Using Big Data to Find Determinants of Health

Mithul S. Nallaka

University of Alabama

Using Big Data to Find Determinants of Health

Abstract

Big Data refers to the analysis of large data sources to find new information. Because this data analysis method is effective in finding obtuse associations, its implications in healthcare are far-reaching. By analyzing both public and private health data, computers may discover new indicators, or determinants, of health. Since personal health is impacted by a wide array of factors from genetics, to social and economic environments, Big Data offers a way to analyze this disparate information to discover new novel associations that may become new determinants of health. It is expected that using Big Data techniques on healthcare data will surmount meaningful discoveries in the future. Exploring this avenue involves applying Big Data techniques to existing health data sets and analyzing the rate and effectiveness that it finds existing and new determinants of health. While there remains the ethical question of allowing a computer to decide what it deems to be a determinant of health on its own, it is important to realize that this is rarely used in lieu of historical medical understanding, but rather as a supplement. In return, however, using Big Data techniques in the healthcare industry to revolutionize personal care.

Introduction

Problem Identification

In modern medicine, the discovery of new determinants of health, which are physical, social, or mental indications of the onset of a disease, is barred by the necessity of analyzing immense volumes of data.

Significance

As such, the applications of Big Data, which allows for the analysis of incredibly large data sets may lead to the discovery of new determinants that would not have been found conventionally. Theoretically, this may lead to more successful clinical outcomes as doctors will be able to use an amalgamation of patient data, genomics, and general trends to more easily predict possible health outcomes. In addition to more meaningful clinical outcomes, the use of Big Data may contribute socially. With meaningful and widespread implementation, the analysis provided by Big Data may even be used by consumers to more effectively understand their health. This in turn leads to more of the general public being aware of their health conditions and the steps that led them to that point (Herland et al., 2014). Also, Big Data may be used to analyze possible steps to recovery or health improvement for those who have health conditions.

Justification

As David W. Bates (2014) discusses in “Big Data In Health Care: Using Analytics To Identify And Manage High-Risk And High-Cost Patients,” Big Data represents a revolutionary technology that can be used to widely benefit the clinical community. This is possible because only Big Data can deal with “the types of insights that are likely to emerge from clinical analytics and the types of data needed to obtain such insights.” In “Predictive Analytics in Healthcare: Emerging Value and Risks,” it is found that the benefits of predictive analysis can be found in operational management, personal medicine, and even public health (Watson, 2019). Finally, in “Development of Big Data Predictive Analytics Model for Disease Prediction using Machine learning Technique” it is found that “big data analysis plays a crucial role to predict the future status of health and offers preeminent health outcomes to people" (Venkatesh, 2019).

One theory that suggests the efficacy of Big Data in healthcare is that it allows for analysis of the ever-growing set of medical data that would not have been previously possible. With the advent of high-level computing and the rise of efficient algorithms and machine learning, the analysis of Big Data is now possible. Because these algorithms can analyze sets of data that would take an eternity to understand conventionally, new processes or developments can be made at a rapid rate. These discoveries could be epidemiological, personal, or even societal, the bounds of the clinical applicability of Big Data are endless.

Literature Review

Introduction

In the ever-evolving field of modern healthcare, various aspects rise as being important to the future growth and success of the industry. This involves the discovery and understanding of health indicators as a means to more personal health, aggregating public and private data sets to offer meaningful opportunities for discovery, and the use of developing technologies such as AI and data mining to find meaningful associations in health data. “Healthcare changes dramatically because of technological developments” and the growing field of data analysis proves to be integral to the future development of health (Thimbleby, 2013). In understanding this possible growth, it is important to understand indicators of health, the discovery of novel associations in data, the importance of public and private health data aggregation, and how indicators of health are used.

Indicators of Health

In discussing indicators of health it is important to recognize that they are trends in various data forms that may lead to an indication of a future health outcome. For example, a

history of congenital heart disease is a general indicator of an individual's likeness to encounter heart disease (Hoffman, 2002). At the onset of modern medicine, there was a prevalence of mortality-rate based interpretation of health indicators, however, there is a growing awareness in the health industry that mortality based indicators severely lack the insight to be accurate or meaningful indicators of health (Hansluwka, 1985). In the height of the information era, the use of other methods to find indicators of health are dependent on large aggregate sources of public and private health data ranging from weight, height, lifestyle, to more sweeping epidemiological understanding.

To highlight the importance of indicators in modern health, the AJPH article "Health Information–Seeking Behaviors, Health Indicators, and Health Risks" writes that many individuals in the past few years actively seek out information that indicates their future health in an effort to become more health-aware (Weaver & Mays, 2011). This in turn leads to individuals privately using health indicators as a measure of their own health as opposed to solely relying on a medical professional. The discovery of new indicators of health thus shifts towards a more personal health care system allowing for individuals to be an integral part of their own health understanding.

