“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

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

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.
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.

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 expedient [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-learn

ing-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 contrasted, 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), UL1TR001000 (LOM), and D43TW007015 (LOM). We thank Lisa Naidoo and Mary Wickline for assistance with article retrieval.

References

13. Rowley J. The wisdom hierarchy: represen -

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. JAHIMA 2012;83(10):38-43; quiz 44.
17. Colpas P, Integration, analytics key to next-gen -
24. Leventhal R. Trend: big data. Big data ana -
39. Bates DW, Spell N, Cullen DJ, Burdick E, Laird N,
Denny JC, Ritchie MD, Basford MA, Pulley JM, Oliveira JL, Lopes P, Nunes T, Campos D, Boyer


