2023

Utilizing Machine Learning in Healthcare in an Ethical Fashion

Nishka Ayyar

Utilizing Machine Learning in Healthcare in an Ethical Fashion

Submitted to
Professor Mark Huber

By
Nishka Ayyar

For
Senior Thesis
Spring 2023
04/24/2023

Table of Contents

Acknowledgements

I would like to thank Professor Mark Huber for his patience, guidance, and support over not only over the past eight months as my thesis advisor, but the past four years as my academic advisor and professor. I've truly enjoyed every class I've taken with him, and I attribute my interest in data science to his wonderful teaching skills.

I would also like to thank my parents and sister for their support and love, and for always pushing me to be the best version of myself. Finally, I would like to thank my friends for making the past four years so exciting and memorable.

Abstract

This thesis paper explores the ethical considerations surrounding the use of machine learning (ML) solutions in healthcare. The background section discusses the basics of machine learning techniques and algorithms, and the increasing interest in their utilization in the healthcare sector. The paper then reviews and critically analyzes four studies that highlight concerns related to using ML in healthcare, including issues of bias, privacy, accountability, and transparency. Based on the analysis of these studies, the paper presents several recommendations for addressing these concerns. The paper concludes with a discussion on the potential benefits of using machine learning technology in healthcare. Ultimately, the purpose of this thesis paper is to show that while the ethical concerns around machine in healthcare are significant, they should not deter the development and adoption of such solutions in the sector, but rather inform their design and implementation to maximize their potential benefits while minimizing their risks and harm.

1. Introduction

The use of machine learning algorithms in healthcare has gained significant attention in recent years due to their potential to revolutionize various aspects of medical practice. Machine learning techniques, a subset of artificial intelligence (AI), are being increasingly applied to vast amounts of healthcare data, including electronic health records, medical images, genetic data, and wearable sensor data, among others. These algorithms have shown promise in areas such as disease diagnosis, treatment planning, drug discovery, personalized medicine, predictive analytics, and more. However, there are many challenges and limitations associated with the use of machine learning models in healthcare. In this paper, I will focus on the ethical concerns posed by the adoption of machine learning algorithms in the healthcare industry. To fully grasp these concerns from the healthcare standpoint, it is important to have a solid understanding of the machine learning component as well. Therefore, I will begin by explaining the technical aspects of machine learning in healthcare, and provide a background of types of data, machine learning techniques, and specific algorithms that are being developed for use in healthcare. I will then review four papers that describe the ethical implications of machine learning across four categories - social justice, electronic health records, algorithm fairness solutions, and intelligent autonomous care provider (IACP) use in behavioral and mental healthcare. Following this review, I will offer a comprehensive list of suggested recommendations that I found through analyzing these papers. Finally, I will conclude by sharing benefits and opportunities related to using machine learning models in the healthcare industry.

2. Background

In this section of the literature survey paper, I will explain types of data commonly used in machine learning algorithms, as well as various machine learning techniques that use this data. I will then describe machine learning algorithms that are specific to the application of machine learning in the healthcare industry.

2.1 Types of Data

Before understanding the algorithms, it is important to understand the types of data available for these algorithms to use [1]. Data can be split into four categories - structured, unstructured, semi-structured, and metadata. Structured data follows a specific structure and order and is very organized and accessible by whoever wishes to use it, and usually stored in a relational database. Examples of structured data include names, addresses, and credit card numbers. Unstructured data, on the other hand, do not follow any order, and this data is much more difficult to analyze than structured data. Examples of unstructured data include blog entries, images, and emails. Semi-structured data borrow components from both unstructured and structured data. This data is not stored in a database as structured data is, but are more organized than unstructured data. Some examples of semi-structured data include HTML files and NoSQL databases. The final data category is metadata, which is different from any of the other categories. Metadata describes the information related to data being used, such as the author's name, the file size, document keywords, etc. Figure 1 below provides a quick overview of the different types of data mentioned above.

Figure 1. Types of data.

Structured Data	• Follows a rigid structure • Easy access • Example: Address data
Unstructured Data	• No order • Difficult to analyze • Example: Email file
Semi-structured Data	• Not stored in a database • Follows a structure • Example: HTML file
Metadata	• Describes relevant data • Example: Keywords of this paper

2.2 Machine Learning Techniques

After the data have been collected, machine learning techniques can be applied to analyze the data and provide useful insights. Machine learning techniques can be divided into four distinct categories: supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning [2].

Supervised learning involves training a model on pre-labeled data to make predictions about new data [2]. The model is fed sample input/output pairs and must learn to map new inputs to their correct outputs from the sample data. For supervised learning models, datasets are split into training and testing sets, and the model is evaluated using performance metrics. An example of a supervised learning algorithm would be text classification. On the other hand, unsupervised learning involves training a model on unlabeled data to discover patterns and group similar data. Because the data are unlabeled, unsupervised learning algorithms do not have a way to evaluate model performance. Instead, unsupervised learning tasks include density estimation, dimensionality reduction, and other data-driven processes.

Then there is semi-supervised learning, which is a combination of supervised and unsupervised learning where both labeled and unlabeled data are used [2]. Semi-supervised learning is the ideal machine learning technique when unlabeled data is more available than labeled data, but both exist. In this case, model performance can often be improved by using the two together. This learning technique is commonly used for outlier detection and is commonly used for trying to achieve a prediction outcome that would be more accurate than just using the limited available labeled data. Credit card fraud detection is an example of semi-supervised learning.

Finally, reinforcement learning involves developing a model that iteratively learns from environmental feedback and uses this learning to self-improve its performance [2]. Reinforcement learning is based on a reward/penalty model, where the goal is to increase reward or minimize risk. One of the benefits of reinforcement learning is that it does not require human interference or assistance, and some example applications are robotics and autonomous driving.

2.3 Relevant Algorithms

Some of the algorithms used most often in healthcare are classification analyses, which is a supervised learning predictive modeling technique. Under the broad term of classification, the most popular algorithms in the healthcare industry are binary classification, naive Bayes, logistic regression, K-nearest neighbors, support vector machines, decision trees, and random forests. In this section, I will briefly describe each of these algorithms.

