International Journal of Big Data and Analytics in Healthcare
Volume 8 • Issue 1

Data Visualization of Big Data for Predictive and Descriptive Analytics for Stroke, COVID-19, and Diabetes

Richard S. Segall, Arkansas State University, USA*
https://orcid.org/0000-0001-7627-2609

Soichiro Takashashi, Arkansas State University, USA

ABSTRACT

Visualization of big data is crucial for meaningful interpretations and especially for healthcare. Brief discussions are made for big data, background for healthcare, and recent work in big data analytics for healthcare. This research pertains to different levels of big data: 5,110 vs. 101,766 vs. 320,200 vs. 1 million data values. Data visualizations and predictive analytics are presented of big data for selected diseases of stroke with 5,110 data values, diabetes with 101,766 data values, and two COVID-19 studies: one with 320,200 data values and another with 1 million data values. Data visualizations are generated for these diseases with big data using Tableau. For stroke patients, an investigation was performed to determine how different living environments affect relationship between strokes. The data visualizations for diabetes showed impact of insulin use yielded reduced hospital stays. Data visualizations for COVID-19 provided temporal trends in confirmed cases, mortality, and recovery rates for 2020-2023. Conclusions and future directions of research are presented.

KEYWORDS

Big Data, Big Data Analytics, COVID-19, Descriptive Analytics, Diabetes, Predictive Analytics, Stroke, Visualization

INTRODUCTION

Big data was originally defined as the collection of datasets whose volume, velocity, or variety is so large that storing, managing, processing, and analyzing the data using traditional databases and data processing tools is complex (Bahga & Madisetti, 2016). According to an estimate in 2017 by IBM, 2.5 quintillion bytes of data is created daily, and 90% of the data in the world today was created in the last two years alone (Perry, 2017). Miele and Shockley (2013) authored a twenty-page IBM Executive Report titled "Analytics: The Real-World Use of Big Data" which is one of the first detailed discussions of big data.

In 2012, the United States (U.S.) government committed $200 million in big data research and development investment (The White House, 2012). Big data applications are estimated to be worth

DOI: 10.4018/IJBDAH.331996 *Corresponding Author

300 billion dollars for the U.S. healthcare industry and 250 billion euros for Europe's public sector administration (Manyika et al., 2011). So, what is big data? The numerical definition of big data is evolving with technological development. A dynamic definition is data that exceeds the capacity of commonly used hardware and software tools to capture, store, and analyze within a tolerable elapsed time is big data (Franks, 2012). Clegg (2017) authored a book on how the information revolution of big data is transforming our lives.

While Segall (2020a) discussed the crucial question "what is big data?," Segall (2020b) discussed open-source software for big data. Segall and Niu (2020) wrote an entire book on open-source software for the statistical analysis of big data. Segall and Cook (2018) completed a two-volume handbook of big data storage and visualization techniques.

This paper specifically addresses big data for three selected diseases of strokes, diabetes, and COVID-19 for databases at different levels of big data (5,110 vs. 101,766 vs. 320,200 vs. 1 million data values).

RESEARCH BACKGROUND AND MOTIVATIONS

Yee et al. (2020) discussed the implications of big data on healthcare and its future steps with uses for clinical decision-making, research and development, population health and surveillance, detecting fraud, prediction capabilities, Google Trends, and preventive measures. They referred to Chen et al. (2016), describing a cognitive computing tool developed by IBM. The tool, named Watson, has been applied to big data challenges in life sciences research by integrating and analyzing big data that includes medical literature, patents, genomics, and chemical and pharmacological data. Chen et al. (2016) specifically discussed the application of IBM Watson to explore big data for cancer kinases.

Healthcare applications that have used big data include those for cancer research, disease detection, and population health. Big data has changed how researchers understand diseases, providing access to patient information, trends, and patterns that were not accessible before. Companies that use big data in healthcare applications include:

- Cancer research carried out by Tempus in Chicago, Illinois (USA) and Flatiron Health in New York City (USA)
- Early disease detection by Pieces in Irving, Texas (USA) and Prognos in New York City (USA)
- Population health research conducted by Amitech in Creve Coeur, Missouri (USA), Linguamatics in Marlborough, Massachusetts, and Socially Determined in Washington, DC (USA). (Schroer, 2023)

Pramanik et al. (2022) provided a comprehensive overview of healthcare big data that extends the traditional 5 V's to 10 V's for healthcare big data: Volume, Velocity, Variety, Veracity, Validity, Viability, Volatility, Vulnerability, Visualization, and Value. Each of these are defined as below in Table 1 where the traditional 5 V's are listed as the first five.

Hogue and Bao (2016) discussed the challenges of big data in health care and analytics tools such as MapReduce, Spark, and Storm. Patil and Vohra (2021) edited a handbook of research that focuses on healthcare applications of data science and analytics.

Varatharajan et al. (2020) discussed an in-depth study of big data analytics in the healthcare industry by analyzing healthcare applications using machine learning. Varatharajan et al. (2020) discussed the need for big data analytics in the healthcare ecosystem with personal health records, electronic health records, health information exchange, and national and international health analytics.

Varatharajan et al. (2020) compared the biggest challenges in big data analytics of data complexity, data volumes, performance, skills, data velocity, and cost for Boston, San Francisco, and Chicago.

Table 1. The 10 V's for big data

Name of "V"	Application for Big Data
Volume	Size of Data
Velocity	The speed at which Data is generated
Variety	Different types of Data
Veracity	Data Accuracy
Value	Usefulness of Data
Validity	Data quality, such as accuracy and correctness
Volatility	How long the data is relevant
Vulnerability	Level of privacy and security of data
Visualization	Presents large and complex clinical reports in meaningful and understandable ways
Viability	Relevancy of data to generate desired and accurate outcomes through analytics and predictive measures.

Note: Derived using Pramanik et al. (2022)

Varatharajan et al. (2020) presented an in-depth discussion of streaming healthcare data using data analytics with stages that are summarized in Table 2.

Varatharajan et al. (2020) also presented several big data platforms and big data tools that are each discussed in more depth with others in Segall & Gao's (2020) work and in a chapter by Segall (2020b) pertaining to open-source software for the statistical analysis of big data. Varatharjan et al. (2020) showed that the three most important roles of big data analytics in healthcare are cost reduction, faster & better decision making, and new products and services.

Sambyal et al. (2019) presented applications, trends, tools, and limitations along future scopes of selected healthcare studies that used big data analytics. Sambyal et al. (2019) categorized the steps of big data analytics as (a) collection, (b) storage, (c) processing, and (d) visualization, and they presented a list of tools and a description for each of the steps.

Panigrahi and Singh (2017) presented a study on big data analytics in bioinformatics that includes the use of genetic algorithms, the development of a novel data life cycle on big data, and details and challenges connected with big data with special relevance to bioinformatics. Other related works about big data and healthcare include Borges do Nascimento (2021), Ranjan et al. (2021), and Sweeney et al. (2023).

Table 2. Stages for generalized healthcare data streaming

Stage	Inputs	Streaming Actions	Results
I	EKG, Arterial Blood Pressure *(*ABP*)*, pulse oximetry, etc.	Speed and volume up to 1000 samples per second per waveform.	Fast Data Ingestion
II	Demography, Labs, Allergies, Medications	Correlate and enrich with electronic health records, patient history, & other available resources.	Situational & Contextual Awareness
III	Linear & Non-linear Multi-Domain Analysis	Advanced custom analytics consumes relevant data and produces insights.	Signal Processing and Feature Extraction
IV	Diagnostic, Predictive & Prescriptive	Triggers best actions, alarms & clinical decision support.	Additional Insights

Note: Derived from figure by Varatharajan et al. (2020)

RESEARCH OBJECTIVES

The primary purpose of this research is to underscore the essential role of big data visualization in healthcare for informed and meaningful interpretations. It aims to explore the application of big data analytics in understanding the dynamics of specific diseases such as strokes, diabetes, and COVID-19, utilizing data sets of varying scales. The research further seeks to investigate the impact of living environments on stroke correlations and the influence of insulin usage on hospital stays for diabetes patients. Lastly, the study intends to provide temporal trends for COVID-19, encompassing confirmed cases, mortality rates, and recovery statistics over a defined period.

Big data analytics are valuable tools that have been used to investigate many studies for various health conditions. Table 3 provides examples from countries around the world.

Other related works pertaining to big data and data analytics for predicting and monitoring of health conditions and diseases include Alqaissi et al. (2022), Bansal et al. (2016), Catalyze (2022), Corsi et al. (2020), Foresee Medical (2022), National Academy of Medicine (2016), Roy et al. (2021), Singh et al. (2021), Sweeney et al. (2023), Venkatesh et al. (2019), Yadav et al. (2021) and Zhang (2020).

Discussions of technologies for using big data not specific to any particular data application include Barik et al. (2018), Jayashree and Swaminathan (2021), and Punia et al. (2021), and each may also be applied to healthcare data. Kahn et al. (2022) discussed systematic analysis of healthcare big data analytics for efficient care and disease diagnosis. Pathak et al. (2021) provided a survey on

Table 3. Studies of various health conditions

Authors	Country	Date	Description of Big Data Study
Chen and Liu	China	2023	Analyzes existing problems in Tumor Big Data visualization.
Dar et al.	India	2021	Deep learning models for detection and diagnosis of Alzheimer's Disease.
Dwivedi et al.	India	2021	Public health surveillance system of infectious diseases.
Gangal et al.	India	2021	Review of Prediction Models for healthcare using machine learning
Kudari	India	2021	Data analytics and machine learning for prediction, detection, and monitoring of chronic autoimmune diseases.
Sharma and Rani	India	2021	Application of genomic data using machine and deep learning with big data microarray data analysis for many cancer types.
Kasson	U. S. A. & Sweden	2020	Annual Review's *Infectious Disease Research in the Era of Big Data*.
Garattini et al.	United Kingdom (U.K.)	2019	Big data analytics, infectious diseases, and associated ethical impacts, including information derived from personal devices.
Wong et al.	Japan & China	2019	Discusses big data analytics and methods of artificial intelligence applicable to infectious disease data.
Lin et al.	China & Norway	2018	Survey of chronic diseases and health monitoring of big data.
Nagavci et al.	Macedonia	2018	Review of prediction of disease trends using big data analytics
Spurlock	U. S. A.	2018	Using data analytics to predict, detect, and monitor chronic autoimmune diseases.

tools for data analytics and data science. Raghupathi et al. (2022) used big data analytic approaches to Cancer blog text analysis.

LITERATURE REVIEW

This section provides an overview of descriptions of and related work on stroke, diabetes, and COVID-19. These three diseases are described, and their related work is studied more deeply with examples of data visualizations and predictive modeling. Tables 4 and 5 compare the big data used for these three diseases used for big data analytics performed and shown in this study.

Stroke

According to the National Heart, Lung, and Blood Institute (NHLBI) of the National Institutes of Health (NIH), a stroke, also known as "Transient Ischemic Attack" (TIA) or cerebrovascular accident, happens when blood flow to the brain is blocked. This blockage prevents the brain from getting oxygen and nutrients from the blood. Without oxygen and nutrients, brain cells begin to die within minutes. Sudden bleeding in the brain can also cause a stroke if it damages brain cells (NHLBI, 2023).

Sanchez-Acevedo et al. (2019) studied the methodology for detecting, preventing, and managing big data analysis for cardiovascular diseases (CVD). Kadam et al. (2019) proposed a framework using a big data approach for CVD prediction using data mining techniques involving Artificial Neural Networks (ANN) and used Apache Spark for the implementation.

Diabetes

According to the Mayo Clinic (2023), diabetes, also called "diabetes mellitus," is a group of diseases that result in too much sugar in the blood and is referred to as "high blood glucose." The most common types are Type 1, Type 2, and Prediabetes. Type 1 diabetes is a chronic condition in which the pancreas produces little or no insulin. Type 2 diabetes is a chronic condition affecting how the body produces blood sugar, known as glucose. Prediabetes is a condition in which blood sugar is high, but not enough for type 2 diabetes.

Rastogi et al. (2021) discussed the surveillance of Type 1 and 2 diabetic subjects on physical characteristics using the Internet of Things (IoT) with big data perspectives.

COVID-19

According to the World Health Organization (WHO, 2023), Coronavirus (COVID-19) is an infectious disease caused by the SARS-CoV-2 virus. WHO first learned of this new virus on 31 December 2019, following a report of a cluster of cases of viral pneumonia in Wuhan, People's Republic of China.

Tenali and Babu (2023) presented a Systemic Literature Review (SLR) and future perspectives for handling big data analytics in COVID-19 diagnosis that also includes a systematic literature review of big data applications for heart disease diagnosis, kidney disease classification, brain tumor, and diabetic disease identification. Tenali and Babu (2023) presented a general architecture of COVID diagnosis using big data analytics using machine learning, deep learning, imaging modalities, and data mining techniques for COVID diagnosis.