Existing Data Analysis in Healthcare

The current implementation of data analysis in healthcare is in improving general decision-making. The healthcare industry already has implemented analysis of data from across the sector to aid in decision making, analyzing cost, and improving treatment decisions (Bjarnadottir & Agarwal, 2014). This implementation has been found to be successful in aiding current healthcare. The analysis of a range of hospital-related data has been proven to improve

both patient outcomes as well as minimizing error and cost. However, in its most relevant current implementation, this form of analysis is both limited and sometimes inadequate. As health becomes increasingly personal and individualized, the unavailability of personal information in this data aggregation results in a possible source of error in current methods. As global health becomes increasingly complex it is important to realize how data stored away imparts a large opportunity cost (Reynolds, 2012). While there is not a shortage of data, there is a consistent lack of proper usage, and as such possible discoveries are not found. As such, the previous statements stand to support the implications of data mining with the inclusion of both private and public health data. After proving essential in the hospital room, the application of data mining and machine learning in a broader health context may lead to significant improvements in general health. This will most effectively be done in discovering new indicators of health through the analysis of data trends in the general health of a population as well as individual characteristics such as medical history and general fitness.

Detecting Novel Associations in Health Data Sets

The aggregation of healthcare data is incredibly important to the general success of the industry. Healthcare organizations already amass large amounts of data on a daily basis and the use of this information can result in meaningful discoveries (Milovic, 2012). Milovic continues that these discoveries include, but are not limited to trends in patient behavior and condition. As the data collected in healthcare is voluminous and heterogenous, it requires an automatic method of understanding and interpreting the data.

Artificial intelligence (AI), and specifically data mining, have also proven useful in discovering novel associations of meaningful value that were not previously discovered in a

study using AI in analyzing diabetic data (Breault et al., 2002). In the cited study, Breault defines AI as the use of various self-learning algorithms that interpret sets of data. This involves the machines processing and analyzing data with given standards and learning from its previous iterations to continually improve itself. Furthermore, the implementation of data mining involved collecting an aggregate set of data and algorithmically separating and sorting the data to create a more effective set of data to analyze. In understanding the successful implementation of AI and data mining in a large application of a diabetic data warehouse it thus can then be applied to a large total aggregate of health data. An example of this novel discovery is exemplified in the article “Detecting Novel Associations in Large Data Sets". As written, the use of a family of statistical analysis known as maximal information-based nonparametric exploration (MINE) proved effective in analyzing indicators of public health from a WHO data set (Reshef & Finucane, 2011). MINE is a particular method of analyzing a data set that does not involve parameterizing the set to a known function. The data is not limited to being statistically represented as a linear, periodic, or even exponential function. Instead, MINE allows for statistical analysis of a data set that does not necessarily have obvious associations. The use of MINE for analyzing health data is effective as health data rarely fits a known statistical function. The effective use of MINE indicates that analyzing large heterogeneous sets of data is not only possible but relatively meaningful as a result of its versatility in searching for meaningful insight or discovery.

Aggregating Public and Private Health Data

The effective use of data mining in the future relies on combining its implementations in both private and public health. As stated previously the most effective means of implementation

thus far is an aggregation of public data and private data. In discussing the aspect of public health data analysis, it was found that academics and data analysts can, and have successfully, used implementations of data mining and machine learning to make an impact on the local and worldwide scale (Santos & Steiner, 2019). This prevalence shows that with available public health information, certain measures or understandings can be made about larger health trends.

However, the use of data mining is not strictly limited to the analysis of public health data and population trends. As written in SAI, machine learning in addition to personalized health information about a patient was successfully used to increase the reliability and accuracy of coronary disease diagnoses and prognosis (Miao, 2018). As supported by Miao and Breault, in addition to analyzing larger public trends such as diabetes and heart disease, machine learning, and data mining has proved their efficacy in analyzing personal and private data to aid diagnosis on an individual basis. This in turn supports the theory that aggregation of sweeping public health data as well as individualized private health data can support the current technological efforts to improve the understanding of health. As supported previously, this can work to find new trends that predict possible health outcomes on both a local and wide scale.

Methods

General Analysis

Analysis Method

Throughout the study of this topic, it will be found to be significantly qualitative. A close analysis of current data mining and processing methods can lead to a qualitative understanding of possible discoveries of determinants of health.

Independent and Dependent Variables

Throughout the experiment, different methods of data mining involving different data sets will be used to find possible determinants of health. With the main differentiation being the type of data mining and machine learning employed. The success of the data mining will then be assessed by how accurately it can associate trends in health data with known health concerns. This in turn will allow it to discover new trends that might also point to these health issues.

