Multidisciplinary Journal of Healthcare (MJH)
ISSN Online: 3078-3011 ISSN Print: 3078-3003

Volume No: 02 Issue No: 01 (2025)

Data Analytics Approach to Population Health Segmentation

Adaeze Ezeogu
Big Data Analytics, Sheffield Hallam University, Sheffield, United Kingdom.
ORCID Number: https://orcid.org/ 0009-0002-7075-4345
Email: Adaezeojinika@gmail.com
Corresponding Author: Adaeze Ojinika Ezeogu

ABSTRACT

This study investigated the segmentation of the population's health using data analytics. It looked at how population health segments were identified, how care was measured for those segments, and how the proper people were given care. It covered how data analytics was applied to care management, socioeconomic determinants of health, and predicting where health resources should be allocated. Patients with the highest risk of disease were identified, clinical treatment decisions were supported, patient identification accuracy and speed were improved, expenses were decreased, preventative measures were promoted, and medical research was supported through data analytics. This research used Sander's notion of onions to understand the technique of understanding the appropriate methodological choice, philosophy, theory creation, and data collection procedures to understand the influence of data-driven approaches to population health segmentation. The survival rate of Covid-19 patients was predicted using Decision Tree, Random Forest, and Logistic Regression models. Qualitative data analysis was also carried out to comprehend the factors influencing the application of data-driven population health segmentation approaches, the role of ineffective healthcare and service delivery, benefits, barriers, and strategies to develop and improve data-driven population health segmentation. The findings revealed a knowledge gap in data analytics, with just half of the respondents knowing and comprehending the frameworks of population health segmentation. The study employed secondary dataset analysis to examine the treatments and methods applied to Covid-19. The investigation identified a knowledge gap in population health segmentation using data-driven methodologies using primary data. To close this gap, the project suggested thoroughly examining medical issues and comprehending their impact on Covid-19 patients based on age. This technique made it possible to comprehend specific medical disorders and how much risk each one posed based on the patient's age.

Keywords: population health segmentation, data analytics, healthcare management, predictive modeling, COVID-19, survival prediction.

INTRODUCTION

Data analytics has been used to manage population health over the years to improve health outcomes, improve care, and address social determinants of health by collecting, analysing, modelling, and storing demographic data. As healthcare's population health management market continues to grow, systems must gather data from multiple sources, apply analytics to the data, and manage population care. Healthcare costs have risen while demand has remained high. Population health

segmentation aims to use scarce health resources better to meet the needs of the vulnerable population (Martino et al. 2022). Data analytics are used in health management to identify populations needing care, measure the care provided to those populations, and deliver care to the right people. The population health management process begins with collecting key demographic and clinical data about patients, typically obtained from electronic health records. Data analytics can thus be used to predict where health resources should be directed, thereby improving care management and addressing social determinants of health.

Kindig and Stoddart (2003) define population health as the health outcomes of individuals and the distribution of those outcomes within the group. It includes health outcomes, health determinant patterns, and policies and interventions that connect the two. Population health realigns a specific group's health determinants and predicts the expected outcomes from these determinants using various data on particular individuals. These groups, defined by health determinants and consequences, can be used to identify a population group. Population health has pillars that aid in categorizing a population's health determinants and outcomes. According to Caldararo and Nash (2017), these pillars are divided into four categories: chronic care management, quality and safety, public health, and health policy. According to the scholar, incorporating these concepts into education and practice and their interactions lays the groundwork for achieving population health goals and strategies. This public health components aid in further segmenting populations based on health determinants and outcomes.

Based on the arguments above, population health has categories unaware of health determinants and outcomes. As a result, population health segmentation divides patients into groups based on specific needs, characteristics, or behaviours that lead to strategic delivery and policy formulation processes tailored to these groups (Chong et al. 2019). Healthcare is complex and has numerous working points, making it difficult for organizations and governments to meet the sector's needs if not segmented. Health is a universal good, and care and services should be tailored to the general public.

Population health segmentation is based on integrating existing healthcare services to fill gaps in a community's social health. According to Yan et al. (2018), there is a growing concern in population health and integrated health systems about maximizing the impacts of healthcare services across the care continuum around the needs of the patients. Providers and physicians must use data to serve a group of people best. To assist large communities of people, big data is frequently used to address population health concerns. Such forces have accelerated the formation of population health segments that meet these standards.

Data analytics uses data insights from data analysis tools to improve visual representation and population appeal. Health systems are distinguished by large and complex datasets that have proven challenging to store or analyse efficiently and economically using traditional data processing systems (Nascimento et al., 2021). Such massive datasets have also resulted in time-consuming management activities, which may have distorted health outcomes in some parts of society (Smeets et al., 2020). With unprecedented challenges such as the covid-19 pandemic, data handling in most healthcare facilities has proven even more challenging to manage within the standard time frame. As a result, the traditional data analysis methods have resulted in slow action and implementation of the recommendations based on their findings.

Furthermore, population health segmentation has been hampered and delayed further by rapid population growth, bringing new and more complex health-related issues. The covid-19 pandemic, one of the world's worst health issues since time

immemorial, has multiplied existing health crises in healthcare facilities to almost unimaginable proportions. Since then, most countries have been left with massive health disparities and despair, which have harmed their socioeconomic fabric.

