OULUN AMMATTIKORKEAKOULU

Mafas Dhaimash

PREDICTIVE ANALYTICS IN HEALTHCARE

Utilising Big Data for disease prevention and treatment

PREDICTIVE ANALYTICS IN HEALTHCARE

Utilising Big Data for disease prevention and treatment

Mafas Dhaimash
Bachelor's Thesis
Spring 2024
Information Technology
Oulu University of Applied Sciences

ABSTRACT

Oulu University of Applied Sciences
Information technology, Web Development

Author: Mafas Dhaimash
Title of the bachelor's thesis: Predictive Analytics in Healthcare: Utilising Big Data for disease prevention and treatment
Thesis examiner(s): Raili Simanainen and Miisa Tanner
Term and year of thesis completion: Spring 2024 Pages: 28

The objective of this thesis was to research the use of predictive analytics in healthcare, specifically focusing on the utilisation of big data for disease prevention and treatment. This thesis highlights the significant role of predictive analytics in healthcare, while studying the potential benefits and challenges related to the use of big data in medical contexts.

The research materials mainly consisted of existing literature on big data in healthcare, including its definition, data sources, benefits and challenges. Additionally, predictive modelling techniques, specifically machine learning algorithms were researched for their efficacy in healthcare. Case studies were analysed to demonstrate successful applications.

The results of this research indicate that while predictive analytics provides significant improvements to healthcare, there are various challenges and concerns to consider. Future development should focus on improving these analytics methods as well as finding solutions for the current challenges.

Keywords: Predictive Analytics, Machine Learning, Big Data Analytics, Healthcare data

CONTENTS

ABSTRACT 3
CONTENTS 4
VOCABULARY 5
1 INTRODUCTION 6
2 BIG DATA IN HEALTHCARE 7
2.1 Big Data defined 7
2.2 Big Data sources in healthcare 8
2.3 Big Data benefits and challenges in healthcare 9
2.4 Ethical and privacy concerns 11
3 PREDICTIVE MODELLING TECHNIQUES 13
3.1 Machine Learning 13
3.2 Disease prediction and prevention 16
4 PERSONALISED TREATMENT OPTIMISATION 18
5 CASE STUDIES AND SUCCESSFUL APPLICATIONS 21
6 CONCLUSION 23
REFERENCES 24

VOCABULARY

ARO	Adjustable Robust Optimisation. A personalised optimisation technique.
CVD	Cardiovascular disease. Refers to heart and blood vessel diseases, such as stroke and thrombosis.
EHR	Electronic Health Records.
FDA	Short for Food and Drug Administration. Ensures that food and drugs are safe for consumption and work as intended.
(n)	This will refer to references. For example (2) in text would refer to reference number 2.

1 INTRODUCTION

Predictive analytics employs statistical techniques, data mining models, and machine learning algorithms for analysing large datasets and predicting future events. In healthcare, this could involve identifying patterns, tendencies, and risk factors associated with different diseases and medical conditions, which in turn would improve patient care and health management. By utilising predictive analytics, healthcare providers can better evaluate patient needs, devise more accurate treatment plans, and distribute resources (e.g. personnel, facilities and equipment) more effectively. As technology has advanced, it has become increasingly important in healthcare. (1.)

This thesis aims to research and provide information regarding the significance of predictive analytics in healthcare, with a particular emphasis on the utilisation of big data.

Big data is an enormous compilation of data. It can be anything between structured and unstructured data. The enormous volumes of data include electronic health records (EHR), data from wearable devices, and medical imaging (e.g. x-ray and ultrasound imaging), among others. Big data analytics allows healthcare organizations to acquire useful and valuable information from the large datasets, which leads to improved decision making and better patient health management. (1.)

2 BIG DATA IN HEALTHCARE

Big data, as the name implies, is a grand compilation of data. Due to being gathered through various different methods, the data can be anything from structured to unstructured. These methods can include, for example sensors (e.g. cameras and GPS-sensors), and text data (e.g. news articles and blog content). In healthcare, these methods include, for instance, medical monitoring devices (e.g. monitoring patients activity levels and vital signs), EHRs (e.g. patients' medical history, diagnoses, and other health related information), and genomics (e.g. analysis and research of gene-related conditions such as diabetes and malignant tumours) (2).

2.1 Big Data defined

Big data refers to a large compilation of information. Originally, big data was described by three Vs: Variety, Velocity, and Volume. As time passed, more Vs were added (which can vary depending on context, needs, and sources), including Value, Veracity, Validity, and Volatility. In this context, the focus will be on the 5 Vs as shown in figure 1.

