Big Data: Case Studies in Healthcare

Joy Carol Ming
Global Health and Health Policy Secondary Candidate

May 2015

1 Introduction

The internet has become so intertwined with our daily lives that most of us interact with the internet on at least an hourly basis, and in some cases, on a basis with a timescale even smaller than that. There are currently over three billion internet users in the world (World Wide Web Consortium, 2015). And these users are constantly interacting with the internet, creating and receiving various types of content. In the realm of the internet, on average, every minute there are four million Google search queries, two and a half million pieces of content shared on Facebook, three hundred thousand tweets, and over two hundred million emails sent (Gunelius, 2014). This is a ridiculous amount of information that is being interacted with. This explosion of internet usage has echoed as a wave of content, of information, of data being generated, shared, and accessed all across the depths and breadths of cyberspace.

What makes this multitude of data unique in this day and age more so than it was before is its novel increase in volume, velocity, and variety (McAfee and Brynjolfsson, 2012). In this case, the *volume*, or amount of data, has increased because of the ability to easily generate data in terms active creation text posts through the social media sites mentioned above or passive recording of information from various distributed sensors, including the powerful mobile devices that are carried around daily. Furthermore, there is increased accessibility of this data in terms of it being shared widely on the internet or stored in large repositories in the cloud. Another note in terms of the data creation and accessibility is the *velocity*, or

the speed at which data travels, looking at enhanced pathways of information traveling at seconds, from its origin to its reception. Finally, it is the *variety* of data that makes this entire analysis extremely interesting. There are so many sources of data that are available, from the text, audio, image, and video streams to the hidden sensors of touch, click, and movement that can be recorded by the input devices of computers, mobile phones, tablets, and even wearables like watches and wristbands in the Internet of Things (Lohr, 2012).

This phenomena of *big data*, or exploring the tons of information that is available, has caught the attention of many in diverse fields. This has resonated in business intelligence, in better understanding what types of information about customers and products are available, allowing for better estimated times of arrival for airlines or speedier and more personalized promotions (McAfee and Brynjolfsson, 2012). In the public sphere, analysis of blog postings, Congressional speeches and press releases, and news articles in an automated fashion can uncover insights into how political ideas spread (Lohr, 2012). There are other examples of big data being used for social impact problems, including poverty, crime, and pollution (Lohr, 2012). More popular culture examples of big data include "Moneyball," using statistical analysis to explore the world of betting on sports (Lewis, 2004).

This is also picking up in the field of healthcare. My thesis topic of analyzing both online medical forum data and electronic health records is an example the use of big data in that I was dealing with on the orders of hundreds of thousands of posts from many different sources (Ming, 2015). This was especially interesting because it required the implementation of advanced and specialized methods to handle large amounts of highly complicated, multifaceted data that has not been otherwise explored. For example, at first, the algorithms I was using were not able to process all of the data even after running for over 72 hours. Therefore, my algorithms had to be optimized to become more tractable and complete in a reasonable amount of time. Additionally, I had to alter the algorithm to deal specifically with the type of data that was used, or the fact that the spread across medically-related concepts was fairly sparse. And, through my thesis, I was able to collect, organize, and analyze online

medical forums and electronic health records in a way that could change the way that autism is understood.

Of course, there are many other applications of big data in healthcare beyond my thesis topic. This extra chapter continues this discussion of big data in health care that was initiated with my thesis topic through case studies, looking at the contexts of these case studies with regards to the sources of data, the technology used, and the application areas that were highlighted. Then, a higher-level analysis of the role of big data in healthcare is discussed with regards to the ethics, challenges, policies, and potential impact. As more data is being created in healthcare, there are increasingly more opportunities to explore and understand new topics.

2 Case studies

2.1 Autism and Online Medical Forums

In this information age, there is a plethora of social data available on the Internet (Chou et al., 2009). However, this data is not being used to its fullest potential, especially by professionals. One example of such data is medical forums, communities where patients and their caregivers discuss symptoms, treatments, and other topics related to a given disorder (Denecke and Nejdl, 2009; Hawn, 2009; Sarasohn-Kahn, 2008). These forums present a wealth of information in looking at pharmacovigilance and adverse drug effects (Almenoff et al., 2005; Chee et al., 2011). Yet this information could be used for more.

