THE SCOPE AND VALUE OF HEALTHCARE DATA SCIENCE APPLICATIONS

Jose Oscar Huerta

Dissertation Prepared for the Degree of

DOCTOR OF PHILOSOPHY

UNIVERSITY OF NORTH TEXAS

May 2021

APPROVED:

Victor Prybutok, Committee Chair and Dean of the Toulouse Graduate School
Gayle Prybutok, Committee Member
Brian O'Connor, Committee Member
Jiangping Chen, Chair of the Department of Information Science
Kinshuk, Dean of College of Information

Huerta, Jose Oscar. *The Scope and Value of Healthcare Data Science Applications*. Doctor of Philosophy (Information Science), May 2021, 122 pp., 23 tables, 12 figures, references, 126 titles.

Health disparities are a recognized public health concern and the need to address these disparities remains worthy of bringing new methods that assist in closing the gap. This research examined the effectiveness of data science to highlight health disparities, and to convey the value of data science applications in related health care applications. The goal of this research was accomplished by undertaking a multi-phased and multi-method approach, best represented in three individual essays. In essay one, a systematic literature review assessed the state in current academic literature of data science applications used to explore health disparities and to determine its applicability. The systematic review was guided by the Preferred Reporting Items for Systematic Review and Meta-Analyses (PRISMA) guidelines. Essay two assessed the capacity of data science software to address the effectiveness of these data science technologies in examining health disparities data. This was conducted using KDnuggets data pertaining to analytics, data science, and machine-learning software. The research in this essay demonstrated the potential utility of leading software to perform the kinds of data science operations that can achieve improved care in healthcare networks by addressing health disparities. Essay 3 provided an appropriate case study to showcase the value data science brings to the healthcare space. This study used a geographic information system to create and analyze choropleth maps to determine the distribution of prostate cancer in Texas. SPSS software was used to assess the social determinants of health that may explain prostate cancer mortality.

ACKNOWLEDGMENTS

Proverbs 3:6 "In all thy ways acknowledge Him, and He shall direct thy paths".

I would like to first acknowledge God for his help and strength during this fulfilling and rewarding endeavor. I would like to express my deepest gratitude to HIM for opening the doors to this opportunity and for aligning me with the right people that would help make this possible.

I would also like to express my heartfelt appreciation to my committee Chair Dr. Victor Prybutok for guiding me in my learning path. His way of reducing the difficult and impossible things into the possible was of great value to me. To my committee members, Dr. Gayle Prybutok and Dr. Brian O'connor, my sincere and heartfelt gratitude for their help and strong support in my efforts toward this goal. I would also like to acknowledge Dr. William Senn for his guidance to me in helping me understand the PRISMA systematic review process.

To my wife, Christina Huerta, I am grateful for her help and support, her patience, and the sacrifices she made to help me accomplish my dream.

To my mother, Maria Pando, the person who first started me on my path toward college, I am blessed to have her and because she believed I could do this.

To my daughters, Alexyss and Skylar Huerta, for their understanding and patience during the times I was gone attending school and for the times I didn't get to fully spend with each of them because of school priorities.

Last but not least, to Dr. Michael Hall, for his Godly advice and encouragement to me.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION

1.1 Background

1.1.1 Origins of Data Science

It is important to understand the origins of data science by exploring the origins of information science, a discipline that is quite similar to data science where efforts are continuously made to extract new insights, and some have determined it to be one of the contributing parents to the discipline of data science. Information science has been around for decades. Although many people, such as those in academia and in business, still question what exactly information science is, the discipline just didn't come to fruition in the past few years. The discipline of information science has been around for quite some time. Information science can be a science in and of itself, meaning that the study of information is not only based and defined by academic terms but in fact, can derive from scientific processes – and these scientific processes work together to bring light to information science as both a science and academic discipline.

The origins of information science could not be possible without the existence of information. Information can be displayed in patterns. These patterns can exist in biological, geological or in unseen forms, such as frequencies that move through the air. Patterns are found all around, and in many cases even in everyday communication between humans or animals (Bates, 2006). Information can have a behavior of its own from the patterns that are identified in the occurred phenomenon; that is to say, the phenomenon of information that is generated and behaves in a way that is worth research and study. The behavior of information

is crucial in the field of information science because information science is the analysis of the properties and actions of information, the powers controlling the distribution of information, and the ways of processing information for optimal accessibility and usability (Borko, 1968).

Information science works within the context of components that serve to build up the body of knowledge – which is a sort of repository where knowledge is stored. These components include, but are not limited to, processes such as collection, dissemination, and the meaningful use of the information produced for the multiple disciplines at work (Borko, 1968). Information science, known as a social science today, began to originate into a generally accepted discipline when theories within the discipline began to emerge. These theories did not emerge for conceptual purposes only, but for the purposes of creating multiple useable applications from the theories - and this type of achievement in the social sciences realized within the discipline of information science, whether recognized or not, was possible because any group practice pretending to be a science must also be realistic (Brookes, 1980).

The discipline of information science was also catapulted forward from other theories that existed around knowledge, or as it is known by some, the world of knowledge. In fact, the phenomenon that occurred around knowledge helped to build a foundation on which information science relied on for its theoretical models and applications – which is to say, that the field of information science needed an objective theory of knowledge at the core of its discipline (Brookes, 1980).

The theory of objective knowledge was important for information science. A significant contribution in this theoretical focus aided in the origination of information science as a discipline, and these components can be found present in Popper's three worlds. Karl Popper, a

lecturer and philosopher, introduced three worlds that helped to delineate how the phenomenon of knowledge can be categorized, and in many ways, perhaps explained. The focus here shall be on the most impactful world for the discipline of information science, which is World 3 of Popper's three worlds. This world can help to explain in ontological form, the world of objective knowledge. In this world, the output from mankind are produced from subjects such as language, art, science, and technology – and eventually can find its way in the form of recorded formats (Brookes, 1980). Think of World 3 as a repository place. A sort of dimension and playground for the information scientists where the information scientist can utilize the collections of records for multiple purposes, which depends on the scope of work or research that is conducted (Brookes, 1980). Furthermore, in regards to the origination of information science, it may be said that the components found in World 3, an important domain from where information science derived and where western thought later was shaped from, were once at the heart and core of thought during the periods when popular philosophers such as Plato, Aristotle, and Euclid were at the pinnacle of their careers (Brookes, 1980). Although much of what has been introduced thus far helped to create some insight on some important milestones that aided in setting the path for making the field of information science a possibility, a true part of the origination of information science came from the documentation movement founded in Europe during the 1890s. It later was a discipline that came to exist by way of transmission from Europe into the United States during the 1930s, where it later evolved due to cost-saving methods that were incorporated to help streamline the storage of items during the war (Lilley & Trice).

Information science continued to grow at a fast pace because of the high impact that

technology had, and continues to have, on the discipline. In essence, this occurred due to the availability of computers and new technologies introduced into the market. Other outside forces that had an impact, and were to some degree associated with a variety of technologies on the discipline, were as follows: forces tied to visionaries that helped to inspire, available funding toward research in information science, and those who dared to be the game changers in earlier years by providing unique ideas that made information science what it is today (Lilley & Trice, 1989). In fact, those earlier years, specifically during the mid-1940s, were to be an evolutionary time for information science. The existence of technologies during the 1940s, such as electric typewriters and xerography copiers made publications easier to produce (Lilley & Trice, 1989). Information science developed as a mature yet still-growing field during the late 1960s and early 1970s. In 1968 the area was shifted from documentation to information science and that name is still applicable today (Lilley & Trice, 1989). These technological availabilities were a part one of four important generations that impacted and gave direction to the field of information science. In addition to the technologies that existed during generation one, the introduction of machines such as Mark I and ENIAC gave way of what was to come in the 2nd generation – the introduction of digital computers during the early 1950s. Soon after the introduction of these digital computers, this second generation generated an explosion in technological growth for a little over a decade. Then during the 1960s and 1970s, the introduction of mainframe computers gave way to a new generation – that is, generation three. However, it was generation four that came in with new technological advancements that changed the world, and more specifically speaking, the introduction of new technical technologies such as the microcomputer in 1971 (Lilley & Trice, 1989). Today, computers have

now become used in schools and knowledge centers. To date, only hardware has been reported, but similar development has also been observed in other media (Lilley & Trice, 1989).

1.1.2 Pertinence of Theoretical Foundation Advancements

Theoretical foundations can be viewed as those components that existed, and were either relied upon or were expanded upon, for the purpose of conjuring up new theoretical frameworks or contributions that led to increased ideas and developments of those theoretical frameworks or models in information science. In this section, theoretical foundations are not discussed as actual theories that existed during phases of theoretical advancements, and although theories are mentioned to some degree, the focus of this section is to focus on those essential components that existed during the important times of theoretical introductions and advancements in information science. In other words, it is important to review how thoughts and existing technologies came to be foundational components to theories in information science.

Prior to World War 2, the field of information science did not officially exist. The closest type of field that existed was the documentation field, and the type of role that existed in that field belonged to that of a documentalist. Most of the significant work that has occurred in regards to the development of theoretical foundations in information science has been post-World War 2. However, this is not to say that the work in the documentation field prior to World War 2 was not significant. In fact, it was significant because information science was derived from this field, which primarily came from Europe (Lilley & Trice, 1989).

Information and technologies that emerged in the 1950s and 1960s helped advance the field of information science (Rayward, 1996). Information science, along with computer science,

originated in the wake of the Second World War. Additionally, this new field of information science could possibly be credited to Cannevar Bush who wrote an article in 1945 called “As We May Think” (Saracevic, 2010). Over the course of time, new theories were introduced.

A mathematical theory of communication by C.E. Shannon became a main component to the information theory that was popular in the early years of the field. In fact, it was the mathematical theory of communication that helped delineate processes that allowed data conversion to a binary format. After information theory was expanded on, the new theory along with the foundational work laid by the introduction of the mathematical theory of communication, helped to research the new world of digital data and information. Additionally, information theory also helped the field of information science launch new theoretical works in its discipline (Ma, 2012). Communication essentially became an important part of information science and it was a focus of study after World War 2. In addition to the development of information theory during this era, other models such as the Data-Information-Knowledge-Wisdom (DIKW) model also held important components that aided in the advancement of information science and other theory developments (Ma, 2012).

The structure of information science was crucial to information science theoretical foundations. There were two main time periods, one from 1972 to 1995 and the other from 1996 to 2006, which were important to these foundations. During 1972 to 1995, the following important milestones and concepts are to be noted: experimental retrieval, citation analysis, practical retrieval, bibliometrics, general library systems, science communication, user theory, online public access catalogs known as OPACs, imported ideas such as information theory and cognitive science, indexing theory, citation theory, and communication theory. During 1996 to

2006 the following were also to be noted as important: user studies, citation analysis, experimental retrieval, webometrics, visualization of knowledge domains, science communication, users' judgment of relevance, information seeking and context, children's information searching behavior, metadata and digital resources, bibliometric models and distributions, and structured abstracts (Saracevic, 2010).

The advancement of new technologies became an important foundation for the theoretical foundations that were present in the development of theories for the past several decades. Eventually, because of these technological advancements, there was an information explosion, which meant that information was produced at a rapid rate and became highly available through new mediums, such as the internet and available applications that interacted with the internet (Saracevic, 2010). In many cases, these applications have been linked to social media. Social media, through the popularization of the internet, has now become one focus of study for the information scientists. Another important component to the theoretical foundation of information science is relevance, especially when information retrieval is considered. In fact, "information retrieval (IR), a major branch of the information science, is about retrieval of relevant information. Thus, the notion of relevance is fundamental to the information sciences" (Sonnenwald & Saracevic, 2016).

1.1.3 Interdisciplinary Nature of Information Science and Multi-Disciplinary Contributions

Integration of ideas and processes that have been brought together for the purposes of creating a specific outcome has been a very popular concept throughout history. These types of integrations can be found in any number of systems that rely on one another in order to work efficiently and to hold a particular value on the purposes it was intended for – and some of

these types of integrations can be found among sewer systems, airport systems, communication systems, and so forth. In regards to the integrations of these systems, many of them relied, and still rely, on information from different disciplinary perspectives to achieve, as mentioned earlier, a common or greater purpose – usually for the purposes of helping mankind. In fact, it is not only the disciplines that have to be integrated, each having its own set of contributions to make, but it is the science of the information that is shared between those disciplines that have to be studied and relied upon to make the contributions work. In many cases, the study of the information contributed by all of these disciplines can be considered segments of the interdisciplinary science discipline at work. One can think of the United States military and determine that if information was not shared among the different branches of the military, there could soon be chaos. It then behooves the military to quickly understand that the information shared between military divisions is essential to military advantages on the battlefield. Information science is not only conducive to creating advantages for the military, but it is conducive to creating advantages to any discipline – and in many cases, the advantage is not created by the information that is static to one discipline, but information that typically is shared from other disciplines, to create an interdisciplinary effect, that then fosters advantages of knowledge and wisdom to the multiple disciplines in focus.

Information science is interdisciplinary in nature because it reaches across disciplinary aisles, regardless of the forces and the behavior of information that takes place under any discipline or situation. It also is involved with a multitude of approaches and techniques that process information for the work and scope of the project that it seeks to research, resolve, investigate, and explain (Borko, 1968). However, the topic of information science, and the fact

that it is considered interdisciplinary in nature, does not go without issues. This is typically due to the fact that information in many ways does not only work from one discipline to the other, but in fact, in many cases, information science works through the unification and integration of data, information, or knowledge (Besselaar, 2001). It is a complex task to try and determine what interdisciplinary is, in regards to its' meaning, within the discipline of information science. This is mainly due to the fact that there are many variations found in the prefixes tied to the concept of disciplinarity; prefixes such as multi, cross, pluri, inter, and trans (Besselaar, 2001). There are multiple forms in which information science can manifest itself among disciplines. In this section's discussion, the focus is on three types of manifestations that can surface in the so-called "disciplinarities" of information science. Although these may be conceptual, in many cases, an information scientist can rely on these concepts to build frameworks and models that are specific and unique to their own and other disciplines. These manifestations are multidisciplinary, interdisciplinary, and transdisciplinary.

A multidisciplinary perspective of information science is valuable in that the key to discerning, evaluating, and investigating key components of subjects or phenomena in information science and the subject matter at hand, can be found through the lenses of perspectives that come from evaluating a variety of disciplines. In the case of information science and a multidisciplinary approach, a key component of such an approach is to work through different frameworks or models that may enhance the value of the segments that are studied within the discipline of information science (Besselaar, 2001). Additionally, information science can benefit from multidisciplinary diverse perceptions. Simply put, with a multidisciplinary approach, the goal is to evaluate the nearness of how multiple disciplines

relate to each other with the context evaluated in the data and information that has been produced for the information scientist to explore (Holland, 2008). An interdisciplinary methodology can have its own analytical, philosophical, and empirical identification. As a consequence, findings from interdisciplinary research are more consistent and integrated (Besselaar, 2001).

What separates interdisciplinarity from multidisiplinarity, is that interdisciplinarity utilizes all key components found across the different disciplines in an effort to bring them together to help synthesize the explored content and to help resolve an issue at hand. Furthermore, the integration does not only happen with the available content explored, but also in the methodologies from other disciplines that are pertinent to the study, project, or problem; that is to say, it is an integration of the respective methods highly known and utilized in those disciplines. Therefore, it is important to note that interdisciplinarity can be an important variable when conveying the findings of any phenomenon that must be explained in a manner that is clear and meaningful to the stakeholders (Holland, 2008).

Information science, in regards to transdisciplinarity, can have an impact through the commonalities that are shared among disciplines - whether they are commonalities in content, methods, or techniques, these types simply permeate into other disciplines. Transdisciplinarity may have some similarities to interdisciplinarity, but it does not integrate content in the way that the latter does because the latter seeks to integrate the content by a technique or approach used for the synthesis of the content. Transdisciplinarities in information science occurs when commonalities are shared with other disciplines - meaning that it shares the permeation of theories from other disciplines that may be present across the disciplines. For

example, information science may share some theories with data science, but both information science and data science may be two distinct fields in its own respect. In other words, methods, techniques, theories, approaches, may have a relation not only with one discipline, but rather with other disciplines (Besselaar, 2001).

Disciplines in robotics, artificial intelligence, and neural networks have begun to have an impact on the formation and evolution of information science. These fields, once possibly multidisciplinary in nature, have begun to integrate with other disciplines, making them interdisciplinary. This means that these fields themselves will begin to integrate data and information from a variety of disciplines and therefore, have the capacity to revolutionize information science from these new emerging fields (Besselaar, 2001). Disciplines that have also contributed to the formation and evolution of information science are mathematics, library or library science, biology, computer science, and physics, with perhaps the most two powerful contributors, arguably speaking, is that of library science and computer science.

Library science was considered an information science and was highly impacted by the introduction of high performing information technologies. Additionally, information science was also its own discipline and was highly impacted by high performing information technologies (Chang & Huang, 2011). Library information science (LIS) has influences of its own in the discipline of information science (IS). Library information science has made a strong contribution to information science mainly because of its capacity to build relationships with other disciplines, mainly those disciplines that are significant in the discipline of LIS, but also contribute to the formation of information science itself because of their respected contributions. These disciplines include, but are not limited to, the general sciences, business,

sociology, and education. All in all, several distinct fields have been part of information science, including library science, information processing, communications, sociology, computer science, and AI, and these disciplines have contributed into information science the methods, techniques, approaches, data, and information that is beneficial to information science because of the achieved respect each discipline holds (Holland, 2008).

1.1.4 Historical Emphasis of Technology and Social Science Positioning of Discipline

The evolution of technology is very much noticeable in today's existing technologies. These technologies are utilized across cultures and societies throughout the world. However, these capacities evolved over time and did not just come to be without one technology building on another one – and it is these technologies that have aided, in more ways than one, the discipline of information science. It is important to note that these technologies have helped to position the discipline of information science as a social science. The information scientist has been given new capacities of study in information with the evolution of these technologies in the social realm. In essence this is because all information, whether structured or unstructured, is at the core of the information science discipline – meaning that essentially, information scientists take on the role of a social scientist due to the nature of the information science discipline, and this takes place in an effort to help the information scientist lay foundational works that are beneficial in presenting theoretical frameworks in the discipline (Brookes, 1980). In an effort to elaborate on the previous remark, information science began to gain greater traction as a social science as technologies allowed the development of new possibilities to emerge. Information science, in many ways, took on a new term that could help to expand the definition and understanding of information science - and that term is informatics.

Informatics is occupied with the research of things that are present, approaches and techniques involved with the phenomenon of information, and the technologies that help generate the information. Essentially, as technologies became more advanced, it also gave users the capacity to facilitate information in a more social manner. This new capacity gave way to the field of social informatics, which in many ways has been incorporated into the field of information science. Additionally, this capacity has increased more in value as new corporations such as Facebook and LinkedIn study acquired data and information for the purposes of enhancing their image and return on investments (Rosenbaum, 2010).

Information and communication technologies, commonly known as ICT's, have helped to shape the way for information science as a social science. ICT's contribute to the way that people use information. This contribution allows for more information to be produced and it increases the possibilities for new theoretical developments within information science - specifically theoretical frameworks that present new sociotechnical theories within information science (Rosenbaum, 2010). ICT's have also allowed for other possible interactions to occur in medicine, communications, education, and business - as well as interactions within commerce, government, and banking. As these interactions increased, so did the transactions. Transactions are data, and the transformation of data into information required the need for the unique disciplines in information science to study the context of these phenomena, and this too positioned information science as a social science (Rosenbaum, 2010).

Ultimately, information science is about people. It is a user center field and aids in giving focus to special concentrations of study in information behavior and social informatics. Additionally, computer science continues to have an effect on information science. Today, new

sub-disciplines, such as data science, are being created within the information science domain - namely because of the effects that computer science has had on the field of information science. Furthermore, the use of the internet, social media, and artificial intelligence, among other new technologies and concepts entering the market, have placed information science in hyper-mode because of the vast amount of data and information that now exists and is generated. Due to this hyperactivity of data and information generation, new technologies have surfaced, such as in the space of data science. The creation of data science software has allowed for new capacities in assessing and analyzing data. The value in these software systems can be priceless for many industries, such as academia, government, and healthcare. In the field of healthcare alone, data science software can help illuminate hidden insights from data.

The purpose of this research segment is to help explore and delineate the value and impact data science software can have through the diverse approaches and applications of healthcare. For example, data science software can have a positive impact if the software is capable of aiding in the reduction of costs, while simultaneously analyzing areas for quality of care improvements. Although this is promising, questions still remain as to what use data science software may really have within healthcare. Therefore, this study will focus on the value data science software applications can have in that focus.

1.2 Problem Statement

Data science can play a vital role in healthcare. To an increasing extent, the industry can benefit from the usefulness in data science technologies that pertain to their respective healthcare disciplines. The challenge is to integrate technology and data science to the benefit of healthcare. Though efforts toward such integration exist, healthcare will benefit from greater

infusion of data science. Ultimately, a greater infusion of data science can help mitigate the lack of possible new insights in healthcare. The word “lack” does not attempt to state that healthcare is behind in technology adoption, but that certain technologies in its arsenal may not be conducive to acquiring new meaningful information, therefore, creating a gap in healthcare information that may be valuable, but often missed. This lack may only continue to grow if healthcare does not attempt to expand its horizon in the data science technology spectrum. Software that does not necessarily align itself with data science processes and functionalities can typically be because of its limitations in evaluating data, metrics, and information that stems from primary, secondary, or tertiary data sets, and can therefore affect the healthcare industry negatively. This dissertation helps to highlight the positives of data science use for healthcare by creating further awareness towards the importance of integrating technology and data science in an effort to help healthcare perceive a better way to process data for the benefit of gathering meaningful insights from its data.

1.3 Research Questions

The initiation of this research was due to the need for assessing data science software as a means to explore and apply its capacities in the healthcare space. The utilization of healthcare data has existed, but to what extent is the data often underutilized, and therefore disregarded by many in the industry. For this study, the general research idea is to explore and assess data science software and its use in healthcare. Therefore, this research overview suggests the following questions that are worth of investigation and act as the main guiding foci for this study:

1. What studies are known about data science in healthcare? A construction of a bibliographic essay on data science in healthcare.
2. What technologies in the form of software exists in data science? A construction of a technologically applied essay on data science in healthcare.

Below are the specific research questions that guided the research for each essay in this dissertation in an effort to create a focus for each specific study in this research:

1. What can be answered about the utilization of data science software to address health disparities in public health?
2. What promising data science software are useful to solve health disparities?
3. Can there be a valuable synergy between data science and healthcare?

Additionally, in this dissertation, the third essay elaborates and demonstrates a full example of tracking and explaining the geography of disease, specifically prostate cancer in the states of Texas for a specified time period.

1.4 Purpose and Contribution

The motivation for this study is to explore how healthcare can benefit from the use of data science software, such as geographical information systems and statistical software, both of which are a type of data science software systems that allow for special insights from the data processed in its programming.

The main goal of this dissertation is to help convey what data science is, where it came from, asses its software, and its application to healthcare. However, the main contribution of this dissertation is to reveal the value data science and data science software can have in addressing the needs for healthcare. For example, users applying data science software to publicly available data sets, in an attempt to understand the geography of disease, may ultimately help determine the potential new insights valuable for the public, government, and

healthcare sectors. Furthermore, it also allows stakeholders opportunities to actionably respond to potential problems gained from the insights acquired. In the example of this dissertation, which is elaborated further in Essay 3, prostate cancer mortality occurrences can be determined and potentially explained geographically, giving opportunity for potential actionable responses to diverse communities of practice. This is important because these insights can help create an agenda for increased treatments or increased screenings to help mitigate the death rate. In short, the results and insights acquired can essentially be utilized to help provide meaningful evidence to stakeholders, if making the case for actionable responses is warranted, whether those actionable responses come from the public, government, or the medical community. In essence, by revealing the value of data science in the healthcare space, it can be deduced that the discipline of data science has made a significant disciplinary contribution to addressing the needs of healthcare. This research will also help position data science as a viable means to solutions and to the creation of new possibilities. Therefore, the purpose of this study is to investigate and research how data science software can be applied to healthcare.

1.5 Organization of the Dissertation

The focus of this dissertation investigates data science in relation to healthcare applications. This manuscript includes five main chapters. Chapter 1 includes the introduction to this research, followed by the literature review in Chapter 2, which discusses integral parts of data science such as the discipline, data production, and its positioning for healthcare data science applications. The three essay methodology used in this dissertation is delineated in

Chapter 3. Chapter 4 consists of the results separated into three essays. Finally, the research and conclusion of this complete dissertation are provided in Chapter 5.

CHAPTER 2

LITERATURE REVIEW

2.1 The History of Data Science and Healthcare

There is a substantial amount of interest on how data science may be used to the good of healthcare organizations. This topic, as it is still being studied, is often researched for its opportunities in medical applications, such as in the field of healthcare, as well as in government applications with links to the field of healthcare. For example, these opportunities can be explored with the use of appropriate data science technologies, such as a geographic information system (GIS), which can help monitor communicable diseases and benefit healthcare organizations. Additionally, public and/or private healthcare systems may currently profit from measuring the impact that data analysis software may have over their organization given the different aspects of data. This can lead to technology implementations that may result in more noteworthy and relatively meaningful healthcare results.

Data science covers a broad variety of strategies and processes. The processes contain a lot of simple and sophisticated ones that transforms data into usable knowledge. Meaningful knowledge is useful in many fields of study; this is particularly true of disciplines that have a lot of details. Data science will help build and establish data systems that are the foundation for the enterprise, through researching, processing and visualizing your data. This style of structure works well for gathering the data that is refreshing information and giving the users access to valuable information. The domain of data science often refers to the world of analysis. It does not have to be restricted to the corporate level, although this is the most frequent usage. Data science may assist in the understanding of phenomenon that has occurred or is occurring, in a

field of a particular study that the scholar is attempting to understand. Data science has developed as a widely respected field, but there is also a great deal of maturing that needs to be achieved all over. Additionally, data science as a discipline is also in its infancy.

Data scientists are individuals who concentrate their career in this area. Data scientists are still being created, and there are more many skills that need to be taught. In general, a data scientist has business intelligence expertise, but not every business intelligence specialist is a data scientist. Data scientists are generally highly educated. According to an EMC Data Science Community Survey (EMC, 2011), data science practitioners were 2.5 times more likely to hold a master's degree and 9 times more likely to hold a doctorate degree than business intelligence professionals.

2.2 The Positioning of Healthcare for Data Science

Healthcare firms have seen progress since they have utilized artificial intelligence in collecting large data to conduct clinical trials (Baptista et al., 2019). The industry in healthcare is preparing itself to retrieve useful information from data science technology and systems, which help in substantial data science use in healthcare applications. Information from electronic health reports and other agencies such as the Center for Medicare and Medicaid Programs (CMS) allows for its usage in many critical environments of healthcare, including at the state level.

Data sets may also be used for studies on population demographics and other dimensions, such as income and lifestyle. In this way, this will help the researchers highlight holes based on their subject (Chase & Vega, 2016). This form of data has ramifications for the field of data analysis, and it is also essential for those who are implementing data science in the

healthcare sector. There will be a major change in the study of data, and more specifically, to the multiple disciplinary facets of organizations such as CMS, which profit from discovering specific and deep perspectives hidden among the complexity of variables and attributes that can occur in their data. Although there are organizational disciplinary factors such as financing, administration, and also legislation; particularly policy that can have a significant effect on certain healthcare disciplines that may conform to CMS' guidelines, this doesn't make these aspects less relevant to procedure efficiency. CMS' initiatives should not only impact doctors and the healthcare staff, they also concern clinics and the public. It is therefore crucial that these influential organizations exploit data science for achieving better insights, particularly because most of healthcare will stand to benefit in changes from new policy put out by organizations such as these.

