Published in final edited form as:

Int J Radiat Oncol Biol Phys. 2016 July 1; 95(3): 890–894. doi:10.1016/j.ijrobp.2015.10.060.

Impending Challenges for the Use of Big Data

Vojtech Huser¹ and James J Cimino²

¹Lister Hill National Center for Biomedical Communications, National Library of Medicine, National Institutes of Health, Bethesda, Maryland, *Highest academic achievement: staff scientist at NIH (NIH academic ranks differ from the traditional structure)*

²Informatics Institute, University of Alabama at Birmingham, Birmingham, Alabama, *Highest academic achievement: tenured research director position: Director of Informatics Institute at University of Alabama at Birmingham (former tenured associate professor at Columbia University, New York, NY)*

Abstract

Big Data in healthcare promise to answer real world questions that can be difficult to address with randomized clinical trials. In this perspective article, we analyze five current Big Data informatics challenges of (1) data volume, (2) research oversight by Institutional Review Boards, (3) semantic data integration (4) data formats, and (5) access to knowledge bases from a biomedical informatics perspective. After briefly describing each challenge, we discuss examples of how are being addressed today. With the ever increasing number of Big Data healthcare repositories, the emphasis is shifting from mere data aggregation to advanced data transformation and analysis.

Introduction

Advances in data storage and data analysis materialized also in healthcare data. In recent years, we have seen an emphasis on using the full potential^{1,2} of these data to answer questions such as: Who were the patients that received radiotherapy as primary treatment? Who among such patients experienced radiotherapy-related complications? Given everything you know about my case, what is the chance that if I choose radiotherapy, I will experience incontinence in the next year? Factors contributing to this trend include more rapid data querying technologies, cheaper data storage, the addition of genomic data to traditional clinical data sets, “meaningful use incentives” for increasing the adoption of electronic health records, and the recent emergence of precision medicine.³

In this perspective paper, we discuss several challenges ahead of big data, some that are being addressed now and others that will need to be addressed in the near future. The list of challenges presented here is not meant to be an exhaustive list but is rather driven by our Big

Corresponding author: Lister Hill National Center for Biomedical Communications, National Library of Medicine, National Institutes of Health, Bethesda, Maryland, Tel: (301) 435-3193; vojtech.huser@nih.gov.
Formerly: Laboratory for Informatics Development, NIH Clinical Center

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Data experience. For each challenge, we provide comments on current approaches to address the challenge.

Challenges ahead

Data Volume

Despite technologic advances, increases in data volume are outstripping institutions' ability to cope with increased demand for data storage (for example in healthcare data warehouses) while maintaining rapid data retrieval. To demonstrate this, we can consider the recent program started in March 2015 by US President Barack Obama: the NIH Personalized Medicine Initiative (PMI).³ A PMI cohort is expected to be of 1 million patients in size and collected data should include clinical data from Electronic Health Record (EHR) systems, patient reported data, and genomic data.⁴

The initiative is considering assembling the cohort by "stitching together" existing cohorts from large institutions or by pursuing a "direct from participant" approach using the patient's right to electronic copies of their health data. Table 1 shows sizes of genomic data compared to some existing clinical data domains. A PMI cohort would require 5.8 PB ($=10^{18}$ B) if complete genomic data (not just differences from the reference sequence) would be included.

Current approaches to this challenge—Table 1 clearly shows that genomic data require the most space. A recent data size analysis that compared data in genomics, astronomy, video (YouTube) and internet data (Twitter), suggest that the term 'genomical' instead of 'astronomical' should be used to refer to enormous data sizes.⁵

One possible solution to the data size challenge is to move the most voluminous data to a secondary storage-optimized platform that is separate from the primary query-optimized data platform. In such solution, however, a new challenge arises for the analytical algorithm to operate across multiple platforms.

IRB challenge

With wider availability of big data to larger groups of researchers, a traditional approach for study approval may have to shift. The current regulatory interpretation applied by institutional review boards (IRBs) uses the principle of individual project approval which is tied typically to a single clinical domain.⁶ Big Data platforms, on the other hand, are designed to formulate generic strategies (for example, retrieve and analyze past similar cases given a patient's scenario) and apply them to all possible contexts (pancreatic cancer, macular degeneration or osteoporosis). This directly challenges the individual project approval principle. Instead, Big Data approaches seem to require a per-method review and approval. For example, for researchers using the de-identified view of the Oncospace database⁷, a streamlined IRB mechanism can lead to an increased research output. The approval process may have to shift to the method level (method X on dataset Y) and in some instances, even to the dataset level (any method on de-identified dataset Y).