Autism spectrum disorders (ASDs) include a variable presentation of neurodevelopmental disorders characterized by impairments in communication, reciprocal social interaction, and restricted behaviors or interests (Faras et al., 2010). Recent prevalence rates for ASDs are now estimated at about one in sixty-eight children in the United States (Baio, 2012). Symptoms for autism typically are apparent before age three and timely diagnosis is important in beginning early intervention (Baio, 2012). However, because ASD remains a diagnosis

that is defined on the basis of behavior, diagnostic assessment is complex (Lord et al., 2013). Furthermore, autism is fairly heterogeneous and has many etiologies, with a wide variety of manifestations of symptoms throughout the spectrum (Geschwind and Levitt, 2007). This means that the same diagnosis of ASDs could include patients with diverse comorbid disorders such as epilepsy or attention deficit hyperactivity disorder as well as different levels of social interaction, from nonverbal to high-functioning (Matson and Nebel-Schwalm, 2007).

Social media provides a unique outlet for characterizing the heterogeneity of autism. Unlike electronic health records, which is the status quo for understanding symptoms of disease and disorders, medical forums can record more detailed descriptions given by caregivers who spend time with autism patients, information that might not be present in clinical records (Frost and Massagli, 2008). In an attempt to make use of unstructured online information and compare it to electronic health records to determine more truths about the diversity of the symptoms of autism, my thesis used topic modeling to explore two different sources of autism data: online medical forums and electronic health records (Ming, 2015). While the thesis was mostly exploratory in terms of methods and data analysis, it seeds an interesting conversation on the role of *informal* information, such as the online health forums, and *formal* information, such as the electronic medical records.

2.1.1 Information source

The source of this information was collected from two areas, online medical records and electronic health records. More specifically, close to two hundred and fifty thousand posts were taken from five different and fairly popular autism forums, including ASD Friendly, ASD Forum, Autism Web, and Talk about Autism (Ming, 2015). In order to get these data sources, the online forums were scraped using the Python implementation of Beautiful Soup, parsing the HTML on the website to collect the specific posts on the page. On the electronic health record side, close to three hundred thousand posts were taken from the Boston Children Hospital electronic medical records for patients with at least one code for

ASD (299.*). This was then filtered with the Boston Children's Hospital informatics and natural processing using CTAKES, and filterend on using only positive concpets in records for patients that were at least 15 years old with at least 4 visits. The sheer volume of this data and the difficulty in dealing with it created the sentiment of big data.

This juxtaposition of online medical forums with electronic health records provided by my thesis brings up an interesting conversation with regards to informal versus formal information. Formal information is generally more structured, specifically looking at structured informational input through the use of billing codes. However, sometimes this is not necessarily the most structured as many different numbers could be correlated with the same disease or disorder, creating confusion in how the information is interpreted, as well as having too much stringent structure that does not allow for more free-form dealing with corner cases properly. On the other hand, extremely informal information systems like social media allow for too much unstructured text, leaving much room for spelling and grammar errors with abbreviations and internet lingo as well as not holding each individual accountable for the information that is inputted. Each of these provides different types of challenges in addition to the new perspectives that they may bring for understanding and analysis.

There are some interesting other applications of data mining and online medical forums that have been explored in the past. One of these is adverse drug side effects (Chee et al., 2011). This makes sense because despite having roots in chemistry, many times side effects are not easily predicted from one person to another. This is true for birth control, in that doctors cannot actually predict how a patient will react, rather, the reaction is more dependent on how the patient reacts to other similar medications (Landry, 2012). For that reason, a lot of women will turn to online medical forums to talk about their reactions to medication and hear how women who have interacted similarly to experienced drugs would interact with a new regimen. Another potential application includes pharmacovigilance, or the use of other spontaneous reporting databases that may signal possible adverse drug reactions (Almenoff et al., 2005).

2.1.2 Analytical methods

In this case, the technology that was used was topic modeling. Topic modeling builds upon the most basic means of analyzing a large corpus of text, which hold the *bag-of-words assumption*, or that the words can be split from each other and the order of the words ignored, as if they were thrown into a random bag (Wallach, 2006). Some of the basic approaches for looking at similar documents include sentiment analysis, which just looks at the count of the number of words and concepts found in the text (Pang and Lee, 2008).

Topic modeling has the assumption that text is generated based on a certain set of themes or topics. Topic models uncover the underlying semantic structure of a document collection based on hierarchical Bayesian analysis of the original texts, discovering patterns of word use and connecting documents that exhibit similar patterns (Blei and Lafferty, 2009). The generative basis of topic models is latent Dirichlet allocation, where each document is created by drawing a topic at random according to the distribution and selecting a word from that topic (Griffiths et al., 2004). More details of the topic model methods are detailed in Ming (2015), my thesis.

Other extensions from topic models include using different assumptions beyond latent Dirichlet allocation, such as probablistic latent semantic analysis (Pritchard et al., 2000). Even more approaches take into account the structure of the text itself, beyond the bag-of-words assumption, especially through the use of neural networks. This is especially salient in the use of deep learning, an especially powerful form of neural networks that is able to classify more information due to the large amount of data, or big data (Ngiam et al., 2011).

