Bryant University
Bryant Digital Repository

Honors Projects in Data Science | Senior Honors Projects

4-23-2023

HIPAA vs. Medical Research: Improving Patient Care Through Integration of Data Privacy and Data Access

Katherine D'Ordine

Follow this and additional works at: https://digitalcommons.bryant.edu/honors_data_science

 Part of the Data Science Commons, and the Other Medicine and Health Sciences Commons

Bryant University

HONORS THESIS

HIPAA vs. Medical Research: Improving Patient Care Through Integration of Data Privacy and Data Access

BY Katherine D'Ordine

ADVISOR • Dr. Suhong Li

EDITORIAL REVIEWER • Dr. Joseph Trunzo

_ Submitted in partial fulfillment of the requirements for graduation
with honors in the Bryant University Honors Program
April 2023

Table of Contents

ABSTRACT

The purpose of this research is to understand the current relationship between data access and data privacy in the health care industry and attempt to find a way that important health care research can still be conducted amidst HIPAA regulations. There is a lack of extensive research on the impacts of data privacy on health care research due to access regulations, so a survey was created regarding current data processes and recommendations for creating a healthier relationship between privacy and access for research. It was distributed to anyone in health care, analytics, or research to get a variety of perspectives. Tweets were also collected based on medical research and privacy regulation key words and analyzed to further understand the sentiment of these topics within the public view, especially in regard to research.

It was found that a well-regulated process to create a partnership between medical professionals and researchers was a viable solution to work with privacy laws, as well as potentially creating a cross-functional team of patients and experts across different involved fields that discusses needs and obstacles in order to overcome them. The data also displayed that recommendation for de-identifying patient medical records for use in research was also a potential contributor to the solution. Analysis also led to the discussion of lacking interoperability in health care and the idea that data quality and structure poses a significant problem for the advancements of medical research, with a recommendation for more interoperability among health care databases to become a priority. Though few are discussing this topic, most analysts, researchers, and medical professionals who participated in the survey agree that the relationship between HIPAA and medical research needs to be discussed and has the potential to lead to legitimate access. Finally, finding a way to get the necessary data to researchers would help to enhance medical treatments and gain a better understanding of patient experiences in order to improve them.

INTRODUCTION

During the age of data everywhere, there are many ways data can be used to significantly benefit people, but it has the potential for being misused. Ethical considerations are imperative to data and are more complex than other ethical discussions since improper use of this data can impact many aspects of the human experience. There is no authority that can give entirely correct information about the proper ways to use data, so it ends up being discerned by the people who use the data to understand the ethics surrounding it. Typically, people air on the side of caution when it comes to risk, but it is important to note that risk and benefit must have a balance. As the article states:

> "As leading data ethicists Floridi and Taddeo put it: 'On the one hand, overlooking ethical issues may prompt negative impact and social rejection …On the other hand, overemphasizing the protection of individual rights in the wrong contexts may lead to regulations that are too rigid, and this in turn can cripple the chances to harness the social value of data science.'" (Hand, 2018)

This essentially states that need for balance and acknowledges that current data may evolve and grow into something that would benefit society, especially since this type of moral dilemma does not have a simple solution; it is more to provide knowledge and resources for people to make their own ethical choices regarding data privacy for both current and future data. Future data uses cannot be known for certain, so it is difficult to categorize all data into public or private. The data itself does not come with ethical considerations; it is more about the way that data is used and what types of analysis data will be subject to (Hand, 2018).

Most of these ethical considerations have to do with human subject data. The E.U. defines this personal data as information that can be connected to a person through direct or indirect means such as through reference of identifying numbers, online presence, location information, socioeconomic details, and other factors. The general data protection regulation (GDPR) states:

> "The protection of natural persons in relation to the processing of personal data is a fundamental… The right to the protection of personal data is not an absolute right… it must be considered in relation to its function in society and be balanced against other

> fundamental rights, in accordance with the principle of proportionality… processing of personal data should be designed to serve mankind." (Hand, 2018)

This part of the regulations shows the difficult decisions about personal privacy and its significance in comparison to promoting a wider good when it may not necessarily be ethical to refrain from giving out readily available data for the benefit of everyone affected (1 Hand, 2018).

An article written to look at literature surrounding ethical data use for future regulations and ethical discussions looked closely into the biomedical side of data with medical information being sensitive and regulated. The article works with main concerns including data protection and anonymization, who owns the data, and how much people know about their data being used. Additional concerns are listed specifically for biomedical data that are important topics of future research in this area, including:

> "The need to distinguish between 'academic' and 'commercial' Big Data practices in terms of potential harm to data subjects[,] future problems with ownership of intellectual property generated from analysis of aggregated datasets[,] and the difficulty of providing meaningful access rights to individual data subjects that lack necessary resources." (Mittelstadt & Floridi, 2016)

These concerns all relate to the usage of biomedical data and indicate that future research about the topic is imperative for any sort of ethical decisions to be made between privacy and access.

When does the benefit to the greater good outweigh an individual's right to privacy of their personal health information? What if there were to be a well-regulated, structured process implemented where research students could partner with medical facilities or organizations to utilize their data for various research projects? Bringing this type of data into research for varying levels of researchers conducting studies may drive improvements in medical and mental health treatments, improving the lives of many in society, but it could also be a cause for concern regarding privacy in used incorrectly; this balance is what needs to be found.

LITERATURE REVIEW

Privacy, its Importance, and How Much We Actually Have

Privacy does not have a universal meaning, but in terms of health care privacy, refers to the collection, storage, and use of information that can be considered identifiable. This includes rules about what data can be collected of an individual in the first place, as well as what that data can be used for and how the person can use their own data. It is important to note that security and confidentiality are also used as a synonym for privacy, but they are not the same thing. Confidentiality is more about the rules for people who receive personal information and what they are allowed to disclose to third parties. Security is more about the technical measures implemented to prevent access or modification of data, prevent denial of services attacks, and physically protect the system or computer in which the data is stored (Institute of Medicine, 2009).

Another important aspect of privacy in the health care industry is whether a patient has willingly authorized specific data to be used for specific things. Patients may be obligated to sign off on the authorization of the data usage or feel as though they must give consent to releasing information to continue with receiving treatment. This can happen with some types of consent forms because they are designed to protect the medical facility from liability of any sort. Even if they signed off, they may not have fully understood what they signed off for considering the complex details. Privacy should be handled with care as it also has value to society, giving people comfort in taking part in research activities and studies without their information being released to the public (Gostin & Nass, 2009).

The amount of privacy that people actually have with their medical records is much more complicated than the idea of just keeping it private. Over the course of an average hospital stay, several hundred different people may see some fragment of a patient's records. Direct access to this data is typically needed for information about the care for the patient. Those directly involved in patient care also have access to patient records in order to provide the proper treatment and have a complete understanding of what a patients need. Other additional services in a medical setting such as labs, therapies, or radiology also need access to this data to perform their jobs safely. On the payment side of medical records, insurance companies

and other third-party billings and accounts have access in order to determine how much money they should pay for specific treatments as well as to see how at risk someone is for illness. Aside from primary users of health information, secondary users sometimes have access to data for supportive services like risk management, medical schools, the support of legitimate medical facilities and individuals, and medical research. More users that indirectly may have access to health information are those who offer management, marketing, or database implementation services. Governments may also receive data to report infectious diseases or discover types of abuse within the community. They may need to disclose allergies to schools or have information in court situations for people under 18 in custody matters for example. Medical information is also making an appearance on the internet most notably through patient portals. These allow patients to be empowered to look at their own health records to understand more about their care and challenge details when they feel something isn't right, but they also allow health care providers to access data on a patient from multiple different locations for quicker care. Even with all these policies in using patient medical records, some places still don't protect patient privacy like they state they do, nor have the proper security implemented to follow through on such claims. This is when privacy breaches may occur, with the occasional story about horrible breaches in patient medical records, but mainly disclosure of personal information such as STDs, mental illnesses, and genetic disorders, which can cause social, economic, or emotional distress. This can lead to discrimination in the workplace or increased insurance rates due to predisposed illness. Some privacy breaches are accidental where records were released on incorrect platforms not behind logins and passwords, but some are malicious and have targeted subjects to harm specific people. There are even some situations regarding people with legitimate access to patient records that use it for the wrong purposes, but this isn't always caught or considered a privacy violation. There is no one right answer to data privacy in the health care industry (Scott, 2000).

Health Research and its Importance

Health research is a clear investigative process that includes the creation, implementation, testing, and evaluation of results in order to create more information for future use and knowledge. Some research is classified as clinical trials where volunteers participate in

studies to test different types of applications using newer medical knowledge. Research also includes data-based methods including biological samples and information from patient records. Both methods have yielded impressive information that detect patterns and form new types of medical treatments. With health care data transitioning quickly into a more electronic platform, research is very possible (Institute of Medicine, 2009).

Just like how privacy holds a high level of importance in society, so does research, as it can provide new information about illness, treatments, care practices, and societal needs. It can help determine what the largest health issues are and provide the necessary data to discover patterns and harbor knowledge that can lead to the invention of life-changing medical interventions for overall increased patient health (Gostin & Nass, 2009).

HIPAA Initiation and Data Regulations

Health care data regulations have become stricter with the implementation of the Health Insurance Portability and Accountability Act (HIPAA) in 1996. This act was created to protect sensitive information in health care so patient data could not be given out without explicit authorization from the patient. The exception to this is that data may, in very specific cases, be released for the purpose of research through the Institutional Review Board (IRB). There are 18 types of data that are considered personal identifiers according to HIPAA and, if not fundamentally essential, should be removed from datasets since they are not always required in medical research. They relate to names and contact information, dates, geographic location, addresses, account and identification numbers, internet locations, and biometric identifiers (Department of Health Care Services, 2022). A full list of identifiers explained in further detail can be found in Appendix A.

The HIPAA Privacy Rule and its Shortcomings

Recently, more attention has been brought to the differences between current and past ethical concerns of personally identifiable information. The current system that regulates the use of health care and patient data is designed to protect patients participating in clinical trials from experiencing physical harm, so the way database research is viewed has shifted. Database research does not rely on accessing human participants; it is much more interested in the data

contained inside of the database. The misuse of this type of data creates risks that are more emotional, social, and economic. When it comes to getting approval from the IRB for a study, different IRBs may give different responses, feedback, and ultimate acceptances or rejections due to their interpretations on the guidance from health care regulations. In some cases, federal regulators get involved to try and accommodate any issues that may arise to allow research to commence, but this can only do so much. A potential new way to handle health research differently than clinical studies could benefit both patients and research (Meslin, 2006).

The HIPAA Privacy Rule aims to protect personal health information and require proper storage while still creating a line of usage in health care settings to provide the best care to the patient. At the time, Congress noticed that patient health records were also used in the process of health research, so they wanted to uphold their law that would allow both privacy and controlled access to be honored. Information that is protected under the Privacy Rule is the information that is covered by health insurance plans or providers, and they may not disclose any information without authorization from the patient or the patient's representative in certain circumstances. It applies to electronic, verbal, or paper health records, but when it comes to health research, they must follow a list of provisions laid out in the HIPAA Privacy Rule. The Institute of Medicine states:

> "The Privacy Rule permits a covered entity to use and disclose PHI for research purposes without an individual's authorization if the covered entity obtains either (1) documentation that an alteration or waiver of the individual's authorization for the use of disclosure of the information has been approved by an IRB or Privacy Boardy, or (2) specified representations from the researchers that the PHI is being used or disclosed solely for purpose preparatory to research, or for research using only the PHI of decedents. A covered entity may also use or disclose PHI without an individual's authorization if the PHI is contained as part of a 'limited dataset' from which specified direct identifiers have been removed, and the researcher enters into a data use agreement with the covered entity." (4 Institute of Medicine, 2009)

In addition to this, the Privacy Rule does not apply to medical records that have been completely de-identified, meaning that all 18 of the HIPAA personal identifiers have been removed properly and there is an insignificant risk that the data would be reidentified in some way during the use of the data (Ness, 2007).

When it comes to privacy regulation flaws, they exist in three main areas: coverage gaps, inconsistencies, and variable interpretation. In terms of coverage gaps, some patients are left without protection when their privacy is invaded. Any information that is stored within noncovered entities is not regulated by any federal process, including database management groups, health agencies, and pharmaceutical entities. While this is the case, the Common Rule from Congress which applies to human subject research, is concerned with studies through the federal government. Most other countries have privacy regulations that are consistent among all health data, whereas the US has missing regulation coverage. For inconsistency, these differences mentioned above create confusing standards regarding consent, data de-identification, and getting patients on board with studies. Under the Common Rule, patients can authorize use of their data in studies that could happen in the future if they are under IRB oversight, but the Privacy Rule contradicts this ability by stating that patients cannot give consent to the use of their data for future research under any circumstances. The Common Rule also allows for a looser standard when it comes to de-identification and personal identity, where the Privacy Rule has a much stricter standard for determining how likely it is that the data will be reidentified. Finally, when it comes to asking patients to participate in studies, "The Privacy Rule creates an artificial distinction between researchers who are internal or external to a covered entity, and offers less protections than the Common Rule," (Gostin & Nass, 2009). On the topic of variable interpretation, federal rules are understood and used in very different ways depending on the board that is deciding upon them. Less people are wanting to serve on IRBs or other privacy boards due to the complex knowledge necessary and the amount of work involved, so regulations and interpretations vary substantially. Research has been both hindered and even discontinued due to overly conservative interpretations from some boards when other boards would have allowed it since they felt the Privacy Rule allowed it. Criteria in these rules use non-measurable words that

don't have a standard meaning across all boards, and therefore confuse what is actually allowed and what shouldn't be (Gostin & Nass, 2009).

What Professionals Say about HIPAA's Impact on Medical Research

A national study of clinical scientists was conducted on the HIPAA Privacy Rule's impact on medical research, and it was determined that most of the surveyed individuals felt that the Privacy Rule had a significant negative impact on medical research with human subjects due to the added time, costs, and increased unknowns. Only 25% felt that the rule had increased patients' privacy in the process (Ness, 2007). Another study aimed to evaluate HIPAA regulations and the impact they pose on those applying for IRB exemptions by looking into time frames where applications were approved or denied and investigating additional factors. They found that HIPAA seemed to get in the way of research using medical databases and increased work for everyone involved in the process, even with the careful consideration of ethical uses of data. More studies ended because they were not able to meet the requirements, and it was not proven if privacy protection truly increased with this new rule. In the past, using medical records in research has been essential for developing new treatments, it was already time consuming then, and recent rules have not helped to improve that process (O'Herrin et al., 2004).

Potential Solutions for HIPAA and Research to Coexist

In a book about the HIPAA Privacy Rule, a committee for health research and privacy worked to develop recommendations in hopes of increasing privacy while still working to create a feasible way to conduct health research with that information. Their main suggestion was that "Congress should authorize HHS [Health and Human Services] and other relevant federal agencies to develop a new approach to ensuring privacy that would apply uniformly to all health research in the United States," (Institute of Medicine, 2009). This involves making sure HHS takes health research out of the umbrella of rules under the HIPAA Privacy Rule, and should instead emphasize secure, accountable, and transparent health care research in a separate entity of rules still monitored appropriately. If no change is going to be made, they at least suggest a revision to the existing Privacy Rule while a longer more thorough process is being executed, which includes the HHS creating more informative materials that give details

to IRBs for consistency purposes as well as revising several details in the rule that prohibit certain research from happening, which are ineffective at proving the intended privacy to patients anyway. In addition to this, the committee also recommended that research institutions storing and using medical records should adopt more security and protection over identifiable information and keep the public informed about the value of research (Institute of Medicine, 2009).

In a journal about changing the HIPAA Privacy Rule to improve both research and privacy, they suggest several revisions and discuss the future with these changes. They suggest that the Department of Health and Human Services (DHHS) should improve their information for privacy in research, promote de-identified data usage, be more communicative about research and its purpose, and stay consistent when it comes to research preparation and subjects. The author also suggests that the DHHS should allow future authorization from patients, allow consent for specific purposes, communicate the benefits of using biospecimens, and finding a way to connect more than one data source to make better use of database capabilities. The newer regulations put in place during the time of this journal did have an impact on the sale of medical records for marketing purposes, but it doesn't currently solve any of the issues with it hindering research. Both can be improved if it becomes more about consistent oversight along with ethical and informed practices so public health can be protected. This important, high-level research can also be conducted to improve knowledge and society (Gostin & Nass, 2009).

A third study was conducted on what patients want with the use of their personal medical information when it comes to research by surveying over 600 people with chronic medical conditions or knowledge of familial conditions about their feelings towards privacy of their medical records. This study provided three recommendations for a public policy that can provide balance between privacy and access. The first lies with those who are in research and policy creation; they need to improve their communication to patients and the public about why this kind of research is important for them or their families, and the important role medical records have in the success of this process. They could explain the benefits with proven examples of its impacts, but also explain the situations where the mishandling of data could happen. This suggestion emphasized better communication and transparency in the

process of using medical records. Second, they recommend bringing up a blanket consent form with patients for them to understand what their medical records could be used for in research. This might bring patients to be more willing to release information when they are aware of it and given guidance in their decision instead of finding out later on that their information was used in a way they didn't want it to be used. That being said, asking for future consent has become a debated topic among professionals, and this should not fully replace obtaining individual consent from a patient for a specific purpose if it is something that can be realistically done. Third, it is understood that some patients will never provide consent to the usage of their data regardless of their level of understanding of the study and how low risk a study may be. This should not be seen as a negative, but as a positive, that the policy keeps both the medical care and research programs in a trusting relationship with one another. It is also important for patients to trust medical professionals that their information is being used for the benefit of everyone and not be suspicious of what is going on with their records. Medical institutions also must take the responsibility seriously and only put through well-planned studies that will produce quality results to avoid the assumption that all studies should be granted data access (Kass et al., 2003).