Binary classification tasks have two labels, usually similar “true” and “false” or “yes” and "no" [3]. One of the labels is the normal state, and the other is the abnormal state. The binary classification algorithm receives a dataset and can label the data as either the normal state or

abnormal state. Naive Bayes is an algorithm based on the Bayes' theorem of probability that assumes feature independence. One of the benefits of naive Bayes is that it only requires a small training set, but this and the assumption of independence may affect performance. The logistic regression model is another probabilistic model, but rather than being based on a theorem like naive Bayes, the logistic regression model uses a logistic function (sigmoid function) for probabilistic estimation. There are many advantages to using the logistic regression model, such as its ability to overfit high-dimensional data, but the main fault is that it assumes independent-dependent variable linearity. Additionally, though the name might suggest otherwise, logistic regression models are more often used for classification problems than regression problems.

The next algorithm, K-nearest neighbors or KNN, is often called a "lazy learning" algorithm [3]. Rather than creating a model using its data, the KNN algorithm classifies the test set using a given similarity measure. The training data points are stored in n-dimensional space, and the proximity to k-nearest neighbors of each point is measured through a majority vote. While this method is relatively unaffected by noisy data, the one drawback is that there is no pre-defined optimal number of nearest neighbors to use, so this must be chosen by the person running the algorithm, which can be tricky. Support vector machines are a versatile model that depend on hyper-planes. They aim to find the best hyper-plane or set of hyper-planes that can separate input data into different classes. These are useful because they have high accuracy and can handle high-dimensional data but are computationally expensive and do not perform well with noisy data. Decision trees and random forests go hand in hand. Decision trees separate input data into smaller subsections based on similarity of features. An example of a decision tree structure can be seen below. Random forests extend the decision tree algorithm by creating an aggregate of many decision trees to achieve a better predictive accuracy than just one single

decision tree. The algorithm does this by either averaging the decision trees or using majority voting. Random forest models can improve decision tree accuracy by minimizing the overfitting that is common with decision trees.

3 Machine Learning Solutions in Healthcare

Healthcare is a complex field with vast amounts of data generated from various sources, including electronic health records, medical imaging, genetic data, wearable devices, and health sensors, among others. The rapid advancement of technology has led to the availability of big data in healthcare, which presents both opportunities and challenges for improving health care delivery. Machine learning, a subset of artificial intelligence (AI), has emerged as a promising approach to harnessing the potential of big data in healthcare. Machine learning algorithms have been increasingly applied to analyze and extract insights from large healthcare datasets, with the goal of improving patient outcomes, optimizing clinical decision-making, and enhancing healthcare delivery.

As machine learning algorithms and artificial intelligence gain traction in several fields, healthcare is an area that has seen rapid growth in using machine learning to solve various issues. These range from data irregularity detection, disease progression prediction, diagnostic assistance, patient data management, and more. In fact, with the emergence of the Covid-19 pandemic over the past three years, machine learning techniques have grown even more prevalent to aid physicians in the management of the new disease. Some of the ways in which ML has been used during the Covid-19 pandemic are for high-risk patient classification, outbreak timing prediction, outbreak location prediction, and for diagnostic and treatment recommendations. In the graphs below, I track the progress of machine learning and healthcare,

as well as the growth during the Covid-19 pandemic specifically, through a keyword search in the PubMed database.

Figure 2. Charting the growth of AI/ML keyword searches related to general healthcare [4].

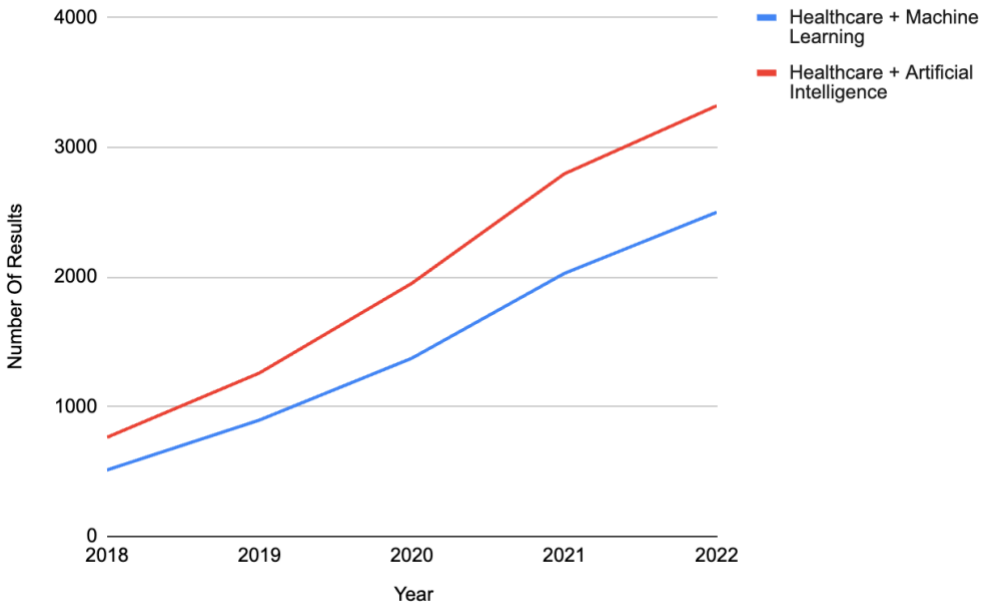
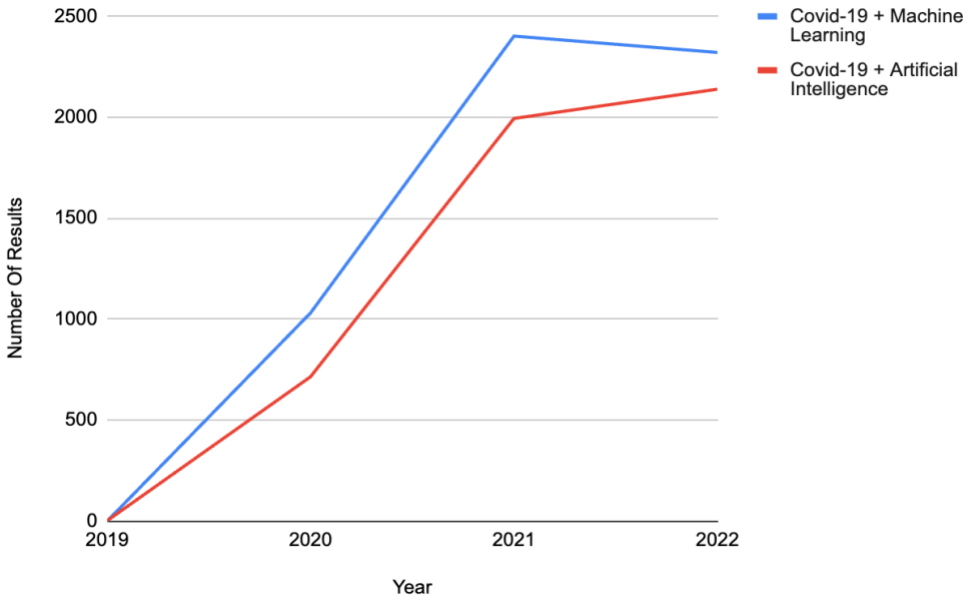


