The Ethical Governance of Artificial Intelligence and Machine Learning in Healthcare

By

Tina Nguyen

The Ethical Governance of Artificial Intelligence and Machine Learning in Healthcare

By Tina Nguyen

This book first published 2023

Ethics International Press Ltd, UK

British Library Cataloguing in Publication Data

A catalogue record for this book is available from the British Library

Print Book ISBN: 978-1-80441-105-6

eBook ISBN: 978-1-80441-106-3

Table of Contents

Acknowledgements

Thank you to my family for always supporting me no matter what the endeavor is, I am forever indebted to you for the love I have and will continue to receive.

Thank you, Dr. Magill, for helping me throughout this process and encouraging me to publish this book.

Much appreciation to faculty and mentors at Duquesne University and UTMB for always lending a helpful ear and providing constructive feedback.

Lastly, thank you to my friends for the unconditional support and keeping me sane!

Abbreviations

ACP	Advanced Care Planning
AD	Advance Directives
AHA	American Health Association
AHIMA	American Health Information Management Association
AI	Artificial Intelligence
AMA	American Medical Association
ANN	Artificial Neural Network
BDAS	Big Data Algorithmic Systems
CCPA	California Consumer Protection Act
CDSS	Clinical Decision Support Systems
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
CSR	Corporate Social Responsibility
DG	Data Governance
DL	Deep Learning
DMBOK	Data Management Body of Knowledge
DNN	Deep Neural Network
EHR	Electronic Health Record
EU	European Union
FDA	Food and Drug Administration
FHIR	Fast Healthcare Interoperability Resources
FTC	Federal Trade Commission
GINA	Genetic Information Nondiscrimination Act
GDPR	General Data Protection Regulation
GWAS	Genome Wide Association Study
HDE	Humanitarian Device Exemption
HHS	Department of Health and Human Services
HIM	Health Information Management
HIPAA	Health Information Portability & Accountability Act
HITECH	Health Information Technology for Economic and Clinical Health
HUD	Humanitarian-Use Device
IDE	Investigational Device Exemption
IRB	Institutional Review Board

IT	Information Technology
JIP	Justice Involved Person
ML	Machine Learning
NB	Notified Body
NLP	Natural Language Processing
OSTP	Office of Science and Technology Policy
PGHD	Patient Generated Health Data
PHI	Protected Health Information
PIPEDA	Personal Information Protection and Electronic Documents Act
PM	Precision Medicine
PMA	Premarket Approval
PMDD	Posthumous Medical Data Donation
PMI	Precision Medicine Initiative
R&D	Research and Development
RCT	Randomized Controlled Trial
RWE	Real World Evidence
SaMD	Software as a Medical Device
UK	United Kingdom
U.S.	United States
VA	Veterans Affairs
VHA	Veterans Health Administration
VUS	Variants of Uncertain Significance
WHO	World Health Organization

This is not a comprehensive list, only frequently used abbreviations are listed

Introduction

The Internet Age has transformed how we interact with technology. With the emergence of 5G (also referred to as fifth-generation mobile network), the usage of Artificial Intelligence (AI) along with Machine Learning (ML) algorithms have only increased since their introduction to society. In fact, it is almost guaranteed that everyone has interacted with AI/ML at some point in their lives. ML algorithms help us choose what to watch on Netflix, while the AI voice-based assistant Siri does menial tasks for us as we scroll through our phones. Essentially, we rely on AI/ML a lot more than we assume. Now before you decide to run off to the woods to live a life away from technology, I want to strongly emphasize that using AI/ML is not necessarily a bad thing. However, this is also not to say that we should automatically default to blindly trusting AI/ML. Some common issues seen in the news about AI/ML are biased results, black box algorithms, and privacy of data. While these are certainly valid concerns, the potential benefits from AI/ML compensates for the risks. Specifically, in areas such as the healthcare industry where the discovery of a treatment for a life-threatening disease or the early detection of cancer can save countless lives.

So, what should be done to control the negative effects of AI/ML without getting rid of it completely? Many scholars and experts in the field are pushing for what is known as AI governance, which is basically managing AI by solely focusing on the safety of the technology. Meanwhile, ethical, legal, and social factors serve as mere elements of AI governance. While this approach guarantees safety, it does not ensure sustainability. The safety of AI/ML is one of the key components for producing and ensuring ethical AI/ML. However, safety should not be the only priority when it comes to ethical AI/ML. AI governance disregards other factors that affect the development and deployment of ethical AI/ML, whereas ethical governance does the exact opposite.

Ethical governance of AI/ML is centered around various stakeholders applying ethics into all aspects of AI/ML starting from the development stage all the way to the deployment stage. This type of governance is important because not only does it ensure that the AI/ML is safe to utilize but is also taking ethical principles into consideration through the whole AI/ML life cycle. Ethical governance of AI/ML also highlights the stakeholders involved, which is not heavily emphasized in AI governance. The people interacting with AI/ML, such as the developers/engineers, regulatory officials, governments, and yes patients, all have a significant role in how ethical AI/ML is developed, improved, and implemented. Another contrasting factor that sets AI governance and ethical governance apart is the discourse surrounding the two—with more attention towards AI governance. Therefore, ethical governance of AI/ML must be advocated for to bring more awareness to this approach.