Evans et al. (2019) argue that with the global encroachment of technology, there has been rapid growth and diffusion of digital media technologies that have changed the landscape of market segmentation in the last two decades. As a result, digital technologies are increasingly used to promote prosocial behaviour change (Evans et al., 2019). On the other hand, Nascimento et al. (2021) discuss how effective data analytics tools have been in dealing with the complex datasets generated by healthcare service delivery. With the rapid increase in the documentation of healthcare services, challenges in dissecting such massive datasets have arisen, necessitating the use of traditional tools that can improve faster understanding of the datasets' insights.

In their systematic review of the literature on population health segmentation, Galetsi et al. (2019) discovered that one of the central values created by the use of data analytics tools is the development of analytical techniques that provide personalized health services to users and supports human decision-making using automated algorithms, challenging power issues in the doctor-patient relationship and creating new working conditions. Population health segmentation is viewed from this perspective as a further intervention in the healthcare sector aimed at ensuring that patients' health issues are handled amicably and that services are rendered pragmatically.

Four components of data analytics are required for population health segmentation. These four components are referred to as descriptive, diagnostic, predictive, and prescriptive analysis by Mehta et al. (2020). Scholars argue that healthcare data is highly diverse due to the diversity of diseases and treatment pathways. As a result, data from the healthcare setting reflect this heterogeneous nature by having different structures. Data ranges in structure from highly structured (demographic data) to semi-structured or weakly structured (diagnostic data) to unstructured (doctor notes) Mehta et al. (2020).

Providers are replacing the "one size fits all" care mentality with value-based care by implementing population health segmentation strategies and data analytics. This improves the patient experience, the health of patient populations, and the cost of care by standardizing the healthcare process. Using data analytics, providers can assess which processes are the most effective methods for wellness and prevention within value-based care models. Organizations can consider physical and social determinants of health that may impact individuals and focus on "well care" rather than waiting for a patient to become ill with population health segmentation. Data analytics and population health segmentation are increasingly being used in collaboration with social determinants of health to collect data to understand better the environmental factors that may influence an individual's health. As data analytics in population health management continues to grow, consumer health informatics is expected to expand in healthcare and the teaching and provider world. As a result, this study focused on the advantages and disadvantages of implementing data analytics approaches used in population health segmentation.

BACKGROUND OF THE STUDY

Global population growth has resulted in the introduction of new healthcare requirements. Several nations have also been subjected to unprecedented health crises, leaving them unsustainable. With the emergence of pandemics like Covid-19, these countries have found it more challenging to meet the individual healthcare needs of their populations. Furthermore, environmental factors such as climate change and sociocultural differences

among many people have significantly impacted and challenged healthcare provision to needy patients. Again, some artificially created human factors, such as immigration caused by wars and other suppressing issues, have made it difficult to fully curb the healthcare challenges they bring to their new settlement areas.

As a result, population health segmentation has been predicted as the healthcare beacon of hope for maximizing limited resources to meet overwhelming healthcare challenges. According to Chong et al. (2019), implementing population health segmentation in the healthcare delivery system will revitalize the population-level strategy by identifying a population's healthcare needs and resource allocations. According to the studies, various governments and organizations can thus implement strategies that elevate healthcare services to each segment, creating and tracking segment-appropriate success metrics.

One of the primary goals of population health segmentation is to improve the efficiency of healthcare service delivery. However, due to the previously mentioned difficulties in achieving this, most governments and organizations have turned to data analytics to help mainstream and target resources based on the existing hierarchy of needs. Such actions have been linked to improved quality of life and increased utilization of healthcare resources. Various studies, however, have looked at data analytics in healthcare from various angles. Unfortunately, most of these studies have never focused on data analytics' actual and specific relevance in population health segmentation. Even though these previous studies have extensive knowledge on the application of data analytics in healthcare systems, they have neglected the population health segmentation aspect in their recommendations for future research or the missing pillar for efficient healthcare service delivery.

According to Mehta et al. (2020), the healthcare industry's focus is shifting toward a patient-centric model, which has since incorporated data-driven ideologies. Harvard Business School examines the significant roles data analytics has played in the healthcare delivery realms, such as improving practitioner evaluation and development, because the data gathered can be used to weigh in on potential areas for improvement, for example.

Dash et al. (2019) conducted another study to assess the impact of big data on healthcare systems. The study acknowledged the efforts made by both the public and private sectors in generating, storing, and analysing the big data required to improve daily diagnosis and prescriptions. The study further argued that massive datasets have emerged due to biomedical research being important to public healthcare.

A survey conducted by Cozzoli et al. (2022) identified the gap in determining the state of the art of big data analytics used by healthcare organizations, the benefits for both health managers and organizations, and the future trajectories of big data analytics research in healthcare. The study shed light on the relationship between big data analytics and healthcare.