Population and Study Type

The particular population in question will be the general adult population of the United States. Due to the nature of the data being collected, the experiment is expected to be longitudinal. The resources required will be several computing machines, and the data required will be private and public health and epidemiological data.

Participants

The population for this study is known to be the wider adult population of the United States. As a result of data mining being far more effective on large data sets, a group as diverse as a large portion of the adult population in the US would offer the best possible data set to analyze with a machine learning model. As the study is not geographically or age-bound, the current study population is very representative of the general population. The sampling of health data from this population will be random to minimize the possibility of rediscovering existing commonalities or associations.

Data Collection and Sources

The data collected will be of available public health data involving geography, population, and epidemiological data. This data can involve CDC records, hospital records, and

any other publicly available information regarding overall patient health trends. This public information does not need to be limited to health. Information such as google searches, general interests, and online presence and activity can also be possibly useful information in the study. Furthermore, more personal data will also be employed. Personal health trackers and health indicators on a personal level will be used under the permission of participants in the study. It should be noted that the collection of this private information will be in accordance with HIPAA regulations and other regulations to protect private health information. These data sources, while incredibly broad, offer the best possible outcome for the experiment. As written by Mauricio Santillana (2015) in an article discussing flu predictions, "Our findings suggest that the information from multiple data sources such as Google searches, Twitter microblogs, nearly real-time hospital visit records, and data from a participatory surveillance system, complement one another and produce the most accurate and robust set of flu predictions when combined optimally" (Santillana et al., 2015). This in turn supports the data set for this experiment as a disparate data set is even more effective in discovering possible determinants of health.

Operationalization of Variables

Independent Variable

The independent variable throughout the course of the experiment will be the form of data mining and machine learning that is used to find possible trends in the data set. These forms will involve multiple different types from basic algorithmic computation and convolutional neural networks to more advanced forms such as random forests, deep learning, and regression analysis as an attempt to close the bounds between machine thinking and human decision making. The growth in complexity in these machine learning types is the level of human

assumptions imparted in their development. As the less and less human intervention is needed, the more complex and thorough the machine learning methods become. These more complex methods often prove to be more successful in analyzing clinical data as Andrew Beam writes for JAMA Network. He continues that because health data sets are so diverse, "algorithms on the high end of the machine learning spectrum have become practical and useful" (Beam & Kohane, 2018).

Dependent Variable

The dependent variable will then be measured as the validity and reliability of the discoveries of these machine learning algorithms in response to known determinants of health. The discoveries of the machine learning algorithm will be qualitatively compared to known determinants to determine how effective it was. This in turn will be an accurate judge of new future discoveries.

Discussion of Validity

One major validity issue in the use of machine learning is that it is nearly impossible for humans to understand the logic or reasoning behind the discoveries of a machine learning algorithm. As Andrew Beam writes, "While algorithms high on the spectrum are often very flexible and can learn many tasks, they are often uninterpretable and function mostly as "black boxes" (Beam & Kohane, 2018). Because the machine is not able to linearly explain its discovery, validating how new discoveries are made will prove to be difficult.

Discussion of Reliability

A considerable reliability issue is that not every computer or machine learning method will interpret a data set the same every time. Due to some variation in how these machine learning algorithms are developed, repeatability is not guaranteed. While the resulting discoveries will

generally prove to be identical or similar, the method by which they are discovered will vary from machine to machine, even when using the same machine learning methods.

Ethics

Due to the nature of the research proposal, it involves the use of private health information (PHI) as well as general public health data. As such, various HIPAA and other regulations that protect PHI will need to be employed in the instance that participants are asked to release medical data on a personal basis. While the ethical questions of using PHI can be solved simply, there lies a broader ethical question of allowing a machine to determine what might be attributed or indicative of our health. While answering this ethical question is not in the scope of this research proposal, it is important to consider that the findings of this study are not to be taken in lieu of historical findings, but rather as a supplement.

Limitations

This research proposal is limited by two key factors: Computing power and data availability. In correspondence to the ethical questions of obtaining PHI, the amount of data available to the research proposal may limit its ability to find new discoveries. In the case of Big Data and data mining assessments, the more available data there is a more likely outcome. In addition to general data availability, the efficiency of a research proposal such as this one is dependent on the available computing power. Analyzing incredibly complex sets of data that have no previously known links requires a large amount of computing power. Thus, when analyzing large sets of data, available computing power becomes a meaningful limitation.

Implications

The implications of this research proposal are deeply integrated into how new discoveries can be employed by healthcare professionals to support existing historical knowledge. As healthcare grows increasingly personalized and complex, the availability and usage of alternate methods to find possible determinants of health can prove to be useful in supporting historical understanding. Finding new determinants that may seem arbitrary to historical analysis, such as how often an individual uses their phone before bed, but are supported quantitatively may offer a new perspective on how healthcare can be approached in the future. This in turn leads to the possible future implications of research such as this leading to a more personal and more effective approach to healthcare.

