Grappling with the Future Use of Big Data for Translational Medicine and Clinical Care

S. Murphy[1, 2, 4], V. Castro[2], K. Mandl[3, 4]
[1] Massachusetts General Hospital Laboratory of Computer Science, Boston, MA, USA
[2] Partners Healthcare Research Information Science and Computing, Somerville, MA, USA
[3] Boston Children's Hospital Computational Health Informatics Program, Boston, MA, USA
[4] Harvard Medical School Department of Biomedical Informatics, Boston, MA, USA

Summary

Objectives: Although patients may have a wealth of imaging, genomic, monitoring, and personal device data, it has yet to be fully integrated into clinical care.

Methods: We identify three reasons for the lack of integration. The first is that "Big Data" is poorly managed by most Electronic Medical Record Systems (EMRS). The data is mostly available on "cloud-native" platforms that are outside the scope of most EMRs, and even checking if such data is available on a patient often must be done outside the EMRS. The second reason is that extracting features from the Big Data that are relevant to healthcare often requires complex machine learning algorithms, such as determining if a genomic variant is protein-altering. The third reason is that applications that present Big Data need to be modified constantly to reflect the current state of knowledge, such as instructing when to order a new set of genomic tests. In some cases, applications need to be updated nightly.

Results: A new architecture for EMRS is evolving which could unite Big Data, machine learning, and clinical care through a microservice-based architecture which can host applications focused on quite specific aspects of clinical care, such as managing cancer immunotherapy.

Conclusion: Informatics innovation, medical research, and clinical care go hand in hand as we look to infuse science-based practice into healthcare. Innovative methods will lead to a new ecosystem of applications (Apps) interacting with healthcare providers to fulfill a promise that is still to be determined.

Keywords

Big Data, healthcare; decision support systems, clinical; genomics

Yearb Med Inform 2017: 96-102
http://dx.doi.org/10.15265/IY-2017-020
Published online August 18, 2017

Introduction

If we are to infuse science-based clinical practice into healthcare, we will need better technology platforms than the current Electronic Medical Record Systems (EMRS), which are not built to support scientific innovation. In different countries, the functionality and overall goals of EMRS can vary, but in the United States (US) the primary purpose of EMRS is to bill for the patient visit [1]. The Epic EMRS, dominant in many US Academic Healthcare Centers and known for its strength of efficient patient billing [2], is mostly closed to outside software additions as part of the quest to develop the most efficient system possible. Other EMRS in the US are perhaps best summarized as vehicles for inter-clinician communication [3]. This communication is mostly facilitated through the exchange of narrative text, and therefore the information which is not coded for billing purposes is left in the unstructured notes exchanged between clinicians [4]. This situation has greatly reduced the appeal of using Big Data for Healthcare in the US. We will largely focus on the loss of this appeal in this report.

At odds with what is presented above as an efficient Healthcare Enterprise Information Technology (IT) management system is data entering the system through channels outside the Healthcare Enterprise, data such as personal health data, patient surveys, patient self-reported outcomes, and patient-initiated genomic testing. Even inside the Healthcare Enterprise, the billing and communication-oriented EMRS is surprisingly disjointed in its ability to handle imaging, waveform, genomic, and continuous cardiac and cephalic monitoring. Much of these data fall under the label of "Big Data" [5].

One defining aspect of "Big Data" is that the data tends to be organized in a fundamentally different way that cannot fit easily and directly into hierarchical and relational databases. Rather, much of the data arrives as (sometimes very large) data objects. In S3 object storing systems (such as https://aws.amazon.com/s3), there is often only minimal metadata regarding what information exists in the data. These data "objects" have their own internal structure and often require a great deal of computation to extract healthcare-relevant features. Support exists for many tools within the cloud-native platform that enable the analysis of the healthcare "Big Data" such as Hadoop, CouchDB, MongoDB, Matlab, R, SAS, and Docker, but the paradigm of computing on data objects rather than examining the data through Structured Query Languages requires new skills and infrastructures which are different from the database-oriented approaches currently employed in most EMRS.

In addition to barriers regarding training people how to use the new tools, the tools to process Big Data and extract healthcare-relevant features must often reside on "Cloud-native" Platforms [6]. This is because they need massive parallel processing to scale their file processing environments. Software services and tools on cloud-native platforms enable scalable provisioning of compute resources, interoperability between

digital objects, indexing to promote discoverability of digital objects, sharing of digital objects between providers and processes, access to and deployment of scientific analysis tools and pipeline workflows, and connectivity with other repositories, registries, and resources. The idea of a Big Data Commons has been advanced at the National Institute of Health (NIH) as "The NIH Commons" (https://datascience.nih.gov/commons) illustrating many of these differences with traditional computing environments.

Understanding the use of Big Data within the policy of healthcare remains a work-in-progress. The cloud-native platform does not need to reside in the public cloud, and similar computational infrastructure can be established in private and public clouds such that parts of the computing infrastructure can move from private to public clouds as needed, especially when considering concerns of data privacy.