Batko and Ślęzak (2022) provide an analytical overview of structured and unstructured data analytics in medical facilities with a case study from Poland. The paper characterized big data use by emphasizing the importance of data analytics in healthcare systems. The article also acknowledged healthcare as a complex system with numerous stakeholders, including patients, doctors, hospital administration, pharmaceutical companies, and healthcare decision-makers. The complexity of the data generated by such a network of partnerships necessitates data analytics to aid decision-making processes.

As a result, the role of data analytics in determining population health segmentation is lacking in various studies. As complex and multifaceted as healthcare is in meeting its demands, population health segmentation is critical for delivering efficient and

impactful services. Other scholars' objectives in determining the impacts of data analytics in population health segmentation and the overall well-being of healthcare services delivery have trumped the link between data analytics and population health segmentation.

PROBLEM STATEMENT

Data analytics is essential in healthcare management. It has also been discovered that data analytics in healthcare systems take different approaches depending on the needs of each facility segment. Furthermore, it can be deduced that the use of data analytics in the healthcare system is gradually but steadily replacing traditional methods that have since become obsolete and irrelevant regarding how healthcare needs are met. However, for these efforts to be realized, data analytics must be used in population health segmentation. Leveraging healthcare gaps will necessitate a strong strategy for ensuring that available resources are used for specific needs by specific population segments.

However, previous research has shown little of how this could be significant. With most studies focusing on general healthcare and data analytics, there is a gap in understanding how important data analytics is in forming population health segmentation. To address this, it was determined that research on how data analytics approaches have mainstreamed population health segmentation was required. To summarize, this study aimed to investigate data analytics approaches and their applications to population health segmentation and to better understand the roles, benefits, and barriers to data analytics implementation in population health segmentation in healthcare systems.

JUSTIFICATION OF THE STUDY

Health is determined through factors that are salient to humans' well-being. As a result, it keeps evolving from one community or society to another. Sociocultural and environmental factors are pivotal in transforming an individual's health needs. A population health segmentation exercise is needed to counter the unprecedented effects of revolving healthcare needs globally (Scheufele et al., 2022). However, with the huge meta data that characterize the current healthcare systems, coming up with population health segments that conform to the actual needs of a population is quite challenging. Thus, data analytics approaches are being sought to aid in reforming population health segmentation. Regrettably, there are no known metrics for determining the significance of this action in terms of impacts on the segmented people. Studies on the general role of data analytics in healthcare management assume the implications of extracting population health segmentation and how it can be revamped through data analytics to improve healthcare services. Thus this study aimed at determining the impacts of data analytics approaches on existing population health segmentation strategies utilizing already existing secondary data and to choose the benefits and barriers to implementation of these data analytics approaches to population health segmentation in healthcare service delivery in the UK among healthcare providers.

RESEARCH AIM

The main aim of this research was to determine the impacts of data analytics on population health segmentation. The use of synthetic patient and population health data for the state of Massachusetts achieved this.

RESEARCH OBJECTIVES

To utilize machine learning to segment an existing population health dataset and model patients' survival risk levels.

To determine the benefits of data-driven population health segmentation on the UK health service delivery.

To identify barriers to implementing data-driven population health segmentation approaches in the UK.

RESEARCH QUESTIONS

How can data analytic tools segment an existing population health data set?

What are the benefits of data-driven population health segmentation approaches on the UK health service delivery?

What are the barriers to implementing data-driven population health segmentation in the UK?

THEORETICAL FRAMEWORK

Recognizing the moral imperative to achieve desired health outcomes, public health practitioners, researchers, and community advocates have begun to advocate for new, data-driven approaches to population health segmentation (McLeroy et al., 1988). This study employs the social-ecological framework (SEF) to address the needs, benefits, challenges, and existing opportunities to develop more efficient integrated data analytics approaches that consider the interface between health protection and health promotion via population health segmentation, with a focus on improving patient care.

The social-ecological framework (SEF) investigates how multiple levels of influence, including intrapersonal, interpersonal, institutional, community/ society, and policy, can affect health outcomes (McLeroy et al., 1988). The primary objective of this research was to determine the effects of data analytics on population health segmentation, which was accomplished through synthetic patient and population health data for Massachusetts.

Using the SEF as a guiding framework and taking into account that population health is influenced at various levels, this study more thoroughly identified the elements of data analytics approaches that are likely to be effective in improving population health segmentation, the benefits of applying data analytics, and the barriers to implementing data analytics approaches to population health segmentation.

LIMITATIONS OF THE STUDY

This study is limited to the use of secondary sources of data. The results can therefore be biased in the context of the current occurrences and dominant factors. The general findings will be based on the reviews of the selected pieces of literature.

LITERATURE REVIEW

The History of Data Analytics

Deryck, (2020) writes in his article "A Brief History of Data Analytics" that recent news stories about data have stoked curiosity in unprecedented ways. The author claims that, in reality, data has long been a part of our culture and has only undergone a few developments and changes in its long history. These developments are intriguing because they each improve on the capabilities established by the forefathers while remaining consistent with what came before. However, data has shifted from being the domain of a few cliques to becoming a mass-market concept, a democratization move that carries both benefits and risks. However, this increased awareness has not been accompanied by a general understanding of these breakthroughs, leading to mistrust, naiveté, and cynicism.