Figure 3. Charting the growth of AI/ML keyword searches related to the Covid-19 pandemic [5].



As seen from these figures, interest in machine learning as applied to healthcare has grown significantly in recent years, particularly in relation to the Covid-19 pandemic in the last three years, showing a real area for research and development.

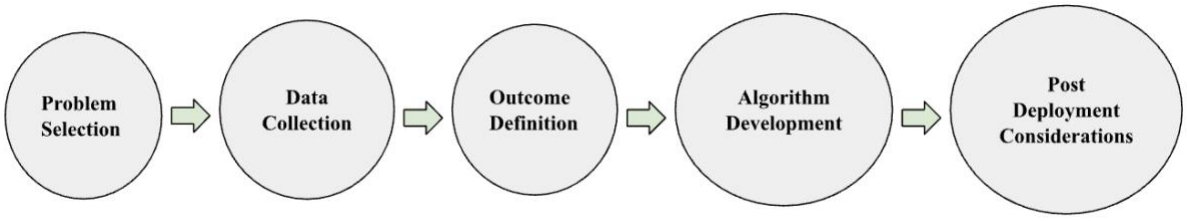
3.1 Ethical Machine Learning in Healthcare

The review paper "Ethical Machine Learning in Healthcare" by Irene Y Chen, Emma Pierson, Sherri Rose, Shalmali Josh, Kadija Ferryman, and Marzyeh Ghassemi describes the ethical concerns related to social justice in healthcare machine learning [6]. In addition to describing the implications, the authors highlight current attempts to resolve these issues, as well as their recommendations to the issues. The paper draws upon a large body of literature to conduct the review.

3.1.1 Model Pipeline

The authors begin by outlining the model development pipeline for healthcare machine learning. In this section, they study the case of death during childbirth of Black women. The inequality studied here is that Black mothers die at a rate of three times that of white mothers, due to a history of unequal reproductive rights of Black women. The pipeline is organized as follows:

Figure 4. Machine Learning Model Development Pipeline.



Problem Selection: Biases during problem selection can occur if the algorithmic problem pertains to an understudied area, or one that requires diversity of thinking, which is lacking in the industry today. For example, one of the reasons that this ethical inequality occurs in the problem of maternal death during childbirth is because this issue does not receive as much research attention as it should.

Data Collection: When dealing with healthcare data, there can be many discrepancies - in this case, many Black women go into labor at predominantly black serving hospitals, but these also have a higher rate of complications during childbirth.

Outcome Definition: After data collection, there are still many ways in which bias can make its way into algorithmic development, such as knowledge bias, healthcare system disparities, social differences, etc. In this problem, simply looking at maternal childbirth deaths would not be helpful when dealing with the mortality rate of Black mothers specifically.

Algorithm Development: Models cannot always account for non-technological biases, such as social or income bias, which is very prevalent when dealing with issues of racial minority communities.

Post-deployment Considerations: after the algorithm is trained and released, considerations about biased predictions may not be considered in clinical healthcare settings. For example, in maternal childbirth deaths in black communities, algorithms may make recommendations without understanding that these policies could end up disadvantaging Black women because they do not pertain specifically to that community.

3.1.2 Ethical Challenges in Pipeline Stages

After outlining the pipeline, the authors then delve into the relevant ethical dilemmas common for each stage of the model pipeline. In this section, I will summarize these stages as described in the paper.

3.1.2.1 Problem Selection

The authors specify four specific areas that cause ethical issues in healthcare machine learning problem selection - global health injustice, racial injustice, gender injustice, and diversity of the scientific workforce. While this section is not directly related to algorithms or development, it is important to understand underlying healthcare biases when learning about the ethical implications of machine learning in the healthcare industry.

Global Health Injustice: This ethical dilemma deals with the fact that there is a great funding disparity in healthcare, and most poverty-stricken countries receive a miniscule amount of funding for their diseases.

Racial Injustice: Racial Bias is a very common ethical issue in healthcare, as diseases that affect white populations on a larger scale, like cystic fibrosis, receive much more funding than those that affect other populations, such as sickle cell diseases.

Gender Injustice: Women's health often receives far less time, attention, and money than illnesses and conditions that affect men. For example, menstrual issues and endometriosis are often stigmatized and people do not have basic knowledge on these issues → this sentiment carries into clinical medicine and women's menstrual health is extremely understudied.

Diversity of the Scientific Workforce: Because of the lack of diversity in the scientific community, research proposals made by non-white scientists are often rejected, even though they often prioritize understudied groups. To combat this, the scientific workforce must be diversified to allow for a more equitable problem selection process.

3.1.2.2 Data collection

Data collection, while a crucial part of the model development process, can cause many different ethical issues if done improperly or if biases are present. This section of the paper deals with the two overarching types of data issues - heterogeneous data loss and population-specific data loss.

Heterogeneous data loss is specific to the type of data being studied. This section discusses four types of data - randomized control trials, electronic health records, administrative health data, and social media data.