Organizing an Approach to Big Data in Healthcare

Big Data faces numerous integration hurdles, both with the EMRS data and with the other Big Data. This is due to several factors, including the relatively unstructured approach to data organization in the cloud-native environment. The approach of S3 file storage is to allow each file to be described with metadata that helps organize the file system, but the internal data structures of the files cannot be directly queried. Various sets of files commonly have different internal structures, defined by a schema that exists in a separate file with a schema definition language. In turn, the files can be read and their content manipulated by the robust, parallel processing computational environment that exists in cloud-native platforms. However, this paradigm is not like the way data is queried in EMRS database relational structures, and this means data integration will need an innovative approach.

Another integration hurdle for Big Data is that the sources are often in data silos that are not directly semantically interoperable, making queries across the systems nearly impossible. For example, patient electroencephalogram waveform data from seizure monitoring is commonly put into files in specialized systems. The metadata of the files are not mapped to standard annotations, or annotations of any kind, that are recognized at the enterprise level. The system itself may use S3 storage, but the internal structures of the data are completely unrecognized by a traditional healthcare enterprise query system. Without a query language to make a semantic link between systems, much of this data is left on the healthcare analytics' cutting room floor.

A third integration hurdle is the necessity to adopt new policies that deal with the sharing of Big Data, which must address at least four issues: 1) what data to focus upon sharing; 2) how to place boundaries on the way data is shared; 3) how to maintain the cost of sharing data; and 4) how to provide adequate motivation for stakeholders to support the sharing of the data.

To focus on what data to share, it is important to understand the use cases behind selecting the data to be shared. For example, the initial focus in a US healthcare data network named PCORNet was to include standard coded healthcare data only, but when a needs' assessment was later undertaken among the network sites, most needs were found to be for more specific data such as cancer registry data, rare laboratory tests, and personal health data [7]. This emphasized that, although "classic" data in the EMRS may be easier to handle than Big Data, use cases are guiding the need to include Big Data, and because this integration will take a significant investment, it will be important to properly assess which Big Data should be prioritized.

Most high priority data will need to be understood at the patient level, that is, data will need to be attributed to specific patients. However, one of the most significant challenges for metadata descriptions of Big Data is that Big Data often arrives in large "blocks" without such attribution. For example, environmental data from area sensors give detailed data regarding the attributes attached to each sensor node, but there is no direct link to a patient's exposure. Even data attributed to people is in large blocks, like the registry of data from people who were participants in a study, or a file of Twitter feeds from the entire population over the past year where people are identified in a coded form. This presents a challenge to placing boundaries on what data is shared, especially regarding consented and unconsented patients. Generally, one must obtain patient consent to associate the internal data with specific patients, although in general, Big Data sets that are shared as "de-identified" pose a significant risk for re-identification due to the incredibly detailed data which may be available per patient [8]. This makes Big Data a challenge for integration into the healthcare enterprise. Most EMRS are not prepared for the challenges of untangling the patient attributions in the data, nor placing boundaries around the granular data elements inside the Big Data.

Although the cost of keeping Big Data in the public cloud is comparable to the cost of local S3 storage, the cost of transferring data can be very high (http://www.networkworld.com/article/3164444/cloud-computing/how-to-calculate-the-true-cost-of-migrating-to-the-cloud.html). This reflects the general principle that costs are minimized by operating on the data locally, and extracting the results of an operation, rather than transferring the entire Big Data set to a remote site for an operation to be performed. For example, although millions of seconds of waveform data may be available, only particular episodes of cardiac arrhythmia or brain epileptiform discharges may be desired. Supplying metadata can direct analysis programs operating at the location of the data to return the most useful data if this function is built into the local processing units. Therefore, an architecture that minimizes data transfer when using Big Data adopts a set of micro-services that allow operations on the data in-place [9]. One must consider that, otherwise, many micro-transactions on very different kinds of Big Data would need to be managed, requiring a sophisticated system for accepting the updating transactions, reconciliation, and failure recovery if one were to attempt to aggregate the Big Data centrally.

Providing the motivation for stakeholders to open their Big Data repositories touches many disciplines, including the cost of preparation, legal ownership, and attribution.

Sharing Big Data can make data provenance and credit attribution to the creators of the data difficult to manage, and several proposals have been made regarding attribution for "block" data sharing, including the FAIR (Findability, Accessibility, Interoperability, and Reusability) paradigm [10]. The FAIR paradigm has been formulated for web services to automatically find and use Big Data. An example of a Big Data index that implements FAIR principles is the NIH Big Data bioCADDIE Application (https://biocaddie.org/).

An Enterprise Big Data Commons could be used to integrate the Big Data for secondary use into translational medicine and clinical care (see Figure 1). An Enterprise Big Data Commons is created in a hospital to accomplish two essential functions. The first one is to employ an indexing strategy that allows data to be searched, preferably in the location it exists rather than copied to a new location. The second function is to make the data available through web services that allow the data on a specific patient to be found, and for specific data to be returned from Big Data repositories. It is less desirable to unnecessarily obtain an entire data set in these cases, which is often the case with Big Data file objects, exposing to unnecessary scrutiny what can often be an entire block of patients. Additionally, a micro-service layer allows a policy to be put into a place that reflects the sensitivity and ownership of the data. The local control of the data in the Big Data repositories allows updates to be managed locally and exposure of new types of data can occur naturally with data policies that are controlled at each site.