In terms of more data driven changes to improve the currently limited knowledge capabilities due to privacy issues, there are two main ways medical research can be modified to comply with certain privacy considerations and regulatory requirements. First, in the data, each record can be assigned a unique identifier that follows the record throughout the analysis and allows for data to be connected from multiple sources. The actual storage of what the identifiers mean is separate and secured. This process is called anonymization, and still requires a level of security to prevent the table of identifiers from being accessed. Secondly, records can be combined into one dataset and then used later for analysis. This process is called aggregation, and as soon as it is complete, data can be sent and published since there will be no way to trace back any of the original records and identify individuals. Regardless of either suggestion, informed consent is still a main aspect of ethical research and should be discussed with patients while data is in the collection phase, if possible, to better communicate the benefits of research (Horner, 1998).

A research article describes how to anonymize medical data and keep it usable for analysis through several different methods in order to overcome the patient privacy concern found in the stronger need for patient data in research. All data for secondary purposes must be anonymized. This includes public data typically associated with census data available for download or long-awaited clinical study results for public access. This also includes quasi-public data with extra stipulations that require a user to sign off that they won't work to re-identify the data, link any data inappropriately, share the data, or contact those who are part of the data set, and that they will register their identity for handling such data. Non-public data is much more restrictive than the previous two types of data, and requires a user to enter into an agreement, have effective privacy and security measures implemented, and allow observation from a data custodian, or data usage monitor, while working with the data. For the three types of data, the more private the information gets, the higher chances of re-identification are and the more restrictive the rules for use are as well. In terms of ways to anonymize patient data for research, the author states:

> "Methods for measuring the risk of re-identification can be used to decide how much to anonymise health data for different types of data release. Perturbation that retains sufficient data quality requires data-centric methods rather than simplistic rules regarding how to generalise fields. Anonymisation methods cannot ensure that the risk of re-identification is zero, but this is not the threshold that is expected by privacy laws and regulations in any jurisdiction. Strong precedents exist for choosing suitable probability thresholds for anonymising data. There is a need for anonymisation standards that can provide operational guidance to data custodians and promote consistency in the applications of anonymisation." (Emam et al., 2015)

If data is handled with care and special attention is paid to making sure each level of data access is processed properly in terms of de-identification, then it should be more feasible to use patient data without it being a breach of privacy.

RESEARCH QUESTIONS AND OBJECTIVES

This research aims to understand the current relationship between data access and data privacy in the health care industry in order to improve the relationship and enhance patient care through research. The current connection is in opposition and preventing researchers from having access to the necessary information to make advancements in the care of patients' mental and physical health while in the hands of medical professionals. There is also an element of data structure obstacles where this research aims to find ways to improve the structure to comply with privacy through a more established process of de-identifying data for usage. A framework for these objectives and hypotheses is provided in Figure 1.1.

Hypotheses:

H1. Data privacy regulations will have a negative impact on data access for legitimate research.

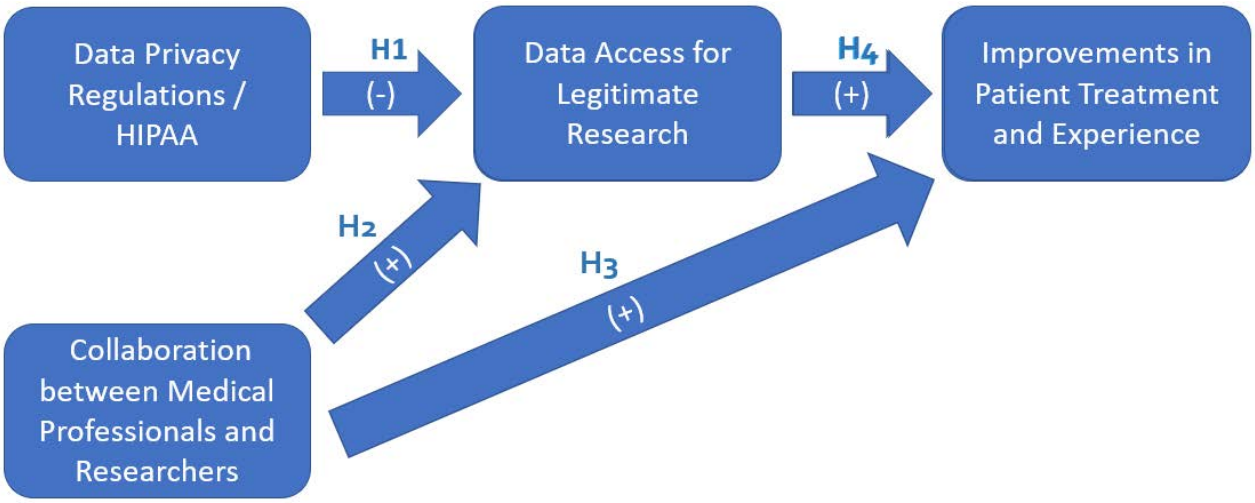
H2. Collaboration between medical professionals and researchers will have a positive impact on data access for legitimate research.

H3. Collaboration between medical professionals and researchers will have a positive impact on patient treatments and experience.

H4. Data access for legitimate research will have a positive impact on patient treatments and experience.

Figure 1.1

Research Framework



The scope of this research is regarding the access to healthcare data, regulations regarding the privacy of that data, and coming up with solutions for using that data properly to benefit medical professionals and the patients in their care. Current research is lacking in extensive studies that aim to improve the quality of care to patients in a medical setting, mainly due to a lack of ability to access biomedical data that is imperative to the success of meaningful research questions. It is also lacking in extensive studies about the relationship between data access and data privacy in terms of the work being done to improve their relationship. This research will provide steppingstones to allowing important research to take place, so patients have better experiences with medical professionals and have a better quality of life post care.

RESEARCH METHODOLOGY

To complete this research, a survey was created and delivered to various medical professionals, people in health care, analysts, and people in research to get a variety of perspectives on the issue. The survey consists of a few demographic and career questions before entering the main portion. It asks about the subject's familiarity with patient medical records in their job and what parts of the data collection and usage process they have participated in to understand their perspective. Then it asks their agreement with various challenges regarding data collection that they have experienced as well as benefits of a well-regulated partnership process between medical professionals and researchers. Towards the end, it asks about their opinions on two recommendations for steps in this overall process, coming from current existing research, that test H2 and H4. The first asks about whether they believe a well-regulated partnership process between medical professionals and researchers would drive better access, and the second asks about whether they believe this legitimate access would actually drive improvements in patient care and treatment options and are referred to as "Research Recommendations" throughout the analysis. All scales that ask for agreement levels are on a 1 to 5 scale, with 1 meaning strongly disagree, 2 meaning disagree, 3 meaning neutral opinion, 4 meaning agree, and 5 meaning strongly agree. Finally, there is an opportunity for the subject to provide other recommendations for overcoming challenges, concerns they have about this process, and anything else they wanted to add to the survey. The goal was to get a wide variety of perspectives on the issue, gauge their knowledge on if given solutions would be successful, and see if any of their recommendations would be feasible to implement and improve the data access and data privacy relationship. The data was collected and transformed for analysis over a period from October-November of 2022.

To enhance the research, 4.1 million Tweets were collected, and basic analysis was done on a subset of these Tweets to understand the point of view and current opinions of the public on this topic. The key words of "medical privacy" and "HIPAA" were used to filter the final dataset from the collected Tweets into a set with roughly 26,000 Tweets. The inclusion of a filter with the word "research" was used for a subset of around 1,000 Tweets for deeper analysis. A Databricks cluster was used for word clouds and to conduct a sentiment analysis to determine if anyone is talking about this issue or doing anything about it.

DISCUSSION AND RESULTS

General Data Demographics and Survey Information

The survey was distributed, accepted responses for just over a month, and 44 responses were received with 41 of them being usable for analysis. Nearly 50% of the participants were above the age of 50 and the data skewed significantly into the age ranges above 40 years old. The gender split was 33 females and 8 males with 0 participants identifying with any other gender. The range of careers was large with Hospital Administrator, Student in research / analytics / health care, and Nurse being the top three careers. Nearly one-third of participants selected more than one career that applied to them, and there were also several write-in responses including Physical Therapist and Insurance Administrator which added other perspectives to the results.

When asked what level of familiarity the participant had with patient medical record collection processes and how the data is structured, only 5 participants said that they didn't have any familiarity, and just about three quarters of participants felt relatively comfortable with the concept ranking their familiarity at a three or above on a 1-5 scale. When asked what stages of the data process they have been part of during their current or former careers, 10 participants had been part of more than one stage, and the most prominent stage with 30 participants was the input stage before data processing as a medical professional, so nearly three quarters of the participants had interacted directly with patients in a medical setting or been part of generating the data from those patients at some point in their careers. The second most prevalent stage with 14 participants was data interpretation and presentation to stakeholders. In terms of the challenges and benefits outlined in the survey, there was a wide range of responses from all career perspectives for each which will be explored in more detail in proceeding sections. It is important to note that in this analysis, there are some categories where a participant could have input numerous options. Each survey was given a unique ID, and the number of participants is the number of distinct IDs present in a given category in a visual. In analysis with those categories, such as benefits or challenges split by career when someone has numerous careers listed, the counts are not inflated with the input of numerous careers. In analysis where opinions need to be totaled by category to see how much is included, such as count of people in each stage of participation in the data process, those

counts do include if someone chose multiple options. All applicable visuals can be found in Appendix C, Figures C.1-C.5.

Correlation

The correlation matrix shown in Figure 2.1 was created as an initial look into the data consisting of all numeric values collected in the survey. This included the participant familiarity with patient medical records, their level of agreement with the effects of various challenges and potential benefits, and how much they agreed that partnership between health care professionals and researchers would improve access and that the access would actually drive improvements in patient treatments and patient experience.

Figure 2.1

Survey Correlation Matrix

	Familiarity	Regulations	Knowledge	Data Structure	Standard Procedures	Data Security	Data Quality	Treatment Options	Patient Experience	Provider Treatment Success	Provider / Researcher Collaboration	Partnership -> Access	Access -> Improvements
Familiarity	1.0000												
Regulations	0.1715	1.0000											
Knowledge	0.0269	0.2848	1.0000										
Data Structure	0.0087	0.0982	0.7315	1.0000									
Standard Procedures	-0.0804	0.1276	0.4353	0.5357	1.0000								
Data Security	0.1438	0.4248	0.3445	0.2790	0.2371	1.0000							
Data Quality	0.0404	0.2925	0.4436	0.4141	0.4749	0.4439	1.0000						
Treatment Options	0.1241	0.2795	0.2539	0.2987	0.3556	0.2173	0.5646	1.0000					
Patient Experience	0.1937	0.4232	0.2333	0.2410	0.2476	0.3507	0.4823	0.8560	1.0000				
Provider Treatment Success	0.1323	0.3409	0.2847	0.3696	0.3391	0.3095	0.4919	0.7629	0.8104	1.0000			
Provider / Researcher Collaboration	0.0555	0.2155	0.2034	0.1770	0.0905	0.2021	0.4108	0.6940	0.7490	0.8078	1.0000		
Partnership -> Access	0.1077	0.4306	0.2004	0.1340	0.1387	0.2439	0.2724	0.2464	0.4118	0.2611	0.4197	1.0000	
Access -> Improvements	-0.0301	0.0019	0.1022	0.1075	-0.1370	0.1865	0.0639	0.0367	0.1028	0.0400	0.0690	0.2613	1.0000

Data quality challenges had a significantly higher correlation to all the potential benefits than all other challenges, showing that the data quality being an issue could be indirectly preventing benefits from being fully experienced. Since participants were asked how much they believe that those benefits could be true with a well-regulated partnership between health care professionals and medical researchers, it shows that data quality might be a more significant challenge standing in the way than previously hypothesized. In terms of other challenges, knowledge of a complex system and knowledge of data structure were the most correlated with a 0.73, followed by the relationship between the understanding of data structure and the standard procedures for inputting data. There were significant correlations

above 0.69 between all the treatment options, meaning that participants felt these options have an overall positive effect on one another and work together in the health care industry. In addition, all benefits correlated highly with one another, showing that one enhanced benefit could produce many more benefits in the industry.

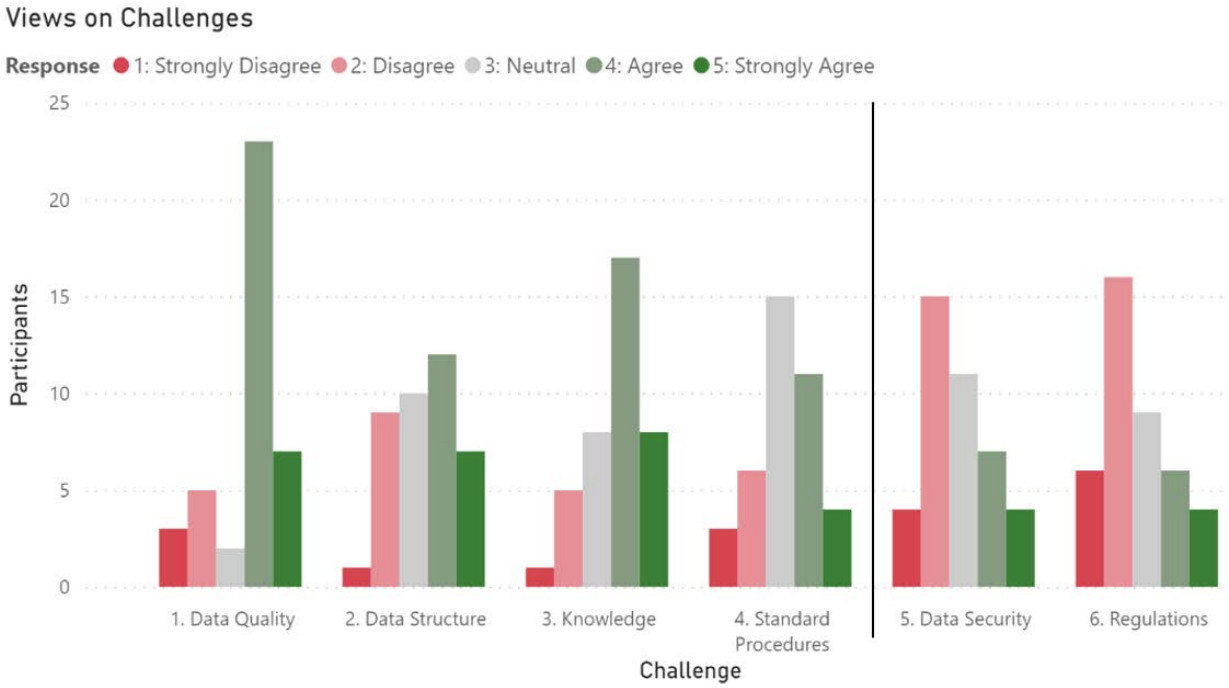
For the level of agreement with partnership meaning better access, regulations / HIPAA was the highest correlated challenge at 0.43, which shows that regulations could be a main contributor to the access issues and that a well-regulated process and partnership could still be seen as a viable solution even with those challenges. Provider / researcher collaboration and patient experience were the highest correlated benefits at 0.42 and 0.41 respectively, showing that these could both be viable benefits with a partnership set up for research to be conducted. There were no significant correlations found between the level of familiarity and any other variables, and the same goes for if the improved health care data access for medical research would actually improve medical and mental health treatments.

Challenges: Patient Medical Record Familiarity and Data Process Stage Participation

As mentioned briefly above, the six main challenges used to investigate H1 through this survey were about data quality and completeness, data security and exposing information, data structure and organization, knowledge of a complex data system, regulations / HIPAA, and standard procedures for entering or storing data in an electronic system. Complete detailed descriptions can be found in the copy of the survey in Appendix B. They are split into two groups to better investigate trends. In Figure 2.2, Group 1 is regarding the data challenges and includes data quality, data structure, knowledge, and standard procedures to the left of the black bar. Group 2 includes the remaining challenges focused on privacy and consists of data security and regulations to the right of the black bar. Overall, people seemed to agree that data quality was one of the largest challenges regarding data in their careers with very few people having a neutral option of it and a significantly higher amount of total agreement and strong agreement than with other challenges.