As such, this book explores the potential of ethical governance of AI/ML in healthcare. Big data and its role in healthcare is discussed in the earlier half of the book with Chapters 1 and 2. Chapter 1 explores big data analytics, which ties into the theme of Chapter 2 on why data governance is needed. Meanwhile, the subsequent chapters explore pivotal issues regarding current and future AI/ML in different areas of healthcare. Chapter 3 discusses the ethical governance of AI/ML in the electronic health record, whereas Chapter 4 addresses ethical governance of AI/ML through precision medicine. Chapter 5 explores proposed AI/ML for decision-making with incapacitated patients. Finally, Chapter 6 examines the regulation of novel technologies and how ethical governance can improve current processes.

The book should preferably be read in a chronological order as the preceding chapters build into the proceeding chapters. However, each chapter can still be read independently from the rest of the book. Each chapter is structured to discuss the ethical issues in different areas of healthcare, followed by recommendations catered to those topics of interest. While the focus on ethical governance of AI/ML is in the U.S.'s perspective, other countries are also examined in the book. The U.S. government recently presented five principles for the development of ethical AI in their "Blueprint for an AI Bill of Rights". I also propose a set of principles, through the recommendations section in each chapter, that coincides with but ultimately surpasses those mentioned in the released blueprint.

This book is written in an accessible manner with clear explanations on complicated concepts for a general audience, such as the intricacies of AI/ML and bioethical principles. The intention of this book is to reach various audiences as AI/ML are not only affecting the healthcare industry but the daily lives of society as well. As such, this book also views ethical governance through a normative ethics approach where the recommendations are not necessarily enforceable, but instead meant to invoke action from the stakeholders addressed on what they ought to do. My hope is that by reading this book, greater awareness is brought to these technologies (existing and proposed) and the inadequacies of the current governance model. Additionally, there are discussion questions to generate robust conversations and further readings are included to supplement readers' understanding of the topics. Discussing these issues will ultimately provide a foundational groundwork for more future research of the ethical governance of AI/ML.

Chapter 1
Ethics of Big Data in Healthcare

Introduction

From suggestions on mobile apps to the forecasting of trends, technology has greatly transcended expectations of what could be accomplished. These innovations have come to fruition because of the data utilized during their infancy stages. While this idea of data having a significant role in creating innovative technology seems abstract to some, data are the one commodity that developers covet the most. In fact, the wearable technology company Fitbit was revealed to have accumulated 150 billion hours of their users' activity in 2018, which included demographic information and sleep pattern data.[1] This data could be utilized to better understand factors that might affect patients' health as well as uncovering any unusual patterns and making connections that were not obvious before. Data allows for these technological advances and deep analyses to happen. Additionally, beneficial machine learning algorithms and deep learning algorithms cannot properly train or provide accurate results without utilizing big data. Therefore, the collection and future application of big data are crucial not only for researchers but more importantly for patients. The potential of these large datasets can be highly beneficial to both parties. However, big data, like the healthcare technology it helps to create, are still susceptible to ethical issues. Some ethical concerns that arise with big data usage are autonomy, transparency, justice, and accountability. These ethical consequences could be mitigated by developing an oversight model and implementing a data governance program. Evidently, big data are needed for the creation and maintenance of innovations. Nevertheless, this does not serve as sufficient justification for the underlying ethical implications that still exist and must be remediated.

Big Data in Healthcare

Big data have a substantial role in healthcare, especially with how interconnected data have become. The introduction of wearable technology and social media platforms have allowed for even more data to be collected in an almost instantaneous manner. According to a 2014 joint report by EMC Corporation and IDC, healthcare data makes up 48% of the digital data per year and its size is expected to increase to 2,314 exabytes by 2020.[2] Meanwhile, by 2030 it was projected that genomic data collection will increase to 25 petabytes on an annual basis.[3] As such, some have argued that not utilizing this data, which is filled with significant information, is a missed opportunity. While others believe that using big data, without obtaining proper consent, will cause irreparable damage to affected parties if the data are applied in a harmful manner. In the proceeding subsections, various concepts about big data will be examined.

A Brief Overview of Big Data

As aforementioned, big data affects all aspects of the public populations' life, from personalized ads to surveillance technology. While these uses of big data have been increasing recently, the magnitude of its implementation in healthcare has rapidly grown throughout the years. Specifically, the Health Information Technology for Economic and Clinical Health (HITECH) Act in 2009, which incentivized hospitals that utilized electronic health records (EHR) instead of paper charts, is often attributed to the mass generation of big data in healthcare.[4] Some examples of healthcare data include, but are not limited to, data from wearables, emails, medical images, video data, and audio data.[5] From this brief list of examples, it is evident that healthcare data comes from several sources and in a multitude of forms. While various definitions for big data exist, they are often either too broad or do not emphasize why big data analytics are needed. Therefore, an ideal definition that addresses the issues of previous definitions is, big data are large datasets that must be analyzed by utilizing advanced processes given that human users cannot realistically be expected to sift through vast amounts of complex data in an efficient manner.[6]