A formulation of an Enterprise Big Data Integration was undertaken in the PIC-SURE (Patient-centered Information Commons: Standardized Unification of Research Elements) Big Data to Knowledge (BD2K) Center of Excellence, sponsored by the National Institutes of Health, where a distributed query system ties together several different platforms commonly used for patient-oriented research into a single system which can be queried to return integrated patient counts and data (http://pic-sure.org). The method of unification for the Big Data platforms is through an ontology-driven approach whereby individual ontology items specify where data resides throughout the system. Once the ontology is used to locate the specified data in the system, services then direct queries against views of database fact and dimension tables [11, 12] derived from the Big Data S3 Repositories. The services combine the results into a single data matrix (table) that appears as though it came from a traditional relational database, but represents the results from Big Data Repositories. This is achieved by incorporating patient features into a common tabular format regardless of the original source. The tabular format can be used for further analysis.

Two innovations have allowed this integration of Big Data and the satisfaction of many of the requirements described above. The first is a web service layer that allows a system such as Informatics for Integrating Biology and the Bedside (i2b2) to act as a supervisory layer for querying the Big Data if the Big Data can be transformed so that at least some of its features can be placed into a patient-centric observation-fact table [12].

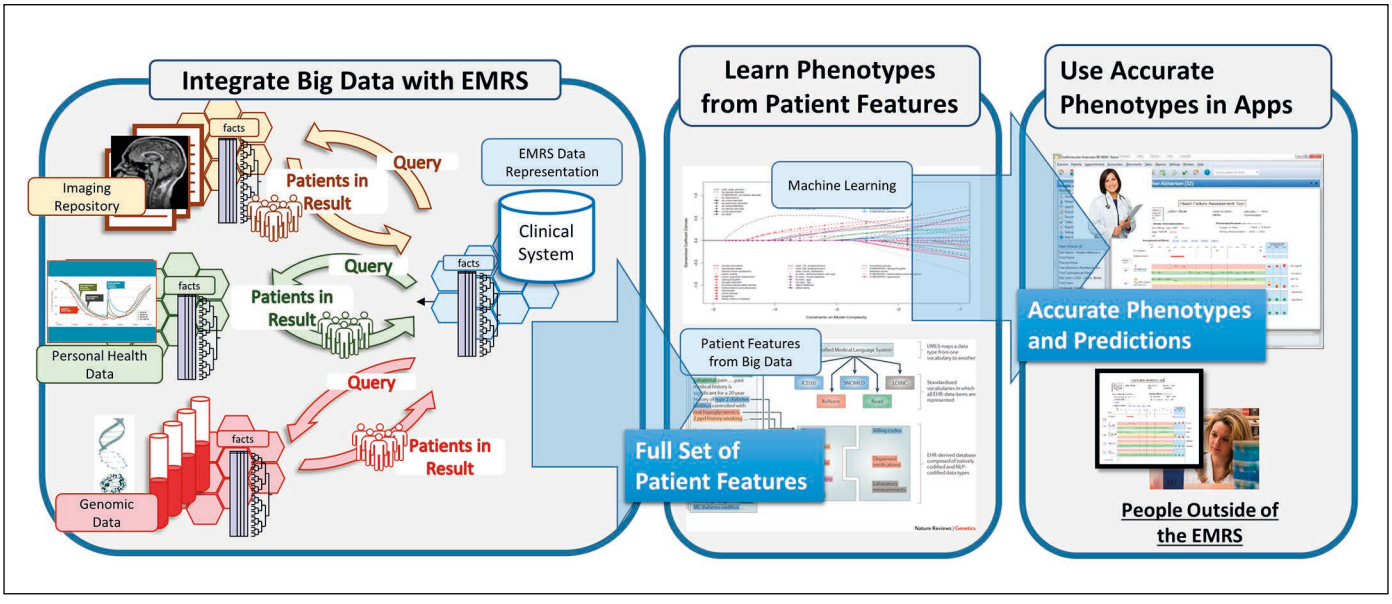

Fig. 1 shows the essential parts of integrating Big Data into Clinical Care. Starting at the left, the figure illustrates independent Big Data repositories of Imaging, Genomics, and Person Health data. Each repository exists as a compendium of files, often in an S3 file repository. The files are formulated with software into a standardized index of observations and facts, such as that which exists in an i2b2 Star Schema. It is the standardized index which is then queried through the EMRS Clinical System and integrated into queries against a similar standardized index created from the EMRS data. This query system for Big Data allows a full set of patient features from the integrated systems to be obtained and subsequently creates and optimized further features through machine learning. As described in the text, this machine learning takes advantage of the Big Data to overcome biases and errors in the clinical EMRS data. Accurate patient phenotypes are integrated into "Apps" which are able to rapidly adapt to the new features available in the Big Data, presenting a complete picture of the patient for decision support and clinical care.