Figure 2.2

Challenges

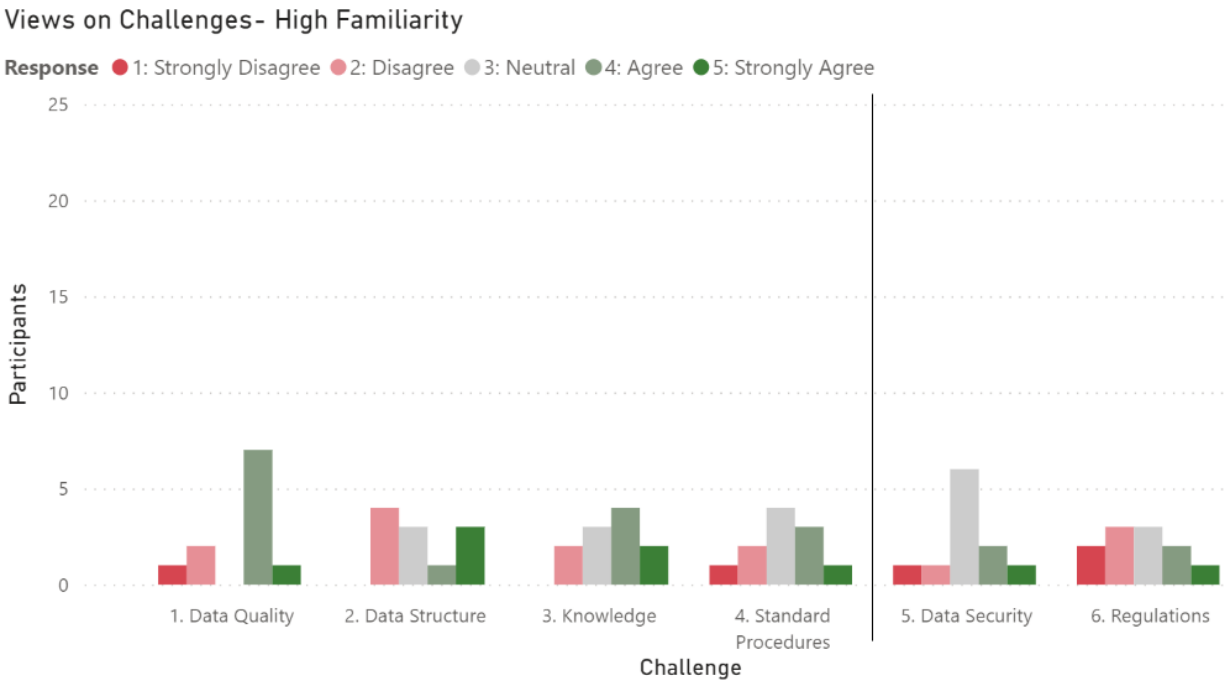


The other data challenges had more agreement than disagreement and more agreement than the privacy challenges. It was hypothesized that privacy challenges were going to be more prominent, but nearly half of the participants disagreed to some degree that both data security and regulations were hindrances in their careers, with close to 10 people on each privacy challenge remaining neutral about it. There was more agreement with the data challenges, which shows it may not always be data privacy causing the separation between medical professionals and the conducting of health care research; it may be that the data doesn't even exist in a structured way that can be analyzed or used for research in the first place. This was surprising due to the high prevalence of health care professionals and people who had worked directly with patients in medical settings where research stated this was the most prominent obstacle. It would also be more expected that analysts or researchers would have this type of general opinion, less so the medical professionals.

When sliced by patient medical record familiarity, more insights are present in regard to the challenges experienced, as displayed in Figure 2.3.

Figure 2.3

Challenges with High Familiarity

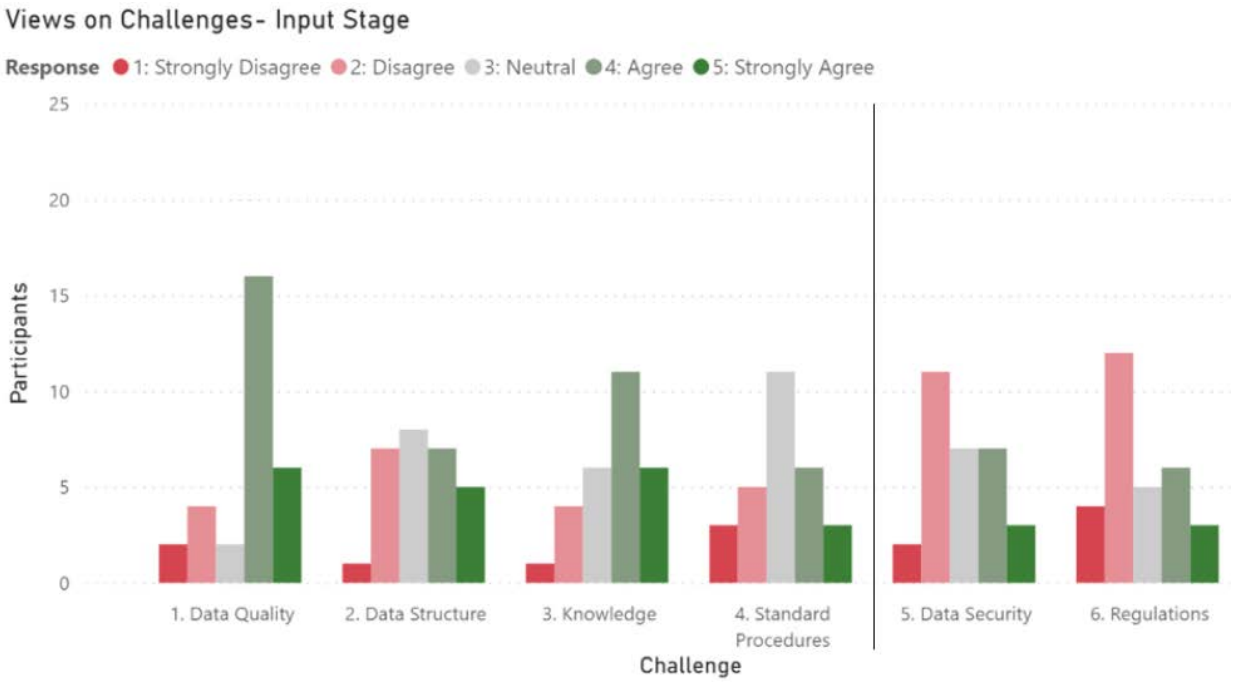


When split by familiarity, the people who had very little to no knowledge of patient medical records (chose a 1 on the scale) had stronger opinions, both strongly agreeing and strongly disagreeing more about how these challenges affect their career. Each challenge had at least one person who strongly disagreed, which could have been attributed to a lack of knowledge on the specific topic. That being said, there were also stronger opinions among the people who had a very high level of understanding of patient medical records (chose a 5 on the scale), also with some disagreement but with more strong agreement than the people with little familiarity. Moving closer to the 2 through 4 range of familiarity, these people seemed to overall have less strong opinions and more neutral opinions about the challenges meaning they might know what they are but aren't sure if those challenges are the ones they are most worried about. It also seems as though the overall trend is that the ratio of people that agreed that the challenges were applicable to them went up with familiarity, which makes sense and shows that people with more knowledge of these patient records understand how complex and challenging this discussion is. This is part of the problem where data may not even be sufficiently structured to use in research, and privacy remains an issue along with it.

In terms of the stages in which people have participated in the data process, input of data and direct communication with patients in a medical setting had significantly more responses than any other stage, shown in Figure 2.4.

Figure 2.4

Challenges with Input Stage



30 people felt this applied to them, and when looking at the data with those participants, they overall represent a similar trend as the entire dataset, including the surprising data point that privacy challenges remain more strongly disagreed with than data challenges, especially since this category is made up of more health care professionals than other groups. The second highest stage represented was the stage including data interpretation and presentation to stakeholders, and more of these people also agreed that data challenges were more prevalent in their work and privacy challenges were less prevalent. This is because more of the data analysts, data scientists, or students in research made up this category and typically have more understanding of the physical data challenges than the health care privacy ones.

Additional Provided Concerns and Challenges

There were spaces provided to participants where they could list additional challenges and concerns regarding the current situations they face and/or regarding potential suggestions or

recommendations provided in the survey. A data analyst in health care who also has experience as a hospital administrator brought up an issue with resources for producing reports containing this type of data, especially since many parts of databases are not accessible to most people and only certified personnel are allowed to write information to those official databases. This perspective shows how even if there was access to the data, it may not be structured or plentiful enough in an organized way for it to be used in research to produce meaningful results. If resources are scarce, then it is less likely to be done in a structured way. A physician and OT were also worried about a similar issue but more regarding errors in data and not having knowledge of how reliable the data input sources are. An analyst/administrator brought up an additional point about research and resources with it being very difficult to provide data to a group that may be working on a similar research effort as another group while keeping the access equal to both groups without causing any issues or competition for resources. All these concerns give way to a lot of additional research that could be done to provide suitable solutions.

On the topic of interoperability and usage of systems throughout different medical facilities, several participants in health care fields brought up concerns with its lack of presence in health care. One stated that different programs contain completely different types of data, such as for emergency departments, inpatient stays, operating rooms, outpatient visits, laboratory notes, or radiology and imaging centers. This requires data to be collected from numerous locations and joined, which begins to bring in the issue of data quality and cleanliness. Another stated that it is very difficult to find all the details needed at the right level since EMR (electronic medical record) systems have so many different styles and fields, that some do not communicate well with one another. A third person mentioned misunderstood data with the complexity that the analysis would bring and was worried it wouldn't even be possible to directly correlate that this research would truly improve patient outcomes. All these concerns speak more to the idea that privacy access is not the only hurdle in this situation; it is also the fact that the data may not even be usable for accurate results even if it were accessible.

A few concerns or other challenges that were brought up were regarding specifics in privacy. A few participants brought up the higher risk of data breaches, ransomware attacks, and

leaking of individual patient information as their main concerns, and others mentioned overall confidentiality and HIPAA especially in mental health data. Two high level professionals in health care companies brought up situations where increased access could cause poor representation in data or some demographics being targeted due to the fact that their health information is now open to different types of sources than it was before, which are factors that would need to be managed properly if access were to be increased without a well-regulated process for specific access purposes. A participant who works in drug development mentioned that some data is proprietary to begin with and there is no way for someone to have access to that data for any purpose yet. This may cause a few issues if access is given to someone's data, but they are part of a clinical study with data that is not accessible. Having separate systems or places where data goes may be a partial solution, but it would take a lot of time to implement effectively into busy hospital systems.

In terms of patient-facing concerns and patient interactions, a medical practice administrator felt as though having a more structured data process and more procedures for inputting data takes away from connecting with patients. If the health care professional is occupied by taking notes and inputting things properly, it may take away from the patient experience and cause them to not feel as though they are being heard and supported through their treatments. Some patients are bringing up concerns about an increase in health care data access or structure because they don't want to feel like a statistic nor feel pressured to be part of research if they feel their options are only to accept it or be "backing away from physician care." This would need to be rectified and methods would be needed to ensure that patient connection and individuality is still maintained in the process.

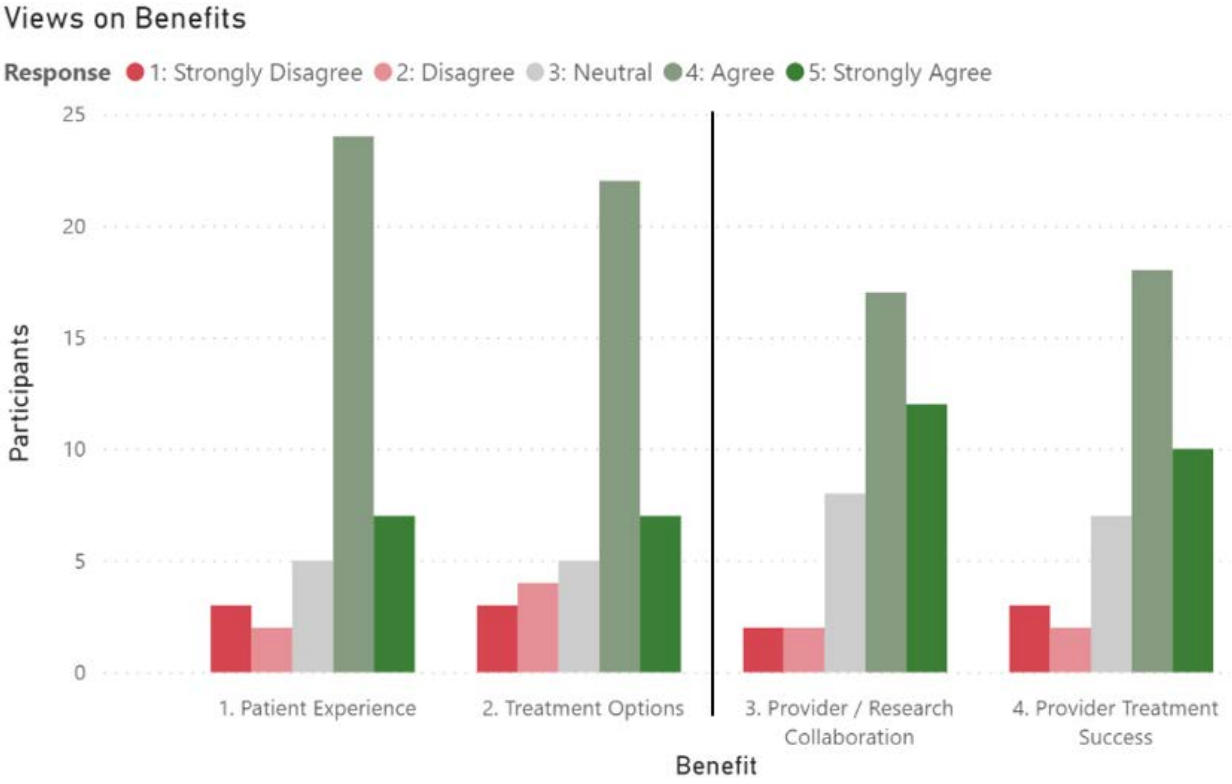
Benefits: Patient Medical Record Familiarity and Data Process Stage Participation

To switch things over to the other side of the analysis where H3 was investigated, there is a lot to be said about the benefits that were asked about in the survey and mentioned by participants. As briefly alluded to in previous sections, there are four total benefits that were included in the survey, fully described in the copy of the survey in Appendix B, and they are split up into two groups. In Figure 2.5, Group 1 is the patient benefits regarding how a patient may be benefited by a well-regulated process between health care professionals and medical

researchers for data access and includes the patience experience and the treatment options as the two benefits to the left of the black bar. Group 2 is regarding the provider benefits for this recommendation and includes provider / researcher collaboration and provider treatment success as the two benefits to the right of the black bar.

Figure 2.5

Benefits

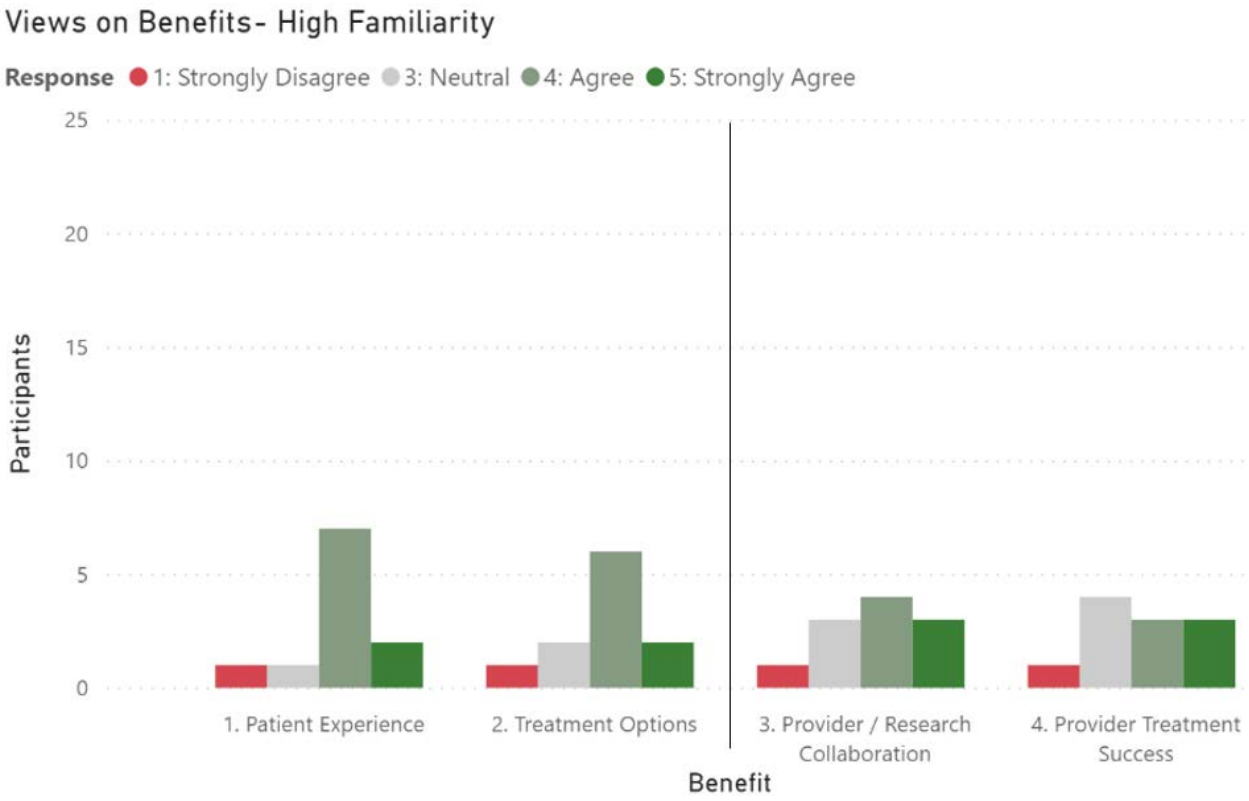


Overall, most participants agreed or strongly agreed with all the benefits. Patient benefits seemed to have a high portion of people who agreed, and there were some people who strongly agreed that there were provider benefits. There were only a few people that disagreed or strongly disagreed with any of the benefits, and it spanned across several types of careers; it was not traced down to a specific category of careers. There were more neutral opinions about provider benefits than neutral opinions about patient benefits, which brings back the idea that there were a lot of patient-facing participants who took the survey, looking out for patients and their needs.