Additionally, several characteristics of big data have been identified as different V's. The core V's comprise of volume, variety, and velocity—with the more recently added veracity rounding off the V's.[7] Although there are additional V's to big data that have been suggested by other experts in the field, such as validity, visualization, value, vulnerability, and viability, the four identified here strongly encompasses the main objectives of big data.[8] As such, only these four V's will be discussed in this book. Volume describes the amount of data that is generated and collected, whereas velocity is the time accumulated to process and analyze big data. Meanwhile, variety deals with "[h]eterogeneity in healthcare data [that] stems from the existence of a large number of formal standards and nonformal standards."[9] Furthermore, veracity, one of the various V's that is typically grouped with the primary three V's, has generated some debates on its definition. Perhaps, due to veracity being a newly identified V descriptor, there are discrepancies on what it entails and its overall role in classifying big data. Given the definition of the term veracity in the field of ethics often equating to "truth telling", it is reasonable to infer that veracity can be described as the caliber of the data. Typically, a person would not associate a morally questionable individual with the attribute of veracity. As such, the inverse of this is also true, where a morally righteous person is expected to practice veracity. Hence, in this book, the concept of veracity in big data will be correlated with the quality of the data. Recognizing veracity as one of the V's displays the significance of how data, despite the intentions of its future use, must be of the highest quality. This characteristic of veracity, among the core V's, is critical for big data analytics. These four V's, in addition to the other V's, showcases the different factors that are utilized to depict big data. While there are contrasting approaches to classifying big data and the definition of veracity, the overall message is that big data needs to be analyzed to provide actionable results.

Big Data Analytics

As crucial as big data are, if this accumulated data is not further evaluated it becomes worthless. Therefore, big data must be studied and analyzed in order to provide beneficial information that can be applied for innovations.

Khanra et al. define big data analytics as, "[…] the analysis of detailed, dynamic, low-cost, massive, and varied data sets to deliver sophisticated solutions."[10] However, in Galetsi and Katsaliaki's literature review on big data analytics in healthcare, they found that 70% of the literature mostly concentrated on clinical data derived from EHRs.[11] While EHR data are the foundation of big data in healthcare, this does not minimize the importance of other data sources' impact on big data analytics. Rather, Pramanik et al. suggest that "[d]ata analytics [,] instead of trying to combine all the available data, target the right data and thus gives the physician the rightful insights."[12] There are several big data analytics tools available, such as Apache Hadoop, Apache Hive, HBase, Apache Zookeeper, Apache Arvo, and Apache Lucene, which we will not delve deeper into as they are beyond the scope of this book.[13] In addition to these big data analytics tools, artificial intelligence (AI) and, more specifically, machine learning (ML) are utilized to analyze big data effectively and efficiently. Machine learning is often described as being "data hungry" due to its need of large datasets for training, which allows for its outputs to improve over time. Furthermore, given the multitude of factors affecting big data analytics, it is not surprising that varying elements are also involved in the process. Given the several processes and interdisciplinary nature of big data, such as the collection and analytics aspects, there is always the participation of multiple parties. Some individuals or groups involved in big data analytics are data scientists, healthcare professionals, and IT personnel. Whether it is interpreting the data itself or making decisions based on the results of big data analytics, each of these groups have specific duties in the big data analytics process.

The possibilities of what can be done with big data analytics are endless, especially in the context of healthcare. According to Dolezel and McLeod, "A McKinsey Global Institute study estimated that effective use of big data analytics could decrease national healthcare expenditures by approximately 8 percent annually."[14] Similarly, big data analytics have been "expected to significantly improve healthcare benefits and reduce costs."[15] With the use of big data analytics, there are no longer unnecessary testing or procedures being done, which will undoubtedly save money for both the healthcare industry and patients. Particularly, the potential for big data analytics is limitless. There are different methods to big data analytics, which include descriptive analytics, diagnostic analytics, predictive analytics, and prescriptive analytics.[16] Each of these approaches to big data analytics have specific functionalities and as such are better suited for certain scenarios. For example, descriptive analytics, "[…] the condensation of big data into smaller meaningful information", would thrive in cases where finding the root cause to a problem are concerned, whereas predictive analytics, which predicts the probability of future events, is more appropriate for the identification and prevention of a disease.[17] By undergoing the descriptive analytics approach, healthcare professionals can answer questions regarding the reasons behind their patients' health or why certain drugs work better for specific patients.[18] Meanwhile, predictive analytics could be used for early detection of epidemics.[19] Before its demise in 2015, Google's Google Flu Trends, which estimated the cases of influenza across the world and shared this data publicly, applied the predictive analytics method in the big data analytics process.[20] Diagnostic analytics takes a step further than descriptive analytics by asking the question of "why did it happen?".[21] Therefore, diagnostic analytics are applied to instances dealing with finding "causes of various disease outbreaks, [or] administrative and patient treatment malpractices."[22] Similarly, prescriptive analytics also expands the method of predictive analytics. Prescriptive analytics is described as "finding the best course of action for the predicted scenario derived from the predictive analysis."[23] Thus, prescriptive analytics is better suited for cases when recommendations are needed based on a predicted outcome derived from predictive analytics. These different approaches to big data analytics all come with their own unique advantages, which ultimately benefits the healthcare industry. While each of these methods have their own merits, the future of big data analytics is much more likely to focus on predictive analytics. According to Kaur

> Big data predictive analytics can lead to not only low attrition rates in research and development but can also lead to development of advanced tools and algorithms to improve clinical trial design and selection of patients to provide customised, personalised treatments[24]

This usage of predictive analytics heavily correlates to healthcare systems moving towards the practice of precision medicine (PM), further discussed in Chapter 4, which seeks to provide prevention and treatment strategies for patients given various contributing factors, such as genetics, environment, and lifestyle. As interest in the precision medicine approach rises, so too will the reliance on predictive big data analytics. Overall, there are several functionalities for big data analytics as well as methods of analytics that are valuable to the healthcare industry.