Randomized control trials are used to try and minimize the existence of bias in treatment. However, these trials often misrepresent populations, and require many assumptions to be made along the way that can skew the results and provide improper treatment that does not represent the entirety of a general patient population.

Electronic health records make up a large percentage of healthcare machine learning data. These records include those of patients, organizations, providers, and other stakeholders. There are many reasons for EHR data biases, including gender discrimination, income disparities, access to care, or EHR system availability.

Administrative health data is another portion of healthcare data that includes insurance documents, survey data, and other administrative work. These records often face ethical problems because there is strong evidence of population discrimination based on sexual orientation, gender, race, and ethnic identity, and spoken language capabilities. As a result, these policies are often unjust, which creates a severe ethical implication for the use of administrative health records.

Social media data, which has grown more important in recent years, can skew analyses and create issues in algorithmic development due to bias. One of the main reasons for this is that when social media data is scraped, the sample of individuals being taken will never be random. Additionally, data samples from social media are often limited to categories like type of device, geographic location, or frequency of use. These restrictions also will lead to skewed data and results.

Another main contributor to data loss is the misrepresentation or underrepresentation of certain communities. This leads to bias because when algorithms output recommendations or results, these groups are then not taken into consideration, and the recommendations will not apply to the general population. The four types of population-specific data loss studied in this paper are low- and middle-income nationals, transgender and gender-nonconforming individuals, undocumented immigrants, and pregnant women.

3.1.2.3 Outcome Definition

The outcome definition section of this paper separates social justice issues into two healthcare outcomes that are generally used for predictive modeling: clinical diagnosis and healthcare costs.

Recommendation algorithms are often used to diagnose patients in a clinical setting. One of the biggest implications with this machine learning approach in healthcare is label noise, which means that people can present in different ways while having the same diagnosis. Healthcare costs are also predicted using machine learning models - the models predict the risk of a patient to the healthcare provider, where high risk patients will likely cost the healthcare provider more. However, there is a high risk of bias in these algorithms, as healthcare costs are heavily influenced by socioeconomic circumstances.

3.1.2.4 Algorithm Development

The next step in the model pipeline is algorithm development. However, algorithms pose many ethical issues as there can be numerous sources of bias at this stage. There are four factors that contribute to ethical issues in the algorithm development stage of the model pipeline: understanding confounding, feature selection, tuning parameters, and defining fairness.

Confounding features are ones who impact the independent and dependent variables. When models are trained, they identify patterns based on what they find in the training data, but confounding features can cause models to incorrectly associate variables together even when there is no relationship. It is not enough to just control for these features. To reduce confounding effects, it is necessary to have the model design be created differently to minimize the impact of having no confounding features.

Feature selection simply refers to the selection of various features to be included when analyzing data in a machine learning model. To prevent model bias, a measure that is sometimes taken is the blind addition of factors like race, gender, and income. However, this can have the opposite effect, and cause diagnostic and treatment inequalities after the model is created and used. A way to assist with this is by making sure to understand the model and those will impact before using automated feature selection.

Tuning parameter biases are often the result of data that lacks diversity. When these parameters are left at default and not adjusting depending on the data, this leads to model overfitting in the training set, which results in the inability to make recommendations to an entire population. This can then harm groups that are underrepresented in the data, such as minorities or marginalized groups. Performance metrics can lead to ethical issues in algorithm development when not enough time or effort goes into carefully choosing which metrics to use. If the incorrect

metrics are chosen, this can lead to the misdiagnosis or mistreatment of a subsection of the population, oftentimes minorities and marginalized groups again.

Defining fairness means essentially what it says. How a model defines fairness can influence which loss function is chosen for the model. If this is done incorrectly, algorithms can violate fairness definitions which then leads to bias, and individuals can once again be misdiagnosed or under- or over-treated.

3.1.2.5 Post-deployment Considerations

The last portion of the pipeline is post-deployment considerations. The ethical implications during this step can be very serious, as they last long past the deployment of a model and extend beyond performance. The four factors that play into post-deployment ethical implications are quantifying impact, model generalizability, model and data documentation, and regulation.

Quantifying impact post-deployment means that models must be regularly audited after they are released, particularly on different populations. Once this is done, areas of concern must be identified for future addressing. Generalization biases have been identified in almost every step of the model pipeline because it is extremely difficult to have a model that is unbiased such that its recommendations can be applied to a general population. Once a model has been deployed, it must continue to be assessed for generalizability, so see what biases are causing this hindrance. Another ethical issue that arises post-deployment is the lack of straightforward documentation. When model documentation is incomplete or inadequate, this can cause ethical issues and lead to performance and bias mishaps. Lastly, the paper mentions regulation. At this point, the FDA regulates all machine learning models used in the healthcare industry. However,

there is no comprehensive regulatory framework that covers every base and deals with the ethical implications of all types of models.

3.1.3 Proposed Recommendations

The final section of this paper offers recommendations for the ethical concerns raised in each stage of the model pipeline. The recommendation section offers five suggestions, one for each step in the pipeline:

Problem Selection - diversity in personnel and frameworks should be increased in the problem selection stage to increase the chance of achieving equity and problems that are generally overlooked should receive time and attention.

Data Collection - Biases and data loss should be a high priority consideration when data is being collected, and the data should be reflective of all population groups, including minorities.

Outcome Definition - If ethical issues and biases are present in outcome labels, then the model design should be improved to accommodate for such biases to minimize and remove them.

Algorithm Development - Algorithm goals should be clearly defined before and during development. Drawbacks should be considered prior to deployment.

Post-deployment Considerations - Audits should be performed continuously after model deployment to examine potential harm and ethical issues.

3.2 Machine Learning in Healthcare

"Machine Learning in Healthcare" by Nigam Shah and Alison Callahan discusses the application of machine learning to electronic health record (EHR) data analysis in clinical settings, highlighting its advantages over traditional analysis methods such as epidemiology [7]. It provides an overview of the different data analysis approaches, including inferential and predictive analyses, and explains how machine learning fits into algorithmic modeling. The paper also discusses the challenges of using machine learning in research and practice, as well as the potential opportunities for impacting health and healthcare delivery. In this section, I will leave out potential benefits described in this paper, as that can be found in section six of my thesis.