The fact table in each Big Data repository serves to associate many of the features of the data with a specific patient, and software can distribute queries to the different patient-centric data repositories and achieve a complex level of inter-system integration.

The second innovation is an ontology specification for every feature that exists in one of the Big Data repositories throughout the enterprise. The ontologies can be used to formulate patient-centric queries such as previously described [12]. The ontologies of the features are published and reconciled at a central site so that the terms can be used in a query. Likewise, the patients are published to a central honest broker and de-duplicated into a set of matched identifiers across the repositories. The honest broker publishes a table of matches and match probabilities, which does not need to contain explicit HIPAA identifiers if the sites can exchange coded patient identifiers. Only the honest broker controls the matching algorithm and the Protected Health Information used for the matching. Once all of this is in place, a query can be formulated at the central site and relevant sub-queries distributed to the Big Data repositories. Each system then runs the query against its fact table(s), returning the set of patient features or patients matching the features as specified by the ontology terms in the query.

Harnessing the Power of Big Data to Find Important Patient Features

The unification of patient data from repositories of genomic data, imaging data, waveform data, and patient reported data can provide value much greater than any single source of data because some of the inherent biases in each source of data can be overcome when these same biases are absent in alternative sources [13-15]. For example, the lack of time dependence for germline genomic data allows it to overcome the sampling bias present in most clinical laboratory data (when tests are taken at a time when patients are believed to be sickest). Of course, genomic data comes with its own biases, but this is different from laboratory sampling biases. Other common sources of secondary healthcare data, such as the coded data obtained from billing, are also inherently biased due, this time, to oversampling of the sickest patients for concurrent diseases [16]. The billing diagnoses are also biased from the coding practices within the business of healthcare, which optimizes the codes that guarantee the greatest reimbursement [17]. For advanced analytics to be performed upon these data, one needs a strategy to eliminate the inaccuracies found in commonly recorded healthcare features.

Harnessing the power of Big Data and integrating it with the data from the EMRS requires methods and tools to identify important features that can be transformed into knowledge to inform clinical decision-making. The EMRS features heterogeneous data sources combined for billing and clinical care. Because the data are not collected primarily for science-based clinical care, the veracity and reliability of the data are inadequate for analysis in its raw form. On the other hand, the scale and diversity of the data provide opportunities for developing machine learning algorithms that combine multiple data types to increase reliability. Below, we present three main use cases for identifying important features in Big Data sources that can power clinical decision support tools and help clinicians make sense of the deluge of data to help care for patients.

Phenotyping

The diversity of Big Data provides an opportunity to characterize complex phenotypes used for discovering genetic associations with disease. A decade ago, genome-wide association (GWAS) studies focused on a binary phenotype (presence of absence of a disease) for identifying pathogenic variants. However, scientists have subsequently discovered that most diseases are not caused by a single gene (Mendelian) but are polygenic with many variants conferring small amounts of risk. Complex genetic traits thus require more power to detect smaller effects [18]. This power is achieved through both larger cohorts and richer and more reliable phenotypic information to quantify subtle difference in disease populations. Linking Big Data EMRS datasets to biological samples and genomic information in Biobanks is an emerging approach for the exploration of genetic causes of both common and rare diseases [19]. Relying solely on ICD-9/10 codes for defining disease cohorts provides inadequate specificity for genetic studies [20]. Machine learning approaches have been utilized to develop high-throughput phenotyping algorithms for characterizing patients with treatment-resistant depression, rheumatoid arthritis, and cerebral aneurysms, among many others [21-23]. Training and validation of algorithms is done through clinician chart review of the medical record as well as with in-person clinical trials [24].

Many important features in the EMRS are "locked" into narrative text [25)]. Natural Language Processing (NLP) tools must be used to leverage the information stored in clinical notes and reports, and patient-related documents, such as blogs and social media. Researchers utilize the UMLS Metathesaurus to map terms that are semantically equivalent and use public reference sources such as Wikipedia and Medscape to identify important text features associated with a disease [26].

Methods for matching controls within hospital-based populations are also improving classification accuracy among comparison cohorts [16]. The portability of these machine learning algorithms across populations and EMRS is an open research question with considerable implications as EMRS are joined together via national clinical research networks [27, 28].

A collinear application of phenotyping is cohort identification for clinical trial recruitment. This use case focuses more on the sensitivity of machine learning algorithms since clinical trials often require large numbers of patients screened to meet enrollment targets. Methods have been developed to match the EMRS data to trial inclusion and exclusion criteria, and real-time alerting of eligible subjects within the EMRS interface [29, 30]. International efforts are underway aiming to make clinical trial participants more representative of the general population using networks of EMRS for recruitment [31, 32].

