International Journal of Big Data and Analytics in Healthcare
Volume 6 • Issue 2 • July-December 2021

Using Data Science Software to Address Health Disparities

Jose O. Huerta, University of North Texas, USA
Gayle L. Prybutok, University of North Texas, USA
Victor Prybutok, University of North Texas, USA

ABSTRACT

The article assesses data science software to evaluate the usefulness of data science technology in addressing concerns such as health disparities. Data science software was analyzed using KDnuggets data related to analytics, data science, and machine learning software. Data science functionalities include computational processes and frameworks that are relevant for healthcare. This study demonstrates the importance of leading applications for conducting data science operations that can improve care in healthcare networks by addressing such factors as health disparities.

KEYWORDS

Analytics Software, Data Science, Data Science Software, Health Analytics Software, Health Disparities

INTRODUCTION

The application of data science in health care was studied by many professionals in the health care 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 health care organizations. Addressing health disparity issues allow health organizations to optimize patient care approaches and improve outcomes. Health care 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 health care. 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

DOI: 10.4018/IJBDAH.20210701.oa4

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.

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.

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). Health care 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 health care 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 health care (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 health care 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.

Data Science and the Data Scientist

To fully tap the potentials of data science, the health care field must develop a sector of well-qualified data science specialists focused on health care 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 health care 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).

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 (Data Science Central) 2014 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).

Another model that shows an overview of the data science process was developed by Cielen, Ali, and Meysman (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, Ali, & Meysman, 2016). Table 1 Data science analysis process in Table 1 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 1. Data science utilizes advanced methods to help determine predictions from the data used (Fayyad, 2012). Figure 1 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 health care data science applications must expect to map health care information needs onto their general outlines. Figure 1 developed from (Cielen, Ali, & Mysman, 2016) delineates the data science process steps map components.

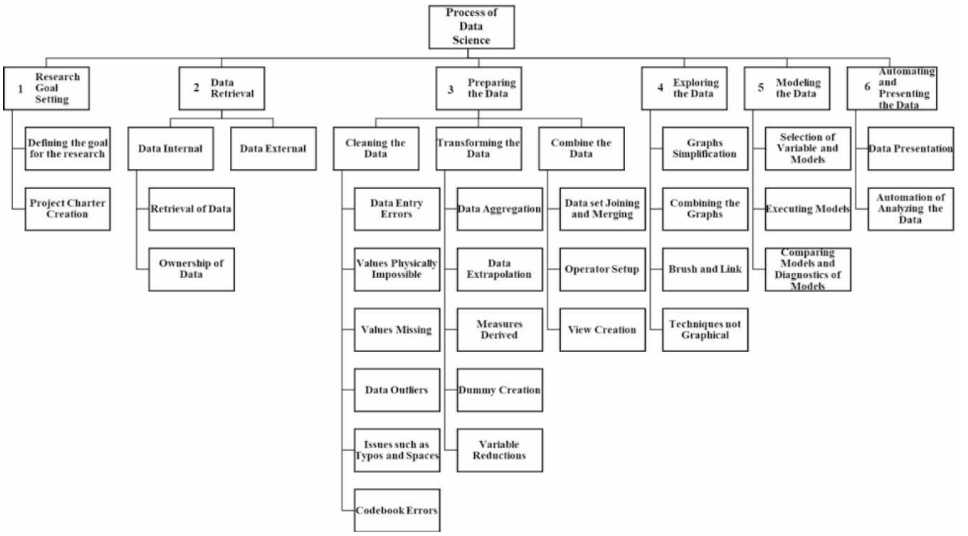
DATA SCIENCE IN HEALTH CARE

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 health care outcomes (Adam, Wieder, & Ghosh, 2017). One goal for data

Table 1. Data science analysis process

Preprocessing Steps	1. Goal Setting
	2. Obtain Data
	3. Data cleaning and formatting
Analysis Steps	4. Data exploration and summary
	5. Analytical methods
	6. Modeling
	7. Data automation and operationalization
Interpretation	8. Presentation
	9. Discussion and interpretation

Figure 1. Data Science Process Steps Map



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 health care. 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 health care 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 health care 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 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 health care, primarily because in the field of data science, most data comes 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.

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 are 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 health care. Disparities in health permeate and continue in diverse type of infirmities and become expensive to health organizations.

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 health care 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.

