Data Mining in Medical Data

Valter Berg Thomas Bye Nilsen

I. Abstract

We observe an ever-increasing quantity of data being produced in medical areas. The data originates from different sources throughout the research field. The diversity of the data and its properties make it challenging to process the data within a reasonable amount of time. To meet the demand for useful results distilled from large collections of data, new techniques are called for. To analyze big data, various data mining algorithms have been applied which extract useful pieces of information that are hidden a large datasets. In this paper we review some of the applied techniques, look for any shortcomings and give suggestions for the remaining challenges. How will the suggested designs fit as theoretical solutions?

II. Introduction

There is a vast and rapidly expanding amount of data being produced digitally. In the US healthcare system alone, in 2011, 150 exabytes of data was stored. Continuing at this rate of growth, the size of the volume will reach the zettabyte scale, and not long after that, the yottabyte scale [14]. When the physical properties of the data increases beyond a critical point, new techniques, architecture, algorithms and visualization of the data are required to use the data in a meaningful context. This has caused a rapid development in computer science research to gain insights into large collections of data, or big data, that have not been achieved earlier. This rapid development has mainly been unfolding within the disciplines like internet business and finance, while in the healthcare area, the progress has not been as widespread. However, there seems to be an increasing attention to big data in the healthcare sector. A survey is predicting that the outcome will be severely beneficial for the sector. [13]

A. Medical big data

Medical big data consists of large quantities of diverse medical data. The data originates from hospitals, pharmaceutical companies and academic environments. The datasets can be categorized based on their origin. Some of the categories are administrative claim records, clinical registers, electronic health records, biometric data and large clinical trials. The value of the data lies in the processing and interpretation of the datasets. The datasets in the different categories, which are likely to be heterogeneous in terms of format and completeness, can be linked together as well. It can be valuable to link heterogeneous datasets together to search for related data that is otherwise inaccessible. [6]

An example where integration of heterogeneous datasets is needed is when it is possible gain insight into the relation between patients' conditions and environmental events. A 360-degree view of a patients integrated with environmental data or even medical metrics yields insight into the relation between the two datasets. Facts regarding epedemic trends described in natural language text can be linked

to a database containing patterns that match the trend. [8]

Medical big data can also be hard to access due the lack of data-sharing intensives and in some cases, the lack of guarantees that untrusted third-parties will not be able to access the data. This is a serious privacy concern.

For patient records, there might be valuable information regarding previous treatments, given parameters like geographical region, age, gender and other biological properties. Relatively large datasets from hospitals across the industry, both public and private, possess such collections. Insights into these datasets could yield information regarding past and present patient treatment, use of medical resources and suggestions to future treatments. By grouping together patients, one could profile patients and review the treatment strategies. Because it is not possible to conduct enough trials to cover all types of patients, it could be crucial that enough evidence is gathered. [10]

Administrative claim records explain how healthcare resources are utilized and the outcomes of such use. Overview of economic status of an institution is important to administer resource usage. By looking at the records, one can see how different research projects and medical treatment projects are funded. Then, it is possible to balance the funding with respect to the funding resources available.

Clinical registers is closely related to patient records. [5] They contain information regarding the health status of patients and the treatments they receive over varying periods of time. They typically focus on the common reason for needing health care and how patients with various characteristics respond to various treatment methods. The clinical records can be compared to find patterns between patients with various characteristics that might prove to be helpful in finding new methods for treatments.

B. Motivation

There are several reasons for using new techniques in medical big data processing. There are several challenges the industry faces that the current technology is not able to overcome. First of all, there are missing values in the case of electronic patient records [6]. When records or attributes within records are missing, their absence yields an incomplete picture and low value of the dataset. This issue is traceable throughout even large datasets, which in turn can concern the industry to a relatively large degree.

In the case of curse of dimensionality has also proven to be a problem. This is the case of an excessive number of dimensions in the datasets. As the number of dimensions increases, the complexity for processing the datasets and the volume for storing the data increases rapidly as well.

Data standardization is also a challenge. The format in which the data is stored and processed can vary across institutions and companies . On a large scale across the industry, even across country borders, the standards can vary to a relatively large degree. The difference in formats can, in turn, prove to be challenging when related datasets are to be processed. Data mining of different datasets will therefore require different algorithms which can be time-consuming for researchers.