When sliced based on patient medical record structure familiarity, a somewhat similar trend happened with benefits as it did with the challenges and is shown in Figure 2.6, but it was a bit less strong since there weren't enough data points to draw complete conclusions.

Figure 2.6

Benefits with High Familiarity



**Note that a location for the Disagree category is not present, and Strongly Disagree is in its place directly to the left on the Neutral category.*

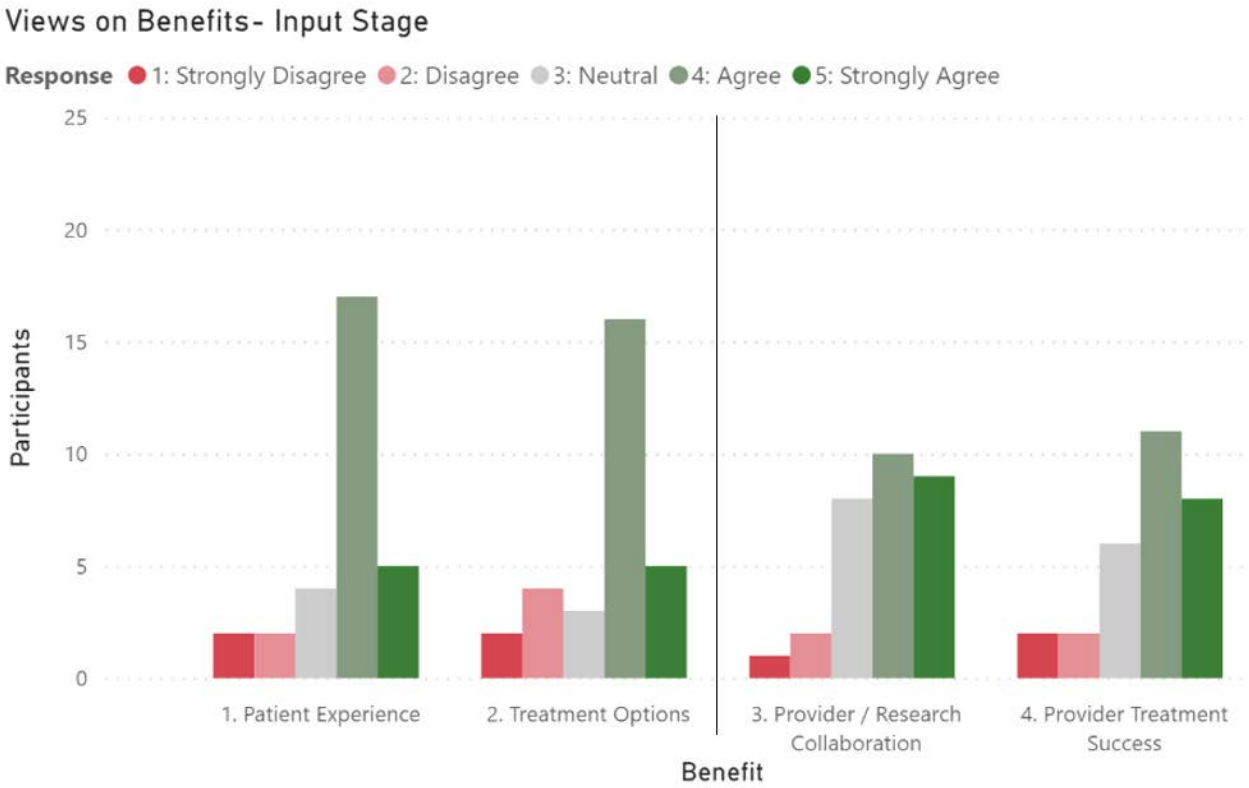
People who had little to no familiarity with patient medical records had more strong opinions in both directions on both provider and patient benefits, but people who were very familiar with them were less strongly opinionated. This could mean that they know a lot about these records, and both want these benefits to be true, but are not sure if they are realistic. This could be a problem if new systems were to be rolled out for better data structure or to enhance research abilities through data access, and people wouldn't think it would work in the first place; it may affect its performance. Similar to the challenges section, familiarity between 2-4

was much less opinionated and had more neutral opinions, but definitely leaned moderately positive throughout.

The benefits shown in Figure 2.7 were also sliced by stage and analyzed in a similar way to how they were in the challenges, kept separated by the black bar between patient benefits on the left and provider benefits on the right.

Figure 2.7

Benefits with Input Stage



With the input and direct patient-facing stage in the data process being most prevalent in the data, its trends similarly represent the overall patient benefits and provider benefits, with the agreement with provider benefits being slightly less prevalent. Considering these are mostly health care professionals directly interacting with patients, they may be focusing more on the patient benefits and agreeing that those could be positive outcomes of better access through a well-regulated data partnership for research. Data interpretation and presentation of data to stakeholders is still the second largest category (it is important to note that it may not have been just because it was the case before, because there are some incomplete responses within

the challenges and benefits data), and people in those roles seem to lean more in agreement with both provider and patient benefits with only a couple disagreeing to any degree.

Research Recommendations

As alluded to earlier, to test H2 and H4, participants ranked how much they agreed that an increased partnership and well-regulated process between health care professionals and people in medical research would increase access to necessary health care data for research, shown in Figure 2.8. From research for this topic, it seems that this might be a good suggestion, so it was important to ask for the feasibility of this. 80% either agreed or strongly agreed with a 4 or 5 respectively, with the remaining portion being neutral and no one disagreeing. They were also asked a second question about whether they thought that this increased access would actually lead to improvements in patient treatment and experience, shown in Figure 2.9. From research it seems like this proper access could potentially have this effect of bringing improvements, and 88% of the participants either agreed or strongly agreed with a 4 or 5 respectively, and the remaining were neutral with only 2 people disagreeing.

Figure 2.8

Level of Agreement with Partnership -> Access

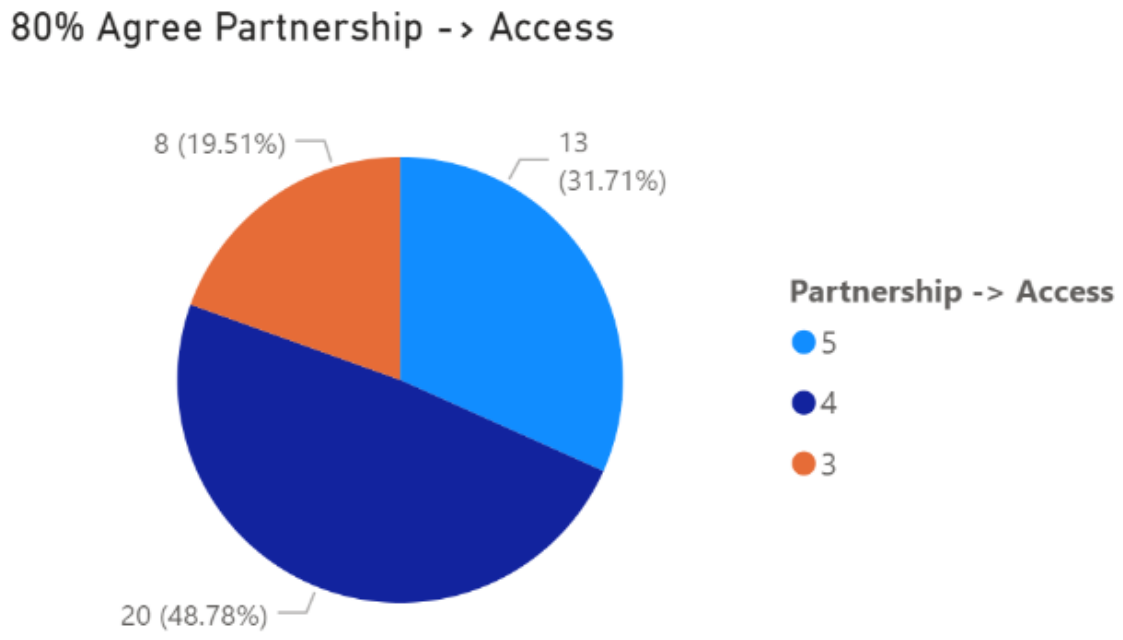
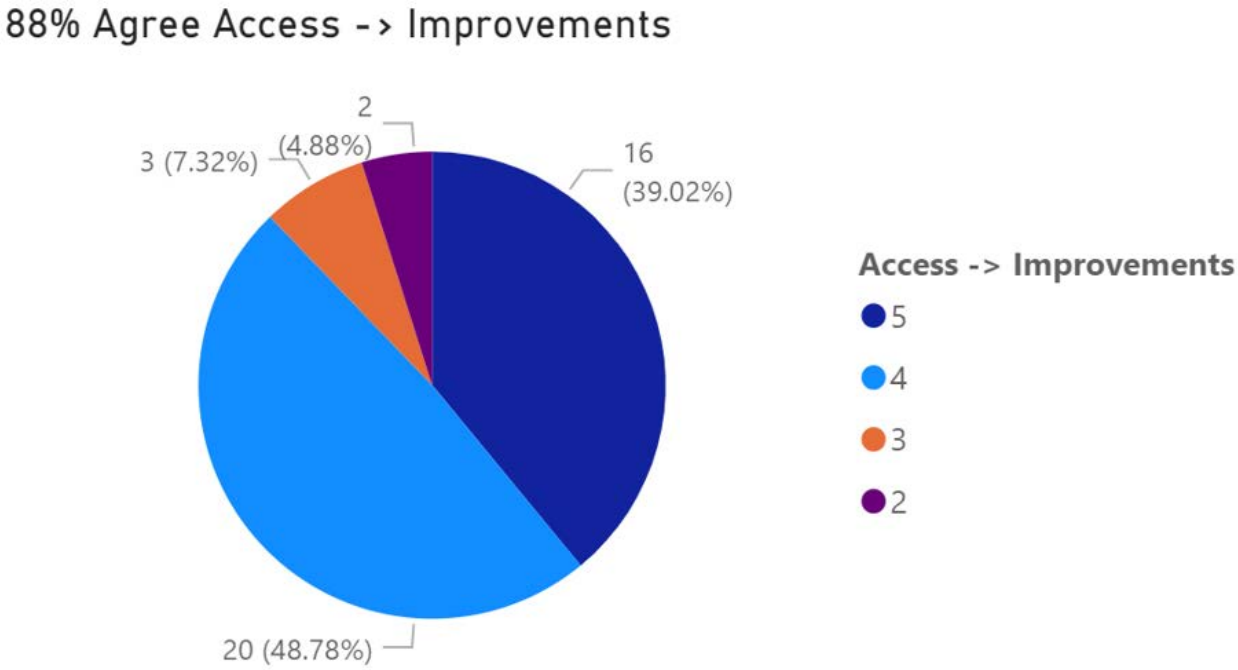


Figure 2.9

Level of Agreement with Access -> Improvements

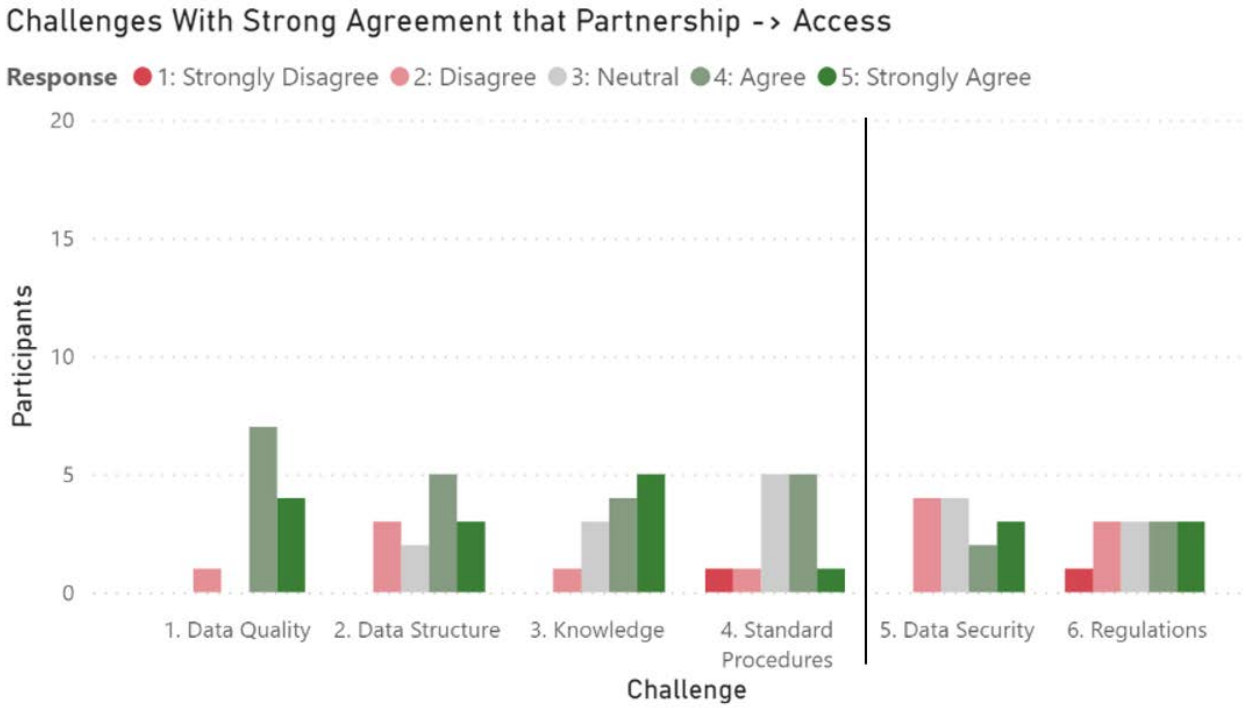


Challenges Related to Research Recommendations

For the first question, when related to challenges, people who strongly agree that a partnership could lead to better access for medical research purposes were the people who also seemed to agree or felt neutral that the challenges listed were applicable to them in their careers, as shown in Figure 2.10.

Figure 2.10

Challenges and Partnership -> Access

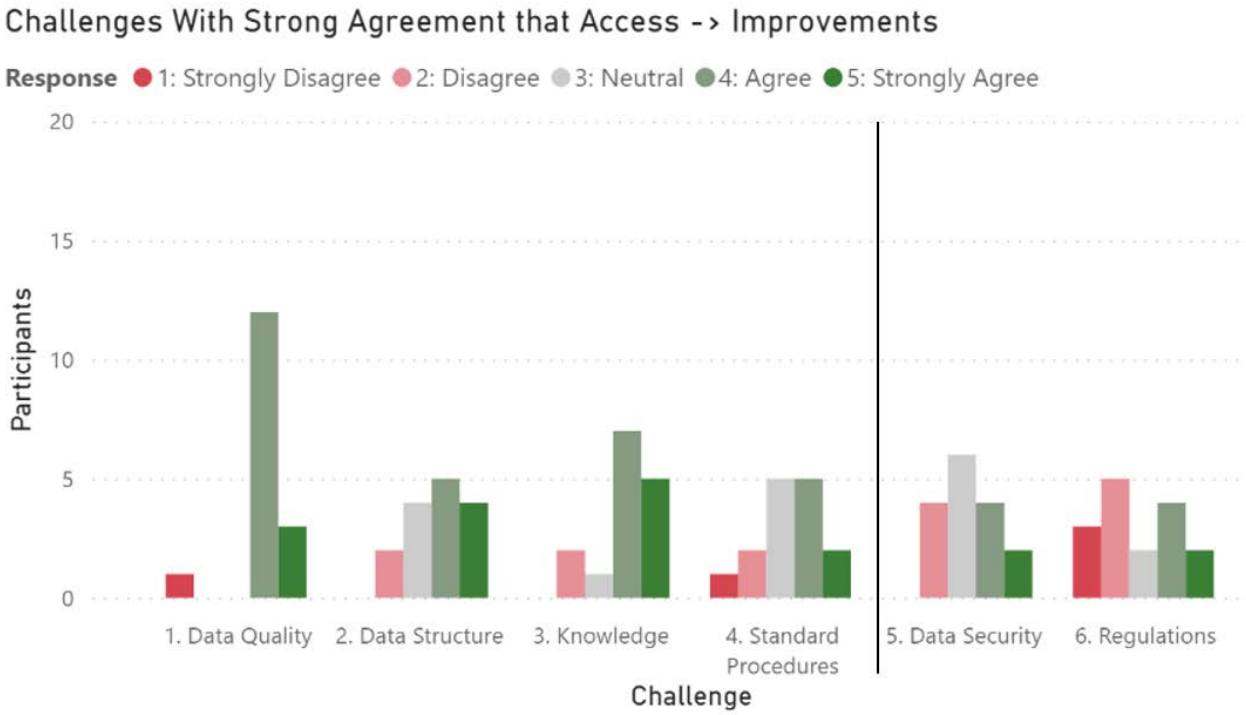


Some people still did disagree with the challenges and thought this would still be a good move for improving access, but it was surprising to see that so many people who experience these challenges thought this could be a viable solution to increasing access. The people who agreed that partnership could be better for access seemed to disagree a little bit more that the challenges affected their careers and still disagreed more with the privacy challenges, and more so than the people who strongly agreed that partnership would drive access. The people who felt neutral about the recommendation had mixed levels of agreement.

The second question follows a similar trend, shown in Figure 2.11. People who strongly agreed that this improved access would actually drive improvements in research, patient experience, and treatment seemed to be the people that agreed more or were neutral about the challenges affecting them in some capacity.