The paper divides data analysis into five overarching categories: descriptive, exploratory, inferential, predictive, and casual. The two categories that are important for this discussion about machine learning in healthcare are inferential and predictive. Inferential analyses measure a model's goodness of fit, while a predictive analysis creates a predictive, statistical model from observations, and uses performance metrics (accuracy, discrimination, recall, calibration, predictive value, specificity) to determine which predictions are correct. This paper mainly focuses on EHR, or electronic health data, to which they claim a machine learning application is the next big step in modern clinical medicine.

One application of machine learning is to classify patients that are "at risk" for a certain medical condition. In the past, scoring systems such as the DLCN criteria, Charlson Comorbidity Index, or APACHE score have been used, but they are not extremely accurate. For brief context, the Dutch Lipid Clinic Network, or DLCN, is a scoring system to diagnose patients as "unlikely," "possible," "probable," or "definite" for likelihood of having familial hypercholesterolemia. The Charlson Comorbidity Index predicts an individual's 1-year risk of

death in relation to their burden of disease. The APACHE score provides a scoring for ICU patients that tells them how severe their disease is. Machine learning algorithms can be utilized based on similar criteria used by these scoring systems and others to create predictive models that can accurately predict risk factors associated with disease. However, the development of EHR-based algorithms and machine learning models poses its own set of limitations and challenges. Firstly, patients are likely only visiting healthcare professionals when they are exhibiting symptoms of illness, which can skew a person's risk assessment score. Additionally, missing data and data loss are a prevalent issue when using EHRs. Patients change between healthcare systems all the time, and records are often lost or re-coded incorrectly. Another major issue in this area is the lack of generalizability, or the ability to generalize the recommendations of a model to an entire population, which results in the ethical dilemma of population sectors (often minority groups) being overlooked or misrepresented. However, to combat these ethical issues, data must be easily accessible, competent machine learning engineers must be able to develop unbiased models and interpret them, and easy integration into the healthcare system must be thought through.

3.3 Ethical Limitations of Algorithmic Fairness Solutions in Health Care Machine Learning

"Ethical Limitations of Algorithmic Fairness Solutions in Health Care Machine Learning" is a paper by McCradden, M. D., Joshi, S., Mazwi, M., and Anderson, J. A. It is widely understood that there are many ethical issues involved with using machine learning in healthcare [8]. As a result, algorithmic fairness solutions have been developed to combat these issues to create neutral models. However, as this paper claims, these solutions can be ethically

and empirically problematic, particularly when there is too much reliance on the idea of technical neutrality. In this section, I will summarize the challenges and limitations of using algorithmic fairness solutions to address pernicious bias in health data, as explained by the paper.

3.3.1 Ethical Challenges and Limitations

One of the first ethical mistakes that arises with algorithmic fairness solutions is the overlooking of biological, environmental, and social elements that influence medical illnesses. These factors play a big role in healthcare and general wellbeing, but their impact is often overlooked and not well-understood Additionally, with biological differences, it is very difficult to distinguish between a biological difference necessitating different treatments and one that will promote unequal treatment if recommendations vary. The authors suggest that only causative associations should result in the incorporation of biological differences in model development. Another issue is that of a disconnect between predicted response to a treatment and actual response. In healthcare, this can have serious implications, and can prevent prompt interference. If this disconnect occurs with a "neutral" model, then the disconnect could go unnoticed by both clinicians and patients. To combat this issue, all medical consequences should be reviewed prior to fairness solution deployment. To minimize pernicious bias in algorithmic fairness solutions, the authors suggest developing a standard set of guidelines for machine learning model reporting as well as studying the resulting real-world consequences.

The paper also emphasizes the importance of transparency, accountability, and clinical trials in machine learning model development and algorithmic fairness solution management. Transparency is needed at every stage of the model pipeline (see figure 4). When this is combined with straightforward model documentation, an adequate accountability system is

developed for fairness solutions. Clinical trials are important in that they can assess model performance before deployment and allow physicians to make data-driven judgements about their care given any drawbacks seen in the model.

3.3.2 Recommendations

The authors also outline six recommendations for combating the ethical bias issue with machine learning models in healthcare. Firstly, machine learning solutions should be thought of subjectively, as to prevent reliance on neutral algorithms. Next, it is important to carefully think through and conceptualize machine learning problems to minimize pernicious, or destructive, bias. Additionally, transparency regarding model development and prediction, as well the physician-driven transparency during patient care is extremely important. There should be a standard reporting mechanism for machine learning models, as well as continuous ML solution auditing both pre- and post-deployment. Finally, it is recommended that when employing machine learning solutions in the healthcare field, to always remember that these are real people being affected, and the potential consequences should not be taken lightly.

3.4 Ethical Issues and Artificial Intelligence Technologies in Behavioral and Mental Health Care

Chapter 11 of "Ethical Issues and Artificial Intelligence Technologies in Behavioral and Mental Health Care" by Luxton, D. D., Anderson, S. L., and Anderson, M. discusses ethical issues concerning using AI technology in behavioral and mental health care [9]. Specifically, the section focuses on autonomous AI, such as robots and virtual caretakers - intelligent autonomous care providers (IACP's). This subfield is known as robo-ethics.

3.4.1 Ethics Background

The authors define the four parts of medical ethics to be respect for autonomy, beneficence, nonmaleficence, and justice. Respect for autonomy refers to a patient's free will, including rights to self-determination and the reception of full disclosure of information. This cornerstone allows for patients to make informed decisions about their healthcare. Beneficence is the assumption that the motive of healthcare providers is to improve the wellbeing of their patients. Nonmaleficence, or do no harm, means that physicians and care providers will not act in a harmful way to patients or society. Finally, justice refers to the fact that patients in similar health situations should be treated similarly. Additionally, under justice, the impact of resource allocation should be constantly assessed and re-assessed.

