What is the Role of Big Data in Health

Matthew Durbin, MD FAAP
Indiana University School of Medicine Department of Pediatrics, Division of NeonatologyfiRiley Hospital for Children
699 Riley Hospital Drive
Indianapolis, Indiana 46202fi
mddurbin@iu.edu

ABSTRACT

MISSING

KEYWORDS

i523
complete

1 WHAT IS THE ROLE OF BIG DATA IN HEALTH

The current state of healthcare system in the United States is often described as a crisis. The term comes with good reason, as spending accounts for 17-18% of GDP, dwarfing other nations, and is exponentially rising at an unsustainable rate. For all of our spending, we have poorer health than most developed and many developing nations. The healthcare industry is behind in technology, with recent adoption of an electronic medical record, and prior reliance on paper charting. Communication is most often by decades old technology including phone or fax. Internet communication between healthcare providers, and with patients, is a recent novelty. We have to poorest health, including obesity due to poor diet, lack of exercise, and substance abuse. We pay more for pharmaceuticals than any other country, and most pharmaceutical budget goes to marketing as opposed to research and development. Meanwhile the business world is far ahead of healthcare.

Big Data has major potential to impact health. Massive data sets related to human health are compiled by insurance companies, pharmaceutical companies, public health institutions and research institutions. Big data will soon have a huge impact on improving the health, but there is a long road ahead. Much of the lag is due to serious issues with privacy and security. The healthcare industry should be able to overcome these obstacles as online banking and financial institutions have done. There is amazing potential with big data and healthcare, but a long way to travel. [4] Healthcare is making strides and big data collection is visible everywhere. The electronic medical record EMR is close to universal and is improving constantly. Medical resources are accessible around the world through smartphones, making medical libraries obsolete. Next generation sequencing technologies are able to measure the genetic contributions to disease that previously a mystery. Wearable technology and fitness tracking apps, nutrition apps are improving personal.

2 COST OF HEALTHCARE

2.1 The Current State

One of the most troubling issues facing the United States, and the world, is the increasing cost of healthcare. The problems are different around the globe. Much of the developing world lacks access to adequate healthcare, which is a serious problem. This paper focuses on a different problem, in the crisis facing the United States. Current healthcare spending is greater than 3 trillion dollars [3]. This makes up 17 percent of GDP. This number grows every year and is unsustainable. This number affects citizens deeply, and currently healthcare costs are responsible for 50% of bankruptcy claims in the United States [4]. All of this extra spending does not equal better health. In most measures of health, from infant mortality to life expectancy, the United States find itself far from the top. There are major issues at play ranging from a massive bureaucracy, to the poor health and obesity of participants.

2.2 The Future

It is projected that the average family will spend over 25% of income on to healthcare [4]. The problem is not projected to improve. As the *baby-boomers* age, the population over 60 with high cost chronic healthcare problems, increases exponentially. In Medical School, we were taught about this *silver tsunami* approaching the US healthcare system (prompting me to go into Pediatrics.) Many individuals, including myself, look to Big Data to uncover these problems and help fix them. Before it is too late. There are technology solutions including the electronic health record, medical reference technology, genomic medicine, telemedicine, wearable health technology, and personalized medicine.

3 ELECTRONIC HEALTH RECORD

3.1 Adoption of and EMR

Throughout history, medical records were taken on paper, but after 2000 the slow transition to electronic records began [7]. The handwritten records were kept in large file cabinets, and when records needed to be shared between physicians or institutions (across the country or across the street), the paper records were faxed over a telephone line. This technology is decades old. As technology raced forward with supercomputers and the worldwide web, medicine continued to use these antiquated forms of communication. Finally, government mandating forced healthcare systems into the modern era and electronic records went online. Currently over 84% of health records are online [4].

3.2 The Current State

A majority of healthcare systems around the world are under a government regulated socialized medical system which comes with a universal health record. The healthcare system in the United states is privatized, therefore the transition to EHR came with individual health entities purchasing a multitude of different EHRs. The problem comes in that a patient presenting to two different healthcare facilities, even if across the street or within the same building, will have two different medical charts that do no communicate with one another. The other problem comes with accessing this information. The two largest companies Epic and Cerner have a commercial interest, with a primary goal to increase revenue to the shareholder. It is exceedingly difficult for the nonprofit entities including academic centers and hospitals to access the patient information within the EHR. There is tremendous potential within the EHR. Beyond data collection, storage, data retrieval, and analysis, we should move towards real time guidance and guidelines for medical decision making to improve health.