2.3 Possibilities from Data Science

Healthcare non-profits of some kind will therefore quickly benefit from the usage of data science and its methods. In essence, the methods of data science rely on gaining a degree of new insight that can be obtained by processing and measuring the data. With the advent of advanced technology and the potential for predicting an outcome based on historical data, the possibilities for predicting the outcome based on historical data are now feasible. The verbal interpretations of this hypothesis would almost release a never stopping new universe of possibility because as evidence becomes increasingly accessible by new technology, the new knowledge may be significant (Spruit & Lytras, 2018). Whatever the case could be with any company or agency who wants to use data science apps, doing what wasn't feasible previous to the usage of data science applications, is still a move forward in the right direction. The aim of

data science is very much to adapt itself to an issue or requirement, be it in the world, or commercially. Major fields of medicine, such as clinical research and medical research, stand to gain great benefits from the use of data science, particularly in deciphering new insights from the data that is generated within this region (Spruit & Lytras, 2018).

New research in data science will spring ahead from the problems data science is experiencing. That is to suggest, the findings of new hypotheses and models will come in from data science due to the amount and variety of data which has been generated by web-based, interactive and ubiquitous technology, and due to the quickness at which data is being produced by large-scale enterprise, social networking and scientific applications (Maneth & Poulovassilis, 2016). Even innovative technologies are created because major businesses like Google and Facebook take advantage of the opportunities to use data in an effort to remain on top of the rivalry or new companies joining the business (Loukides, 2011). This too would further extend and describe the possibilities of data science.

2.4 Data Output

Making and utilizing data has been central to many disciplines, but evaluating the data is of vital significance to many fields. While many possess basic statistics, some rely on digital data, which are growing in significance for research, and the methods to interpret those data need modern data analytics (Westra et al., 2017). Data science can affect innovations of philosophy and economy that is focused on data; in other words, data-driven theory and data-driven economy. To avoid this from occurring negatively, we must cooperate between various disciplines, including the healthcare sector, the computer field, and information science field (Cao, 2017). There are many other significant applications of data science technology, such

as ridding the planet of disease epidemics. The International Society for Disease Surveillance (ISDS) organize regular conventions for the agency. The 15th annual conference theme was titled, "New Frontiers in Surveillance: Data Science and Health Security", and it was based on two key items. The first major move in utilizing health records was to make awareness from data created from health information systems and the second, was to allow the reaction of nations to the outbreaks of disease (Dixon, 2017). These types of conferences have allowed experts in the field of data science and the arts of mathematics to begin to recognize the value of data science and computation art and their capacity to extract new information. More and more disciplines are raising the consciousness of data science among the general public and find that it is imperative to the potential vision of the arts of their respective discipline.

2.5 Graphical Information Systems

Graphical information systems are important instruments that data scientists may apply to their tool set for monitoring diseases and evaluating possible causes leading to disease. Health is a major concern in today's global society, particularly in public health. Transactional data that can be gathered from patients from their electronic health record systems can help evaluate several important variables beyond detection of a disease, such as for example, susceptibility, suicide threats, and other factors which will possibly drive direct clinical treatment and study to enhance healthcare (Chiu & Li, 2018). In addition to the goals of physicians and scientists, health organizations want to provide knowledge on when a specific illness is taking place like prostate cancer. This also helps in the monitoring and clarification of the spread of the disease (De la Torre Diez, Cosgaya, Garcia-Zapirain, & Lopez-Coronado, 2016). In essence, many resources and technologies can be used to support data scientists in their

respective field, particularly in the field of health but it is important to take in that data scientists are an essential component of the data analysis method. The data scientists use data mining to extract valuable knowledge. Data scientists can help to educate the decision making and function in the medical sector. Additionally, a wide-range of industries can benefit from the expertise of the data scientists after a thorough review of their data (Power, 2016).

2.6 Big Data Emerges

The world in healthcare is already creating loads of data and there is already a modern trend of healthcare, and that is the concept of big data. Big data is sometimes misused as a buzz word to indicate actually processing massive datasets. The word Big Data means a very wide volume of data (Rumbold & Pierscionek, 2017). Large data may also see a huge application in healthcare. In the case of GIS systems, big data will improve public health monitoring by integrating spatial variables and social determinants of health (Zhang et al., 2017). Allen et al. (2016) used GIS methodologies and data mined from social network sites, and then leveraged techniques in machine learning, a part of data science, to sort through the data before study. In the healthcare sector, Big Data and its interpretation are helping to reduce expenses in the clinical analysis of electronic medical information (De la Torre Diez, Cosgaya, Garcia-Zapirain, & Lopez-Coronado, 2016). This is only a tiny sampling of the various forms of data science applications and frameworks that currently operate and can be used for healthcare enhancements.

Data science can be an essential aspect of healthcare and can be used in clinical informatics. Information technology has proven helpful in rising productivity in healthcare. For example, clinical informatics software is a mix of information technology and implementation in

the distribution of health information (Alexander, 2015). The healthcare sector has acknowledged the value of informatics and it has invested extensively in it (HIT). This has prompted an increase in research in this area (Detmer & Shortliffe, 2014). However, informatics alone cannot solve all the issues in healthcare and data science may offer a solution, although many data science activities may not achieve full maturity for many years to come. In essence, there are computational and technological capacities aligned with informatics that may be useful and desirable elements of data science.

Technology is essential to physicians because it helps in reducing needless health treatment, reducing medical care costs, and enhancing patient safety. Data science may help assess the quality of treatment needed for the well-being of the patient. The details can emerge from recent perspectives from data analysis methods for patient health improvement.

Knowledge and health-related analysis have been going on for a time. Until electronic health records, medical details were being assessed by providers. Sadly, the research was constrained due to the lack of technical capacities at the time. They tried to enhance the welfare of the patients, but there were drawbacks such as an overabundance of evidence that might be overlooked while the patient's status is measured. Electronic health record systems were developed as a part of this competition. Additionally, Congress has been active in the marketing of health information technology since 2004 - when Congress adopted legislation to use health information technologies and health information sharing programs (Marchibroda, 2007). There are states which have rendered it a top priority to utilize such technologies. This is an essential move in healthcare, especially because, in the area of data analysis, most data come from an archive or computer framework of some kind. Furthermore, the Department of

Health of New York decided the complete implementation of HIT and HIE, which has advantages to the medical community as a whole. The state of New York in 2006 introduced the Healthcare Quality and Affordability Legislation for New Yorkers (HEAL NY), a grant focused initiative that focuses on three areas: 1) electronic health record (EHR) implementation, 2) electronic prescribing (ePrescribe), and the creation and implantation of exchange partnerships in the community (Kern & Kaushal, 2007). All these initiatives helped with the digitization of more health data.

2.7 Data Science

Data science is really special. It is essentially a blend of theory and science approach, which usually utilize a package of data, tools and functionalities within its scope, such as the simulation of findings (Broome, 2016). Data science goes deeper than the average analysis. Data Science is the way in which historical understandings of what is, would be, and has been gathered is used to discover better alternatives for the future, utilizing technology, in a mathematical application scientists' fashion. Data science can be thought of as the various forms of implementing reasoning through a sequence of acts, which are then able to generate new information regarding data, which could assist in improved decision making (Power, 2016).

Data science is the use of sophisticated techniques to predict data from the data (EMC, 2011). These techniques are symbolic of the activities a data scientist would perform while approaching data. There are several components to data science, which vary from study-based disciplines, such as economy, social, and census research, to financial, scientific, advisory, industry, and media disciplines (EMC, 2011). Perhaps, the best way to explain data science is that it is a combination of data inference, algorithm creation, and technology to solve

analytically complex problems (Data Science Central, 2016). Data science is extracted from the processes of gathering, planning, processing, visualizing, handling, and storing data (Stanton & De, 2013). Data science is not only organized by the participation of these systems in operation, but rather research and process, as a definition, as well as their respective process elements, each with their own emphasis (Smith, 2006). Data science can be applied to several kinds of applications, and hence it is essential to provide domain-specific applications to derive useful knowledge from data science (Liu et al., 2009). Data science has also helped to progress in other fields of medicine, particularly healthcare. The modern data science innovations are allowing potential findings from the data. Data science involves understanding how to lay the foundation of the data and know how to collect the data. It is essential to understand the methods to better analyze and evaluate the data or portray the data. It is therefore necessary to ensure the data cannot be destroyed in the future or at least to store the data because it is accessible for decades.

As data analytics is becoming more and more sophisticated, there is a growing need from companies for individuals with specialized data analysis expertise. Individuals ought to have an in-depth understanding of data science strategies and principles, particularly in the realm of large data, since this topic relates to data in the "V" characteristics, such as length, velocity, variety, meaning and veracity (Maneth & Poulovassilis, 2016). Data scientists grasp data science principles. They have the expertise and experience to effectively use data science techniques. The word data scientist is not well-defined. It is a widely debated concept and several have tried to describe it. The challenge to describe this concept would be close to the complexity in describing “information scientist”. In order to differentiate the field, it is

proposed that “data scientists”, is a mash-up of statistician, observer, and code maker (EMC, 2011). Even as a description of a data scientist cannot be made, there are several questions that will help in describing the word. That is to suggest, we may identify a data scientist from their curiosity in data science. This implies that if term “data scientist” is found to be the true essence, then perhaps it is well-defined. To obtain an understanding of data scientists, one may deduce that they collect data, massage it into a tractable shape, allow it say its tale, and present that story to others (Loukides, 2011). As a tentative concept, data scientists are those dedicated to the analysis of data, metadata, quick retrieval, archiving, sharing, mining to discover unexpected information and data connections, simulation in two and three dimensions like movement, and management (Liu, et al., 2009).

Data scientists are acquainted with libraries such as Python, Perl, R studio, Hadoop, SQL, deep learning applications, and the like. The R statistical kit, Python, and Perl are among the most popular software utilized in data science occupations, with about 25% of professionals utilizing these tools (EMC, 2011).

Data scientists may use artistic abilities through their careers to help generate a message from the data (Loukides, 2011). To be a good data scientist, one must be attentive, perceptive, and have a good imagination (Patil, 2011). A data scientist should be able to take a topic and integrate diverse solutions for the particular issues relevant to the multiple challenges in the big challenge at hand (Loukides, 2011). Data scientists have the potential to learn a variety of different abilities. To do so, data scientists must mix entrepreneurship with patience. They must be eager to develop these devices incrementally. They are able to explore their ideas before refining them into a proper product. And of course, they must continually be able to

iterate and improve their own products (Loukides, 2011). That is to suggest, a data scientist will profit from studying the skills that are relevant to computer science or mathematics. A data scientist must have the potential to learn the technology domain, interact with data consumers, have attentive exposure to the big picture of a complicated structure, have competent information about how data can be interpreted, converted and evaluated, have the capacity to visualize and present data, have the ability to have ethical thinking skills (Stanton & De, 2013).

2.8 History of Data Science

The fundamental elements and methods in data science have been in use for several years. While data science has recently become common among industries and among a variety of disciplines, there were pioneers who helped data science surface from an unseen discipline to a recognizable one. Dr. Vincent Granville, an influential scholar in the area of data analysis, developed Data Science Central, a well-regarded platform for those interested in analytics and the science of data. He has also been regarded by Forbes and CNN as one of the main influencers in the domain of data, particularly big data (Data Science Central, 2013).

Data science didn't gain legitimacy until the early 2000s and only gained legitimacy when the “Data Science Journal” provided recognition to the word (Smith, 2006). Data science gradually came into being by the evolution phase owing to various fields and the evolution of scientific advancements over the decades. The data science pioneers realized the value of enhancing how we share knowledge. They didn't know precisely how it would be. Data science was starting to develop thanks to the combined efforts of early pioneers. John W. Tukey published an essay on data analysis's prospects in the early 1960s and highlighted that statistics are highly significant in interpreting results. Paul Naur wrote a book in the early 1970s. In his

novel, the author used the word ‘data science', but no one knew it for what it was about to become, a new discipline from the viewpoint of an earlier field and probably a collection of analytical processes.

Knowledge Discovery in Databases became a hot topic in the late 1980s. Finally, in 1996, the term “data science” was used in a title at a conference in Kobe, Japan when the member of the International Federation of Classification Societies (EFCS) met together (Gil Press, 2013). Today, data science has evolved to hold influence through the Data Science Journal, and it continues to grow in importance. Technically, as technologies evolve, data science also evolves. Data science is now a multinational concept. It is becoming well known that "data science" is a growing area that is attracting a wide variety of scientists and will certainly continue to expand (Liu et al., 2009).

2.9 Frameworks

The main themes and conceptual guidelines of data science components and processes can be found in two generally accepted data science models. The first model is the data science project life-cycle and consists of 7 cycles or components, and are as follows: 1. data acquisition, 2. data preparation, 3. hypothesis and modeling, 4. evaluation and interpretation, 5. deployment, 6. operations, and 7. optimization (Manna, 2014). The model is shown in Figure 2.1.

Another model shows an overview of the data science process. The process involves six steps and is as follows: 1. Setting the research goal, 2. Retrieving data, 3. Data preparation, 4. Data exploration, 5. Data modeling, and 6. Presentation and automation (Cielen et al., 2016). The model is shown in Figure 2.2.

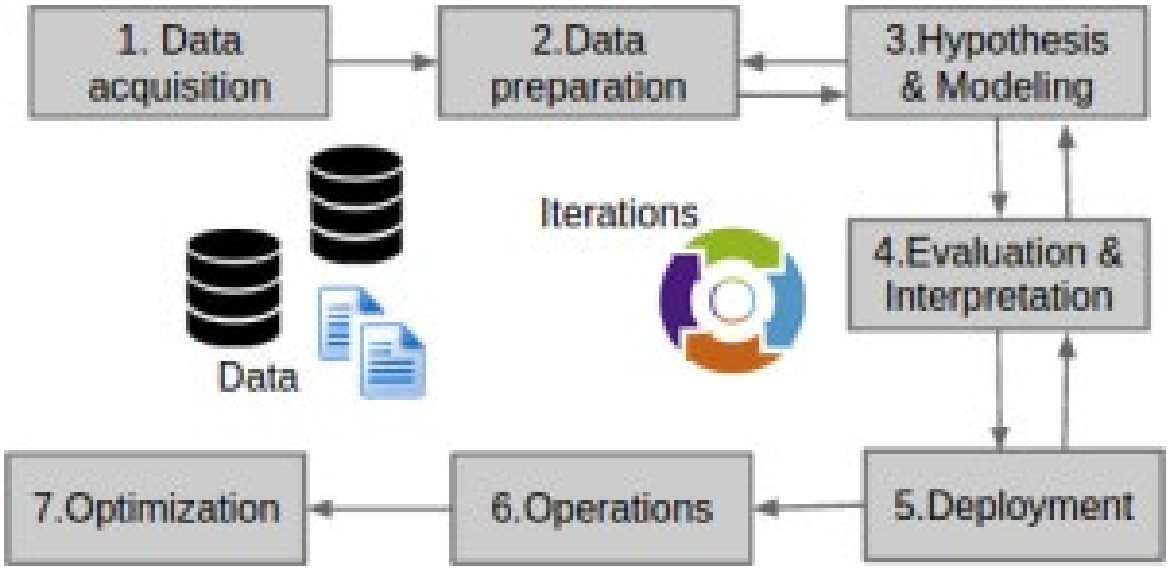

Figure 2.1: Data science project life-cycle (Manna, 2014).

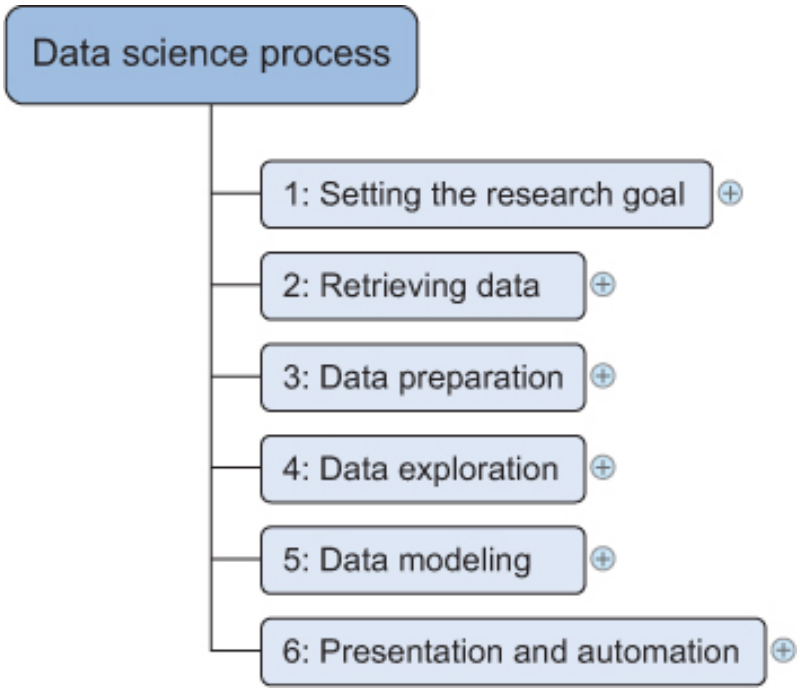

Figure 2.2: The six steps of the data science process (Cielen et al., 2016)

In the Figure 2.3, we find the targets, method, subprocesses, and subprocesses of each main phase in the data science process.

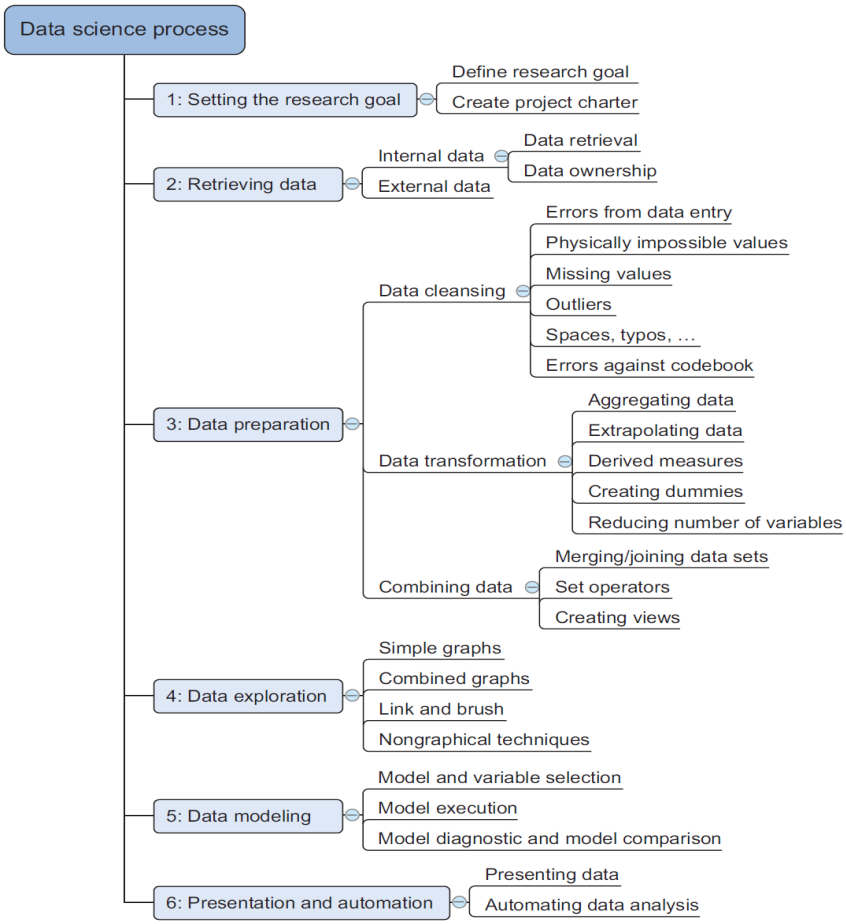

Figure 2.3: Detail mapping of the six steps of the data science process (Cielen et al., 2016)

2.10 Healthcare and Data Science

Demand for data scientists in healthcare has developed over the past few years. In healthcare there are many fields that utilize data and thus the knowledge stored in healthcare networks has grown over the past ten years (De la Torre Diez, Cosgaya, Garcia-Zapirain, & Lopez-Coronado, 2016, p. 1). Healthcare data should be used in a meaningful way, to be usable for improvement. It lets data scientists develop new insights and can be helpful in building new technologies that eventually benefit all players in healthcare, from the patient to the treating doctors (Adam, Wieder, & Ghosh, 2017). Additionally, the health and medical science field can

benefit from the potentials of data science. Not only does data science allow for many types of developments, such as professional development, but also in building data-driven theories that would be conducive to making data science a more advanced analytics discipline that involves data science processes in the healthcare field (Cao, 2017). The primary benefit of utilizing data science processes such as deep learning and graph analysis when evaluating large data is that the projections of patient results will help us discern the future outcomes of some population classes of patients. This, in essence, enables for the best treatment treatments to arise, and for fresh insights to emerge for better quality patient care results (Adam, Wieder, & Ghosh, 2017).

The ultimate aim of data science in healthcare is to extract new knowledge that can help us make smarter choices on quality of treatment for patients (Adam, Wieder, & Ghosh, 2017). While the application of data analytics may be helpful, there are still problems in the domain of human health. Challenges involve data consistency, insufficient data, and standardizing data (Delaney & Westra, 2016). There are dynamic technological problems within healthcare that are covering many areas of medicine. In healthcare, there are specialties that have often had issues with data. The difficulties are in collecting, exchanging, and processing data (Dunn & Bourne, 2017).

2.11 Summary

The area of data science, together with the data scientist, use various statistical methods for evaluating data in any domain or industry, such as healthcare. For example, the area of data science is continually rising, shifting, and evolving. Not only is data science its own field in and of itself, but it is a “working” paradigm, whether conceptual or not, for the various disciplines that occur. As a field, data science is looked at as a sequence of measures that can

be used to find the answers of how data can turn hidden phenomena to impactful strategies that help solve problems.

Healthcare will be preserved for the advancement of data analysis, which will further boost the industry as a whole. In a sense, data science is several distinct disciplines brought together as one. This will cause data science to penetrate the walls of all scientific disciplines. Several organizations in this sector are beginning to venture into the area of software sciences to explore creative algorithms and methods that can assist with improved standard of care of cognitive sciences and clinical outcomes. Data science is a valuable competence to be learned by those interested in the healthcare industry, and to learn new knowledge in this ever-growing multidisciplinary sector would be beneficial. The growing role of data science in the world implies that the concept, discipline and science itself has a promising future and the potential for improving new systems and methods for healthcare.

Data science analytics are useful tools that can reduce dramatically the costs of healthcare services, but so much more research is important before we determine completely the effect that this can have on the healthcare sector. Assuming that current available research show that for the above reasons there is no significant impact, more study on this subject will be required.

CHAPTER 3

METHODOLOGY

The purpose, problem and research questions set forth in the dissertation is addressed in three distinct essays and the results present the findings from the essays. Essay 2 extends Essay 1, and Essay 3 extends both Essay 1 and Essay 3. Each essay has a specific motivation and contributes to the overall research question and extends the body of knowledge in the discipline.

3.1 Essay 1

It is not a relatively recent phenomenon to address health disparities in public health, but there are limitations in recognizing the broad spectrum of those disparities in an attempt to better overcome them. The application of resources of data science can highlight the ability to solve health inequalities in public health. The purpose of this study was to undertake a systematic review of the literature to help describe micro and macro themes. To help determine this, a thematic analysis was conducted for the studies selected. The systematic review was guided by the Preferred Reporting Items for Systematic Review and Meta-Analyses guidelines and were used to systematically review the application of data science software and its role in helping to address health disparities in the public domain. In the discovery process, 564 references were found. Also, 387 abstracts were checked for research relevance after duplicates were eliminated, of which 316 documents were omitted during the screening. There were 71 full text articles that were collected and checked for eligibility. Of the 71 full text articles, there were 22 eligible articles that were chosen for further full text evaluation and 49 were removed with reason. After thematic analysis, in addition to the implementation of

methodological techniques across various computing systems, the findings demonstrated a strong convergence of geographical analytical approaches. This highlights the need for integrated geo- and statistical methodologies for healthcare data science to better promote further study and technology use in resolving health inequalities in public health.

3.2 Essay 2

The article assesses data science software to examine the effectiveness of data science technologies to address problems such as health disparities. An assessment of data science software was conducted using KDnuggets data pertaining to analytics, data science, and machine-learning software. Data science functionalities provide analytical processes and applications suitable for healthcare. Tensorflow and Python can automate and model the analysis of income, education, race, age, while cross-referencing such variables to outcomes in patient care and finance, revealing health disparities. This study demonstrates the utility of leading software to perform data science operations that can improve care in healthcare networks by addressing such factors as health disparities. Such findings document the process by which how the healthcare community can continually and iteratively evaluate data science software. Integration and modification of methodologies in this paper allows users to consider the evaluation of data science software for healthcare applications, particularly around health disparities.

3.3 Essay 3

A small study was conducted that incorporates data science software utilization in an effort to showcase significant use of data science software, namely geographical information

systems (GIS) and statistical package for the social science (SPSS) in the healthcare spectrum. The study has a problem, three main hypothesis, research questions, and results that drive the study. A literature review of the topic was also be conducted, taking into consideration multiple factors that contribute to new insights or solutions to problems. The study looks at the geography of prostate cancer mortality in Texas between 1999 to 2009. It leverages the International Classification of Disease, Tenth Revision (ICD-10) codes for Prostate Cancer (C61). In an effort to show the distribution or geography of the disease, choropleth maps were used. Two datasets were also used in the study, specifically from VitalWeb to show mortality data, and Texas Health Rankings to analyze explanatory variables in the study.

This study uses a geographic information system to create and analyze choropleth maps determining the distribution of prostate cancer in Texas and uses SPSS software to analyze social determinants of health that may explain prostate cancer mortality. The data, collected for period 1999–2009, was furnished by the Texas Health Rankings and VitalWeb. The dataset was for 1999–2004 and 2004–2009. It was comprised of age-adjusted data specific to the 2000 US Standard Population data, based on an age-distributed and -weighted methodology to create age adjustments for statistical purposes. The study found there was a statistically significant ($P < .05$) percentage of African Americans with age-adjusted prostate cancer mortality, but no statistically significant correlations were found in other races. The study indicates a number of ways medical communities and public health agencies can employ GIS and SPSS to screen for and treat prostate cancer more effectively.

CHAPTER 4

RESULTS AND DISCUSSION

4.1 Essay 1: Addressing Health Disparities in Public Health through the Application of Data Science Software in the Last Five Years: A Systematic Review*

4.1.1 Introduction

Where there are inequalities in variables such as social class, race or gender, both of which may be disruptive and expensive to public health, the standard of treatment and health conditions deteriorate (Demeester et al., 2017). These deterioration and drawbacks are recognized as health disparities or health inequalities and present obstacles for organizations to fix them due to the fact that health disparities are barely observed several times and appear to hide in stealth mode under the radar. This then opens up opportunities for analyzing technologies in the space of data science and their potential to solve health inequalities in public health. Data science has a broad variety of approaches, from text mining to spatial models and statistical processing (O'Connor, 2018).

Healthcare data science can be restricted in certain respects, but it can still help to solve critical and popular healthcare issues, especially where disparities in health may be present under the radar. These health inequalities can occur across organizations in healthcare. Data science can help to address these gap problems and optimize patient care and enhance performance. Nevertheless, despite attempts to resolve health inequalities in the public domain, there is already a disparity in efforts to address them, which can be attributed to the lack of effective tools to define and resolve them.