The chapter also outlines the history of medical ethics, of which the first formal ethical code was published in 1847 by the American Medical Association. There were a few follow-up codes by the British Medical Association and World Medical Association, and in 1949, the WHO (World Health Organization) released the International Code of Medical Ethics. Many other medical associations have since followed suit, including the American Psychiatric Association, American Psychological Association, and American Counseling Association.

The authors then go on to define roboethics and machine ethics. Roboethics is the subfield of ML ethics that has to do with the creation and design of robots, as well as how people interact with robots. Some ethical concerns relating to roboethics include the treatment of robots, whether they have rights, and whether mistreating robots will lead to the eventual, subconscious mistreatment of humans. Machine ethics, on the other hand, involves building IACP's in a way that forces them to act ethically towards humans and other machines, and gives them the

capability to make ethical decisions. Machine ethics is important because it reduces the need for damage control by installing preventative measures in robots through built-in ethical principles.

3.4.2 Ethical Challenges

Next, the paper details specific ethics challenges: Therapeutic Relationships and Emotional Reactions, Competence of Intelligent Machines, Patient Safety, Respect of Privacy and Trust, Deception and Appearance, and Responsibility. In this subsection, I will describe each of these challenges.

Therapeutic Relationships: These are the professional relationships between healthcare providers and their patients. Because physicians and providers are in positions of power, there is a possibility for patients to be exploited or harmed. In mental healthcare specifically, which is one of the main topics of this paper, patients often experience high emotions, and it is important for professionals to respect these emotions without violating the patient's trust or invalidating their feelings. When IACP's replicate the position of these professionals, designers must consider the ethical involvement of the robots and virtual humans and be able to design them in such a way that this does not become an issue. An important implication of this design process that has arisen through the increased use of IACP's in mental healthcare is the failure for the machines to be sensitive to emotional reactions from patients - handling these situations incorrectly could harm the patient and those around them. Similarly, it is important in a clinical psychology setting for the physician and patient to establish empathetic understanding. An ethical implication associated with using machine learning technology in this sense is called the ELIZA effect, when people mis-perceive machines to possess humanistic qualities such as emotional intelligence. For background, ELIZA was an MIT-engineered chatbot program that

allowed for natural language conversation between humans and computers [10]. ELIZA was a very intelligent program that could simulate real conversations between people, entrancing users. However, many people abused this program because they mis-perceived the intrinsic, non-human characteristics of ELIZA to be human. As a result, they would develop a dangerous relationship with the chatbot, often revealing very intimate, personal details about themselves and others. While having very humanistic computer programs can be beneficial in some cases, when the goal is to completely replace the human component of a medical interaction, but often will cause more harm than benefit. This is because the ELIZA effect can lead to patients becoming too attached to the IACPs, to a dangerous extent, and if they perceive the relationship as becoming negative in some way, have the potential to harm themselves or others.

Competence of Intelligent Machines: a general assumption when meeting with a mental healthcare professional is their competence, or ability to appropriately do their job. Incompetent medical professionals could lead to a patient being harmed and an entire profession or organization losing its credibility. When dealing with human professionals, healthcare organizations have sections in their ethical codes that highlight the importance of professional competency. However, IACP systems are not always designed to perform to the highest level of competence, which poses a serious ethical problem when dealing with the medical system.

Patient Safety: In clinical settings, patient safety can be compromised in many ways - the paper describes a situation where a patient discloses intent to harm another person, and the physician must follow duty-to-warn and notify authorities and others. However, this can compromise patient confidentiality, and many ethical dilemmas arise when trying to tread the fine line between these. The same issue is present when using IACP's, particularly because IACP's lack the human capabilities required to help make such judgements.

Respect of privacy and trust: With a human healthcare professional, privacy and trust can always be violated and misused. When data is involved, this issue becomes even more common, as the violation no longer needs to be intentional. Private data can be accidentally left unsecured, or intentionally accessed by hackers. Additionally, conversations between patients and IACP's have the potential to be recorded and saved, whereas this is less common in in-person settings. Additionally, the dual-use feature capability of such machines could result in the loss of human trust. Dual-use means that a single machine can be used for multiple purposes - for example, a machine designed for mental healthcare can be repurposed and used to interrogate a high-security prisoner. As a result, the person using the machine for mental health assistance may have decreased trust in its effectiveness and capability to keep information private and secure.

Deception and Appearance: The main issue with deception and appearance discussed in this article is called The Wizard-of-Oz effect. This occurs when patients, particularly those with mental issues of delusion or psychosis may believe that a machine has more capabilities than it does or is "alive". This can be harmful both to the patient and the machine. The patient may develop an emotional relationship with the machine, which can grow unsettling and dangerous very quickly. Additionally, patients may begin enacting force upon the machine if they believe it is alive, which could cause the machine to work less effectively.

Responsibility: The final implication posed by the paper is that of responsibility, or more specifically, who should be held responsible for the actions of IACP's. In healthcare, when machines are put in control of making healthcare decisions for patients, there can be a severe implication of responsibility if anything goes wrong, because the machines cannot be held responsible as doctors can. However, when a machine is referred to as "autonomous", this transfers some responsibility to the IACP, as it is thought of as a decision-making tool with

agency. But machines do not exhibit moral consequential feelings, such as guilt, humiliation, stress, etc. Therefore, the question becomes whether the responsibility should lie with the designers, the users, or a hybrid of both, and how such ethical issues should be dealt with.

3.4.3 Ethical Challenge Recommendations

To address issues with therapeutic relationships and emotional reactions, the authors suggest that ethical codes and guidelines be updated to consider these relationships, their issues, and subsequent consequences.

The paper advises that the best course of action to handle the ethical issue of competency will be to train the users and adjust the design of IACP systems. Users must be aware of what they are signing up for when using IACP's and understand the drawbacks of these systems. Additionally, systems should be designed such that they are acting ethically towards humans, and trained to specifically handle cases that they are dealing with. The paper provides an example about IACP's only being accessible to those with depression if the machine is specifically designed to care for those people. Additionally, manufacturers should make sure to display the capabilities and qualifications of their machines very clearly so that the users can make informed decisions about whether to use the IACP's. Additionally, there should be an existing process for users to submit issues and have them resolved.