Challenges of Big Data in Healthcare

Supporters of big data usage believe that this large quantity of data is the answer to all the current issues in healthcare. Although big data can be utilized for many purposes, some existing challenges hinder big data from being faultless. According to Chen and Quan-Hasse, "Bigger does not always mean better, accessible does not always mean ethical, and convenience does not always mean efficient."[25] Again, the mass implementation of EHRs have become a significant source of big data. Some examples of different types of data present in EHRs include lab data, billing data, medical imaging data, medication records, and clinical notes.[26] These contrasting types of data are accumulated to make the big data present in EHRs. Specifically, these data can be classified into three contrasting data types, known as unstructured, structured, and semi-structured data. It was reported that 80% of the data in the EHR are usually categorized as unstructured, whereas structured data completes the rest of this statistic amounting to 20%.[27] While both structured and semi-structured data, which at least includes already arranged data, are more manageable, unstructured data are infamous for complicating the data analytics process.[28] Unstructured data consists of data that does not have precise formatting, such as physician's notes or medical imaging. The variability of these data makes it difficult for even advanced systems to sort through.[29] Thus, unstructured data, which comprises most of the data found in the EHR, becomes a barrier for performing effective big data analytics. According to Morr and Hassan, "Data can be biased, incomplete, or filled with noise; indeed, healthcare data scientists and analysts spend more than 60% of their time cleaning the data."[30] Effective and efficient outcomes cannot be easily achieved from unstructured data, heavily affecting the performance of big data analytics. Having data scientists/analysts sift through big data defeats the whole purpose of big data analytics tools and methods, which are meant to ease the process of clinical decision-making by having readily available data.

An additional obstacle to the efficient usage of big data in healthcare is interoperability. The main purpose of interoperability is to allow data to seamlessly flow between different organizations. This is a highly valuable functionality given that numerous populations of patients often visit multiple hospitals. Interoperability also has rather crucial implications for global healthcare. In the context of global healthcare, interoperability can assist healthcare professionals in providing effective and efficient care to patients in underdeveloped areas.[31] Therefore, the lack of interoperability directly affects clinicians' workflows as well as patients' care since their information could flow into the system as either missing or incorrect. While some might argue that these institutions could try to bypass these discrepancies by directly sending the patient's information over, this suggestion not only defeats the purpose of interoperability (the instantaneous sending of this pertinent information) but temporarily puts the patient's protected health information (PHI) in a defenseless state. The reason for the lack of interoperability is often attributed to insufficient standards.[32] According to Satti et al.,

> Healthcare organizations tend to move towards standards that are easy to use and cost effective. While, this is usually not a problem when medical components have to be made interoperable within the organization boundary, interoperability between different, often competing, healthcare organizations is a major challenge[33]

Similarly, lacking standardized interoperability becomes an increasingly pressing matter because it also affects big data analytic capabilities. As previously mentioned, effective big data analytics is dependent on the four V's. The four V's are not present when there is lacked interoperability, which results in big data analytics becoming compromised. Due to standardization not being enforced, these healthcare organizations will not feel the need to work with other organizations and instead continue to create barriers for true interoperability to happen. As such, without proper standardization, interoperability cannot realistically be achieved.

Moreover, despite the various benefits of having interoperability, the push for this to happen seems to be one sided. Rehman et al. state that "[t]he US Health Department is aiming for interoperability between disparate EHRs by 2024."[34] As 2024 is approaching, these other problems with interoperability remain unresolved.