RESEARCH METHODOLOGY

For this study, we used several publicly available datasets. Each was selected to demonstrate different characteristics. These datasets were then analyzed using appropriate methods to produce insight and visualizations of value for healthcare professionals. The study is exploratory and illustrative of the types of approaches that can be employed in this sector. Each of the three diseases studied has a separate section with more complete descriptions of the specific datasets, analytics, and visualizations.

EXPERIMENTAL DESIGN AND PERFORMANCE EVALUATION

This section provides the design background to show how big data is useful for each of these diseases in the context of healthcare professionals. For each case, we provide further insight on the datasets and analytics/visualizations.

Datasets Collection

This section outlines the datasets used for the analysis, allowing big data comparisons used for the diseases of stroke, diabetes, and COVID.

Table 4. Diseases and their characteristics of big data used in this research

Name of Disease	Number of Data Values	Characteristics of Data
Stroke	5,110	Globally gathered Focused on gender, age, smoking, and environmental statements Body indexes such as BMI, glucose, and blood pressure are also focused on
COVID-19 (Example 1)	320,200	Data is collected worldwide The major characteristics are confirmed cases, deaths, new cases, and hospitalized patients Data about the vaccinations is also collected, such as the number of vaccinated populations and the number of either positives or negatives
COVID-19 (Example 2)	1,000,000	Data is collected worldwide The significant difference feature from the other COVID-19 dataset is including the latitudes and longitudes, which are helpful to visualize Geography The data about the recovery rates from COVID-19 is also included
Diabetes	101,766	This dataset encompasses ten years (1999-2008) of clinical care within a network of 130 hospitals and integrated healthcare delivery systems in the United States Comprises a comprehensive set of over 50 patient demographics and hospital outcomes attributes.

Table 5. Sources of big data used for data visualization

Name of Disease	Date of Data Creation	URL of Data Source
Stroke	1 July 2022	https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset
COVID-19 (Example 1)	31 December 2019	World Health Organization (WHO) https://covid19.who.int/data
COVID-19 (Example 2)	21 January 2021	Bing collects data from multiple trusted, reliable sources, including the World Health Organization (WHO), Centers for Disease Control and Prevention (CDC), National/Regional and State Public Health Departments, BNO News, Wikipedia, and 24/7 Wall St. WHO: https://www.who.int/emergencies/diseases/novel-coronavirus-2019 CDC: https://www.cdc.gov/coronavirus/2019-ncov/index.html BNO News: https://bnonews.com/index.php/2020/04/the-latest-coronavirus-cases/ Wikipedia: https://en.wikipedia.org/wiki/COVID-19_pandemic 24/7 Wall St.: https://247wallst.com/
Diabetes	1990-2008	https://www.kaggle.com/datasets/brandao/diabetes/versions/1?resource=download

Data Visualization and Analytics of Big Data for Stroke (Cardiovascular Disease)

Dataset Description

As per the World Health Organization (WHO), strokes are the second most prevalent global mortality factor, contributing to approximately 11% of the overall disease count. The dataset is leveraged to forecast the probability of a stroke occurrence in patients by analyzing input parameters encompassing gender, age, diverse ailments, and smoking habits. Each entry within the dataset furnishes pertinent details concerning the individual patient. There are 12 columns and 5,110 rows.

Data Visualization for Stroke

Figure 1 depicts the relationship between the age distribution of stroke patients and the frequency of patients within each age category. It uses an area graph, representing data by filling in areas and allowing for easy observation of patterns and trends. The horizontal axis represents different age categories, while the vertical axis represents the frequency of stroke patients. Additionally, the graph includes information about cigarette smoking in relation to stroke, with different categories such as "formerly smoked," "never smoked," "smoke," and "unknown" represented by different colors.

By analyzing the graph in Figure 1, it becomes clear that there is a correlation between age and stroke occurrence. The number of stroke patients increases with age, and older individuals are more susceptible to experiencing strokes. However, an intriguing finding is the high count of patients aged 14, which surpasses the figures observed in individuals in their 20s and below. This odd finding implies that strokes can also impact younger people, but more research is needed to determine why this difference exists.

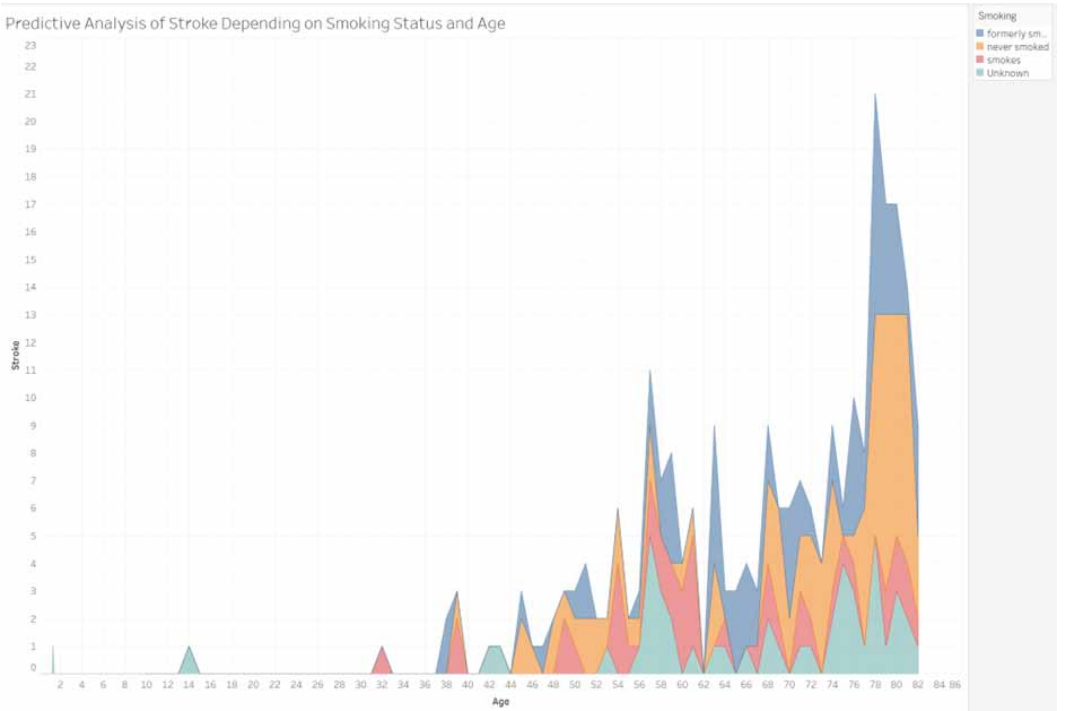
Moreover, Figure 1 provides insights into the relationship between cigarette smoking and stroke. The "never smoke" category shows a higher frequency of stroke patients compared to the "smoke" category. This suggests that continuous smoking may not significantly impact stroke occurrence, while non-smokers have a lower likelihood of experiencing strokes. It is important to note that a category labeled as "unknown" is represented by the green color. This category might include individuals whose smoking status is not known or not disclosed and whose stroke occurrence is not clearly determined.

Table 6. Attribute information for stroke big data

Column No.	Attributes	Description
1	Id	Unique identifier
2	Gender	Male, female, or other
3	Age	Age of patient
4	Hypertension	0 if the patient does not have hypertension, 1 if the patient has hypertension
5	Heart Disease	0 if the patient does not have heart disease, 1 if the patient has heart disease
6	Ever Married	Yes, or no
7	Work Type	Children, government job, never-worked, private, or self-employed
8	Residence Type	Rural or urban
9	Average Glucose Level	The average glucose level in blood
10	BMI	Body mass index
11	Smoking Status	Formerly smoked, never smoked, smoke, or unknown
12	Stroke	0 if the patient does not have stroke, 1 if the patient has stroke

Source: Fedesoriano (2021).

Figure 1. Predictive analysis of stroke depending on smoking status and age



In summary, the area graph of Figure 1 allows for a comprehensive understanding of the correlation between age distribution, smoking, and stroke. It highlights the increased risk of stroke with advancing age and suggests that consistent smoking may not be a significant contributing factor.

This data visualization comprises horizontal bar charts, which are extensively acknowledged for their aptitude in effectively conveying data in a lucid and comprehensible manner. Horizontal bar graphs excel at enhancing data visibility and facilitating information interpretation. Adjusting bars' length and position facilitates the identification of data trends, and the evaluation of relative relationships.

In Figure 2, the number of patients with hypertension, which causes stroke, and the number of patients with heart disease overall are categorized by gender. Age is also color-coded, with darker colors representing the elderly and lighter colors representing younger people; the darker colors in all three graphs suggest that these diseases are more common among the elderly. Also, regardless of gender, the highest percentage of patients appear to have hypertension compared to stroke and heart disease.

Furthermore, when the three bars in Figure 2 are compared by gender, there is a significant difference in the number of stroke and heart disease patients. In women, the number of stroke patients exceeds that of heart disease patients, while the opposite is true in men. Therefore, it is difficult to say that there is a direct causal relationship between heart disease and stroke. However, the present data clearly show a strong correlation between the prevalence of hypertension and the incidence of strokes in both men and women. This observation indicates that hypertension is a primary underlying condition causing stroke predisposition.

Figure 3 presents a data visualization performed to produce a graphical analysis that illustrates how different living environments affect the occurrence of strokes. The living environment was categorized based on whether the respondent was married or unmarried, whether they lived in the city or the country, and their occupation. Married or unmarried is indicated by "yes" or "no," and

Figure 2. Relationship between stroke, hypertension, and heart disease

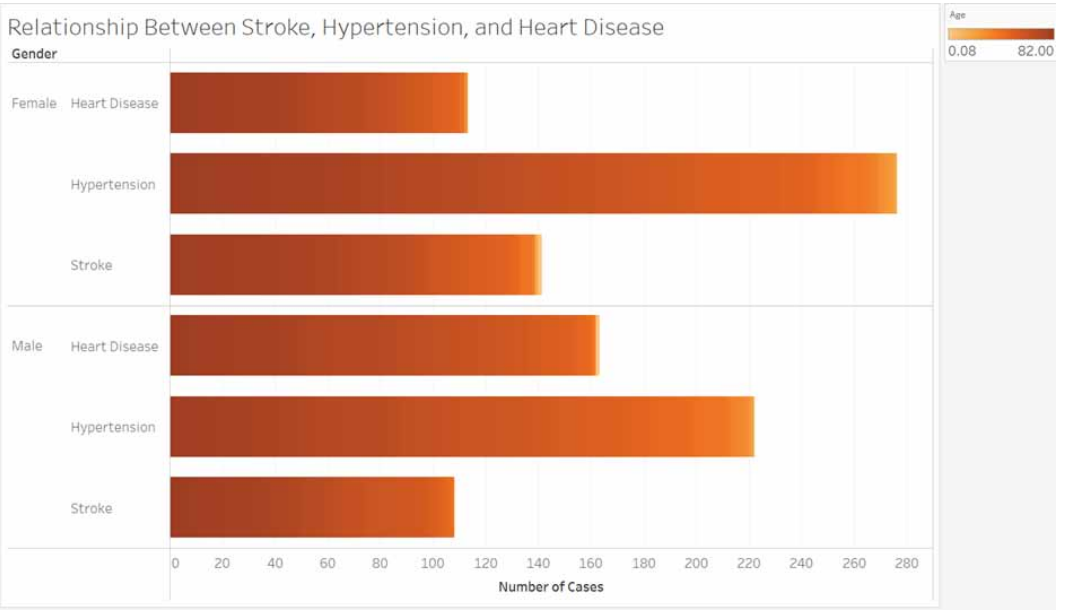
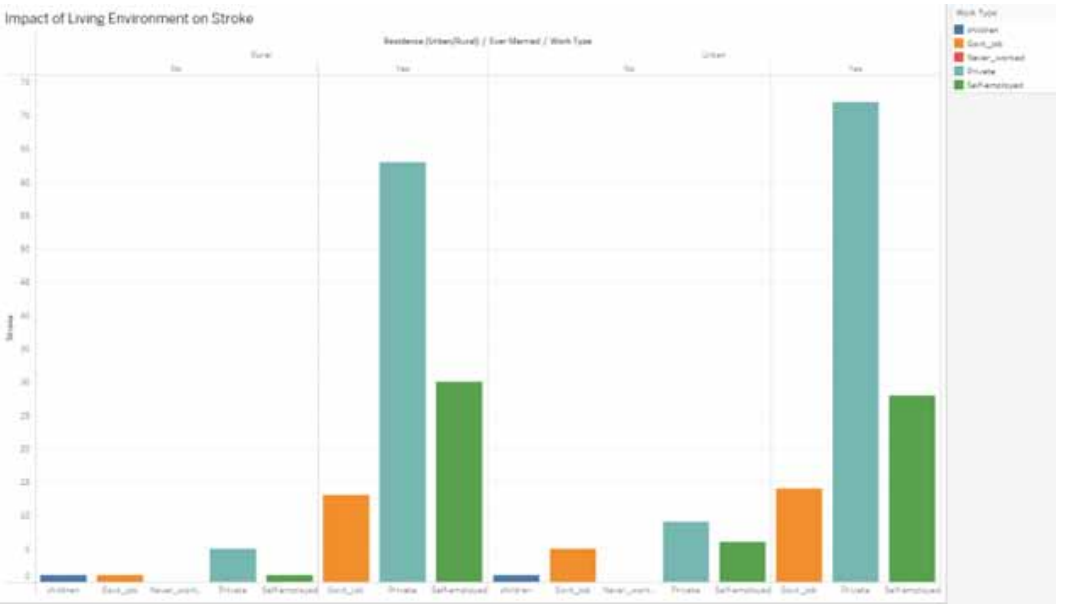


Figure 3. Impact of living environment on strokes



the type of work is indicated by “children” in blue, “government job” in orange, “never worked” in red, “private” in light blue, and “self-employed” in green.