Figure 2.11

Challenges and Access -> Improvements



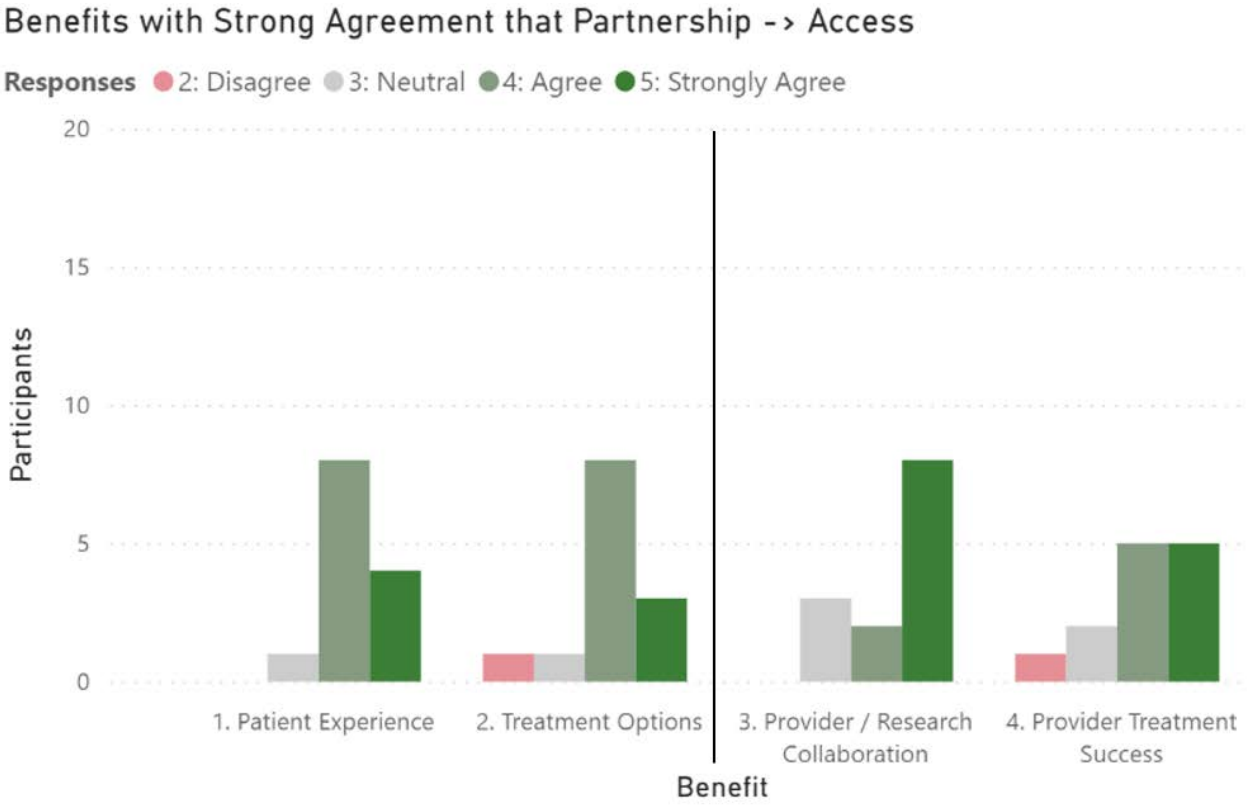
Those who agreed (chose a 4) this access would yield improvements seemed to have higher rates of disagreement that those challenges were applicable to them. This combination could be explained by the fact that people who don't feel these challenges apply to them may be more likely to agree that both questions, partnership driving access and access driving improvement, could be feasible ways to handle this situation. For the people who strongly agree, it seems like they may see the challenges and still want something to be done about it to improve this relationship for health care research to happen. Finally, the people who felt neutral about access leading to improvements did not have any strong opinions about the challenges, and the two people who disagreed about it were in strong agreement with most of the challenges, feeling as though they saw those challenges and didn't think this would be a viable solution. With these two people in upper-level management, it makes sense that they would understand a lot about this process and raise different types of concerns than other participants.

Benefits Related to Research Recommendations

Returning to the two big questions about partnership equaling access and access equaling actual improvements, Figure 2.12 shows benefits seen by people who strongly agree partnership could equal access.

Figure 2.12

Benefits and Partnership -> Access



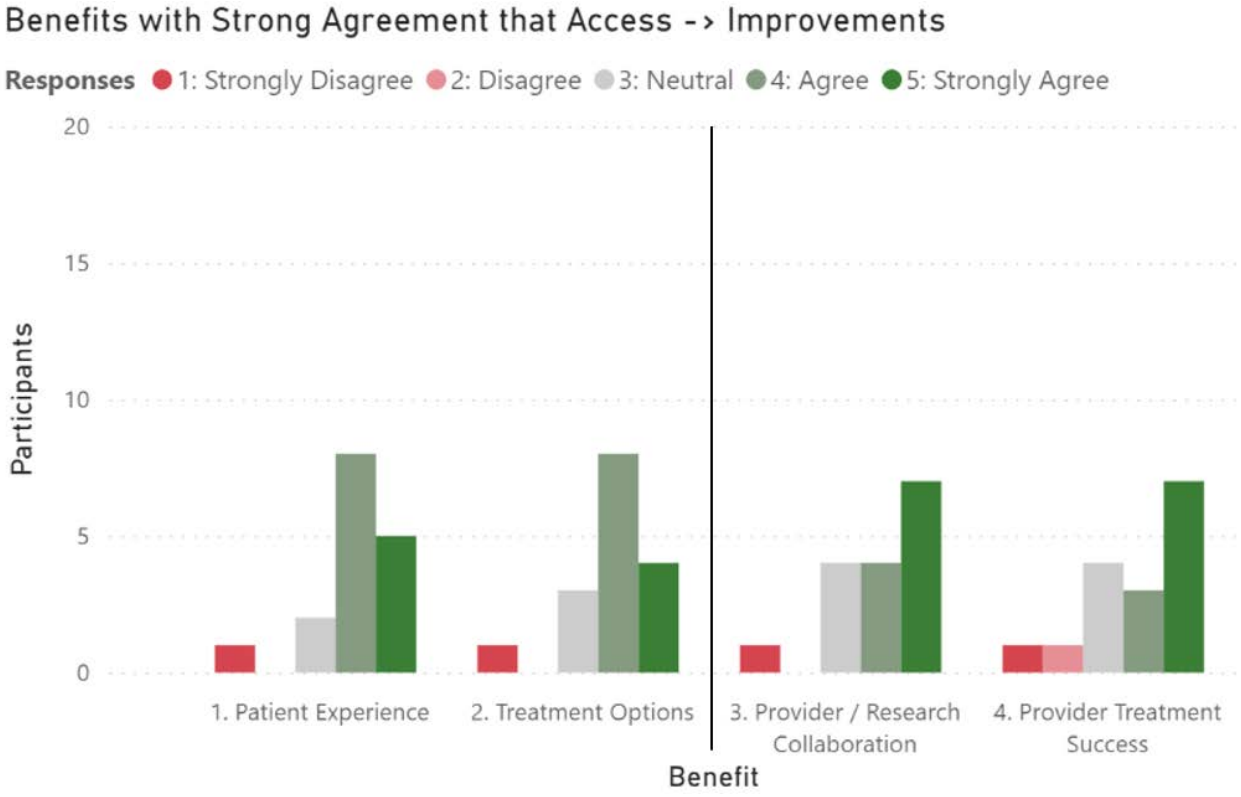
For this first question about the well-regulated partnership for research, almost all participants who strongly agreed that this could be a viable solution to the researched problem strongly agreed, agreed, or were neutral about all the benefits. This shows that there are some analysts, health care professionals, and researchers who see the benefits in having a process implemented and believe that the partnership could lead to access for research. People who agreed (chose a 4) on this question followed a similar trend, but there were a few more instances of disagreement with benefits that followed with the higher number of people who fell into the "agree" category for this question. Most of the people who started to disagree with the provided benefits at this level were in the medical field, and may see this partnership

as feasible, but don't see how it would actually provide benefits. That would be an issue that would need to be cleared up to ensure medical professionals are not discouraged from future changes in these processes if they don't think something is feasible at this time. People who were neutral about this question had a wide range of opinions, but there was a higher ratio of disagreement with both provider and patient benefits than before. These people spanned different careers, but most were not directly in contact with patients on a daily basis and may both not see as much of the benefits and do not know if a partnership will help with access.

For the second question about if this increased data access for medical research would actually drive improvements in research and patient experience, there were still hardly any responses that didn't either agree or strongly agree with it, and the comparisons are provided in Figure 2.13.

Figure 2.13

Benefits and Access -> Improvements



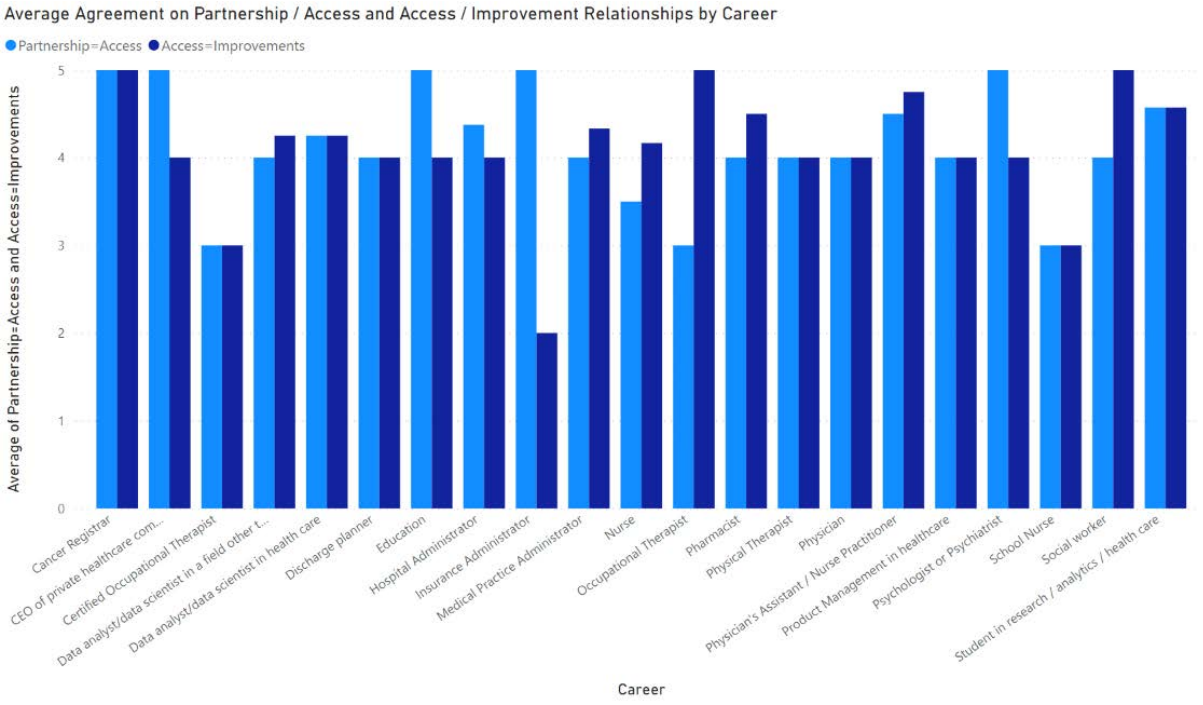
As for people who strongly agreed that access would lead to improvements, only two people answered anything that wasn't either neutral or a level of agreement when asked about the benefits. These people seem to agree that these benefits would be applicable and could make become true in the medical world if this health care access partnership were to actually drive improvements. That would make it very encouraging for a system to be implemented for research if a high proportion of participants believe the research would have the intended purpose of benefiting both patients and providers since both seemed to have similar ratios. For the people who chose 4 and agree that access would drive improvements, there were a few more disagreements with benefits, but still a higher proportion of agreement or at least neutral opinions. There were more patient benefits that were disagreed with at this level than provider benefits, but most of those participants were patient-facing.

Research Recommendations by Career and Data Process Stage Participation

To round out the discussion of why some participants feel differently about these two research recommendations, Figure 2.14 provides a direct comparison by career.

Figure 2.14

Average Agreement with Research Recommendation Relationships by Career

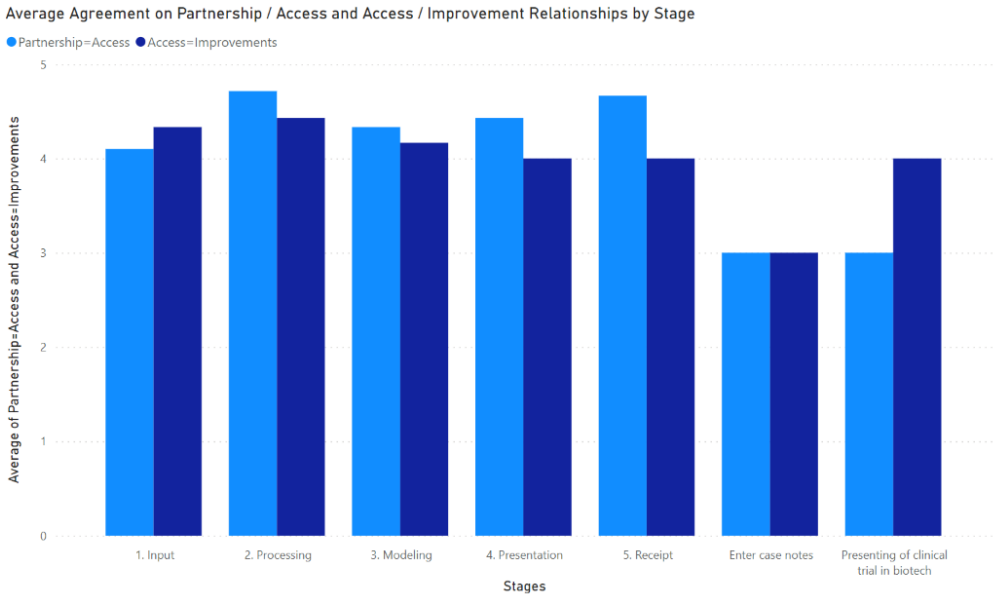


Separated by career, the two questions have been placed next to one another to show what stage of these high-level recommendations people are finding to be feasible and which they are hesitant to agree with for one reason or another, whether they think it can't happen or are worried about what might happen if it does. Several of these values are the same for a given profession, but a few stick out such as the insurance administrator and occupational therapist. While there is only one of each, the insurance administrator sees how the partnership could lead to better access but hesitates when it comes to that access actually having an impact on research and improving patient care. This could be due to position and what they are able to see from their perspective of the health care industry. The occupational therapist felt the opposite, where they are not sure if they can see how a partnership would be the right way to go about gaining access to data but is in strong agreement with the idea that access could actually play a role in improving medical research and patient outcomes.

On a stages of data participation basis shown in Figure 2.15, the patient focused stage (the data input stage in direct contact with patients) is the one where there is a slightly higher average in agreement that data access would drive improvements in research and patient care than the agreement with the partnership being the way to get access, though they are still both in the range of agreement (around a 4 on the scale).

Figure 2.15

Average Agreement with Research Recommendation Relationship by Stage



This could be because they see how it could lead to improvements but may have other ideas that are more realistic for that to happen based on what they know from their expertise. All other standard data process stages typically carried out by analysts, people in research, or others in similar non-patient-facing fields have a slightly higher agreement that this partnership is a viable solution to gaining access, and a slightly lower agreement that the access will actually lead to research that improves patient outcomes. Again, these are both still in the "agree" category around 4 on the scale, some agreement with partnership driving access being closer to a 5 for strong agreement. This may be explained by their expertise in how data access could or should be happening, and they have had experience where access hasn't been effective (for example, if the data isn't usable or readable/bad quality) and they don't know if it would be similar with the notorious complexity of health care data. What all this hesitation shows is that nobody is an expert at everything, and each person probably has hesitations based on their own industry knowledge. This can lead to less progress being made in the industry or cross-functionally; and it is worth looking into whether something can be done about it to provide more communication between groups.

Additional Provided Comments and Recommendations

Before jumping into the conclusions, in the additional comments and recommendations section of the survey, a few participants chose to provide a little bit of recommendation or extra information. For example, a pharmacist who strongly agreed with most of the challenges mentioned that even they don't have access to a lot of information, it is only what insurance companies give them since they are not connected to medical offices. However, through partnerships with insurance companies, pharmacies can get more information about patient treatments and medications to "suggest additional therapies, thus making more profit for the pharmacy." Though not directly related to the questions at hand, it brings up an interesting suspicion about if medical data is as private as it is discussed to be since it may be shared for the profit of another segment of the health care industry. That being said, it is also data that when in the right hands, could potentially improve the regimen or treatment for a patient. Since it can go both ways, this topic is something that would be very interesting to explore in future research.

A medical practice administrator stated that they agree that a collaborative effort to increase access for legitimate research would be extremely beneficial to both patients and providers / medical practices. Regarding a well-regulated and structured partnership process, one participant with knowledge from several applicable careers felt that even though it is currently tough for people in research to get access to the protected data they need to provide meaningful improvements to the health care industry, this type of program could be a great option and is feasible. Another person summed up an aspect of this research quite well in stating, "Risks aside, I think democratization of data accessibility can help harmonize practices and costs across the US health care system." There may even be unintended positive effects to this type of change to health care and medical research industries that may be indirect, but still significant. Another professional with several levels of health care career experience gave a general recommendation for how to go about successfully getting data for legitimate research and having it positively impact patient care. They stated:

> "The most important one I can think of is educating and informing healthcare providers on the utility and application of the data. Also, in the mental health field - you may want to work with large state-run credentialing boards to reach a broader audience of mental health workers. Many are in small to private practice and their clinical data would vary significantly from say college counseling centers, inpatient units, IOPs or community mental health centers."