2.1.3 Application area

This specifically looks at autism spectrum disorder, a disorder that has constantly held a strong sense of enigma. This disorder was selected first and foremost because of its increasing awareness, more press that has built to the point that autism has built an "epidemic of panic" with the number of diagnoses (Konner, 2011). Another important aspect of autism is its

pervasive uncertainty, where attempts at hormonal, genetic, neurological, and environmental explanations of autism resulted in the declaration that it is "time to give up on a single explanation of autism" with the realization that there will be no single genetic or cognitive cause for the diverse symptoms of autism (Happé et al., 2006). Finally, autism is based mainly on a behavioral diagnosis of the Diagnostic and Statistical Manual of Mental Disorders (DSM) which makes it extremely volatile and difficult to pinpoint during the extremely early critical period (American Psychiatric Association and others, 2013).

These are all factors that make autism an interesting application area to explore in terms of looking at both informal and formal information sources. The thesis project aimed to better understand this enigmatic disorder, harnessing the increased awareness and discussion, specifically in relation to the parents, the caregivers that are interacting with children with autism. Taking this dialogue around autism to better understand the behaviors associated with it can contribute to another facet to the behavior-based diagnosis process or explore insights into the understanding of treatments. Not only will this approach leverage the existing trends in the increasing prevalence and discussion, it also explores a new dimension into an age-old problem.

2.2 Google Flu Trends

Influenza is a contagious respiratory illness caused by influenza viruses A and B that cause mild to severe illness–serious outcomes can result in hospitalization or death, especially for those at high risk such as older people, young children, and people with certain health conditions (Centers for Disease Control and Prevention, 2014). History has shown the flu to be be a major concern in the past epidemics in various locations with various strains. Season influenza epidemics are a major public health concern, causing tens of millions of respiratory illnesses and two hundred and fifty thousand to five hundred thousand deaths worldwide each year (World Health Organization, 2003). Mutations of the strains can create extremely potent symptoms and manifestations, including the most recent 2009 H1N1 viruses (Clark

and Lynch 3rd, 2011). The best way to prevent the flu is by getting vaccinated each year as the strains mutate fairly quickly (Centers for Disease Control and Prevention, 2014).

Another important aspect of flu is its epidemiology, or an understanding its spread. The Centers for Disease Control and Prevention (CDC) track severity of the flu principcally through its national Influenza Surveillance System that monitors key indicators like the percentage of deaths resulting from pneumonia or influenza, rates of influenza-associated hospitalizations, pediatric deaths, and the percentage of visits to outpatient clinics for influenza-like ilness (Centers for Disease Control and Prevention, 2014). This surveillance data is augmented with the use of mathematical models to fill in the picture of the disease burden and the impact of the influenza immunization programs.

However, the Centers for Disease Control and Prevention data is not necessarily comprehensive in understanding the spread of the disease because it is reliant on surveillance, which has some innate lag. Ginsberg et al. (2009) attempted to use an interesting approach that explores a new set of data as a proxy for flu spread in the creation of Google Flu Trends. This project analyzes what they call "health-seeking behavior" in terms of search queries to online search engines to track influenza-like illnesses in a population. The authors find that the frequency of certain queries is highly correlated with the percentage of physician visits in which a patient presents influenza-like symptoms. The model that is developed is relatively effective, accurately estimating the current level of weekly influenza activity in each region of the United States with a reporting lag of about one day (Carneiro and Mylonakis, 2009).

This works because there is an extremely large amount of information, with millions of users around the world are contributing to this data source in a way that benefits them and they therefore do not necessarily have any incentive to input data. Furthermore, this new proxy seems to do an adequate approach in that it is able to achieve a slight advantage over the Centers for Disease Control and Prevention estimates, though the actual worth of these predictions are debated. Some claim that the data exhibits big data hubris (Lazer et al., 2014). Others have improved upon the models and methods provided in the original

paper in order to create more accurate predictions (Yang et al., 2015). However, in the end, if integrated with public health practitioners and tools, this type of unique and innovative strategy can be one step closer to true real-time outbreak surveillance (Carneiro and Mylonakis, 2009).

2.2.1 Information source

What makes this project extremely unique is the use of passive information collection, or tracking and utilizing information from a process or action that people already engage in. This is interesting because there is less of a need for explicit incentivization, or having people respond to some prompting of information input, which makes collecting and generating the information often the most difficult step. Getting this information effectively "for free" lessens the burden on the research in that regard. All the researchers have to do is figure out exactly which questions to ask in order to determine which data to analyze and how to explore the data.

