. Med Care 2013;51(9):761-6.
85. Cholleti S, Post A, Gao J, Lin X, Bornstein W, Cantrell D, et al. Leveraging derived data elements in data analytic models for understanding and predicting hospital readmissions. AMIA Annu Symp Proc 2012;2012:103-11.
86. Mathias JS, Agrawal A, Feinglass J, Cooper AJ, Baker DW, Choudhary A. Development of a 5 year life expectancy index in older adults using predictive mining of electronic health record data. J Am Med Inform Assoc 2013;20(e1):e118-24.
87. Collins SA, Cato K, Albers D, Scott K, Stetson PD, Bakken S, et al. Relationship between nursing documentation and patients' mortality. Am J Crit Care 2013;22(4):306-13.
88. Jackson HA, Cashy J, Frieder O, Schaeffer AJ. Data mining derived treatment algorithms from the electronic medical record improve theoretical empirical therapy for outpatient urinary tract infections. J Urol 2011;186(6):2257-62.
89. Wagholikar KB, MacLaughlin KL, Kastner TM, Casey PM, Henry M, Greenes RA, et al. Formative evaluation of the accuracy of a clinical decision support system for cervical cancer screening. J Am Med Inform Assoc 2013;20(4):749-57.
90. van den Branden M, Wiratunga N, Burton D, Craw S. Integrating case-based reasoning with an electronic patient record system. Artif Intell Med 2011;51(2):117-23.
91. Gotz D, Stavropoulos H, Sun J, Wang F. ICDA: a platform for Intelligent Care Delivery Analytics. AMIA Annu Symp Proc 2012;2012:264-73.
92. Wang HQ, Li JS, Zhang YF, Suzuki M, Araki K. Creating personalised clinical pathways by semantic interoperability with electronic health records. Artif Intell Med 2013;58(2):81-9.
93. Steinhubl SR, Muse ED, Topol EJ. Can Mobile Health Technologies Transform Health Care? JAMA 2013.
94. Steinberg DM, Levine EL, Askew S, Foley P, Bennett GG. Daily text messaging for weight control among racial and ethnic minority women: randomized controlled pilot study. J Med Internet Res 2013;15(11):e244.
95. Shi HJ, Jiang XX, Yu CY, Zhang Y. Use of mobile phone text messaging to deliver an individualized smoking behaviour intervention in Chinese adolescents. J Telemed Telecare 2013;19(5):282-7.
96. Quantified Self. Http://quantifiedself.com. Accessibility verified April 20, 2014.
97. Dobkin BH. Wearable motion sensors to continuously measure real-world physical activities. Curr Opin Neurol 2013;26(6):602-8.
98. de Nazelle A, Seto E, Donaire-Gonzalez D, Mendez M, Matamala J, Nieuwenhuijsen MJ, et al. Improving estimates of air pollution exposure through ubiquitous sensing technologies. Environ Pollut 2013;176:92-9.
99. Jovanov E, Milosevic M, Milenkovic A. A mobile system for assessment of physiological response to posture transitions. Conf Proc IEEE Eng Med Biol Soc 2013;2013:7205-8.
100. Dobkin BH, Dorsch A. The promise of mHealth: daily activity monitoring and outcome assessments by wearable sensors. Neurorehabil Neural Repair 2011;25(9):788-98.
101. Guilbert TW, Arndt B, Temte J, Adams A, Buckingham W, Tandias A, et al. The theory and application of UW ehealth-PHINEX, a clinical electronic health record-public health information exchange. WMJ : Official publication of the State Medical Society of Wisconsin 2012;111(3):124-33.
102. Lin G, Ma J, Zhang L, Qu M. Linking cancer registry and hospital discharge data for treatment surveillance. Health Informatics J 2013;19(2):127-36.
103. Fernandez-Luque L, Karlsen R, Bonander J. Review of extracting information from the Social Web for health personalization. J Med Internet Res 2011;13(1):e15.