Figure 3 reveals two primary differences between urban and rural living environments, and each item is classified according to whether the respondent is married or unmarried. Interestingly, regardless of whether one lives in the city or the countryside, there is a marked difference between

married and unmarried people. Unmarried persons have a lower incidence of stroke than married persons. Furthermore, when looking at work type, the incidence of stroke is much higher among married people than among other types of work. In particular, "private" workers are the most likely to have a stroke in any living environment. On the other hand, the "never-worked" type workers had the lowest incidence of stroke.

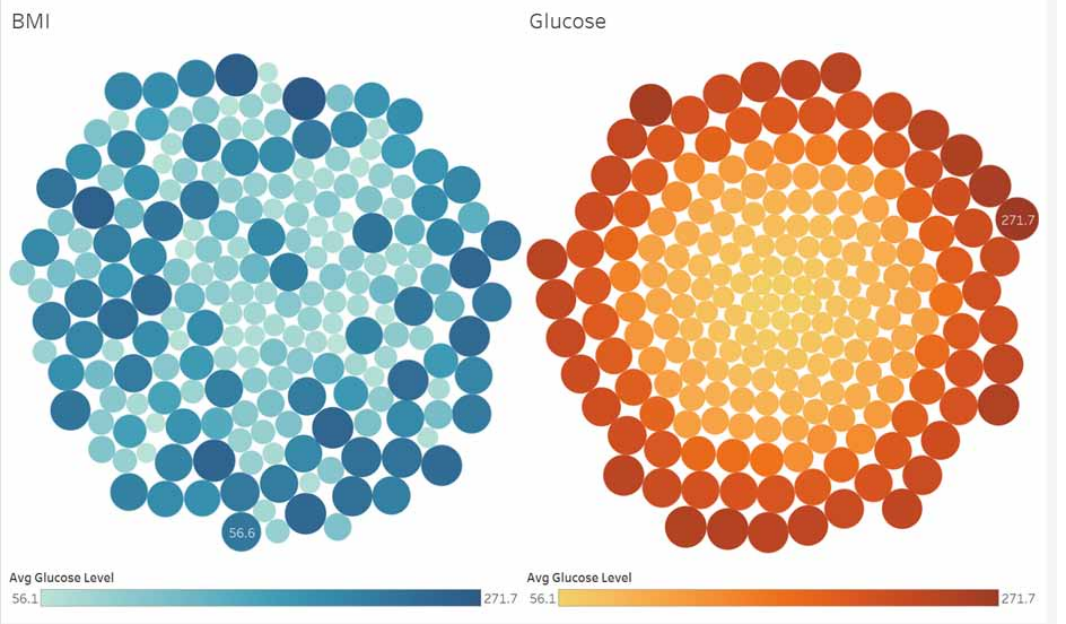
These findings of Figure 3 indicate that occupational stress plays a significant role in the etiology of stroke. Nevertheless, when examining the residential context, no discernible disparities of a comprehensive nature emerged between urban and rural locales concerning stroke prevalence. Therefore, the proposition that your living environment—urban or rural—is a major factor in stroke incidence, appears improbable.

Figure 4 is a visualization of two indices of physical health in people who have suffered a stroke. Each index is represented in a graph, reflecting the characteristics of stroke patients. The darker color of the circles on the graph indicates the magnitude of each value. The lighter the color, the smaller the value; the darker the color, the larger the value.

The left graph of Figure 4 is based on Body Mass Index (BMI). BMI measures the relationship between body weight and height and is widely used as an indicator of health risk. Observation of the graph for stroke patients reveals variations in the data, even after filtering. This variation suggests that BMI may not be the primary factor directly influencing stroke. The right graph of Figure 4 is based on the blood glucose levels of individual patients. Blood glucose levels are expressed as a glucose index. Observation of this graph reveals a regularity in which the values increase from the inside to the outside. This result suggests that blood glucose levels are more likely to cause stroke when compared to BMI results. In other words, the lower the blood glucose level, the lower the risk of stroke may be.

These findings indicate that blood glucose could serve as a more significant parameter than BMI when evaluating the well-being of individuals who have experienced a stroke. In forthcoming investigations, regulating blood glucose levels could be crucial in preventing and treating strokes.

Figure 4. Association between body health indicators: BMI and glucose level



Predictive Analytics for Stroke

Predictive analytics for the stroke data of 5,110 data values are described in Table 6 using Minitab statistical software (Minitab LLC, 2023).

The most pronounced deviations from a linear relationship in Figure 5 occur for the average glucose level, indicating it does not fit a normal distribution very well. Body Mass Index (BMI) fits a linear relationship up to a level of about 99%, after which deviations occur, indicating a fit to normal distribution up to about 99%. The remaining variables in Figure 5 fit a normal distribution very well, as indicated by the near-perfect linear relationship.

Figure 6 is the Analysis of Variance (ANOVA) table for the sources of data from Table 6 and Figure 5. In the ANOVA table of Figure 6, it can be noted that the attributes of average glucose level and stroke have the most significant F-values. In contrast, attributes of BMI, gender, and work type have the least significant F-values.

Data Visualization of Big Data for COVID-19: Example #1

Dataset Description

The dataset for our first example of COVID-19 disease is maintained by "Our World in Data." (Caesar, 2023). This dataset was updated daily during the COVID-19 pandemic. The dataset is available in CSV, XLSX, and JSON formats, with the CSV and XLSX files providing one row of data per location and date. The COVID-19 dataset contains essential data related to COVID-19, such as confirmed cases, deaths, hospitalizations, tests, and other relevant variables. The dataset encompasses a comprehensive collection of 67 distinct rows, each offering unique data points. The initial record dates back to January 3, 2020, while the most recent corresponds to May 17, 2023. Notably, the dataset comprises a vast accumulation of 320,200 rows of extensive data about research-related attributes, as delineated in

Figure 5. Probability plot of hypertension, heart disease, glucose, body mass index (BMI), and stroke for 5110 patients

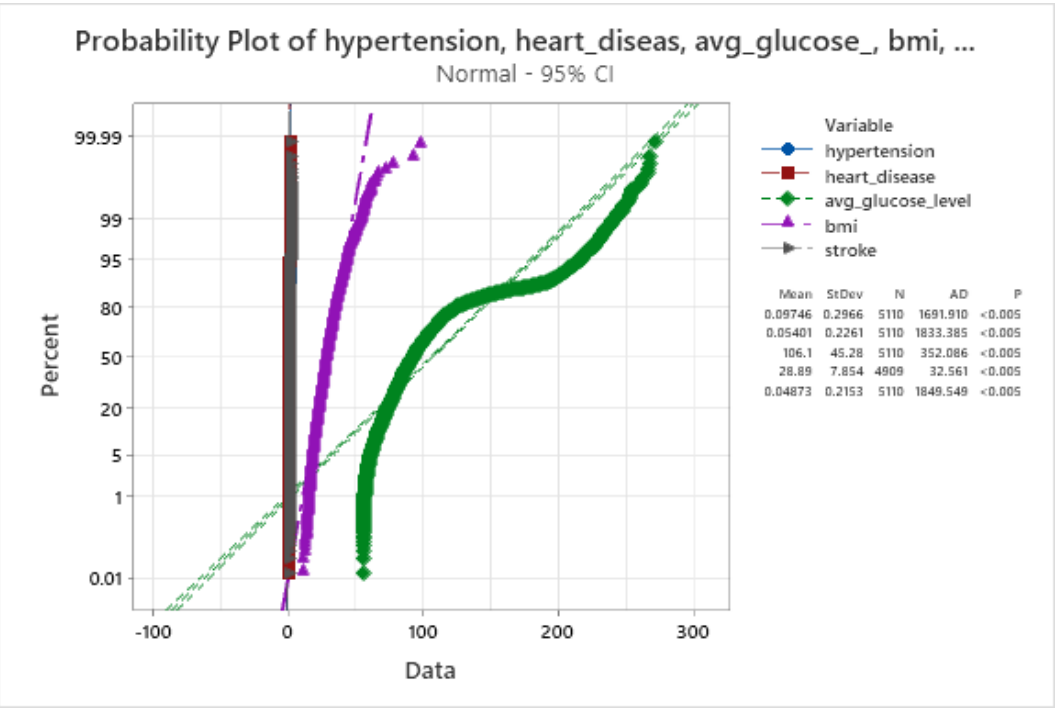


Figure 6. Analysis of variance (ANOVA) results for attributes of big data for stroke

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	10	13.078	1.30784	29.40	0.000
hypertension	1	0.988	0.98755	22.20	0.000
avg_glucose_level	1	2.971	2.97142	66.79	0.000
bmi	1	0.151	0.15080	3.39	0.066
stroke	1	2.393	2.39263	53.78	0.000
gender	2	1.521	0.76048	17.09	0.000
work_type	4	1.765	0.44123	9.92	0.000
Error	4898	217.893	0.04449		
Lack-of-Fit	4895	217.393	0.04441	0.27	0.990
Pure Error	3	0.500	0.16667		
Total	4908	230.971			

the ensuing tables. Only those attributes most essential for this study from this vast amount of data are listed in Table 7.

Data Visualizations for Example 1 of COVID-19

Figure 7 shows a time series of the number of COVID-19 cases and fatalities from 2020 to 2023 for the period of record. The graph is color-coded for each of the six continents, visually indicating the scale of COVID-19 damage on each continent.

As seen in Figure 7, the increase and decrease trends for cases and fatalities show a similar pattern over time. Specifically, the European continent, delineated in a light blue hue, exhibits the highest degree of impact, surpassing the remaining five continents consistently throughout the months spanning from 2020 to 2023. Asia, North America, and South America are next most affected, in that order. On the other hand, the Oceania region shows significantly less damage than the other regions. This phenomenon can be partly attributed to the relatively modest population of Oceania.

Moreover, comparing the incidence of cases to the fatality rates across continents in Figure 7, it is generally observed that the number of cases tends to be greater than the number of fatalities, except in the African and North and South American continents. In these continents, the mortality rate is observed to surpass the incidence rate.

Figure 8 has been created to examine the proportion of individuals who have attained the entire course of COVID-19 vaccination across various continents. Furthermore, a line graph has been included to depict the fluctuations in population size per continent, thereby enabling an assessment of the ratio between the number of vaccinated individuals and the overall population.

Figure 8 shows the number of cases and deaths where the European continent has a higher percentage than the other continents, except for Asia. The figure reveals that the Asian continent in orange has a more significant percentage, thus indicating a higher awareness of prevention against COVID-19. This awareness suggests that individuals residing within the Asian continent proactively receive vaccinations, and the outcomes are conspicuous.

However, when we analyze the population proportion, it is evident that a comparable proportion of individuals successfully undergo vaccination in relation to the European populace. This observation does not imply an absence of preventive consciousness among the European population; rather, attention could be directed towards evaluating the efficacy level of the vaccine.

Figure 9 focuses on the effectiveness of the COVID-19 vaccine and allows analysis of the extent to which the vaccine has reduced the damage caused by COVID-19. Figure 9 comprises two components,

Table 7. Attributes information for COVID-19 with 311,511 rows

Column No.	Attributes	Description
1	total_case	Total confirmed cases of COVID-19. Counts can include probable cases, where reported.
2	new_case	New confirmed cases of COVID-19. Counts can include probable cases, where reported. In rare cases where our source reports a negative daily change due to a date correction, we set this metric to NA.
3	total_cases_per_million	Total confirmed cases of COVID-19 per 1,000,000 people. Counts can include probable cases, where reported.
4	new_case_per_million	New confirmed cases of COVID-19 per 1,000,000 people. Counts can include probable cases, where reported.
5	total_deaths	Total deaths attributed to COVID-19. Counts can include probable deaths, where reported.
6	new_deaths	New deaths attributed to COVID-19. Counts can include probable deaths, where reported. In rare cases where our source reports a negative daily change due to a data correction, we set this metric to NA.
7	total_deaths_per_million	Total deaths attributed to COVID-19 per 1,000,000 people. Counts can include probable deaths, where reported.
8	new_deaths_per_million	New deaths attributed to COVID-19 per 1,000,000 people. Counts can include probable deaths, where reported.
9	icu_patients	Number of COVID-19 patients in intensive care units (ICUs) on a given day.
10	hosp_patients	Number of COVID-19 patients in hospital on a given day.
11	total_tests	Total tests for COVID-19.
12	new_tests	New tests for COVID-19 (only calculated for consecutive days).
13	positive_rate	The share of COVID-19 tests that are positive, given as a rolling 7-day average.
14	total_vaccinations	Total number of COVID-19 vaccination doses administered.
15	people_vaccinated	Total number of people who received at least one vaccine dose.
16	people_fully_vaccinated	Total number of people who received all doses prescribed by the initial vaccination protocol.
17	iso_code	ISO 3166-1 alpha-3 – three-letter country codes. OWID-defined regions (e.g., continents like 'Europe') contain the prefix 'OWID_'.
18	continent	Continent of the geographical location
19	location	Geographical location
20	date	Date of observation
21	population	Population (latest available values). See https://github.com/owid/covid-19-data/blob/master/scripts/input/un/population_latest.csv for a full list of sources.
22	female_smokers	Share of women who smoke, most recent year available.
23	male_smokers	Share of men who smoke, most recent year available.