* Essay 1 is presented in its entirety from Huerta, J., Senn, W., & Prybutok, V. and is pending submission to the Journal of Decision Systems.

Topics found in published research, through conducting conceptual observations of the data, as well as through means of reviewing patterns that are then grouped together for classification purposes, can be useful for research and practitioners in that it allows guidance for building the different types of data scientific approaches that are important in addressing health disparities in the public domain.

The use of data science software to address and resolve health disparities in public health has earned relatively few systematic reviews. The goal of this paper is to conduct a systematic review of publications related to data science and health inequalities in an attempt to recognize micro-level and macro-level themes that will assist prospective scholars in the subject with various facets of approaches and information regarding the subject matter.

4.1.2 Methods

This study adopted the Preferred Reporting Items for Systematic Review and Meta-Analysis (PRISMA) Statement (Liberati et al., 2009) as a guide. In December 2020, using three databases from the Ebscohost platform, a comprehensive search of literature for the application of data science software and its role in addressing health disparities was conducted.

4.1.2.1 Eligibility Criteria

Initial inclusion criteria included studies that were scholarly peer reviewed, published between January 2015 to December 2020, and were printed in the English language. The rationale for the eligibility criteria was to explore current studies in English to help English based universities and organizations. Additionally, there was high interest in exploring reputable articles that tend to come from peer-reviewed journals. Studies were excluded in the

comprehensive search that did not meet these criteria. All peer reviewed literature describing health disparities and data science software applications published based on the criteria delineated were included for further analysis.

4.1.2.2 Information Sources

To locate published peer-reviewed studies exhibiting the application of data science software to help address public health disparities, a search strategy was constructed. Three main electronic databases were searched. These included Medline, Health Source: Nursing/Academic Edition, and Academic Search Complete.

A brief analysis of the reference lists and citations of the included papers were carried out in an effort to find any additional literature on the subject matter not originally found in the main search.

4.1.2.3 Search Strategy

On December 10, 2020, the main search was carried out. Search strategies were developed for all electronic database utilized for this systematic review to include the following search terms:

1. Data science, analysis, analyze, machine learning, big data, artificial intelligence, predictive, prediction, inferential.
2. Application, software, freeware, shareware, R, python, sql, Hadoop, statistics.
3. Health disparity, public health, population health, community health, primary care, patient-centered, medical care, healthcare, patient care.

The three delineated levels above document targeted search terms and were combined with Boolean/phrase search modes to capture the best return within each database. The following example of the search query was separately initiated for each database:

AB (“data science” OR “datascience” OR “analy*” OR “machine learning” OR "big data" OR “AI” OR “artificial intelligence” OR "predict*" OR "infer*") AND AB (“application*” OR “software” OR “freeware” OR “shareware” OR “R” OR “python” OR “sql” OR “hadoop” OR "statistics") AND AB (“health disparity” OR “health disparities” OR “disparities in health” OR "disparities") AND AB ("public health" OR "population health" OR "community health" OR "primary care" OR "patient-centered" OR "medical care" OR "healthcare" OR "healthcare" OR "patient care")

Table 4.1: Systematic Review Search Strategy

Search ID	Database	Search Date Conducted	Search Syntax	Field Option	Search Options	Date Range Search	Number of Results
S1	Ebscohost Medline	12/10/20	AB (“data science” OR “datascience” OR “analy*” OR “machine learning” OR "big data" OR “AI” OR “artificial intelligence” OR "predict*" OR "infer*") AND AB (“application*” OR “software” OR “freeware” OR “shareware” OR “R” OR “python” OR “sql” OR “hadoop” OR "statistics") AND AB (“health disparity” OR “health disparities” OR “disparities in health” OR "disparities") AND AB ("public health" OR "population health" OR "community health" OR "primary care" OR "patient-centered" OR "medical care" OR "health care" OR "healthcare" OR "patient care")	AB Abstract	Limiters - Scholarly (Peer Reviewed) Journals; Date of Publication: 20150101-20201231; English Language Expanders - Apply equivalent subjects Search modes - Boolean/Phrase	1/1/2015 to 12/31/2020	332
S2	Ebscohost Health Source: Nursing/Academic Edition	12/10/20	AB (“data science” OR “datascience” OR “analy*” OR “machine learning” OR "big data" OR “AI” OR “artificial intelligence” OR "predict*" OR "infer*") AND AB (“application*” OR “software” OR “freeware” OR “shareware” OR “R” OR “python” OR “sql” OR “hadoop” OR "statistics") AND AB (“health disparity” OR “health disparities” OR “disparities in health” OR "disparities") AND AB ("public health" OR "population health" OR "community health" OR "primary care" OR "patient-centered" OR "medical care" OR "health care" OR "healthcare" OR "patient care")	AB Abstract or Author-Supplied Abstract	Limiters - Scholarly (Peer Reviewed) Journals; Published Date: 20150101-20201231 Expanders - Apply equivalent subjects Search modes - Boolean/Phrase	1/1/2015 to 12/31/2020	56
S3	Ebscohost Academic Search Complete	12/10/20	AB (“data science” OR “datascience” OR “analy*” OR “machine learning” OR "big data" OR “AI” OR “artificial intelligence” OR "predict*" OR "infer*") AND AB (“application*” OR “software” OR “freeware” OR “shareware” OR “R” OR “python” OR “sql” OR “hadoop” OR "statistics") AND AB (“health disparity” OR “health disparities” OR “disparities in health” OR "disparities") AND AB ("public health" OR "population health" OR "community health" OR "primary care" OR "patient-centered" OR "medical care" OR "health care" OR "healthcare" OR "patient care")	AB Abstract or Author-Supplied Abstract	Limiters - Scholarly (Peer Reviewed) Journals; Published Date: 20150101-20201231; Language: English Expanders - Apply equivalent subjects Search modes - Boolean/Phrase	1/1/2015 to 12/31/2020	176
						Total Number of Results	**564**

Table 4.1 documents author described categories: search id, database, search date conducted, search syntax, field option, search options inclusive of limiters; expanders; search modes; date range search, and number or results. After all criteria was applied to the comprehensive search on the Ebscohost platform, search S1 Medline yielded the most results (332), followed by S3 Academic Search Complete (176), and then S2 Health Source:

Nursing/Academic Edition (56). The total number of yielded results for all three databases was 564.

4.1.2.4 Data Collection and Study Selection Process

All records searched were retrieved using the Ebscohost platform to connect to each individual database. Using Ebscohost record saving functionality, all records searched for each database were saved in unique record folders and downloaded to a master Microsoft Excel spreadsheet. Data collected for the research included information on the following attributes: article title, author, journal title, publication date, volume, issue, first page, page count, abstract, analysis methodologies and outcomes.

Using the inclusion criteria delineated by the authors, all documents were screened independently of each other. The measures taken for the selection of papers were the initial evaluation of significance using the abstracts of the listed sources. Abstracts were coded using the following coding process: 0 = exclude, 1 = include, 2 = maybe. Papers coded as 0 were deemed meaningless and were discarded from selection consideration in the screening process. Published articles coded as 1 = include or 2 = maybe were retrieved. The articles retrieved were evaluated for systematic review inclusion based on the criteria for selection. After duplicates removed, there were 31 includes, 40 maybe, and 316 exclude.

4.1.2.5 Quality of Included Studies and Risk of Bias Assessment

The research was largely descriptive and retrospective, so risk of bias was mitigated through the examination of the validity of these findings by two expert reviewers. By thoughtful conversation and debate, any disagreements between the reviewers were settled. The research

also did not attempt to discriminate between qualitative and quantitative in the study designs of the publications studied. Quality of included studies was addressed through choosing only peer-reviewed publications.

4.1.2.6 Synthesis of Results

Synthesis of data from the studies allowed for a sensible categorization of the findings into themes both at the micro-level and macro-level. At the micro-level, the following themes surfaced: statistical methods, health disparity, geo analysis, quantitative study, application, software, public health, qualitative study. At the macro-level, three main global themes surfaced: geographic data science, healthcare software applications, healthcare applications.

Classifications of topics is a method with great potential, especially in the area of subject analysis. When seen as a means of aligning commonly used terms for a more global understanding, classifying topics can prove to be useful in systematic reviews. It is crucial to understand that it provides consumers with a logical base of coding processes in order to understand the central principle of classification, which will essentially contribute to the creation of informative topics. The codes can be evaluated after the coding process has been completed to determine the organization of the patterns gathered into categories, typically homogeneously grouped; this can be simplified to a process known as thematic analysis.

There are many possible uses of thematic analysis and for multiple purposes. In this study, a comprehensive and transparent thematic analysis was conducted for arranging, explaining, exploring and interpreting the content reviewed and was used as a general guide in this systematic analysis (Clarke & Braun, 2014; Nowell, Norris, White, & Moules, 2017). Studies that were coded into specified criteria were selected and classified into themes, reviewed, and

further classified at a more global level.

4.1.3 Results

In an effort to understand the systematic review, it is important to also understand the review approach in an effort to get a more thorough understanding of the results. Additionally, the understanding of special skills is not needed, but knowledge of the content studied, and the necessary expertise of the discipline's data that is being analyzed would prove beneficial to researchers interested in this space.

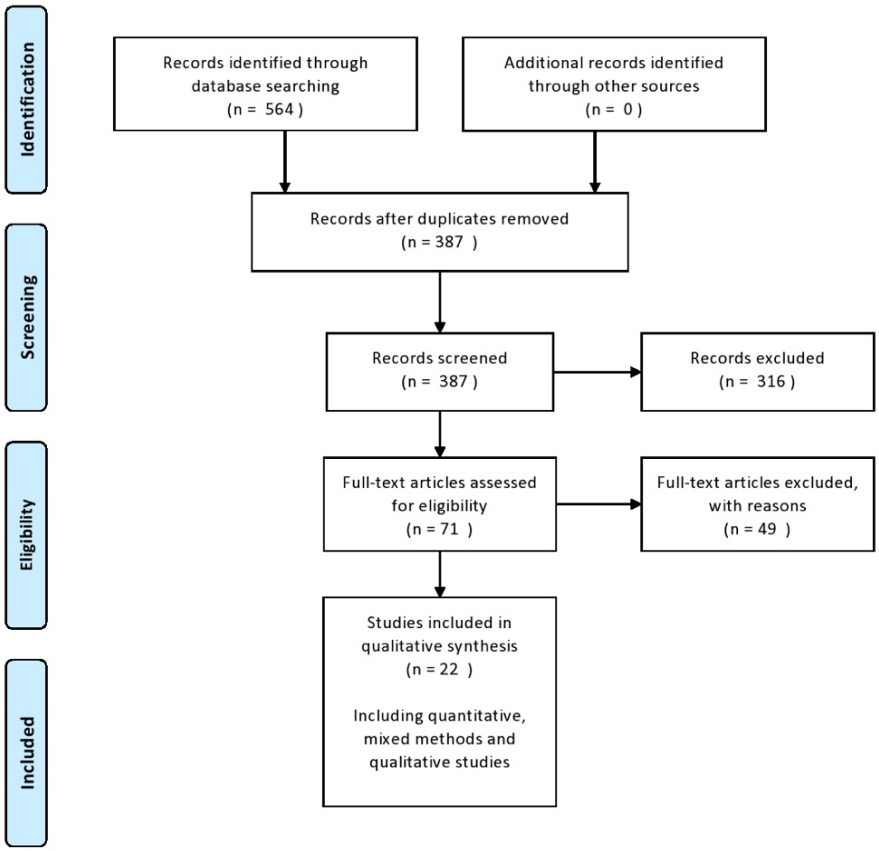

Figure 4.1: Total systematic review

The systematic review exhibits that there were 564 references found in the searching strategy. After duplicates were removed, there were 387 abstracts that were reviewed for

study validity, of which 316 records were excluded in the initial screening. After the initial part of this screening, there were 71 full text articles obtained and reviewed for eligibility, of which there were 22 qualifying articles that were selected for full text review and 49 were omitted with reasons. The 49 omissions with reasons are delineated in the following: 22 traditional academic analysis not inclusive of study, 15 does not align with data science, 12 does not address health disparity. For the full systematic review results, see Figure 4.1.

4.1.3.1 Study Characteristics

Out of the 22 publications that were included in this study, there were multiple themes that emerged. Of the micro-level themes that emerged there were 13 statistical methods, 16 health disparity, 9 geo analysis, 17 quantitative study, 17 application, 21 software, 6 public health, and 2 qualitative study. Geo analysis was determined as a viable theme for topics that emerged in the geographical application of visual and statistical methods for discovery. Further synthesis showed that of the micro-level themes, there were three macro-level themes: 9 geographic data science, 11 healthcare software applications, 2 healthcare applications (see Table 4.2).

Table 4.2: Micro-Level Themes

Micro-Level Themes	n	Macro-Level Themes	n
Statistical Methods	13	Geographic Data Science	9
Health Disparity	16	Healthcare Software Applications	11
Geo Analysis	9	Healthcare Applications	2
Quantitative Study	17		
Application	17		
Software	21		
Public Health	6		
Qualitative Study	2		

4.1.4 Discussion

Twenty-two studies were reviewed in this systematic analysis of the English published literature on the application of data science software and its role in addressing health disparities in public health. Previous systemic reviews are uncommon or have not been performed on the use of data science to solve health inequalities in public health. Consequently, the results of this study are new and can be used to advise on the role that data science applications play in resolving public health inequalities, as well as to assist prospective future scholars in evaluating the thematic framework for future studies on this topic. The variability, diversity, and uniqueness of the intent and nature of the studies involved make it challenging to synthesize results and draw reasonable inferences or deductions. The heterogeneity, however, did offer valuable insights into the thematic structure that can ultimately aid in determining contextual themes to look for when considering publications into the various applications of data science software and their role in solving and addressing components of health disparities.

4.1.4.1 Micro-Level Themes

In reviewing the literature on the use of data analytics to address health inequalities in public health, interesting findings were gathered around the thematic study, indicating cross-sectional associations between health disparities, geographical research, predictive methodologies, and the use of multiple software applications. While in these associations, thematic research revealed strengths, the analysis also showed a lack of qualitative studies in the results and weak results aimed at true public health problems.

4.1.4.2 Macro-Level Themes

Further evaluation of the literature into thematic structure indicated that the micro-level themes could further be consolidated into macro-level themes (see Table 4.2). These themes also helped to determine that, through thorough in-depth synthesis, the relationships between the micro-level themes could establish conditions for the global fusion of themes. The macro-level themes were of value in that applications toward addressing health disparities in public health tend to incorporate a geographical approach in addition to utilizing other software for healthcare applications. This is an important find as all micro-level themes tended to only address a more local synthesis and the macro-level themes exhibited a global aspect, as well as showing an additional aspect in geographical methodologies along with healthcare applications for the publications reviewed.

4.1.4.3 Implications for Practice and Research

There have been few comprehensive studies on the use of data science to resolve health inequalities in the field of public health. The numerous aspects of the studies chosen in this study were analyzed by the thematic overview of the studies published. The various components found in the studies selected showed a variation of components used to help address health inequalities through data science and were highlighted in the study. Thus, opportunities exist for fellow scholars to examine more studies to help determine the full scope of methodologies in addressing health disparities with healthcare data science tools. In respect to implications for practice, public health administrators and analysts in charge of analytical development can be informed through the study about the knowledge and general approaches involved when aiming to address health disparities. Such a creation can be beneficial to

researchers, essentially by establishing a suitable approach to resolve health inequalities that can arise from data science insight and by offering a means to comprehensively interpret the thematic results performed in this systemic study, which they can leverage using these conclusions.

4.1.4.4 Strengths and Limitations

Since this systematic review was intended to collect publications on applications of data science to evaluate their role in resolving health inequalities, not all published findings that may translate into the outcomes of the thematic analysis are likely to have been captured. The analysis is also able to show strengths within the papers, amid these limitations. Reviewed strengths reflect prospects for potential analysis. Following thematic review, the results revealed a clear integration of regional empirical methods, in addition to applying computational strategies through diverse software applications. This highlights the need integrated geo- and statistical methodologies for data science in healthcare to help facilitate further research and use of technologies to solve health gaps in public health.

4.1.5 Conclusion

In healthcare, data science can play a critical role in solving health disparities. Increasingly, the sector can benefit from the utility of data science software. The objective of this paper was to systematically review literature on the use of data science in the public space to address health disparities. In addition, the use of thematic analysis was included in the study to better explain micro- and macro-level themes that emerged from the studies locally and globally. Data science's ability to better solve health inequalities and enhance patient treatment

and increase outcomes has been noticed. The findings of this study showed the need for systematic adoption of data science software using geographical and predictive approaches to better promote further testing and technological usage in resolving health inequalities in public health through data science.

4.2 Essay 2: Using Data Science Software To Address Health Disparities*

4.2.1 Introduction

The application of data science in healthcare was studied by many professionals in the healthcare space to forecast its value and particular uses. Although data science is a beneficial tool for new knowledge and insights in healthcare, there exist challenges to its application in the domain. These challenges include data accuracy, missing data, and standardizing of data (Delaney & Westra, 2016). Although these are very important challenges to address, an important axiom to keep in mind is that the underlying information complexity to be achieved would have a major effect on the information system structure most appropriate for achieving the desired information outcome (Murphy, Murphy, Buettner, & Gill, 2015). In addition to these challenges, healthcare specialties such as the biomedical field have had challenges acquiring, sharing, and analyzing data (Dunn & Bourne, 2017). Therefore, data science in healthcare may in some ways be limited, but it is nonetheless useful to help solve significant and common healthcare problems. One such problem is that of health disparities found across healthcare organizations. Addressing health disparity issues allow health organizations to optimize patient

*Essay 2 is reproduced in its entirety from "Using Data Science Software to Address Health Disparities," International Journal of Big Data and Analytics in Healthcare, forthcoming, https://www.igi-global.com/article/a-novel-framework-of-health-monitoring-systems/268414, with permission from IGI Global.

care approaches and improve outcomes. Healthcare organizations can benefit through the impact data science software can have on their organizations and the multiple ways data science can lead to important findings in healthcare. For instance, the Covid-19 pandemic media coverage has reported mortalities among blacks in the United States at a higher rate compared to Caucasians (Shelby Lin Erdman, 2020). Is this due to disparities in socioeconomic issues and healthcare access that ultimately may lead to the mortality rate? While not the focus of the paper, it is important to recognize the potential that is driving the development and examination throughout the paper and where data science can offer some promise. The present study will review data science in relation to addressing health disparities in healthcare. It assesses data science software to examine the effectiveness of data science technologies that may be used to address problems such as health disparities.

4.2.2 Literature Review

Applications of data science are evident in numerous fields, ranging from research-based disciplines such as market, social, and census research to financial, technical, consulting, business, and media disciplines (Fayyad, 2012). The field of healthcare has begun to benefit from data science amid acquisition of new healthcare technologies. These new technologies also make available new opportunities for data science exploration, which can lead to intriguing discoveries from the data collected. For example, data science can be an important component of health informatics. Although viewed with some skepticism initially, health informatics has been embraced by the healthcare industry over time through vital investments in health information technology (HIT), increasing exploration of its utility (Detmer & Shortliffe, 2014). Data science may have similar adoption challenges, but as data begins to increase at a rapid

rate, embracing data science as a discipline and new technology will soon begin to make sense. For example, data science software can be important to clinicians because it can reduce unnecessary expenses in patient care, improve care quality and patient safety, and streamline the patient care process. Additionally, data science can help to determine the level of care or the level of care transitions that must occur for the well-being of the patient. Such information can come from new insights surfaced in the application of data science to patients' health improvement.

4.2.2.1 Data Production

Data science has come to cohere as a recognized field internationally, crossing numerous disciplines over decades, and evolving to respond to new data technologies (Liu et al., 2009; Press, 2013; Smith, 2006). As the healthcare field has met new data challenges in recent years, data science has offered powerful tools. It is of high value to note that big data and its application in the healthcare industry help to cut costs from analysis performed from electronic medical records (De la Torre Diez, Cosgaya, Garcia-Zapirain, & Lopez-Coronado, 2016). Such benefits have been made possible by innovations in managing large data sets. The importance of digital data for science is growing, and methods for analyzing these data need new data analytics (Westra, 2017).

The field of healthcare is witnessing an ever-increasing generation of large and complex data sets, commonly called *big data*, a term that functions as a shorthand for the diverse objects of data science (Rumbold & Pierscionek, 2017). Healthcare has experienced big data increase and therefore makes data science approaches to information promising. For example, in GIS applications, big data can boost monitoring of public health by combining spatial

variables and social health determinants (Zhang et al., 2017). More specifically, Allen, Tsou, Aslam, Nagel, & Gawron (2016) conducted a study that utilized geographical information systems (GIS) methodologies using data mined from social media platforms, leveraging techniques in machine learning, a component of data science, to filter through the data before analysis. Data science features a broad variety of techniques including mining text, visualizing data, geospatial modeling, machine learning, and predictive analysis. (O'Connor, 2018). Healthcare has seen the advancement of data science due to the following: big data, new data produced from sources that emerge from clinical trials and research, and the new technological capacities available for creating and deciphering data, whether structured or unstructured (Baptista et al., 2019). The industry of healthcare is positioning itself to retrieve valuable insights from data science technologies and processes, which help to produce noteworthy value, aiding in the significant utilization of data science methods and data science software for healthcare applications.

For example, information from electronic health records and other organizations such as the Center for Medicare and Medicaid Services (CMS) produce clinical data sets that allow for its use across multiple important settings in healthcare (Chase & Vega, 2016). Data sets can store information particular to the population, such as demographics, which can help aid in research when incorporating other factors such as income into the study. In turn, this can help researchers highlight gaps based on the subject matter content of the study (Chase & Vega, 2016). These types of health-centric data are necessary for healthcare data science applications, and there can be a significant improvement in the analysis of data. Organizations such as CMS can benefit from finding relevant and deep insights buried among the complexity

of variables and attributes that can exist in their data. Other healthcare organizations that work closely with CMS do so through multi-disciplinary aspects that exist in many forms, such as that of finance, management, and even policy; especially policy that can have a major impact on many healthcare disciplines that must adhere to CMS standards. For example, many of CMS' policies affect hospitals, providers, and the public. It is therefore imperative that these powerful organizations leverage data science for achieving better insights, especially since much of healthcare can stand to gain improvements from new policies set forth by organizations such as these.

4.2.2.2 Data Science and the Data Scientist

To fully tap the potentials of data science, the healthcare field must develop a sector of well-qualified data science specialists focused on healthcare data issues. The field of data science benefits from recruiting individuals that have unique data mining and analytical skills. Individuals that are interested in the field of data science should also have an in-depth understanding of data science techniques and concepts, especially in the domain of big data. Data science area concerns techniques for the extraction of information from various data, with a specific emphasis on 'Big' data displaying 'V' attributes such as veracity, value, variety, velocity, and volume (Maneth & Poulovassilis, 2016).

Data scientists possess an in-depth understanding of data science concepts and the necessary skill sets and knowledge to utilize data science techniques. There are a number of hallmarks of an effective data science practitioner, which should inform the successful future development of the healthcare data science sector. First, data scientists collect data, manipulate it in a tractable form, tell the tale and present the tale to others (Loukides, 2011). In

an effort to “traditionally” define the term *data scientist*, authors Liu, et al., (2009), proposed a tentative definition as a scientist committed to the study of data collection, analysis, metadata, rapid retrieval, archiving, sharing, mining to discover unexpected information and data relationships, two- and three-dimensional visualization including movement and management. Second, data scientists are normally familiar with toolkits popular in data science such as Python, Perl, R studio, Hadoop, SQL, machine learning software, and the like. Open source software, such as the R statistics kit, Python, and Perl are used by one in five data science professionals (Fayyad, 2012). Third, the data scientist benefits from artistic skills in the data science profession because it allows them to help paint a picture from the phenomena in the data (Loukides, 2011). A data scientist should have technical expertise, be curious and clever, and have the ability to tell a story through data (Patil, 2011). A data scientist should have the capacity to take an issue and incorporate multiple solutions for the different difficulties of the major problem at hand (Loukides, 2011). The skills necessary for a data scientist can vary in range. That is, a data scientist possesses skills acquired in computer science or mathematics.

Finally, in addition, a data scientist should be familiar with the four A’s of data, which are architecture, acquisition, analysis, and archiving. Ultimately, it is important to note that data scientists combine creativity with persistence, the desire to incrementally create data items, the ability to experiment and the ability to iterate on a solution (Loukides, 2011). Data scientists also benefit by skills in the following areas: a) the capacity to learn the application domain, b) the ability to communicate with data users, c) attentive insight into the big picture of a complex system, d) knowledge of how data can be represented, transformed and analyzed,

e) the capacity to visualize and present data, f) attention to quality, and g) ethical reasoning abilities (Stanton & De, 2013).

4.2.2.3 Models

Applications to healthcare must recognize that the essential components and processes of today's data science can be found in two generally accepted models. A data science project life cycle was proposed with 7 components, as follows: 1. acquisition of data, 2. preparation of data, 3. model and hypothesis building, 4. interpret and evaluate, 5. implementation, 6. operationalize, and 7. optimize (Manna, 2014).

Table 4.3: Data Science Analysis Process

Preprocessing Steps	• Goal Setting • Obtain Data • Data cleaning and formatting
Analysis Steps	• Data exploration and summary • Analytical methods • Modeling • Data automation and operationalization
Interpretation	• Presentation • Discussion and interpretation

Another model that shows an overview of the data science process was developed by Cielen et al. (2016), which propose six steps, as follows: 1. research goal setting, 2. data retrieval, 3. preparing the data, 4. exploring the data, 5. modeling the data, and 6. automating and presenting the data (Cielen et al., 2016). Table 4.3: delineates the steps that build upon that foundation.

Each major step in the data science process model is comprised of goals and other processes, each respective to their major step, as shown in Table 4.3. Data science utilizes

advanced methods to help determine predictions from the data used (Fayyad, 2012). Figure 4.2 shows the decisions that a data scientist undertakes when approaching data and the data scientist starts at the top of the figure making decisions that branch down to the granular level in each of the paths. As these two models are the dominant, organizing conceptual schema of the data science discipline, the development of healthcare data science applications must expect to map healthcare information needs onto their general outlines. Figure 4.2 developed from Cielen et al. (2016) delineates the data science process steps map components.