Hypothesis Generation

Big Data techniques provide unique opportunities to conduct the unsupervised analysis of the data to identify novel features correlated to diseases and outcomes. These approaches can help elucidate common biological pathways among diseases. For example, clusters of comorbidities have been used to identify population subtypes of autism, schizophrenia, and inflammatory bowel disease using EMRS data [33, 34]. Unsupervised methods have been developed to identify new indications for existing medications – so called drug repurposing [35] – as well as to detect new adverse drug interactions that may put patients at risk [36]. In the area of health services utilization, unsupervised approaches can help detect patterns in the delivery of care that can inform future machine learning algorithm development [37]. Unsupervised approaches have been criticized as difficult to reproduce and filled with spurious associations [38]. Indeed, these results must be considered as only the first steps in identifying an important question or hypothesis, which can then be confirmed with controlled observational and interventional studies.

Prediction – Clinical Decision Support

Accurate classification of disease cohorts is a precursor for identifying predictive features in Big Data that could lead to interventions that prevent disease, hospital re-admissions, and adverse events associated with drugs [39]. Naïve Bayes models have been used to assign risk scores based on a combination of features for the prediction of future suicide attempts and domestic abuse in children [40, 41]. Other modelling approaches have focused of quantifying the cognitive burden of patients' medications to predict future falls and hospital re-admissions [42, 43].

In the US, the shift from volume to value-based payment models creates incentives for providers to promote good outcomes. These improved outcomes include preventing re-admissions and emergency room visits and overall reduction in costs incurred by high-risk patients [44]. Initial outcome studies have shown some success in achieving improved outcomes through Big Data analytics and clinical decision support after surgical interventions and the early detection of sepsis [45, 46].

The increasingly complex health care environment presents clinicians with an amount of data that in some cases exceeds a person's ability to organize and make sense of the data. These growing demands beg for clinical decision support tools that help clinicians make sense of all the data. These support tools will rely on aggregated data from heterogeneous data sources in a common information model, which can be used to train and deploy machine learning algorithms to identify patient cohorts and predict future outcomes. The next section describes a new breed of applications in healthcare, the focused "App" which by using innovative visualization techniques can present this information in a clinically-relevant way and unlock the value of EMRS utilization of Big Data.

Apps to Connect the Healthcare Big Data Commons to Clinical Care

To complete the infusion of research and innovation into healthcare outside the constraints of the EMRS, a new ecosystem of "Apps" is being developed that may provide a way to bridge Big Data into the EMRS workflow and therefore allow its translation into science-based clinical care. Historically, the point of care has been a walled garden. EMRS principally displays to clinicians the information they entered previously, but not the wide range of data and Web-based services that could and should drive cost-efficient care and decision-making [47]. These barriers have limited the health care encounter from taking advantage of features derived from Big Data sources.

But now there is an opportunity to expose the point of care to third party apps that can bridge clinical activities with Big Data. Apps are a particularly good fit with Big Data because of the ability to compute upon the variety of features found in Big Data in a widely distributable application. Both Meaningful Use Stage 3 and the 21st Century Cures Act require that certified health information technologies have Application Programming Interfaces (API) to bring the full power of the Web to patient interactions, including external services and data.

In healthcare, APIs are opening the clinical encounter to third-party IT innovation and redefining clinical interoperability in terms of substitutability -- apps that can be added to or deleted from EMRS as easily as from a smartphone [48]. In 2009, Mandl and Kohane proposed that the EMRS should be re-imagined as "iPhone-like" platforms supporting a selection of 'substitutable' modular third-party applications (Apps). Substitutable apps connected to the EMRS bring the full power of the Web to patient interactions, including external services and data [49]. Subsequently, the SMART Health IT project was funded by the Office of the National Coordinator of Health Information Technology (ONC) under the Strategic Health IT Advanced Research Projects (SHARP) program as a part of the HITECH Act.

Technically, SMART leverages Health Level Seven's (HL7) Fast Health Interoperability Resources (FHIR) to help standardize communication between apps and the EMRS, several health care vocabularies to ensure common nomenclature (including RxNorm for medications, LOINC for laboratory test, SNOMED CT for clinical terminology), and a universal web standard (OAuth 2) for authorizing the apps to access health data. Together, these technologies form a robust apps API.

By defining an API that consistently presents well-specified data, SMART Health IT (a) enables purchasers, users, and administrators of platform-based systems to be able to install and subsequently substitute apps from different vendors without software programming and (b) creates a broad market for app developers across multiple systems, including the EMRS, patient-facing apps, and health information exchanges. Fostering third-party apps creates a market where innovations compete with each other for purchase and use by providers (and patients), thus reducing dependency on updates and specific functions made by an EMRS vendor [50]. Recently the ONC funded an

Apps Gallery [51] for end users—clinicians, researchers, and patients at home to locate apps that can be connected to the EMRS via SMART and FHIR.