Conclusion

As health becomes increasingly complex and personal it is important to utilize growing technologies to complement existing health standards. One of the most complex aspects of healthcare is finding determinants of health, or what in our lives impacts our health in meaningful ways. Due to the abstract nature of how many things can impact our health, analysis by means of the growing field of Big Data and data mining can prove to be an effective way to find new novel associations. These associations may be difficult to discover otherwise and may prove to be quantitatively meaningful determinants of health. The efficacy of this method of implementation would lead to a much more complex and nuanced understanding of health both from a professional and personal point of view. Individuals can learn that healthcare is specific to them and their lifestyle, and professionals can use this method to support historical medical understanding. In conclusion, using Big Data and data mining techniques in the healthcare

industry can revolutionize the way in which we perceive health by finding new, unforeseen novel associations that may indicate new determinants of health.

References

Bates, D. W., Saria, S., Ohno-Machado, L., Shah, A.,; Escobar, G. (2014). Big data in healthcare: Using analytics to identify and MANAGE high-risk and High-cost Patients. Health Affairs, 33(7), 1123-1131. doi:10.1377/hlthaff.2014.0041

Beam, A. L., & Kohane, I. S. (2018). Big data and machine learning in health care. JAMA, 319(13), 1317. doi:10.1001/jama.2017.18391

Breault, J. L., Goodall, C. R., & Fos, P. J. (2002). Data mining a diabetic data warehouse. Artificial Intelligence in Medicine, 26(1-2), 37-54. doi:10.1016/s0933-3657(02)00051-9

Dos Santos, B. S., Steiner, M. T., Fenerich, A. T., & Lima, R. H. (2019). Data mining and machine learning techniques applied to public health problems: A bibliometric analysis from 2009 to 2018. Computers & Industrial Engineering, 138, 106120. doi:10.1016/j.cie.2019.106120

Hansluwka, H. E. (1985). Measuring the health of populations, indicators and interpretations. Social Science & Medicine, 20(12), 1207-1224. doi:10.1016/0277-9536(85)90374-0

Herland, M., Khoshgoftaar, T. M., & Wald, R. (2014). A review of data mining using big data in health informatics. Journal Of Big Data, 1(1), 2. doi:10.1186/2196-1115-1-2

Hoffman, J. I. E., & Kaplan, S. (2002). The incidence of congenital heart disease. *Journal of the American College of Cardiology*, *39*(12), 1890–1900. https://doi.org/10.1016/s0735-1097(02)01886-7

Marconi, K., Lehmann, H., Bjarnadóttir, M. V., & Agarwal, R. (2014). Big data and health

analytics. Improving Decision-Making Using Health Data Analytics, 24. doi:10.1201/b17945

Miao, K. H., & H., J. (2018). Coronary heart disease Diagnosis using deep neural networks. International Journal of Advanced Computer Science and Applications, 9(10). doi:10.14569/ijacsa.2018.091001

Milovic, B. (2012). Prediction and decision making in health care using data mining. International Journal of Public Health Science (IJPHS), 1(2). doi:10.11591/ijphs.v1i2.1380

Nutley, T., & Reynolds, H. (2013). Improving the use of health data for health system strengthening. Global Health Action, 6(1), 20001. doi:10.3402/gha.v6i0.20001

Reshef, D. N., Reshef, Y. A., Finucane, H. K., Grossman, S. R., McVean, G., Turnbaugh, P. J., . . .Sabeti, P. C. (2011). Detecting novel associations in large data sets. Science, 334(6062), 1518-1524. doi:10.1126/science.1205438

Santillana, M., Nguyen, A. T., Dredze, M., Paul, M. J., Nsoesie, E. O., & Brownstein, J. S. (2015). Combining search, social media, and traditional data sources to Improve Influenza Surveillance. PLOS Computational Biology, 11(10). doi:10.1371/journal.pcbi.1004513 Song, X., Mitnitski, A., Cox, J., & Rockwood, K. (2004). Comparison of machine learning techniques with classical statistical models in predicting health outcomes. Studies in health technology and informatics, 107(Pt 1), 736–740.

Thimbleby, H. (2013). Technology and the future of healthcare. Journal of Public Health Research, 2(3), 28. doi:10.4081/jphr.2013.e28

Venkatesh, R., Balasubramanian, C., & Kaliappan, M. (2019). Development of Big Data Predictive Analytics Model for Disease Prediction using Machine learning Technique. Development of Big Data Predictive Analytics Model for Disease Prediction Using Machine Learning Technique. doi:10.1007/s10916-019-1398-y

Watson, K. (2019). Predictive analytics in healthcare Emerging value and risks. Deloitte Insights.