Meanwhile, other types of data not originating from EHRs, such as patient generated data and genomic data, also have their own issues—namely, privacy concerns.[35] These types of data are needed to allow for more detailed big data analytics to occur. Although EHRs are a land mine filled with personal information that can be compromised if appropriate security and privacy measures are not followed, there are at least some levels of protection in comparison to other sources of data. Patient generated data typically comes from wearables with smart watches being a popular example. According to Garattini et al., "Wearable devices can transmit vast amounts of data sometimes in constant live streaming modes which raises important ethical issues regarding privacy and security."[36] Unlike the data in EHRs, wearable data is not protected under the HIPAA Privacy Rule because manufacturers of wearables are not classified as one of the covered entities, which include health plans, healthcare clearing houses, and healthcare professionals.[37] Even though wearable data are patient generated data, the manufacturers are the ones who own this data, not the patient. This is concerning as wearable devices collect highly sensitive and real-time information, such as the patient's location.[38] Not only can this data be utilized without the consent of patients, given that wearable data does not explicitly belong to them, but also has the potential to track their whereabouts. Additionally, genomic data, which can be stored in EHRs but technically does not originate from them, are another cause of concern regarding privacy. According to Dash et al., it is "estimate[d] [that] the number of human genomes sequenced by 2025 could be between 100 million to 2 billion."[39] This statistic clearly shows that genomic data is one of the more notable contributors to big data in terms of volume. While having more genomic data is helpful for discovering and potentially preventing specific diseases, this data contains substantially sensitive information that not only directly affects the patient themselves but also any related individuals. Thus, if a genomic data leak were to happen for a single patient, then the amount of people affected is expected to be much more than with other types of data breaches (e.g.) a single EHR being compromised). Likewise, given the intimate nature of genomic data, as this type of information directly originates from the patient's body, it is expected that there are proper security measures ensuring the privacy of these data. However, large datasets, particularly those including genomic data, are more susceptible to being misused and thus negatively impacting "consumer trust".[40] There are currently no concrete laws protecting against genomic data breaches in the U.S., rather only laws that protect against discrimination because of one's genetic information. Even then, the Genetic Information Nondiscrimination Act (GINA) also has its own faults and does not cover as much as expected.[41] The lack of privacy protections for patient generated data and genomic data further complicates the big data analytics process since patients might become wary and decide they no longer want to produce or provide these data. From these various examples, it is evident that large amounts of data, as required by the continuous application of big data analytics, does not always correlate to effective care. These challenges in healthcare regarding big data act only as an assortment of obstacles for big data analytics to reach its fullest potential.

Ethical Concerns of Big Data

In addition to the current challenges of big data in healthcare, ethical issues also threaten the greater utilization of big data. Some ethical concerns with big data usage include autonomy, transparency, justice, and accountability. While there might be other ethical concerns that are not addressed in this section with big data usage, these four seemingly cover the main aspects most stakeholders are worried about. These issues must be addressed and corrected for big data analytics to be considered as ethical.

Autonomy

The ethical principle of autonomy allows patients to self-determine what happens to their body. As a condition of their autonomy, patients can choose to either agree with the proposed treatment or forgo treatment if they find it to be too burdensome. Fundamentally, autonomy is the patient's right to decide for themselves what they want in terms of their own health. A highly related concept to autonomy is the process of informed consent. Elements for the standard of informed consent involve the patient understanding the risks and benefits to treatment/ nontreatment, deciding what course of action to take without any coercion, and having the ability to communicate their decision. While the informed consent process for treatment/nontreatment is rather straightforward as patients are often given adequate information about the procedure and their prognosis, other aspects of the patients' care are not as clear cut. For instance, when patients are admitted into hospitals, they technically provide consent to the hospital to use their data. This is a violation of patients' autonomy as some patients might not have explicitly consented to their data being collected or utilized in ways they did not agree to. Similarly, Lacroix argues that "[t]he issue of meaningful informed consent [...] arises because big data analytics involves data that may be a continuous collection over time and the intended consequences are not known or fully understood at the time of collection."[42] As such, patients' autonomy are violated once again because what they could have consented to previously does not match with what is currently being done with their data. Besides the collection of data being an issue, data sharing without the patient's knowledge also breaches the standard of informed consent and to an even further extent autonomy. According to McCoy et al., "[...] even if patients are able to authorize sharing of their data, they are rarely given the information and opportunity to ask questions needed to give meaningful informed consent to future uses of their data."[43] Patients are often not aware of what is implied when they receive care at a hospital, especially regarding what happens to their data. The same situation pertains to the data that comes from wearables. Despite the intentions of wearables to promote autonomy in the sense of patients taking control of their health, wearables violate patients' autonomy since they discretely bypass the consent process to collect the users' data.[44] Additionally, some patients do not want to be compared and believe that their conditions are vastly different than others. As mentioned earlier, the application of predictive big data analytics will increase with the advent of precision medicine. Therefore, in the context of precision medicine, which aims to find therapeutic strategies by grouping patients together by shared characteristics, using big data analytics could be seen as an intrusion on the patient's autonomy. With big data, patients no longer have control over their decisions; instead, they inadvertently lose their autonomous rights to other stakeholders. As Lacroix states, "Control requires awareness of the use of personal data and real freedom of choice."[45] Control, and thus the patients' autonomy, cannot be fully granted if patients are deceived by healthcare organizations and other entities.