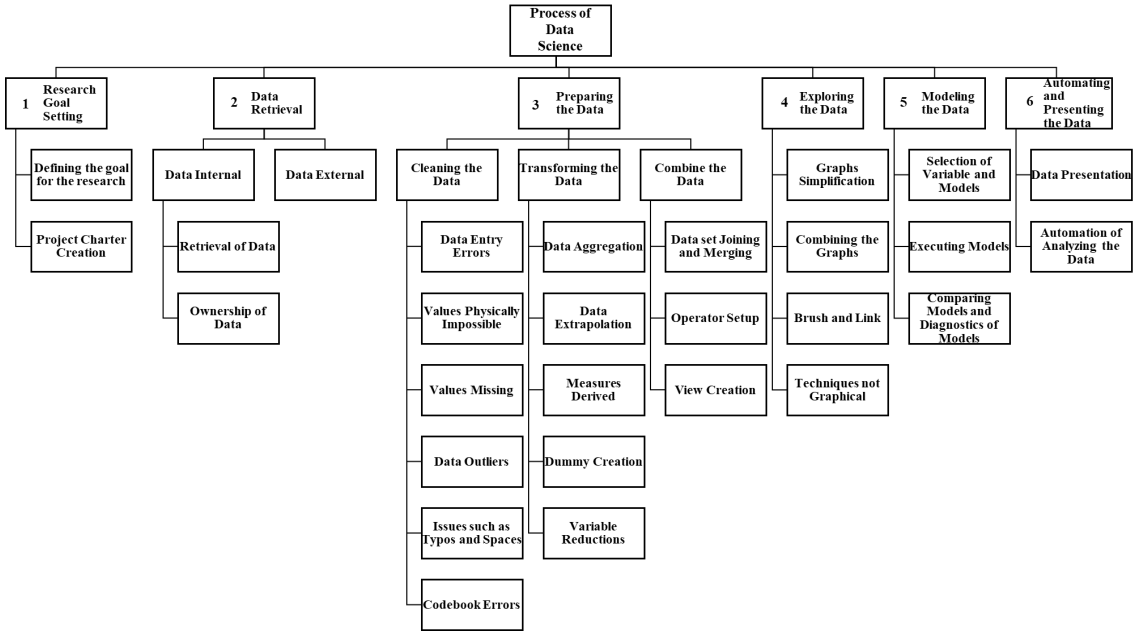

Figure 4.2: Data science process steps map (Cielen et al., 2016)

4.2.2.4 Data Science in Healthcare

A high demand for data scientists in the field of healthcare has emerged and in the last 10 years, the information collected in healthcare systems has increased, making Big Data in healthcare possible (De la Torre Diez, Cosgaya, Garcia-Zapirain, & Lopez-Coronado, 2016). In response, healthcare needs new models to make information fully meaningful and actionable. Data scientists can contribute new knowledge to building innovative solutions that ultimately

help all stakeholders in healthcare, from the patient to the treating physicians (Adam, Wieder, & Ghosh, 2017). Data science allows for the construction of data-driven theories conducive to advanced analytics in the healthcare field (Cao, 2017). One advantage of using data science processes, such as machine learning and graph analytics for deciphering big data, is that analyzing large health datasets can help in the prediction of patient outcomes. This, in turn, allows for the right clinical interventions to occur, and for new insights to surface for higher quality healthcare outcomes (Adam, Wieder, & Ghosh, 2017). Ultimately, the goal for data science in healthcare is to extract new insights that will support better decisions, leading to reduced costs and the improvement of targeted quality of care for patients (Adam, Wieder, & Ghosh, 2017).

Furthermore, data science can be applied to the integrated analysis of data across fields related to healthcare. For example, collaboration among disciplines such as healthcare, computing, and informatics can produce innovations in data-driven theory and data-driven economy (Cao, 2017). It is essential, however, that fully trained data scientists undertake the operation of data science software in such collaborations. In this way, data scientists can help in decision-making, and leaders working in the healthcare industry can benefit from the insights extracted by data scientists after careful analysis of their data (Power, 2016).

Health information and health data analysis have been central to the healthcare sector for many years. In most cases, before the electronic health record system era, patient data were being assessed by providers, but unfortunately, the analysis was limited due to the lack of technological capacity. As is the case today, providers' goal was to improve the health of patients, but that presented challenges, such as an overload of information that could possibly

be missed during initial assessment of the patient. This challenge ultimately helped to set the stage for the creation and use of electronic health record systems. Additionally, the United States Congress has been involved in marketing the use of health information technologies since 2004, when Congress began to introduce bills for the utilization of health information technologies (HIT) and electronic health information exchange systems (HIE) (Marchibroda, 2007).

Some states have made the use of such technology a top priority. This is an important step in healthcare, primarily because in the field of data science, most data come from a repository or database system of some sort. The state of New York has determined there are benefits to healthcare following full adoption of HIT and HIE. In 2006, in support of the state's hope for adoption by the healthcare community, the state of New York initiated the Healthcare Efficiency and Affordability Law for New Yorkers (HEAL NY), a grant-based program that focuses on three things: 1) electronic health record (EHR) adoption, 2) electronic prescribing (ePrescribe), and the development and implantation of clinical data exchanges throughout the community (Kern & Kaushal, 2007).

HIT has allowed for the collection of protected health information (PHI). Such information includes information surrounding socioeconomic status, sexual orientation, religion, location, race, ethnicity, gender, and mental health. Collection of such information can prepare for focused datasets that can allow for applications of data science to help determine disparities in health among types of groups in the dataset population.

4.2.2.5 Health Disparities

The quality of care and outcomes in health deteriorate when there are disparities in

elements such as socioeconomic status, race or ethnicity, all of which can be devastating and costly to public health. Outcomes in health can be affected not only by cultural ignorance and callousness by health practitioners, but more broadly by social and economic inequities within the habitat of the population (Demeester et al., 2017). Health dissimilarities or differences that are associated with disadvantages in social, economic, and environmental settings are known as health disparities.

People are typically affected negatively in their health because of the disparate challenges they encounter around race, religion, income status, gender, age, mental health, and the like (Office of Disease Prevention and Health Promotion, n.d.). Social disadvantages are usually associated with structured differences in the healthcare system that tend to lead to health disparities (West et al., 2017). For many years people in America have tended to suffer in their health due to disparities in income, education, race, and location. Recently, there has been an effort at local, state, and regional levels to reformulate healthy standards through various determinants of health efforts (Trujillo & Plough, 2016). The Institute of Medicine has deemed such inequalities in the services and outcomes provided by health organizations as key issues to address. Contributions to such health disparate circumstances are influenced by factors in the healthcare system, such as factors in that exist in the elements of culture, provider, and those of the patient (McQuaid & Landier, 2017).

In efforts to address health disparities, health organizations have intensified their approach to social determinants of health (SDOH). SDOH is defined by the World Health Organization (WHO) as conditions in living specific to a person's environment made up of components such as birthplace, habitat or neighborhood life, age, and other factors that

contribution to such conditions of living. The intensified approach by health organizations target lowering negative threats to health and focus on enhancing positive outcomes in health (Hughes et al., 2019). Healthcare is faced with excessive costs in healthcare services and such services can become wasteful, inefficient, and ineffectual due to the disparities that exist in health (King, 2016; Chin, 2016). Past studies examining disadvantaged groups have included the recognition of components that tend to influence disparities in outcomes and access in healthcare. Disparities in health permeate and continue in diverse type of infirmities and become expensive to health organizations.

4.2.2.6 Health Equity

Addressing health disparities through mitigation efforts leads to improving health equity (Anderson et al., 2018). Health equity can be defined as health excellence achieved through the eradication of disparities in health (Office of Disease Prevention and Health Promotion, n.d.). Therefore, in an effort to pursue improvements in health equity, the use of data science in healthcare should be to aid in the reduction of health disparities. The use of data science software can help analyze factors associated to health disparity. It can also aid healthcare organizations such as hospitals, clinics, provider practices, community, and public health officials find common health disparities that can help emphasize possible interventions for mitigation purposes. Although the following evaluation does not specifically treat healthcare data, it evaluates a number of software applications suitable to the kinds of data science operations healthcare organizations need to undertake to address issues such as health disparities.

4.2.3 Methods and Data Sources

KDnuggets is a top influential site for artificial intelligence, data science, and machine learning and has received numerous academic citations (KDnuggets, 2020). An assessment of data science software was conducted in the study using KDnuggets data. Figure 4.3 presents how several software products reflect the available programs in data science.

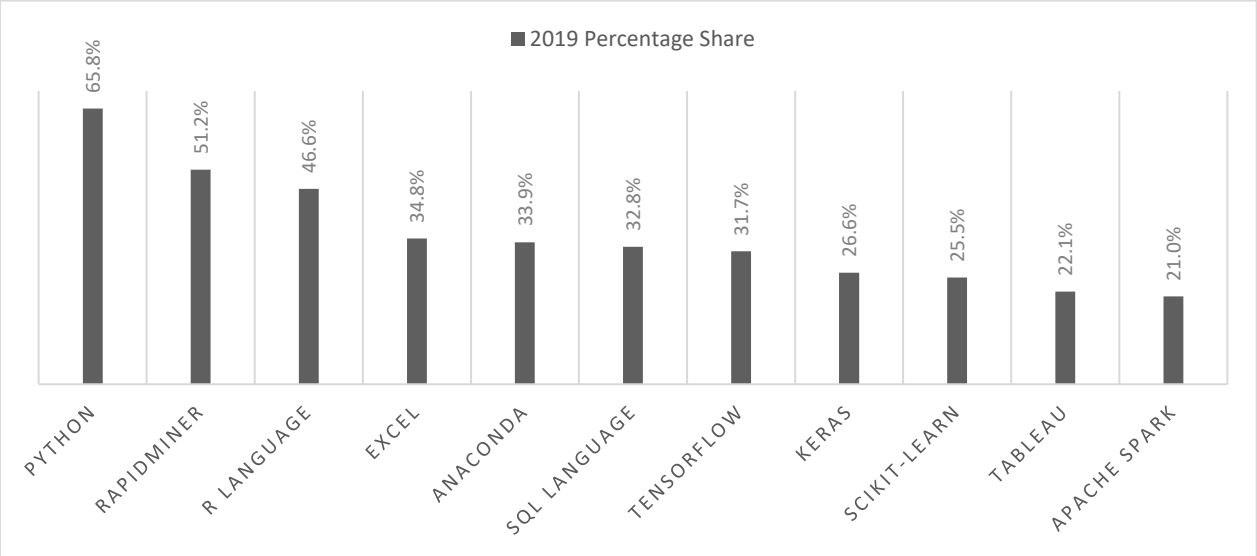

Figure 4.3: Software 2019 Percentage Share

Among the software listed, only software included in KDnugget's poll that categorically pertained to analytics, data science, and machine-learning software with a 30% or greater percentage share during the poll year 2019 were selected for the study. The poll conducted by KDnuggets sought to identify and measure utilization of analytical, data science, and machine learning software among the participants polled. The goal of the approach for this study is to use the top utilized software identified by KDnuggets to conduct an assessment of the criteria that should be present to leverage data science processes through the utilization of data science software that may be used to address health disparities. Data for Figure 4.3 is sourced from Piatetsky (2019)..

Table 4.4: Software Selection Sub-Criteria

Performance	Functionality	Auxiliary Task Support	Software Quality Characteristics	Critical Vendor Criteria	Software and Hardware Criteria
• Sturdiness • Time Behavior	• Openness • Completeness • Adaptability • Interoperability • Procedures • Security Levels • Simultaneous users • Big Data Processing • Data Sampling	• Data Cleansing • Data Filtering • Binning • Record Deletion • Handling Blanks	• Vertical Solution • Interface Type • DBMS Standard • Error Reporting • User Interface • Technique Suite • Graphic Capabilities • Data Visualization • Platform Independence • Platform Variety • Action History • Ease of Use • Domain Variety • Technical Source	• User manual & tutorial/training • Maintenance and upgrading • Consultancy • Product Established • Indirect Benefits	• Internal and external memory • Source Code

The study incorporated a software selection criteria framework based on the following criteria elements: performance, functionality, auxiliary task support, software quality characteristics, critical vendor criteria, and software and hardware criteria (Bhargava, 2013). Sub-criteria were modified in an effort to meet the needs for data science software assessments. Although this model was originally created for data mining software, we found the framework applicable to data science software. Table 4.4 delineates the sub-criteria assessed for each categorical segment of the data science data selection criteria framework.

Additionally, a project management software scoring model was adopted and the scoring criteria was modified to align with evaluation needs for data science software. The original software scoring model included performance indicators of poor (1), bad (2), good (3), and excellent (4) (Gharaibeh, 2014). Table 4.5 exhibits the new scoring model modified for data science software scoring.

Table 4.5: Data Science Software Scoring Model

Score	Performance	Condition
1	Poor	None
2	Ok	Partial
3	Good/Excellent	Full

4.2.4 Results

Table 4.6 exhibits the individual results for the data science software evaluation based on the following new framework criteria and sub-criteria. It is important to note that Anaconda is a distributor platform and does not necessarily compute data science algorithms. However, it is an important part of a data scientist's toolkit and can facilitate and integrate other important data science applications into its platform. This limitation resulted in Anaconda's score to be lower than the other data science applications evaluated.

Table 4.6: Data Science Software Selection Criteria Framework

Criteria	Criteria Group	Criteria Meaning	Python	RapidMiner	R Language	Excel	Anaconda	SQL Language	Tensorflow
Performance									
Sturdiness	Reliability	Performs without crashes	3	3	3	3	3	3	3
Time Behavior	Efficiency	Speed of computational results	3	2	3	2	2	3	3
Functionality									
Openness	Functional	Accessible for more development	3	3	3	1	2	3	3
Completeness	Functional	The extent of software required functions met	3	3	3	2	2	3	3
Adaptability	Functional	Customizable for industries or companies	3	3	3	2	2	3	3
Interoperability	Functional	Capacity to integrate with other applications	3	2	3	2	3	3	3
Procedures		Has suite of procedures for data science	3	3	3	3	2	2	3
Security Levels	Functional	Policy exists for security application of software such as identification of users and encrypting data	3	3	3	3	2	3	3
Simultaneous users	Functional	Can handle simultaneous users on the system	2	2	2	3	3	1	2
Data type Flexibility		Supports a variation of types of data	3	3	3	2	2	3	3
Big Data Processing		Capacity for processing high data volumes	3	3	3	1	2	3	3
Data sampling		Data sampling capacity at random for predictive models	3	3	3	3	2	3	3
Auxiliary Task Support									
Data Cleansing		Data modification of values for cleaning data	3	3	3	3	2	3	3
Data Filtering		Capacity to filter data based on a set of selections defined by user	3	3	3	3	2	3	3
Binning		Improved efficiency by allowing binning of data that is continuous	3	3	3	3	2	3	3
Record Deletion		Biased or unbiased record deletion capacity	3	3	3	3	2	3	3
Handling blanks		Blank handling capacity on entries	3	3	3	3	2	3	3
Software Quality Characteristics									
Vertical Solution	Personalization	Software package customized version to help meet specific industry requirements	3	3	3	3	3	3	3
Interface type	Personalization	Package type is user interface based	3	3	3	3	3	3	3
DBMS standard	Portability	Other types DB software packages such as SQL server and Oracle can be accessed by the software	3	3	3	3	3	3	3
Error reporting	Usability	Ability to message and report on errors	3	3	3	3	3	3	3
User interface	Usability	User interface ease of utilization	2	3	2	3	3	2	2
Technique Suite		Capacity to employ techniques such as time series and modeling	3	3	3	3	3	2	3
Graphic Capabilities		High graphic visualization quality for viewing such as decision trees	3	3	3	3	2	2	3
Data visualization	Usability	Effective data representation capacity	3	3	3	3	2	2	3
Platform Independence		Capacity to add other models and/or functionalities	3	2	3	3	3	2	3
Platform variety	Portability	Software can be used on a variety of platforms	3	2	3	3	2	3	3
Action history	Usability	In data science processes, software allows to modify action history	3	3	3	3	2	3	3
Ease of use	Usability	Users can easily learn and operate the software	3	2	2	3	3	2	2
Domain variety	Usability	Software is domain diverse and capable of being tailored to other industry for business problem solving	3	3	3	3	3	3	3
Technical Source	Opinion	Other vendors and in-house experts and consultants opinion on software	2	2	2	2	3	2	2
Critical Vendor Criteria									
User manual & tutorial/training	Vendor	Manuals, guidelines, tutorials, and other learning material available to users	3	3	3	3	3	3	3
Maintenance and upgrading	Vendor	Contracts and available for upgrades based on annual agreement as maintenance program	3	3	3	3	3	3	3
Consultancy	Vendor	Technical support availability to users	2	2	2	3	2	2	2
Product Established		Maturity of the software product	2	2	2	3	2	3	2
Indirect benefits	Benefits	Customer service improvement	1	2	1	2	1	1	1
Software and Hardware Criteria									
Internal and external memory	Hardware	Package run based on storage that is primary and secondary	2	2	2	2	3	2	3
Source code	Software	Source code availability	3	3	3	1	3	3	3

There are 6 total categories and a total of 38 sub-criteria categories. Table 4.7 shows the results for sub-criteria as it pertains to the total number of types of functionality met. Python and Tensorflow met the highest number of full functionality sub-criteria components, followed by R language, Rapidminer and Excel, SQL language, and Anaconda. Among partial functionality types, Anaconda had the highest met followed by Rapidminer, SQL language, Excel, R language, Python and Tensorflow. For those with no functionality, the highest number met was Excel followed by SQL language, a tie among Python, R language, Anaconda, and Tensorflow. Rapidminer had zero in this category.

Table 4.7: Number of Functionality Requirements Met

Type Functionality	Python	RapidMiner	R Language	Excel	Anaconda	SQL Language	Tensorflow
Full	31	27	30	27	17	26	31
Partial	6	11	7	8	20	10	6
None	1	0	1	3	1	2	1

Table 4.8 exhibits the overall scored results for each software ranked from highest to lowest.

Table 4.8: Top Ranked by Score Total

Software	Total Score
Tensorflow	106
Python	106
R Language	105
RapidMiner	103
SQL Language	100
Excel	100
Anaconda	92

The highest rank software programs in Table 4.8 indicate that Tensorflow and Python met the majority of the sub-criteria components. There were no major differences between tensorflow and python. R language was ranked second followed by RapidMinder, SQL language

and Excel, and Anaconda. There were no meaningful differences noted between the top four software ranked software based on their capacity to analyze structured and unstructured data. Although a powerful data extractor and data manipulator language, the SQL language in comparison to the four top-ranked software, did not fully meet the technique suite sub-criteria and lacked in data visualization capabilities. However, SQL should be integrated with software platforms to optimize data processes important to data science workflows. Excel showed to be a competitor among the software assessed but lacked in its capacity to fully allow big data processing and it is not considered an open source software limiting valuable contributions from the development community. As noted earlier, Anaconda is a distribution platform and acts as a gateway platform to multiple data science software. Although it scored the lowest due to only meeting partial functionality criteria through its capacity to integrate software to its platform, it is worth noting that it allows for better efficiencies and access to data science software.

4.2.5 Conclusion

The field of data science utilizes various methodological approaches for analyzing data in any domain or sector, including healthcare. The healthcare sector has not seen the full benefits of data science. However, this sector is beginning to dive into the field to explore new algorithms and methods that will aid in higher quality of care and quality outcomes. With the creation of new technologies and their capacities of creating data, possibilities into predicting probable outcomes based on historical data are now possible (Spruit & Lytras, 2018). Such innovations are especially likely, as this paper has argued above, in relation to healthcare sector networks connected through CMS and state initiatives such as HEAL NY.

The evaluation insights gained from this study based on the Data Science Software Selection Criteria Framework delineate how data science functionalities can help aid healthcare in approaching analytical processes with new analytical applications suitable for healthcare. For example, based on the highest ranked software in the study, Tensorflow and Python both have the capacity of automating and modeling the analysis of variables such as income, education, race, age, and cross-referencing such variables to outcomes in patient care and finance to determine outcomes that reveal health disparities. This paper documents a process that potentially can be used to address health disparities. Rankings should constantly be revisited due to advancements and development of new software and changes within the discipline of data science. Furthermore, contributions in this work allow the healthcare community to continually and iteratively evaluate data science software, as progressions are made, using the methods in this research.

This paper has demonstrated the data science capabilities through exhibiting the potential utility of leading software to perform the kinds of data science operations that can achieve improved care within such networks by addressing such factors as health disparities.

4.3 Essay 3: Application of GIS and SPSS for Prostate Cancer and Health Disparity Detection in Texas*

4.3.1 Introduction

4.3.1.1 Prostate Cancer in the United States

Prostate cancer has increased across the United States. During 2011–2015, prostate

*Essay 3 is reproduced in its entirety from "Application of GIS and SPSS for prostate cancer and health disparity detection in Texas," International Journal of Healthcare Technology and Management, forthcoming, with permission from Inderscience.

cancer was the most diagnosed cancer among males, followed by lung and bronchus cancer. The age-adjusted incidence of prostate cancer in the United States was 109.0 per 100,000, which consisted of 953,204 cases (U.S. Department of Health and Human Services, 2018) (see Table 4.9 and Figure 4.4).

Table 4.9: Top 10 Cancers by Rates of New Cancer Cases United States, 2011-2015

Cancer Type	Age-Adjusted Rate	Case Count	Population
Prostate	109.0	953,204	778,060,201
Lung and Bronchus	70.8	572,602	778,060,201
Colon and Rectum	45.1	365,934	778,060,201
Urinary Bladder	35.4	275,807	778,060,201
Melanomas of the Skin	27.3	219,303	778,060,201
Non-Hodgkin Lymphoma	22.8	182,273	778,060,201
Kidney and Renal Pelvis	22.1	184,358	778,060,201
Leukemias	17.7	139,112	778,060,201
Oral Cavity and Pharynx	17.6	151,268	778,060,201
Pancreas	14.4	117,201	778,060,201

Source: https://gis.cdc.gov/Cancer/USCS/DataViz.html U.S. Cancer Statistics: The Official Federal Cancer Statistics, Centers for Disease Control and Prevention. Rate per 100,000 men

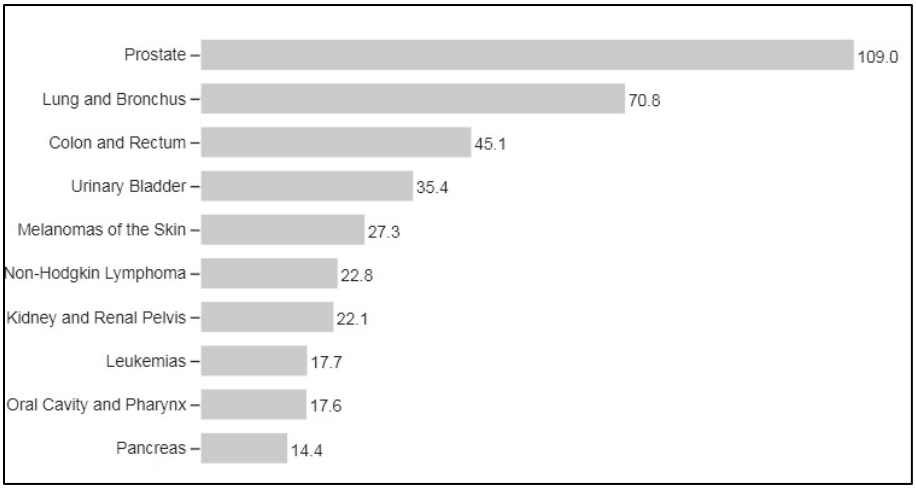

Figure 4.4: Top 10 cancers by rates of new cancer cases

Source: https://gis.cdc.gov/Cancer/USCS/DataViz.html. U.S. Cancer Statistics: The Official Federal Cancer Statistics, Centers for Disease Control and Prevention

Prostate cancer is prevalent in the US, and the analysis of spatial patterns of prostate cancer distribution, along with an examination of their changes over time, promises significant insights into how the disease spreads geographically over time. Tools that analyze such patterns have helped determine the spatial distribution of disease and its geographic spread (Bui and Pham, 2016). This study examines the geography and spatiotemporal patterns of prostate cancer in Texas counties, drawing upon county-level, mortality-rate data during the decade 1999–2009. Through the utilization of Statistical Package for the Social Sciences (SPSS) software and geographical information system (GIS) technologies, the study leverages analytical, mapping, and visualization functionalities, providing new insights that can help explain health disparities in prostate cancer. Other GIS-like applications, such as web-based spatial processing tools, have been used successfully in other countries to measure spatial patterns in an effort to track disease incidences (Bui and Pham, 2016). In addition, social determinants of health such as race, socioeconomic status (SES), and healthcare accessibility (Wilkinson and Marmot, 2003) are evaluated in an attempt to explain the existence and geographical distribution of the disease.

Other studies analyzing the role of social determinants in healthcare studies (Shulan, Gao, and Moore, 2015) have shown contradictory results regarding social determinants such as SES in explaining health disparities in prostate cancer, as well as its prevalence (Cheng et al., 2009). Nevertheless, these factors remain important to study further in cancer because in other geographic settings, the prevalence of malaria and other diseases have been found to be correlated with environmental and socioeconomic factors (Bui and Pham, 2016). In this study, the term *health disparities* refers to the differences found in incident cases, deaths, and

healthcare access due to variables such as socioeconomic status, settlement or habitation, gender, or ethnic and racial makeup (LaVeist and Pierre, 2014).

To mitigate the problem of health disparities, the US Office of Disease Prevention and Health Promotion initiative Healthy People 2020 has determined to track rates for the following components of disease: illness, mortality, long-lasting conditions, and other factors in health outcomes that may correlate with factors such as race and ethnicity, gender, geographic location, and the like ("Disparities"). This study advances that initiative, focusing particularly on delineating and understanding prostate cancer mortality geographically in the state of Texas in relation to social determinants such as race, socioeconomic status, and healthcare access. The study specifically addresses how such factors influence disparities in the disease. Although gender is a social determinant, it was not evaluated because prostate cancer affects only males. The study attempts to answer the following research questions: 1) What is the geographic distribution of prostate cancer deaths across Texas? 2) Why are prostate cancer deaths geographically distributed in that way? 3) How does the geographic, spatial-temporal pattern of the disease change over time?

4.3.2 Background

Prostate cancer affects the prostate gland cells, usually in the form of high cell-growth rate, and the risk of prostate cancer increases with age (Klassen and Platz, 2006). In the United States, one in six men over the age of 50 will be diagnosed with prostate cancer (Penson and Chan, 2007). It is number two in both incidence and mortality (Figure 4.7 and Table 4.10). During 2011–2015, about 953,204 new cases of prostate cancer were reported, and 140,086 men died in the United States (Figures 4.5 and 4.6).