However, while this idea is quite clever, there are some weaknesses to this approach. The main problem is that the connection between Googling symptoms and actual manifestation of the flu can be slightly tenuous. For example, it is hard to fully adjust for people who are searching for symptoms even though they do not have the flu or people who have the flu and do not search for the symptoms. With regards to the latter concern, there is the question of internet connectivity, whether this is due to digital literacy or simple internet access, this may affect the types of populations that are looking for flu symptoms and therefore the groups of people that are recorded in these estimations. Furthermore,

The presentation of the Google Flu Trends project has sparked a lot of debate with regards to the power of big data. The service not only wildly overestimated the number of flu cases in the United States in the 2012-13 flu season and has consistently overshot in the last few years (Lohr, 2014). While this false positive is slightly less problematic than the other options of false negatives, it still demonstrates the difficulty of managing this type

of loosely connected data. One of the stronger criticisms is against this idea of "big data hubris," or the assumption that big data sets trump traditional data collection and analysis (Lazer et al., 2014).

2.2.2 Analytical methods

The methods that the original authors utilize for their analysis mainly involve a regressive model. They first find the set of queries in an unsupervised manner with no previous knowledge of influenza, selecting from fifty million candidate queries separated to produce a list of highest scoring queries. Then these queries are parsed to exclude marginally seasonal terms such as 'high school basketball' which coincide by time but not by symptom. These are used to create a linear model fitting the percentages reported by the Centers for Disease Control and Prevention and then validated on previously untestsed data (Ginsberg et al., 2009).

In line with the idea of "big data hubris" is the discovery that often times it is necessary to use more than one source of data, specifically with regards to the data that has been mined or collected from informal sources (Lohr, 2014). Yang et al. (2015) has found that utilizing an ensemble of methods as well as a variety of different data sources to add even more dimensions onto the data analysis. They present a self-correcting, robust, and scalable autoregression that uses publicly available online search data, which is effectively lower quality than the data used in the original Google Flu Trends approach but performs better and is able to incorporate the seasonability in influenza epidemics but also captures the changes in people's online search behavior over time.

Looking at these different approaches, it is important to note the potential for overfitting, which is creating a model that is too well-adjusted to a given set of data and does not perform well in a new or slightly different environment. Some of the warning signs for overfitting include models that perform suspiciously well, meaning that lower performances may sometimes be conversely better in light of robustness. There are machine learning and statistical methods to account for overfitting, including rigorous cross-validation and

bootstrapping, or using parts of the data as test sets in addition to those used as training sets (Kohavi et al., 1995). There are also robustness checks that can determine how fluctuations and changes in data can influence the way the model performs. However, the integration of multiple modes of data, as seen in the case of Yang et al. (2015) can potentially help with this possibility of overfitting. However, while more validation must be done, there is a limited amount of data that can be used as the ground truth in the cases of epidemiology because it is hard to track in the first place.

Additional methods that could add layers onto the analysis of this data could involve visualization of this spread. Currently, the results look mainly at the potential incidence of flu at a given time. Having a better understanding of potentially how these trends move over time would be interesting for the predictive side of the analysis for future flu trends. Moreover, exposing this type of analysis to a tool for public health officials and researchers to use, potentially to allow them to visualize the changes over time on a geographic scale, would be an interesting integration of human and computer interaction to solve a fairly difficult problem from many different angles at a time.

2.2.3 Application area

This case study looks specifically at a more epidemiological impact of big data in healthcare. In this case, this is especially fitting, because the Google search queries are able to determine the frequency of the searches, which is the primary metric with regards to flu incidence. Additionally, there is some reasonable ability to determine the location of each of the queries, as this is one of the parameters that Google tracks with each of the searches conducted by the users. This combination of information is especially important with this specific problem.

These parameters of frequency and location are interesting in flu. While the actual biology of the disease is fairly straightforward in that much of it is understood and that new trends can be quickly studied in the lab setting, what makes influenza difficult to manage and understand is its spread. Even though vaccines can be developed that somewhat predict

the strains in a given season, unless they are taken by the population, the spread of flu cannot be fully handled (Fiore et al., 2008). Furthermore, there are more complicated forces at play, including the highly contagious nature of flu as well as ideas such as herd immunity (Piedra et al., 2005). Influenza can only evolve into an epidemic when the contagious nature has developed past a point of control by all efforts from public health officials (Glezen, 1982). Therefore, epidemiology and spread of the flu are some of the more pertinent aspects of the disease and having some sense of understanding or predictability attained by this method can help in the control of this disease.