Transparency

As previously mentioned, proper informed consent to use their data is not explicitly obtained from patients regarding what happens with their EHR data. Not being transparent about patients' data could invoke the feeling that their care is not as safe as it should be. A rather recent example of this is Ascension and Google's "Project Nightingale" in 2019, which was meant to "[…] store patient information in Google's cloud, then apply machine-learning technology to deliver recommendations or predictions to clinicians and administrators at Ascension."[46] While this project had good intentions, patients were unaware not only of this partnership but what was being done with their data. However, Wachter and Cassel explain that "[t]his form of data sharing is entirely legal" because "virtually all hospitals and clinics enter into what are commonly known as business associate agreements to share certain data with outside vendors who agree to keep data secure as they analyze it for various purposes […]."[47] Although these organizations are technically safe under these laws, the trust between patients and their respective healthcare organizations is heavily tested regarding the collection and usage of big data, especially given the absence of informed consent. Similarly, a lack of transparency, in any case and not just with big data, could cause patients to mistrust their physicians. If patients discover their physicians were cognizant of big data analytics on a much wider scale is occurring and decided not to disclose this information with them, then some patients might feel deceived. This is damaging not only for that specific patient-physician relationship but also for subsequent encounters the patient may have. The same logic also applies for healthcare organizations, where patients could potentially avoid seeking healthcare from any organization. Therefore, transparency of big data, regarding the collection and subsequent usage of it especially in terms of major projects, should be disclosed to patients. A lack of transparency is also related to the ethical principle of non-maleficence, where it is the duty of the healthcare professional to not cause harm to the patient. While physical pain is not as likely to occur with big data analytics, the emotional distress some patients could have towards the gathering and usage of big data is just as harmful as malpractice. Once again, this correlates to the feelings of deception and the resulting loss of trust that patients might be experiencing with the absence of transparency. Regaining the patient's trust after such a distressing event could potentially be a difficult task if not approached in a delicate manner. According to Ballantyne and Stewart, "Transparency can help justify public confidence in institutions, [...] ensure accountability and facilitate public debate."[48] As such, it is expected that these institutions would strive to be transparent about big data and big data analytics. However, Lacroix states that "[a]ccountability has been championed over transparency, which to date is known to have many limitations in protecting an individual's right to privacy."[49] While accountability will be discussed further in depth in a proceeding section, this debate on which of the two ethical challenges should be weighted more is counterproductive to resolving the current issues with big data. Not being transparent affects the patients' view of the healthcare institution itself and gradually the healthcare industry as a whole. This not only negatively affects the healthcare organization's reputation but their business as well since patients are less likely to seek care from what they view as a morally questionable organization. Furthermore, transparency promotes autonomy as patients can inquire institutions' policies/responses regarding big data and decide where they want to have their care. Likewise, having transparency ensures that patients are at least generally aware of the collection, usage, and security of their data. Therefore, insinuating that transparency constrains autonomy and more specifically the patient's right to privacy is heavily misguided. Being transparent does not mean that the patient's data will be more easily susceptible to privacy concerns, but instead involves the patient in understanding the process of big data analytics. In fact, deficient transparency causes additional problems and as such should not be swept aside to deal with later or be completely disregarded.

Justice

In addition to autonomy and transparency, big data also violates the ethical principle of justice. This ethical principle deals with equity and equality in care, whether through access to healthcare services or insurance coverage. In reference to Project Nightingale, McCoy et al. suggest that "[p]atients lack control […] because they may have no option other than to seek care in a health system that plans to share their data."[50] This leaves patients in a vulnerable situation as they might not have the means or ability to go to another healthcare organization. Similarly, there is an imbalance in the patient and healthcare organization relationship as patients are prone to being exploited since they have more to lose in their relationship with healthcare organizations.[51] Specifically, healthcare organizations have financial gain over patients' data as big data analytics has been known for saving costs. Meanwhile, patients are not only unaware of their data being collected and used but how there are potential incentives from their data. Institutions also tend to not disclose where and what is being done with the savings from big data analytics. Additionally, low-middle income countries do not have the same resources that first world countries, such as the U.S. and the United Kingdom (UK), typically have. As aforementioned, a source of big data is patient generated data which usually comes from EHRs and wearable technology. Some patients from lower socioeconomic situations, even those in the U.S. and UK, are more likely to not prioritize buying wearables or even having regular check-ups. Therefore, this disadvantage leaves some groups of patients to contribute even less sources for big data, which leads to inconsistencies and potentially biased outcomes. Patients with the most contributions will benefit from big data analytics, whereas powerless patients will likely have these findings enforced on them. Particularly, implicit bias can also occur with big data and further negatively affect already vulnerable groups. For example, using collected big data from "justice-involved persons (JIP)" for public health purposes.[52] Formerly incarcerated individuals are typically classified as a vulnerable population as they are often racial minorities and are from/or end up in a low-socioeconomic status. Similarly, these are individuals who generally already lack access to healthcare services. Rosen et al. argue that "[c]reating public health systems that rely on this tracking [of JIP] may further perpetuate the disproportionate surveillance of persons of color as well as that of low-income persons, who may be least able to curate their digital footprint".[53] This relates back to the argument that vulnerable groups are unable to contribute enough data for big data analytics and as a result are not actually gaining any of the benefits. Furthermore, research projects are also subject to providing unequal outcomes for their participants. Notably, if the recruitment of participants does not take the cultural values and nuances of minority patients into consideration, then prospective participants are excluded from partaking in research that might be beneficial for them or similar individuals. Bakken and Koleck list some examples of these nuances:

> (a) the severity of illness and sociodemographic composition of patients represented in EHR data vary by type and location of the healthcare organization, (b) Latinos are less likely than Whites or Blacks to use an app for health tracking, (c) and racial and ethnic minorities are less likely to participate in biobanks[54]

While Bakken and Koleck provide rather specific scenarios, these examples show that unless the research team was already mindful of accounting for these nuances in various groups, then their research project is subjected to unintentionally discriminating against participants. Not only will these participants not be able to fully contribute to the research study, but they also will not receive relevant benefits. Big data analytics are supposed to help control health disparities as several connections between the data are discovered. However, it seems that big data are doing the complete opposite and instead is reinforcing inequities in healthcare.