To deal with patient safety, the authors state that IACP's must be capable of monitoring and assessing the risk of patient harm and be able to apply decision making skills and judgment towards deciding ethical courses of action. Additionally, back-up measures should be instituted should an IACP fail in such a situation. Standard guidelines for clinical best practices should be followed by IACP's as well.

To address privacy and trust concerns, it is necessary for machines to be equipped with detailed information on its purpose, scope, and capabilities. Additionally, details about data collection and storage should be shared with IACP users, as well as whether the machines adhere to safety and privacy guidelines.

Deception and appearance issues can be ameliorated through decreasing how much a machine resembles a human, and how realistic the outward design is. Many of the deception and appearance issues caused by IACP's occur because the machines resemble a human and patients are more likely to treat the machine as such if this is the case. Although the ideal level of realism is yet to be determined, these ethical considerations are very important when determining the design of an IACP used in healthcare.

There are currently no specific measures in place to deal with the ethical implication of responsibility in IACP's. A recommendation provided by the authors is that the legal burden should lie with the designers of machines but extend to the user if the technology is used inappropriately. The authors also suggest that as autonomous machines are used more frequently and for more complex situations, the decision processes of these machines should be properly detailed if there are ever ethically murky situations, as well as legal restrictions on use.

3.4.4 Design and Testing Recommendations

This paper highlights two recommendations that involve testing and analysis that would aid with the moral and ethical dilemmas caused using IACP's in healthcare. These are called the Turing test and GenEth.

The Turing test measures whether a machine is of similar intelligence to a human and performs in a way that mimics the intelligent behavior of a human. To combat ethical issues, a

Moral Turing Test (MTT) is referenced in this paper. To deal with IACP's, the normal Turing Test would be insufficient - the test would need to have some aspect of ethical examination. In this version, both the machine and the human subject or doctor would be faced with an ethical dilemma, and their responses and actions to the dilemma would be recorded. The judge would then blindly compare the actions to see whether the machine is performing actions that are ethically responsible when compared to the human. The drawback to this test is that critics have said it is not enough to prove whether a machine is moral or not. Another version of the MTT, developed by the Andersons, compares the actions of a machine in an ethical dilemma to that of an ethicist when faced with the same dilemma, rather than a doctor or other human. In this version, if a machine's actions match the actions of the ethicist, the machine can be considered ethically responsible.

The previously mentioned Andersons also developed an analysis tool, GenEth, that analyzes and codifies ethical principles for any ethically harmful situation. Using inductive logic as the basis for its engineering, GenEth can not only respond to real situations, but can also make predictions and determinations about untested situations. Although GenEth has not yet entered the mental healthcare domain, its decision-making aid would be very beneficial to ethical dilemmas in the mental health space.

5 Final Recommendations

Through my review of existing literature across four subsections, I have identified the following issues to be the most prominent ethical concerns when utilizing machine learning algorithms in the healthcare industry: fairness, safety and privacy, transparency, and accountability. Below, I have created a table summarizing recommendations from the four

research papers regarding how to minimize consequences when faced with ethical dilemmas in these four areas.

Figure 5. Comprehensive overview of proposed recommendations from literature survey.

Fairness	• Increase personal and framework diversity in problem selection • Choose data that reflects all subsections of a population • Design machine learning solutions that are flexible enough to accommodate for outcome label bias • Create technology that allows for generalizability so that minority populations are not overlooked
Safety and Privacy	• Continuously monitor for data loss • View machine learning solutions as subjective rather than objective • Train users on machine best practices • Publicize details about machine safety and privacy guidelines
Transparency	• Clearly define algorithm goals • Create a standard reporting mechanism for models, as well as a standard set of performance metrics (to be adjusted for the various avenues of healthcare) • Communicate clearly with patients about the purpose and drawbacks of machine learning technology
Accountability	• Continue to audit technologies post-release in the case of patient harm • Educate the healthcare community about machine learning applications in the industry, to allow for seamless integration into healthcare systems

6 Benefits of Machine Learning in Healthcare

Although most of this thesis paper focuses on the issues associated with using machine learning in the healthcare industry, it is important to note that there are many benefits to this application. These are outlined best by the papers “Artificial Intelligence in Healthcare and Medicine: Promises, Ethical Challenges and Governance”, “Machine Learning in Healthcare” and “Open-Source Clinical Machine Learning Models: Critical Appraisal of Feasibility, Advantages, and Challenges”.

6.1 Artificial Intelligence in Healthcare and Medicine: Promises, Ethical Challenges and Governance

In “Artificial Intelligence in Healthcare and Medicine: Promises, Ethical Challenges and Governance” by Jian Guan, AI in medicine is organized into three categories: virtual, physical, and a combination of both [11].

Virtual AI, using machine learning, can detect patterns in data to predict data trends or facilitate decision-making under uncertain conditions. This technology has proven valuable in genetics, molecular medicine, and precision medicine, allowing for the discovery of diagnostic biomarkers and therapeutic targets. It can also optimize clinical trials and simplify process management while reducing costs. Physical AI, represented by robots, assists elderly patients or attending surgeons in aged care settings or complex surgeries using a minimally invasive approach. Brain-Computer Interfaces (BCIs) form a communication pathway between the central nervous system and an output, improving the quality of life for patients with neurological disorders such as spinal cord injuries, stroke, and amyotrophic lateral sclerosis. In 2018, the FDA approved the first AI-based diagnostic system, IDx-DR, which detects diabetic retinopathy and makes screening decisions independently from a clinician. AI has shown great potential in prediction, imaging, pathological diagnoses, and treatments. Researchers are also developing biohybrids by engineering miniaturized interfaces between living and artificial systems, learning from nature to develop innovative AI solutions. Communication and collaboration between AI and the neuroscience field have become commonplace in the era of digital healthcare.