Data analytics has existed since the dawn of civilization. He asserts that the earliest records of writing that have survived are examples of data analytics rather than ethereal writings such as poems, great speeches, love letters, or novels. In Sumeria, an ancient region of Iraq, scribes created the first database. They compiled a list of state-employed ploughmen and recorded it on clay tablets. Furthermore, these inscriptions used this raw data to calculate their pay, giving rise to data analytics. Over time, technology-enabled the replacement of clay with papyrus and, later, paper, simplifying and lowering the cost of capturing and retaining information. The development of algebra and the decimal system in the ninth century CE (also in Iraq) improved computation and data structuring. However,

the author claims that more inventions had to wait another millennium.

Deryck (2020) asserts that before computers' ability to analyze larger amounts of data in the 1950s, 1960s, and 1970s, such analytics degrees remained largely static. He claims that the first databases were built from information previously stored in paper journals, allowing for an expansion of the breadth and depth of what was published. The development of mainframe computers enabled programmers to produce reports that businesses could use. As a result, the "canned report" was born. More significant than what the Sumerians created 3000 years ago, but still somewhat rigid. If you didn't like the report, it's unfortunate, in the author's opinion. And basic search and customization were later added to these static reports, but in reality, they were still only limited by the programmer's perception of what information should be displayed.

Early on, the data in the report was frequently linked to the prefabricated report (Deryck, 2020). As a result, its scope was effectively limited to what the analyst could see on the screen. As a result, the data that supported your national report, which included a list of facts about countries, only included information about the countries in question. This was fine when there were only a few reports, but as they grew, so did the number of data sets, causing a maintenance problem (which, for many businesses, is still an issue after 50 years!). To address this, these small data sets were combined into larger structures, and "normalization" was used to remove overlaps, duplicates, and inconsistencies. In general, however, the desire for well-managed and structured databases that enabled people to save more triumphed. This immediately created friction between individualization and uniformity, which is still visible today.

As storage capacity increased, it became more difficult to query databases, necessitating using a consistent method (Deryck, 2020). This also applied to the increasing number of database vendors. Thus, in the 1970s, SQL (Structured Query Language, or "SEE-KWEL"), a logical language, was developed. To achieve the desired results, proficiency in the language, as with any other, was required, so the official discipline of data analysis was established. Unfortunately, as the databases grew larger and more complicated, this became more error-prone, more difficult to verify, and took longer to complete. As personal computers replaced mainframes, the major flaw in the prefabricated report became clear. Business users preferred to view data how they saw fit rather than how the programmer intended. MS Windows' visual features also raised expectations for data presentation.

According to author Deryck (2020), the major issue with scripted reports became clear when personal computers replaced mainframes. According to the author, business users wanted to view data how they wanted to, not how the programmer intended. The MS-Windows aesthetic features raised expectations for data presentation as well. They also desired to see the big picture, which made it difficult to physically connect data from various firm divisions to the group level and aggregate it across them. Furthermore, they desired to view data at a more granular level rather than simultaneously viewing a flat, one-dimensional picture (for example, defined per country) (say, per region or town). In the writer's opinion, this "drill up and down" was the first significant development.

According to Deryck (2020), the problem with this level of flexibility is that the perfectly standardized data for the prepared reports couldn't keep up, resulting in the creation of a brand-new "denationalization" procedure based on how a specific person wanted to see it. Deryck also claims that data was organized into "cubes" rather than tables, with dimensions for time, product category, region, and other variables. This significantly advanced the data analytics field's development because it required technical expertise to run SQL and some creativity and intelligence to select the

appropriate dimensions and measurements that benefited the cube.

No matter how simple it was to create the reports, according to Deryck (2020), the analysis was always retroactive and relied on events from the past. According to the author, data cubes could only display what already existed and could not provide insight into what was yet to happen. Although the past could be used to predict the future, the data could not openly express what was not yet present. With its seemingly limitless supply of new data, the Internet was the biggest surprise to business intelligence in the 1990s. Data on clay tablets, columns and rows in tables, cubes. This was insufficient to deal with the situation. Big data was first conceived.

According to Deryck (2020), big data is an architectural framework that allows for storing and retrieving massive amounts of unstructured data. He claims that the last feature enabled the most innovation because, in the past, massive data storage required structure, which limited its scalability. After all, the difficulty of implementing this standardization increased proportionally to its size.

Furthermore, the author observes that, in a comparable time frame, the concept of allowing others to store your data—such as with Microsoft Azure and Google's cloud storage—worked to support this design because any element of your business could access all components of this large (and virtual) data store. Again, the open-source nature of these allowed for significant scalability.

To sum it up, Deryck (2020) contends that data analysts can begin to gain more from data by developing a scenario or hypothesis that forecasts a specific outcome of seemingly unrelated behaviour in the past and comparing how this held true in reality during the present using a small test set. In theory, this could forecast the future, draw conclusions, inform decision-making, and do other things. This was akin to alchemy from a business, social, and political standpoint. But how do you choose which scenarios to develop? There is far too much data to manage and far too many possible data analytics combinations.