Accountability

Along with the other ethical issues previously discussed, the action of appropriately attributing accountability to different groups regarding concerns about big data is an ongoing one. Currently, in the U.S. there are relaxed restrictions on big data as well as government oversight being rather lenient in comparison to other countries.[55] While this was meant to promote the usage of big data for research and development (R&D) of innovations, these less than restrictive guidelines can actually lead to the misuse of big data.[56] Similarly, the lack of regulations on the data implies that other countries have access to U.S. citizens' information, which then raises concerns about security and privacy. Thus, the issue now becomes who should be responsible for protecting big data? Patients are left susceptible to having their genetic information and PHI, amongst other types of data, exposed to parties they are unaware of. Despite the intention of the Open Science Initiative to propel R&D efforts, it is still the duty of the U.S. government to protect its citizens from the potential harm of having their data misused.[57] Besides the international implications of big data usage, statewide data sharing is another concern. While a push for better interoperability has been advocated for due to its benefits of allowing for continuous health data exchange, this could also lead to an exacerbation of security and privacy issues. According to Watson and Payne,

> By reducing barriers to sharing data, EHRs increase the likelihood that personal information provided to or created by health-care providers can be combined with other clinical or publicly available data, even that created by the data subject, and used to invade patient privacy[58]

Therefore, proponents of interoperability are not thinking about the repercussions that instantaneous data sharing can have on patients. If the support for prompt data sharing, without considerations for data privacy and security, comes to fruition, then patients are essentially used as a means for healthcare organizations and professionals to achieve outcomes. Although enhanced interoperability would undoubtedly improve care, advocating for better data sharing without adequate research on the implications will harm patients in other ways. It is expected that there will be no consequences on those who advocated for advanced but unsafe data sharing because they can hide behind the argument that their intentions were to help patients.

In addition to data sharing, the responsibility of data breaches is often left unaccounted for. According to Mabee, "[…] it is even more concerning that experts believe and surveys reveal that the actual number of breaches across all sectors is underreported, even if the entity is legally required to report the incidents."[59] This not only shows a lack of transparency from companies/organizations but also accountability. Healthcare organizations are willing to deceive patients by not disclosing data breaches that have occurred if it results in not having to take responsibility for said breaches. From Ponemon Institute's report, it was discovered that while 51% of data breaches were from hackers, "[…] 25% were due to a system glitch and 24% occurred due to human error."[60] While system glitches are inevitable, the fault for data breaches (hackers and human errors) are potentially attributed to those involved in protecting patients' data. Healthcare organizations are supposed to ensure appropriate measures are in place to protect their patients' information. Meanwhile, it could also be argued that healthcare professionals are careless during their daily workflow, especially when it comes to administrative duties. For example, hackers often target healthcare professionals' emails and as such if a clinician absentmindedly clicks onto a suspicious link, then the whole organization can become compromised. Another potential source of human error are IT professionals who either did not test the newly installed functionalities correctly, which makes the system vulnerable, or were careless on the backend, such as clicking on phishing emails or sending PHI through unsecured emails for troubleshooting. Therefore, all three parties have rather equal responsibility for data breaches. Additionally, while data breaches in terms of EHR data is already bad enough, the unauthorized disclosure of genomic data is even worse. Genomic data does not only affect the individual themselves, but any related individuals as well. As mentioned earlier, the lack of transparency is closely related to a violation of non-maleficence since patients are harmed due to mistrust. If a data breach of genetic information were to happen and they were not immediately informed of this, then it is expected that patients will feel highly distressed. Delaying and withholding the parties that are accountable for the data breach of genetic information does nothing to reassure the patients that the healthcare industry has their best interests in mind. Juzwishin states that "[…] organizations have only recently come to publicly apologize for negative consequences of a patients' interaction with the health care system"[61] Receiving an apology is one of the first of many steps to corrective action in cases of data breaches. However, it has also been suggested to "[…] adopt the same approach as the airlines have no blame and open expression of airline transport accidents."[62] Although this approach will certainly alleviate anxieties of the responsible stakeholders, the effects on patients and their families are potentially irreparable. Clearly, no group wants or is willing to take full accountability for the several concerns regarding big data. It is also unfair to hold just one group accountable for the misuse of big data or when a data breach occurs. Nevertheless, patients who have been harmed by the misapplication of big data deserve to know that those involved are held responsible and disciplinary action is taken. They must also be assured that the appropriate safeguards regarding big data are in place at their organization. Therefore, there needs to be a proper process for accountability in order to regain the trust of patients as well as to provide ethical care.

Recommendations

The various ethical challenges, in addition to other barriers, discussed have clearly displayed the negative side of big data. It is important for those involved to acknowledge and subsequently address how these ethical issues will be mitigated. Juzwishin argues that "[b]ig data cannot be successful in addressing the sharing of data unless legislation, regulations and policies are revised to encourage integration without compromising the security, privacy and confidentiality of health data and information."[63] However, legislative bodies should not practice ethics washing by making "unsubstantiated or misleading claims about, or implementing superficial measures in favour of, the ethical values and benefits of digital processes, products, services, or other solutions in order to appear more digitally ethical than one is."[64] Therefore, healthcare organizations and legislative bodies need to cooperate and understand the current ethical issues regarding big data as well as how to resolve them. This can be achieved by developing an oversight model and creating a data governance program.