These apps not only enable local customization of the EMRS to clinical tasks, but also allow the connection between the EMRS and external data and services. To advance genomic medicine, traditional EMRS cannot incorporate, for example, whole exome sequences into their data core. But SMART enables the EMRS to readily query an external FHIR genomics server to provide decision support to a treating clinician [52, 53]. An example is the Precision Cancer Medicine (PCM) app, designed to present patients' genomic test results to oncologists in real time as a component of clinical practice, as well as provide links to external knowledge bases. Because the app was developed against the SMART API, even though the initial deployment was at Vanderbilt, it can easily be deployed to other EMRS.

An emerging feature of the SMART ecosystem is the adoption of a standardized, service-oriented architecture (SOA) for clinical decision support (CDS)—SMART CDS Hooks [54]. This approach separates CDS rules from the EMRS itself. Instead, the actionable knowledge is hosted on a rules engine made accessible by a third party. EMRS vendors, including Cerner and Epic, are embedding hooks, or triggers in the EMRS, to launch third-party decision services and SMART apps after key events—for example, medication prescribing. This approach is highly promising for bringing predictive analytics to the point of care. The computation of risk scores in real time (e.g. risk of readmission after discharge for congestive heart failure) or firing of standardized rules (e.g. immunization recommendations from the Advisory Committee on Immunization Practices) can occur at a single point of control, but influence care everywhere.

Conclusion

Infusing Big Data into the healthcare encounter has several barriers. First, the data are often positioned outside the EMRS with their own distinct semantic and technical structures. Integrating big data into clinical systems requires web services organized around a common information model to reach out to the Big Data Repositories as they exist in cloud-native infrastructures. Second, features that are important to healthcare often need to be computed from, rather than being readily available from, the Big Data. Finally, until the development of SMART App standards, there were few ways to present features of Big Data into the patient-clinician encounter.

What is essential in representing Big Data is a permissive data structure that can represent features, including computed features, of the data which can dynamically expand using an ontology driven-approach. This model can be linked to indexing and data retrieval paradigms through web services. With a dynamically expandable ontology-driven data model, new types of features of Big Data can be accommodated. Local groups can be involved in creating specialized data sets. Different standards can be accommodated as agreed upon by specialized groups. Various levels of data granularity can be accommodated, so that the data presentation can more precisely adhere to privacy needs to allow a more targeted answer for what a clinician "needs to know" rather than receiving large data sets with multiple patients which is the hallmark of most Big Data. The ability to lose the data chain of custody is minimized and attribution can be specifically assigned to data holders.

Apps that will run inside and outside the EMRS can provide support for complex decisions and workflows that involve genomics, imaging, and personal health repositories. Accurate phenotyping can become a routine part of clinical care. An infrastructure based on the SMART API will allow an ecosystem of apps to be shared across healthcare institutions using web service standards such as FHIR. This provides a vision for a new type of EMRS, which is expected to play out over the next five to ten years.

Acknowledgments

Thanks to Christopher Herrick for development of figure 1. Funding from NIH awards U54 HG007963 and RO1 HG009174 contributed to these concepts.

References

1. Haneuse S, Daniels M. A General Framework for Considering Selection Bias in EHR-Based Studies: What Data Are Observed and Why? EGEMS (Wash DC) 2016;4(1):1203.
2. Johnson RJ. A Comprehensive Review of an Electronic Health Record System Soon to Assume Market Ascendancy: EPIC. J Healthc Commun 2016;1(4).
3. Ash JS, Bates DW. Factors and Forces Affecting EHR System Adoption: Report of a 2004 ACMI discussion. J Am Med Inform Assoc 2005;12(1):8-12.
4. Johnson SB, Friedman C. Integrating Data From Natural Language Processing into a Clinical Information System. Proc AMIA Annu Fall Symp 1996:537-41.
5. Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, Efron MJ, et al. Big Data: Astronomical or Genomical? PLoS Biol 2015;13(7):e1002195.
6. Stine M. Migrating to Cloud-Native Application Architectures: OReily; 2015.
7. Berchuk A, Kahn M, Rusincovitc S, Meeker D, Murphy S, Bhosale R, et al. Editor Assessment and Planning for the PCORnet Common Data Model. AMIA 2017 Joint Summits on Translational Science; 2017.
8. Cimino JJ. The False Security of Blind Dates: Chrononymization's Lack of Impact on Data Privacy of Laboratory Data. Appl Clin Inform 2012;3(4):392-403.
9. Knowledge Representation for Health Care, HEC 2016 International Joint Workshop. New York, NY: Springer Berlin Heidelberg; 2017.
10. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for Scientific Data Management and Stewardship. Sci Data 2016;3:160018.
11. Kimball R. The Data Warehouse Toolkit : Practical Techniques for Building Dimensional Data Warehouses. New York: John Wiley & Sons; 1996.
12. Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S, et al. Serving the Enterprise and Beyond with Informatics for Integrating Biology and the Bedside (i2b2). J Am Med Inform Assoc 2010;17(2):124-30.
13. Kruse CS, Goswamy R, Raval Y, Marawi S. Challenges and Opportunities of Big Data in Health Care: A Systematic Review. JMIR Med Inform 2016;4(4):e38.
14. Belle A, Thiagarajan R, Soroushmehr SM, Navidi F, Beard DA, Najarian K. Big Data Analytics in Healthcare. Biomed Res Int 2015;2015:370194.
15. Brookhart MA, Sturmer T, Glynn RJ, Rassen J, Schneeweiss S. Confounding Control in Healthcare Database Research: Challenges and Potential Approaches. Med Care 2010;48(6 Suppl):S114-20.
16. Castro VM, Apperson WK, Gainer VS, Ananthakrishnan AN, Goodson AP, Wang TD, et al. Evaluation of Matched Control Algorithms in EHR-based Phenotyping Studies: a Case Study of Inflammatory Bowel Disease Comorbidities. J Biomed Inform 2014;52:105-11.
17. O'Malley KJ, Cook KF, Price MD, Wildes KR, Hurdle JF, Ashton CM. Measuring diagnoses:

ICD Code Accuracy. Health Serv Res 2005;40(5 Pt 2):1620-39.
18. Solovieff N, Cotsapas C, Lee PH, Purcell SM, Smoller JW. Pleiotropy in Complex Traits: Challenges and Strategies. Nat Rev Genet 2013;14(7):483-95.
19. Gainer VS, Cagan A, Castro VM, Duey S, Ghosh B, Goodson AP, et al. The Biobank Portal for Partners Personalized Medicine: A Query Tool for Working with Consented Biobank Samples, Genotypes, and Phenotypes Using i2b2. J Pers Med 2016;6(1).
20. Sinnott JA, Dai W, Liao KP, Shaw SY, Ananthakrishnan AN, Gainer VS, et al. Improving the Power of Genetic Association Tests with Imperfect Phenotype Derived from Electronic Medical Records. Hum Genet 2014;133(11):1369-82.
21. Perlis RH, Iosifescu DV, Castro VM, Murphy SN, Gainer VS, Minnier J, et al. Using Electronic Medical Records to Enable Large-Scale Studies in Psychiatry: Treatment Resistant Depression as a Model. Psychol Med 2012;42(1):41-50.
22. Liao KP, Cai T, Savova GK, Murphy SN, Karlson EW, Ananthakrishnan AN, et al. Development of Phenotype Algorithms using Electronic Medical Records and Incorporating Natural Language Processing. BMJ 2015;350:h1885.
23. Castro VM, Dligach D, Finan S, Yu S, Can A, Abd-El-Barr M, et al. Large-scale Identification of Patients with Cerebral Aneurysms Using Natural Language Processing. Neurology 2017;88(2):164-8.
24. Castro VM, Minnier J, Murphy SN, Kohane I, Churchill SE, Gainer V, et al. Validation of Electronic Health Record Phenotyping of Bipolar Disorder Cases and Controls. Am J Psychiatry 2015;172(4):363-72.
25. Hripcsak G, Albers DJ. Next-Generation Phenotyping of Electronic Health Records. J Am Med Inform Assoc 2013;20(1):117-21.
26. Yu S, Liao KP, Shaw SY, Gainer VS, Churchill SE, Szolovits P, et al. Toward High-Throughput Phenotyping: Unbiased Automated Feature Extraction and Selection from Knowledge Sources. J Am Med Inform Assoc 2015;22(5):993-1000.
27. Mandl KD, Kohane IS, McFadden D, Weber GM, Natter M, Mandel J, et al. Scalable Collaborative Infrastructure for a Learning Healthcare System (SCILHS): Architecture. J Am Med Inform Assoc 2014;21(4):615-20.
28. Weber GM, Murphy SN, McMurry AJ, Macfadden D, Nigrin DJ, Churchill S, et al. The Shared Health Research Information Network (SHRINE): a Prototype Federated Query Tool for Clinical Data Repositories. J Am Med Inform Assoc 2009;16(5):624-30.
29. Goodwin T, Harabagiu SM, editors. Automatic Generation of a Qualified Medical Knowledge Graph and its Usage for Retrieving Patient Cohorts from Electronic Medical Records. 2013 IEEE Seventh International Conference on Semantic Computing (ICSC). IEEE: 2013.
30. Ennis C, Snyder D, Ainsworth T, Stacy M, Sanderson I, editors. Utilization of the EPIC Electronic Health Record System for Clinical Trials Management at Duke University. 2014 IEEE International Conference on Healthcare Informatics (ICHI). IEEE: 2014.
31. Geifman N, Butte AJ. Do Cancer Clinical Trial Populations Truly Represent Cancer Patients? A Comparison of Open Clinical Trials to the Cancer Genome Atlas. Pac Symp Biocomput 2016;21:309-20.
32. Porter M, Ramaswamy B, Beisler K, Neki P, Single N, Thomas J, et al. A Comprehensive Program for the Enhancement of Accrual to Clinical Trials. Ann Surg Oncol 2016;23(7):2146-52.
33. Doshi-Velez F, Ge Y, Kohane I. Comorbidity Clusters in Autism Spectrum Disorders: an Electronic Health Record Time-Series Analysis. Pediatrics 2014;133(1):e54-63.
34. Ananthakrishnan AN, Gainer VS, Cai T, Perez RG, Cheng SC, Savova G, et al. Similar Risk of Depression and Anxiety Following Surgery or Hospitalization for Crohn's Disease and Ulcerative Colitis. Am J Gastroenterol 2013;108(4):594-601.