Using Data Analytics to Create an Intelligence-Led Approach to Population Health Management

Population health management has become increasingly important for health systems as the industry focuses on value-based care. Because health systems have access to vast amounts of data, effectively utilizing that data to inform population health management strategies is critical. According to Shaban-Nejad et al. (2018), data analytics is required for population health management to identify populations and their healthcare needs, assess the quality of care delivered to these populations and ensure the appropriate individuals receive the appropriate care. They also argue that, regardless of a health system's resources, population health management should begin with a three-party data review strategy to help health system leaders understand the pulse of their population and inform future strategies. Gamache et al. (2018) state that population health management systems can carry out the following three intelligence-led approaches to population health management through the use of data analytics: collecting data from various sources and converting this data into a valid form, applying analytics to the data – metrics, reports, trends, graphs, and word lists; and managing the care for the population – work lists for care managers, alerts and reminders for providers, postcards to patients, and reminder calls.

Population health management systems can use data analytics to gather data from diverse sources and transform it into a useful form. According to Raghupathi and Raghupathi (2018), physician offices, hospitals, and laboratory firms are the main separate, independent healthcare suppliers. Although each participant has a unique patient data set, this data is often incomplete. The scholars emphasize that data must be integrated from many entities,

including payers (claims), physician practices, and hospitals, to lay the groundwork for data-driven population health management (medical records). Atasoy et al. (2019) note that the data formatting into a standard structure, the matching of terms and codes, and the mapping of patient and provider identification are transformations that require data from various segments of population health management can be made possible using data analytics.

Atasoy et al. (2019) argue in their study that population health management systems can use data analytics to generate intelligence-led population health segmentation measurements, trends, reports, infographics, or task lists. Using standard quality metrics or developing population health segmentation-specific metrics and care management components, such as software that generates work lists for patients who need to be contacted for an intervention (such as a phone call, education session, or home visit) and tools to track the care provided to patients.

According to Raghupathi and Raghupathi (2018), data analytics can be used for work lists for care managers, notifications and reminders for clinicians, postcards to patients, and reminder calls to manage population care. For example, data analytics techniques enable risk stratification within each population health segment and identification of patients (or population members) who require special attention for better care management. According to the researchers, risk stratification is a clinical activity that helps identify which parts of the population have chronic diseases and require better care management, rather than just a financial exercise to determine which members cost the most or use the most resources. For instance, data analytics techniques enable the stratification of risk within each population health segment and the identification of patients (or population members) who require special attention for better care management. According to the researchers, risk stratification is a clinical activity that helps identify which parts of the population have chronic diseases and require better care management, rather than just a financial exercise to determine which members cost the most or use the most resources.

Identifying Effective Interventions using Data Analytics

The application of data analytics in medical research, according to Maryville University (2022), can enhance public health institutions' ability to forecast disease outbreaks, increase disease prevention, improve the quality of life, and lengthen lifespans. And that the most critical health issues that the world's populations face are being addressed through research that data analytics supports:

Data scientists at Blue Cross Blue Shield and the analytics company Fuzzy Logix have discovered 742 risk indicators that can predict someone's likelihood of taking opioids with high accuracy.

The Cancer Moonshot program, started by former President Barack Obama in his second term, uses cutting-edge data analytics tools to uncover the patterns and cancer treatments used worldwide with the best success rates.

Doctors can consult a database of more than 30 million electronic health records (EHRs) produced by the research cooperation company Optum Labs to support their treatment choices. The database has proven highly beneficial when treating individuals with complicated medical histories or several ailments.

With the use of algorithms that can more precisely identify patterns that point to a specific diagnosis, AI-based analytics holds the potential to enable radiologists to "read" images.

With the help of the Mental Health Research Network, Kaiser Permanente could accurately identify patients most likely to attempt suicide by analyzing their electronic health records (EHRs) and the answers to a standard depression questionnaire.

According to, Data analytics aid in early disease identification, better patient documentation, the provision of telehealth services, the removal of human bias for better outcomes, and the management of financial mistakes and risks (Built-In & HealthITAnalytics, 2022). The potential benefits of data analytics in healthcare are seen in seven key areas, including supporting clinical treatment decisions made by doctors and other healthcare professionals, improving the accuracy and speed of identifying patients at the highest risk of disease; providing more detail in patient EHRs; making healthcare delivery more efficient, which lowers costs; promoting preventive measures by giving patients greater insight into their health; and, finally, improving the accuracy and speed of identifying patients at highest risk of disease (Maryville University, 2022).

According to a report from Maryville University (2022), data analytics helps analyse clinical data to promote medical research. The report further highlights that data analytics enhance scientific research in numerous health-related fields by collecting and analysing clinical data from diverse sources. EHRs, electronic medical records, personal health records, and public health records are some of the best resources for clinical data that can be analysed using data analytics techniques, as detailed below:

EHRs incorporate test results, diagnoses, treatment plans, allergies, X-rays, and other medical images for a patient in standardized digital forms. This facilitates information sharing and creates privacy and legal compliance constraints that restrict how the data may be used.