Develop Oversight Models

According to Woolley,

> Much policy literature is premised upon the assumption that trustworthy policy and governance mechanisms do or will exist for [big data analytics], but comparatively little guidance is offered as to

> what those mechanisms are, what normative rationale underwrites them, or how policy should establish or support them[65]

More will be discussed on data governance in the proceeding section, but the policy literature clearly shows that there is insufficient oversight on big data analytics. This is a significant issue because without oversight models, there is no one in charge of regulating and keeping close surveillance over big data. By having oversight models, healthcare organizations and research institutions are ensuring that they are utilizing big data in an appropriate manner. Due to the distinctness and the amount of healthcare organizations as well as institutions, there is not a single board that can realistically manage each of their big data processes. Thus, various oversight models with different duties and members will be explored in this section. Cohen and Mello propose creating an "[…] institutionally based, broadly representative data access committees that include patients specifically trained for their oversight roles."[66] This is an interesting proposition as this oversight group includes patients, the group that is directly affected by big data analytics. Involving patients into the oversight of big data will enhance the goal of achieving ethical use of their data since the group is able to advocate for their peers and themselves. Wachter and Cassel also suggest a similar method to Cohen and Mello's approach. Wachter and Cassel believe that developing a "visible patient-community advisory board" will promote transparency from healthcare organizations.[67] Although this model was not expanded on, working off of Wachter and Cassel's idea, this patient-community advisory board should be a part of meetings regarding big data usage/plans of mergers and projects with big tech companies. Including the patient-community advisory board will not only ensure patients that there is a group without any financial gain being involved but also allow these individuals to bring up any concerns that patients might have to the healthcare organization. These individuals will most likely be volunteers and ideally, they should have a basic knowledge of big data analytics. However, individuals who are not well versed in big data should not be immediately excluded as they can provide another perspective for the advisory board. Relatedly, McCoy et al. recommend a community benefit model, where "[…] financial benefits of data sharing [are] with communities rather than individuals […]."[68] While the community benefit model is theoretically not an oversight model, it promotes transparency, justice, and accountability—which are crucial to the oversight process. Perhaps having this community benefit model will motivate organizations to better manage their big data processes as well as being transparent with their patients. Meanwhile, Paxton argues that:

> [a]s academic researchers collaborate more with companies on interesting and complex datasets, the ethics boards governing academic research must begin to raise real questions about the risks posed even by data collected by business as a matter of course or through internal experimental work[69]

Evidently, there are no concrete guidelines on the ethical usage of big data for research studies. This lack of oversight on research involving big data, particularly the issue of consent, will create potentially dangerous and unethical outcomes for research participants. According to Paxton,

Researchers should have a duty—not only to specific participants in a study but also to the scientific community and the public—to conduct ethical research: Using ethically obtained data (no matter who obtained it) must remain a pillar of that ethical obligation[70]

To ensure that researchers are thinking in these terms, the Institutional Review Board (IRB) should redefine “minimal risk” and “private” data in order to promote ethical use of big data in research.[71] As mentioned previously, de-identified data can be re-identified and as such open-source data (meant for research and is de-identified) can potentially put participants in a risky situation. This relates back to the U.S. having relaxed regulations on who can access open-source data. While this sharing of big data is meant to progress research and development efforts, there needs to be better oversight from not only the IRB but the U.S. government as well. Therefore, redefining “minimal risk” and “private” data is a solid starting point. Moving forward, the U.S. government must take an active role in overseeing the activities concerning big data and ensuring its citizens that they are protected. Overall, oversight models, whether newly developed or an already existing body, need to create or modify current regulations for big data to be appropriately used in healthcare.

Create a Data Governance Program

Furthermore, a data governance program must either be established, if an organization previously did not have one, or reworked, for organizations that already have an existing program. Here, data governance programs are briefly discussed, but for a more detailed exploration of these programs along with their ethical issues refer to Chapter 2. According to Ladley, data governance is "[…] the organization and implementation of policies, procedures, structure, roles, and responsibilities which outline and enforce rules of engagement, decision rights, and accountabilities for the effective management of information assets."[72] In this definition, Ladley not only describes the main objectives of data governance but also makes the clear distinction of data being an asset. As such, data governance, if properly executed, can provide various benefits for the organization. Most significantly, data governance can save organizations from having to deal with costs relating to data breaches by having preventive measures in place that effectively avoids data security and privacy issues.[73] Despite the numerous advantages of having a data governance program, many organizations do not have one implemented or their current program is rather insufficient.[74] Therefore, the push for data governance must be established by the organizations' leaders.

Before completely advocating for data governance, some challenges that current data governance programs face now should be discussed first. As the data governance process involves the whole organization and not just the higher executive employees, it is expected that all employees abide by the organization's policies, which would lay out the foundation for sustainable data governance. Specifically, an effective data governance program will ensure that current policies are not only followed but is also reflected in the organization as a whole. However, Mittelstadt states that "[t]rust is often cited as a key value in data governance policy and oversight mechanism[s],yet [it] is often poorly grounded in a philosophical sense."[75] Framing trust as the center for data governance policies while not actually practicing in that manner is deceiving and does the exact opposite of "good" data governance. While some might argue that there is room for interpretation regarding the goals of data governance, organizations should strive to do more than just meet the minimum conditions with their