4 KNOWLEDGE

Only 10-20 years ago, Hospital libraries and medical school libraries were once filled with books and journal articles. If a healthcare practitioner wanted information relevant to clinical care, they went to libraries to pour through the resources with exhaustive efforts. Today, those libraries are mostly void of books. Almost every individual in western medicine has access to a computer, and usually to a handheld device, capable of accessing far more information than could ever be stored in a library. There are massive information sources, such as PubMed, a gigantic repository of journal articles and books that is constantly being updated with new information. And Up To Date, a point of care medical reference commonly used on a handheld deceive, with evidence based clinical guidelines contributed by over 5,000 physicians [9]. The massive amount of data now accessible to most healthcare providers and scientists is changing healthcare rapidly. Still, there is much room for improvement as care is commonly delivered based on anecdotal evidence, and cost and quality should continue to improve.

5 NEXT GENERATION SEQUENCING

5.1 The Human Genome

The first human genome was sequenced in 2003[2]. This colossal global effort took over 10 years and thousands of scientists working at great expense. In the end, a private and public group collectively sequenced the first genome. Initially, the technology was extremely expensive and took great deal of time. Through technological advancements including sequencing cores and big data, the cost of the genome has plummeted. The 1000-dollar genome project is an attempt to make sequencing more affordable [4]. We are a long way away form being able to utilize the genome to deliver care. Bioinformatics expertise has lagged behind technology. Groups still do not agree on a standard way to process the information. Still this technology improves papidly, and recently a group published 24-hour genome sequencing for intended us in clinical decision making. Soon it may be a reality for physicians to utilize genomic information, whether about drug susceptibility, or prognosis, to guide medical care.

5.2 Beyond DNA

Initial estimates placed the number of genes at ¿100,000 [1]. Looking at the massive amount of diversity and the billions of unique human beings on this earth, this was a appropriate estimate. The current number is estimated somewhere around 20,000. The question is what accounts for the rest of phenotypic diversity and disease. The human genome project utilized whole exome sequencing. Whole exome sequencing involves sequencing the entire coding region, or exome, of the genome. This consists of around 20,000 genes and over 30 million nucleotides. The exome, though massive, consists of only 1% of the total genomic DNA. Many genetic diseases involve alteration of this coding exome but we are discovering that many diseases are due to problems outside of this coding region. Sequencing only 1% of the genomic material is a fraction of the time, cost, and burden of analysis, compared with whole genome sequencing, but we must move towards whole genome sequencing to capture all disease states. We have also come to realize that splicing and other post transactional regulation introduces much diversity. We have the technology to sequence the entire RNA transcriptome and the proteome as well. This produces a data set which dwarfs the genome and genomic DNA sequence information. These technologies are currently only utilized in the research setting. Despite our advanced technology, we have very little idea of how to interpret the data in a clinical setting. Again the bioinformatics expertise lags behind. There is amazing potential to advance knowledge and study human disease and a tremendous amount of big data analytics along the way.

6 WEARABLE TECHNOLOGY, NURITION AND WELLNESS APPS

Massive data sets exist, collected by insurance companies, in electronic health records, by pharmaceutical companies and by research institutions. There is another very exciting source of big data on the horizon, in personal wearable technologies, and also fitness, wellness and nutrition apps [4]. Individuals wearing FitBits, with fitness apps on their mobile devices, wearing smartwatches, etc. can track health and wellness measures in ways that once required inpatient hospital monitoring and sophisticated research lab settings. They track sleep and activity throughout the day and night. In addition, there are countless apps which track nutrition and health. People log meals and nutrition to keep accountable. Often these apps work with time tested and well researched diets including weight watchers, etc. This technology has already changed the way many individuals look at health and wellness. This exciting new dataset has great potential to advance human health and improve disease that may be the root cause of our healthcare epidemic.