Like EHRs, electronic medical records only contain data from patients' paper charts made in medical offices, clinics, and hospitals. Their greatest significance is in tracking a patient's healthcare through years of visits and screenings. They are mainly utilized for diagnosis and treatment.

Patient-maintained personal health records, as opposed to those kept by healthcare providers, preserve a patient's medical care history. The documents are not designed to replace the medical records that healthcare practitioners keep legally; they are meant to aid patients in managing their health.

One of the most promising sources of healthcare data for medical research is public health records. An illustration of a cloud-based data science platform is the Cancer Research Data Commons (CRDC) of the National Cancer Institute, which connects data analytics tools with data repositories that house genomic, proteomic, comparative oncology, imaging, and other types of data.

The report from Maryville University (2022) further reveals that patient data can be used to enhance health outcomes thanks to data analytics. And that in healthcare settings, quality improvement aims to provide patients with safe, efficient care while decreasing the trauma connected to such care. To accomplish this, healthcare professionals gather and analyze patients' data, increasingly in real-time, to better understand today's complex healthcare environments, to create and implement a systematic approach to improve patient outcomes and to continuously develop, test, and implement improvements to healthcare processes. Moreover, healthcare professionals may decrease readmission rates, reduce errors, and more accurately pinpoint populations at risk by evaluating patient data. These studies use patient data such as blood sugar levels, body temperature, blood test findings, and the patient's care preferences.

Roles of data analytics in the effective planning of health resources and service delivery

Pressures to lower costs, improve coordination and outcomes, provide more services with fewer resources, and become more patient-centric present a challenge for health institutions worldwide (Bardsley et al., 2019). They emphasize, however, that there is mounting evidence that the sector is being made more difficult by ingrained

inefficiencies and subpar clinical outcomes. According to the researchers, developing analytics expertise can help these health institutions use "big data" to provide actionable insights, establish long-term goals, improve outcomes, and shorten time to value, as detailed below:

Data analytics aids in gaining better insights for efficient planning of healthcare resources and services, which in turn can aid in demonstrating value and producing better results, such as the development of new treatments and technologies (Bardsley et al., (2019). They claim that informed and educated customers can become more responsible for their health by gaining knowledge. Kumar and Singh (2018) assert in their study that analytics can help to address everything from minor details to complex processes, support exploration and discovery, aid in the design and planning of policies and programs, improve sustainability, improve service delivery and operations, and provide a way to measure and evaluate critical organizational data. Among its most significant advantages are the ability to increase healthcare access, align compensation with performance, and support cost growth control.

Palanisamy & Thirunavukarasu (2019) state that new analytics techniques can generate clinical and operational improvements to ensure efficient healthcare resources and service delivery planning. According to the researchers, analytics in healthcare is progressing from a traditional baseline of transaction monitoring using simple reporting tools, spreadsheets, and application reporting modules to a model that will eventually incorporate predictive analytics and enable health institutions to "see the future," create more individualized population health management systems, allow dynamic fraud detection, and predict patient behaviour.

METHODOLOGY

Sander's research onions concept was utilized to explain the methodology for this research work, identifying the suitable methodological choice, philosophy, theory development, and data collection procedures.

Saunders et al. (2007) divided the research onion into three decision levels. The first two outer rings are Research philosophy and Research approach; the second is Research design, which comprises methodological choices, research plan, and time horizon; and the third is data collection and analysis techniques.

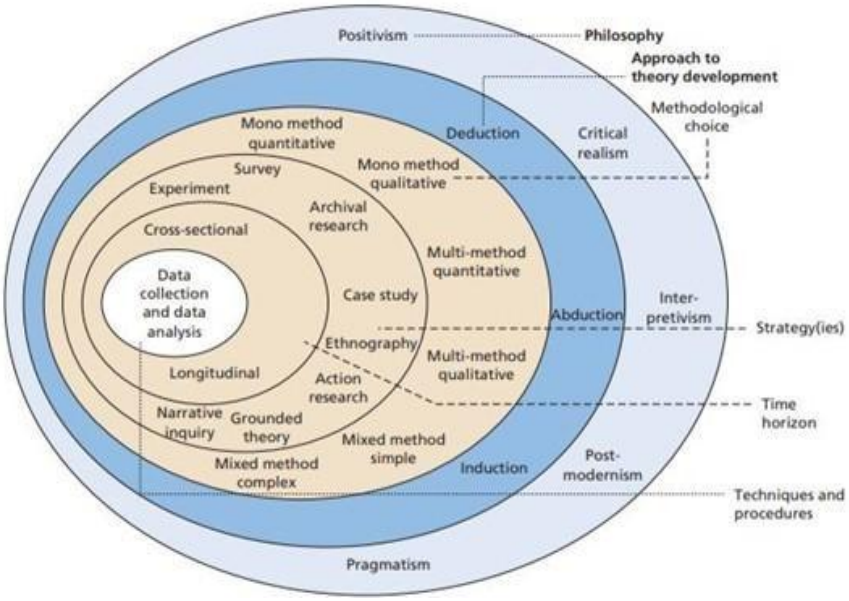

Figure 1: Research Onion
Source: Saunders et al. 2019.

