Big Data Analytics in a Healthcare Organization

John Wensink

MIS440 Cloud Computing and Big Data

Colorado State University-Global Campus

Dr. Mazen Alkhatib

February 22, 2020

Big Data Analytics in a Healthcare Organization

A vast abundance of data is available to support healthcare professionals, but how can we as an organization make use of this data while still protecting patient’s privacy and maintaining compliance within governmental regulations? Our organization stands to gain from the wealth of electronic data available such sources as digital health records, data from smartwatches, medical Internet of Things (IoT) devices, and other wearables as well as recent advances in genomic sequencing, multidimensional medical imaging, pharmaceutical research, and the sharing of records from other healthcare organizations (New England Journal of Medicine, 2018). The current state of big data analytics and cloud computing now allows our organization to gain meaningful and valuable insights into relationships that exist within the data, both obvious and hidden. Healthcare information is a unique set of data types characterized by an incredibly high volume of information, the velocity of the data is fast-moving, and the variety of personal healthcare data is extraordinarily broad encompassing the full spectrum of the digital information landscape. The use of wearable medical technology and the Internet of Things, if properly analyzed, is able to bring our level of insight down to the scale of individual patient’s biological trends. The value of this information is tremendous, and we need to be certain of the data’s authenticity and veracity. Traditional informational analytics falls short of this data’s potential to revolutionize the healthcare industry. As such, our firm has a unique opportunity to utilize a new to healthcare information management by leveraging cloud technology to facilitate the acquisition, storage, analysis, processing, and transmitting this valuable data, with emphasis placed on the patient’s right to understand how their personal healthcare information is being used

**Making Use of Cloud Services**

The magnitude of computational resources that will be required to make this project a success is staggering. Fortunately, modern on-demand cloud services are able to provide this capacity we are seeking without the need to own these resources outright. The National Institute of Standards and Technology (NIST) defines a cloud service as it pertains to five feature attributes. On-demand self-service, broad network access, resource pooling, rapid elasticity, and resource optimization as a measured service (Mell & Granst, 2011). The NIST researchers go on to further elaborate their findings by comparing and contrasting different cloud deployment and service models.

**Cloud Deployment Model**

The cloud deployment model which fits our mission best is that of the hybrid cloud, which will be made available to employees within different areas of our organization, as well as a portal for the patients themselves (Mell & Grance, 2011). This community cloud will be managed by specialists within our own organization who have a deep understanding of the possibilities and requirements that managing such a system entails. Emphasis must be placed on acquiring and retaining individuals with a high degree of technical literacy as it pertains to cloud-based data management in the field of medical information. Third-party providers and pre-existing applications are to be avoided wherever possible. Although the use of original applications and local employees will come at a higher cost to the organization than reliance on third-party contractors and programs, there are tangible benefits to using this approach. Some examples of these benefits include better accountability with personal healthcare information security, high standards of productivity are easier to enforce and obtain, and locally sourced professionals will be more personally invested in providing a high-quality resource to their own community which stands to benefit from the job creation this project will require (Albanese, 2018). A hybrid cloud will be the appropriate selection for the deployment model, as it encompasses each of the other deployment models (Public, Community, and Private) which will all have a use within our big data environment. Public clouds will be utilized by our patients in order to access their health records, prescriptions, treatment plans, and scheduling tools. Private clouds will be necessary for internal use in our organization as it relates to data analytics and forecasting, and community clouds will be of great importance for other research institutes to collaborate on problems and share in our successes, promoting a bright future for the entire healthcare industry. For all of these applications to work together seamlessly, a hybrid cloud model is recommended.