To some degree related to the lack of data standardizations, about 50% of all data in medical datasets are considered to be unstructured [6]. The unstructured data mainly consist of free-text reports and is used in communication between healthcare providers. The free-text can be from nurse notes, blood tests, or physiological data and it has no standard template on its appearance.

This heterogeneity in free-text data has to be considered when extracting information from it with data mining techniques. Although there have been more research applied to the field of data mining in medical health data, there is still room for more development, as depicted in this paper. Increased funding in medical big data will improve medical treatments and benefit patients for generations to come. Real-Time analytics in medical big data is a goal that benefits medical health care by providing results at a relatively high rate. [11] One of today's solution is to use cloud computing services provided by third-party vendors. The reason is that medical institutions do not intend to invest in software, hardware and trained personnel when there are "pay-as-you-go" services available [**?**]. The privacy aspect is elevated because a third-party vendor is involved. This can pose a threat to patient privacy and security because the data is in another party's possession.

III. Techniques

For some issues listed above, there exist solutions that partly overcome the challenges. For missing values, it is proposed to omit the data point. However, this may lead to loss in information and tampering of integrity. However, this method is overused in a practical context. [6] A suggestion is to use different techniques used in statistics. One technique is to estimate variance and covariance derived from maximum likelihood methods. This method is based on observing previous events. Previous events might involve other patients' records. The other algorithm is listwise detection. These two methods are used in datasets that suffer from MAR (missing at random) values. [7] The patients' privacy concern becomes relevant at this point. There must be no connection between two patients' identity, nor it can be possible to backtrack the calculation that points to a particular patient.

Another technique is imputation. Imputation is based on estimating substitute values based on other, for example from other similar datasets. [2]. The datasets can be patients with similarities.

Several techniques have been proposed to solve curse of dimensionality. [3] lists several methods to that can be used. One is DQC which has the ability to expose hidden structures and determine their significance in high-dimensional big data. It outlines correlated variables in the higher dimensions. One can look at it as a summery of the higher-dimensions. Another is feature hashing (FH). It is based on hashing the ID of the higher-dimensional feature. In this way, it is possible to reduce the storage space in the lower dimensions.

The issue regarding data standardization can be solved by agreeing on a standardized format that is traceable throughout the industry and research area. [1] established a forum consisting of over 200 individuals from 54 countries contributed to the dialogue on standardization and interoperability. In other words, there have been willingness to achieve this goal, yet there seems to be vastly more work to do. Representatives from institutions and companies must come together and agree on a set of formats compatible with existing technology and algorithms, which also retains patients' privacy rights. The challenge of standardizing is amplified by the presence of unstructured data. A potential solution can be applied to the input and output of the data mining algorithm. At the input side, a format can be determined so that a larger percent of the input data is structured. At the output end, the data can also be formatted in a determined format adhering to some standard.

Regarding the challenges with extracting infor-

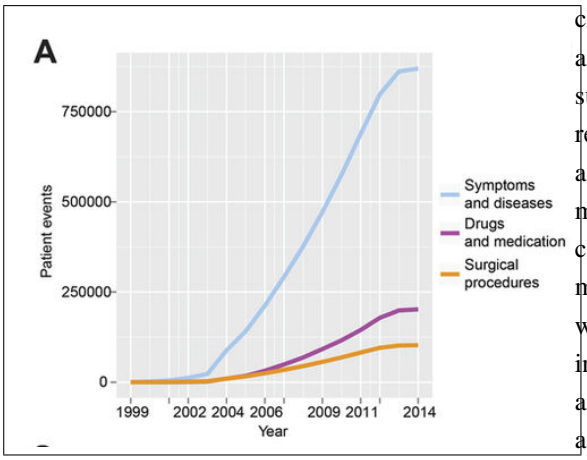

Figure 1. Patient events in terms of symptoms and diseases (blue), drugs, medication (magenta) and surgical procedures (orange) accumulated over time in the EHRs at UNN. [9]

mation from free-text in medical data the research group Jensen, K. et al. [9] worked on identifying cancer patient trajectories with analysing free text in electronic health records. Their goal was to identify cancer patient trajectories in need of resource demanding treatment and repeated hospital admissions. They concluded that their tool for identifying patient trajectories through free text analyses would be sufficient to use as a data-driven support tool in the complete cancer trajectory. They also claimed that use of their tool could decrease adverse events and readmissions and improve the quality of cancer care. TODO: DESCRIBE FIGURE 1 HERE SOMEWHERE.