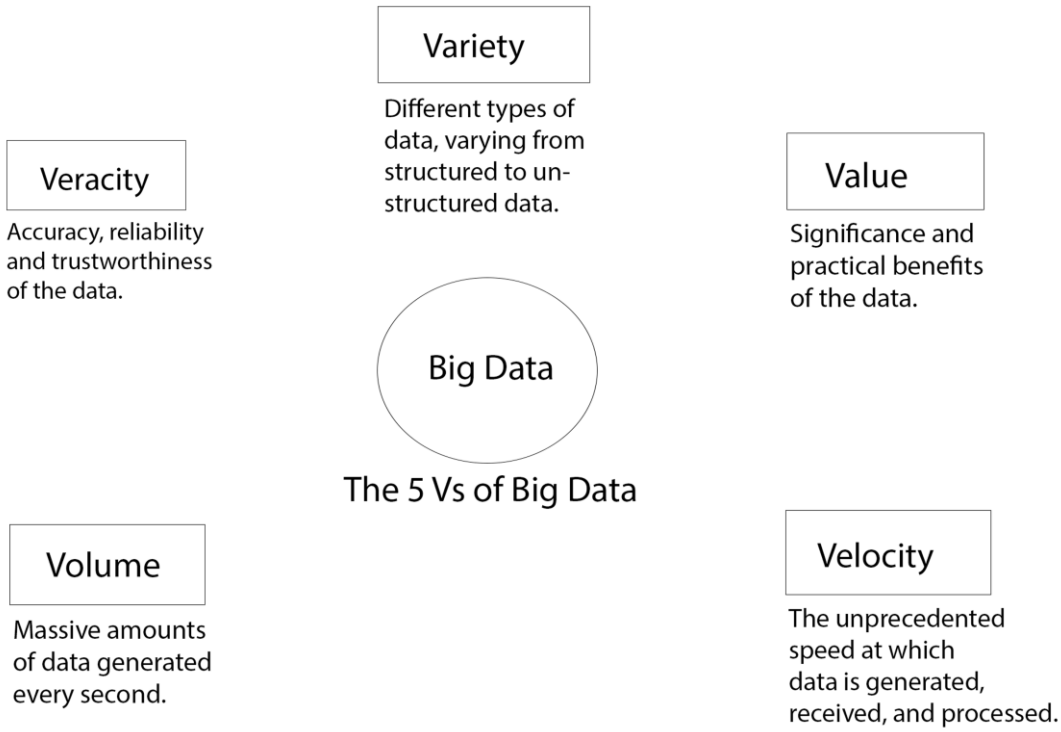


FIGURE 1. The 5Vs of Big Data (3).

Volume signifies the vast quantities of healthcare data generated, such as EHRs and genomic data. Velocity is the speed of information generated in healthcare, including patient monitoring devices and wearable technology. Variety indicates the diversity of information, including EHRs and laboratory results (structured data), as well as diagnostic imaging and clinical notes (unstructured data) (4). Moreover, ensuring the veracity of data is crucial in healthcare, to prevent possible negative outcomes. Additionally, the extraction of useful information from data (value) is also important for improving healthcare. (3; 5.)

2.2 Big Data sources in healthcare

As previously explained, there are various methods for gathering big data. In this chapter, some of the sources will be explained more thoroughly. For example, there are clinical, financial, biometric, patient generated, social media, and health research data.

Clinical data is gathered through various ways, for instance from genomic data (e.g. data collected from genes and DNA related study), imaging centres (e.g. X-rays and ultrasound images), and laboratories (e.g. blood tests and tissue samples). Clinical data aids with patient care (e.g. diagnosis, and management of diseases), clinical research (e.g. studying diseases, and testing new treatments), and healthcare management (e.g. monitoring population health), among others. (1; 6.)

Biometric data is collected from different devices that track patients' activity level, heart rate, and blood pressure, among other parameters. For instance, body-worn trackers can detect irregular heartbeats, potentially saving lives. Biometric data enables patients to continuously monitor their health, supporting preventive care and disease management. (1; 7.)

Patient generated data refers to data produced, logged, or collected from patients with health concerns. For example, it could be data from self-monitoring systems (e.g. sleep, eating habits, and physical activity), symptom logs (e.g. pain, fatigue, and other symptoms), and medication as well as treatment logs (e.g. records of medication taken, dosage and commitment to treatment plans). This data can aid both patients and healthcare providers monitor health and manage chronic conditions as well as overall health. It also assists healthcare providers with decisions regarding treatment, which is important for more personalised treatment. (1; 8.)

Data from social media includes information gathered from social media platforms, such as health-related discussions, patient generated content, and healthcare related advertisements. This data can give information regarding patients' opinions, behaviours, attitudes, and preferences related to healthcare, giving healthcare providers insight on patient needs which can improve the overall patient experience. (1; 9.)

Financial data is information regarding financial activity, including income, costs, and other transactions. Revenue provides insight about the sources of income, which can assist in determining where improvement is needed. Revenue data can also reveal new profitable income sources. Analysing expenses (e.g. employee salaries, device maintenance, and other expenses) can assist in identifying where expense reduction is possible. (10; 11.)

Health research data includes for example new treatment methods, drug research, and disease trends. This data helps researchers understand which treatments are the most effective, and how to improve overall healthcare. For instance, clinical trials are performed, where new medicine is tested on people, to determine its effectiveness and potential side effects. Once the medicine is licensed by the Food and Drug Administration (FDA), it is also important to observe its effects on the masses. In conclusion, health research is important to the advancement of medical knowledge, and innovation in healthcare. (1; 12.)

2.3 Big Data benefits and challenges in healthcare

In the past, treatment decisions were mostly made by doctors. In the present day, big data allows for more precise diagnoses and enhanced decision making, which also enables less expensive treatment methods. With data-driven decision making, it becomes easier to detect diseases early on, making more personalised treatment plans (including medicine), and even preventing diseases altogether (for example by recognising early warning signs, such as certain patterns). Utilising big data can also assist in monitoring patients at home, reducing the need for hospital visits, and with the previously mentioned benefits leading to overall improved population health management. Enhanced patient outcomes refer to better disease management, symptom relief, and rates of recovery. Through the assistance of big data, medical professionals can optimise resource allocation,

minimising the amount of wasted resources, and therefore reducing the overall expenses. Additionally, big data can advance the speed of new medicine creation, and the advancement of medical research (e.g. getting new insights regarding diseases). (1; 13.)

Data related to healthcare is highly sensitive information, as it contains private health-related details. Healthcare data is often kept in one place (centralised), since it can enhance the efficiency of healthcare. However, centralisation makes data more vulnerable to attacks. This makes it more crucial to protect the data and to prevent unauthorized access. Big data has multiple different sources. Healthcare related data can for example have different structures, use different terminology, and be in varying formats. This causes difficulties in synchronisation of different systems and technologies. (13.)

Due to the vast amounts of data, it can pose analytical and technical challenges, as well as expense related difficulties. For example, transferring, storing, and securing the large quantities of data can be very expensive (1). However, according to the research of Wullianallur Raghupathi and Viju Ragupathi, it is primarily the unstructured data which is posing challenges: "*Structured data is data that can be easily stored, queried, recalled, analyzed and manipulated by machine.*" and "*Already, new data streams—structured and unstructured—are cascading into the healthcare realm from fitness devices, genetics and genomics, social media research and other sources. But relatively little of this data can presently be captured, stored and organized so that it can be manipulated by computers and analyzed for useful information.*" (14). As structured data was previously explained to be easily analysed, it can be concluded that the second part is about unstructured data. Another challenge introduced by the enormous amounts, as well as the large variety of data, is related to data quality and accuracy. It is important that the data is as accurate and precise as possible, this being especially true in the field of healthcare. Poor data quality could lead to inaccurate conclusions and unreliable outcomes, which can be detrimental to patient care and treatment decisions (13).