METHODOLOGY 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 2 presents how several software products reflect the available programs in data science. Among the software in the table, only software included in KDnugget's poll that categorically pertained to analytics, data science, and machine-learning software with a 30 percent 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 2 is sourced from (Piatetsky, 2019).

Subsequently, 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 2 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 were comprised of scores of 1 to 4 and included performance indicators of poor (1), bad (2), good (3), and excellent (4) (Gharaibeh, 2014). Tables 3 thru 6 are briefly mentioned and described in more detail prior to their insertion in the manuscript. Table 3 exhibits the new scoring model modified for data science software scoring. Table 4 exhibits the breakdown of the full assessment scored. Table 5 shows the number of functionality requirements met by functionality type. Total scores in Table 6 exhibit efforts to rank software derived from total scores.

Figure 2. Software 2019 Percentage Share

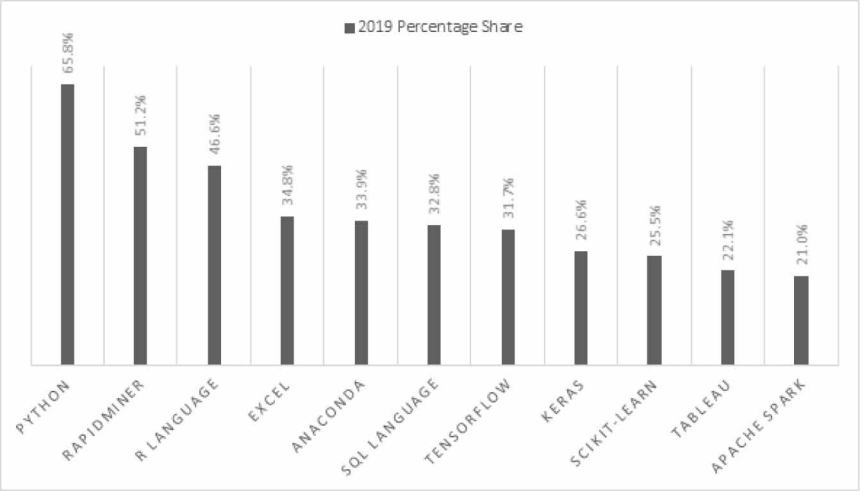


Table 2. Software Selection Sub-Criteria

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

Table 3. Data Science Software Scoring Model

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

RESULTS

Table 4 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.

There are 6 total categories and a total of 38 sub-criteria categories. Table 5 shows the results for sub-criteria as it pertains to the total number of types of functionality met. Python and Tensorflow met

Table 4. 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

Table 5. 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

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

The highest rank software programs exhibited in Table 6 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.

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

Table 6. 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 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 provides an opportunity 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.

REFERENCES

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 PMID:28122121

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. doi:10.1371/journal.pone.0157734 PMID:27455108

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. doi:10.1377/hlthaff.2017.1301 PMID:29505363

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), 41. Advance online publication. doi:10.1007/s10916-019-1162-3 PMID:30637593

Bhargava, N., Aziz, A., & Rajiv, A. (2013). Selection criteria for data mining software: A study. *IJCSI International Journal of Computer Sciences, 10*(3).

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

Chase, J. D., & Vega, A. (2016). Examining health disparities using data science. *Research in Gerontological Nursing*, *9*(3), 106–107. doi:10.3928/19404921-20160404-01 PMID:27210530

Chin, M. H. (2016). Creating the business case for achieving health equity. *Journal of General Internal Medicine*, *31*(7), 792–796. doi:10.1007/s11606-016-3604-7 PMID:26883523

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

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), 209. Advance online publication. doi:10.1007/s10916-016-0565-7 PMID:27520614

Delaney, C. W., & Westra, B. (2016). Big data. *Western Journal of Nursing Research*, *39*(1), 3–4. doi:10.1177/0193945916671687 PMID:30208772

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), 1133–1139. doi:10.1377/hlthaff.2016.0979 PMID:28583973

Detmer, D. E., & Shortliffe, E. H. (2014). Clinical informatics. *Journal of the American Medical Association*, *311*(20), 2067. doi:10.1001/jama.2014.3514 PMID:24823876

Dunn, M. C., & Bourne, P. E. (2017). Building the biomedical data science workforce. *PLoS Biology*, *15*(7), e2003082. doi:10.1371/journal.pbio.2003082 PMID:28715407

Erdman, S. L. (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. doi:10.4236/jsea.2014.71004