Source: Caesar (2023) Our World in Data - COVID-19

employing data pertaining to the count of individuals who have received vaccination as the shared variable. The upper graph of Figure 7 compares the number of vaccinated individuals with the tally of COVID-19-related fatalities. In contrast, Figure 9 juxtaposes the number of vaccinated individuals with the incidence of COVID-19 cases. In the color coding within Figure 9, green indicates the number of vaccinations, orange indicates the number of deaths, and blue indicates the number of cases.

Comparing the two variables in this way reveals significant differences. Figure 9 depicts the statistically significant relationship between the quantity of COVID-19 cases and the number of individuals receiving vaccinations. There is a discernible upsurge in the count of vaccinations commencing in March 2021, which decelerates the rate of growth in COVID-19 cases and mitigates

Figure 7. Monthly COVID-19 cases and deaths across six continents (January 2020–May 2023)

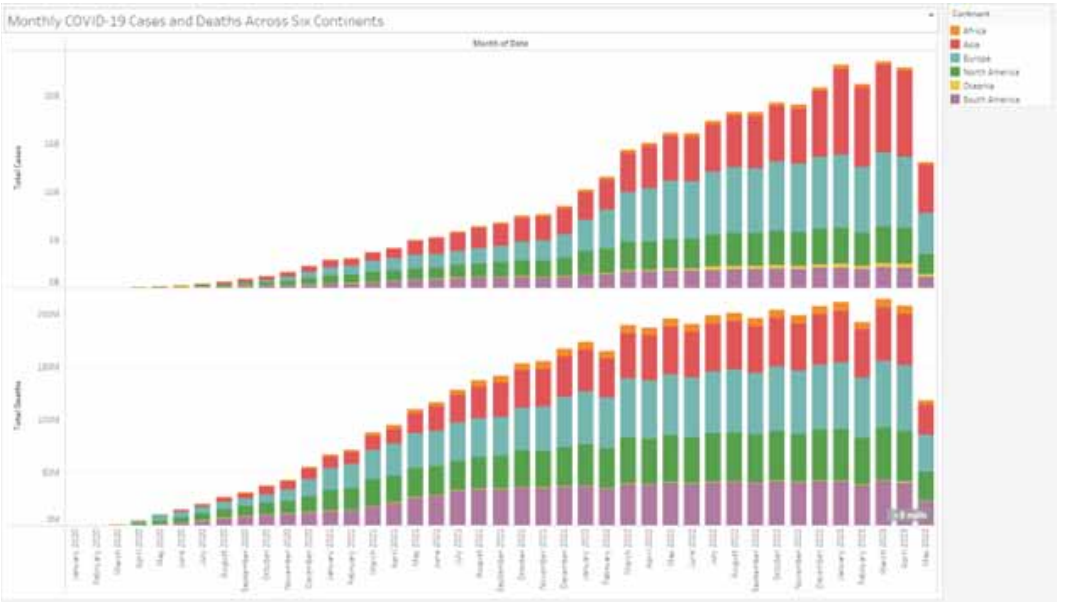
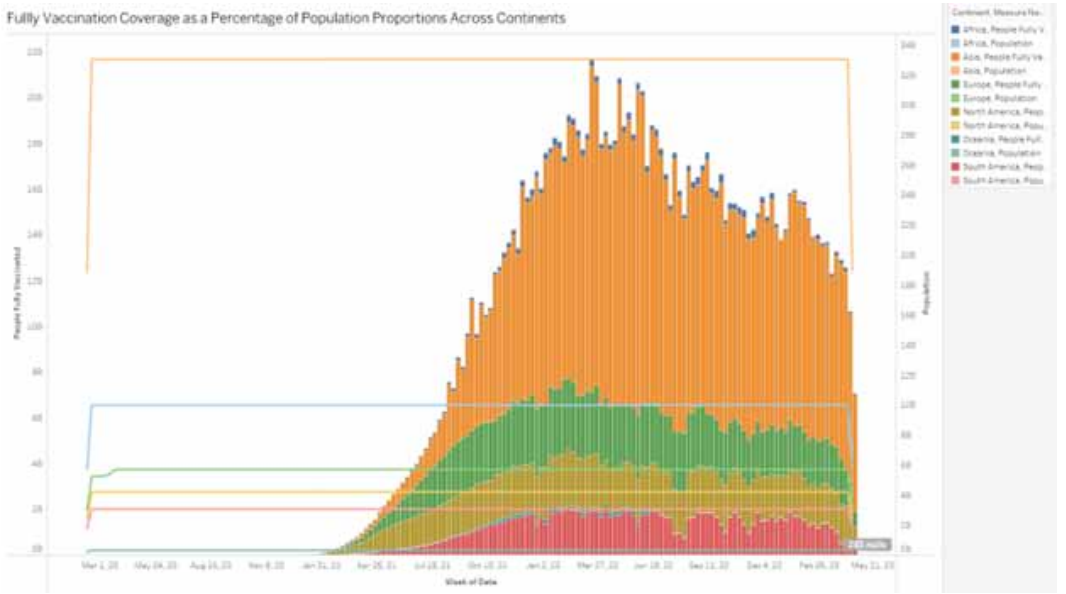


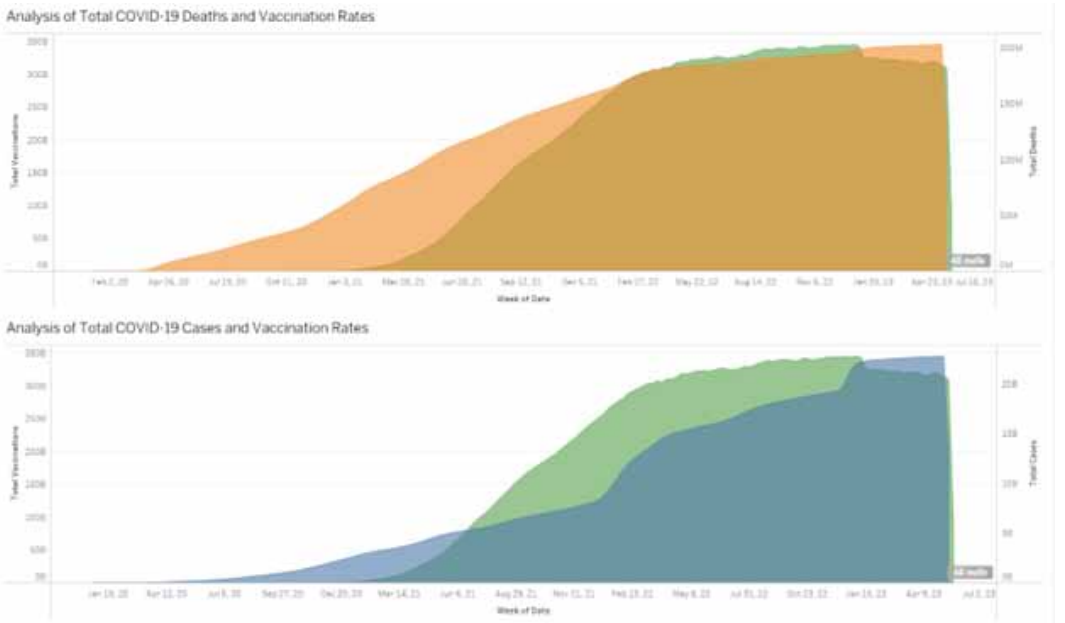
Figure 8. COVID-19 full vaccination coverage as a percentage of population proportions across continents



their overall magnitude. However, analysis of the top graph of the number of deaths vs. the number of vaccinations reveals that the number of deaths due to COVID-19 continues to increase beyond when vaccination was initiated.

These disparities suggest that the primary impact of the vaccine lies in diminishing the incidence of COVID-19 infection, primarily serving to regulate the occurrence of fresh infections. Conversely, its influence on mitigating the prevalence of severe cases or fatalities may be indirect or limited.

Figure 9. Analysis of total COVID-19 deaths, cases, and vaccination rates (February 2020–April 2023)



Data Visualization and Analytics for Big Data for COVID-19: Example #2

Dataset Description

Microsoft has made the dataset available for scholarly and pedagogical endeavors via the Bing web browser listed in Table 5. The collection of COVID-19 data commenced on the 11th of May 2020 and encompasses the latest information up until the present year of 2023 and consists of one million data values. The dataset used is a representative sample of the one million data values available and encompasses a magnitude of roughly 125,000+ rows distributed across 17 columns. The subsequent table presents meticulous attribute information of each column.

Data Visualizations for Example 2 of COVID-19

Figure 10 illustrates the temporal progression of COVID-19 incidence across multiple nations from 2020 to 2023. The color gradation conveys meaningful information, with lighter hues representing lower COVID-19 case counts and darker shades denoting higher case volumes.

First, let us discuss the United States, which has the most prominent cases. The number of COVID-19 cases in the U.S. is the most severely affected worldwide, as indicated by the darker color on the graph. This color intensity is determined based on indicators such as the number of cases and deaths. The damage in the U.S. is alarmingly extensive compared to India, Brazil, and European countries. The large population of the U.S. is thought to be a factor in the rapid spread of the infection and the burden on the medical system.

Also, when analyzed by continent, the African continent has suffered relatively minor damage compared to countries on other continents. There are multiple possible reasons for this. The relatively low population density of the African continent compared to other continents and the implementation of strict countermeasures and early quarantine measures may have contributed to this. At the same time, countries in some parts of the African continent are suffering average damage. The differences indicate that each country strives to control the spread of the infection through various

Table 8. COVID-19 dataset of 125,000+ rows description

Column No.	Attribute Name	Description of Attribute
1	id	Unique identifier
2	updated	The as at date for the record
3	confirmed	Confirmed case count for the region
4	confirmed_change	Change of confirmed case count from the previous day
5	deaths	Death case count for the region
6	deaths_change	Change of death count from the previous day
7	recovered	Recorded count for the region
8	recovered_change	Change of recovered case count from the previous day
9	latitude	Latitude of the code identifier
10	longitude	Longitude of the centroid of the region
11	ios2	2 letter country code identifiers
12	ios3	3 letter country code identifiers
13	country_region	Country/region
14	admin_region_1	Region with country_region
15	admin_region_2	Region with country_region_1
16	ios_subdivision	Two-part ISO subdivision code
17	load_time	The date and time the file was loaded from the Bing source on GitHub.

Source: Kemp & Mabee (2023, May 5)

countermeasures. However, since the situation differs from country to country, each country may be adopting its own strategies.

In general, Figure 10 visually represents the damage caused by COVID-19 by country and shows that the situation is particularly severe in the United States and some other countries. Interpreting these results in more detail through further research and data analysis is important.

In Figure 11, the measure names corresponding to different colors used are (a) Confirmed, (b) Deaths, (c) Recovered, and (d) Recovered (2). Examining Figure 11, which juxtaposes the data of the incidence of COVID-19 cases and fatalities against the proportion of individuals recuperating, the prevalence of COVID-19 cases and the mortality rate exhibit a persistent upward trajectory. As shown in Figure 11, the percentage of deaths is greater than the number of infected persons.