Utilising the maximum effect of big data requires specific skills, alongside the constantly advancing technology (including big data related technologies and tools), there is a growing demand for data analysists and data scientists (15). When handling health related information, there are many legal and ethical aspects to consider. For example, it is important to acquire permission, inform patients

about how their data will be utilised (e.g. for analysis purposes), and to be honest about the practises in handling their private data. Inability to meet these regulations could lead to criticism, damaged business reputation, and even legal problems (16). (1; 13.)

TABLE 1. A Few benefits and challenges of Big Data in healthcare.

Benefits of Big Data in healthcare	**Challenges of Big Data in healthcare**
Better decision making	Data privacy and security concerns
Early disease detection and prevention	Synchronisation and compatibility issues
Personalization of medicine and treatment	Analytical and technical challenges
Population health management	Data quality and accuracy
Enhanced patient outcomes	Lack of human resources
Expense savings and efficiency gains	Ethical and regulatory considerations
Research and innovation opportunities	

2.4 Ethical and privacy concerns

Utilising predictive analytics provides numerous benefits to healthcare. However, health related data is private information, and there are significant concerns regarding its privacy and ethical considerations. It is important to take these issues into consideration, to ensure data safety, patient trust, as well as regulatory compliance, among others.

It is essential to obtain patient consent before using their data. This includes ensuring that patients are fully aware of how their data will be used and giving them the opportunity to decide about its use. For example, patients should be able to choose whether their information can be applied for medical research, or different intentions. This can be challenging, especially in situations where patients may not completely understand the risks of sharing their data. (17.)

Another important consideration is related to data security. Health related data is sensitive information which must be protected from data misuse and unauthorised access, among other potential threats. Healthcare organisations must implement security measures that can protect patient data throughout different phases of data management including gathering, storing, analysing, and deleting the data. (18; 19.)

Additionally, it is crucial to verify that all personnel handling patient data are aware of their responsibilities regarding the protection of private patient information. This brings another security challenge, data anonymisation. To protect patient privacy, their data must be anonymised before being utilised for predictive analytics. Anonymising can include deleting names and addresses, among other identifiers. However, it is possible to re-identify individuals, if there are enough similarities in their data, such as unique medical conditions. Therefore, anonymising data while keeping it useful for healthcare analytics is a significant concern. (20.)
In addition, it is essential for healthcare institutions to verify that their data usage practises comply with relevant regulatory requirements. For example, in Europe there is the General Data Protection Regulation (GDPR), meanwhile in the United States there is the Health Insurance Portability and Accountability Act (HIPAA). These regulations define rules for the management of private data, including within healthcare organisations. (17; 21.)

In conclusion, healthcare providers must notify patients about regarding the utilisation of their data, and the benefits and risks while asking for consent of its use. Additionally, organisations must be honest about their data management practises, in compliance with the relevant regulatory requirements. All personnel handling data must be well trained regarding data security. These are important concerns to consider, to ensure patient trust is retained, privacy is protected, and ethical standards are maintained.

3 PREDICTIVE MODELLING TECHNIQUES

Predictive modelling is a data analytics technique, which utilises historical data (for example patient data) to predict future events, such as health outcomes. In healthcare, predictive modelling encompasses a variety of tasks, including forecasting disease progression and outcomes for conditions such as cancer and diabetes (22). For example, predictive modelling often includes machine learning methods, which have a significant role in healthcare analytics. By examining broad datasets, machine learning methods can uncover patterns and provide new insights on data, which can improve treatment methods, and enhance diagnosis accuracy, among others (23). This chapter will focus on machine learning techniques and disease prediction within the field of predictive modelling.

3.1 Machine Learning

Machine learning refers to the creation of statistical models and algorithms that allow computers to study and learn to enhance their efficiency on a particular assignment without being specifically programmed for it. In healthcare, these tasks may include improving patient outcomes, disease prevention, and treatment strategies. Machine learning techniques are becoming more common in healthcare to analyse large and intricate datasets, and obtain valuable information. This valuable information is then utilised for developing predictive models that can support clinicians in improving diagnostic accuracy, forecasting treatment results, and personalising treatment schemes. These techniques are used for analysing different medical data, including EHRs, medical imaging, and genetic data, among others, to recognise trends and patterns that might not be so evident to professional analysts. Machine learning techniques can be categorised into supervised, and unsupervised learning. (23; 24.)

Supervised learning is one method of machine learning in which the model is developed using categorised data, meaning the data provided is connected to the correct result. These methods are used when the result is known, and the algorithm's task is to learn the connection between the input and the output (25). Under supervised learning, two main tasks are commonly performed: classification and regression. Classification algorithms are used when the output variable is categorical. For example, they are useful for disease diagnosis, such as predicting whether a patient

has a specific condition based on their symptoms and health record. Additionally, classification algorithms are utilised in various other tasks, such as image classification where medical images are categorised into different classes, such as benign (harmless) or malignant (harmful) tumours. Regression algorithms are utilised when the output variable is either a continuous (e.g. estimating an individual's life expectancy based on age, lifestyle, and medical history) or a binary value (e.g. estimating whether or not an individual has had lung cancer before, based on various factors such as age and smoking status) (26). Supervised learning algorithms include decision trees (tree-like model of decisions and their possible outcomes, as can be seen in figure 2), random forest (consists of multiple decision trees, combining the predictions of decision trees making more accurate predictions), and neural networks (consists of interconnected nodes that perform tasks by processing input data, making predictions, and learning from feedback, as illustrated in figure 3), among others (27). (23; 24.)