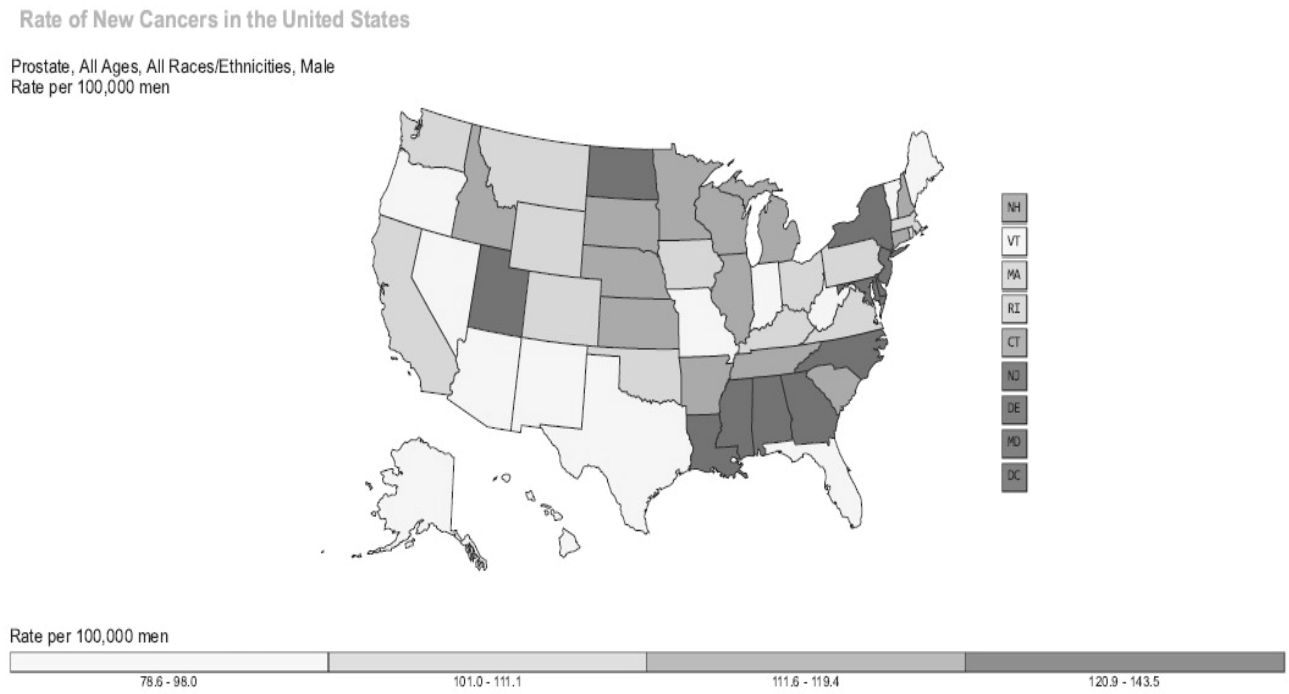

Figure 4.5: Rate of new cancers in the United States

Source: https://gis.cdc.gov/Cancer/USCS/DataViz.html. Centers for Disease Control and Prevention, U.S. Cancer Statistics: The Official Federal Cancer Statistics

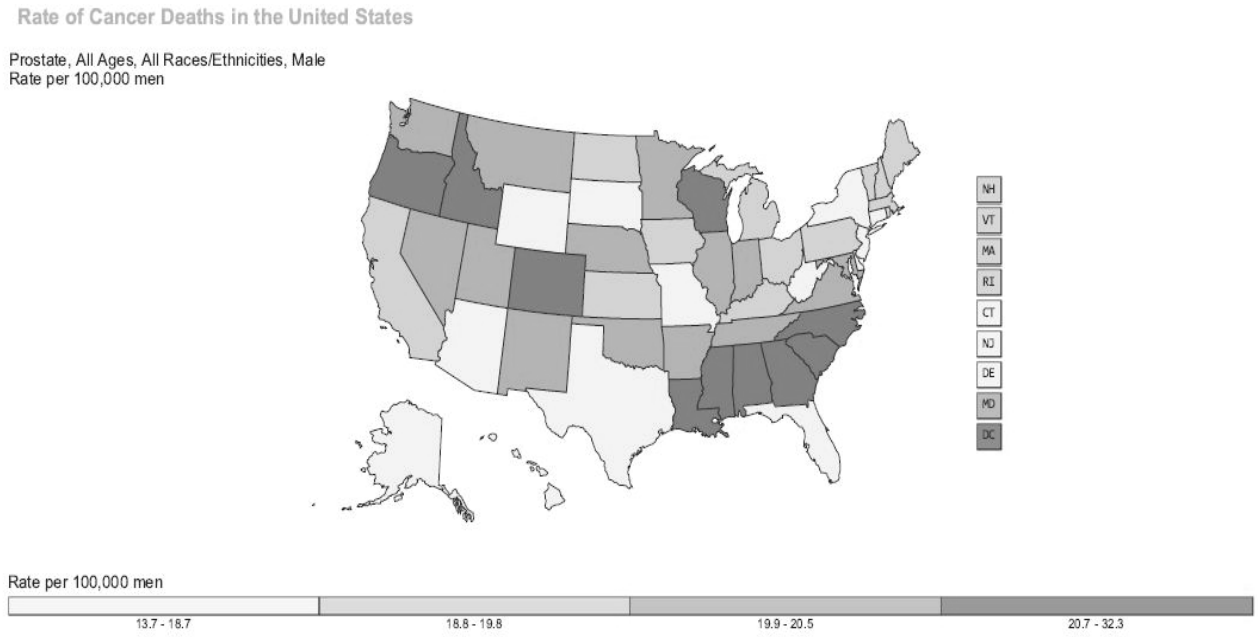

Figure 4.6: Rate of cancer deaths in the United States

Source: https://gis.cdc.gov/Cancer/USCS/DataViz.html. U.S. Cancer Statistics: The Official Federal Cancer Statistics, Centers for Disease Control and Prevention

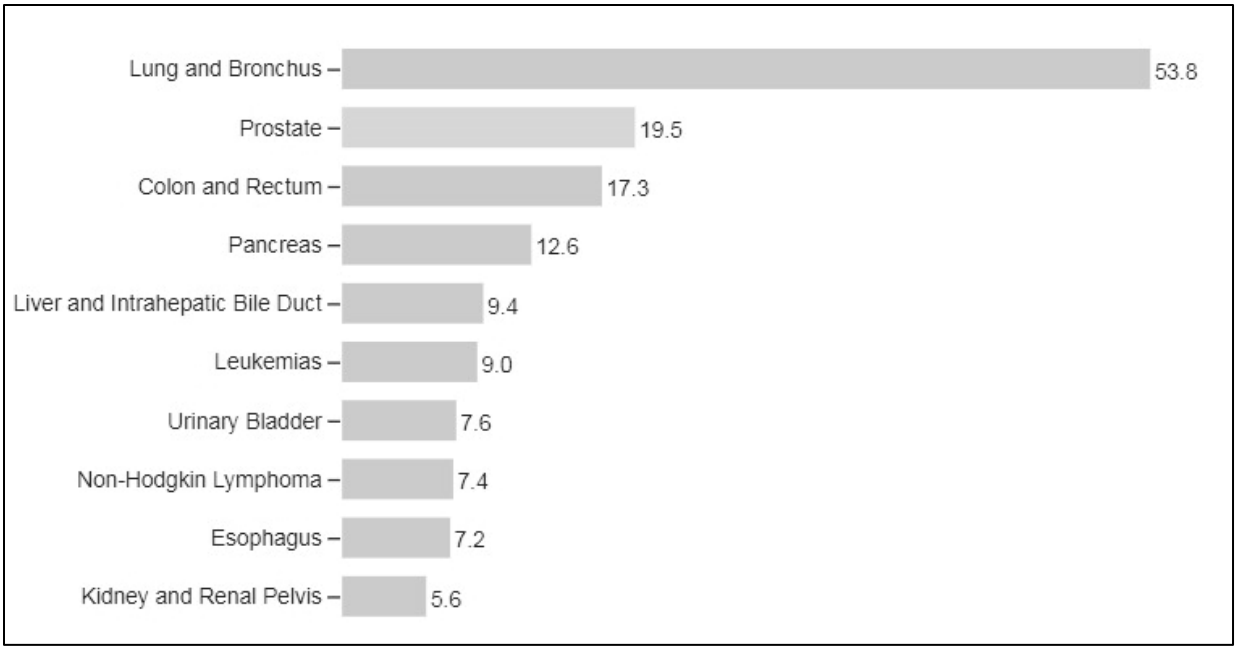

Figure 4.7: Top 10 cancers by rates of cancer deaths United States

Source: https://gis.cdc.gov/Cancer/USCS/DataViz.html. U.S. Cancer Statistics: The Official Federal Cancer Statistics, Centers for Disease Control and Prevention

Table 4.10: Top 10 Cancers by Rates of Cancer Deaths United States, 2011-2015

Cancer Type	Age-Adjusted Rate	Death Count	Population
Lung and Bronchus	53.8	427,587	778,060,201
Prostate	19.5	140,086	778,060,201
Colon and Rectum	17.3	135,542	778,060,201
Pancreas	12.6	100,599	778,060,201
Liver and Intrahepatic Bile Duct	9.4	80,526	778,060,201
Leukemias	9.0	67,201	778,060,201
Urinary Bladder	7.6	55,652	778,060,201
Non-Hodgkin Lymphoma	7.4	56,402	778,060,201
Esophagus	7.2	59,082	778,060,201
Kidney and Renal Pelvis	5.6	45,076	778,060,201

Source: https://gis.cdc.gov/Cancer/USCS/DataViz.html. U.S. Cancer Statistics: The Official Federal Cancer Statistics, Centers for Disease Control and Prevention. Rate per 100,000 men.

4.3.2.1 Prostate Cancer in Texas

The Texas Cancer Registry (TCR) is a population-based registry that provides data and cancer measures. TCR has technical and functional capacities leveraging geographical maps. This registry has assisted in trending prostate cancer morbidity and mortality data [Texas Department of State Health Services (TDSHS) (a) and (b)]. According to the Texas Health and Human Services Cancer Registry dataset [TDSGS (a)], during 2011–2015, there was an annual average of 11,572 new cases of prostate cancer in the state of Texas. Non-Hispanic Whites comprised an annual average of 7,367 incidents registered. Blacks comprised 1,807. Asian/Pacific Islanders comprised 188. American Indian/Alaska Natives comprised 30. Hispanics comprised an annual average of 2,059. Overall, prostate cancer was more conspicuous among those aged 50 and older [TDSHS (a)]

As in the United States as a whole, of new cancer cases, prostate cancer was the number one cancer in Texas, followed only by lung and bronchus cancer, but the age-adjusted rate in Texas was slightly lower than the national rate, at 95.4 per 100,000 population, which consisted of 57,860 cases (Table 4.11 and Figure 4.8).

Table 4.11: Top 10 Cancers by Rate of New Cancer Cases United States, 2011-2015

Cancer Type	Age-Adjusted Rate	Case Count	Population
Prostate	95.4	57,860	65,783,771
Lung and Bronchus	65.5	36,272	65,783,771
Colon and Rectum	45.7	26,655	65,783,771
Urinary Bladder	26.9	14,150	65,783,771
Kidney and Renal Pelvis	24.4	14,675	65,783,771
Non-Hodgkin Lymphoma	21.3	12,208	65,783,771
Melanomas of the Skin	17.8	10,173	65,783,771

Cancer Type	Age-Adjusted Rate	Case Count	Population
Leukemias	17.5	9,986	65,783,771
Liver and Intrahepatic Bile Duct	17.2	11,048	65,783,771
Oral Cavity and Pharynx	16.8	10,481	65,783,771

Source: https://gis.cdc.gov/Cancer/USCS/DataViz.html. U.S. Cancer Statistics: The Official Federal Cancer Statistics, Centers for Disease Control and Prevention. Rate per 100,000 men.

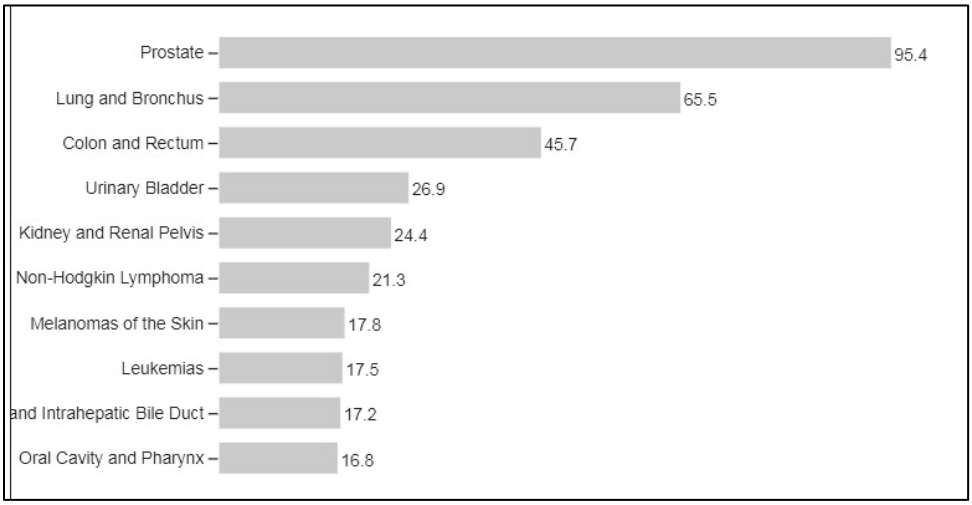

Figure 4.8: Top 10 cancers by rates of new cancer cases Texas

Source: https://gis.cdc.gov/Cancer/USCS/DataViz.html. U.S. Cancer Statistics: The Official Federal Cancer Statistics, Centers for Disease Control and Prevention

The Texas Health and Human Services Cancer Registry dataset [TDSHS (b)] indicates that during 2011–2015, there was an annual average of 1,695 deaths. Prostate cancer mortality rates were higher among Blacks (35.7), followed by Whites (17.3), Hispanics (15.2), Non-Hispanic Asian/Pacific Islanders (7.9), and American Indian/Alaska Natives (5.8). The higher rates of prostate cancer mortalities recorded were among age 50 and older. Among those aged 50–59, prostate cancer mortalities were almost six times higher than those aged 40–49. The age-adjusted rate for prostate cancer deaths in Texas was 18.1, which consisted of 8,519 deaths (Figure 4.9 and Table 4.12).

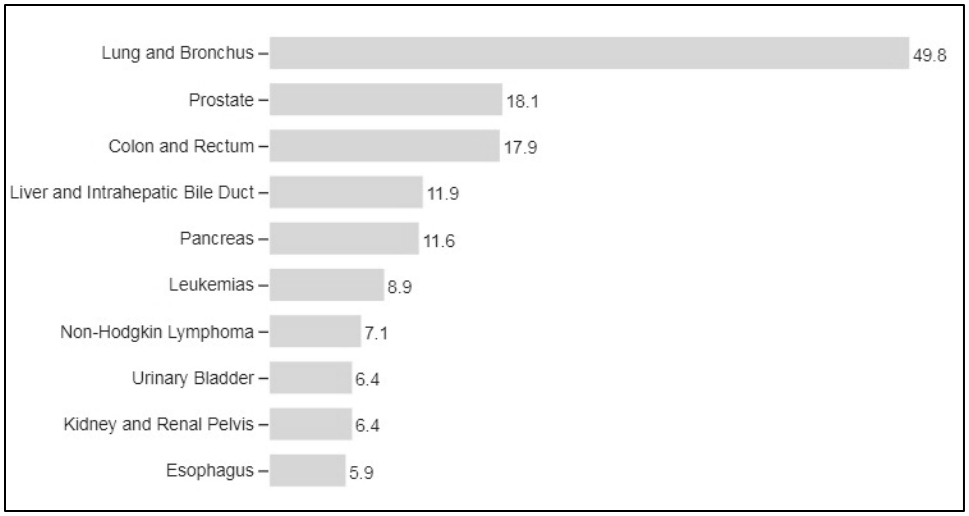

Figure 4.9: Top 10 cancers by rates of cancer deaths Texas

Source: https://gis.cdc.gov/Cancer/USCS/DataViz.html. U.S. Cancer Statistics: The Official Federal Cancer Statistics, Centers for Disease Control and Prevention

Table 4.12: Top 10 Cancers by Rates of Cancer Deaths United States, 2011-2015

Cancer Type	Age-Adjusted Rate	Death Count	Population
Lung and Bronchus	49.8	26,917	65,783,771
Prostate	18.1	8,519	65,783,771
Colon and Rectum	17.9	9,897	65,783,771
Liver and Intrahepatic Bile Duct	11.9	7,267	65,783,771
Pancreas	11.6	6,432	65,783,771
Leukemias	8.9	4,579	65,783,771
Non-Hodgkin Lymphoma	7.1	3,663	65,783,771
Kidney and Renal Pelvis	6.4	3,558	65,783,771
Urinary Bladder	6.4	3,105	65,783,771
Esophagus	5.9	3,357	65,783,771

Source: https://gis.cdc.gov/Cancer/USCS/DataViz.html. U.S. Cancer Statistics: The Official Federal Cancer Statistics, Centers for Disease Control and Prevention. Rate per 100,000 men.

4.3.2.2 Geographic Impact and Social Determinants in Health

Geography is essential to understanding disease and its spread. The geographical connection between people and their environments is comprised of many components that

affect the social, economic, and physical aspects of people's lives (Klassen and Platz, 2006). Social determinants such as race, environment, and socioeconomics have been determined to have correlations to health (LaVeist and Pierre, 2014). The present study articulates three main hypotheses to help validate the involvement of social determinants related to health disparities in prostate cancer.

4.3.3 Hypotheses

4.3.3.1 Hypothesis 1

There is a relationship between minority race and the geography of prostate cancer mortality in Texas. A positive relationship is hypothesized between the percentage of African Americans and the age-adjusted death rate (AADR) of prostate cancer. The same is expected for the percentage of Hispanics and the percentage of other races combined.

4.3.3.2 Hypothesis 2

There is a relationship between socioeconomic status (income level) and prostate cancer mortality. A negative relationship is hypothesized between income and age-adjusted death rate (AADR) prostate cancer mortality. A positive relationship is hypothesized between healthcare costs, unemployment, uninsured adults and age-adjusted death rate (AADR) prostate cancer mortality. That is, as each individual variable increases, AADR prostate cancer mortality also increases.

4.3.3.3 Hypotheses 3

There is a relationship between healthcare access and prostate cancer mortality. A negative relationship is hypothesized such that an increase in healthcare access, namely access

to primary care physicians, results in a decrease in prostate cancer mortality.

4.3.4 Literature Review

4.3.4.1 Prostate Cancer

Three factors are considered risks for prostate cancer: age, race, and heredity (Attard, et al., 2016). These factors are also important to assessing the distribution of the disease geographically. That is, the variables may be used to map out incidences and mortalities of prostate cancer using geographical components, such as demographics, to visualize where the disease occurs. For example, a person that has the disease can be categorized in a specific ethnic group or race, and race can be used to help determine the spread of the disease through the geographical space of interest. Race and ethnicity are based on genetic variations inherited from an individual's parents (Pearce, Foliaki, Sporle, and Cunningham, 2004). With these genetic variations, individuals can inherit other traits that increase the risk of diseases such as prostate cancer (Rebbeck, 2017). For example, prostate cancer is dominant in black men of African lineage. In fact, the highest mortalities occur among men with Afro-Caribbean and sub-Saharan African descent. In 2008–2011, the mortality rate for black men was 43 per 100,000, followed by whites (19.8), Hispanics (17.8), and Asians/Pacific Islanders (9.4) (Rebbeck, 2017). Therefore, because the disease has high hereditary risks, race/ethnicity is an important factor to study in prostate cancer.

4.3.4.2 Health Disparities

Health disparity exists when health differences impact a group negatively. Measures of health disparity can encompass the portion of a population affected by disease, its severity, its

symptoms, and its mortalities ("Disparities in Health and Healthcare: Five Key Questions and Answers"). Differing healthcare access and capacity to receive disease screening are important factors ("Health Disparities: MedlinePlus"). Such disparities effect racial and ethnic minorities most severely, but socioeconomics, gender, age, geography and disability are also significant components (Braveman, 2014; Kumar et al., 2018). Healthy People 2020 highlighted the effects of these and other such forces, including mental health and religion ("Disparities").

4.3.4.3 Social Determinants of Health

Social determinants of health involve conditional settings and circumstances that affect populations. They may include factors such as birthplace, where people grow up, where they live, and where they work. Other factors include age, money, and political conditions. Social determinants can negatively affect health, producing health inequalities, which can be defined as unfairness in the health status of individuals or groups ("About Social Determinants of Health").

There are five main categories in social determinants of health: economic stability, education, health, community, and neighborhood. Elements such as impoverishment and deprivation, stability in the housing market, security of food, and the stability and quality of employment make up economic stability. A degree from high school or higher education, childhood education access that is of high caliber, and the ability to read are components of education, and any of these can affect health results. Elements of health include health insurance, literacy in relation to health terms or health literature, and healthcare access. Discrimination, workplace conditions, civic participation, community belonging, and incarceration are elements of community. Quality of air and water, neighborhood safety, access

to healthy foods, transportation, and housing are neighborhood elements. Anytime these are deficient, health is affected negatively (Henry, 2019).

The social ladder can determine the life expectancy for groups. In this study, the term *social ladder* is a shorthand to describe a population's socioeconomic status, encompassing the typical social determinants that accompany a particular position. The lower on the social ladder, the greater risk for disease, and the lower expectancy for lifespan. Impoverishment in the social and economic dimensions of peoples' lives can affect quality of health. For this reason, it is imperative that health policies focus on changing negative outcomes surrounding determinants of health. Disadvantages can come in the form or poor education, employment insecurities, limited job mobility, poor housing conditions, family responsibilities in difficult social conditions, and insufficient retirement resources. The often unmanageable stress and anxiety that such disadvantages produce can lead to compromised health or early death. Diet and food supply are important factors shaped by social ladder position. Food shortage can lead to diseases of malnutrition, while overeating can lead to diseases such as cancer and diabetes. Security and satisfaction in employment can be conducive to good health, while high unemployment rates are associated with sickness and early death. In addition, financial difficulties stemming from employment insecurity can generate psychological stressors that affect health negatively (Wilkinson and Marmot, 2003).

4.3.4.4 Healthcare Access

Healthcare access is an important factor in prostate cancer mortality. Access to healthcare facilities that have the clinical and technical capacities to conduct prostate-specific antigen (PSA) testing for early detection and treatment are imperative to fighting the disease (Major et

al., 2012). The availability of primary care physicians or providers that specialize in prostate cancer is also important because a disproportionate availability of prostate cancer specialists throughout a geographic area is directly correlated to the mortality rate (Kim et al., 2017).

To help facilitate the discovery of healthcare access among researchers, geographic information systems can help to produce maps, not only of the disease but also of other important healthcare variables. This type of mapping can be a simple choropleth map, which is normally used to visualize the geography of variables such as healthcare accessibility (Sherman et al., 2014).

4.3.4.5 Mode of Spread

The geographic distribution of prostate cancer has been analyzed to determine the mode of spread across geographic regions. Typically, datasets from two or more time periods are used to analyze how the spatial pattern of the disease is changing over time. Other factors that are taken into consideration in the geographic distribution analysis are location, at-risk population, number of cases or incidences, and sometimes an age-adjusted rate (Gregorio et al., 2004). Tracking a spatiotemporal pattern can help researchers understand where the disease occurs and why. The findings of such studies can then help to pinpoint other factors that may be associated with increase or decrease of the disease. For example, prostate cancer mortalities may decrease as healthcare access and rate of PSA screening increase.

4.3.5 Methodology and Data Sources

The International Classification of Diseases, Tenth Revision (ICD-10-CM) code used for the study is C61, defined in this study as prostate cancer. The ICD-10-CM is based on a

classification logical system that helps healthcare practitioners and healthcare workers identify disease through coding processes to determine disease diagnosis (https://www.cdc.gov/nchs/icd/icd10cm.htm). GIS-produced choropleth maps were used to visualize the geography of the variables. Two main datasets were obtained for the study. Both datasets were specific to Texas. The first dataset was obtained through VitalWeb, an online user-intuitive website that houses large and complex datasets and leverages the health data analysis software Vitalnet to conduct analysis (https://www.ehdp.com/vitalnet/overview.htm). The dataset obtained was comprised of age-adjusted prostate cancer mortality data per 100,000 from 1999 to 2009, and the variables used for the study were name of county (Name) and age-adjusted death rate (AADR). The calculated state average was 23.0. The age-adjustment was set to the 2000 U.S. Population standard.

The second dataset contained the explanatory variables, or social determinants, and was obtained from the 2012 Texas Health Rankings published dataset. The published data was provided and compiled by the County Health Rankings & Roadmaps program, which was created as a central hub for reliable community health data (http://www.countyhealthrankings.org/about-us).

The explanatory variables used to address Hypothesis 1 were % African American, % Hispanic, and % other combined races. The variable “other combined races” is defined as % American Indian and Alaskan Native, % Asian, and % Native Hawaiian/other Pacific Islander. The source of data was U.S. Census Bureau records for 2009. The compiled dataset was most applicable to explain prostate cancer mortality for 1999–2009, the time period used in the study. Additional races could not be assessed and are addressed in the limitations section of

this study. A Pearson bivariate correlation analysis was conducted and analyzed against the AADR prostate cancer variable obtained from the VitalWeb dataset.

Variables analyzed for Hypothesis 2 were median household income, % uninsured adults, healthcare costs, and % unemployed. The variables with their respective data sources and data year were: median household income (Small Area Income and Poverty Estimates [SAIPE], 2010); % uninsured adults (Small Area Health Insurance Estimate [SAHIE], 2009); healthcare costs (Health Resources and Services Administration [HRSA], 2007); and % unemployed (Local Area Unemployment Statistics, Bureau of Labor Statistics, 2010). Pearson's statistical bivariate correlation was applied to analyze the relationship between the variables and AADR prostate cancer.

To address Hypothesis 3, the variable of primary care physicians was used. The source of the data was Heath Resources and Services Administration, Area Resource File (ARF) for 2009. This variable was used to determine healthcare access to counties. As in the statistical analysis of Hypothesis 1 and 2, a Pearson bivariate correlation analysis was conducted between PCP rate and AADR prostate cancer to measure the relationship between both variables.

4.3.6 Results

4.3.6.1 Brief Notable Findings

Counties with the highest mortality rates and above the Texas state mean (23.0) were Menard (35.0), Crockett (34.9), Dimmit (43.2), Refugio (38.3), Harrison (26.5), Delta (35.7), Jack (38.5), Haskell (37.0), King (39.7), Garza (38.7), Floyd (44.2), Bailey (43.2), Cochran (43.2), and Mitchell (34.3). Similarly, very high concentrations showing high mortality rates were found in smaller, dispersed clusters, mainly in northwestern Texas, exemplified by Bailey (43.2), Cochran

(43.2), Floyd (44.2), King (39.7), Haskell (37), Garza (38.7), and Mitchell (34.3) counties. In contrast, counties with the lowest rates of prostate cancer mortality and below the state mean were Loving (0), Glasscock (0), Oldham (0), Roberts (0), Briscoe (0), Kenedy (0), Zapata (6.8), La Salle (4.1), and Edwards (5.3), located in the panhandle and southern Texas.

The eastern region showed Harrison, Rusk, Cherokee, Anderson, Houston, Shelby, Nacogdoches, Sabine, Jasper, Tyler, Hardin, Jefferson, Liberty, Chambers, Galveston, Brazoria, Matagorda, Wharton, Colorado, Washington, Waller, Robertson, and Falls counties with high concentrations, and Panola county with a very high concentration. Likewise, AADR ranged above the state mean and was between 25.6 to 33.5 in western counties such as El Paso, Reeves, Ward, and Pecos. Brewster showed high concentrations, and Crockett showed a very high concentration. Additionally, high concentrations were also found in Dallam, Hansford, Wheeler, Donley, Hall, Swisher, Lamb, Crosby, Lynn, Dawson, Gaines, and Stonewall counties. The central region had high concentrations among Wise, Palo Pinto, Parker, Eastland, Erath, Comanche, Brown, Mills, Coryell, Runnels, Concho, Kimble, Sutton, and Gillespie, with very high concentrations showing for Jack and Menard counties. Eastern Texas had more counties with a greater concentration of prostate cancer in comparison to the western and central regions.

4.3.6.2 Change Map Analysis

The main time period of 1999–2009 was broken into two portions for further analysis in this section. Time Period 1 was 1999–2004 (see Figure A.11), and Time Period 2 was 2005–2009 (see Figure A.12). Prostate cancer mortality rates during Time Period 1 were mostly moderate for the eastern region. However, Time Period 2 showed a heavier presence for the eastern region.

In an effort to review the changes in the disease mortality rate through time, a new map was created, namely a choropleth map charting the difference between the two time periods. The new change map showed the areas where mortality rates tended to improve and areas where the disease tended to worsen (see Figure A.13).

The greatest improvements between Time Period 1 and Time Period 2 were around the Panhandle region in counties such as Dallam, Cochran, and Borden. In the Permian Basin, counties such as Crane and Reagan also showed great improvements. Additionally, there were improvements in the southeastern region such as in Goliad County.

Areas that tended to worsen over both time periods were in the western region, in counties such as Hudspeth and Brewster. The Permian Basin region also showed Ward County worsening. Counties surrounding the Panhandle region that worsened were Sherman, Hartley, and Hansford. The north-central region showed Stephens, Shackelford, and Throckmorton counties worsening over the two time periods.