This makes this application especially suited for the methods that are described above. The Google Flu Trends platform has since been abstracted into a Google Trends platform, or a timely, robust, and sensitive surveillance system for epidemics and diseases with high prevalances in developed countries or those with large populations of Web search users (Carneiro and Mylonakis, 2009). A similar approach has been used on dengue, also a highly contagious disease that has a large amount of spread (Gubler and Clark, 1995). Perhaps in the future this could also be applied to other contagious diseases that have specific symptoms that can serve as anchor words to alert to a manifestation of a given disease. Another consideration is that the disease has to have a wide enough population of infection that these levels of search queries will register a large enough volume to create a marginal difference that influences the information count.

Furthermore, other future efforts could include the integration of other forms of social media data, such as Twitter. Some researchers have pitched the idea of using Twitter for influenza surveillance to account for the concerns of replicability, overfiting, construct validity, granularity, and temportal confounds (Broniatowski et al., 2014). Other similar arguments could be made for more general sources of information like Facebook or increasingly specific information sources such as WebMD or other medical search engines.

3 Discussion

Big data is becoming an increasingly relevant topic for many different types of applications and healthcare is not by any means excluded. There are many examples of big data being used for healthcare in the current sphere, varying across the different information sources, analytical methods, and application areas.

The two case studies present examples of ways in which large amounts of information, or big data, can be useful in some aspect of public health, whether it is to better understand the symptoms of an otherwise engimatic disorder like autism or to better track and visualize the spread of a highly contagious and potentially fatal disease like influenza. Of course, these are not the only examples of big data in healthcare. There are other instances of big data being used, from using similar data sources such as online medical forums to understand new application areas, such as adverse drug effects, as well as using new data sources, such as Twitter to understand similar application areas, such as epidemiology of influenza and other contagious diseases. The work in big data and healthcare is currently developing and gaining momentum.

As technology progresses, there will be even more instances of bigger and bigger data. People will be using more specialized searching options and more specific data input options such as personally controlled online health data, increasing agency in their health records (Steinbrook, 2008). More importantly, there will be sensors everywhere, from the mobile devices in pockets to wearables to items in the household that are all part of the Internet of Things. Even though there will be an increase of data, there will still be many challenges to the full integration of big data in healthcare. On the flip side, there is also the possibility for many impacts. These challenges and impacts are described in the section below.

3.1 Challenges

3.1.1 Internet connectivity

Even though a lot of what is considered big data is generated by online interactions such as posts on social media, there are other realms of big data that use different types of interactions, especially on a global scale. Internet access is not completely equitable all over the world and over sixty percent of the world's population still does not have access to the internet (Tharoor, 2014). A lot of this inequality is seen in the differences between urban and rural areas and can be extended into distinguishing between developed and developing countries.

One of the biggest problems is making sure that the populations without internet access still are able to have their voices heard, are still able to be counted in terms of their instances of diseases as well. It is easy to look only at the averages, to cover up the specific minority populations. And it is even easier to look only at the existing data and forget about the negative space, or the instances of missing data.

Of course, the most long-lasting solution to this problem is to increase technological access. This would have many positive externalities in addition to allow for more expansive data collection across all populations. However, in light of the difficulty in doing so, some other methods could serve as intermediate steps. There are some active data collection methods such as phone calls or text messages that still leverage technology but more accessible technology. This has been used in both developed and developing countries in the field of mHealth or mobile health in data collection and analysis (Eysenbach and Group, 2011). Other more passive methods of data collection also should be explored in potentially looking at purchase patterns of flu-related supplements in local drug stores as a potential proxy for the spread of flu.

3.1.2 Methodological capability

The previous case studies explore two of the basic methods, including topic modeling and regression analysis. In the discussion of these case studies, other methods are mentioned, including sentiment analysis and deep learning. These are all examples of data science and machine learning approaches to analyzing large amounts of text data. There are even more methods in natural language processing such as clustering or categorization such as part-of-speech tagging (Collobert and Weston, 2008).

However, understanding text is an extremely difficult problem. While generating text with a human stringing sentences together based on certain rules of grammar, trying to derive those rules and the meanings using a computer is a very difficult an unsolved problem. It is even harder to get the computer to then do even more with this data, whether it is to categorize it with other input data or to classify it based on certain innate properties (Spyns, 1996). There is still much to be done in terms of technology and the methods that are used to analyze the large amounts of data.

In the field of healthcare, it is important to achieve some level of correctness, however that may be defined. As discussed earlier on, having false positives might be slightly excusable because that means that the patient would falsely believe they have some disease or disorder that they would then seek further diagnosis or treatment for. This could be an issue if the patient is able to act outside of the formalized medical sector such as with home remedies that can have deleterious side effects. However, having a false negative, or having a patient not realize that they had a disease or disorder could be even more disastrous. As of the current state of technology, this level of correctness has not been able to be achieved yet.