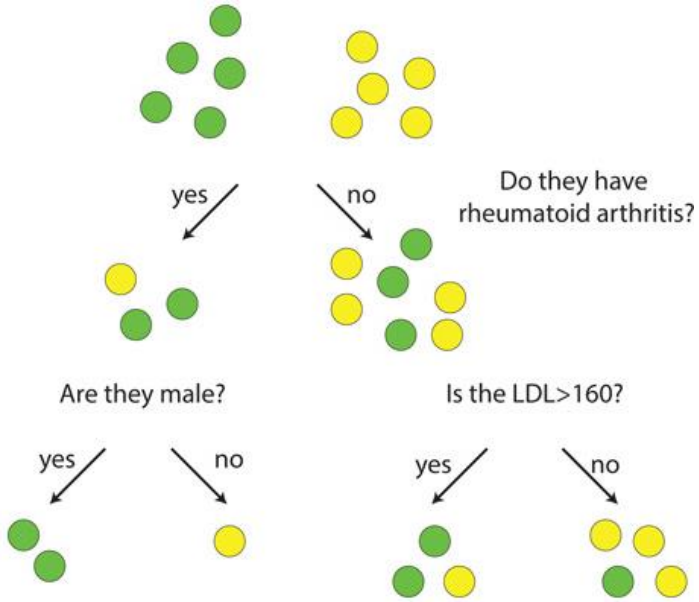


FIGURE 2. "Machine learning overview." By Rahul C. Deo, National Library of Medicine (24).

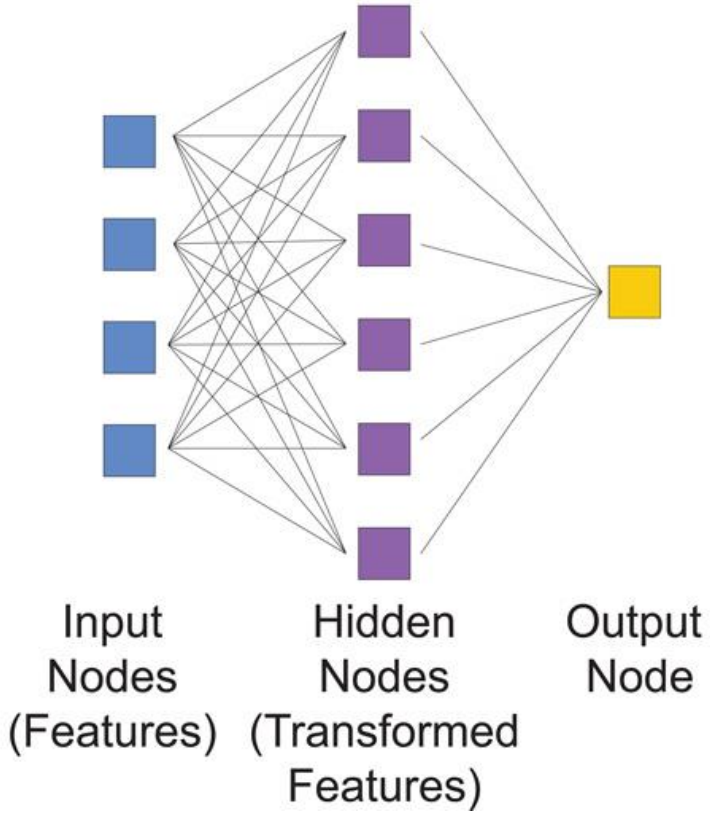


FIGURE 3. "Machine learning overview." By Rahul C. Deo, National Library of Medicine (24).

Unsupervised learning is a different method of machine learning, in which the algorithms are developed with uncategorised data, in other words the input data is not connected to any specific output. These methods are used when the output is unknown, and the algorithm's task is to learn the structures and patterns in data without guidance. In the context of unsupervised learning, a couple of main tasks are commonly performed: clustering and dimensionality reduction. Clustering algorithms are used to gather similar data points together based on their features, without predefined output variables. For instance, these algorithms can be used to identify various subgroups within people by deriving data from their medical records, lifestyle, and additional health characteristics. Dimensionality reduction techniques are utilised to decrease the quantity of attributes within a dataset, while retaining its crucial information. By decreasing the dimensions of the data, these techniques assist in the discovery of hidden patterns. Focusing on the more important parts can lead to earlier diagnosis and better treatment outcomes. Unsupervised learning algorithms include K-means clustering (separates the dataset in pre-defined (k) clusters according to the similarity of the data points) (28), and principal component analysis (PCA, simplifies complex datasets while retaining the important information, as shown in figure 4), among others (29). (23; 25.)

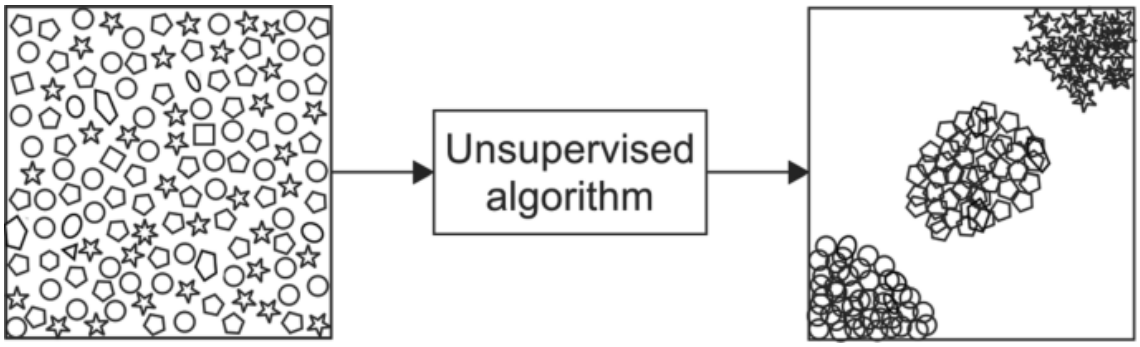


FIGURE 4. "A visual illustration of an unsupervised dimension reduction technique" by Jenni A. M. Sidey Gibbons and Chris J. Sidey-Gibbons, BMC Medical Research Methodology (23).