Current approaches to this challenge—One approach to streamlining the IRB process is for data that are external to the organization to be made available to researchers as pre-de-identified. In this case, local IRB treats the data as non-human research and excludes them from requiring IRB review.

Another approach is to file a separate IRB project that governs creation of a Big Data resource. Such a protocol would provide a mechanism for mini-reviews on the repository level (submitted by data requestors to the repository stewards and not to an IRB). The approval time would typically be much shorter compared to an IRB initial review. The local IRB still retains review capacity by reviewing annually (in batch mode) all mini-reviews that occurred in the past year and guiding the repository steward in providing an appropriate review.

Despite some advances and innovation in IRB review,⁸ the challenge of data de-identification remains. For example, there is lack of consensus on setting a threshold value for some common data usage contexts (e.g., a possible threshold rate of 0.01% of records possibly containing protected health information (PHI) or personally identifiable information (PII) for testing data visualization methods deployed within a firewall of a covered entity). It is possible to draw an analogy of the risk of re-identification to the risk of harm from taking a falsely effective drug. In approving drugs to market, we accept a 5% level of type I error (chance of a drug to be falsely thought to be effective). A set of thresholds and contexts similar to this threshold for falsely effective drug risk could be applied to re-identification risk. Although, most likely the privacy risk threshold would be orders of magnitude lower than 5%, for example, 0.01%. The caveat of this analogy is that possible harms resulting from taking a falsely effective drug are different from harms resulting from being re-identified. The drug risk represents a direct health-related risk limited to the single patient while the privacy risk may affect patient's personal or professional life in general and have possibly impact on multiple people, such as patient's family.

Semantic data integration

It is relatively easy to amass a large volume of data using a common information model or data structure (syntactic integration), but many Big Data datasets are suffering from lack of semantic integration (data meaning).⁹ For example, it is possible to have a large number of laboratory results data, but linking equivalent tests from different sources requires context and domain knowledge, and tedious mapping effort. The semantic integration problem typically arises in two possible scenarios. The first scenario is combining routine healthcare data from multiple institutions (for example, CancerLinQ¹⁰), while the second, smaller scenario is combining data from multiple studies conducted at different times or at different institutions (for example efforts to combine data from multiple radiogenomics trials).¹¹ Standards such as the Logical Observation Identifiers Names and Codes (LOINC) and the Systematized Nomenclature of Medicine - Clinical Terms (SNOMED-CT) try to address the problem for routine data. For research data collection, general efforts include various Common Data Elements (CDE)^{12,13} initiatives and radiation oncology specific initiatives, such as Radiation Oncology Ontology¹⁴ or DICOM-RT.¹⁵

Current approaches to this challenge—There are two general approaches to the challenge with very different cost to the Big Data repository custodian. The first approach is to ignore the problem at the data repository level and let the data recipient deal with the semantic heterogeneity problem. The advantage of this approach is that each individual data recipient is most knowledgeable about his/her data integration needs and will only integrate data that matter to his/her analysis and do it in a manner that best fit his/her needs. This conserves central resources. One disadvantage of this approach is that a data recipient may not understand the meanings implied by the data (e.g., “Is this a serum sodium or a urine sodium result?”). A second disadvantage is that if any two or multiple data recipients have overlapping data integration needs, redundant and repeated integrations are carried out. A mechanism to share the data integration scheme or mappings either among the users or between the user and the repository custodian can be very useful. This sharing mechanism, however, has its own many challenges (such as training the data users on numerous standard terminologies and possible mapping formalism) and is difficult to implement in practice.

The second approach is that the repository custodian assumes the responsibility for semantic integration. However, deciding centrally and prior any data analysis which data domains to semantically integrate is difficult and there may be inadequate central resources to tackle all possible domains in need of semantic integration. Another risk is that the data integration done centrally may not fit the needs of every data recipient. Even more complex risk is that the recipient may not be aware of inappropriate data integration done for him/her centrally and may unknowingly generate false analytical results. A combination of the two approaches is what most custodians pursue in practice. However, drawing a clear cut boundary between them remains challenging.

Databases versus data files

Advances in data science have brought about changes to how data analyses are conducted. Part of the Big Data movement, in addition to increasing the available data (more data), is to empower more researchers with smaller resources and programming skills to utilize available Big Data (more users) and allow the user to intuitively access a wider range of analytical methods (more questions).