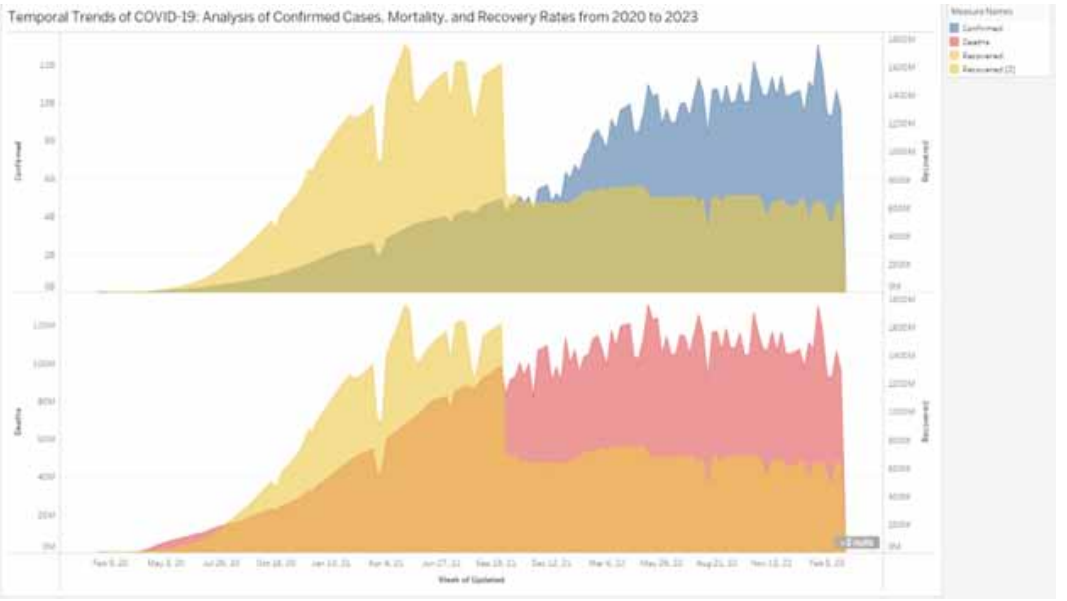
Upon analyzing the trajectory of recovered cases, it becomes evident that the trend in Figure 11 exhibits a favorable pattern until September 2021, characterized by a significant volume of recovered cases concerning the number of infected cases and the fatality rate. A substantial proportion of individuals afflicted with COVID-19 have successfully recuperated.

Since September 2021, in Figure 11, there has been a noticeable decline in the number of individuals in the process of recovering, thus providing clear evidence that more than half of these individuals have faced job losses. After this point, the number of recovered individuals has remained consistently stable without significant fluctuations. This observation implies a concerning correlation between the decline in recovery rates and the adverse impact on employment for the affected individuals.

Concerning the status of COVID-19 infections, the number of infections and deaths has been increasing while the number of recovered persons has been decreasing. This shift indicates that the spread of COVID-19 infection is continuing. Future measures need to be taken to reduce the spread of infection.

Figure 10. Mapping the distribution of confirmed COVID-19 cases

Figure 11. Temporal trends of COVID-19: Analysis of confirmed cases, mortality, and recovery rates weekly from February 2020 to February 2023



Predictive Analytics for Example #2 of COVID-19

Using Minitab statistical software (Minitab, 2023), linear regression modeling was performed using 591,101 data values for the number of deaths due to COVID-19 and the number recovered from COVID-19 to yield the following equation and Model Summary as shown in Figure 12 and Analysis of Variance as shown in Figure 13. The regression equation is

Figure 12. Model summary statistics for regression equation of number of deaths and number recovered from COVID-19

Model Summary

S	R-sq	R-sq(adj)
23402.0	76.22%	76.22%

Figure 13. Analysis of variance for linear regression equation of number of deaths and numbered recovered from COVID-19

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	1.04611E+15	1.04611E+15	1910163.22	0.000
Error	596100	3.26457E+14	5.47655E+08		
Total	596101	1.37257E+15			

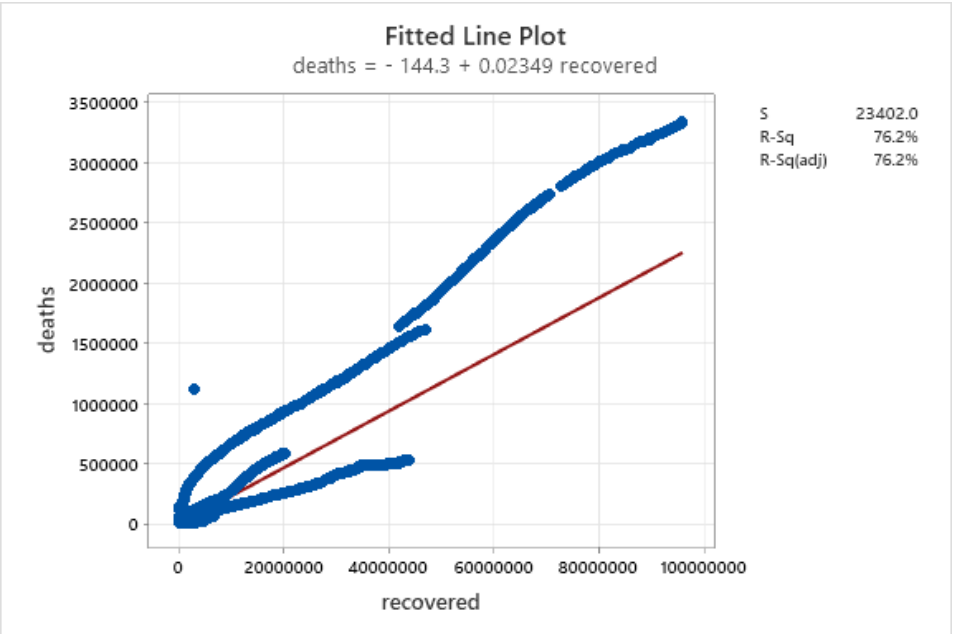
deaths = - 144.3 + 0.02349 recovered

It can be noted from Figure 12 that the R-square value for this fitted linear regression is 76.22%, indicating that there is a moderate but not strong linear relationship between the number of deaths and the number recovered from COVID-19 for the 596,101 data values due to big data number of values.

It can be noted from the Figure 13 Analysis of Variance (ANOVA) that the large F-test statistic value yields a P-value of 0.000. According to Google (2023b), a P-value of zero and those close to 0 indicate that the observed difference is unlikely to be due to chance.

Figure 14 above shows that the data points do not fit a linear relationship well, as indicated by the R-Squared value of 76.2%, and that the number recovered from COVID-19 far exceeds the number of deaths due to COVID-19.

Figure 14. Fitted line plot for regression of number of deaths and number of recovered from COVID-19



Data Visualization and Analytics of Big Data for Diabetes

Dataset Description

The utilized dataset is from the University of California at Irvine (UCI) Machine Learning Repository. It encompasses a comprehensive depiction of clinical care spanning a decade (1999-2008) across 130 U.S. hospitals and integrated delivery networks. Specifically focused on inpatients diagnosed with diabetes, the dataset comprises over 50 distinct characteristics elucidating patient-specific and hospital-related outcomes.

The data includes information on laboratory tests and medications administered during the patient's hospital stay, as well as detailed attributes such as patient number, race, gender, age, admission type, length of stay, admission physician specialty, number of tests performed, HbA1c test results, diagnosis, number of medications, diabetes medications, number of outpatient, inpatient, and emergency visits in the year prior to admission. The report includes detailed information on attributes as below.

This diabetes dataset encompasses ten years (1999-2008) of clinical care within a network of 130 hospitals and integrated healthcare delivery systems in the United States. It comprises a comprehensive set of over 50 patient demographics and hospital outcomes attributes.

Table 9. Diabetes big data description of attributes

Attributes	Data Type	Description
Encounte_id	Alphanumeric	Unique identifier of an encounter
Patient_id	Alphanumeric	Unique identifier of a patient
Race	Categorial	Values: Caucasian, Asian, African American, Hispanic, and other.
Gender	Categorial	Values: male, female, and unknown/invalid
Age	Categorial	Grouped in 10-year intervals: [0, 10), [10, 20) ..., [90, 100)
Weight	Categorial	Weight in pounds.
Admission_type_id	Categorial	Integer identifier corresponding to 9 distinct values, for example, emergency, urgent, elective, newborn, and not available
Discharge_diposition_id	Categorial	Integer identifier corresponds to 29 distinct values, for example, discharged to home, expired, and unavailable.
Admission_source_id	Categorial	Integer identifier corresponding to 21 distinct values, for example, physician referral, emergency room, and transfer from a hospital.
Time_in_hopsital	Numerical - Discrete	Integer number of days between admission and discharge.
Payer_code	Categorial	Integer identifier corresponds to 23 distinct values, such as Blue Cross/Blue Shield, Medicare, and self-pay.
Medical_Specialty	Categorial	Integer identifier of a specialty of the admitting physician, corresponding to 84 distinct values, for example, cardiology, internal medicine, family/general practice, and surgeon.
Number_outpatient	Numerical	Number of outpatient visits of the patient in the year preceding the encounter.
Number_emergency	Numerical	Number of emergency visits of the patient in the year preceding the encounter.
Number_inpatient	Numerical	Number of inpatient visits of the patient in the year preceding the encounter.

Table 10. Diabetes dataset information

Name of Disease	Number of Data Values	Years of Data	URL of Data Source
Diabetes	101,766	1990-2008	https://www.kaggle.com/datasets/brandao/diabetes/versions/1?resource=download

Source: Clore et al. (2014)

Table 11. Information of the value of insulin in Figure 15

Fig 15 Row 1 Left	Down	Dosage was increased during the encounter
Fig. 15 Row 1 Right	No	Dosage was decreased
Fig. 15 Row 2 Left	Steady	The dosage did not change
Fig. 15 Row 2 Right	Up	The drug was not prescribed

Data Visualizations for Diabetes

Figure 15 is a ribbon chart, capable of superimposing multiple data series and emphasizing the comparative variations within those series. The advantage of this technique is that multiple data series can be compared on a single chart to identify trends and relative differences quickly. It is also possible to visualize correlations and interactions by observing features such as overlap and intersection of series. This ribbon chart analyzes insulin administration data and displays two data series—Inpatient and Outpatient—differentiated by the degree of insulin administration. The graphs are also divided by age to reveal in which age group the diabetes epidemic is concentrated.

The analysis results indicate that the highest number of patients is in the late 50s to 80s age group. In contrast, a relatively small number of patients was observed in the earlier age groups. Furthermore, when focusing on the degree of insulin administration, the "No" and "Steady" data were overwhelmingly more common than in the other two graphs. Hence, although the insulin administration demonstrates overall efficacy, the sample size of younger patients is limited, and the data indicates minimal fluctuations (particularly within the age range of 20 to 40 years). Considering these viewpoints, it is evident, as shown in Figure 15, that the efficacy of insulin exhibits substantial variability based on the patient's age.

Figure 16 presents a comprehensive examination of the overall diabetic population, encompassing a comparative analysis of individuals receiving inpatient and outpatient care. The investigation also

Figure 15. An analysis of age and insulin usage in relation to the combined sum of inpatient and outpatient data

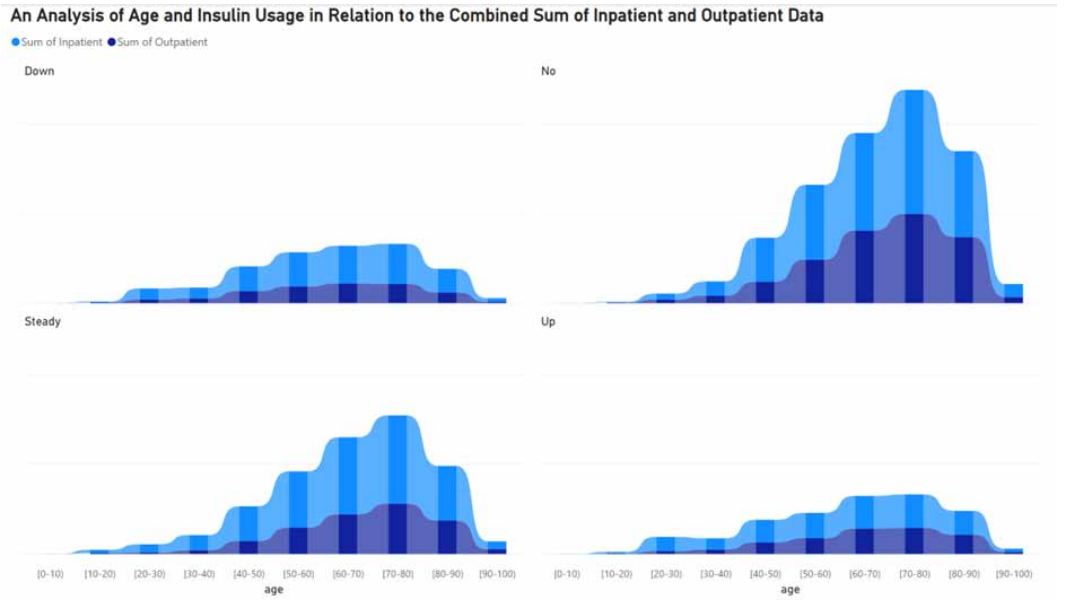
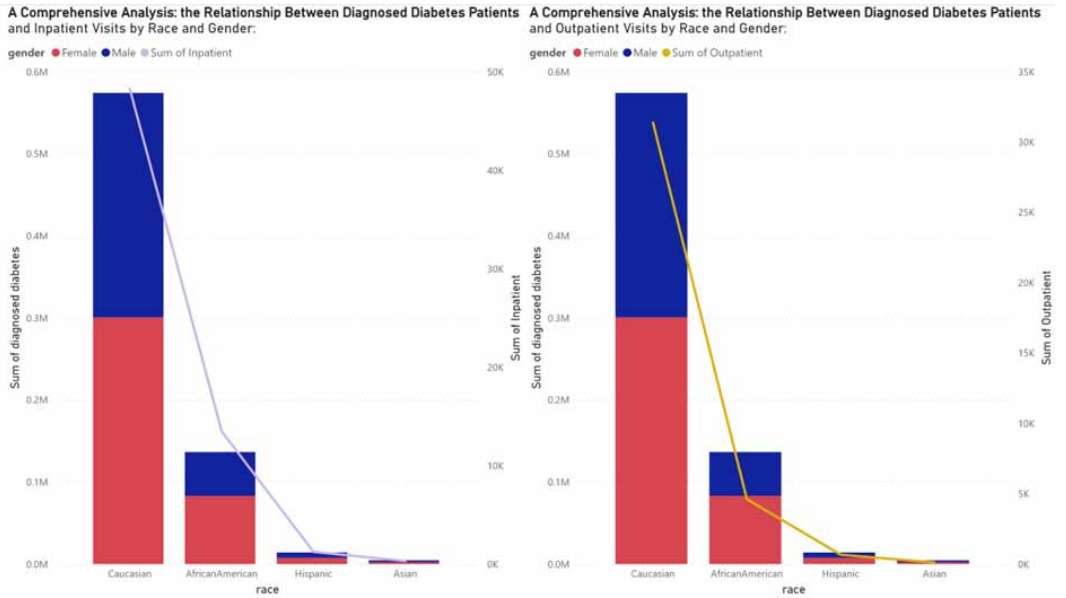


Figure 16. A comprehensive analysis: The relationship between diagnosed diabetes patients and inpatient/outpatient visits by race and gender



highlights distinct racial demographics—namely Caucasians, African Americans, Hispanics, and Asians—and reveals outcomes based on their attributes.