Overall, the majority of the regions showed slight improvements. Regional improvements appeared in areas of western Texas such as El Paso County, areas of the Permian Basin such as Andrews and Ector counties, areas of the Panhandle region such as Ector and Moore counties, areas of the southern region such as Starr and Cameron counties, areas of the Eastern region such as Harris and Fort Bend counties, areas of the North region such as Cook and Grayson counties, areas of the North East region such as Bowie and Cass counties, and areas of the central region such as San Saba, Mason, and Llano counties.

4.3.6.3 Race/Ethnicity and Prostate Cancer Mortality

The overall age-adjusted death rate (per 100,000) of prostate cancer from 1999 to 2009

for all races combined, including Whites, was 9.5, a total of 18,315 deaths across the state (Table 4.13). Based on this data, there were 9.5 deaths per every 100,000 population. Blacks had the highest rate among the group at 18.8, followed by Whites (9.0), Hispanics (7.9), and Others (3.7).

Table 4.13: Death Rate for Prostate Cancer, 1999-2009, in Texas, by Race

Race	Age-Adjusted Rate	Deaths
White	9.0	12,456
Black	18.8	3,059
Hispanic	7.9	2,646
Other	3.7	154
Total	9.5	18,315

Source: https://www.ehdp.com/vn/rw/txu1/eg2/8a17xgqn-tbl.htm. Texas VitalWeb ICD-10 Underlying Cause of Death. Rate per 100,000 men.

The explanatory variables from the Health Rankings dataset were used to analyze several minority group populations to help explain the high concentrations of prostate cancer mortality and its geography. These groups were African Americans, Hispanics, and Other Races Combined, which were American Indian, Alaskan Native, Asian, Native Hawaiian, and Other Pacific Islanders.

4.3.6.4 Analysis of African Americans

The first minority group analyzed in relation to geographic concentrations of prostate cancer mortality were African Americans. There was a high to very high population rate and concentration of African Americans in the eastern part of Texas (Figure A.3). A comparison of the geography of the percentage of African Americans (Figure A.3) and the geography of prostate cancer (Figure A.1) showed a visual similarity between the two. That is, for eastern

Texas, there was a high concentration of Blacks, which appeared geographically similar to the high concentrations of prostate cancer. Counties with percentage of Blacks to AADR (per 100,000) are shown in Table 4.14.

Table 4.14: % of Blacks to AADER (per 100,000)

Counties	% of Blacks	AADR
Jefferson	34.7	31.3
Waller	26.0	28.3
Houston	25.9	32.0
Anderson	22.5	27.9
Harrison	22.4	26.5
Robertson	22.2	29.4
Rusk	18.1	31.3
Shelby	18.0	29.3
Washington	17.7	27.4
Jasper	17.2	30.9
Nacogdoches	16.6	28.4
Wharton	14.7	32.8
Cherokee	14.6	30.5
Mitchell	14.4	34.3
Galveston	14.3	27.9
Colorado	14.2	27.7
Liberty	12.1	29.0
Matagorda	11.8	26.2
Tyler	11.8	26.8
Brazoria	11.4	26.9
Chambers	10.8	29.4
Sabine	9.3	27.4
Cochran	7.2	43.2
Hardin	7.2	29.4
Garza	6.2	38.7
Floyd	4.6	44.2

Counties	% of Blacks	AADR
Haskell	4.3	37.0
King	3.8	39.7
Bailey	2.0	43.2

In contrast, other pockets showing very high concentrations of prostate cancer did not show high concentrations of Blacks for the region. Examples were Bailey (2.0% of Blacks to 43.2 AADR), Cochran (7.2% of Blacks to 43.2 AADR), Floyd (4.6% of Blacks to 44.2 AADR), King (3.8% of Blacks to 39.7 AADR), Haskell (4.3% of Blacks to 37 AADR), Garza (6.2% of Blacks to 38.7 AADR), and Hardin (7.2% of Blacks to 29.4 AADR) county.

4.3.6.5 Analysis of Hispanics

The second minority group analyzed in relation to the geography of prostate cancer mortality was Hispanics. There was a high to very high concentration of Hispanics in the western, northwestern, and southern regions of Texas, especially along the Mexico-Texas border-state line (Figure A.4). A comparison of the southern geography of the percentage of Hispanics (Figure A.4) and the geography of prostate cancer (Figure A.1) did not show a visual similarity between the two. In fact, prostate cancer mortality in south Texas showed the lowest AADR (Figure A.1), while Hispanics were a dominant group in the region (Figure A.4). Counties with percentage of Hispanics to AADR (per 100,000) are shown in Table 4.15.

Table 4.15: % of Hispanics to AADR (per 100,000)

Counties	% of Hispanics	AADR
Starr	97.2	11.8
Brooks	90.3	13.9
Hidalgo	89.8	14.5
Zapata	88.9	6.8

Duval	87.4	9.8
Willacy	86.8	14.5
Cameron	86.6	17.3
La Salle	77.1	4.1
Kenedy	68.8	0.0
Bailey	56.8	43.2
Floyd	49.6	44.2
Cochran	48.7	43.2
Garza	41.2	38.7
Mitchell	33.3	34.3
McMullen	33.1	13.4

In contrast, there were pockets in the western and northwestern regions with high concentrations of prostate cancer and moderate concentrations of Hispanics. Examples were Bailey (56.8% of Hispanics to 43.2 AADR), Cochran (48.7% of Hispanics to 43.2 AADR), Floyd (49.6% of Hispanics to 44.2 AADR), Garza (41.2% of Hispanics to 38.7 AADR), and Mitchell (33.3% of Hispanics to 34.3 AADR) counties.

Since mortality is reported by place and time of death, it is imperative to take into consideration the migration of these groups to locations where better healthcare access exists. For example, Hispanics in southern Texas may have migrated for more specialized cancer care to eastern Texas, where renowned cancer care facilities exist. Likewise, Hispanics in the western region may have migrated for more specialized cancer care to the far western region, namely El Paso County.

4.3.6.6 Analysis of Other Races

The combination of remaining nonwhite races was categorized as Other Combined or Other Races Combined in the analysis. Whites could not be analyzed due to the limitations of

the study (see Research Limitations). This combination was comprised of American Indian, Alaskan Native, Asian, Native Hawaiian, and Other Pacific Islanders. There was a high to very high concentration of small clusters for Other Races Combined in the northern, central, and eastern parts of Texas (Figure A.5). A comparison of the northern geography of the percentage of other races combined (Figure A.5) and the geography of prostate cancer (Figure A.1) showed a visual similarity between the two. That is, northern Texas, specifically Denton, Collin, Tarrant, and Dallas counties, showed a high concentration of other races combined, which appeared geographically similar to the high concentrations of prostate cancer for their respective counties. Counties with percentage of other races combined to AADR (per 100,000) are shown in Table 4.16.

Table 4.16: % of Other Races (excluding Whites) to AADR (per 100,000)

Counties	% of Other Races	AADR
Fort Bend	15.8	22.1
Collin	11.0	22.1
Harris	6.7	24.8
Denton	6.6	25.4
Travis	6.4	22.0
Dallas	5.7	24.3
Brazoria	5.5	26.9
Tarrant	5.4	25.4
Williamson	5.1	17.6
Calhoun	4.9	23.5
Brazos	4.8	22.0
Bell	4.5	24.9

In contrast, the central region showed high concentrations of other races combined, while prostate cancer concentrations were shown to be light to very light. Examples were Bell (4.5% of other races combined to 24.9 AADR), Williamson (5.1% of other races combined to

17.6 AADR), and Travis (6.4% of other races combined to 22.0 AADR) counties. There was also a moderate concentration of other races combined in eastern counties such as Brazos (4.8% of other races combined to 22.0 AADR) and Harris (6.7% of other races combined to 24.8 AADR). The eastern region was not shown to have a strong similarity in the comparison of the geography of prostate cancer and percentage of other races combined.

4.3.6.7 Hypothesis 1 Findings

Hypothesis 1 of this study posited a positive relationship between race and the geography of prostate cancer mortality, specifically among the percentage of African Americans, Hispanics, and other races combined. A Pearson bivariate correlation matrix was conducted to analyze the relationship between three main minority groups and the age-adjusted death rate (AADR) in prostate cancer specific to the state of Texas (Table 4.17).

Table 4.17: Correlations - Hypothesis 1 Findings

		AADR	African American	Hispanic	Other Combined
AADR	Pearson Correlation	1	.245**	-.097	.043
	Sig. (2-tailed)		.000	.125	.491
	N	254	254	254	254
African American	Pearson Correlation	.245**	1	-.393**	.194**
	Sig. (2-tailed)	.000		.000	.002
	N	254	254	254	254
Hispanic	Pearson Correlation	-.097	-.393**	1	.023
	Sig. (2-tailed)	.125	.000		.717
	N	254	254	254	254
Other Combined	Pearson Correlation	.043	.194**	.023	1
	Sig. (2-tailed)	.491	.002	.717	
	N	254	254	254	254

**. Correlation is significant at the 0.01 level (2-tailed). Table shows percentage for African American, Hispanic, and Other Combined. AADR represents age-adjusted death rate.

A close examination of the data indicates that the percentage of African Americans in Texas shows a weak but positive correlation with the percentage of African Americans and AADR. This positive correlation was statistically significant at the .001 level, demonstrating a correlation between higher death rates among the African American Population in Texas. Percentage of Hispanics shows a slight decrease in AADR, while the percentage of other combined shows a slight increase in AADR.

4.3.6.8 Socioeconomics and Prostate Cancer

Socioeconomic factors were analyzed to help explain the prostate cancer mortality rate in geographic setting and status. These factors were household median income, healthcare costs, unemployment, and uninsured adults.

4.3.6.9 Household Median Income

The 2012 Texas Health Rankings used data from the 2010 Small Area Income and Poverty Estimates (SAIPE) to determine the overall median household income of $48,622 for Texas (Table 4.18). The study analyzed the geography of household median income.

Table 4.18: Texas Median Household Income

Years of Data Used:	2010
Range in Texas (Min-Max):	$22,948 - $81,113
Overall in Texas:	$48,622

Source: (http://www.countyhealthrankings.org/app/texas/2012/measure/factors/63/data) .County Health Rankings & Roadmaps program: 2012 Texas Health Rankings Compilation

Household median income was below the Texas median household income across the southern region, especially regions adjacent to the Mexico-Texas border (Figure A.6), an area with a high population of Hispanics (Figure A.4), but mostly low to moderate concentrations of

prostate cancer (Figure A.1). High levels of household median income were shown in the eastern parts of Texas, with other high levels clustering in central and northern Texas, as well as the western regions known as the Permian Basin. With the exception of the eastern region, regions showing high concentrations of household median income, such as the northern region, showed only low to moderate concentrations of prostate cancer. There were small pockets of clusters in the western (Permian Basin), northwestern (Panhandle), central, northern, and eastern parts of Texas that showed household median income above the state overall household median income.

4.3.6.10 Healthcare Costs

The 2012 Texas Health Rankings used data from the Dartmouth Atlas of Healthcare 2007 data to determine the overall healthcare costs of $10,889 for Texas (Table 4.19). The table shows the price-adjusted Medicare reimbursements per enrollee.

Table 4.19: Texas Overall Healthcare Costs

Years of Data Used:	2007
Range in Texas (Min-Max):	$5,999 - $15,429
Overall in Texas:	$10,889

Source: http://www.countyhealthrankings.org/app/texas/2012/measure/factors/86/data?sort=desc-0. County Health Rankings & Roadmaps program: 2012 Texas Health Rankings Compilation

Healthcare costs were at the highest levels in the southern region of Texas (Figure A.7). This region was a heavily Hispanic-populated area with low household median income levels (Figure A.6). However, prostate cancer mortality rates were low in this region (Figure A.1), and healthcare costs may not have been a contributing factor of mortality rates recorded from this region. However, this may have been due to migration to more specialized cancer care facilities in the eastern region.

4.3.6.11 Unemployment

The 2012 Texas Health Rankings used the Bureau of Labor Statistics 2010 data to determine the overall 8.2% of unemployment for the state (Table 4.20). The table shows percentage of population ages 16 and older unemployed but seeking work.

Table 4.20: Texas Overall Unemployment

Ranking Methodology	
Years of Data Used:	2010
Summary Measure:	Health Factors - Social & Economic Factors (Employment)
Weight in Health Factors:	10%
Summary Information	
Top U.S. Performers:	5.4% (10th Percentile)
Range in Texas (Min-Max):	4.1% - 17.9%
Overall in Texas:	8.2%

Source: http://www.countyhealthrankings.org/app/texas/2012/measure/factors/23/map?sort=desc-0. County Health Rankings & Roadmaps program: 2012 Texas Health Rankings Compilation

Areas with high rates of prostate cancer mortality had low rates of unemployment (Figure A.4). Unemployment percentages were lowest in the northwestern parts of Texas, specifically within the panhandle region (Figure A.8). Moderate to high levels of unemployment were in the southern to eastern part of Texas, along the Texas state line, and in the Gulf of Mexico region. Prostate cancer mortality rates recorded for southern Texas region may not have been correlated to unemployment, assuming migration has been excluded, namely due to Hispanics showing a low rate for prostate cancer mortality in that region (Figure A.4). However, African Americans showed a high concentration in eastern Texas (Figure A.3) and may have been correlated to high concentration of prostate cancer mortality rates (Figure A.1), namely because of the genetic factor in prostate cancer.

4.3.6.12 Uninsured Adults

The 2012 Texas Health Rankings used data from 2009, provided by the Small Area Health Insurance Estimate (SAHIE), to determine the overall 31% of uninsured adults for Texas (Table 4.21). The table shows percentage of adults under age 65 without health insurance.

Table 4.21: Texas Overall Uninsured Adults

Years of Data Used:	2009
Range in Texas (Min-Max):	19% - 51%
Overall in Texas:	31%

Source: http://www.countyhealthrankings.org/app/texas/2012/measure/factors/3/data?sort=desc-0. County Health Rankings & Roadmaps program: 2012 Texas Health Rankings Compilation

Uninsured adults were highest among southern Texas counties along the southern state border (Figure A.9), where there was a high concentration of Hispanics (Figure A.4). Additionally, there was a high concentration of uninsured adults in northwestern Texas, specifically the Panhandle, where there was a low percentage of unemployment. It is important to note that this area is agricultural, and employers may not have provided insurance for employees in this region, as employment may have been seasonal. This assumption helps to explain why a low percentage of unemployment region had a high concentration of uninsured adults.

4.3.6.13 Hypothesis 2 Findings

Hypothesis 2 of this study posited that other factors may contribute to prostate cancer mortality, specifically household median income, healthcare costs, unemployment, and adults who are uninsured, all of which are assumed to show a negative relationship between these variables and AADR, with the exception of household median income, which is assumed to

show a positive correlation. Therefore, a Pearson bivariate correlation matrix was conducted to analyze the relationship between these socioeconomic factors and the age-adjusted death rate (AADR) in prostate cancer (Table 4.22).

Table 4.22: Correlation - Hypothesis 2 Findings

		AADR	Household Income	Healthcare Costs	Percent of Unemp	Percent of Uninsured Adults
AADR	Pearson Corr	1	-.077	.058	.075	-.002
	Sig. (2-tailed)		.219	.360	.235	.980
	N	254	254	251	254	254
Household Income	Pearson Corr	-.077	1	-.008	-.277**	-.705**
	Sig. (2-tailed)	.219		.898	.000	.000
	N	254	254	251	254	254
Healthcare Costs	Pearson Corr	.058	-.008	1	.208**	-.020
	Sig. (2-tailed)	.360	.898		.001	.757
	N	251	251	251	251	251
Percent of Unemployed	Pearson Corr	.075	-.277**	.208**	1	.159*
	Sig. (2-tailed)	.235	.000	.001		.011
	N	254	254	251	254	254
Percent of Uninsured Adults	Pearson Corr	-.002	-.705**	-.020	.159*	1
	Sig. (2-tailed)	.980	.000	.757	.011	
	N	254	254	251	254	254

**. Correlation is significant at the 0.01 level (2-tailed). *. Correlation is significant at the 0.05 level (2-tailed).

Table 4.22 demonstrates there were no statistically significant correlations between AADR and the listed variables. However, inferences can be made from the table. An increase in household income showed a very slight decrease in AADR prostate cancer mortality, indicating a weak negative relationship between both variables. Healthcare costs and AADR showed a very weak positive relationship, showing that as healthcare costs increased, there was a very

slight increase in AADR. Likewise, as the percentage of unemployment increased, AADR increased slightly but still showed a weak positive relationship. Additionally, as the percentage of uninsured adults increased, AADR decreased slightly, indicating a very weak negative relationship.

4.3.6.14 Healthcare Access and Prostate Cancer Primary Care Physicians

Primary care physicians were analyzed to help determine the accessibility of healthcare in counties. High concentrations of primary care physicians were in the eastern, central, and northern region of the state. There was especially a very high concentration in the eastern region. In like manner, the high concentration of healthcare access was comparable to the findings of prostate cancer regional concentrations. That is to say, a high concentration of prostate cancer and primary care physician access were regionally comparable.

4.3.6.15 Hypothesis 3 Findings

Hypothesis 3 of this study posited that there is a negative relationship between healthcare access and prostate cancer. That is to say, greater access to healthcare would indicate a decline in prostate cancer mortality. This is measured in the form of available primary care physicians in the area, who have legal authority to prescribe PSA testing for patients.

Table 4.23 shows a very weak positive correlation between the rate of primary care physicians and AADR. This was a surprising finding since the assumption was that the more available access to healthcare, the greater impact on the mortality rate. However, this correlation does not appear to be statistically significant and would need further review and analysis.

Table 4.23: Correlations - Hypothesis 3 Findings

		AADR	PCP Rate
AADR	Pearson Correlation	1	.108
	Sig. (2-tailed)		.105
	N	254	226
PCP Rate	Pearson Correlation	.108	1
	Sig. (2-tailed)	.105	
	N	226	226

4.3.7 Discussion

Communities that stand to benefit from the insights acquired in this study are several. First, public health officials can find them useful because they are able to determine the geographic locations of mortalities occurring from prostate cancer in the state of Texas. Second, the study can help the medical community, including hospitals, clinics, and providers, to create interventions that can help further define and determine necessary interventions and treatments to save lives. The medical community can find opportunities to work with public officials at the state, county, or city level, specifically to help increase prostate-specific antigen (PSA) testing to find prostate cancer in its early stage. Furthermore, it can help the medical communities initiate further investigations into the high mortality rates among the African American population and to further determine if the high mortality rate among blacks is specific to late PSA testing or other external factors that were not covered in this study. Third, the public in general benefits from this study, especially those among minority populations, because it helps to increase awareness and can drive a positive reaction among the high-risk population to get screened more often. Other groups that may benefit from the study are academic researchers who would like to know the benefits of utilizing data science software to

explore prostate cancer disease. Additionally, insurance groups may also be interested in the results of the study because it can help them mitigate costly treatment in the future by intervening early.

4.3.8 Conclusion

This study documents the value data science programs such as geographical information systems and statistical software have for communities in the public and medical sector. This study assessed concentrations of prostate cancer mortality against social determinants in health to help explain the health disparities of prostate cancer. The study encompassed included race/ethnicity, socioeconomic status, and healthcare access factors, in an attempt to explain the geography of the disease and the reason for its distribution throughout the state of Texas. Two main conclusions can be drawn from the study. The first is that high concentrations of prostate cancer mortality were found mainly in the eastern and central areas of Texas. In like manner, smaller clusters of high concentrations existed in the West Texas Permian Basin and the Panhandle. This is an unusual paradox because there were large numbers of Hispanics living in these regions, but the study did not make a statistically significant finding of correlation between prostate cancer mortality and Hispanic race/ethnicity. This may be because Hispanics tended to migrate, seeking better healthcare treatment. The Hispanic paradox is important to note and requires further research. A deeper study and comparison of the prostate cancer mortality rate among Blacks and Whites in these regions may help solve the paradox. The second conclusion is that there was a high mortality rate among Blacks in the eastern part of Texas, despite renowned healthcare facilities such as MD Anderson Cancer Center in the region. Because the percentage of Blacks in this study was found to be statistically significant when

analyzing prostate cancer mortality rate for Texas, more research is suggested because African Americans had poorer access to prostate cancer treatment and not necessarily a more aggressive form of the disease ("Black Race Not Associated with Worse Prostate Ca Outcomes," 2018). In actuality, race discrimination may have been an important contributor. Benjamins and Whitman (2013), in a recent study on the relationships between discrimination in healthcare and healthcare outcomes, found that African Americans commonly suffered discrimination in healthcare. One study found that there were notable racial disparities in Texas and that both black and Hispanic men were more likely to die of prostate cancer than white men (White, Coker, Du, Eggleston, and Williams, 2010). Minorities, especially blacks, may not receive adequate treatments for disease. Other opportunities for future research include genetics and prostate cancer, specialized medicine, specific migration assessments, environmental considerations, and political factors, whether at the local, regional, or state level. The social determinants in health explanatory variables that were assessed were socioeconomic and healthcare access domains. They did not show a statistically significant correlation with prostate cancer. New variables, and combinations thereof, from these domains are in order for future research. For example, in lieu of assessing primary care physicians against prostate cancer mortality in general, a deeper study examining insured and non-insured African Americans and prostate cancer mortality may produce a fuller assessment of healthcare access. Additionally, other socioeconomic variables such as occupation, education, wealth, and place of residence can be helpful.

4.3.9 Research Limitations

The dataset acquired from the Texas Health Rankings, specifically the measure for

primary care physicians, had missing values. This resulted in an incomplete choropleth map, showing holes in the map for several counties. Additionally, for demographics measure, the percentage of whites was missing from the dataset. Although this was a limitation, and perhaps worthy of mention in the study at a more detailed level, this research study focused on three minority groups, which excluded whites. This exclusion was based on the fact that whites already have lower rates than African Americans and were not a variable of interest in the study. Another limitation of the study was that the unemployment data was comprised of teenagers and younger adults. Employment tends to be tied to other factors such as the ability to afford healthcare and employer-provided insurance. Therefore, unemployment data may misrepresent the relationship between employment and prostate cancer mortality. Finally, AADR was a limitation for counties with small populations, especially in rural areas. For example, two deaths from Hispanics that occur from prostate cancer could show a 50% mortality rate if there were only four Hispanics in the area.

4.3.10 Appendix: Additional Figures for Essay 3

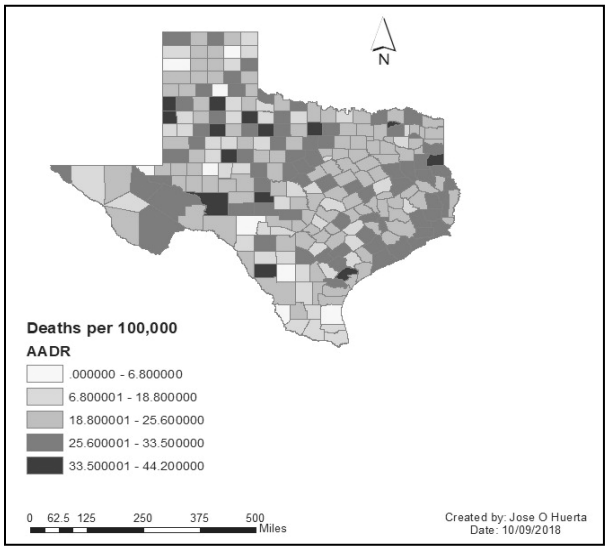

Figure A.1: Age-adjusted prostate cancer mortality in Texas counties, 1999 to 2009

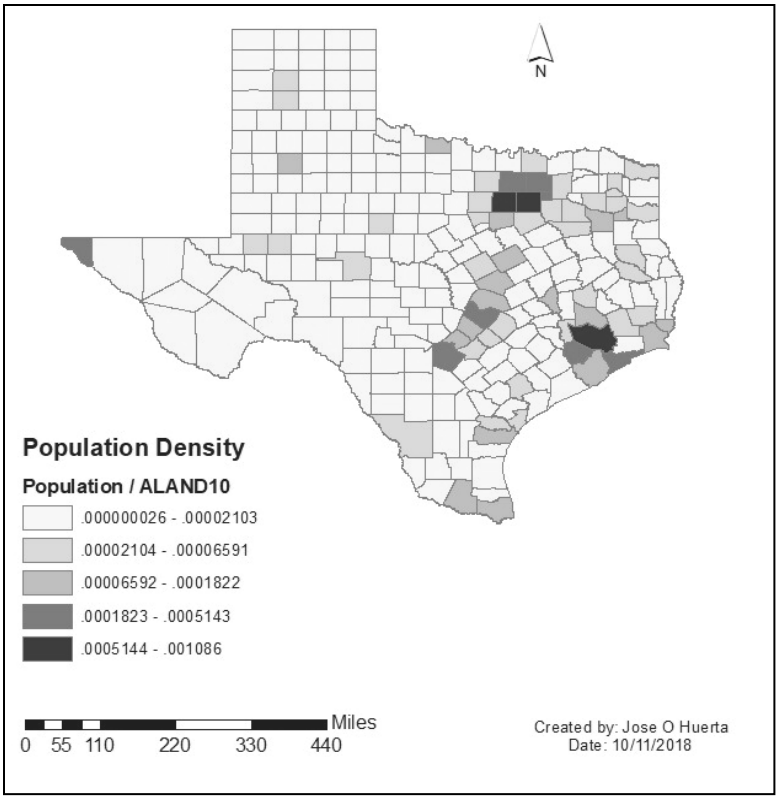

Figure A.2: Geographic distribution of 2009 Texas U.S. Census (population density)

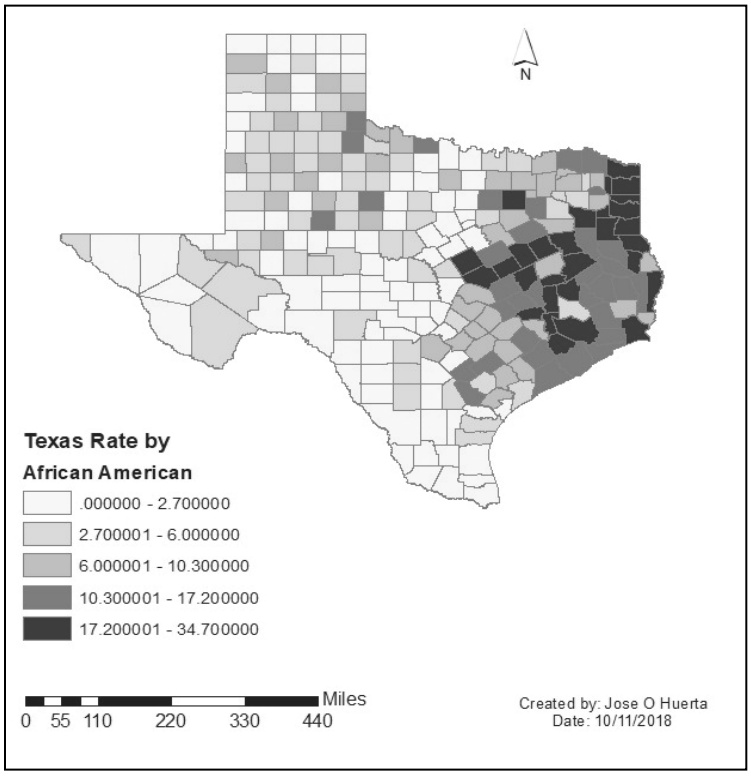

Figure A.3: Geographic distribution of 2009 Texas U.S. Census of African Americans