**Cloud Service Model**

The scope of this project calls for the use of the Platform as a Service (PaaS) cloud service model, as it emphasizes the creation of a new web application aimed at revolutionizing the healthcare industry by the use of big-data analysis. Although we are creating new web applications, we don’t need to reinvent the wheel and will want to utilize existing application development toolkits that have proven themselves to be secure, reliable, and efficient as they pertain to eliminating repetitive tasks that are commonly associated with web-app development. Google’s AppEngine (GAE) seems to be the obvious choice for our PaaS development goals as it contains tools for both developing as well as hosting our healthcare information system. GAE offers scalability as well as the efficiency with their on-demand fee structure, with the ability to commission and decommission dynamic hardware resource requirements. The utilization of Google’s Tensorflow when integrated with Apache’s Spark provides an open-source solution for our organization to create our own machine learning applications. The Tensorflow on Spark framework is able to provide the varying degrees of abstraction required for supervised learning, unsupervised learning, and artificial neural networks. Our organization will also need to investigate the benefits and drawbacks of using a Docker-based containerized infrastructure versus the use of hypervisor-based virtual machines to determine which approach will most closely fit our data isolation needs as a portable healthcare data information system (Schlosser, 2020). We understand that our organization’s resources are not infinite, and to that end, recommend the utilization of other open-source analysis tools including Apache’s Hadoop for gaining insights into the data collection which is sure to span vast data clusters (Apache, 2018). Using open-source tools is one way that our application will seek to balance the cost-benefit relationship which will carry a high-cost burden from the avoidance of third-party contractors and Machine Learning (ML) applications.

This project has great potential to revolutionize the way the healthcare industry uses big-data analytics to make intuitive tools that can provide our professionals with insights into treatment possibilities. It is of vital importance to remember our organization’s mission to improve the lives of the patients we treat while acting ethically and in compliance with existing and future governmental regulations. For this reason, we recommend that the search for high-quality data professionals begin without delay so as to gain insights from their experience as we begin to build this application which holds tremendous potential to improve the lives of the people we serve.

**Information Security, Patient Privacy, and Information Ethics**

Our dedication to protecting our company, employees, and patients, and intellectual property will be a key metric of the success or failure of our big data analytics initiative. Obtaining this data ethically is another complex challenge that needs to be addressed. In the United States, The Health Insurance Portability and Accountability Act (HIPAA) is an act passed by Congress and signed into law in 1966 (DCHS, 2019). HIPAA provides regulations regarding the safeguarding of confidential patient information, along with hefty fines for non-compliance. HIPAA, along with other state-specific regulations demands that our team pay great attention to the project’s information security concerns, and rollout the program only after rigorous testing from independent security firms. Other concerns in this domain are related to the physical sensors that will be collecting personal healthcare information and the networks that they utilize to transmit this data to our secure network. With the growing prevalence of IoT wearables as well as more specialized wearable medical devices, it is critical that the information gathered by such sensors is collected with the informed consent of the patient, transmitted to our organization securely, and utilized ethically in the furtherance of our organization’s mission. Training our employees in the subjects of information security and informed medical consent will go a long way towards ensuring our organization complies with these requirements, however, information security is a complex subject that requires professional supervision. We will need to allocate substantial monetary resources towards offering competitive salaries that will attract and retain well-qualified information security professionals to our organization. Information security, patient privacy, and analytical ethics will need to be a cornerstone of this project if we are going to have long term success in this area of healthcare bioinformatics

**Information Security**

The basis of Information Security (InfoSec) is to utilize a standard of protective measures designed to prevent the unauthorized use of electronic information. This is a highly technical field of study that should not be managed without the proper credentials. Fortunately, the proliferation of cloud applications has caused such concerns to be addressed on the global scale required to offset many of more common threats. A sub-specialization of Software as a Service (SaaS), Security as a Service (SECaaS) allows businesses to integrate their information systems with security services that can be more effective, up-to-date, and economical than relying solely on in-house security professionals. Utilizing the benefits SECaaS can offer our organization protection from maligned actors in areas such as “identity and access management, data loss prevention, encryption capabilities, vulnerability management, and email security” (Musa, 2017, p.350). For this reason, it is recommended that our project make use of SECaaS solutions, to be employed by a team of in-house professionals so that the benefits of these services can be utilized al the while being supervised by a team of professionals who can provide services custom-tailored to our unique situation. We would be wise to look at the impact past data breaches can have on healthcare businesses, with one example being the attack on Dominion National, a Virginia based health insurance administrator. In April 2019 it was discovered that the company’s patient information servers had been breached by an unauthorized user, and had been breached since August of 2010, nearly ten years (HIPAA Journal, 2019). The effect of this attack was that over 2.9 million members had sensitive information such as bank account and routing numbers, social security numbers, taxpayer ID’s, as well as more basic information such as names and addresses, etc. Although it is too early to know what kinds of penalties lie in store for Dominion National, the firm is already facing class-action lawsuits with litigants being represented by Kehoe Law Firm of Philidelphia (KehoeLawFirm, n.d.).