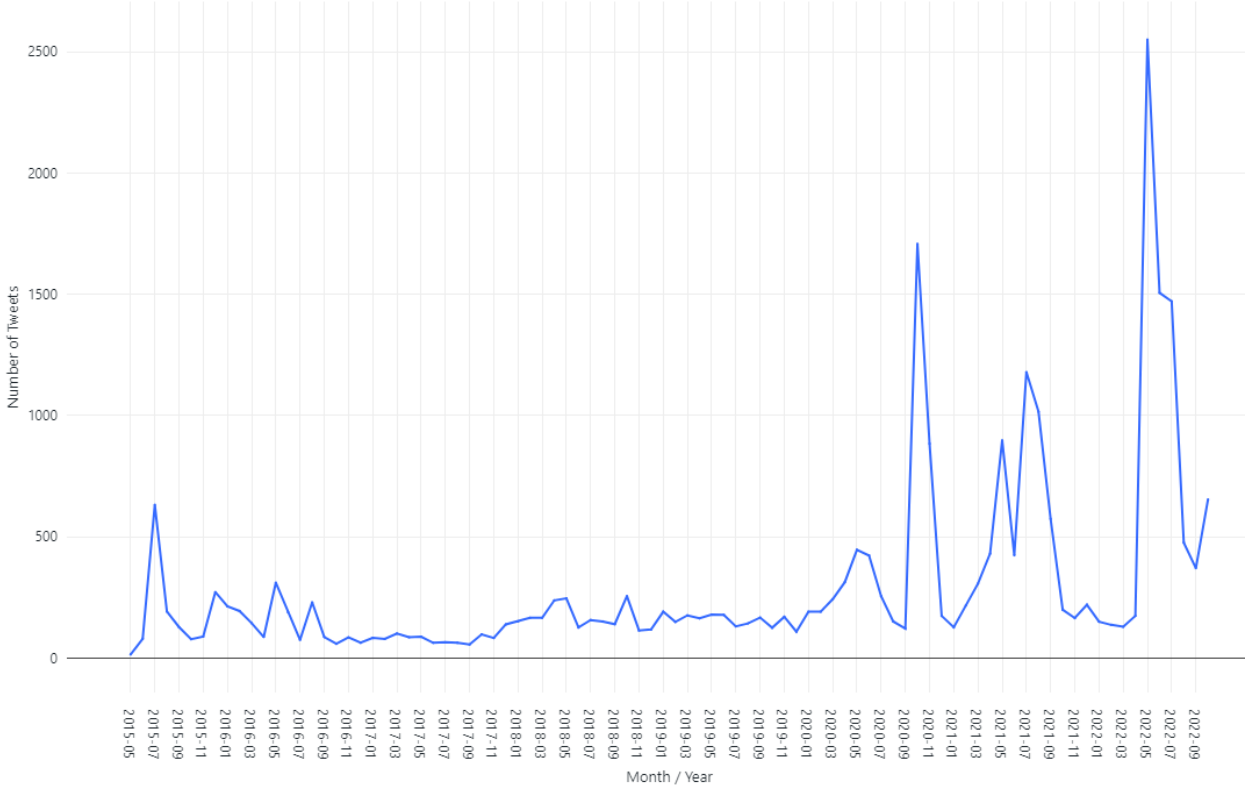
This brings up an important point about who to work for especially in terms of improving mental health outcomes of patients and alludes more to the point mentioned earlier where communicating with professionals from numerous fields cross-functionally would allow experts to share knowledge and create a solution.

Secondary Twitter Analysis about Medical Privacy and HIPAA with Medical Research

A secondary analysis was conducted with Tweets to investigate the public opinion of various concepts contained in this research. The 26,000 Tweets mentioned earlier were utilized for the purposes of this analysis, all of them were relating to either "HIPAA" or "medical privacy." Figure 3.1 shows a time series of the total Tweets collected by month.

Figure 3.1

Time Series for Number of Tweets Discussing HIPAA or Medical Privacy



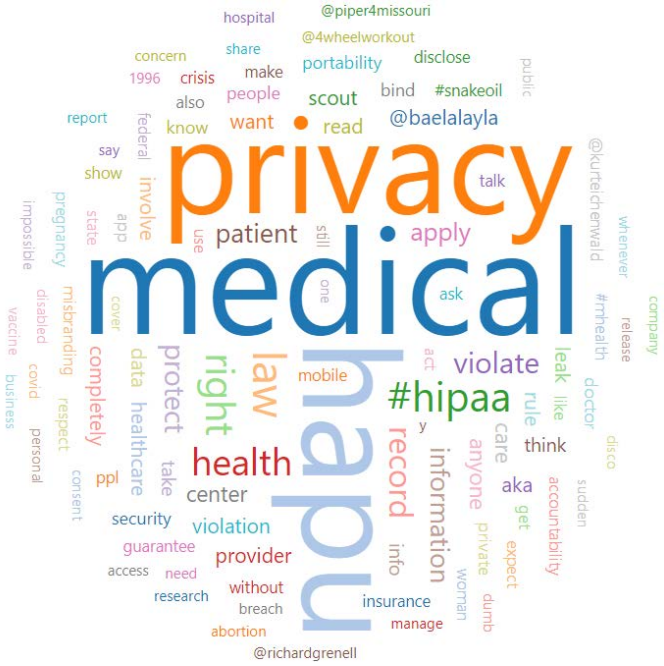
The number of Tweets collected across the years showed a significant upward trend when 2020 started and continued climbing even more in 2021 and 2022, showing that the COVID-19 pandemic caused a lot more discussion of medical privacy and HIPAA among the public. This was a bit surprising given that the total number of Tweets collected in the 4.1 million decreased after 2020, so it seems as though there is a higher ratio of discussion about medical privacy in general than before; it is not just an inflation of the number of overall Tweets from the pandemic or other worldly events.

Text Analysis with Word Clouds

The 26,000 Tweets were cleaned and processed for use in several word clouds including a general one in Figure 3.2.

Figure 3.2

Unfiltered Overall Most Prevalent Terms



First, the top 100 words found among all the Tweets were listed out and made into a word cloud, showing that the top words were "medical," "privacy," and "hapu." "Hapu" was found to be "HIPAA" in text form but processed differently due to HIPAA not being recognized in the English Language for Tweet processing. Since these three words were the terms used for creating the dataset, they were removed in Figure 3.3 and throughout the remainder of the analysis to provide more meaningful results regarding the rest of the words present.

Figure 3.3

Filtered Overall Most Prevalent Terms

After filtering, it was clear to see the words that appeared the most regarding HIPAA and medical privacy were "law," the use of the "#hipaa" hashtag, "right," "health," "record," "violate," and "protect." When looking at some of the full Tweet text that contained these words, there are differing opinions of whether medical privacy and HIPAA actually protect patients, whether it is being violated in regard to certain events such as COVID-19, and what the laws actually say about patient medical record protection. Overall, it provided knowledge that people were in fact discussing these topics, but not a lot could be said yet about how people felt about medical privacy in terms of research.

To dive a little deeper into the word usage across different times frames, the word clouds were split by year to determine trends based on real world events. "#Hipaa" was the number one term found in Tweets from 2015 all the way through 2019, but then became much less significant from 2020 to 2022 and seemed to drop off during the start of COVID-19. Figure 3.4 highlights the earliest terms that were prevalent in 2015 to show later how much they changed over time.

Figure 3.4

Earliest Most Prevalent Terms from 2015

Law was prevalent in the first couple of years but became the most common word in 2021 when many laws were being discussed about the pandemic and other controversial health care topics. Something interesting to note was that in 2015, the discussion seemed to be more heavily around violation and invasion of privacy and the rights people had to their privacy, and 2022 focused heavily on laws and protection of health records, while most other years focused on general patient record discussion. When the pandemic began in 2020, there was more discussion of respect, expectations, and wanting to talk about what was going on. In 2020, a user became the most common word in the text content of these Tweets, "@kurteichenwald" run by Kurt Eichenwald, a writer that has been extremely active in health care matters especially during COVID-19, and many people were communicating with him through Twitter about the events.

In 2022, the most prevalent words were very different than the collection of words from previous years, and were concerned more about rights, violations, laws, and data leaks as shown in Figure 3.5.

Figure 3.5

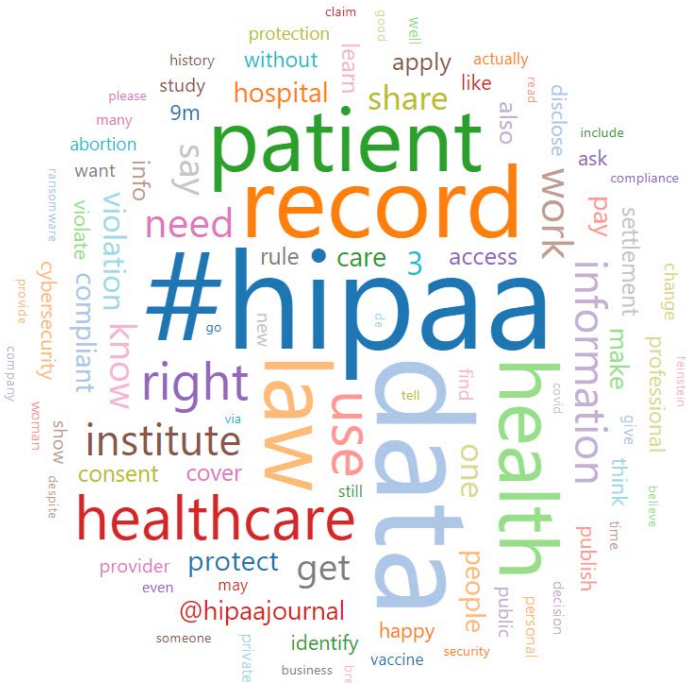
Latest Most Prevalent Terms from 2022

This change makes sense considering the number of controversial health care topics in laws and privacy being discussed. That being said, there were still no particularly telling details about research, studies, or clinical trials and its relationship to data access; nothing had appeared in word clouds or listed in any other aggregate analysis. This could mean that even though they have a lot to say about medical privacy and HIPAA, as expected by the small amount of peer reviewed research on this topic, not a lot of people have been talking about its connection to medical research and improvements in patient care, at least on Twitter. The remaining word clouds for years within the range of 2016-2021 can be found in Appendix D, Figures D.1-D.6.

When a specific subset of data was used that contained anything regarding research, there were less than 1,000 Tweets out of over 26,000 from the main dataset. The key words removed in previous analysis were still removed in this analysis in Figure 3.6 to avoid cluttered results.

Figure 3.6

Most Prevalent Terms in Research-Related Tweets



A word cloud about the most common words used in research-based Tweets was the best way to visualize them. In addition to the hashtag "#hipaa," some of the most common words included "patient," "record," "data," "healthcare," "law," "health," "right," "use," "need," "information," "share," and "access." A portion of these words have been most prevalent in all Tweets about medical privacy and HIPAA, but some specific to the research-geared Tweets included "need," "use," "right," "share," and "access." This is the kind of information that was being investigated, and it seems as though the people who are discussing research might be mentioning it in reference to access and the need to use or share health care data. This also means it could be in reference to people who do not believe access should be given for medical research, but a more in-depth analysis would need to be done with a significantly greater quantity of Tweets to be certain. In a cursory look at the text fields themselves, it was clear that though there was a low number of people who were discussing research and privacy / access, they had very strong and differing opinions on it. Some brought up that HIPAA laws allow for research and that they don't believe these records should be accessed for this research, while most others brought up that HIPAA has created unnecessary roadblocks to improving medical treatments and improving patient care / quality of life. This is where the

sentiment analysis comes in to get a general idea of how people feel about HIPAA and medical privacy.

Sentiment Analysis and Views on Research

Back in the main dataset with all medical privacy and HIPAA related Tweets, an initial pie chart in Figure 3.7 shows that across all years there is slightly more emphasis on negative Tweets than positive Tweets, making up about 40% and 34% respectively, with neutral Tweets taking up just over a quarter of Tweets. This was also broken down by year in Figure 3.8 in a similar way that the word clouds, to see if specific time frames had more frequent words or strong opinions.

Figure 3.7

Overall Sentiment Ratio

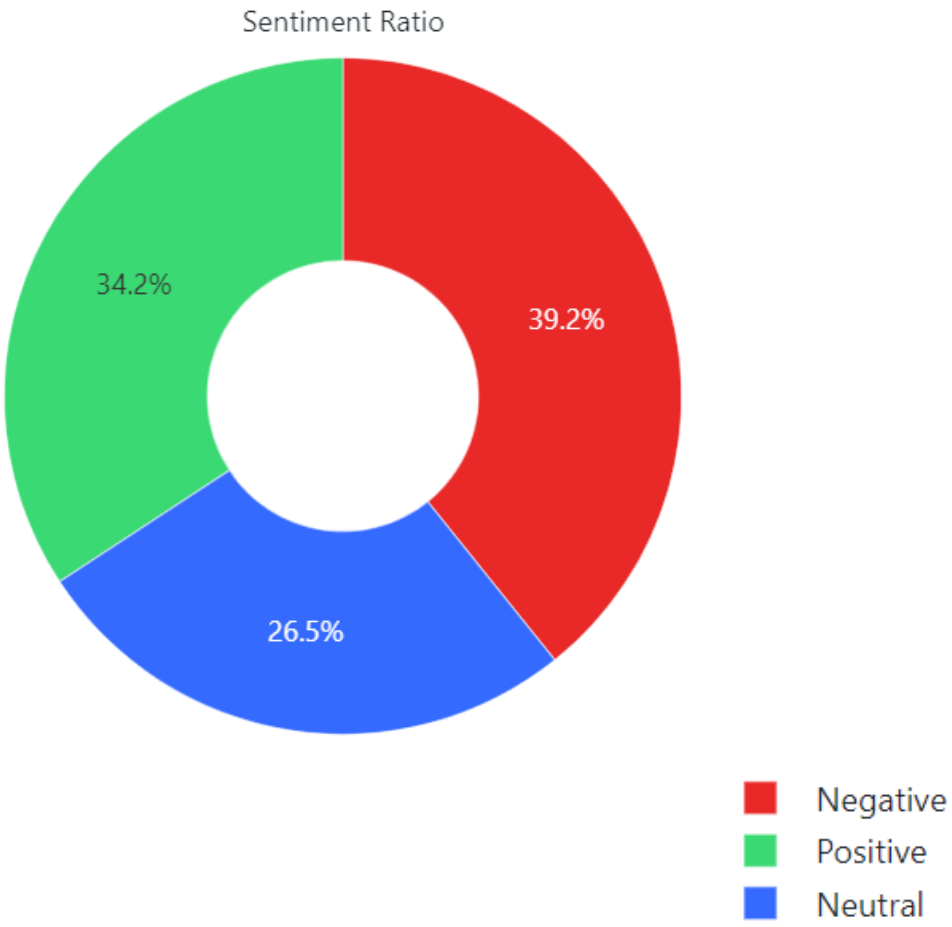
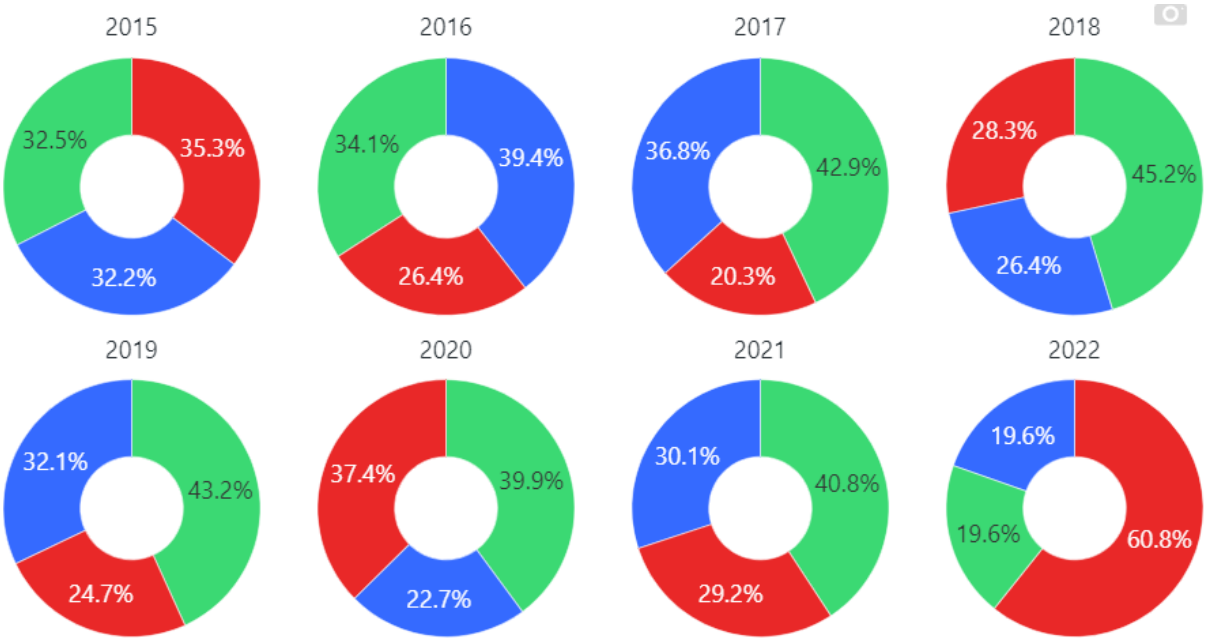


Figure 3.8

Sentiment Ratio by Year



Overall, from 2015 to 2019, the ratios were relatively split in thirds give or take a few points, with positive Tweets becoming more prominent from 2017-2019. 2020 showed significantly less neutral Tweets as people became polarized by the pandemic, and there were more negative Tweets than were taken away from the neutral Tweet percentage, showing a significant decrease in positive opinions towards medical privacy and HIPAA. The ratios returned to a similar level in 2021 as they were in 2019, but then in 2022, the largest change occurred with over 60% of the Tweets in 2022 being negative regarding medical privacy and HIPAA laws, with an even split between positive and neutral Tweets for the remaining portion. Several controversial health care topics were prevalent in the news and media at this time which could explain the dramatic change. Several other related word clouds describing this relationship are in Appendix D, Figures D.7-D.9.