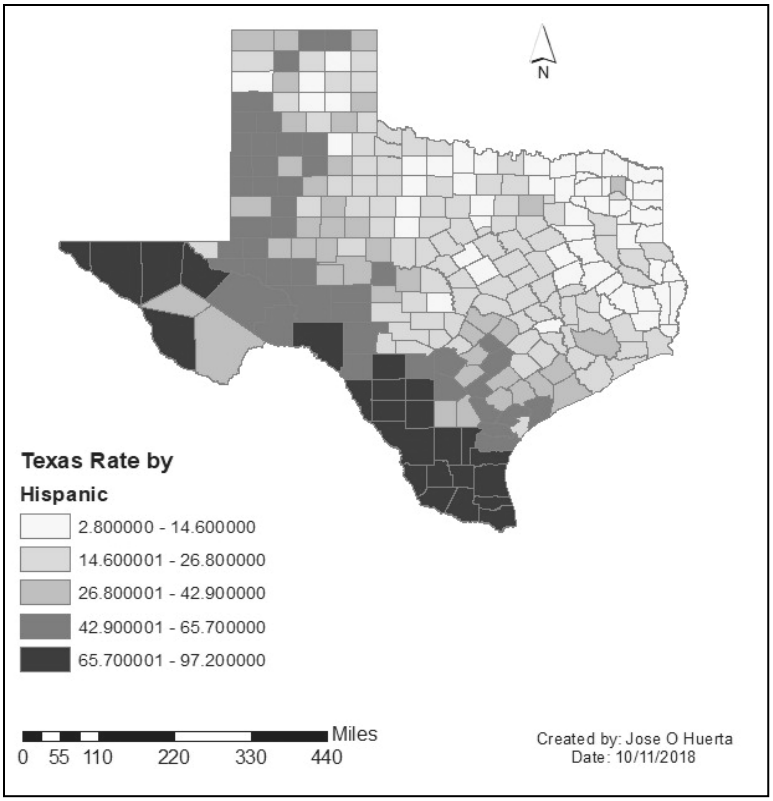

Figure A.4: Geographic distribution of 2009 Texas U.S. Census of Hispanics

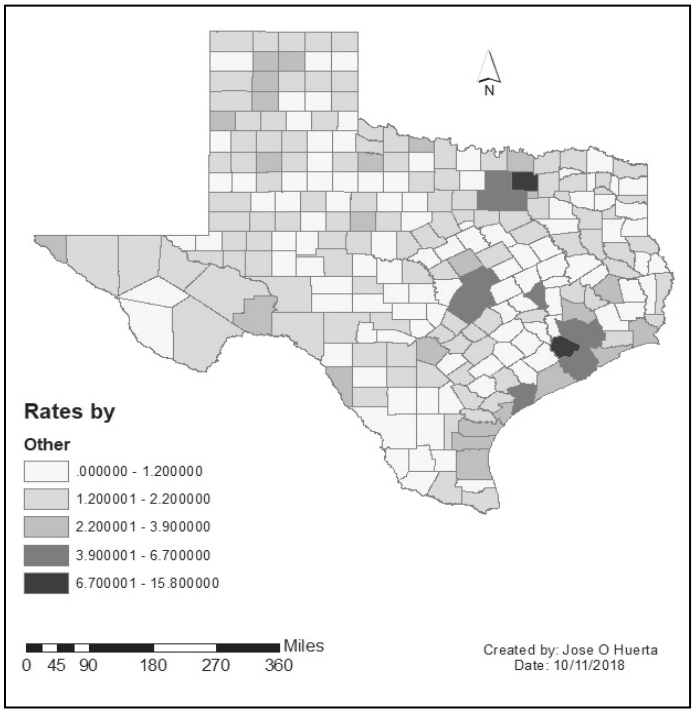

Figure A.5: Geographic distribution of 2009 Texas U.S. Census of other races combined: American Indian, Alaskan Native, Asian, Native Hawaiian, Other Pacific Islanders

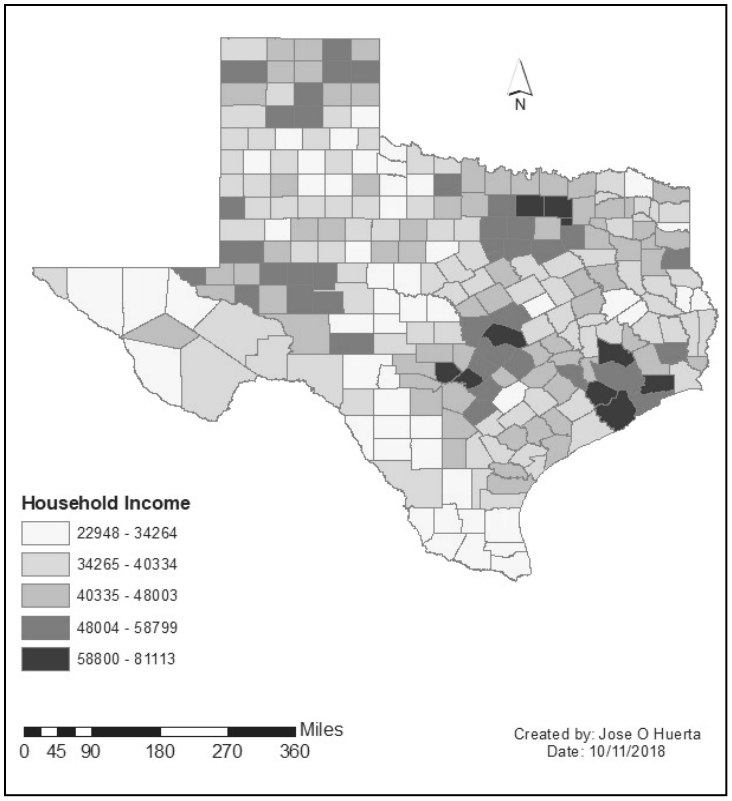

Figure A.6: Geographic distribution of Texas household median income

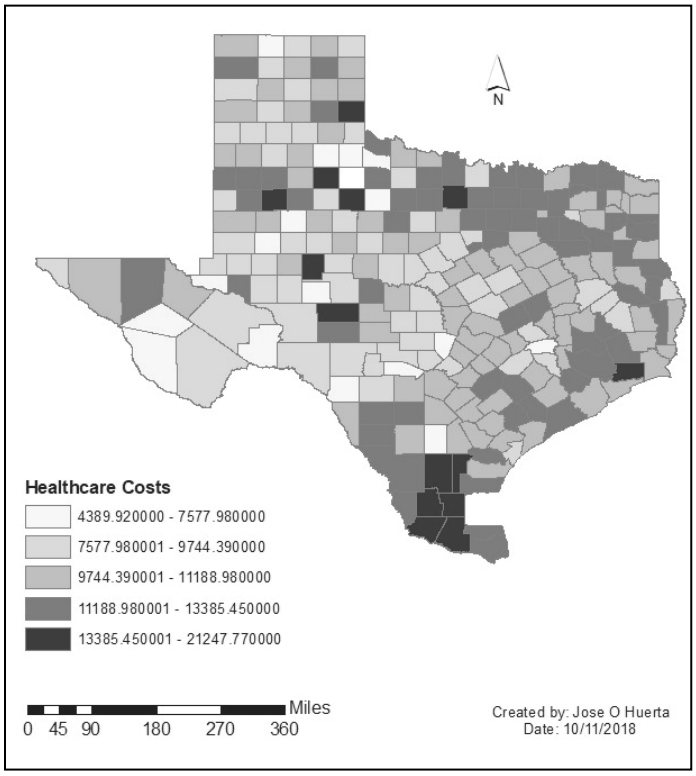

Figure A.7: geographic distribution of Texas healthcare costs

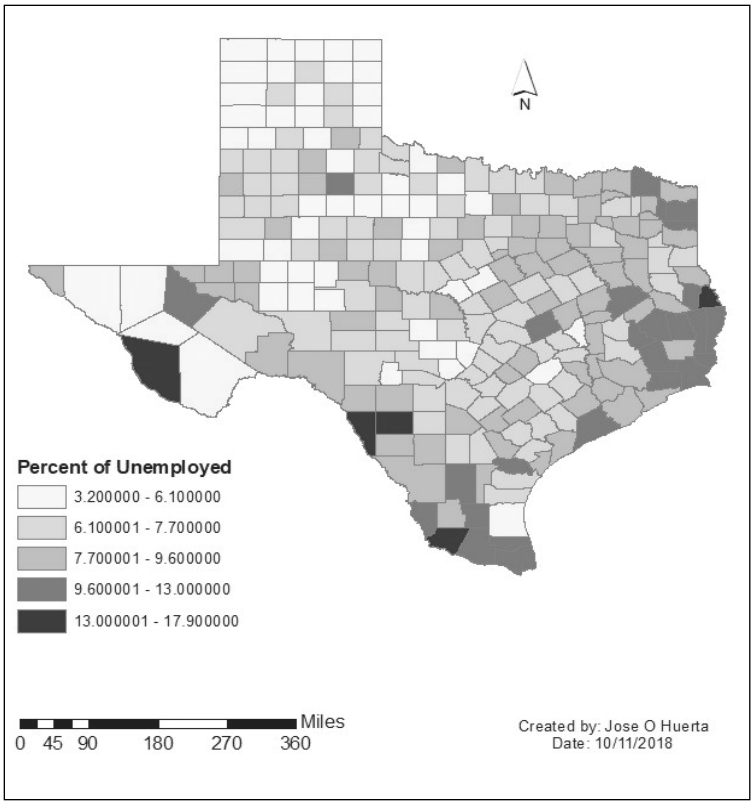

Figure A.8: geographic distribution of unemployment in Texas

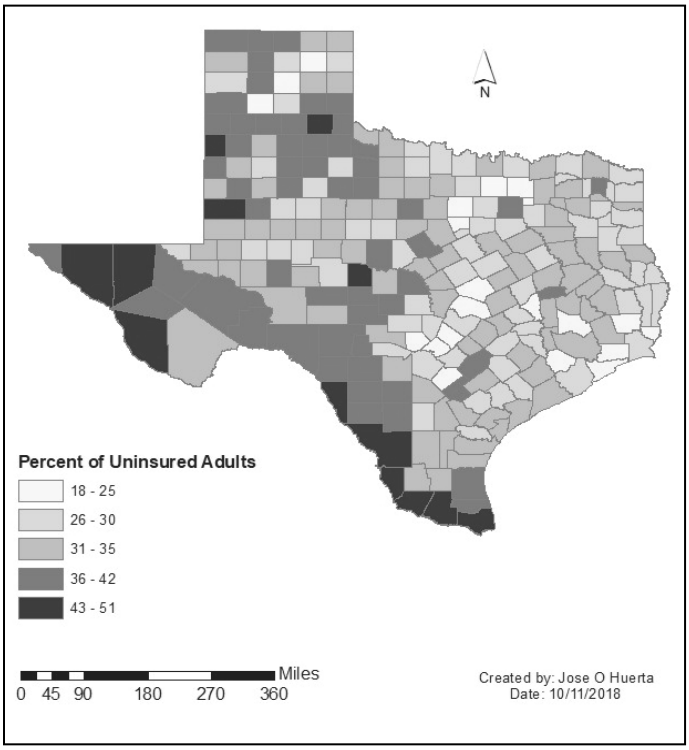

Figure A.9: Geographic distribution of uninsured adults in Texas

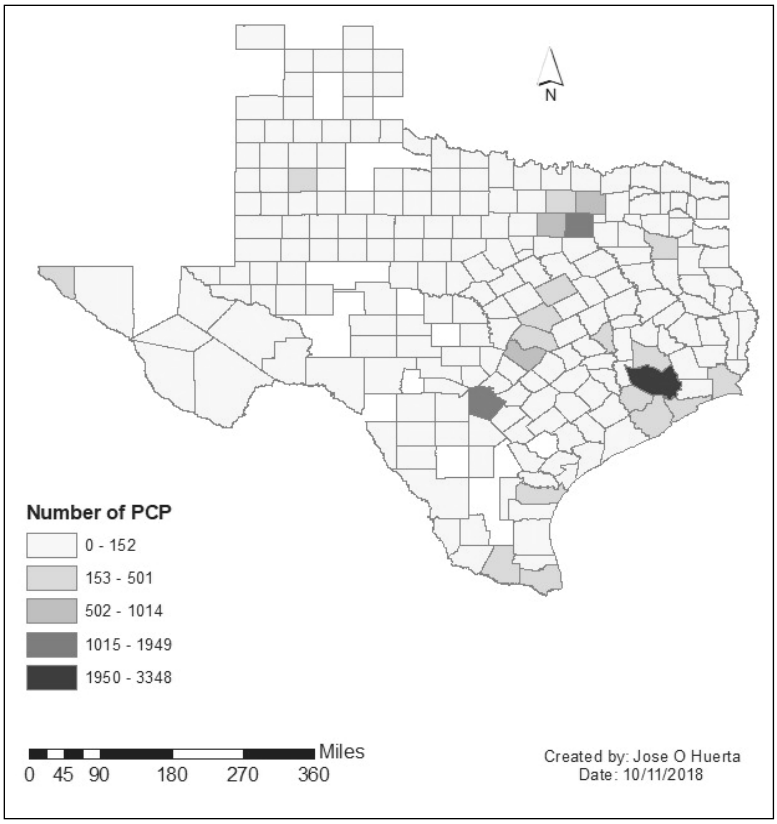

Figure A.10: Geographic distribution of primary care physicians in Texas

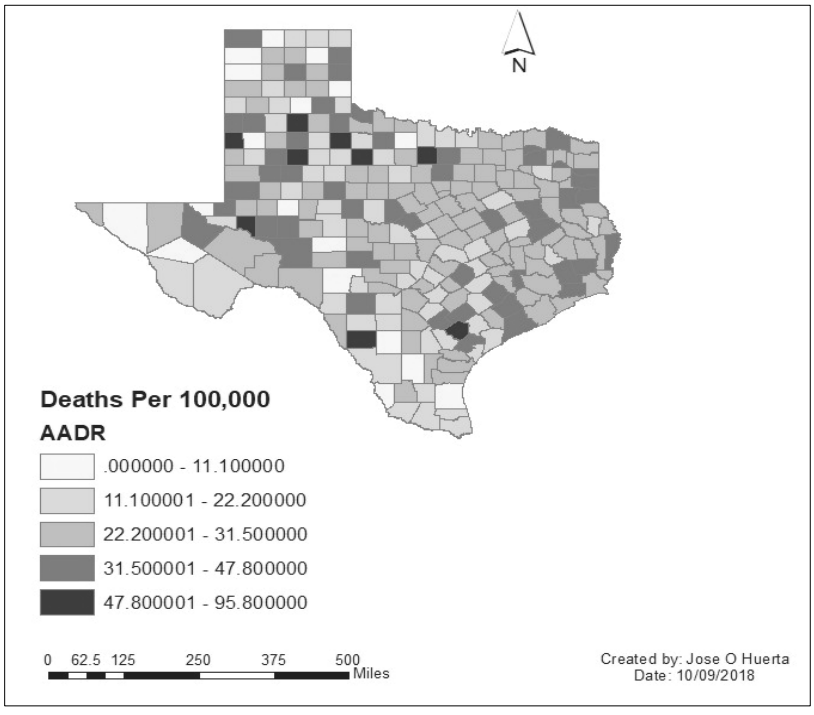

Figure A.11: Age-adjusted prostate cancer mortality in Texas counties, 1999 to 2004

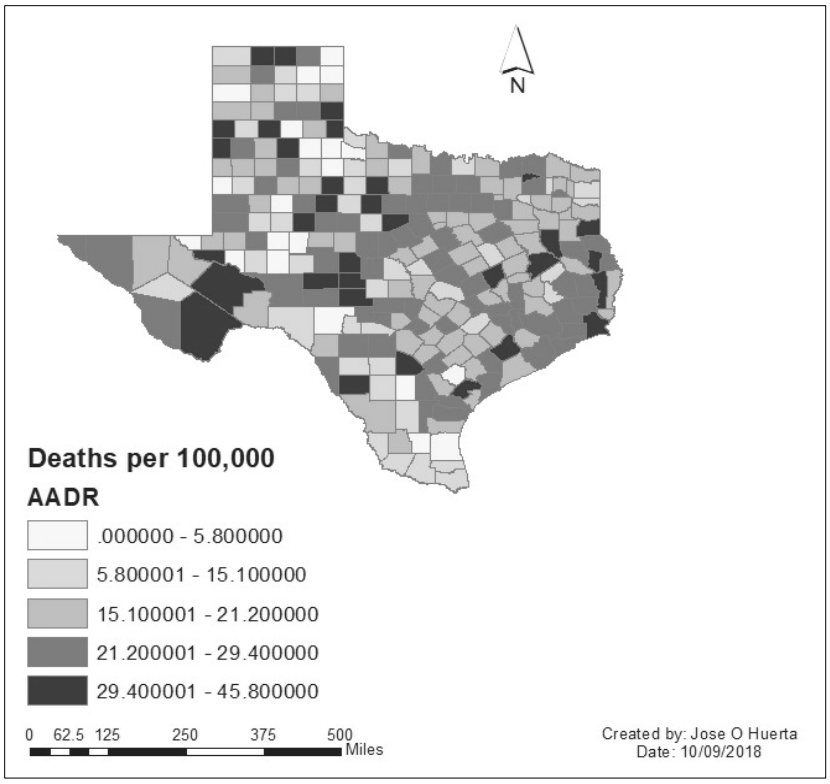

Figure A.12: Age-adjusted prostate cancer mortality in Texas counties, 2005 to 2009

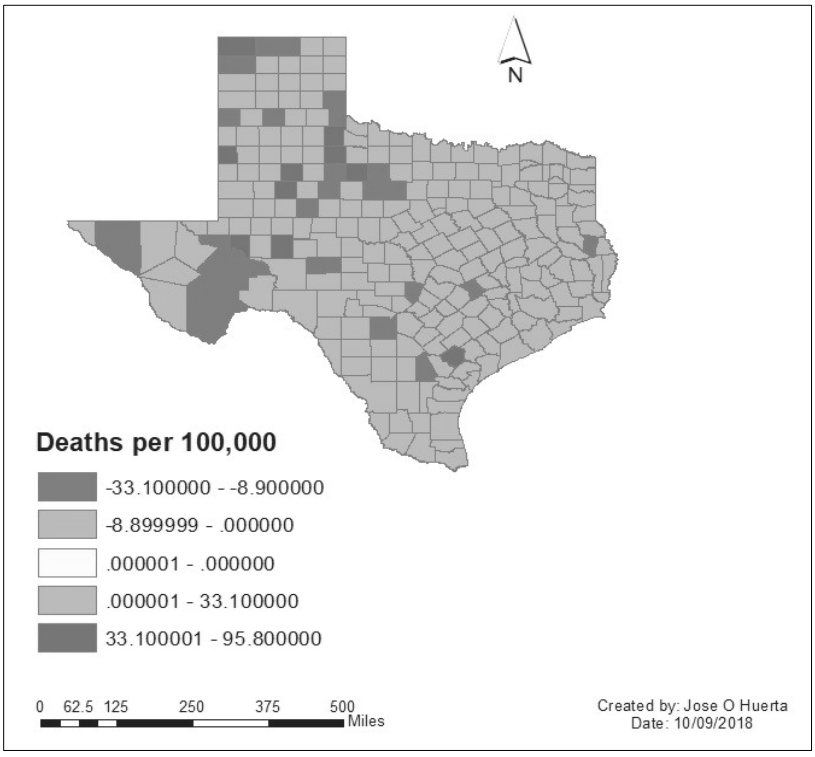

Figure A.13: Age-adjusted prostate cancer mortality in Texas counties, change map

CHAPTER 5

DISCUSSION, CONCLUSION, STUDY LIMITATIONS, AND FUTURE WORK

5.1 Discussion

This research in full addressed the research questions set forth in the study and provided its results in three distinct essays. To address the first research question, what can be answered about the utilization of data science software to address health disparities in public health, a systematic review was conducted on studies that addressed health disparities in public health through healthcare data science applications to determine its impact in the sector. The systematic review was guided by the Preferred Monitoring Items for Systematic Evaluation and Meta-Analysis (PRISMA). Additional analysis was conducted through thematic analysis of the literature to help produce insights regarding the approaches to solving health inequalities in the studies. The discoveries in this study illustrated the importance of further incorporating geographical and data scientific methodologies for healthcare to help encourage the further usage of data science software to help fix complex health inequality problems in the healthcare industry.

To address the second research question, what promising data science software are useful to solve health disparities, an assessment of data science software was conducted. The study sought to discuss and analyze the efficacy of technologies in data science in reducing health inequalities. Additionally, this research illustrates the usefulness of leading data science software that have the appropriate functionality and operations to resolve health disparities. Furthermore, this study shows how the healthcare sector should continuously and iteratively test data science software due to its benefits. The study also helps users in the general or public

health sector to recognize the assessment and opportunities of data science software for healthcare applications.

Finally, to address the third research question, can there be a valuable synergy between data science and healthcare, a small research case study was performed by integrating data science tools into healthcare, demonstrating how to use geographic information systems (GIS) and SPSS. The research looked at prostate cancer mortality in Texas from 1999 to 2009. The diagnosis leverages the ICD-10 codes for Prostate Cancer (C61). In African Americans, there was a statistically significant percentage of prostate cancer mortality, but no statistically significant associations were observed among other races. The research explores ways medical and public health services may test for and manage prostate cancer more efficiently.

5.2 Conclusion

This dissertation reviewed healthcare data science applications and offered evidence for the idea that data science will play a major role in healthcare by demonstrating the significance of utilizing data science to increase the use of healthcare services to become more effective with data. Additionally, the usage of data regarding healthcare will enhance various aspects of healthcare. Therefore, the research sought to aid in determining the degree to which healthcare data is being reviewed and investigated aspects of data science applications in healthcare.

The results of this study have contributed to the body of research about data science applications and exhibits data science applications in healthcare by showing it is useful for solving multiple healthcare challenges such as solving health inequalities. Additionally, this work helped settle the opportunity data science has at being a feasible tool for healthcare

applications. The perspectives obtained from this analysis allows for further contribution to the body of research by demonstrating how data science assists in preparing and applying data-driven approaches to solving analytic challenges in healthcare. From the top ranking applications, Tensorflow and Python each have the capacity to automate and model variables such as salary, education ethnicity, age, and cross-referencing variables to results in medical care and finance to assess outcomes that expose health inequalities. Therefore, Essay 2 of this work introduces a method that, theoretically, can be utilized to resolve health inequalities.

The perspectives obtained from this study can help many groups. A mapping tool is beneficial for public health authorities since it indicates the origin of prostate cancer deaths. Second, the research will help the medical world, including hospitals, facilities, and suppliers, identify and assess appropriate procedures and therapies to save lives. The medical profession will see ways to collaborate with state and county policymakers to promote additional PSA research to find prostate cancer in its early stages. Additionally, it may help to analyze and decide whether the high mortality rate among African Americans were specific to late PSA testing or other external influences not addressed in this report.

Public gain also comes from this research because it raises visibility and may drive a favorable response within high-risk groups to get screened more regularly. Additionally, academic students will profit from the analysis if they would like to use data science to investigate prostate cancer in their own state. Insurance groups might also be interested in the outcomes of the analysis because it lets them foresee potential expensive care.

5.3 Study Limitations

There were limitations to this study. In regards to the systematic review of Essay 1, its

aim was to compile all publications on applications of data science and assess their position with respect to health inequality. It is conceivable that the results were not completely collected within our examined public and private health study data accessible in the systematic analysis. Furthermore, health disparities in public health are a significant obstacle to analytical science. This demonstrates the possible need for widely accessible public health data sets, and for the collaboration of various public health research groups. Other limitations of the research, was not being able to assess the capacity of healthcare organizations to adopt open source data science software and platforms. This can prove to be detrimental to exposing protected health information. In Essay 2, 30% usage was selected because a marginally higher percentage resulted in too few software packages and a marginally lower percentage resulted in too many to allow meaningful and reasonable comparisons.

For Essay 3, there were incomplete values in the primary care provider details for some counties in Texas. This culminated in an imperfect choropleth chart with gaps for some counties. Additionally, for demographics, the percentage of whites was absent from the dataset. While this analysis had drawbacks, it is hard to interpret data regarding a population that was not represented in the survey. As a predictor of concern, blacks still have higher rates than whites, so whites were omitted from the report. Another limitation was that the data are biased because it contains data on teens and younger people, much younger than the demographic that gets impacted by prostate cancer. Therefore, unemployment data can underreport the association between employment and prostate cancer mortality. Lastly, AADR was a weakness for small counties and particularly in remote regions. For example, two deaths among Hispanics will indicate a 50% mortality rate if there were only four Hispanics in the area.

5.4 Future Work

In this report, the existing results of the need for widespread implementation of data science software utilizing spatial and statistical methods to further encourage further research and technical use in addressing health disparities in public health by data science are clearly mentioned.

In regards to open source data science software programs, future studies in healthcare, especially hospitals, are needed to determine how protected health information can be kept secure with the adoption of data science platforms. Further studies can be conducted in the field of data science to determine the capacities of healthcare organizations to implement open source programming systems.

Since African Americans were shown to have poorer access to prostate cancer care and a more serious type of the disorder, more research is proposed in the public health sector. Other potential applications for data science software beyond what was demonstrated in Essay 3 include research on biology and prostate cancer, research on medical specialties that can help make associations to other medical disorders or particular diseases, research into the human migration trends, and research into the impact of municipal and state politics.

REFERENCES

About Social Determinants of Health. [online] https://www.who.int/social_determinants/sdh_definition/en/ (accessed 9 September 2019).

About us (n.d). Retrieved from http://www.countyhealthrankings.org/about-us

Adam, N. R., Wieder, R., & Ghosh, D. (2017). Data science, learning, and applications to biomedical and health sciences. *Annals of the New York Academy of Sciences*, *1387*(1), 5-11. doi:10.1111/nyas.13309

Alexander, G. L. (2015). Building Evidence in Health Informatics. *Western Journal of Nursing Research, 37*(7), 839-841. doi:10.1177/0193945915576692

Allen, C., Tsou, M., Aslam, A., Nagel, A., & Gawron, J. (2016). Applying GIS and machine learning methods to Twitter data for multiscale surveillance of influenza. *PLOS ONE*, *11*(7), e0157734. https://doi.org/10.1371/journal.pone.0157734

Anderson, A. C., O'Rourke, E., Chin, M. H., Ponce, N. A., Bernheim, S. M., & Burstin, H. (2018). Promoting health equity and eliminating disparities through performance measurement and payment. *Health Affairs*, *37*(3), 371-377. https://doi.org/10.1377/hlthaff.2017.1301

Attard, G., Parker, C., Eeles, R. A., Schröder, F., Tomlins, S. A., Tannock, I., . . . Bono, J. S. (2016) 'Prostate cancer', *The Lancet,* Vol. 387, No. 10013, pp.70-82. doi:10.1016/s0140-6736(14)61947-4

Baptista, M., Vasconcelos, J. B., Rocha, Á., Silva, R., Carvalho, J. V., Jardim, H. G., & Quintal, A. (2019). The impact of perioperative data science in hospital knowledge management. *Journal of Medical Systems*, *43*(2). https://doi.org/10.1007/s10916-019-1162-3

Basu, S., Faghmous, J. H., & Doupe, P. (2020). Machine learning methods for precision medicine research designed to reduce health disparities: A structured tutorial. *Ethnicity & Disease*, *30*(Suppl 1), 217-228. doi:10.18865/ed.30.s1.217

Bates, M. J. (2006). Fundamental forms of information. *Journal of the American Society for Information Science and Technology*, *57*(8), 1033-1045. doi:10.1002/asi.20369

Bauer, J., Groneberg, D. A., Maier, W., Manek, R., Louwen, F., & Brüggmann, D. (2017). Accessibility of general and specialized obstetric care providers in Germany and England: An analysis of location and neonatal outcome. *International Journal of Health Geographics*, *16*(1). doi:10.1186/s12942-017-0116-6

Benjamins, M. R., and Whitman, S. (2013). 'Relationships between discrimination in healthcare and healthcare outcomes among four race/ethnic groups', *Journal of Behavioral Medicine*, Vol. *37*, No. 3, pp.402-413. doi:10.1007/s10865-013-9496-7

Besselaar, P. V. D., & Gaston, H. (2001). Disciplinary, Multidisciplinary, Interdisciplinary – Concepts and Indicators. *Paper for the 8th conference on Scientometrics and Informetrics.* Sydney, Australia.