Hughes, M. C., Baker, T. A., Kim, H., & Valdes, E. G. (2019). Health behaviors and related disparities of insured adults with a health care provider in the United States, 2015–2016. *Preventive Medicine*, *120*, 42–49. doi:10.1016/j.ypmed.2019.01.004 PMID:30639668

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), S17–S20. Advance online publication. doi:10.1016/j.jbi.2007.08.010 PMID:17945542

King, C. (2016). Disparities in access to preventive health care services among insured children in a cross sectional study. *Medicine*, *95*(28), e4262. doi:10.1097/MD.0000000000004262 PMID:27428239

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?* O'Reilly Media. https://www.oreilly.com/data/free/what-is-data-science.csp

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

Manna, M. (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), S11–S16. Advance online publication. doi:10.1016/j.jbi.2007.08.008 PMID:17981099

McQuaid, E. L., & Landier, W. (2017). Cultural issues in medication adherence: Disparities and directions. *Journal of General Internal Medicine*, *33*(2), 200–206. doi:10.1007/s11606-017-4199-3 PMID:29204971

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*, 63-109. 10.1093/comjnl/bxw07310.28945/2289

O'Connor, S. (2018). Big data and data science in health care: What nurses and midwives need to know. *Journal of Clinical Nursing*, *27*(15-16), 2921–2922. doi:10.1111/jocn.14164 PMID:29148112

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

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

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

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

Press, G. (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

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), 27. Advance online publication. doi:10.1186/s12910-017-0184-y PMID:28388916

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

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 (Atlanta, Ga.)*.

Trujillo, M. D., & Plough, A. (2016). Building a culture of health: A new framework and measures for health and health care in America. *Social Science & Medicine*, *165*, 206–213. doi:10.1016/j.socscimed.2016.06.043 PMID:27405727

West, K. M., Blacksher, E., & Burke, W. (2017). Genomics, health disparities, and missed opportunities for the nation's research agenda. *Journal of the American Medical Association*, *317*(18), 1831. doi:10.1001/jama.2017.3096 PMID:28346599

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. doi:10.1016/j.outlook.2016.11.021 PMID:28057335

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. doi:10.18865/ed.27.2.95 PMID:28439179

Jose O. Huerta is an experienced healthcare information technology professional with over 20 years of experience in the healthcare information technology (HIT) space. He holds a Master of Science (MS) degree in business management from Troy University and a Bachelor of Science (BS) degree in business management from Park University. He holds two certifications in electronic health record technology from the American Health Information Management Association (AHIMA). He is a board member of the Texas Health Information Management Association (TXHIMA) and received his PhD in May 2021 at the University of North Texas in Denton, Texas.

Gayle Prybutok is an Assistant Professor, Health Services Administration in the Department of Rehabilitation and Health Services, College of Health and Public Service, at the University of North Texas. She served as the Coordinator for the Health Services Administration Master's Degree and was instrumental in creating that program and the Ph.D. in Health Services Research. Dr. Gayle Prybutok holds a Bachelor's in Nursing from Thomas Jefferson University, an MBA from Texas Woman's University and a Ph.D. in Information Science with a focus on Health Informatics from the University of North Texas. She formerly served as the Chief Nursing Officer of a local hospital, Director of home health and hospice agencies, and was the Executive Director of a national non-profit funded by NIH to procure human tissue for research. Her research interests include online health communication, health care quality improvement, and health education via the Internet.

Victor R. Prybutok is a Regents Professor of Decision Sciences in the Information Technology and Decision Sciences Department in UNT's G. Brint Ryan College of Business, Vice Provost for Graduate Education, and Dean of the Toulouse Graduate School at the University of North Texas. He received, from Drexel University, his B.S. with High Honors in 1974, an M.S. in Bio-Mathematics in 1976, an M.S. in Environmental Health in 1980, and a Ph.D. in Environmental Analysis and Applied Statistics in 1984. Dr. Prybutok is an American Society for Quality certified quality engineer, certified quality auditor, certified manager of quality / organizational excellence, and an accredited professional statistician (PSTAT®) by the American Statistical Association. He has authored over 200 journal articles, more than 350 conference presentations/proceedings, and several book chapters. In 2017 he received the American Society for Quality Gryna Award for a co-authored manuscript and in 2018 was awarded the Decision Sciences Institute Lifetime Distinguished Educator Award. Most recently he was awarded the 2020 Distinguished Service Award by the Southwest Region of the Decision Sciences Institute.