**Patient Privacy**

The digitization of medical records offers the capability to analyze vast amounts of personal healthcare information, but these datasets are typically quite large and have a high degree of complexity. The scope of such records presents a problem in effectively managing the information contained within with regards to individual patient’s privacy wishes. Kamble, Gunasekaran, Goswami, & Manda (2018 p.231) found that implementing specific requirements of database capabilities was more likely to succeed in implementing and maintaining patient privacy in a healthcare data environment. The list of criteria includes best practices for data governance, security analysis, and specific analytics for the preservation of privacy. The study found that the end-to-end authentication of users, as well as encryption of patient data that has been stored and maintained with respect to local, national, and international standards not only lead to increased patient privacy but also improved the performance and efficiency of the information system. Kamble et al. credits these improvements to implementing the use of blockchain technology not dissimilar to the way cryptocurrencies such as Bitcoin maintain security and privacy across the global digital currency network. Yue, Wang, Jin, Li, & Jiang (2016) describe in detail how a blockchain-based system where individual patients maintain ownership of their own data can be implemented into healthcare data analytics to create a system that guarantees trustless and auditable computation capabilities by the use of an encrypted public ledger system. The research even presents the idea of Secure Multi-Party Computing (MPC) which could enable untrusted third-parties to make use of sensitive information without sacrificing privacy.

**Information Ethics**

When considering the ethical implications of the collection and analysis of healthcare data Ienca et al. (2018) discovered through a meta-analysis of 1,093 publications that “findings suggest that while big data trends in biomedicine hold the potential for advancing clinical research, improving prevention and optimizing healthcare delivery, yet several epistemic, scientific and normative challenges need careful consideration. (p.1)” In the recent past, informed consent had been the gold standard of such practices as codified in The Federal Policy for the Protection of Human Subjects, otherwise known as the Common Rule. This statute served the industry well until the explosion of IoT and big data in the late 2000s began to show that the Common Rule was generally not compatible with such practices. In January of 2019, the Revised Common Rule was passed which replaced the idea of a patient’s informed consent regarding identifiable personal healthcare information, with broad consent, which allows more generalized use and re-use of a patient’s data. Although the Revised Common Rule raised the ceiling to which such data could be used for modernized analytics, unintended side effects such as genomic testing corporations selling identifiable personal data to third parties emerged (Froomkin, 2019). With regards to our own healthcare information system, a patient simply signing a broad consent form is not enough to guarantee that we are doing everything we can to act ethically with their personal healthcare data. The patient needs to know what they are signing, specifically opt-into the research program (rather than opt-out), and have the ability to withdraw their consent for future projects at any time. It is essential that we differentiate different levels of patient consent with regards to how a patient’s data will be used. For example, a patient should be able to grant consent to the collection of their data as it pertains to their own treatment while opting out of research projects aimed at using such data for studies not directly related to their care. Other areas of consideration outside the scope of this proposal will need to be considered including areas such as discovering unborn children predisposed to genetic disorders, use of smartphone cameras for optical medical imaging (Rat et al., 2018), end-of-life issues, and countless others will need to be examined and planned for.

**Big Data Analytics and Machine Learning**

In 2020, with the high degree of cloud computational and processing resources at our disposal, our organization is in a unique position to take full advantage of the breakthroughs in big data analytics and cloud computing the last few decades have created. By applying supervised, semi-supervised, and unsupervised machine learning algorithms including cognitive computing and neural networks tremendous capabilities of classification and prediction lie at our fingertips. In 2016 the IBM computer company stated that “the amount of health information is doubling every three years; by 2020 it is estimated to double every 73 days” (IBM, 2016).

With the exponential explosion of the creation of new data, due in part to the internet of things, in order to efficiently utilize these information resources, it is critical that we integrate mined information across multiple data types while making efforts to preprocess this information. Preprocessing of this mined big data is critical to avoid inefficiencies due to irrelevant or redundant data. However, it is also to maintain a cautious approach to doing so with the understanding that improper filtering has the potential to affect the output information of these systems. With high-quality, labeled input data, we are able to take advantage of such techniques as regression, k-nearest neighbor, and Bayesian classification techniques to solve such problems as outcome probabilities, genomics, and epidemiology to name a few. Other machine learning techniques such as clustering, association analysis, and dimensional reduction can be useful tools to help identify relationships in data where no correlation had been identified before. Finally, with ample tuning and weighting of these algorithms, visualization of these insights can be efficiently created more efficiently by recent advances in GPU image rendering and massively parallel processing infrastructures in the cloud.