In conclusion, supervised learning algorithms can assist clinicians with monitoring patient progress, assessing patient risk, and making personalised treatment decisions, among other tasks. On the contrary, unsupervised learning algorithms assist in identifying hidden patterns and discovering new insights, leading to earlier diagnosis and better treatment outcomes. By utilising these machine learning techniques, healthcare can be significantly improved.

3.2 Disease prediction and prevention

Machine learning has an important role in disease prediction and prevention, providing various methods to analyse large quantities of data and identify patterns that can assist predicting and preventing different diseases, such as cardiovascular diseases and cancer. Cardiovascular diseases (CVD), which includes heart attack, stroke, and coronary artery disease can all be predicted through machine learning models, and then preventive measures can be employed to minimise the risk of developing said diseases. Predictive models can analyse variables including age, gender, blood pressure, and lifestyle habits to predict an individual's risk of developing CVD. People identified with high-risk of developing CVD can then take preventive measures such as lifestyle changes (e.g. exercise and diet), medication, and scheduled monitoring to reduce the risk of developing the disease. Similarly, predictive modelling can be utilised for forecasting and preventing various other diseases, for instance cancer, as previously mentioned. (30.)

An application of these disease prediction methods can be seen in telemedicine. Telemedicine refers to remote healthcare services, which can be remote monitoring (e.g. wearable heart rate monitors that forward the data to medical professionals), virtual meeting (e.g. video call or phone

call), or online instructions (e.g. video instructions for a medical device), for instance (31). By utilising predictive analytics in telemedicine, healthcare providers can remotely analyse patient data to predict the risk of potential diseases. This not only improves healthcare services, but also their accessibility. Patients have better access to monitoring and treatment even in more remote locations, such as countryside. Additionally, telemedicine enables continuous patient observation, allowing for faster measures in times of risk and emergencies. (32.)

4 PERSONALISED TREATMENT OPTIMISATION

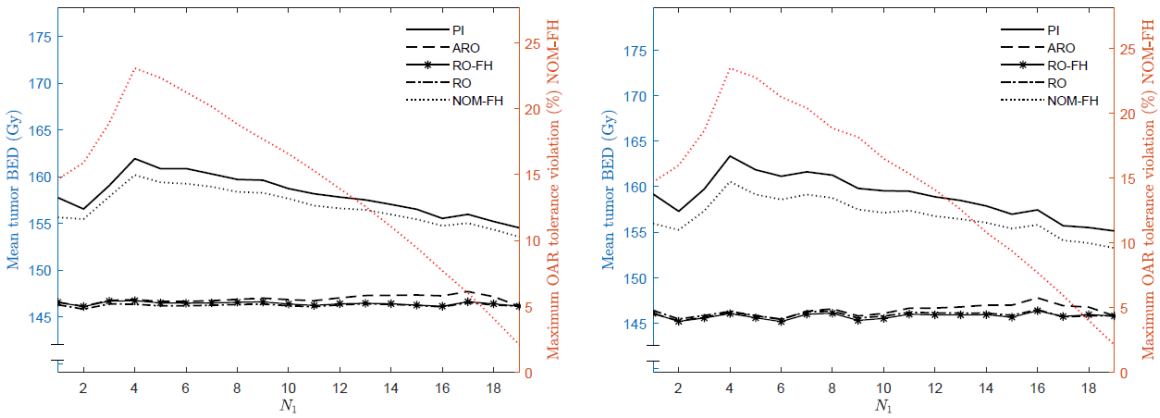
Personalised treatment optimisation refers to customising medical treatments based on each patient's personal characteristics, including genetics, lifestyle, and medical history. People have varying body types, meaning some may react to specific medicine or treatment differently than others, even if they have similar physiques (33). For example, patients with allergies or medical conditions may require different dosages or even alternative treatment options from others, to avoid negative side effects (34).

Machine learning algorithms can identify important risk factors that can assist in choosing the correct treatment options for patients. Traditional treatment methods often disregard the possible variances in patients, which results in unsatisfactory outcomes. By utilising predictive analytics to analyse health related data, healthcare providers can ensure that patients receive suitable treatments for their needs. This approach can lead to more successful treatment outcomes, such as higher survival rates, reduced disease progression, and improved quality of life. (34; 35.)

Analysing large datasets (big data), allows medical professionals to identify patterns and trends regarding patient care. This information can be utilised to predict treatment outcomes which, for instance, will assist medical professionals anticipate how patients will react to specific treatment methods. It is important to ensure that patients receive the most effective treatment, with minimal negative side effects. (34.)

For example, personalised treatment optimisation is utilised in oncology (cancer diagnosis and treatment) to improve radiation therapy through advanced techniques such as Adjustable Robust Optimisation (ARO) and Bayesian learning. Treatment plans are instantaneously modified by ARO based on variable biomarker data, as can be seen in figure 5, which illustrates a couple of uncertainty sets. These sets represent the initial variability in patient data, such as biomarkers or early imaging results, and updated data collected during treatment, reflecting ongoing changes and uncertainties. Biomarker data includes measurable biological indicators, such as molecules, that provide information about the body's condition, indicating whether there are unusual signs or not, or how the body reacts to a particular treatment method (36). The ARO technique ensures that as a patient's condition changes, the treatment plan is modified to maintain effectiveness. This method

assists in optimising the therapy in real time. Bayesian learning is a method that continuously updates the parameters of tumour response during the treatment, therefore enhancing the accuracy and effectiveness of the treatment plans. This process is shown in figure 6 (split into two parts for better visibility), where new data from treatment is utilised to improve predictive treatment models. Personalised treatment optimisation in oncology is enhanced by these methods, which consider the patient's initial conditions and adjust the treatment in real time for better treatment outcomes. (37.)



(a) DQ_1 (convex), uncertainty set 1 (b) DQ_1 (convex), uncertainty set 2

FIGURE 5. "Adjustable Robust Optimization to account for dynamic biomarker uncertainty". By MGH Radiation Oncology Physics Division (37).