This lack of relationship to research meant that diving deeper into the research specific Tweets and the common words found in these Tweets broken up by positive, negative, and neutral was crucial to finding any type of relationship between medical privacy and research. A general count showed that research had more positive Tweets than negative Tweets, with the smallest portion being neutral Tweets. A word cloud for each sentiment was then created

with proper terms filtered out to avoid cluttered results, The positive word cloud shown in Figure 3.9 has a heavy emphasis on "data," followed by "patient," "health," "law," "record", "information," "share," and "need."

Figure 3.9

Most Prevalent Terms Surrounding Positive Research Tweets

These terms could be interpreted as more associated with those who see research as a benefit to the medical care of patients and the health care industry as a whole and believe that it should be shared to improve patient health. Some of these people may also mention how current laws prevent records from being utilized for research, but these are the positive sentiment Tweets in this category, so this may be less prevalent. For the neutral word cloud shown in Figure 3.10, "data" was once again the top word with other top words including "law," "health," "use," "healthcare," "@hipaajournal," "record," and "one."

Figure 3.10

Most Prevalent Terms Surrounding Neutral Research Tweets

While some of these are similar to the above words, there are some neutral words included, and maybe these people who have neutral Tweets aren't really discussing this topic of debate. In terms of the negative Tweets in Figure 3.11, the top word was "record" followed by "patient," "institute," "law," "3," "data," "pay," "violation," and "settlement."

Figure 3.11

Most Prevalent Terms Surrounding Negative Research Tweets

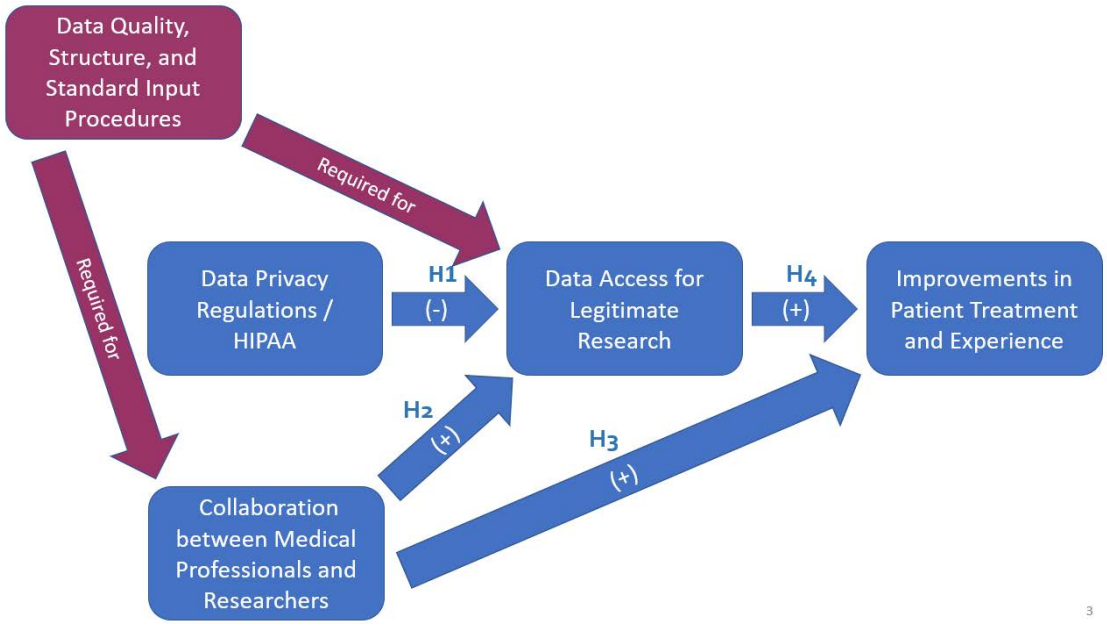
These few more negative words could be contributing to it being Tweets with a negative sentiment, but the fact that they are still being discussed at least a little bit in health care privacy / HIPAA means that people have strong opinions of HIPAA. Again, this could mean people disagree with how HIPAA and privacy affect research because it could hinder important advancements, but it could also be about people not liking that HIPAA could be violated in problematic ways with the increase in research using patient medical records. Further research would need to be conducted to establish a stronger relationship in the view of the public on Twitter, but this at least shows there are strong opinions and people discussing this relationship, whether they believe it needs a change or is going in a problematic direction.

CONCLUSIONS AND RECOMMENDATIONS

The proper balance for the relationship between HIPAA / medical privacy and data access for medical research is extremely important to find, but also a difficult one with people having very strong opinions about how an improvement in its relationship should be executed if at all. Since they are so different and not fully communicated, this brings the problem that there isn't currently much being done about it yet. It was concluded that a viable way to go about this type of massive industry change would be to have a well-regulated structured process to create a symbiotic relationship between medical professionals and researchers that primarily improves patient care and outcomes. That being said, many participants displayed concern with certain data-specific challenges, and they contributed equally if not more than data access due to data privacy in terms of hindering advancements in medical research. Without structured data and interoperability, there is no data to be shared. Without data in an analyzable state, nothing can be used in the first place; this is what needs to be fixed first. Overall, the four hypotheses were proven correct, with an additional finding holding true for H1 regarding the data quality and structure challenges being more prevalent than the privacy challenges, though the privacy challenges do still exist. These relationships are displayed in Figure 4.1.

Figure 4.1

Research Framework Revisited



A potential solution to both being contributing obstacles would be the implementation of a process at initial point of data collection that was more structured, interoperable with other related health care facilities, and set up with the 18 HIPAA personal identifiers flagged in databases / data repositories where they can be removed for purposes of research and assigned a unique, unrelated identifier. This implementation process would come with the need to increase communication with anyone affected by the new systems or processes to inform them and increase their trust in the new system. Many survey participants at varying levels of knowledge of this topic and participation in these kinds of issues in their career agreed that this process would bring a variety of benefits to both patients and providers. Most also agreed that the partnership side would lead to access, and that access would lead to improvements in medical research, which shows the viability of this recommendation as well as its potential to be effective in improving patient outcomes both physically and mentally / emotionally. These effects could span as far as mitigating issues or helping patients manage life after medical care if they develop any form of medical PTSD.

In addition to the outlined recommendations, aggregating all the challenges and concerns listed by survey participants shows that there may be an industry wide solution regarding cross-functional teams. The plan would include making a strong team of experts in their field to get together and spend a significant amount of time explaining perspectives, bringing up concerns, discussing obstacles and goals for their given careers or departments, and working through challenges all together to find a solution that benefits all parties involved either directly or indirectly. This team would include health care professionals, administrators, patient-facing careers like doctors and nurses, researchers with medical knowledge, analysts and data scientists, legal professionals / privacy law lawyers, patient advocates to have their perspective, and anyone behind the scenes related to these fields of study that may have a contributing perspective on the matter. Many of these challenges are industry wide and even cross-industry related, and trying to overcome them in an organized and feasible manner is far too daunting to be done by one group, so communication with others could improve these relationships. This would allow teams to begin the necessary steps to find and successfully implement a solution to the two main reasons of data privacy and data quality / structure for the separation between data access and medical research.

LIMITATIONS AND FUTURE RESEARCH

Due to the nature of this research, there were some limitations encountered throughout the process that may have affected the results and conclusions elaborated on above. First, there were not a significant number of informational sources or studies found that were already discussing this specific type of relationship, which was the main reason why other proposed research was unable to be completed. The few studies found regarding this topic were used, but the remaining information contained came more from research on specific one-dimensional portions of these overall hypotheses. Second, in terms of the primary source of information, only 41 usable surveys were used to conduct analysis, and having at least 100 would have led to stronger results and clearer delineations between opinions. None of these participants were patients or patient advocates partially due to the presence of these data access issues, so that side of the argument may not have been represented as well. Third, there were not a lot of Tweets collected that discussed medical privacy and research, which in and of itself is a data point that it is not discussed by as wide a range of people as previously predicted when these topics are most likely being discussed more individually unrelated to one another. Twitter is also public, and it was difficult to find out the background of the users in the dataset to find out where their opinions were stemming from, unlike in the survey data where there was a lot more information about who the person was and why they felt the way they did. A side note, the spelling of HIPAA is typically incorrect and written as HIPPA, so any Tweets about HIPAA that were spelt incorrectly were not included in my dataset and could be something to note for the future. That being said, there still may not have been enough data points in the Twitter analysis to come to substantial population wide conclusions.

Future research can take this idea and produce a more in-depth study with more questions about the data structure obstacle instead of mainly about data privacy and HIPAA and could even include new obstacles or general views about new solution proposals as they are created and tested. Opening this data collection to a much wider population and gaining significantly more participants responding to these issues over time would help gain more perspective and a more comprehensive idea of what is being discussed or proposed as a solution to these complex issues. Doing a more extensive Twitter analysis with more Tweets could also lead to strong results and more patterns that could also be investigated in future research. A second

main research avenue that would be necessary for making substantial progress in this theory would be to look more into the privacy relationship with research in comparison to the data structure and quality relationship with research. This was a big part of the survey results and produced more questions about which obstacles are larger and how many others are out there that may not have been previously identified.

For future research down the line if these discussed obstacles are overcome and research can occur for medical advancements and an increase in understanding of patient outcome trends, it is suggested to investigate predicting risk factors that put someone at risk for developing medical PTSD post hospital care. This entire thesis became about the lack of data access for research from privacy and structure issues after no data for the original proposed thesis was available from any direct or indirect sources for unconfirmed reasons, suspicioned to be due to privacy or structure concerns. The original proposed thesis was about predicting medical PTSD in surgery patients, and working to understand what risk factors impact a person's chance of having medical trauma from medical care and experiencing lifelong negative impacts from it. It would also have led to potential solutions to implement in order to prevent this from happening in the first place or manage it effectively if it develops, improving patient care both physically and mentally / emotionally as well as patient long term quality of life post medical treatment. Since no data was available, the thesis evolved into investigating the reason why it could not be successfully completed. Having this highly important medical PTSD research, which was even less discussed than the issues described in this thesis, be completed after potential solutions from this thesis are successful would be an incredible use of resources to help improve patient quality of life. It could even develop into more research and overall improvement in health care and surrounding industries for the benefit of all involved, especially those treated in medical facilities who deserve more positive treatment outcomes than what they may have experienced.

REFERENCES

Emam, K. E., Rodgers, S., & Malin, B. (2015). Anonymising and sharing individual patient data. *BMJ*, *350*, h1139. https://doi.org/10.1136/bmj.h1139

Gostin, L. O., & Nass, S. (2009). Reforming the HIPAA Privacy Rule: Safeguarding Privacy and Promoting Research. *JAMA*, *301*(13), 1373–1375. https://doi.org/10.1001/jama.2009.424

Hand, D. J. (2018). Aspects of Data Ethics in a Changing World: Where Are We Now? *Big Data*, *6*(3), 176–190. https://doi.org/10.1089/big.2018.0083

Horner, J. (1998). Research, ethics and privacy: The limits of knowledge. *Public Health*, *112*(4), 217–220. https://doi.org/10.1016/S0033-3506(98)00235-2

Kass, N. E., Natowicz, M. R., Hull, S. C., Faden, R. R., Plantinga, L., Gostin, L. O., & Slutsman, J. (2003). The Use of Medical Records in Research: What Do Patients Want? *The Journal of Law, Medicine & Ethics : A Journal of the American Society of Law, Medicine & Ethics*, *31*(3), 429–433.

Medicine, I. of, Services, B. on H. C., Policy, B. on H. S., & Rule, C. on H. R. and the P. of H. I. T. H. P. (2009). *Beyond the HIPAA Privacy Rule: Enhancing Privacy, Improving Health Through Research*. National Academies Press.

Meslin, E. M. (2006). Shifting Paradigms in Health Services Research Ethics. *Journal of General Internal Medicine*, *21*(3), 279–280. https://doi.org/10.1111/j.1525-1497.2006.00391.x

Mittelstadt, B. D., & Floridi, L. (2016). The Ethics of Big Data: Current and Foreseeable Issues in Biomedical Contexts. In B. D. Mittelstadt & L. Floridi (Eds.), *The Ethics of Biomedical Big Data* (pp. 445–480). Springer International Publishing. https://doi.org/10.1007/978-3-319-33525-4_19

Ness, R. B., & Joint Policy Committee, S. of E., for the. (2007). Influence of the HIPAA Privacy Rule on Health Research. *JAMA*, *298*(18), 2164–2170. https://doi.org/10.1001/jama.298.18.2164

O'Herrin, J. K., Fost, N., & Kudsk, K. A. (2004). Health Insurance Portability Accountability Act (HIPAA) Regulations. *Annals of Surgery*, *239*(6), 772–778. https://doi.org/10.1097/01.sla.0000128307.98274.dc

Scott, C. (2000). Is Too Much Privacy Bad for Your Health—An Introduction to the Law, Ethics, and HIPAA Rule on Medical Privacy. *Georgia State University Law Review*, *17*(2), 481–530.

Services, D. of H. C. (2022). *List of HIPAA Identifiers*. Retrieved November 1, 2022, from https://www.dhcs.ca.gov/dataandstats/data/Pages/ListofHIPAAIdentifiers.aspx

APPENDICES

Appendix A: 18 HIPAA Personal Identifiers

1. "Names	2. All geographic subdivisions smaller than a State, including street address, city, county, precinct, zip code, and their equivalent geocodes, except for the initial three digits of a zip code if, according to the current publicly available data from the Bureau of the Census: a) The geographic unit formed by combining all zip codes with the same three initial digits contains more than 20,000 people; and b) The initial three digits of a zip code for all such geographic units containing 20,000 or fewer people is changed to 000.	3. All elements of dates (except year) for dates directly related to an individual, including birth date, admission date, discharge date, date of death; and all ages over 89 and all elements of dates (including year) indicative of such age, except that such ages and elements may be aggregated into a single category of age 90 or older;
4. Telephone numbers	5. Fax numbers	6. Electronic mail addresses
7. Social security numbers	8. Medical record numbers	9. Health plan beneficiary numbers
10. Account numbers	11. Certificate/license numbers	12. Vehicle identifiers and serial numbers, including license plate numbers
13. Device identifiers and serial numbers	14. Web Universal Resource Locators (URLs)	15. Internet Protocol (IP) address numbers
16. Biometric identifiers, including finger and voice prints	17. Full face photographic images and any comparable images	18. Any other unique identifying number, characteristic, or code, except as permitted by paragraph"

(3 Department of Health Care Services, 2022)

Appendix B: The Full Distributed Survey

Health Care Data Access for Analytical Research

INFORMED CONSENT

You are invited to participate in a research study survey about data access and data privacy in the health care industry. This would provide potential solutions for health care professionals and researchers to improve data access and structure for health care advancements. This survey will ask about your career background, your opinions on various data topics, and your recommendations for ways to make advancements. This study is being conducted by Katherine D'Ordine, advised by Dr. Suhong Li, for an undergraduate data science thesis at Bryant University in Smithfield, RI. Participation in this study is voluntary. This survey is anonymous and takes less than 5 minutes to complete. All responses will be aggregated to help us understand more about health care data from a variety of perspectives. You may skip any questions that you do not feel comfortable answering, and you may end the survey at any time. If you have any questions about this survey, please contact Katherine D'Ordine at kdordine@bryant.edu. Thank you for your time and participation!

Age

- 21-30
- 31-40
- 41-50
- 51-60
- > 60

Gender

- Female
- Male
- Transgender
- Non-binary/non-conforming
- Prefer not to answer
- Other

What positions have you had regarding research, analytics, or health care? Choose all that apply.

- o Data analyst/data scientist in health care

- o Data analyst/data scientist in a field other than health care
- o Hospital Administrator
- o Insurance Administrator
- o Medical Practice Administrator
- o Nurse
- o Pharmacist
- o Physician
- o Physician's Assistant / Nurse Practitioner
- o Psychologist or Psychiatrist
- o Student in research, analytics, or health care
- o Other: _____

How familiar are you with the patient medical record collection process and how the data is structured?

Not at all familiar 1 2 3 4 5 Very Familiar

What stages in the data process have you participated in? Choose all that apply.

- o Before data processing as a medical professional: generating or recording data in a hospital or medical office directly with patients
- o Data processing / data engineering
- o Model building / refinement
- o Data interpretation / visualization / presentation to stakeholders
- o Receipt and implementation of new recommendations as a stakeholder
- o Other: _____

Challenges: Do you feel that the following are challenges in your work place regarding initial data collection processes?