7 TELEMEDICINE

Telemedicine involves a virtual visit between a physician and patient [6]. There are obvious benefits, especially when a patient population is spread across a wide geographic space either due to a high level of physician specialization, or a rural patient population. Highly specialized, but critical subspecialists are often in great shortage. This places a great burden on the available providers, with often unsustainable schedules. Video technology allows doctors, nurses and practitioners to visualize patients, perform a limited

physical, and to communicate with individuals at a distance. There is great potential to improve cost and reduce burden. There are limitations. Many physician specialists are values for their technical, hands on skills. Telemedicine is not much of a help, the technical procedures, such as inserting airways into the trachea of small babies, and insert central arterial lines into major vessels to deliver lifesaving medications, require hands on skills. The same goes for surgeons and other highly skilled technical professions. Interventional techniques and robotics are increasingly being used to perform procedures, but while these operations are performed, a surgeon needs to very close, in case unforeseen accidents problems necessitate a conventional correction. Procedural specialties are the greatest expense to our healthcare system and their procedural skills are a long way from being performed through telemedicine or robotics.

8 SOCIAL MEDIA

One interesting trend is the multitude of health information shared over social media networks. Blogs, columns, and posts providing information about nutrition and wellness, news stories, and information sharing. The story reporting googlefis flu prediction trends ahead of the CDC, based on search history, spread virally over facebook [5]. The field will continue to expand. wonderfully

9 PERSONALIZED MEDICINE

Wikipedia summarized personalized medicine as: "a medical procedurethat separates patients into different groupsfi!?with medical decisions, practices, interventions and/or products being tailored to the individual patient based on their predicted response or risk of disease." [8] In a way the culmination of big data and health is with personalized medicine. In a hopefully not so distant future the electronic health record, pharmaceutical data and genomic data will provide a more tailored, affordable, and high-quality approach to healthcare. Hopefully healthcare will catch up with financial and ecommerce and in their ability to harness big data for good.

10 CONCLUSION

This paper highlights just a handful of technology driven big data solutions to our healthcare crisis. As Congress debates legislature to face this crisis, big data more harmoniously moves towards solutions. Better health without economic ruin is a reality and big data will play a major role. Much work is left to be done

ACKNOWLEDGMENTS

Thank you to Dr. Geoffrey Fox, Gregor von Laszewski, and all of the course instructors for an excellent introduction to Big Data and Data Science.

REFERENCES

[1] [n. d.]. ([n. d.]). Vanderbilt University: Introduction to Bioinformatics Course Lectures.

[2] Francis S Collins, Michael Morgan, and Aristides Patrinos. 2003. The Human Genome Project: lessons from large-scale biology. *Science* 300, 5617 (2003), 286–290.

[3] Centers for Medicare & Medicaid Services et al. 2014. National health expenditures 2012 highlights. *Online verfügbar unter http://www.cms. gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/National-HealthExpendData/Downloads/highlights. pdf* (2014).

[4] Geoffrey Fox. [n. d.]. Unit 6 Lectures. ([n. d.]).

[5] Jeremy Ginsberg, Matthew H Mohebbi, Rajan S Patel, Lynnette Brammer, Mark S Smolinski, and Larry Brilliant. 2009. Detecting influenza epidemics using search engine query data. *Nature* 457, 7232 (2009), 1012–1014.

[6] Maria Hernandez, Nayla Hojman, Candace Sadorra, Madan Dharmar, Thomas S Nesbitt, Rebecca Litman, and James P Marcin. 2016. Pediatric critical care telemedicine program: A single institution review. *Telemedicine and e-Health* 22, 1 (2016), 51–55.

[7] Erik WJ Kokkonen, Scott A Davis, Hsien-Chang Lin, Tushar S Dabade, Steven R Feldman, and Alan B Fleischer. 2013. Use of electronic medical records differs by specialty and office settings. *Journal of the American Medical Informatics Association* 20, e1 (2013), e33–e38.

[8] Wikipedia. [n. d.]. Personalized Medicine. ([n. d.]). https://en.wikipedia.org/wiki/Personalized_medicine

[9] Wikipedia. [n. d.]. UpToDate. ([n. d.]). https://en.wikipedia.org/wiki/UpToDate Wikipedia: The Free Encyclopedia. Wikimedia Foundation, Inc. 22 July 2004. Web. 2 Sept. 2016.