The analysis in Figure 16 reveals that while there were no significant differences in the overall number of patients, there were significant differences in the Inpatient and Outpatient figures for Caucasians and African Americans compared to other racial groups. In particular, the number of Inpatients with diabetes tended to be particularly large compared to the number of Outpatients.

The findings from Figure 16 indicate that there may be variations in the medical requirements and treatment alternatives for hospitalized individuals with diabetes based on their racial background. Moreover, the notable disparities in the number of Caucasian and African American hospitalized patients in comparison to other racial cohorts underscore concerns regarding healthcare accessibility and the existence of health disparities.

Figure 17 comprehensively examines the correlation between patient counts and glucose levels among diabetic inpatients and outpatients. It encompasses two axes: one representing the number of patients and the other depicting glucose values. The visual representation employs bar and line graphs to depict distinct glucose value ranges.

The first notable result of Figure 17 is that the glucose value "None" is significantly larger than the other three values. "None" means that no glucose was detected in the diabetic. This result indicates that maintaining a certain blood glucose level is very important to prevent diabetes.

In the remaining three categories of glucose levels, there was a certain level of consistency in the patient population. Specifically, it was demonstrated that a consistent number of patients were present across all scenarios. However, in the glucose range below 200, a marginally reduced patient count was observed compared to the other two ranges.

Furthermore, the analysis focused on gender, and interesting characteristics were observed in the distribution of diabetic patients by gender. It was found that in the case of patients in their 20s with glucose levels below 300, the number of female patients was overwhelmingly higher, accounting for more than 90% of the total number of patients.

Figure 17. Analyzing the combined total of inpatient and outpatient cases based on age, gender, and glucose levels

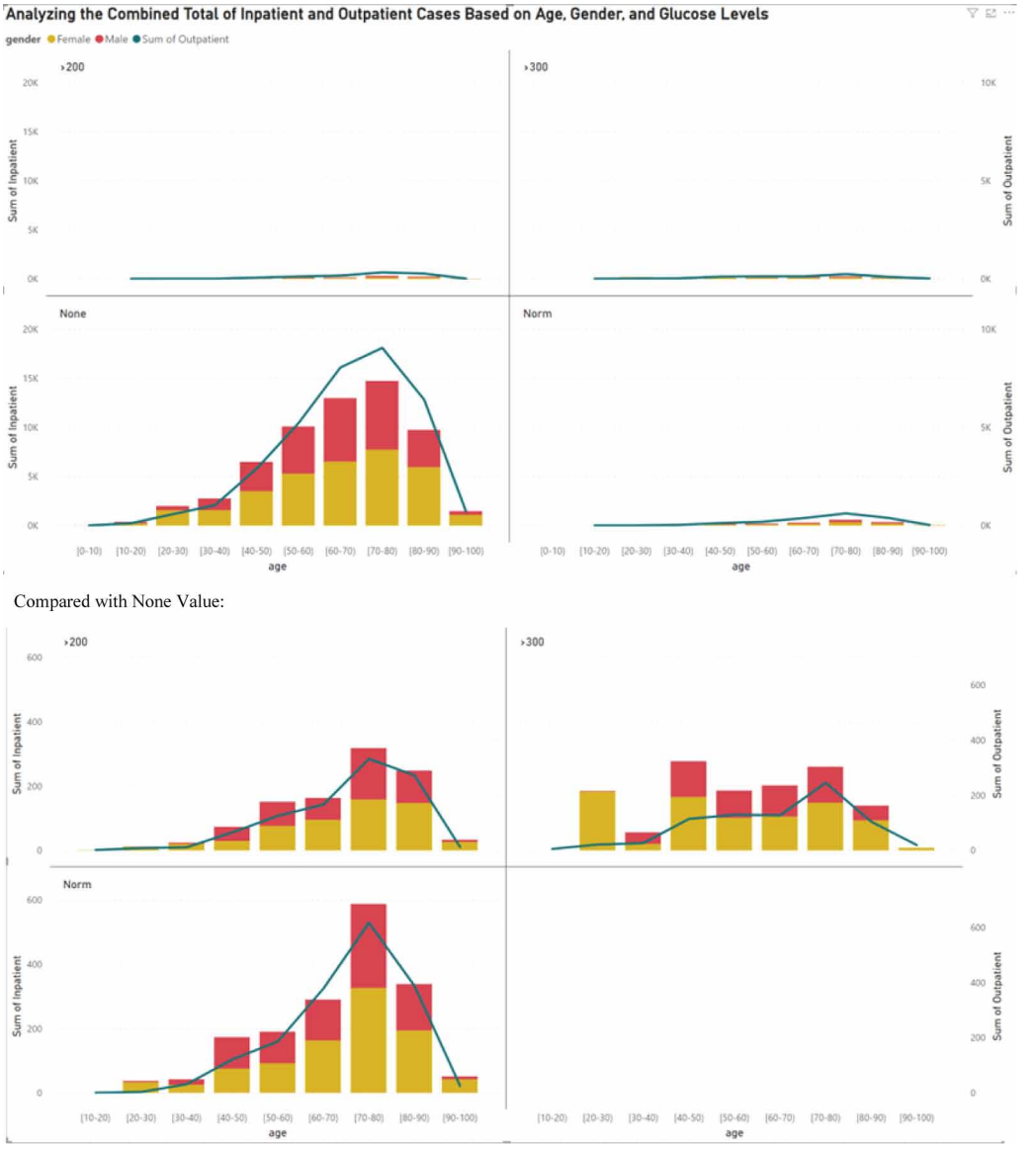


Table 12. Information of the value of glucose levels

Row 1: > 200:	Glucose level is less than 200.
Row 2: > 300	Glucose level is less than 300.
Row 3: None	Glucose is not measured.
Row 4: Norm	Glucose level is at a normal level.

Note: The final block is compared without “None” value.

The findings of this study highlight the significance of maintaining optimal blood glucose levels in the management of diabetes, as well as the association between patient distribution and gender. These results are anticipated to provide valuable insights for developing effective strategies in treating and preventing diabetes.

In this study, as shown in Figure 18, a detailed analysis was conducted to investigate the impact of insulin levels on the length of hospital stay among hospitalized diabetic patients. While emphasizing the significance of insulin efficacy in diabetes treatment, an examination was also conducted to assess potential variations in effectiveness based on gender.

Initially, an analysis was conducted from a proportional perspective. Consequently, noteworthy gender disparities in the efficacy of insulin among male and female patients were not observed. The graphical representations consistently displayed similar trends in the proportions of male and female patients, signifying the absence of gender-based variations in insulin effectiveness.

Next, the focus was on the degree of insulin administration. Patients were categorized in Figure 13 as either "No" (not dosed) or "Up," "Down," or "Steady" (different doses). The results showed that patients in the "No" group, who were not receiving insulin, clearly had a longer hospital stay than those who were receiving insulin. These results suggest that insulin is an essential treatment for improving diabetes.

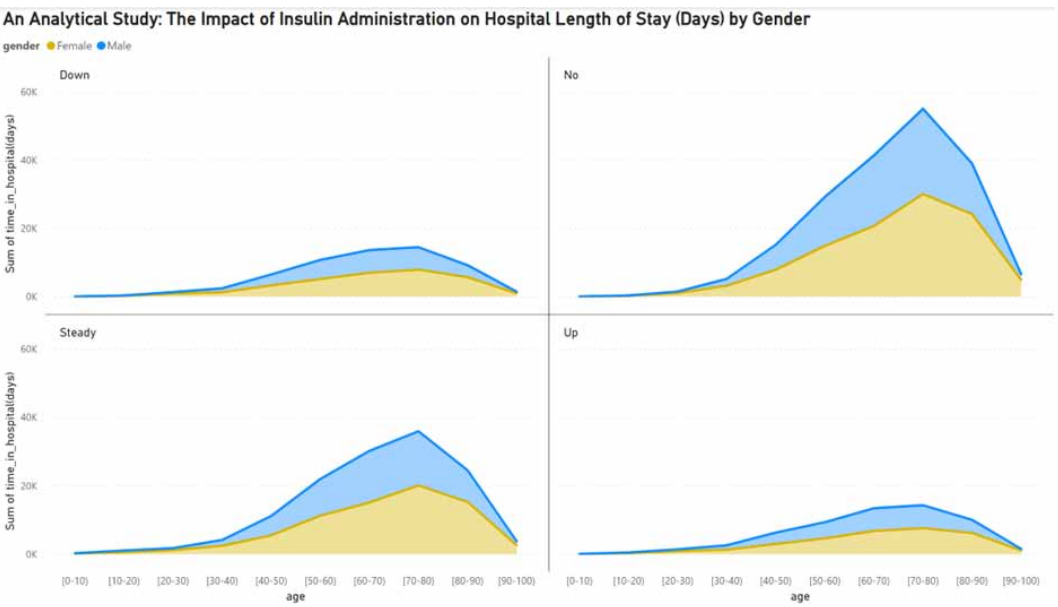
Furthermore, comparisons, as shown in Figure 18, were made within the "Up," "Down," and "Steady" groups. The results showed that patients in the "Steady" group tended to require comprehensively longer hospital stays than the other groups. In other words, the results suggest that stable administration of a constant insulin dose may have a lasting effect.

The findings of this study demonstrate the significance of insulin in enhancing diabetes management, with no discernible disparities in insulin efficacy between genders. Moreover, the results imply that the observed effects are anticipated to be enduring rather than transient.

Diabetes Predictive Analytics

A study of 101,766 data values with 47 attributes for diabetes was conducted by the University of California at Irvine (UCI) Machine Learning Repository (2014) for diabetes at 130 hospitals in the U.S.

Figure 18. An analytical study: The impact of insulin administration on hospital length of stay (days) by gender (Note: Female is indicated by orange and male is indicated by blue. The figure uses the same key as Figure 15 for Down, No, Steady, and Up.)



for years 1999-2008. Minitab (2023) was used to perform regression using CART (Classification and Regression Trees for Machine Learning). Regression illustrates important patterns and relationships between a continuous response and significant predictors within highly complicated data without using parametric methods.

Unlike linear regression techniques, CART analysis does not assume a particular form of relationship between the independent and dependent variables. Therefore, CART can often be used even in cases where data are not suitable for analysis by linear regression (Google, 2023c). A more detailed discussion of CART is provided by Brownlee (2020).

Figure 19 below shows the response information for performing dataset of 101,766 data values using a 70% training and 30% test using CART.

CART was performed for this Diabetes data set of 101,766 data values, and a tree was constructed using Minitab. However, Minitab stopped growing the tree at 1024 notes due to default limits. The CART tree is shown in Figure 20, placed in the Appendix in vertical format to show how complicated and complex CART can be for big data.

CONCLUSION

In this paper, we first provided a brief background of big data and its presence in various big data studies of various health conditions, including generalized healthcare data streaming. Applications of big data to healthcare are presented using representative data for stroke (cardiovascular disease), COVID-19, and diabetes. COVID-19 is a contagious disease, while strokes and diabetes are not. Heart disease and diabetes can be hereditary diseases.

Research Contribution

For stroke, data visualization for big data of 5,110 data values for 2021-2022:

- Highlighted the increased stroke risk with advancing age and suggested that consistent smoking may not be a major contributing factor.
- Showed that hypertension is a primary underlying condition that predisposes one to stroke.
- Indicated that occupational stress plays a significant role in the etiology of stroke.
- Revealed that blood glucose could serve as a more significant parameter than Body Mass Index (BMI) when evaluating the well-being of individuals who have experienced a stroke.