Research Philosophy This study adopted the pragmatism philosophy which approaches research from a practical point of view, where knowledge is not fixed, but instead is constantly questioned and interpreted, consisting of an element of the researcher's involvement and subjectivity, specifically when drawing conclusions based on study participants' responses and decisions (Saunders et al., 2007). Pragmatism is based on the proposition that researchers should use the philosophical and methodological approach that works best for the research problem being investigated (Tashakkori et al., 1998).

In this study, the researcher used existing secondary data best suited to answering the research questions, as in the case of pragmatic philosophy. Conclusions were drawn from the analysed secondary data and qualitative data from interviews. This way, the researcher could focus on presenting the study's outcome more than the philosophical ideas behind the study.

RESEARCH APPROACH

This study adopted both the inductive and deductive approaches. The inductive approach was used to develop the research topic and construct a working theory for the study. This was then followed up with a reasoned approach to validate the conclusion. Saunders et al. (2007) assert that combining the inductive and deductive methods is a better way to mitigate the risk of research bias in the study.

RESEARCH DESIGN

The study adopted a descriptive survey with a mixed research design approach. Kothari (2004) defines a descriptive study with a mixed research design approach as a type of research in which a researcher blends part of qualitative and quantitative research methodologies with the comprehensive objective of knowledge and confirmation. This aided in acquiring in-depth insights on the data analytics approaches to population health segmentation and their impact on healthcare service delivery. The method's capacity to elicit a wide variety of baseline data makes it the most suitable for this study (Creswell, 2003).

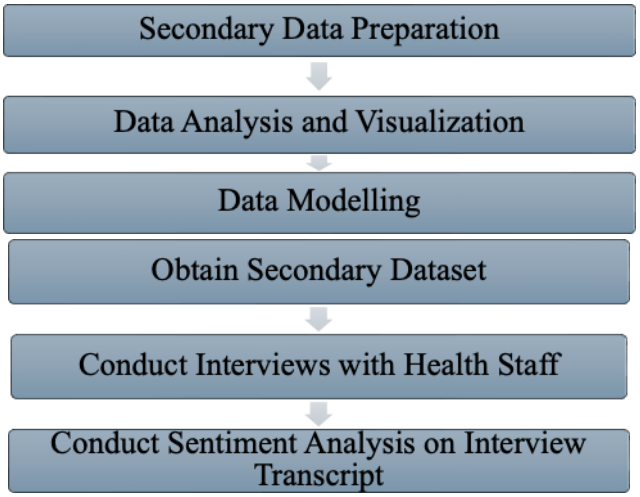

Figure 2: Methodology Process outline

LOCATION OF STUDY

This research was carried out in Sheffield. It is a city in the county of South Yorkshire. With a population of approximately 556,500, the city is located 47 kilometres south of Leeds, 51 kilometres east of Manchester, and 53 kilometres north of Nottingham (Census, 2021). Sheffield Teaching Hospitals NHS Foundation Trust, Sheffield Children's NHS Foundation Trust, and Sheffield Health and Social Care NHS Foundation Trust provide social and health

care services in the city. Sheffield Teaching Hospitals NHS Foundation Trust provides healthcare to residents of Sheffield and South Yorkshire (primarily adults). The Royal Hallamshire Hospital, Weston Park Hospital, Charles Clifford Dental Hospital, and The Northern General Hospital are all part of the trust. Sheffield Children's NHS Foundation Trust is responsible for providing healthcare to children in Sheffield. Sheffield Health and Social Care Services NHS Foundation Trust provides mental health services, learning disability services, substance abuse services, and long-term neurological conditions. The Sheffield Institute for Motor Neuron Disease is part of this trust (also known as Sheffield Institute for Translational Neuroscience) (Census, 2021).