104. Vickey TA, Ginis KM, Dabrowski M. Twitter classification model: the ABC of two million fitness tweets. Transl Behav Med 2013;3(3):304-11.
105. Myslin M, Zhu SH, Chapman W, Conway M. Using twitter to examine smoking behavior and perceptions of emerging tobacco products. J Med Internet Res 2013;15(8):e174.
106. Agno CF, Guo KL. Electronic health systems: challenges faced by hospital-based providers. Health Care Manag (Frederick) 2013;32(3):246-52.
107. Nazi KM, Hogan TP, McInnes DK, Woods SS, Graham G. Evaluating patient access to Electronic Health Records: results from a survey of veterans. Med Care 2013;51(3 Suppl 1):S52-6.
108. Dwork C, Pottenger R. Toward practicing privacy. J Am Med Inform Assoc 2013;20(1):102-8.
109. Sweeney L. k-anonymity: A model for protecting privacy. Int J Uncertain Fuzz 2002;10(5):557-70.
110. Machanavajjhala AG, J.; Kifer, D. I-diversity: Privacy Beyond k-anonymity. ACM Trans Knowl Discov Data 2007.
111. Dwork C. Differential privacy. Automata, Languages and Programming, Pt 2. 2006;4052:1-12.
112. Machanavajjhala A, Kifer D, Gehrke J, Venkitasubramaniam M. l-diversity: Privacy beyond k-anonymity. ACM Trans Knowl Discov Data 2007;1(1):3.
113. Yoo S, Shin M, Lee D. An Approach to Reducing Information Loss and Achieving Diversity of Sensitive Attributes in k-anonymity Methods. Interact J Med Res 2012;1(2):e14.
114. Gardner J, Xiong L, Xiao Y, Gao J, Post AR, Jiang X, et al. SHARE: system design and case studies for statistical health information release. J Am Med Inform Assoc 2013;20(1):109-16.
115. Xiao XW, G.; Gehrke, J. Differential Privacy Via Wavelet Transforms. IEEE Trans Knowl Data Eng 2011;23(8):1200-14.
116. Meslin EM, Alpert SA, Carroll AE, Odell JD, Tierney WM, Schwartz PH. Giving patients granular control of personal health information: Using an ethics 'Points to Consider' to inform informatics system designers. Int J Med Inform 2013;82(12):1136-43.
117. Malin B, Nyemba S, Paulett J. Learning relational policies from electronic health record access logs. J Biomed Inform 2011;44(2):333-42.
118. Boxwala AA, Kim J, Grillo JM, Ohno-Machado L. Using statistical and machine learning to help institutions detect suspicious access to electronic health records. J Am Med Inform Assoc 2011;18(4):498-505.
119. Ferrandez O, South BR, Shen S, Friedlin FJ, Samore MH, Meystre SM. BoB, a best-of-breed automated text de-identification system for VHA clinical documents. J Am Med Inform Assoc 2013;20(1):77-83.
120. Regola N, Chawla NV. Storing and using health data in a virtual private cloud. J Med Internet Res 2013;15(3):e63.
121. Alabdulatif A, Khalil I, Mai V. Protection of electronic health records (EHRs) in cloud. Conf Proc IEEE Eng Med Biol Soc 2013;2013:4191-4.
122. Ohno-Machado L, Bafna V, Boxwala AA, Chapman BE, Chapman WW, Chaudhuri K, et al. iDASH: integrating data for analysis, anonymization, and sharing. J Am Med Inform Assoc 2012;19(2):196-201.
123. Ohno-Machado L. To share or not to share: that is not the question. Sci Transl Med 2012;4(165).

Correspondence to:
Lucila Ohno-Machado
Division of Biomedical Informatics
9500 Gilman Drive, MC 0505
La Jolla, California, 92037-0505, USA
Tel: +1 858 822 4931
E-mail: machado@ucsd.edu