11 BIBTEX ISSUES

Warning–no key, author in vanderbilt

Warning–no author, editor, organization, or key in vanderbilt

Warning–to sort, need author or key in vanderbilt

Warning–no key, author in vanderbilt

Warning–no key, author in vanderbilt

Warning–no key, author in vanderbilt

Warning–no author, editor, organization, or key in vanderbilt

Warning–empty author in vanderbilt

Warning–empty year in vanderbilt

Warning–no number and no volume in centers2014national

Warning–page numbers missing in both pages and numpages fields in centers2014national

Warning–empty year in fox6

Warning–empty year in wiki-personalized

Warning–empty year in wiki-uptodate

(There were 14 warnings)

12 ISSUES

DONE:

Example of done item: Once you fix an item, change TODO to DONE

12.1 Assignment Submission Issues

Do not make changes to your paper during grading, when your repository should be frozen.

12.2 Uncaught Bibliography Errors

Missing bibliography ffile generated by JabRef

Bibtex labels cannot have any spaces, _ or & in it

Citations in text showing as [?]: this means either your report.bib is not up-to-date or there is a spelling error in the label of the item you want to cite, either in report.bib or in report.tex

12.3 Formatting

Incorrect number of keywords or HID and i523 not included in the keywords

Other formatting issues

12.4 Writing Errors

Errors in title, e.g. capitalization

Spelling errors

Are you using *a* and *the* properly?

Do not use phrases such as *shown in the Figure below.* Instead, use *as shown in Figure 3*, when referring to the 3rd ffigure

Do not use the word *I* instead use *we* even if you are the sole author

Do not use the phrase *In this paper/report we show* instead use *We show.* It is not important if this is a paper or a report and does not need to be mentioned

If you want to say *and* do not use *&* but use the word *and*

Use a space after . , :

When using a section command, the section title is not written in all-caps as format does this for you

`\section{Introduction} and NOT \section{INTRODUCTION}`

12.5 Citation Issues and Plagiarism

It is your responsibility to make sure no plagiarism occurs. The instructions and resources were given in the class

Claims made without citations provided

Need to paraphrase long quotations (whole sentences or longer)

Need to quote directly cited material

12.6 Latex Errors

Erroneous use of quotation marks, i.e. use “quotes” , instead of ” ”

To emphasize a word, use *emphasize* and not “quote”

When using the characters & # % _ put a backslash before them so that they show up correctly

Pasting and copying from the Web often results in non-ASCII characters to be used in your text, please remove them and replace accordingly. This is the case for quotes, dashes and all the other special characters.

12.7 Structural Issues

Acknowledgement section missing

Incorrect README file

In case of a class and if you do a multi-author paper, you need to add an appendix describing who did what in the paper

The paper has less than 2 pages of text, i.e. excluding images, tables and figures

The paper has more than 6 pages of text, i.e. excluding images, tables and figures

Do not artifficially inffate your paper if you are below the page limit

12.8 Details about the Figures and Tables

Capitalization errors in referring to captions, e.g. Figure 1, Table 2

Do use *label* and *ref* to automatically create figure numbers

Wrong placement of figure caption. They should be on the bottom of the figure

Wrong placement of table caption. They should be on the top of the table

Images submitted incorrectly. They should be in native format, e.g. .graffle, .pptx, .png, .jpg

Do not submit eps images. Instead, convert them to PDF

The image files must be in a single directory named ”images”

In case there is a powerpoint in the submission, the image must be exported as PDF

Make the ffigures large enough so we can read the details. If needed make the ffigure over two columns

Do not worry about the ffigure placement if they are at a different location than you think. Figures are allowed to ffoat. For this class, you should place all ffigures at the end of the report.

In case you copied a ffigure from another paper you need to ask for copyright permission. In case of a class paper, you must include a reference to the original in the caption

Remove any ffigure that is not referred to explicitly in the text (As shown in Figure ..)

Do not use textwidth as a parameter for includegraphics

Figures should be reasonably sized and often you just need to add columnwidth

e.g.

`/includegraphics[width=\columnwidth]{images/myimage.pdf}`