35. Dang T-T, Ouankhamchan P, Ho T-B, editors. Detection of New Drug Indications from Electronic Medical Records. 2016 IEEE RIVF International Conference on Computing & Communication Technologies, Research, Innovation, and Vision for the Future (RIVF). IEEE: 2016.
36. Tatonetti NP, Denny JC, Murphy SN, Fernald GH, Krishnan G, Castro V, et al. Detecting Drug Interactions from Adverse-Event Reports: Interaction Between Paroxetine and Pravastatin Increases Blood Glucose Levels. Clin Pharmacol Ther 2011;90(1):133-42.
37. Weber GM, Kohane IS. Extracting Physician Group Intelligence from Electronic Health Records to Support Evidence Based Medicine. PLoS One 2013;8(5):e64933.
38. Open Science C. PSYCHOLOGY. Estimating the Reproducibility of Psychological Science. Science 2015;349(6251):aac4716.
39. Walsh C, Hripcsak G. The Effects of Data Sources, Cohort Selection, and Outcome Definition on a Predictive Model of Risk of Thirty-day Hospital Readmissions. J Biomed Inform. 2014;52:418-26.
40. Reis BY, Kohane IS, Mandl KD. Longitudinal Histories as Predictors of Future Diagnoses of Domestic Abuse: Modelling Study. BMJ 2009;339:b3677.
41. Barak-Corren Y, Castro VM, Javitt S, Hoffnagle AG, Dai Y, Perlis RH, et al. Predicting Suicidal Behavior From Longitudinal Electronic Health Records. Am J Psychiatry 2016:appiajp201616010077.
42. McCoy TH, Castro VM, Cagan A, Roberson AM, Kohane IS, Perlis RH. Sentiment Measured in Hospital Discharge Notes Is Associated with Readmission and Mortality Risk: An Electronic Health Record Study. PLoS One 2015;10(8):e0136341.
43. Castro VM, McCoy TH, Cagan A, Rosenfield HR, Murphy SN, Churchill SE, et al. Stratification of Risk for Hospital Admissions for Injury Related to Fall: Cohort Study. BMJ 2014;349:g5863.
44. Burwell SM. Setting Value-based Payment Goals--HHS Efforts to Improve U.S. Health Care. N Engl J Med 2015;372(10):897-9.
45. Evans RS, Benuzillo J, Horne BD, Lloyd JF, Bradshaw A, Budge D, et al. Automated Identification and Predictive Tools to Help Identify High-risk Heart Failure Patients: Pilot Evaluation. J Am Med Inform Assoc 2016;23(5):872-8.
46. Erskine AR, Karunakaran B, Slotkin JR, Feinberg DT. Harvard Business Review [Internet] 2016. Available from: https://hbr.org/2016/12/how-geisinger-health-system-uses-big-data-to-save-lives.
47. Weber GM, Mandl KD, Kohane IS. Finding the Missing Link for Big Biomedical Data. JAMA 2014;311(24):2479-80.
48. Mandl KD, Mandel JC, Kohane IS. Driving Innovation in Health Systems through an Apps-Based Information Economy. Cell Syst 2015;1(1):8-13.
49. Mandl KD, Kohane IS. No Small Change for the Health Information Economy. N Engl J Med 2009;360(13):1278-81.
50. Bloomfield RA, Jr., Polo-Wood F, Mandel JC, Mandl KD. Opening the Duke Electronic Health Record to Apps: Implementing SMART on FHIR. Int J Med Inform 2017;99:1-10.
51. SMART Apps Gallery 2017 [Available from: https://gallery.smarthealthit.org/.
52. Warner JL, Rioth MJ, Mandl KD, Mandel JC, Kreda DA, Kohane IS, et al. SMART Precision Cancer Medicine: a FHIR-based App to Provide Genomic Information at the Point of Care. J Am Med Inform Assoc 2016;23(4):701-10.
53. Alterovitz G, Warner J, Zhang P, Chen Y, Ullman-Cullere M, Kreda D, et al. SMART on FHIR Genomics: Facilitating Standardized Clinico-genomic Apps. J Am Med Inform Assoc 2015;22(6):1173-8.
54. Berman JJ. Concept-match medical data scrubbing. How Pathology Text can be used in Research. Arch Pathol Lab Med 2003;127(6):680-6.

Correspondence to:
Shawn Murphy MD, Ph.D.
HMS Professor of Neurology and Partners' Healthcare
Chief Research Information Officer
Laboratory of Computer Science
50 Staniford Street, 7th floor
Boston, MA 02114, USA
E-mail: murphy.shawn@mgh.harvard.edu