3.1.3 Technological acceptance

In addition to infrastructural barriers to the use of big data, there are some mental blocks to the use of big data. There is the technological understanding on the side of the patients or users. Even looking at the example of the Google queries, some patients are less likely to

be the users that input Google search queries for their flu-like symptoms, opting for more traditional methods of phone calls or doctors visits. There is a spectrum of technological acceptance, on the other end, where there are users who are prolific Google query makers.

This barrier increases with even more technological interaction necessary. Even less users are comfortable using technology with regards to inputting or submitting information on Twitter or Facebook, specifically health-related or other sensitive information. And with the development of the Internet of Things, there are patients who are less likely to download an app, carry around a mobile device, or wear something like a watch or wristband that will be tracking all of their vital and health information.

However, an even more insurmountable difficulty is with regards to the clinicians and researchers themselves. There is overwhelming desire to continue using entrenched systems that are not quite suited for the novel and innovative techniques of big data. People are generally used to the way things are done and often times the initial, clunky iterations of electronic health records are difficult to transition to, resulting in resistance on that front. This user experience and lack of usability is both a mindset and a technological design problem. Moreover, there is a need for the change in the education and the mindset in how healthcare is run in terms of more software that will enhance but not replace the role of the doctors.

In addition, there is a paradoxical cycle of not trusting the technology enough to rely on it for some healthcare decision making. Since the technology and methods are not entirely perfect yet, it is difficult to put all trust and responsibility on a computer program or mathematical model. However, it will be difficult to develop more reliable systems without demonstrated demand or understanding in part of the clinicians. Some solutions to this would involve more communication and information dissemination between medical professionals and statisticians, mathematicians, and computer scientists. An even more tractable solution to this problem would be for individuals with these backgrounds to work at least with collaborators in the other fields and at most to make a career doing work in the other

fields. This cross-pollination is important for the success of big data and its role in healthcare.

3.1.4 Privacy and security

One of the biggest debates with regards to information is privacy. In many of the examples above, the information is being collected without the explicit knowledge of the individuals. For example, when users are on Google looking up flu-like symptoms, they rarely think about the fact that they are contributing to the data that is collected by the search engine about them and their searching habits. While this is effective on the side of information collection, this may potentially be problematic from the perspective of the user. It is possible to take seemingly diverse and disparate information and determine a user's identity or at least to learn a lot about the user that was not explicitly stated.

The ethics of using this information is debatable. More policy, guidelines, and regulations need to be developed to match up with the novel data sources and the newly developed and implemented methods for approaching this big data. In the meantime, the current standards of ethics should still persist. Similar levels of respect for the privacy and information of the individual should be upheld and information should not be attempted to be uncovered unless necessary for analysis. This is especially important given the increasingly powerful methods for information exploration through data mining and which increases the potential ability to accidentally uncover more information about a user or patient than is needed.

Another important and related point is with regards to security of information. Since there is now more sensitive information stored and available in large amounts, security needs to be increased for all static stores of information as well as the fast-moving information streams. This requires more technological developments in robustness of the security systems. This also requires more awareness of the need for this security and for the potential problems or gaps in the security system that might be exposed.

These are both very important aspects of big data and are especially tantamount in the realm of healthcare. The data that is being dealt with can often result in a life or death situa-

tion for the patients. Furthermore, health is especially sensitive and identifiable information. Therefore, the highest precautions must be taken when working with this information. In order to understand and develop approaches to these problems, it is important to build teams of multidisciplinary individuals, matching the best of the privacy and security engineers with the brightest in healthcare information input and analysis.

3.2 Impact

3.2.1 Better understanding of disease

As explored in the idea of the online medical forums to inform more about the diagnosis and treatment of autism, big data has the potential to provide better understanding of disease. Big data allows for the integration of many different types of information, informal and formal, that can paint an even more comprehensive picture of the disease overall by engaging the conversations that are held in a clinical office as well as in the public sphere.

It is important to integrate these novel perspectives in a way that cannot be done with just a few data points. Rather, it is the volume, the large amounts of data that create more legitimacy for the data sources. It is the idea of the wisdom of the crowd, that one opinion could not stand alone, but put into context with other opinions and data points, echoed and supported by all the other individuals in the online space creates more potential for a given statement (Kittur et al., 2007).

3.2.2 Population healthcare

The idea of Google Flu Trends looks at big data and its impact on population healthcare. In the case of Google Flu Trends, the original study, Ginsberg et al. (2009), looks specifically at the spread of the flu in the United States, matching it up to the Centers for Disease Control and Prevention. Following iterations have been able to look at the spread of flu and even dengue on a global scale, which has a wider reach than intended due to the plethora of data, or big data.