**Data Mining, Integration, and Preprocessing**

Getting into the details of what we want out of our cloud-based healthcare analytics system includes examining the sources for the data that will be mined. Smartphones and general wearable technology such as smartwatches are common among the American population and are able to provide a great deal of information regarding metrics such as GPS location, physical activity, heart rate, oxygen saturation, and mood. While these devices are plentiful, the data they produce is less precise than what a physician or information analyst would prefer. Health sensor data is not limited to consumer products. Researchers at MIT’s computer science and artificial intelligence laboratory are in the process of creating a passive device called ReMix that has the ability to use ingestible implants to track the small movements of cancerous tumors during the course of radiation treatment (Gordon, 2018). Other potential sources of sensor data will likely proliferate as wireless communications providers begin to implement 5G service across the country. 5G communication will allow for the transmission of very large quantities of data from various sensors at incredible speeds. The 5G revolution offers limitless potential in terms of low latency, high-reliability sensor communication with the possibilities emerging for medical care for areas such as remote surgery, remote emotional pacification, augment reality diagnoses, and even lie detection (Chen, Yang, Hao, Mao, & Hwang, 2017). A wealth of unstructured data can be found by analyzing an individual’s social media presence. Information such as Facebook posts Twitter feeds, and even user reviews of businesses can help healthcare providers gain insight on an individual’s day-to-day lifestyle (Raghupathi, 2017). With all of this information available both structured and unstructured, there is a real challenge and need to integrate and pre-process data from disparate sources so that machine learning algorithms are able to make use of the information in an efficient manner. Data that is incomplete, inconsistent, or statistically noisy will need to be cleaned by filling in missing entries, normalized and aggregated, as well as integrated by using data cubes, multiple databases, and appropriate file types (Markov, n.d.). The saying ‘garbage in equals garbage out’ has profound application when we are working with healthcare data if we are expecting useful results.

**Machine Learning Algorithms**

Now that we have high-quality inputs for our healthcare information system to utilize, we are ready to begin examining different Machine Learning (ML) algorithms that have proven themselves to be useful as they relate to healthcare data. In addition to the algorithms themselves, visualization solutions will be needed to make the algorithm’s outputs simple to understand for their human operators. Some examples of visualization tools that are prevalent today include D3, HighCharts, Leaflet, Vega, and Power BI (Chou, 2019). Each of these tools serves a unique purpose whether it be graphics presentation, interactive charting, or mobile compatibility. Having ML outputs without a way to visualize them would be akin to having an encyclopedia of knowledge written in a foreign language.

**Supervised machine learning algorithms.** Some examples of supervised machine learning techniques (where labeled training data is used to tune the ML algorithm) include regression techniques such as decision trees and support vector networks, probabilistic classification algorithms such as the Markov model and native bayesian, as well as non-parametric statistical analysis techniques such as K-Nearest Neighbour (KNN) and Kuiper’s test for seasonal distribution. These supervised machine learning techniques can be useful for learning the features of a large healthcare dataset, performing repetitive tasks such as analyzing scan results, and help with providing doctors with the most recent clinical procedures (Sciforce, 2019). Regression techniques, one of the most widely used ML strategies, can be applied to healthcare analytics by determining the probabilities of one outcome or another such as whether or not a skin spot is likely to be malignant cancer. The naive Bayes classifier is useful in healthcare research to determine how a particular research article should be classified and what keywords might cause it to appear in a search result. Support vector machines, while also useful for regression and classification serve a unique role in the diagnosis of cancer or neurological disease from irregular data that can have missing datapoints (Sciforce, 2019).