In the past, large scale analysis often required storage of the data using a databases engine (e.g., MySQL) hosted on a dedicated server to perform large analytical tasks because regular personal computers were not powerful enough for such an analysis. With larger capacities for personal computers, many data scientists can now analyze datasets locally, saving time and programming effort by skipping the need to first import the data into a database platform. Moreover, the availability of simple file manipulation tools (especially command line based tools) has proven increasingly popular. While the use of files created new innovative approaches, Big Data users now have to master more possible data format technologies and data manipulation techniques, and that is core of the third challenge we present here. Additional trend that makes the data analysis and data format issue even more complex is the emergence of alternative database platforms such as NoSQL platforms (databases using “not only SQL [Structured Query Language]” as a main query paradigm), columnar databases (databases that optimize storage by data columns instead of rows), and

Resource Description Framework (RDF) triple stores (databases organized to store trios of subject-predicate-object; for example ‘Thickness of surface bolus’ [SNOMED CT: 228843002] is-a ‘Radiotherapy beam modifier dimension’ [SNOMED CT:228840004]).

Current approaches to this challenge—In the future, file-based approaches and complex command-line-based pipelines will offer convenient simple alternatives to traditional database-based data analysis. Seamless integration of traditional database-hosted data with simple-to-distribute data files is important. Data analysis libraries (such a ‘dplyr’ and tidy data¹⁶ packages in R) provide a unified syntax for accessing a wide spectrum of databases, including access to in-memory data loaded from data distributed as comma separated value files or other simple formats (referred to as “flat files”). The new term “data wrangling”, defined as the process of “making data more usable by changing their form but not their meaning” highlights this common data task performed by Big Data analysts. An example of a data wrangling tasks that is increasingly being made more intuitive for non-expert data consumer is “long-to-wide” data transformation where “long format” of patient data consisting of one data row per subject and data element (often used by large databases) is transformed into a “wide format” with one row per subject that contains multiple data elements pre-selected by the user for easier data review.

Access to knowledge bases

In the Big Data space, the distinction between raw patient level *fact data* (e.g., “John Doe’s chemotherapy event log in 2013”) and necessary *knowledge* (e.g., “sunitinib is a chemotherapy agent first approved in 2006”), typically stored in knowledge bases, becomes very important. The term ‘database’ is often used ambiguously and it is not clear whether database of facts or database of knowledge is meant. The term knowledge base is sometimes preferred to refer to databases of knowledge.

For complex analyses that sift through all permutations of a given problem, dynamic access to knowledge is key to the data analysis or data manipulation. Analyses often need access to multiple knowledge bases and easy convenient access is often the decisive factor in utilizing or not utilizing a given knowledge base. Historically, many analysts would try to import whole given knowledge base on a scheduled basis into their local environment. For example, a weekly data import script may load the most recently registered trials from ClinicalTrials.gov registry or from dbSNP knowledge base into the analytical platform.

Current approaches to this challenge—A partial solution to the repeated download of external knowledge bases is remote access through Application Programming Interfaces (APIs).¹⁷ An API is a language and message format used by an application to communicate with another program or external server. By using APIs, it is possible to modularize a larger task into smaller pieces and make remote resources appear as if they are readily available locally. This is only possible, however, if the amount of data acquired from external source is reasonably small and the access is fast. Service uptime and API use policies that do not substantially restrict external users (e.g., number of API calls per minute or hour, or data transfer volume limitations) are important feasibility factors for replacing regular downloads with API calls. While APIs offer incremental advantages, they come in many flavors and

they still lack further standardization. Big data analysts often have to deal with tens of different APIs each requiring a different set of functions and API specific data format adjustments (e.g., transforming JavaScript Object Notation (JSON) format to XML.) A promising further solution to format standardization, although only applicable to knowledge bases in the Resource Description Framework format, is making a publicly available SPARQL endpoint (reverse acronym for SPARQL Protocol and RDF Query Language)¹⁸ that unifies the query interface and allows integration of internal and external servers in a single query.

Conclusion

The growing phenomenon of Big Data is encouraging and draws a much larger community of researchers and the public to consider the benefits of maintaining large healthcare data repositories. The challenges for the newly trained data scientists are shifting from data acquisition to data transformation. The emergence of the concept of data wrangling is reflecting these new challenges.

Acknowledgments

We would like to thank Justin Kirby and Bhadrassain Vikram for providing comments on an earlier version of this manuscript.