This brings into light another benefit of having large amounts of data, specifically with regards to geographic data. Being able to visualize the spread and epidemiology is especially important for epidemiology and population health. By looking at the sheer volume and diversity of data, it only makes sense that there are trends, there are spreads over the geographic dimension as well as along time dimension, and more. And this is important in looking at healthcare at a macro-level, at the level of populations, of communities, of countries. Therefore, the volume of the data, specifically geographic data, can help inform efforts to control or understand epidemics.

3.2.3 Personalized healthcare

Another area of healthcare that could benefit from big data is personalized healthcare. As mentioned before with the increased Internet of Things, there are going to be more and more ways to collect information personal healthcare. For example, take a diabetes patient. Cameras can track the nutritional intake of an individual and accelerometers, gyroscopes, and global positioning service can track the movement of the person down to the number of steps taken and the change in elevation (Noronha et al., 2011). There are even plugins into the cell phone that can measure blood glucose levels (Patrick et al., 2008). Then, the cell phone can give more personalized information about what the individual should do in terms of diet and exercise to positively alter their blood glucose level.

Another facet of having large amounts of data is that there is information about all different types of data is its ability to, interestingly, become more personalized. Because there are just so many different types of data from different people, more clustering of different trends to create various profiles is possible. Creating more comprehensive models to account for different types of people is important to understand diversity of symptoms but also to allow for more personalized suggestions for healthcare. This type of attention is going to be a positive step towards how healthcare is delivered and experienced.

References

June Almenoff, Joseph M Tonning, A Lawrence Gould, Ana Szarfman, Manfred Hauben, Rita Ouellet-Hellstrom, Robert Ball, Ken Hornbuckle, Louisa Walsh, Chuen Yee, et al. Perspectives on the use of data mining in pharmacovigilance. *Drug safety*, 28(11):981–1007, 2005.

American Psychiatric Association and others. *Diagnostic and statistical manual of mental disorders, (DSM-5®)*. American Psychiatric Pub, 2013.

Jon Baio. Prevalence of autism spectrum disorders: Autism and developmental disabilities monitoring network, 14 sites, united states, 2008. morbidity and mortality weekly report. surveillance summaries. volume 61, number 3. *Centers for Disease Control and Prevention*, 2012.

David M Blei and John D Lafferty. Topic models. *Text mining: classification, clustering, and applications*, 10:71, 2009.

David Andre Broniatowski, Michael J Paul, and Mark Dredze. Twitter: Big data opportunities. *Science*, 345(6193):148, 2014.

Herman Anthony Carneiro and Eleftherios Mylonakis. Google trends: a web-based tool for real-time surveillance of disease outbreaks. *Clinical infectious diseases*, 49(10):1557–1564, 2009.

Centers for Disease Control and Prevention. Influenza (flu). 2014.

Brant W Chee, Richard Berlin, and Bruce Schatz. Predicting adverse drug events from personal health messages. In *AMIA Annual Symposium Proceedings*, volume 2011, page 217. American Medical Informatics Association, 2011.

Wen-Ying Sylvia Chou, Yvonne M Hunt, Ellen Burke Beckjord, Richard P Moser, and

Bradford W Hesse. Social media use in the united states: implications for health communication. *Journal of medical Internet research*, 11(4), 2009.

Nina M Clark and JP Lynch 3rd. Influenza: epidemiology, clinical features, therapy, and prevention. In *Seminars in respiratory and critical care medicine*, volume 32, pages 373–392, 2011.

Ronan Collobert and Jason Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167. ACM, 2008.

Kerstin Denecke and Wolfgang Nejdl. How valuable is medical social media data? content analysis of the medical web. *Information Sciences*, 179(12):1870–1880, 2009.

Gunther Eysenbach and CONSORT-EHEALTH Group. Consort-ehealth: improving and standardizing evaluation reports of web-based and mobile health interventions. *Journal of medical Internet research*, 13(4), 2011.

Hadeel Faras, Nahed Al Ateeqi, and Lee Tidmarsh. Autism spectrum disorders. *Annals of Saudi medicine*, 30(4):295, 2010.

Anthony E Fiore, David K Shay, Karen Broder, John K Iskander, Timothy M Uyeki, Gina Mootrey, Joseph S Bresee, and Nancy S Cox. Prevention and control of influenza: recommendations of the advisory committee on immunization practices (acip), 2008. *MMWR. Recommendations and reports: Morbidity and mortality weekly report. Recommendations and reports/Centers for Disease Control*, 57(RR-7):1–60, 2008.