**Unsupervised machine learning algorithms.** While supervised machine learning techniques are extraordinarily useful when it comes to classification problems, there are other problems that our healthcare information system seeks to resolve. Using unsupervised machine learning techniques we are able to learn about the inherent structure of our data without providing labeled data. Some examples of useful unsupervised machine learning techniques include clustering techniques such as hierarchical clustering and k-means analysis, dimensional reduction techniques such as principal component analysis and factor analysis as well as stochastic neighbor embedding. Without a priori information regarding an object’s class, unsupervised learning algorithms can discover unknown relationships between objects. For example, k-means analysis can help separate diabetes patients into different risk categories as well as discovering risk factors regarding the concentration of fluoride in drinking water (Ogbuabor & Ugwoke, 2018). Other techniques such as principal component analysis which has shown to be effective in determining the causes of systemic inflammation caused by the presence of various inflammatory cytokines, that were shown to be correlated in many different dimensions. Using a principal component analysis algorithm showed the researchers that “the variance of these ten cytokines can be accounted for by three principal components (PC). As a result, the model was remarkably simplified” (Zhang & Castelló, 2017).

**Artificial Neural Networks and Deep Learning.** Artificial Neural Networks (ANN) seek to reproduce the efficiency and breadth of biological information processing. These systems attempt to ‘learn’ how to complete tasks by drawing on examples, without hard rules how to arrive at a conclusion. The classic example is the identification of cats in images and videos without any prior conditioning about what a cat is or is not. Rather, using labeled training data, ANN seeks to identify certain characteristics of images that are (or are not) common to those training images of cats. ANN relies on analyzing vast multitudes of data and becomes more reliable the more it is put to use. Some examples of ANN include restricted Boltzmann machines, deep belief networks, propagation training, and convolutional computing. ANN, as it applies to the analysis of healthcare data, can be useful in areas such as rational drug design, biochemical analysis, medical image analysis, and diagnostic testing. We are only just beginning to see the capabilities that are possible using ANN. As such, it should be the stated goal of our organization to become a leader in this field.

**Conclusions**

The explosion in the quantity of healthcare data that has become available in the previous decades presents a tremendous opportunity for our organization, but it is not without its challenges. In order to make use of all this data, it is imperative that we create a secure information network with the capability to harness this tremendous volume of data in a secure, ethical, and efficient way paying special attention to patient’s privacy concerns. This data can come from a wide variety of sources but must be transformed into something that is usable and relevant to the growing variety of machine learning algorithms at our disposal. The recent advances in cloud computing such as PaaS, SECaaS, and IaaS need to be leveraged in parallel with highly skilled in-house professionals that are able to use this tremendous resource responsibly. There is no telling what the future holds for big-data analytics and cloud computing and we are learning more every day at an accelerating pace. Although the ambition of this project greatly exceeds anything our organization has done before, the rewards that will come of it in the form of quality of care, treatment outcomes, and profits for our shareholders is without limit and work should be started immediately to make this dream a reality for our organization.

References

Albanese, J. (2018, July 30). Relying on the Business of Strangers: How to Manage Third-Party Vendors, Suppliers, and Contractors. Retrieved February 2, 2020, from [https://www.inc.com/jason-albanese/relying-on-business-of-strangers-how-to-manage-thi rd-party-vendors-suppliers-contractors.html](https://www.inc.com/jason-albanese/relying-on-business-of-strangers-how-to-manage-thi)

Apache. (2018). Apache Spark™ - Unified Analytics Engine for Big Data. Retrieved February 2, 2020, from <https://spark.apache.org/>

Chicco, D. (2017). Ten quick tips for machine learning in computational biology. *BioData Mining*, *10*(1). Retrieved from <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5721660/pdf/13040_2017_Article_155.pdf>

Chen, M., Yang, J., Hao, Y., Mao, S., & Hwang, K. (2017). A 5G Cognitive System for Healthcare. *Big Data and Cognitive Computing*, *1*(1), 2. Retrieved from <https://www.researchgate.net/publication/315928517_A_5G_Cognitive_System_for_Healthcare>

Chou, L. (2019, September 11). 9 Data Visualization Tools That You Cannot Miss in 2019. Retrieved March 8, 2020, from <https://towardsdatascience.com/9-data-visualization-tools-that-you-cannot-miss-in-2019-3ff23222a927>

DHCS. (2019). Health Insurance Portability & Accountability Act. Retrieved February 23, 2020, from <https://www.dhcs.ca.gov/formsandpubs/laws/hipaa/Pages/1.00WhatisHIPAA.aspx>

Froomkin, M. (2019). Big Data: Destroyer of Informed Consent. *Yale Journal of Health Policy, Law, and Ethics*. Retrieved from <https://poseidon01.ssrn.com/delivery.php?ID=818078115066107107004067120121124122127040038027075004027084014009125024118094114109050006019042105058018091098087078115001029117037004023036118119078114104074028011015097123112093122000080112087084010010083022069115004100012017118026002087114085&EXT=pdf>

Hwang, K., & Chen, M. (2017). *Big-data analytics for Cloud, IoT, and cognitive computing*. Chichester, West Sussex: Wiley Blackwell.