6.2 Machine Learning in Healthcare

As mentioned in section 3.2, “Machine Learning in Healthcare” by Alison Callahan and Nigam H. Shah focuses on the use of electronic health records (EHRs) as a technological development in healthcare [7]. The “Opportunities for machine learning in healthcare” section of this paper discusses the potential advantages of using EHR data for predictive modeling. The authors explain that a great opportunity in this field is the adoption of machine learning models that can categorize patients into different risk groups. This has the potential to greatly impact healthcare value and bring clinical practice closer to precision medicine. Additionally, identifying high-risk and high-cost patients in time for targeted intervention will become increasingly necessary as healthcare providers take on more financial risk in treating their patients. This can be done with predictive modeling, which has already begun to successfully be implemented in medical practice, resulting in more efficient and better-quality care. These models have also been applied to hospital and practice management, streamlining operations, and improving patient outcomes. The paper states that there is ample opportunity for predictive modeling to enable proactive treatment, more efficient use of resources, and deliver better care at a lower cost, as seen in other industries that have incorporated machine learning into their workflows.

6.3 Open-Source Clinical Machine Learning Models: Critical Appraisal of Feasibility, Advantages, and Challenges

“Open-Source Clinical Machine Learning Models: Critical Appraisal of Feasibility, Advantages, and Challenges” is a paper by Keerthi B Harish, BA; W Nicholson Price, JD; Yindalon Aphinyanaphongs, MD, PhD [12]. The authors identify four main advantages to using open-source machine learning in healthcare. First, the transparency of open-source data and

software allows for the widespread assessment of a specific model's performance, which allows for many people to validate these models. Currently, many models that are used in healthcare require proprietary software, whereas the open-source tools would allow for a faster and more accurate ability to improve technological performance and capabilities.

Next, the use of open-source technology allows hospitals to customize models to fit their specific population, which is important because ML tools are developed using data sets from large cohorts and applied to individuals, resulting in varying levels of safety and efficacy across different populations. This issue is referred to as the "curly braces problem" in medical informatics, where model performance degrades when implemented in new settings. However, by using open-source models, facilities can adjust the model weighting to optimize performance for their unique patient populations. This customization can increase safety and efficacy, as different departments within a hospital may require slightly different model calibrations. The ability to personalize care through the adjustment of source code is a valuable benefit of open-source technology.

The third advantage to open-source options is that they include lower cost options which could allow for a quicker and easier integration process of machine learning in clinical practice. The paper explains that the concept of open-source models is characterized by the availability of their code. Transparency is a hallmark of these models, which have typically been less expensive than proprietary options in the past. However, despite significant investment from developers and investors, the healthcare sector has been slow to adopt machine learning (ML) technologies due in part to concerns by clinicians and administrators about potential financial and value-based repercussions. This hesitancy has hindered the implementation of ML tools, particularly among

hospitals that are new to this technology. To encourage experimentation, reducing the financial risks associated with adoption may be helpful.

The final potential benefit of (low-cost) open-source options is that they may increase competition, leading to changes in both pricing and functionality. While open-source models have become more popular, proprietary technology is still being developed. However, facilities may have to choose between proprietary and open-source options when using deep learning. The availability of open-source models may push proprietary developers to improve their offerings in order to justify their higher prices. This could lead to improvements in areas such as user interface, integration with existing IT systems, or implementation guidance and maintenance.

7 Conclusion

Incorporating machine learning techniques in the healthcare industry poses a wide array of potential benefits, but it is important to acknowledge that there are many challenges that arise with this technological advancement. This paper provides a background for the machine learning knowledge needed to understand the various algorithms being introduced in the healthcare industry, as well as four areas in which advancements are currently being made. To do so, I analyzed the ethical implications posed by four different papers, as well as the recommendations suggested by the authors. Through this, I was able to compile a comprehensive list of recommendation suggestions in four areas of ethical concern. I concluded this paper on a more promising note, with a brief analysis of the benefits offered by utilizing machine learning solutions in healthcare.

8 References

1. Tondak, A. (2022, December 06). Structured, semi structured and unstructured data. Retrieved April 21, 2023, from https://k21academy.com/microsoft-azure/dp-900/structured-data-vs-unstructured-data-vs-semi-structured-data/
2. K. Shailaja, B. Seetharamulu and M. A. Jabbar, "Machine Learning in Healthcare: A Review," 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, India, 2018, pp. 910-914, doi: 10.1109/ICECA.2018.8474918.
3. Ayodele, T. O. (2010). Types of machine learning algorithms. *New advances in machine learning*, *3*, 19-48.
4. (healthcare)+and+(machine+learning) - search results - pubmed. (n.d.). Retrieved April 21, 2023, from https://pubmed.ncbi.nlm.nih.gov/?term=%28healthcare%29%2BAND%2B%28machine%2Blearning%29
5. (COVID-19)+and+(machine+learning) - search results - pubmed. (n.d.). Retrieved April 21, 2023, from https://pubmed.ncbi.nlm.nih.gov/?term=%28healthcare%29%2BAND%2B%28machine%2Blearning%29
6. Chen, I. Y., Pierson, E., Rose, S., Joshi, S., Ferryman, K., & Ghassemi, M. (2021). Ethical machine learning in healthcare. *Annual review of biomedical data science*, *4*, 123-144.
7. Callahan, A., & Shah, N. H. (2017). Machine learning in healthcare. In *Key Advances in Clinical Informatics* (pp. 279-291). Academic Press.
8. McCradden, M. D., Joshi, S., Mazwi, M., & Anderson, J. A. (2020). Ethical limitations of algorithmic fairness solutions in health care machine learning. *The Lancet Digital Health*, *2*(5), e221-e223.
9. Luxton, D. D., Anderson, S. L., & Anderson, M. (2016). Ethical issues and artificial intelligence technologies in behavioral and mental health care. In *Artificial intelligence in behavioral and mental health care* (pp. 255-276). Academic Press.

10. Hall, D. (2021, July 07). The Eliza Effect. Retrieved April 21, 2023, from https://99percentinvisible.org/episode/the-eliza-effect/
11. Guan, J. (2019). Artificial intelligence in healthcare and medicine: promises, ethical challenges and governance. *Chinese Medical Sciences Journal*, *34*(2), 76-83.
12. Harish, K. B., Price, W. N., & Aphinyanaphongs, Y. (2022). Open-Source Clinical Machine Learning Models: Critical Appraisal of Feasibility, Advantages, and Challenges. *JMIR Formative Research*, *6*(4), e33970.