Jeana H Frost and Michael P Massagli. Social uses of personal health information within patientslikeme, an online patient community: what can happen when patients have access to one another's data. *Journal of Medical Internet Research*, 10(3), 2008.

Daniel H Geschwind and Pat Levitt. Autism spectrum disorders: developmental disconnection syndromes. *Current opinion in neurobiology*, 17(1):103–111, 2007.

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

W PAUL Glezen. Serious morbidity and mortality associated with influenza epidemics. *Epidemiologic reviews*, 4:25–44, 1982.

Thomas L Griffiths, Mark Steyvers, David M Blei, and Joshua B Tenenbaum. Integrating topics and syntax. In *Advances in neural information processing systems*, pages 537–544, 2004.

Duane J Gubler and Gary G Clark. Dengue/dengue hemorrhagic fever: the emergence of a global health problem. *Emerging infectious diseases*, 1(2):55, 1995.

Susan Gunelius. The data explosion in 2014 minute by minute, infographic. 2014.

Francesca Happé, Angelica Ronald, and Robert Plomin. Time to give up on a single explanation for autism. *Nature neuroscience*, 9(10):1218–1220, 2006.

Carleen Hawn. Take two aspirin and tweet me in the morning: how twitter, facebook, and other social media are reshaping health care. *Health affairs*, 28(2):361–368, 2009.

Aniket Kittur, Ed Chi, B Pendleton, Bongwon Suh, and Todd Mytkowicz. Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie. *World wide web*, 1 (2):19, 2007.

Ron Kohavi et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, volume 14, pages 1137–1145, 1995.

Melvin Konner. Epidemiology: Epidemic of panic. *Nature*, 469(7331):468–469, 2011.

Lauren Landry. Sick of playing the hormonal lottery? one harvard student is helping women find birth control that fits. 2012.

David M Lazer, Ryan Kennedy, Gary King, and Alessandro Vespignani. The parable of google flu: traps in big data analysis. 2014.

Michael Lewis. *Moneyball: The art of winning an unfair game*. WW Norton & Company, 2004.

Steve Lohr. The age of big data. *New York Times*, 11, 2012.

Steve Lohr. Google flu trends: The limits of big data. *New York Times*, 2014.

Catherine Lord, Edwin H Cook, Bennett L Leventhal, and David G Amaral. Autism spectrum disorders. *Autism: The Science of Mental Health*, 28:217, 2013.

Johnny L Matson and Marie S Nebel-Schwalm. Comorbid psychopathology with autism spectrum disorder in children: An overview. *Research in developmental disabilities*, 28(4): 341–352, 2007.

Andrew McAfee and Erik Brynjolfsson. Big data: the management revolution. *Harvard business review*, (90):60–6, 2012.

Joy Carol Ming. Topic model analysis of multimodal autism data. 2015.

Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 689–696, 2011.

Jon Noronha, Eric Hysen, Haoqi Zhang, and Krzysztof Z Gajos. Platemate: crowdsourcing nutritional analysis from food photographs. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*, pages 1–12. ACM, 2011.

Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135, 2008.

Kevin Patrick, William G Griswold, Fred Raab, and Stephen S Intille. Health and the mobile phone. *American journal of preventive medicine*, 35(2):177–181, 2008.

Pedro A Piedra, Manjusha J Gaglani, Claudia A Kozinetz, Gayla Herschler, Mark Riggs, Melissa Griffith, Charles Fewlass, Matt Watts, Colin Hessel, Julie Cordova, et al. Herd immunity in adults against influenza-related illnesses with use of the trivalent-live attenuated influenza vaccine (caiv-t) in children. *Vaccine*, 23(13):1540–1548, 2005.

Jonathan K Pritchard, Matthew Stephens, and Peter Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959, 2000.

Jane Sarasohn-Kahn. *The wisdom of patients: Health care meets online social media*. California HealthCare Foundation Oakland, CA, 2008.

P Spyns. Natural language processing. *Methods of information in medicine*, 35(4):285–301, 1996.

Robert Steinbrook. Personally controlled online health data-the next big thing in medical care? *New England Journal of Medicine*, 358(16):1653, 2008.

Ishaan Tharoor. Worldviews map: The world without the internet. 2014.

Hanna M Wallach. Topic modeling: beyond bag-of-words. In *Proceedings of the 23rd international conference on Machine learning*, pages 977–984. ACM, 2006.

World Health Organization. Influenza fact sheet. 2003.

World Wide Web Consortium. Internet live statistics. 2015.

Shihao Yang, Mauricio Santillana, and SC Kou. Argo: a model for accurate estimation of influenza epidemics using google search data. *arXiv preprint arXiv:1505.00864*, 2015.