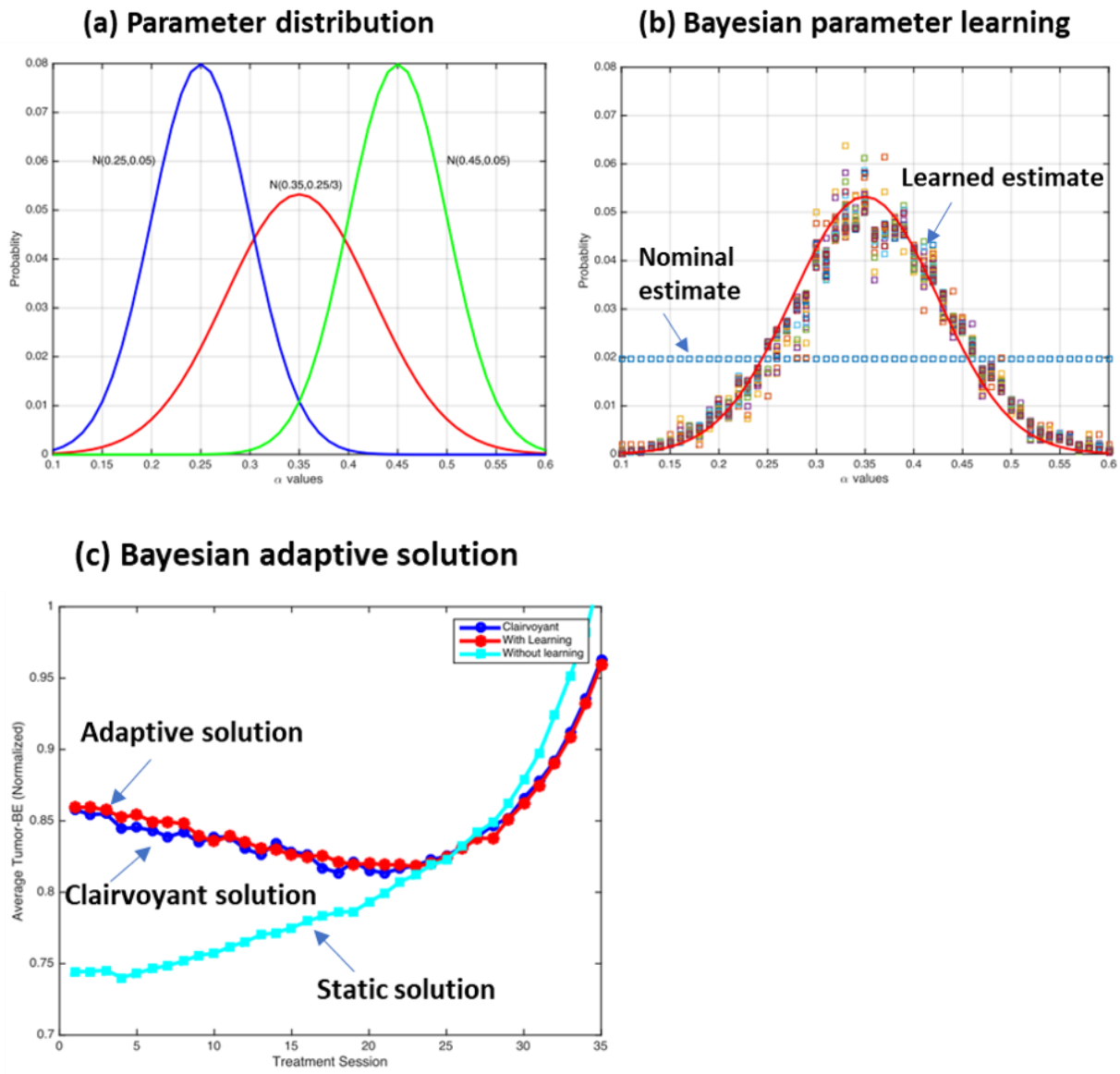


FIGURE 6. This image illustrates the Bayesian learning process that continuously updates tumour response parameters during treatment. By MGH Radiation Oncology Physics Division (37).

5 CASE STUDIES AND SUCCESSFUL APPLICATIONS

Numerous healthcare institutions have successfully applied predictive analytics to improve decision making, patient outcomes, and healthcare in general. Such institutions include Geisinger Health System, Mayo Clinic, and University of Chicago Medicine.

Geisinger Health System is focused on improving healthcare through the utilisation of predictive analytics. For instance, the organisation applies predictive analytics models to analyse EHRs, among other patient data, identifying patients at risk of manifesting chronic diseases including diabetes and CVD. Once high-risk individuals are identified, preventive measures are employed. These measures may be for instance, lifestyle changes (e.g. healthier lifestyle), or regular screenings and tests. Overall, the use of predictive analytics by Geisinger improves patient outcomes, reduces healthcare costs (e.g. through decreased hospitalisations), and enhances the accuracy of healthcare services. (38.)

Mayo Clinic utilises predictive analytics models to enhance patient care and treatment results. Mayo Clinic has developed a Mayo Clinic Early Warning Score (MC-EWS) which is a machine learning algorithm designed to predict the worsening condition of a patient in general hospital settings. By analysing patient data, such as laboratory tests and vital signs, the MC-EWS identifies patients at risk of developing life threatening conditions, such as sepsis and cardiac arrest. Early prediction of these conditions significantly improves the patients' chances of survival, as it allows for early intervention. (39.)

The University of Chicago Medicine (UCM) utilises predictive analytics to improve patient care, such as predicting the risk of rehospitalisation, and the cause of rehospitalisation. For instance, in a research publicly available in the National Library of Medicine (40), UCM developed machine learning models to forecast the rehospitalisation and its cause among individuals experiencing severe worsening of chronic pulmonary disease (COPD). With accurate predictions, UCM can identify patients with high-risk which allows for early countermeasures and better patient results. This includes optimised discharge planning and continued post-release support, to reduce rehospitalisation rates. Overall, this leads to better quality of care and lower healthcare costs. (40.)