Bhargava, N., Aziz, A., & Rajiv, A. (2013). Selection criteria for data mining software: A study. *IJCSI International Journal of Computer Sciences*, *10*(3), No. 2, 1694-0784.

Black Race Not Associated with Worse Prostate Ca Outcomes. [online] https://www.medpagetoday.com/meetingcoverage/astro/75896 (Accessed 24 October 2018)

Borko, H. (1968). Information science: What is it? *American Documentation*, *19*(1), 3-5. doi:10.1002/asi.5090190103

Boslett, A. J., Denham, A., Hill, E. L., & Adams, M. C. (2019). Unclassified drug overdose deaths in the opioid crisis: Emerging patterns of inequity. *Journal of the American Medical Informatics Association*, *26*(8-9), 767-777. doi:10.1093/jamia/ocz050

Braveman, P. (2014) 'What are health disparities and health equity? We need to be clear', *Public Health Reports*, Vol. *129*, No. 1, Suppl. 2, pp.5-8. doi:10.1177/00333549141291s203

Brookes, B. C. (1980). The foundations of information science. Part I. Philosophical aspects. *Information Scientist*, *2*(3-4), 125-133. doi:10.1177/016555158000200302

Brookes, B. C. (1980). The foundations of information science. Part II. Quantitative aspects. *Classes of things and the challenge of human individuality*, *2*(5), 209-221. doi:10.1177/016555158000200502

Broome, M. E. (2016). Big data, data science, and big contributions. *Nursing Outlook*, *64*(2), 113-114. doi:10.1016/j.outlook.2016.02.001

Buajitti, E., Watson, T., Norwood, T., Kornas, K., Bornbaum, C., Henry, D., & Rosella, L. C. (2019). Regional variation of premature mortality in Ontario, Canada: A spatial analysis. *Population Health Metrics*, *17*(1). doi:10.1186/s12963-019-0193-9

Bui, T. Q., and Pham, H. M. (2016) 'Web-based GIS for spatial pattern detection: application to malaria incidence in Vietnam', *SpringerPlus*, Vol. *5, No.* 1. doi:10.1186/s40064-016-2518-5

Cao, L. (2017). Data science: A comprehensive overview. *ACM Computing Surveys, 50*(3), 1-42. https://doi.org/10.1145/3076253

Cecchetti, A. A., Bhardwaj, N., Murughiyan, U., Kothakapu, G., & Sundaram, U. (2020). Fueling clinical and translational research in Appalachia: Informatics platform approach. *JMIR Medical Informatics*, *8*(10), e17962. doi:10.2196/17962

Centers for Disease Control and Prevention and National Cancer Institute, U.S. Cancer Statistics Working Group. (2018) *U.S. Cancer Statistics data visualizations tool*. https://gis.cdc.gov/Cancer/USCS/DataViz.html (Accessed 18 October 2019).

Chan, J., Polo, A., Zubizarreta, E., Bourque, J., Hanna, T. P., Gaudet, M., … Abdel-Wahab, M. (2019). Access to radiotherapy and its association with cancer outcomes in a high-income country: Addressing the inequity in Canada. *Radiotherapy and Oncology*, *141*, 48-55. doi:10.1016/j.radonc.2019.09.009

Chang, Y., & Huang, M. (2011). A study of the evolution of interdisciplinarity in library and information science: Using three bibliometric methods. *Journal of the American Society for Information Science and Technology*, *63*(1), 22-33. doi:10.1002/asi.21649

Chase, J. D., & Vega, A. (2016). Examining health disparities using data science. *Research in Gerontological Nursing*, *9*(3), 106-107. https://doi.org/10.3928/19404921-20160404-01

Chen, B. K., Cheng, X., Bennett, K., & Hibbert, J. (2015). Travel distances, socioeconomic characteristics, and health disparities in nonurgent and frequent use of hospital emergency departments in South Carolina: A population-based observational study. *BMC Health Services Research*, *15*(1). doi:10.1186/s12913-015-0864-6

Cheng, I., Witte, J. S., McClure, L. A., Shema, S. J., Cockburn, M. G., John, E. M., and Clarke, C. A. (2009) 'Socioeconomic status and prostate cancer incidence and mortality rates among the diverse population of California', *Cancer Causes & Control*, Vol. *20, no.* 8, pp.1431-1440. doi:10.1007/s10552-009-9369-0

Chin, M. H. (2016). Creating the business case for achieving health equity. *Journal of General Internal Medicine*, *31*(7), 792-796. https://doi.org/10.1007/s11606-016-3604-7

Chiu, H., & (Jack) Li, Y. (2018). Improving healthcare management with data science. *Computer Methods and Programs in Biomedicine*, *154*, A1. doi:10.1016/s0169-2607(17)31508-0

Choi, S., Yoo, J., Park, J., Lee, H., Tran, H. T., Lee, J., & Oh, J. (2018). Manifestations of socioeconomic status and its association with physical child punishment– results from the multi-indicators cluster survey in Viet Nam, 2006–2014. *Child Abuse & Neglect*, *85*, 1-8. doi:10.1016/j.chiabu.2018.08.022

Cielen, D., Meysman, A.D.B., and Ali, M. (2016). *Introducing data science: Big data, machine learning, and more, using Python tools*. Shelter Island, NY: Manning Publications.

Clarke, V., & Braun, V. (2014). Thematic analysis. *Encyclopedia of Critical Psychology*, 1947-1952. doi:10.1007/978-1-4614-5583-7_311

County Health Rankings & Roadmaps. Healthcare Costs. [online] http://www.countyhealthrankings.org/app/texas/2012/measure/ (Accessed 18 October 2019).

Data Science Central (2013). My Data Science Book - Table of Contents. Retrieved from http://www.datasciencecentral.com/profiles/blogs/my-data-science-book

Data Science Central. (2014). The data science project lifecycle. Retrieved from http://www.datasciencecentral.com/profiles/blogs/the-data-science-project-lifecycle

De la Torre Díez, I., Cosgaya, H. M., Garcia-Zapirain, B., & López-Coronado, M. (2016). Big data in health: A literature review from the year 2005. *Journal of Medical Systems*, *40*(9). https://doi.org/10.1007/s10916-016-0565-7

Delaney, C.W., & Westra, B. (2016). Big data. *Western Journal of Nursing Research*, 39(1), 3-4. https://doi.org/10.1177/0193945916671687

DeMeester, R. H., Xu, L. J., Nocon, R. S., Cook, S. C., Ducas, A. M., & Chin, M. H. (2017). Solving disparities through payment and delivery system reform: A program to achieve health equity. *Health Affairs*, *36*(6),11331139. https://doi.org/10.1377/hlthaff.2016.0979

Detmer, D. E., & Shortliffe, E. H. (2014). Clinical informatics. *Jama, 311*(20), 2067. https://doi.org/10.1001/jama.2014.3514

Disparities in Health and Healthcare: Five Key Questions and Answers. [online] https://www.kff.org/disparities-policy/issue-brief/disparities-in-health-and-health-care-five-key-questions-and-answers (Accessed 18 October 2009).

Dixon, B. E. (2017). 2016 International Society for Disease Surveillance Conference New Frontiers in Surveillance: Data Science and Health Security. *Online Journal of Public Health Informatics*, *9*(1). doi:10.5210/ojphi.v9i1.7791

Dunn, M. C., & Bourne, P. E. (2017). Building the biomedical data science workforce. *PLOS Biology*, *15*(7), e2003082. https://doi.org/10.1371/journal.pbio.2003082

Dwyer-Lindgren, L., Bertozzi-Villa, A., Stubbs, R. W., Morozoff, C., Mackenbach, J. P., Van Lenthe, F. J., … Murray, C. J. (2017). Inequalities in life expectancy among US counties, 1980 to 2014. *JAMA Internal Medicine*, *177*(7), 1003. doi:10.1001/jamainternmed.2017.0918

EMC. 2011. Data science revealed: A data-driven glimpse into the burgeoning new field. Retrieved from www.emc.com/collateral/about/news/emc-data-science-study-wp.pdf.

Erdman, Shelby Lin. (2020, May 6). *Black communities account for disproportionate number of COVID-19 deaths in the US, study finds*. CNN.

https://www.cnn.com/2020/05/05/health/coronavirus-african-americans-study/index.html

Fayyad, U. (2012, July 4). *Data science revealed: A data-driven glimpse into the burgeoning new field*. https://fayyad.com/data-science-revealed-a-data-driven-glimpse-into-the-burgeoning-new-field/

Gharaibeh, H. M. (2014). Developing a scoring model to evaluate project management software packages based on ISO/IEC software evaluation criterion. *Journal of Software Engineering and Applications*, *07*(01), 27-41. https://doi.org/10.4236/jsea.2014.71004

Gil Press. 2013. A Very Short History of Data Science. Retrieved from https://www.forbes.com/sites/gilpress/2013/05/28/a-very-short-history-of-data-science/#1161694a55cf

Gregorio, D. I., Kulldorff, M., Sheehan, T., and Samociuk, H. (2004) 'Geographic distribution of Prostate Cancer incidence in the era of PSA testing, Connecticut, 1984 to 1998', *Urology*, Vol. *63, No.* 1, pp.78-82. doi:10.1016/j.urology.2003.08.008

Harris DE, Aboueissa AM, Baugh N, Sarton C. Impact of rurality on maternal and infant health indicators and outcomes in Maine. Rural Remote Health. 2015 Jul-Sep;15(3):3278. Epub 2015 Jul 21. PMID: 26195158.

Health Disparities. [online] https://medlineplus.gov/healthdisparities.html (Accessed 18 October 2019).

Henry, T.A. (2019). *Social determinants of health: What medical students need to know*. [online] https://www.ama-assn.org/delivering-care/patient-support-advocacy/social-determinants-health-what-medical-students-need-know?matchtype=e&network=g&device=c&adposition=1t3&keyword=social%20determinants%20of%20health&utm_source=google&utm_medium=ppc&utm_campaign=pe-digital-ads-membership&utm_effort=GG0001&gclid=CjwKCAjwnMTqBRAzEiwAEF3ndo4yIcrpKp0qwljLmkfoTzcdUQEE0ZAY_1n5_IxFcNNCwf0yqvTfzhoCL9UQAvD_BwE (Accessed 18 October 2019).

Holland, A. G. (2008). Information science: an interdisciplinary effort? *Journal of Documentation*, *64*(1), 7-23. doi:10.1108/00220410810844132

Hughes, M. C., Baker, T. A., Kim, H., & Valdes, E. G. (2019). Health behaviors and related disparities of insured adults with a healthcare provider in the United States, 2015–2016. *Preventive Medicine*, *120*, 42-49. https://doi.org/10.1016/j.ypmed.2019.01.004

KDnuggets. (n.d.). *About KDnuggets.* https://www.kdnuggets.com/about

Kern, L. M., & Kaushal, R. (2007). Health information technology and health information exchange in New York State: New initiatives in implementation and evaluation. Journal of Biomedical Informatics, 40(6). doi:10.1016/j.jbi.2007.08.010

Kim, E., Moy, L., Gao, Y., Hartwell, C. A., Babb, J. S., & Heller, S. L. (2019). City patterns of screening mammography uptake and disparity across the United States. *Radiology*, *293*(1), 151-157. doi:10.1148/radiol.2019190647

Kim, J. H., Sun, H. Y., Kim, H. J., Ko, Y. M., Chun, D., and Park, J. Y. (2017) 'Does uneven geographic distribution of urologists effect bladder and Prostate Cancers mortality? National health insurance data in Korea from 2007-2011', *Oncotarget*, Vol. *8, No.* 39. doi:10.18632/oncotarget.18036

King, C. (2016). Disparities in access to preventive healthcare services among insured children in a cross sectional study. *Medicine*, *95*(28), e4262. https://doi.org/10.1097/md.0000000000004262

Klassen, A. C., and Platz, E. A. (2006) 'What can geography tell us about prostate cancer?', *American Journal of Preventive Medicine*, Vol. *30, No.* 2, pp.S7-S15. doi:10.1016/j.amepre.2005.09.004

Kumar, S., Singh, R., Malik, S., Manne, U., and Mishra, M. (2018). 'Prostate cancer health disparities: An immuno-biological perspective', *Cancer Letters*, Vol. *414*, 153-165. doi:10.1016/j.canlet.2017.11.011

LaVeist, T.A., and Pierre, G. (2014). 'Integrating the 3Ds—social determinants, health disparities, and health-care workforce diversity', *Public Health Reports,* Vol. 129, No. 1, Suppl. 12, pp.9-14. doi:10.1177/00333549141291s204

Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gotzsche, P. C., Ioannidis, J. P., … Moher, D. (2009). The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: Explanation and elaboration. *BMJ*, *339*(jul21 1), b2700-b2700. doi:10.1136/bmj.b2700

Lilley, D. B., & Trice, R. W. (1989). Information Science 1945-1968. *A history of information science 1945-1985*. San Diego, Calif: Academic Press.

Lilley, D. B., & Trice, R. W. (1989). Information Science 1945-1968. *Online Activity 1945-1985*. San Diego, Calif: Academic Press.

Linton, K. D., & Catto, J. W. (2013). Prostate Cancer. *Surgery (Oxford)*, *31*(10), 516-522. doi:10.1016/j.mpsur.2013.08.001

Liu, L., Zhang, H., Li, J., Wang, R., Yu, L., Yu, J., & Li, P. (2009). Building a Community of Data Scientists: An Explorative Analysis. *Data Science Journal*, *8*, 201-208. doi:10.2481/dsj.008-004

Loukides, M. (2011) *What is Data Science?* Sebastopol, CA: O'reilly Media, Inc. Retrieved October 8, 2017, from http://www.oreilly.com/data/free/what-is-data-science.csp

Ma, L. (2012). Meanings of information: The assumptions and research consequences of three foundational LIS theories. *Journal of the American Society for Information Science and Technology*, *63*(4), 716-723. doi:10.1002/asi.21711

Major, J. M., Oliver, M. N., Doubeni, C. A., Hollenbeck, A. R., Graubard, B. I., and Sinha, R. (2012) 'Socioeconomic status, healthcare density, and risk of prostate cancer among African American and Caucasian men in a large prospective study', *Cancer Causes and Control* Vol. 23, No. 7, pp.1185–1191. doi: 10.1007/s10552-012-9988-8

Maneth, S., & Poulovassilis, A. (2016). Data science. *The Computer Journal, 60(3), 285-286*. https://doi.org/10.1093/comjnl/bxw073

Manna, Maloy. (2014, December 18). *The data science project lifestyle*. Data Science Central. https://www.datasciencecentral.com/profiles/blogs/the-data-science-project-lifecycle

Marchibroda, J. M. (2007). Health information exchange policy and evaluation. *Journal of Biomedical Informatics*, *40*(6). https://doi.org/10.1093/comjnl/bxw07310.1016/j.jbi.2007.08.008

McLoughlin, G. M., McCarthy, J. A., McGuirt, J. T., Singleton, C. R., Dunn, C. G., & Gadhoke, P. (2020). Addressing food insecurity through a health equity lens: A case study of large urban school districts during the COVID-19 pandemic. *Journal of Urban Health*, *97*(6), 759-775. doi:10.1007/s11524-020-00476-0

McQuaid, E. L., & Landier, W. (2017). Cultural issues in medication adherence: Disparities and directions. *Journal of General Internal Medicine*, *33*(2), 200-206. https://doi.org/10.1007/s11606-017-4199-3

Misnevs, B., & Yatskiv, I. (2016). Data Science: Professional Requirements and Competence Evaluation. *Baltic J. Modern Computing, 4(3)*.

Murphy, W. F., Murphy, S. S., Buettner, R. R., & Gill, G. (2015). Case study of a complex informing system: Joint interagency field experimentation (JIFX). *Informing Science: The International Journal of an Emerging Transdiscipline, 18*, 063-109. https://doi.org/10.1093/comjnl/bxw07310.28945/2289

Nowell, L. S., Norris, J. M., White, D. E., & Moules, N. J. (2017). Thematic Analysis: Striving to Meet the Trustworthiness Criteria. *International Journal of Qualitative Methods*, *16*(1), 160940691773384. doi:10.1177/1609406917733847

O'Connor, S. (2018). Big data and data science in healthcare: What nurses and midwives need to know. *Journal of Clinical Nursing*, *27*(15-16), 2921-2922. https://doi.org/10.1093/comjnl/bxw07310.1111/jocn.14164

Office of Disease Prevention and Health Promotion. (n.d.). *Disparities*. HealthyPeople.gov. https://www.healthypeople.gov/2020/about/foundation-health-measures/Disparities

Overview—Vitalnet Data Warehouse Software. [online] https://www.ehdp.com/vitalnet/overview.htm (Accessed 18 October 2019).

Patil, D. J. (2011). *Building data science teams*. O'Reilly Media.

Pearce, N., Foliaki, S., Sporle, A., and Cunningham, C. (2004) 'Genetics, race, ethnicity, and health', *BMJ*, Vol. *328, No.* 7447, pp.1070-1072. doi:10.1136/bmj.328.7447.1070

Penson, D. F., and Chan, J. M. (2007) 'Prostate Cancer', *The Journal of Urology,* Vol. 177, No. 6, pp.2020-2029. doi:10.1016/j.juro.2007.01.121

Piatetsky, G. (2019). *Python leads the 11 top data science, machine learning platforms: Trends and analysis*. KDNuggets. https://www.kdnuggets.com/2019/05/poll-top-data-science-machine-learning-platforms.html

Posted by Zeeshan Alam on May 4, 2017 at 1:00am View Blog. (n.d.). What is Data Science and why Data Science is in demand. Retrieved October 13, 2017, from http://www.datasciencecentral.com/profiles/blogs/what-is-data-science-and-why-data-science-is-in-demand

Powell Doherty, R., Müller-Demary, D., Hosszu, A., Duminica, A., Bertke, A., Lewis, B., & Eubank, S. (2018). A survey of quality of life indicators in the Romanian Roma population following the 'Decade of Roma inclusion'. *F1000Research*, *6*, 1692. doi:10.12688/f1000research.12546.2

Power, D. J. (2016). Data science: supporting decision-making. *Journal of Decision Systems*, *25*(4), 345-356. doi:10.1080/12460125.2016.1171610

Press, Gil. (2013, May 28). *A very short history of data science*. Forbes. https://www.forbes.com/sites/gilpress/2013/05/28/a-very-short-history-of-data-science/#1161694a55cf

Rayward, W. (1996). The history and historiography of information science: Some reflections. *Information Processing & Management*, *32*(1), 3-17. doi:10.1016/0306-4573(95)00046-j

Rebbeck, T. R. (2017) 'Prostate cancer genetics: variation by race, ethnicity, and geography', *Seminars in Radiation Oncology*, Vol. *27, No.* 1, pp.3-10. doi:10.1016/j.semradonc.2016.08.002

Rosenbaum, H. (2010). Social Informatics. *Encyclopedia of Library and Information Sciences, Third Edition*, 4814-4819. doi:10.1081/e-elis3-120043526

Rumbold, J. M., & Pierscionek, B. K. (2017). A critique of the regulation of data science in healthcare research in the European Union. *BMC Medical Ethics*, *18*(1). doi:10.1186/s12910-017-0184-y

Saha, K., Kim, S. C., Reddy, M. D., Carter, A. J., Sharma, E., Haimson, O. L., & De Choudhury, M. (2019). The language of LGBTQ+ minority stress experiences on social media. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), 1-22. doi:10.1145/3361108

Saracevic, T. (2010). Information science. *School of Communication and Information. Rutgers University, News Brunswick, New Jersey, U.S.A.*

Schootman, M., Chien, L., Yun, S., & Pruitt, S. L. (2016). Explaining large mortality differences between adjacent counties: A cross-sectional study. *BMC Public Health*, *16*(1). doi:10.1186/s12889-016-3371-8

Sherman, R. L., Henry, K. A., Tannenbaum, S. L., Feaster, D. J., Kobetz, E., and Lee, D. J. (2014) 'Applying Spatial Analysis Tools in Public Health: An Example Using SaTScan to Detect Geographic Targets for Colorectal Cancer Screening Interventions', *Preventing Chronic Disease*, Vol. *11*. doi:10.5888/pcd11.130264

Shulan, M., Gao, K., and Moore, C. D. (2015) 'Thirty-day all-cause hospital readmissions - racial and income disparities and risk factors in a veterans integrated healthcare network', *International Journal of Healthcare Technology and Management*, Vol. *15, No.* 2, p.112. doi:10.1504/ijhtm.2015.074539

Smith, F. J. (2006). Data science as an academic discipline. *Data Science Journal*, *5*, 163-164. doi:10.2481/dsj.5.163

Smith, F. J. (2006). Data science as an academic discipline. *Data Science Journal*, *5*, 163-164. https://doi.org/10.2481/dsj.5.163

Song, I., & Zhu, Y. (2015). Big data and data science: what should we teach? *Expert Systems*, *33*(4), 364-373. doi:10.1111/exsy.12130

Sonnenwald, D. H, & Saracevic, T., (2016). Theory development in the information sciences. Chapter 8: *Relevance: In Search of A Theoretical Foundation.* Austin, TX: University of Texas Press.

Soylu, T. G., Elashkar, E., Aloudah, F., Ahmed, M., & Kitsantas, P. (2018). Racial/ethnic differences in health insurance adequacy and consistency among children: Evidence from the 2011/12 national survey of children's health. *Journal of Public Health Research*. doi:10.4081/jphr.2018.1280

Spruit, M., & Lytras, M. (2018). Applied data science in patient-centric healthcare: Adaptive analytic systems for empowering physicians and patients. *Telematics and Informatics*, *35*(4), 643-653. doi:10.1016/j.tele.2018.04.002

Stanton, J. M., & De, G. R. (2013). *An introduction to data science*. Sage.

Texas Department of State Health Services [TDSHS] (a). "Cancer incidence leading sites 2011-2015". [online] dataset, *Age-Adjusted Cancer Incidence in Texas, 2011-2015 Leading Sites by Sex and Race/Ethnicity*. https://www.dshs.texas.gov/tcr/data/incidence-and-mortality.aspx (Accessed 15 October 2018).

Texas Department of State Health Services [TDSHS] (b). "Cancer mortality leading causes 2011-2015". [online] dataset, *Age-Adjusted Cancer Mortality in Texas, 2012-2-16 by Sex and Race/Ethnicity*. https://www.dshs.texas.gov/tcr/data/incidence-and-mortality.aspx (Accessed 15 October 2018).

Texas Department of State Health Services [TDHSH] (c). *Welcome to the Texas Cancer Registry*. [online] https://www.dshs.texas.gov/tcr/home.aspx (Accessed 15 October 2018).

Thomas, J., & Harden, A. (2008). Methods for the thematic synthesis of qualitative research in systematic reviews. *BMC Medical Research Methodology*, *8*(1). doi:10.1186/1471-2288-8-45

Thurman, D. J., Kobau, R., Luo, Y., Helmers, S. L., & Zack, M. M. (2016). Health-care access among adults with epilepsy: The U.S. national health interview survey, 2010 and 2013. *Epilepsy & Behavior*, *55*, 184-188. doi:10.1016/j.yebeh.2015.10.028

Trujillo, M. D., & Plough, A. (2016). Building a culture of health: A new framework and measures for health and healthcare in America. *Social Science & Medicine*, *165*, 206-213. https://doi.org/10.1016/j.socscimed.2016.06.043

U.S. Cancer Statistics Working Group. U.S. Cancer Statistics Data Visualizations Tool, based on November 2017 submission data (1999-2015): U.S. Department of Health and Human Services, Centers for Disease Control and Prevention and National Cancer Institute; www.cdc.gov/cancer/dataviz, June 2018.

Upchurch, D. M., Krueger, E. A., & Wight, R. G. (2016). Sexual orientation differences in complementary health approaches among young adults in the United States. *Journal of Adolescent Health*, *59*(5), 562-569. doi:10.1016/j.jadohealth.2016.07.001

Van Rheenen, S., Watson, T. W., Alexander, S., & Hill, M. D. (2015). An analysis of spatial clustering of stroke types, in-hospital mortality, and reported risk factors in Alberta, Canada, using geographic information systems. *Canadian Journal of Neurological Sciences / Journal Canadien des Sciences Neurologiques*, *42*(5), 299-309. doi:10.1017/cjn.2015.241

Vierboom, Y. C., & Preston, S. H. (2020). Life beyond 65: Changing spatial patterns of survival at older ages in the United States, 2000–2016. *The Journals of Gerontology: Series B*, *75*(5), 1093-1103. doi:10.1093/geronb/gbz160

Wambui, K. M., & Musenge, E. (2019). A space-time analysis of recurrent malnutrition-related hospitalisations in Kilifi, Kenya for children under-5 years. *BMC Nutrition*, *5*(1). doi:10.1186/s40795-019-0296-5

West, K. M., Blacksher, E., & Burke, W. (2017). Genomics, health disparities, and missed opportunities for the nation's research agenda. *JAMA*, *317*(18), 1831. https://doi.org/10.1001/jama.2017.3096

Westra, B. L., Sylvia, M., Weinfurter, E. F., Pruinelli, L., Park, J. I., Dodd, D., Keenan, G., Senk, P., Richesson, R., Baukner, V., Cruz, C., Gao, G., Whittenburg, L., & Delaney, C. W. (2017). Big data science: A literature review of nursing research exemplars. *Nursing Outlook*, *65*(5), 549-561. https://doi.org/10.1016/j.outlook.2016.11.021

White, A., Coker, A. L., Du, X. L., Eggleston, K. S., and Williams, M. (2010). 'Racial/ethnic disparities in survival among men diagnosed with prostate cancer in Texas', *Cancer*, Vol. *117, No.* 5, pp.1080-1088. doi:10.1002/cncr.25671

Wilkinson, R. G., and Marmot, M. G. (2003) *Social Determinants of Health: The Solid Facts*, World Health Organization, Geneva, Switzerland.

Yadeta, T. A., Mengistu, B., Gobena, T., & Regassa, L. D. (2020). Spatial pattern of perinatal mortality and its determinants in Ethiopia: Data from ethiopian demographic and health survey 2016. *PLOS ONE*, *15*(11), e0242499. doi:10.1371/journal.pone.0242499

Zhang, X., Pérez-Stable, E. J., Bourne, P. E., Peprah, E., Duru, O. K., Breen, N., Berrigan, D., Wood, F., Jackson, J. S., Wong, D. W. S., Denny, J. (2017). Big data Science: Opportunities and challenges to address minority health and health disparities in the 21st century. *Ethnicity & Disease*, *27*(2), 95. https://doi.org/10.18865/ed.27.2.95