For stroke, the predictive analytics yielded either a very good or strong fit to the normal distribution except for the variable of average glucose level that had a non-linear relationship.

For COVID-19, data visualizations for big data for COVID-19 for 2020 to 2023 showed that:

- The total cases and mortality rates occurred from most to least for the continents of Africa, Asia, Europe, Oceania and South America from April 2020 to May 2023.

Figure 19. Response Information for 101,766 data set for diabetes

Response Information

Data Set	N	% of N	Mean	StDev	Minimum	Q1	Median	Q3	Maximum
Training	71241	70.0	4.39773	2.98003	1	2	4	6	14
Test	30525	30.0	4.39191	2.99697	1	2	4	6	14

- COVID-19 full vaccination coverage as a percentage of population proportions across continents occurred in the order of Africa, Asia, Europe, North America, Oceania, and South America from April 2020 to May 2023.
- The number of vaccinations exceeded the number of deaths and cases for February 2020 to July 2023.
- The number of recovered cases exceeded the number of confirmed cases until mid-September 2021.

For COVID-19, the predictive analytics for Example #2 with 596,101 data values indicated a moderate but not strong linear relationship between the number of deaths due to COVID-19 and the number recovered with an upper bound of 100 million, as shown in Figure 14.

For diabetes, data visualizations for big data for diabetes using UCI Data for 1999-2008 showed that:

- The sum of inpatient insulin use age surpassed that of outpatients.
- The age group 70-80 had the most prominent steady use of insulin.
- Caucasian/White had the most considerable incidence of diagnosed diabetes as both inpatients and outpatients, with equal distributions of males and females.
- The combined totals of inpatient and outpatient cases based on age, gender, and glucose levels had the maximum for the age group of 70-80, regardless of gender.
- The impact of insulin administration on hospital length of stay showed that steady insulin administration occurred with maximum hospital length of stay for the 70-80 age group.
- For diabetes, the predictive analytics for the data set from the UCI Machine Learning Repository of 101,766 data values with 47 attributes that were modeled using the CART algorithm with a 70%/30% ratio of Training data/Test data yielded a sizeable branching tree that was too large to print but presents the many possible branching options for analyzing this big data set.

The research has illustrated that the application of big data analytics to big data of different levels of aggregation for these selected diseases is resilient and not sensitive to increasing levels of big data.

Future Work and Research Limitations

The future directions of the research include investigating the use of big data analytics to other diseases selected from the four categories of (a) infectious diseases, (b) deficiency diseases, (c) hereditary diseases, and (d) physiological diseases. Future studies would include comparisons within and between these disease categories depending on data availability.

COMPETING INTERESTS

The authors declare there are no competing interests.

ACKNOWLEDGMENT AND FUNDING

Dr. Richard Segall acknowledges support of this research through the Seed Money Grant Award by Arkansas Biosciences Institute (ABI) and the support of facilities of the Neil Griffin College of Business at Arkansas State University in Jonesboro to conduct this research. Undergraduate student Soichiro Takahashi acknowledges receipt of funding provided by the Arkansas Biosciences Institute (ABI) for the Undergraduate Summer Internship Program to support this research.

REFERENCES

Alqaissi, E. Y., Alotaibi, F. S., & Ramzan, M. S. (2022). Modern machine-learning predictive models for diagnosing infectious diseases. *Computational and Mathematical Methods in Medicine*, *6902321*, 1–13. Advance online publication. doi:10.1155/2022/6902321 PMID:35693267

Bahga, A., & Madisetti, V. (2016). *Big data science & analytics: A hands-on approach*. Big-Data-Analytics Book Company.

Bansal, S., Chowell, G., Simonsen, L., Vespignani, A., & Viboud, C. (2016). Big data for infectious disease surveillance and modeling. *The Journal of Infectious Diseases*, *214*(4, suppl 4), S375–S379. doi:10.1093/infdis/jiw400 PMID:28830113

Barik, R. K., Priyadarshini, R., Dubey, H., Kumar, V., & Mankodiya, K. (2018). FogLearn: Leveraging fog-based machine learning for smart system big data analytics. *International Journal of Fog Computing*, *1*(1), 15–34. doi:10.4018/IJFC.2018010102

Borges do Nascimento, I., Marcolino, M., Abdulazeem, H., Weerasekara, I., Azzopardi-Muscat, N., Gonçalves, M., & Novillo-Ortiz, D. (2021). Impact of big data analytics on people's health: Overview of systematic reviews and recommendations for future studies. *Journal of Medical Internet Research*, *23*(4), e27275. doi:10.2196/27275 PMID:33847586

Brownlee, J. (2020). *Classification and regression trees for machine learning*. https://machinelearningmastery.com/classification-and-regression-trees-for-machine-learning/

Caesar, M. (2023). *Our World in Data – COVID 19.* https://www.kaggle.com/datasets/caesarmario/our-world-in-data-covid19-dataset

Catalyze. (2022). *Discovering the role of big data in driving early infectious disease detection*. https://www.catalyze-group.com/discovering-the-role-of-big-data-in-driving-early-infectious-disease-detection/

Chen, X., & Liu, B. (2023). Research progress of tumor big data visualization. *Electronics (Basel)*, *12*(3), 743. doi:10.3390/electronics12030743

Chen, Y., Argentinis, E., & Weber, G. (2016). IBM Watson: How cognitive computing can be applied to big data challenges in life sciences research. *Clinical Therapeutics*, *38*(4), 688–701. doi:10.1016/j.clinthera.2015.12.001 PMID:27130797

Clegg, B. (2017). *Big Data: How the information revolution is transforming our lives*. Icon Books.

Clore, J., Cios, K., DeShazo, J., & Strack, B. (2014). *Diabetes 130-US hospitals for years 1999-2008*. UCI Machine Learning Repository. https://archive.ics.uci.edu/dataset/296/diabetes+130-us+hospitals+for+years+1999-2008

Corsi, A., de Souza, F. F., Pagani, R. N., & Kovaleski, J. L. (2020). Big data analytics as a tool for fighting pandemics: A systematic review of literature. *Journal of Ambient Intelligence and Humanized Computing*, *12*(10), 9163–9180. doi:10.1007/s12652-020-02617-4 PMID:33144892

Dar, G. M., Sharma, A., & Singh, P. (2021). Deep learning models for detection and diagnosis of Alzheimer's Disease. In M. Roy & L. Gupta (Eds.), *Machine learning and data analytics for predicting, managing, and monitoring disease* (pp. 140–149). IGI Global. doi:10.4018/978-1-7998-7188-0.ch011

Dwivedi, M. K., Pandey, S. K., & Singh, P. K. (2021). Public health surveillance system: Infectious diseases. In D. Yadav, A. Bansal, M. Bhatia, M. Hooda, & J. Morato (Eds.), *Diagnostic applications of health intelligence and surveillance systems* (pp. 201–220). IGI Global. doi:10.4018/978-1-7998-6527-8.ch010

Fedesoriano. (2021, January 26). *Stroke prediction dataset*. Kaggle. https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

Foresee Medical. (2022). *Big data healthcare analytics*. https://www.foreseemed.com/blog/big-data-analytics-in-healthcare

Franks, B. (2012). Taming the big data tidal wave: Finding opportunities in huge data streams with advanced analytics (Wiley & SAS Business Series). John Wiley & Sons.

Gangal, A., Kumar, P., Kumari, S., & Saini, A. (2021). Prediction models for healthcare using machine learning: A review. In G. Rani & P. Tiwari (Eds.), *Handbook of research on disease prediction through data analytics and machine learning* (pp. 70–91). IGI Global. doi:10.4018/978-1-7998-2742-9.ch005

Garattini, C., Raffle, J., Aisyah, D. N., Sartain, F., & Kozlakidis, Z. (2019). Big data analytics, infectious diseases and associated ethical impacts. *Philosophy & Technology*, *32*(1), 69–85. doi:10.1007/s13347-017-0278-y PMID:31024785

Google. (2023a). *Stroke*. https://www.google.com/search?client=firefox-b-1-d&q=stroke

Google. (2023b). *P-value.* https://www.Google.com

Google. (2023c) *CART Regression*. https://www.google.com/search?client=firefox-b-1-d&q=CART+regressio+minitab

Hoque, M. R., & Bao, Y. (2016). Application of big data in healthcare: Opportunities, challenges and techniques. In Information Resources Management Association (Ed.), Big data: Concepts, methodologies, tools, and applications (pp. 1189–1208). IGI Global. doi:10.4018/978-1-4666-9840-6.ch053

Jayashree, K., & Swaminathan, B. (2021). Big data in cloud computing. In S. Goundar & P. Rayani (Eds.), *Applications of big data in large- and small-scale systems* (pp. 77–84). IGI Global. doi:10.4018/978-1-7998-6673-2.ch005

Kadam, K., Kamat, P. V., & Malav, A. P. (2019). Cardiovascular disease prediction using data mining techniques: A proposed framework using big data approach. In H. Bouarara, R. Hamou, & A. Rahmani (Eds.), *Advanced metaheuristic methods in big data retrieval and analytics* (pp. 159–179). IGI Global. doi:10.4018/978-1-5225-7338-8.ch007

Kasson, P. M. (2020). Infectious disease research in the era of big data. *Annual Review of Biomedical Data Science*, *3*(1), 43–59. doi:10.1146/annurev-biodatasci-121219-025722

Kemp, S., & Mabee, D. (2023, May 5). *Bing Covid-19 tracker*. Microsoft Learn Azure Open Datasets. https://learn.microsoft.com/en-us/azure/open-datasets/dataset-bing-covid-19?tabs=azure-storage

Khan, S., Khan, H. U., & Nazir, S. (2022). Systematic analysis of healthcare big data analytics for efficient care and disease diagnosing. *Scientific Reports*, *12*(1), 22377. doi:10.1038/s41598-022-26090-5 PMID:36572709

Kudari, J. M. (2021). Data analytics to predict, detect, and monitor chronic autoimmune diseases using machine learning algorithms: Preventing diseases with the power of machine learning. In M. Roy & L. Gupta (Eds.), *Machine learning and data analytics for predicting, managing, and monitoring disease* (pp. 150–182). IGI Global. doi:10.4018/978-1-7998-7188-0.ch012

Lin, R. H., Ye, Z., Wang, H., & Wu, B. (2018). Chronic diseases and health monitoring big data: A survey. *IEEE Reviews in Biomedical Engineering*, *11*, 275–288. doi:10.1109/RBME.2018.2829704 PMID:29993699

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & McKinsey Global Institute. (2011). *Big data: The next frontier for innovation, competition, and productivity*. McKinsey Global Institute.

Mayo Clinic. (2023). *Diabetes*. https://www.mayoclinic.org/diseases-conditions/diabetes/symptoms-causes/syc-20371444?utm_source=Google&utm_medium=abstract&utm_content=Diabetes-mellitus&utm_campaign=Knowledge-panel

Minitab, L. L. C. (2023). *Minitab statistical software*. https://www.minitab.com/en-us/

Nagavci, D., Hamiti, M., & Selimi, D. (2018). Review of prediction of disease trends using big data analytics. *International Journal of Advanced Computer Science and Applications*, *9*(8), 46–50. doi:10.14569/IJACSA.2018.090807

National Academy of Medicine. (2016 May 10). *Workshop on big data and analytics for infectious disease research, operations, and policy*. https://nam.edu/event/workshop-on-big-data-and-analytics-for-infectious-disease-research-operations-and-policy/

NHLBI. (2023). *What is a stroke?* National Health, Lung, and Blood Institute. National Institutes of Health. https://www.nhlbi.nih.gov/health/stroke

Owid. (n.d.). *Covid-19-data/public/data at master owid/covid-19-DATA*. GitHub. https://github.com/owid/covid-19-data/tree/master/public/data

Panigrahi, P. P., & Singh, T. R. (2017). Data mining, big data, data analytics: Big data analytics in bioinformatics. In S. Ram (Ed.), *Library and information services for bioinformatics education and research* (pp. 91–111). IGI Global. doi:10.4018/978-1-5225-1871-6.ch005

Pathak, P., Iyengar, S. P., & Abhyankar, M. (2021). A survey on tools for data analytics and data science. In B. Patil & M. Vohra (Eds.), *Handbook of research on engineering, business, and healthcare applications of data science and analytics* (pp. 28–49). IGI Global. doi:10.4018/978-1-7998-3053-5.ch003

Patil, B., & Vohra, M. (Eds.). (2021). *Handbook of research on engineering, business, and healthcare applications of data science and analytics*. IGI Global. doi:10.4018/978-1-7998-3053-5

Perry, J. S. (2017). What is big data? *IBM Developer Blog*. https://developer.ibm.com/blogs/what-is-big-data-more-than-volume-velocity-and-variety/

Pramanik, P. K., Pal, S., & Mukhopadhyay, M. (2022). Healthcare big data: a comprehensive overview. In Information Resources Management Association (Ed.), Research anthology on big data analytics, architectures, and applications (pp. 119–147). IGI Global. doi:10.4018/978-1-6684-3662-2.ch006