1: Strongly Disagree 2: Disagree 3: Neutral 4: Agree 5: Strongly Agree

Regulations/Privacy/HIPAA
Knowledge of a complex data system
Data structure and organization
Standard procedures for inputting data
Data breach, data being exposed, or losing certain information in a new system
Data quality (completeness, duplicates, etc.)
Other, please describe. Your answer

Benefits: Do you feel that the following would be benefits for your work or for patient experience if there was a well-regulated process where research students could partner with medical facilities or organizations to utilize their data for various research projects?

1: Strongly Disagree 2: Disagree 3: Neutral 4: Agree 5: Strongly Agree

Patient treatment options
Patient treatment experience
Provider experience and their ability to treat patients
Interaction and collaboration between health care professionals and academic researchers
Other, please describe. Your answer

Do you believe having a well-regulated partnership process between researchers and analysts would drive better access to data?

Strongly Disagree 1 2 3 4 5 Strongly Agree

Do you believe better access to health care data for analytical research would drive improvements in medical and mental health treatments?

Strongly Disagree 1 2 3 4 5 Strongly Agree

Please provide any additional recommendations for overcoming the challenges in accessing health care data for analytical research.

Please provide any concerns you may have about increasing health care data access.

Is there anything else you would like to add to this survey?

Appendix C: Demographics and Supplementary Visuals from Survey Analysis

Figure C.1

Age Distribution

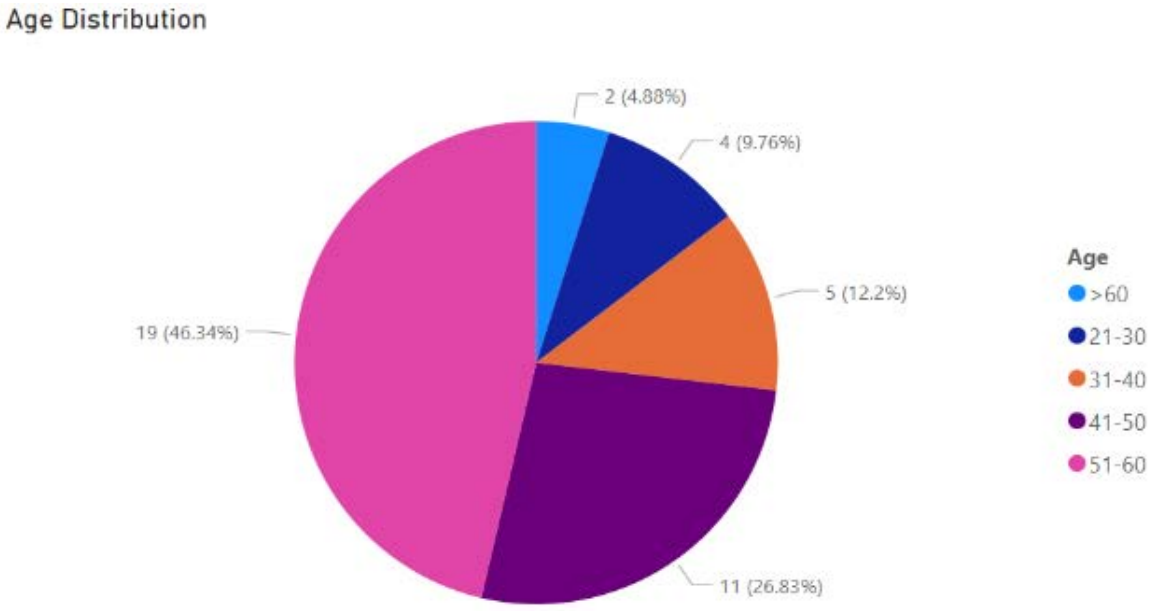


Figure C.2

Gender Demographic

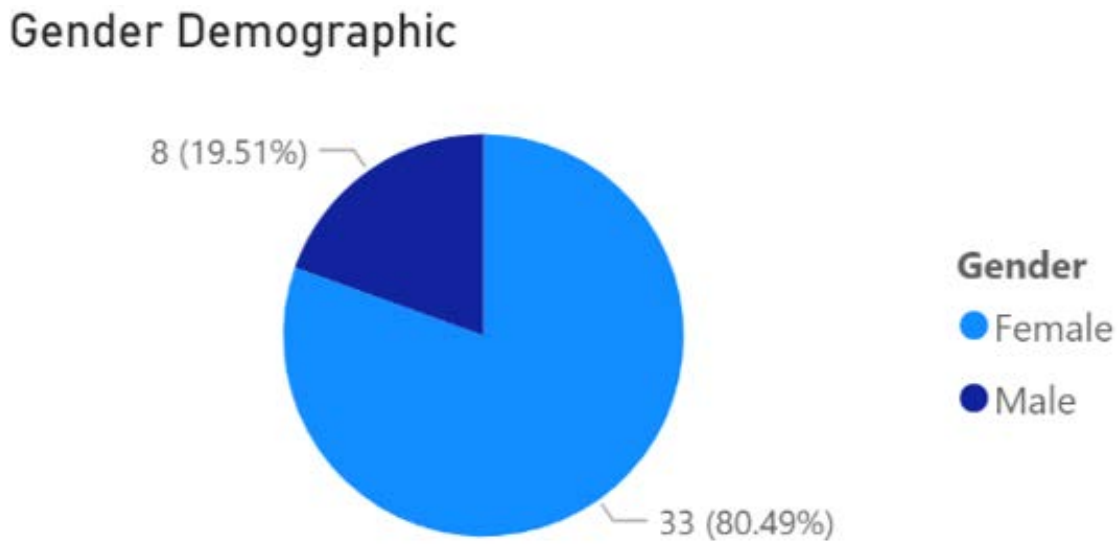


Figure C.3

Career Prevalence

Career	Participants
Hospital Administrator	8
Student in research / analytics / health care	7
Nurse	6
Data analyst/data scientist in a field other than health care	4
Data analyst/data scientist in health care	4
Pharmacist	4
Physician	4
Physician's Assistant / Nurse Practitioner	4
Medical Practice Administrator	3
Cancer Registrar	1
CEO of private healthcare company (Optometry)	1
Certified Occupational Therapist	1
Discharge planner	1
Education	1
Insurance Administrator	1
Occupational Therapist	1
Physical Therapist	1
Product Management in healthcare	1
Psychologist or Psychiatrist	1
School Nurse	1
Social worker	1

Figure C.4

Level of Familiarity with Patient Medical Records

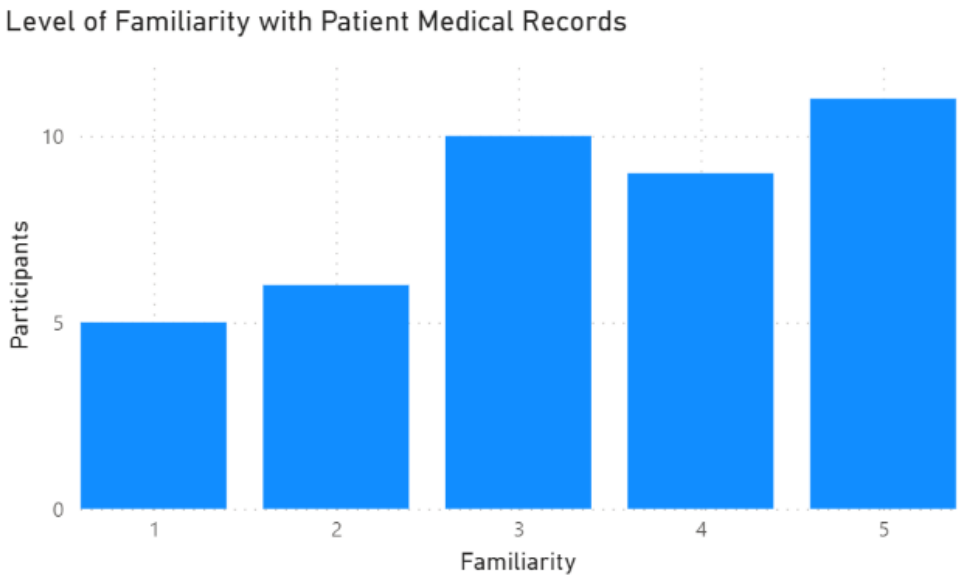
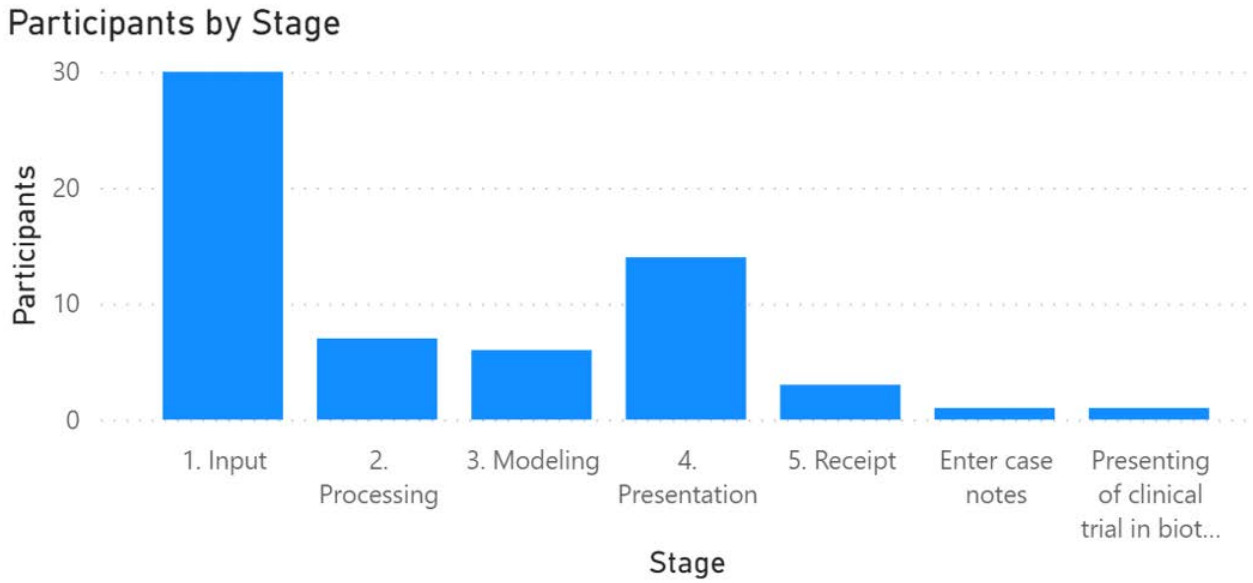


Figure C.5

Stages of Participation in the Data Process



Appendix D: Supplementary Visuals from Secondary Twitter Analysis

Figure D.1

Most Prevalent Terms from 2016

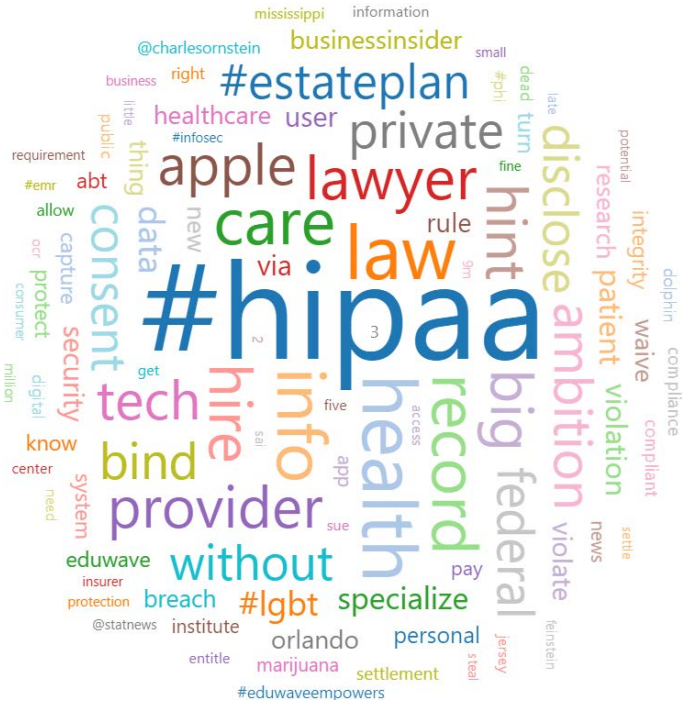


Figure D.2

Most Prevalent Terms from 2017

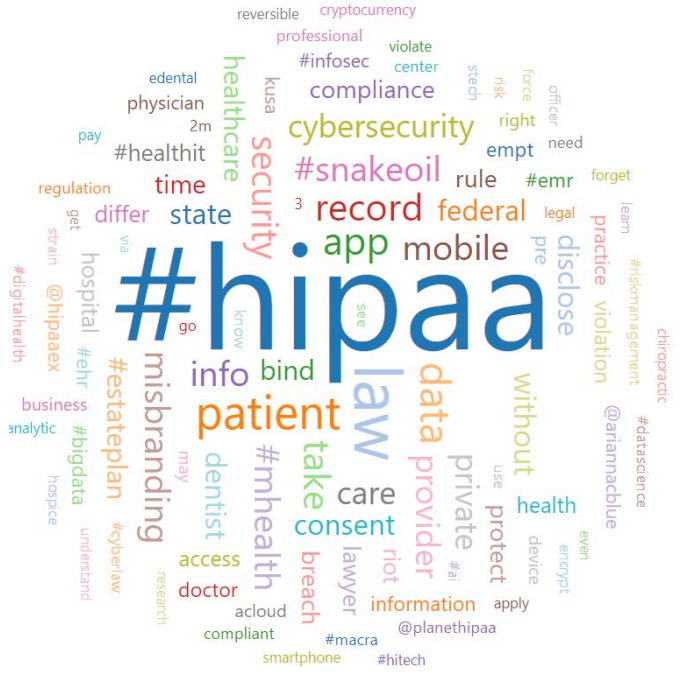


Figure D.3

Most Prevalent Terms from 2018

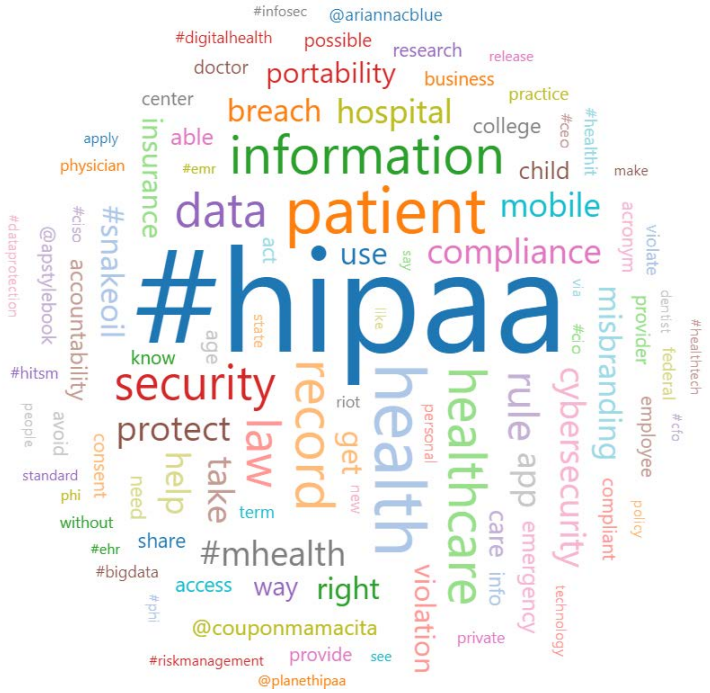


Figure D.4

Most Prevalent Terms from 2019

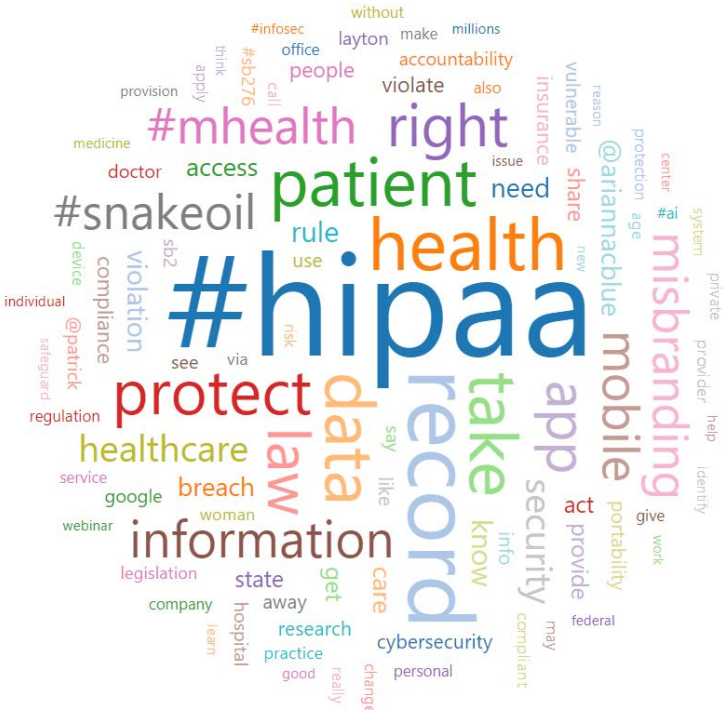


Figure D.5

Most Prevalent Terms from 2020

Figure D.6

Most Prevalent Terms from 2021

Figure D.7

Most Prevalent Terms for All Positive Tweets

Figure D.8

Most Prevalent Terms for All Neutral Tweets

Figure D.9

Most Prevalent Terms for All Negative Tweets