Regarding real-time analytics, both security and financial requirements must be retained. The data contains IDs of real people that need to be protected against unauthorized users. While it is cheaper to make use of third-party cloud vendors [4], the privacy concern is still to discuss.

Here is a nice technique discussed. [12]

IV. Discussion

As we have seen, there are still some challenges that need to be addressed. Overcoming the challenges needs to be prioritized by governments across country borders. Despite the fact that the suggestions are theoretical, they are supported by reasoning. In the years to come, as the population and medical research grow in size, investments must be done. We have covered a subset of the challenges there is; there are more to shortly mention. One is regulatory compliance issues [11] which involves legal challenges that can surface in the process of storing and processing large amounts of data. Lack of appropriate skills can act as a barrier in implementing big data systems [11]. With change in technology, techniques and treatment standards, the personnel operating the systems must have up-to-date knowledge and skills. Even though there have been decades of research, there is still a risk of adverse events and mortality in cancer treatment today. With continuously assessing patient related risk factors the treatment can be optimized, adverse events avoided and number of hospital readmissions decreased. As mentioned the research from Jensen, K. et al. adapted to this concern with assessing patient related risk factors. As their results revealed, they achieved a mechanism that can be used as a tool when conducting patient trajectories. This is a valid example on how the data stored in EHR in the healthcare sector can be utilized to improve the medical treatment structure.

V. Conclusion

Although there have been more research applied to the field of data mining in medical health data, there is still room for more development, as depicted in this paper. Increased funding in medical big data will improve medical treatments and benefit patients for generations to come. As we have seen, there is room for improvements. As in any other area, there will only become more data to store and process. Addressing the challenges is the first step towards implementing

solutions. Several papers, on which this paper is based, have addressed these challenges. It shows that there is willpower to solve the remaining issues.

REFERENCES

[1]

[2] Imputation. https://en.wikipedia.org/wiki/_(statistics). Accessed: 2018-04-311.

[3] Rehman, m.h., liew, c.s., abbas, a. et al. data sci. eng. (2016) 1: 265. https://doi.org/10.1007/s41019-016-0022-0. Technical report.

[4] Maya Hao Li Jean Stanford-David Koester Patti Reynolds Arnon Rosenthal, Peter Mork. Cloud computing: A new business paradigm for biomedical information sharing. Technical report.

[5] American Medical Association. Waht is clinical data registry. Technical report, American Medical Association, 2014.

[6] Hyung-Jin Yoon Choong Ho Lee. Medical big data: promise and challenges. Technical report, Kidney Res Clin Pract, 2017.

[7] Craig K. Enders. A primer on maximum likelihood algorithms available for use with missing data, structural equation modeling, 8:1, 128-141. Technical report, 2001.

[8] Alexandros Labrinidis Yannis Papakonstantinou Jingesh M. Pate Raghu Ramakrishnan Cyrus Shahabi H.V. Jagadish, Johannse Gehrke. Big data and its technical challenges. Technical report, 2014.

[9] Karl Oyvind Mikalsen Rolv-Ole Lindsetmo Irene Kouskoumvekaki Mark Girolami Stein Olav Skrovseth Knut Magne Augestad Kasper Jensen, Cristina Soguero-Ruiz. Analysis of free text in electronic health records for identification of cancer patient trajectories. *Scientific Reports volume 7, Article number: 46226*, 2017.

[10] Harlan M. Krumholz. Big data and new knowledge in medicine: The thinking, training, and tools needed for a learning health system. Technical report, 2014.

[11] Clemens Scott Kruse. Challenges and opportunities of big data in health care: A systematic review. Technical report, 2016.

[12] Ernestina Menasalvas Myra Spiliopoulou, Pedro Pereira Rodrigues. Medical mining. *KDD '15 Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 2325–2325, 2015.

[13] Allan S. Detsky Travis B. Murdoch. The inevitable application of big data to health care. *American Medical Association*, 2013.

[14] Viju Raghupathi Wullianallur Raghupathi1. Big data analytics in healthcare: promise and potential. Technical report, 2014.