Punia, S. K., Kumar, M., Stephan, T., Deverajan, G. G., & Patan, R. (2021). Performance analysis of machine learning algorithms for big data classification: ML and AI-based algorithms for big data analysis. *International Journal of E-Health and Medical Communications*, *12*(4), 60–75. doi:10.4018/IJEHMC.20210701.oa4

Raghupathi, V., Zhou, Y., & Raghupathi, W. (2022). Exploring big data analytic approaches to cancer blog text analysis. In Information Resources Management Association (Ed.), Research anthology on big data analytics, architectures, and applications (pp. 1843–1863). IGI Global. doi:10.4018/978-1-6684-3662-2.ch090

Ranjan, J., & P., M. J. (2021). Big data analytics in the healthcare industry. In P. Smith & T. Cockburn (Eds.), *Global business leadership development for the fourth industrial revolution* (pp. 134–154). IGI Global. 10.4018/978-1-7998-4861-5.ch006

Rastogi, R., Chaturvedi, D. K., & Singhal, P. (2021). Surveillance of Type I and II diabetic subjects on physical characteristics: IoT and big data perspective in Healthcare@NCR, India. In S. Velayutham (Ed.), *Challenges and opportunities for the convergence of IoT, big data, and cloud computing* (pp. 277–313). IGI Global. doi:10.4018/978-1-7998-3111-2.ch016

Roy, M., & Gupta, L. R. (Eds.). (2021). *Machine learning and data analytics for predicting, managing, and monitoring disease*. IGI Global. doi:10.4018/978-1-7998-7188-0

Sambyal, N., Saini, P., & Syal, R. (2019). Big data analytics: Applications, trends, tools, and future research directions. In B. Gupta & D. Agrawal (Eds.), *Handbook of research on cloud computing and big data applications in IoT* (pp. 67–81). IGI Global. doi:10.4018/978-1-5225-8407-0.ch004

Sánchez-Acevedo, M. A., Acosta-Chí, Z. A., Sabino-Moxo, B. A., Márquez-Domínguez, J. A., & Canton-Croda, R. M. (2019). Big data analysis for cardiovascular diseases: Detection, prevention, and management. In Information Resources Management Association (Ed.), Coronary and cardiothoracic critical care: Breakthroughs in research and practice (pp. 60–77). IGI Global. doi:10.4018/978-1-5225-8185-7.ch004

Schroeck, M., Shockley, R., Smart, J., Romero Morales, D., & Tufano, P. (2012). *Analytics: The real-world use of big data.* IBM Global Business Services. https://www.bdvc.nl/images/Rapporten/GBE03519USEN.PDF

Schroer, A. (2023). *20 Big data in healthcare examples and applications, built in*. https://builtin.com/big-data/big-data-in-healthcare

Segall, R. S. (2020a). What is open source software (OSS) and what is big data? In R. Segall & G. Niu (Eds.), *Open source software for statistical analysis of big data: Emerging research and opportunities* (pp. 1–49). IGI Global. doi:10.4018/978-1-7998-2768-9.ch001

Segall, R. S. (2020b). Open source software (OSS) for big data. In R. Segall & G. Niu (Eds.), *Open source software for statistical analysis of big data: Emerging research and opportunities* (pp. 50–72). IGI Global. doi:10.4018/978-1-7998-2768-9.ch002

Segall, R. S., & Cook, J. S. (Eds.). (2018). *Handbook of research on big data storage and visualization techniques* (Vols. 1–2). IGI Global. doi:10.4018/978-1-5225-3142-5

Segall, R. S., & Niu, G. (Eds.). (2020). *Open source software for statistical analysis of big data: Emerging research and opportunities*. IGI Global. doi:10.4018/978-1-7998-2768-9

Sharma, A., & Rani, R. (2021). Machine learning perspective in cancer research. In G. Rani & P. Tiwari (Eds.), *Handbook of research on disease prediction through data analytics and machine learning* (pp. 142–163). IGI Global. doi:10.4018/978-1-7998-2742-9.ch008

Singh, S., & Shrivastava, V. (2021). The analysis of machine learning techniques for heart disease prediction. In S. Purohit, D. Singh Jat, R. Poonia, S. Kumar, & S. Hiranwal (Eds.), *Proceedings of the International Conference on Communication and Computational Technologies* (pp. 137–143). Springer. doi:10.1007/978-981-15-5077-5_12

Spurlock, C. (2018). Using data analytics to predict, detect and monitor chronic autoimmune diseases. *Towards Data Science*. https://towardsdatascience.com/using-data-analytics-to-predict-detect-and-monitor-chronic-autoimmune-diseases-ca062e9cb12c

Sweeney, S. M., Hamadeh, H. K., Abrams, N., Adam, S. J., Brenner, S., Connors, D. E., Davis, G. J., Fiore, L., Gawel, S. H., Grossman, R. L., Hanlon, S. E., Hsu, K., Kelloff, G. J., Kirsch, I. R., Louv, B., McGraw, D., Meng, F., Milgram, D., Miller, R. S., & Srivastava, S. et al. (2023). Challenges to using big data in cancer. *Cancer Research*, *83*(8), 1175–1182. doi:10.1158/0008-5472.CAN-22-1274 PMID:36625843

Tenali, N., & Babu, G. R. (2023). A systematic literature review and future perspectives for handling big data analytics in COVID-19 diagnosis. *New Generation Computing*, *41*(2), 243–280. doi:10.1007/s00354-023-00211-8 PMID:37229177

The White House. (2012, March 29). *Obama administration unveils "big data" initiative: Announces $200 million in new R&D investments.* https://obamawhitehouse.archives.gov/the-press-office/2015/11/19/release-obama-administration-unveils-big-data-initiative-announces-200

University of California Irvine (UCI) Machine Learning Repository. (2014). *Diabetes 130-US hospitals for years 1999-2008*. https://archive.ics.uci.edu/dataset/296/diabetes+130-us+hospitals+for+years+1999-2008

Varatharajan, A. C. S., & Varatharajan, V. (2020). Big data analytics in the healthcare industry: An analysis of healthcare applications in machine learning with big data analytics. In A. Haldorai & A. Ramu (Eds.), *Big data analytics for sustainable computing* (pp. 160–178). IGI Global. doi:10.4018/978-1-5225-9750-6.ch010

Venkatesh, R., Balasubramanian, C., & Kaliappan, M. (2019). Development of big data predictive analytics model for disease prediction using machine learning technique. *Journal of Medical Systems*, *43*(8), 272. doi:10.1007/s10916-019-1398-y PMID:31278468

Wong, Z. S. Y., Zhou, J., & Zhang, Q. (2019). Artificial intelligence for infectious disease big data analytics. *Infection, Disease & Health*, *24*(1), 44–48. doi:10.1016/j.idh.2018.10.002 PMID:30541697

World Health Organization. (2023). *Coronavirus Disease (COVID-19).* https://www.who.int/emergencies/diseases/novel-coronavirus-2019/question-and-answers-hub/q-a-detail/coronavirus-disease-covid-19

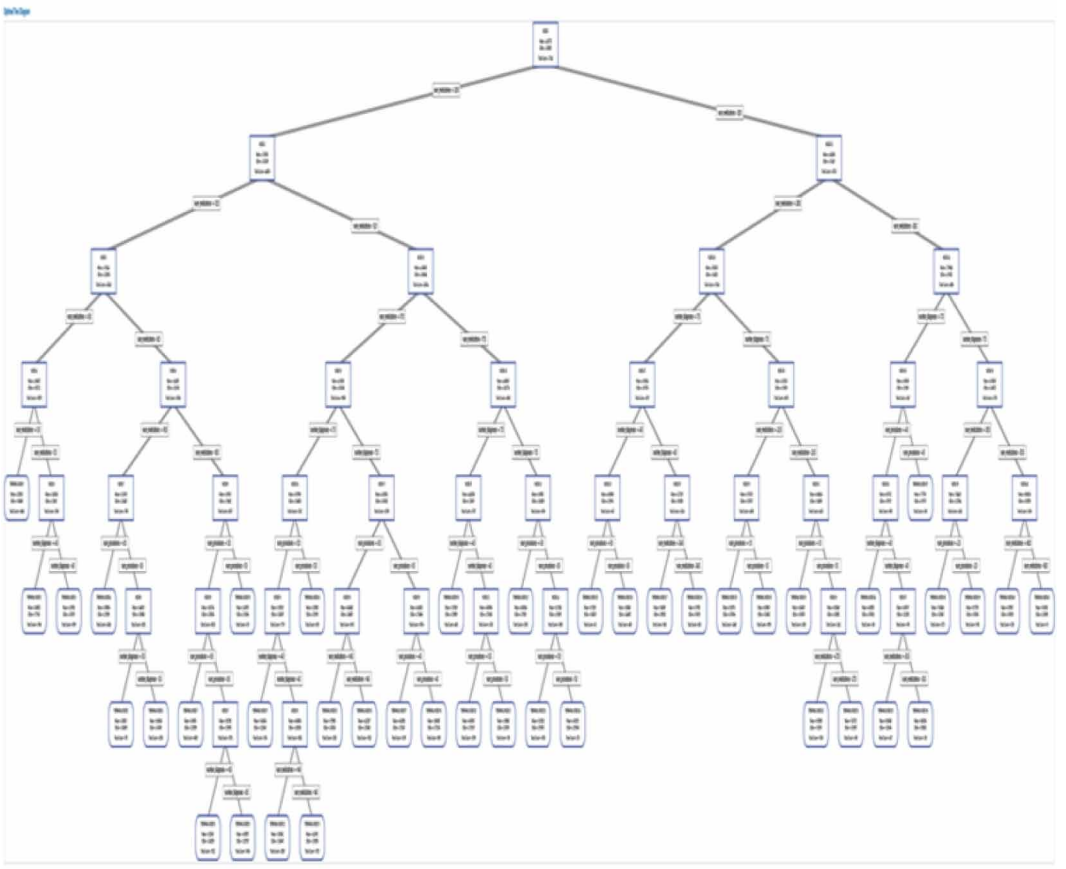
Yadav, D. C., & Pal, S. (2021). Analysis of heart disease using parallel and sequential ensemble methods with feature selection techniques: heart disease prediction. *International Journal of Big Data and Analytics in Healthcare, 6*(1), 40–56. 10.4018/IJBDAH.20210101.oa4

Yee, S. W., Gutierrez, C., Park, C. N., Lee, D., & Lee, S. (2020). Big data: Its implications on healthcare and future steps. In R. McHaney, I. Reychev, J. Azuri, M. McHaney, & R. Moshonov (Eds.), *Impacts of Information Technology on Patient Care and Empowerment* (pp. 82–99). IGI Global. doi:10.4018/978-1-7998-0047-7.ch005

Zhang, Z. (2020). Predictive analytics in the era of big data: Opportunities and challenges. *Annals of Translational Medicine*, *8*(4), 68. doi:10.21037/atm.2019.10.97 PMID:32175361

APPENDIX

Figure 20. Regression tree created by CART using Minitab for diabetes data from University of California Irvine (UCI) Machine Learning Repository with 101,766 data values (* WARNING * Minitab stopped growing the tree at 1024 nodes due to default limits)



Richard S. Segall is Professor of Information Systems & Business Analytics in Neil Griffin College of Business at Arkansas State University in Jonesboro. He holds BS/MS in mathematics, a MS in operations research and statistics from Rensselaer Polytechnic Institute in Troy, New York, and a PhD in operations research form University of Massachusetts at Amherst. He has served on the faculty of Texas Tech University, University of Louisville, University of New Hampshire, University of Massachusetts-Lowell, and West Virginia University. His research interests include data mining, big data, text mining, web mining, database management, and mathematical modeling. His funded research includes that by U.S. Air Force, NASA, Arkansas Biosciences Institute (ABI), and Arkansas Science & Technology Authority (ASTA). He was a member of former Arkansas Center for Plant-Powered-Production (P3) and is a member of Center for No-Boundary Thinking (CNBT), serves on the editorial boards of the International Journal of Data Mining, Modelling and Management (IJDMMM), International Journal of Data Science (IJDS), and International Journal of Fog Computing (IJFC), and is co-editor of five books: (1.) Biomedical and Business Applications Using Artificial Neural Networks and Machine Learning, (2.) Open Source Software for Statistical Analysis of Big Data, (3.) Handbook of Big Data Storage and Visualization Techniques, (4.) Research and Applications in Global Supercomputing, and (5.) Visual Analytics of Interactive Technologies: Applications to Data, Text & Web Mining.

Soichiro Takahashi is an undergraduate student enrolled at Arkansas State University in Jonesboro, currently pursuing a bachelor's degree in Computer & Information Technology. Originally from Aichi, Japan, he also currently works as a Student Research Technician at the U.S. Department of Agriculture located in Arkansas Biosciences Institute (ABI). His research interests predominantly pertain to data science, encompassing data mining, data visualization, data automation, and data analysis.