Ienca, M., Ferretti, A., Hurst, S., Puhan, M., Lovis, C., & Vayena, E. (2018). Considerations for ethics review of big data health research: A scoping review. *Plos One*, *13*(10). Retrieved from <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6181558/pdf/pone.0204937.pdf>

Islam, M., Hasan, M., Wang, X., Germack, H., & Noor-E-Alam, M. (2018). A Systematic Review on Healthcare Analytics: Application and Theoretical Perspective of Data Mining. *Healthcare*, *6*(2), 54. Retrieved from <https://www.mdpi.com/2227-9032/6/2/54/pdf>

Kamble, S. S., Gunasekaran, A., Goswami, M., & Manda, J. (2018). A systematic perspective on the applications of big data analytics in healthcare management. *International Journal of Healthcare Management*, *12*(3), 226–240. Retrieved from <http://web.b.ebscohost.com.csuglobal.idm.oclc.org/ehost/pdfviewer/pdfviewer?vid=1&sid=7e343a9c-e309-4f43-8341-e4cc40bdda15@pdc-v-sessmgr05>

KehoeLawFirm. (2019, June 28). Dominion National Data Breach - Data Compromise Recently Discovered -. Retrieved March 8, 2020, from <https://kehoelawfirm.com/active-investigations/dominion-national-data-breach>

Mell, P., & Grance, T. (2011, September). The NIST Definition of Cloud Computing. Retrieved February 2, 2020, from <https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-145.pdf>

Musa, K. (2017). Evaluating Security-as-a-Service (SECaaS) Measures to Increase the Quality of Cloud Computing. *International Journal of Science and Engineering Applications*, *6*(12), 350–359. Retrieved from <https://ijsea.com/archive/volume6/issue12/IJSEA06121001.pdf>

New England Journal of Medicine. (2018, January 1). Healthcare Big Data and the Promise of Value-Based Care. Retrieved February 2, 2020, from <https://catalyst.nejm.org/doi/full/10.1056/CAT.18.0290>

Ogbuabor, G. F., & Ugwoke, F. N. (2018). Clustering Algorithm for a Healthcare Dataset Using Silhouette Score Value. *International Journal of Computer Science and Information Technology*, *10*(2), 27–37. Retrieved from <http://aircconline.com/ijcsit/V10N2/10218ijcsit03.pdf>

Raghupathi, W. (2014). Big data analytics in healthcare: promise and potential. *Health Information Science and Systems*, *2*(1). Retrieved from <https://link.springer.com/article/10.1186/2047-2501-2-3>

Rat, C., Hild, S., Sérandour, J. R., Gaultier, A., Quereux, G., Dreno, B., & Nguyen, J.-M. (2018). Use of Smartphones for Early Detection of Melanoma: Systematic Review. *Journal of Medical Internet Research*, *20*(4). Retrieved from <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5923035/>

Schlosser, H. (2020, January 30). Docker vs. Virtual Machine: Where are the differences? Retrieved February 2, 2020, from <https://devopscon.io/blog/docker/docker-vs-virtual-machine-where-are-the-differences/>

Sciforce. (2019, June 3). Top AI algorithms for Healthcare. Retrieved March 8, 2020, from <https://medium.com/sciforce/top-ai-algorithms-for-healthcare-aa5007ffa330>

Yue, X., Wang, H., Jin, D., Li, M., & Jiang, W. (2016). Healthcare Data Gateways: Found Healthcare Intelligence on Blockchain with Novel Privacy Risk Control. *Journal of Medical Systems*, *40*(10). Retrieved from <https://www.semanticscholar.org/paper/Healthcare-Data-Gateways:-Found-Healthcare-on-with-Yue-Wang/208735a6c437b8ae3efba01693c3e8a06289c3dd>

Zhang, Z., & Castelló, A. (2017). Principal components analysis in clinical studies. *Annals of Translational Medicine*, *5*(17), 351–351. Retrieved from <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5599285/>