In conclusion, the implementation of predictive analytics has proven to be crucial in healthcare, as demonstrated by Geisinger Health System, Mayo Clinic, and the University of Chicago Medicine. By utilising predictive analytics, these institutions have been able to detect people prone to developing chronic diseases or worsening conditions, allowing for earlier intervention and preventive measures. This improves patient care and outcomes, leading to reduced healthcare costs.

6 CONCLUSION

This thesis has provided research on the usage of predictive analytics in healthcare, with a particular emphasis on big data for disease prevention and treatment. As the research indicates, predictive analytics can significantly improve healthcare by optimising treatment plans, personalising patient care, and enabling early disease prediction, among other benefits. Predictive analytics techniques are employed to accurately predict diseases such as cancer, which enables the possibility for prevention or early treatment, improving patient outcomes.

However, the research has also shown that there are several challenges and considerations that must be taken into account when utilising predictive analytics in healthcare. Health related data is private information, therefore addressing security and privacy issues is essential. Ethical concerns regarding the use of patient data must also be managed. Additionally, ensuring compliance with regulations such as the HIPAA and GDPR must be maintained. Furthermore, data professionals must continually upgrade their skills to keep up with the constantly advancing technology.

Another important consideration is the fact that machines can also make mistakes. While predictive analytics offers various benefits and valuable insights, it is essential to remember that these methods are not perfect. They do not have the right answers to everything and should not be blindly relied upon, especially in healthcare. Instead, they should be regarded as advanced tools that assist healthcare professionals in their decision making. The healthcare professionals should always exercise caution and use their own judgement to make the final decision. With human oversight, predictive analytics can be utilised to its best potential, minimising the risks associated with its use.

In summary, the research indicates that predictive analytics can significantly improve healthcare by enabling more accurate disease prediction and treatment plans. However, to get the full advantage of predictive analytics, various challenges must be addressed first. Future research should focus on developing methods for managing these challenges.

REFERENCES

1. Batko, Kornelia & Ślęzak, Andrzej 2022. The use of Big Data Analytics in healthcare. Journal of Big Data 9 (3). Search date 14.4.2024. https://doi.org/10.1186%2Fs40537-021-00553-4

2. National Human Genome Research Institute 2018. Genetics vs. Genomics Fact Sheet. Search date 14.4.2024. https://www.genome.gov/about-genomics/fact-sheets/Genetics-vs-Genomics

3. Smowltech 2023. The 5 Vs of Big data: what they are and how to apply them correctly. Search date 20.4.2024. https://smowl.net/en/blog/big-data-5v/

4. Eastwood, Brian 2023. How to Navigate Structured and Unstructured Data as a Healthcare Organization. Search date 20.4.2024. https://healthtechmagazine.net/article/2023/05/structured-vs-unstructured-data-in-healthcare-perfcon

5. Touro University Illinois 2021. Applications and Examples of Big Data in Healthcare. Search date 20.4.2024. https://illinois.touro.edu/news/applications-and-examples-of-big-data-in-healthcare.php

6. VIE Healthcare. Clinical Analytics For Hospitals. Search date 19.4.2024. https://vie-healthcare.com/healthcare-data-analytics/clinical-data/

7. Aetna International. Technology and information is impacting every aspect of our lives – including our health and the care we receive. Search date 19.4.2024. https://www.aetnainternational.com/en/about-us/explore/future-health/how-biometrics-can-keep-you-healthy.html

8. Mars, Maurice & Scott, Richard E. 2022. Chapter 1 Electronic Patient-Generated Health Data for Healthcare. Digital Health [Internet]. Linwood, Simon Lin, editor. Brisbane: Exon Publications. Search date 19.4.2024. https://doi.org/10.36255/exon-publications-digital-health-patient-generated-health-data

9. Chen, Junhan & Wang, Yuan 2021. Social Media Use for Health Purposes: Systematic Review. Search date 19.4.2024. https://doi.org/10.2196%2F17917

10. McElhaney, Trevor 2023. Unveiling the Pulse of a Medical Practice with Financial Analysis. Search date 18.4.2024. https://www.doctorsmanagement.com/blog/a-comprehensive-financial-analysis-unveiling-the-pulse-of-a-medical-practice/

11. Kazmi, Robert 2023. Utilizing Financial Data in Healthcare. Search date 18.4.2024. https://www.koombea.com/blog/financial-data-in-healthcare/

12. Institute of Medicine (US) Committee on Health Research and the Privacy of Health 2009. Chapter 3 The Value, Importance, and Oversight of Health Research. Beyond the HIPAA Privacy Rule: Enhancing Privacy, Improving Health Through Research. Nass, Shary J., Levit, Laura A. & Gostin, Lawrence O., editors. Washington DC: National Academics Press. Search date 19.4.2024. https://doi.org/10.17226/12458

13. Kruse, Clemens Scott; Goswamy, Rishi; Rawal, Yesha & Marawi, Sarah 2016. Challenges and Opportunities of Big Data in Health Care: A Systematic Review. JMIR Med Inform 4 (4). Search date 15.4.2024. https://doi.org/10.2196/medinform.5359

14. Raghupathi, Wullianallur & Ragupathi, Viju 2014. Big data analytics in healthcare: promise and potential. Health Information Science and Systems 2 (3). Search date 15.4.2024. https://doi.org/10.1186%2F2047-2501-2-3

15. Ganesh Y S 2023. Challenges Faced by Business Enterprises on big data analysis. Search date 18.4.2024 https://www.linkedin.com/pulse/challenges-faced-business-enterprises-big-data-analysis-ganesh-y-s

16. International Association of Business Analytics Certification 2023. The Ethical Implications of Big Data Analytics. Search date 18.4.2024. https://iabac.org/blog/the-ethical-implications-of-big-data-analytics

17. Andreotta, Adam J., Kirkham, Nin & Rizzi, Marco 2021. AI, big data, and the future of consent. AI & Society 37 (4), 1715-1728. Search date 10.5.2024. https://doi.org/10.1007/s00146-021-01262-5

18. VIE Healthcare. Pros & Cons of Big Data in Healthcare. Search date 10.5.2024. https://viehealthcare.com/healthcare-data-analytics/big-data/pros-cons/