References

1. Hripcsak G, Duke JD, Shah NH, et al. Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. *Studies in health technology and informatics*. 2015; 216:574–578. [PubMed: 26262116]
2. Moore KL, Kagadis GC, McNutt TR, Moiseenko V, Mutic S. Vision 20/20: Automation and advanced computing in clinical radiation oncology. *Med Phys*. 2014; 41(1):010901. [PubMed: 24387492]
3. NIH. [Accessed June 15, 2015] Precision Medicine Initiative. 2015. <http://www.nih.gov/precisionmedicine>
4. Precision Medicine Initiative (PMI) Working Group Report (Sep 17, 2015). 2015. <http://www.nih.gov/precisionmedicine/09172015-pmi-working-group-report.pdf>
5. Stephens ZD, Lee SY, Faghri F, et al. Big Data: Astronomical or Genomical? *PLoS Biol*. 2015; 13(7):e1002195. [PubMed: 26151137]
6. Kano M, Getrich CM, Romney C, Sussman AL, Williams RL. Costs and inconsistencies in US IRB review of low-risk medical education research. *Med Educ*. 2015; 49(6):634–637. [PubMed: 25989411]
7. Oncospace consortium website. 2015. <https://oncospace.radonc.jhmi.edu/>
8. Rosenbloom ST, Harris P, Pulley J, et al. The Mid-South clinical Data Research Network. *J Am Med Inform Assoc*. 2014; 21(4):627–632. [PubMed: 24821742]
9. Santanam L, Hurkmans C, Mutic S, et al. Standardizing naming conventions in radiation oncology. *Int J Radiat Oncol Biol Phys*. 2012; 83(4):1344–1349. [PubMed: 22245204]
10. Schilsky RL, Michels DL, Kearbey AH, Yu PP, Hudis CA. Building a rapid learning health care system for oncology: the regulatory framework of CancerLinQ. *J Clin Oncol*. 2014; 32(22):2373–2379. [PubMed: 24912897]
11. Kerns SL, Ostrer H, Rosenstein BS. Radiogenomics: using genetics to identify cancer patients at risk for development of adverse effects following radiotherapy. *Cancer Discov*. 2014; 4(2):155–165. [PubMed: 24441285]

12. NLM. [Accessed Aug 12, 2015] Common Data Element (CDE) Repository. 2014. <http://cde.nlm.nih.gov>
13. PROMIS. [Accessed Aug 14, 2015] Patient Reported Outcomes Measurement Information System: Sample Instrument. 2015. <http://www.nihpromis.org/measures/SampleQuestions#physicalhealth>
14. CancerData.org. Radiation Oncology Ontology. <https://www.cancerdata.org/roo>
15. Law MY, Liu B. Informatics in radiology: DICOM-RT and its utilization in radiation therapy. *Radiographics*. 2009; 29(3):655–667. [PubMed: 19270073]
16. Wickham H. Tidy data. *Journal of Statistical Software*. 2014; 59(10)
17. Wikipedia. [Accessed Sep 10, 2015] Application programming interface. 2015. https://en.wikipedia.org/wiki/Application_programming_interface
18. Tilahun B, Kauppinen T, Kessler C, Fritz F. Design and development of a linked open data-based health information representation and visualization system: potentials and preliminary evaluation. *JMIR Med Inform*. 2014; 2(2):e31. [PubMed: 25601195]

Summary

We discuss five current Big Data informatics challenges of (1) data volume, (2) research oversight by Institutional Review Boards, (3) semantic data integration, (4) data formats, (5) access to knowledge bases, and from a biomedical informatics perspective.

Table 1

Data type	Data elements	Single patient (average)	Cohort of 1 million patients
Clinical reports	Text	10 MB	10 TB
Lab results	Value, units, flag	0.3 MB	0.3 TB
Administrative + EHR data	Dx, Proc, Rx	2 MB	2 TB
Exome genomic data (variants) (VCF)	Position, type, base(s)	125 MB	125 TB
Imaging data	Multiple image formats	421.9 MB *	421.9 TB *
TOTAL		559.2 MB	559.2 TB
Raw exome genomic data (BAM)	Position, base, quality	5.7 GB	5.7 PB
GRAND TOTAL		6.3 GB	6.3 PB

MB: megabyte (10^6), GB: gigabyte (10^9), TB: terabyte (10^{12}), PB: petabyte (10^{15}), EB: Exabyte (10^{18})

* Imaging data estimate does not represent an average patient but is based on cancer patient cohort in the Cancer Imaging Archive⁴ (13.5TB of image data for approximately 32 thousand cancer patients; data as of April 2015)