19. Temidayo Jacob 2023. The Impact of Big Data on Healthcare Decision Making. Analytics Vidhya. Search date 10.5.2024. https://www.analyticsvidhya.com/blog/2023/01/the-impact-of-big-data-on-healthcare-decision-making/

20. Office of the Data Protection Ombudsman. Pseudonymised and anonymised data. Search date 10.5.2024. https://tietosuoja.fi/en/pseudonymised-and-anonymised-data

21. Basu, Treena; Engel-Wolf, Sebastian & Menzer, Olaf 2020. The Ethics of Machine Learning in Medical Sciences: Where Do We Stand Today? Search date 10.5.2024. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7640783/

22. Yang, Christopher C. 2022. Explainable Artificial Intelligence for Predictive Modeling in Healthcare. Journal of Healthcare Informatics Research 6 (2), 228–239. Search date 27.4.2024. https://doi.org/10.1007%2Fs41666-022-00114-1

23. Sidey-Gibbons, Jenni A. M. & Sidey-Gibbons, Chris J. 2019. Machine learning in medicine: a practical introduction. BMC Medical Research Methodology 19 (64). Search date 27.4.2024. https://doi.org/10.1186/s12874-019-0681-4

24. Deo, Rahul C. 2015. Machine Learning in Medicine. Circulation 132 (20), 1920–1930. Search date 27.4.2024. https://doi.org/10.1161%2FCIRCULATIONAHA.115.001593

25. Mrukwa, Grzegorz 2023. Supervised and Unsupervised Machine Learning - Types of ML. Search date 27.4.2024. https://www.netguru.com/blog/supervised-machine-learning

26. Harris, Jenine K. 2021. Primer on binary logistic regression. Family Medicine and Community Health 9 (1). Search date 27.4.2024. https://doi.org/10.1136%2Ffmch-2021-001290

27. harkiran78 2024. Top 10 Machine Learning Algorithms | Data Science for Beginners. GeeksforGeeks. Search date 27.4.2024. https://www.geeksforgeeks.org/top-10-algorithms-every-machine-learning-engineer-should-know/

28. GeeksforGeeks 2024. K means Clustering – Introduction. Search date 28.4.2024. https://www.geeksforgeeks.org/k-means-clustering-introduction/?ref=lbp

29. Jaadi, Zakaria. A Step-by-Step Explanation of Principal Component Analysis (PCA). Modified by Whitfield, Brennan on 23.2.2024. Search date 28.4.2024. https://builtin.com/data-science/step-step-explanation-principal-component-analysis

30. Peng, Mengxiao; Hou, Fan; Cheng, Zhixiang; Shen, Tongtong; Liu, Kaixian; Zhao, Cai & Zheng, Wen 2023. Prediction of cardiovascular disease risk based on major contributing features. Scientific Reports 13 (4778). Search date 3.5.2024. https://doi.org/10.1038/s41598-023-31870-8

31. MedlinePlus [internet]. Telehealth. Search date 28.5.2024. https://medlineplus.gov/telehealth.html

32. Shaid, Talha & Graepel, Thore. Harnessing the Power of AI in Healthcare: Remote Patient Monitoring, Telemedicine, and Predictive Analytics for Improved Clinical Outcomes. Search date 28.5.2024. http://dx.doi.org/10.13140/RG.2.2.12844.68481

33. Saptarshi Dutta 2021. Personalized Medicine through Machine Learning. Analytics Vidhya. Search date 9.5.2024. https://www.analyticsvidhya.com/blog/2021/06/personalized-medicine-through-machine-learning/?trk=article-ssr-frontend-pulse_little-text-block

34. Kumar, Deepak; Pawar, Priyanka P.; Gonaygunta, Hari; Nadella, Geeta Sandeep; Meduri, Karthik & Singh, Shoumya 2024. Machine learning's role in personalized medicine &

treatment optimization. World Journal of Advanced Research and Reviews 21 (2), 1675-1686. Search date 9.5.2024. https://doi.org/10.30574/wjarr.2024.21.2.0641

35. Black Peak Technologies 2023. Personalized Medicine: Leveraging The Power of Machine Learning for Modern Treatment. Search date 9.5.2024. https://blackpeaktechnologies.com/blog/personalized-medicine-machine-learning/

36. National Cancer Institute. Biomarker. Search date 4.6.2024 https://www.cancer.gov/publications/dictionaries/cancer-terms/def/biomarker

37. MGH Radiation Oncology Physics Division. Treatment Personalization (Optimization). Search date 10.5.2024. https://gray.mgh.harvard.edu/research/osrt/303-treatment-personalization-optimization

38. McCarthy, Douglas; Mueller, Kimberly & Wrenn, Jennifer 2009. Geisinger Health System: Achieving the Potential of System Integration Through Innovation, Leadership, Measurement, and Incentives. Search date 3.5.2024. https://www.commonwealthfund.org/sites/default/files/documents/___media_files_publications_case_study_2009_jun_mccarthy_geisinger_case_study_624_update.pdf

39. Romero-Brufau, Santiago, Whitford, Daniel, Johnson, Matthew G., Hickman, Joel, Morlan, Bruce W., Therneau, Terry, Naessens, James & Huddleston, Jeanne M. 2021. Using machine learning to improve the accuracy of patient deterioration predictions: Mayo Clinic Early Warning Score (MC-EWS). Journal of the American Medical Informatics Association 28 (6), 1207-1215. Search date 4.5.2024. https://doi.org/10.1093%2Fjamia%2Focaa347

40. Bonomo, Matthew; Hermsen, Michael G.; Kaskovich, Samuel; Hemmrich, Maximilian J., Rojas, Juan C.; Carey, Kyle A.; Venable, Laura Ruth; Churpek, Matthew M. & Press, Valerie G. 2022. Using Machine Learning to Predict Likelihood and Cause of Readmission After Hospitalization for Chronic Obstructive Pulmonary Disease Exacerbation. International Journal of Chronic Obstructive Pulmonary Disease 17, 2701-2709. Search date 4.5.2024. https://doi.org/10.2147/COPD.S379700