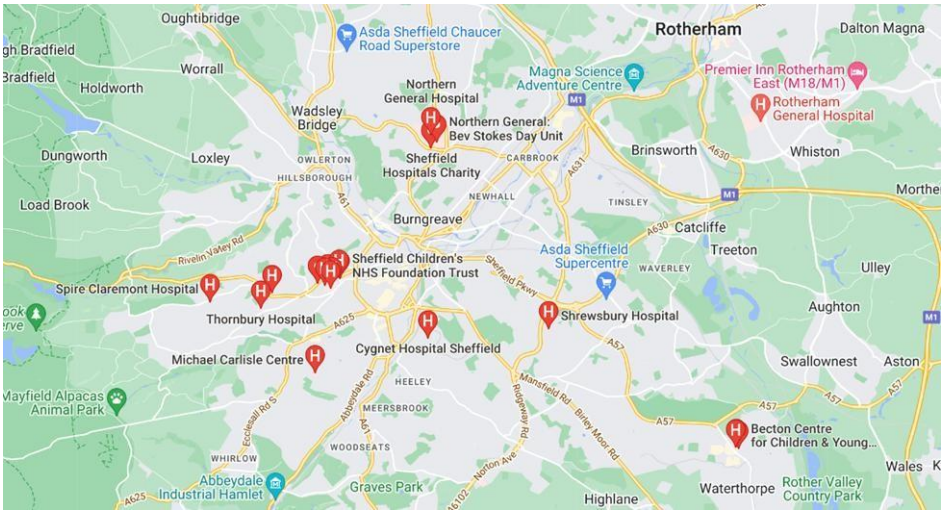

Figure 3: Map of Sheffield
Source: Google Maps

TARGET POPULATION

The study's target population were key healthcare workers in the hospitals, including medical directors, health records personnel, and monitoring and evaluation experts. These key healthcare workers redefine the development and implementation of approaches to the medical care of the public and inform the development of policies and strategies in achieving universal healthcare coverage for the population. To determine their influence and role in implementing the data analytics approaches to population health segmentation, the research focused on the healthcare professionals with key roles in implementing approaches to treatment and handling data within the chosen study sites. Inclusion Criteria All medical directors have been practising for at least 3 years and influence hospital policy development and implementation. Health records officers within hospitals handle and secure patient data in the respective hospitals. Monitoring and evaluation officers stationed in the hospitals participating in the implementation of health frameworks from the Ministry of Health Exclusion Criteria All other clinical staff was excluded from the study.

SAMPLING DESIGN AND PROCEDURE

The hospitals in Sheffield were organized into strata, i.e., Sheffield Teaching Hospitals NHS Foundation Trust, Sheffield Children's NHS Foundation Trust, and Sheffield Health and Social Care NHS Foundation Trust. Key informant interview participants were selected using the purposive sampling method. The ability to communicate approaches used in population health segmentation, the applications of data analytics on population health segmentation, the working experience (more than three years) in the field to express community experience and practice, and the willingness

to participate are all characteristics that will be used to identify study participants.

Kothari (2004) points out that the purposive sampling method ensures that the chosen cases help answer the research questions and achieve the research objectives. It was best suited for this study because it discovered the meaning and implication of the different myths and misconceptions and benefited from an intuitive approach. It is also cost-effective and time effective. Resident enumerators were used to selecting study participants according to the inclusion criteria. Three medical directors and one nurse in the five hospitals run by the NHS Foundation Trusts who influence policy development and implementation within their hospitals were the subjects of key informant interviews. The medical directors gave important information on the benefits and barriers to implementing data-driven population health segmentation approaches in the UK health service delivery. Monitoring and evaluation experts conducted in-depth interviews to explore the various opinions and concerns about data-driven approaches to population health segmentation. They were chosen based on their first-hand knowledge of data analytics and population health segmentation.

RESULTS AND DISCUSSION

Secondary Data Results and Analysis

An analysis should visualize the dataset obtained (Chen et al., 2022), which allows us to understand and gain insight into the segmented population health dataset. The results and visualization displayed were after the data pre-processing, as shown below.

```
dummy_cleaned_dataset.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 51251 entries, 0 to 57074
Data columns (total 32 columns):
 #   Column                                Non-Null Count  Dtype
---  ------                                --------------  -----
 0   Age                                   51251 non-null  float64
 1   Acute deep venous thrombosis          51251 non-null  uint8
 2   Acute pulmonary embolism              51251 non-null  uint8
 3   Acute respiratory distress syndrome   51251 non-null  uint8
 4   Acute respiratory failure             51251 non-null  uint8
 5   Anemia                                51251 non-null  uint8
 6   Body mass index 30+ - obesity         51251 non-null  uint8
 7   COVID-19                              51251 non-null  uint8
 8   Cough                                 51251 non-null  uint8
 9   Dyspnea                               51251 non-null  uint8
 10  Hypertension                          51251 non-null  uint8
 11  Hypoxemia                             51251 non-null  uint8
 12  Injury of heart                       51251 non-null  uint8
 13  Loss of taste                         51251 non-null  uint8
 14  Pneumonia                             51251 non-null  uint8
 15  Prediabetes                           51251 non-null  uint8
 16  Respiratory distress                  51251 non-null  uint8
 17  Sepsis caused by virus                51251 non-null  uint8
 18  Septic shock                          51251 non-null  uint8
 19  Suspected COVID-19                    51251 non-null  uint8
 20  Wheezing                              51251 non-null  uint8
 21  Survived                              51251 non-null  uint8
 22  AdmittedWithCovid                     51251 non-null  uint8
 23  AdmittedWithCovidOnICU                51251 non-null  uint8
 24  AdmittedCovidPatientsOnVentilator     51251 non-null  uint8
 25  AdmittedCovidSurvivors                51251 non-null  uint8
 26  Asian                                 51251 non-null  uint8
 27  Black                                 51251 non-null  uint8
 28  Native                                51251 non-null  uint8
 29  Other                                 51251 non-null  uint8
 30  White                                 51251 non-null  uint8
 31  GENDER                                51251 non-null  uint8
dtypes: float64(1), uint8(31)
```

Figure 4: Pre-processed Dataset

Target Variable

A dataset's target variable is the feature of a dataset about which we want to understand more. A supervised machine learning algorithm learns patterns from existing data and discovers relationships between various features of the dataset and the objective (Antognini et al., 2021). So, we started by examining the target or dependent variable, i.e., Survival status. This feature is a categorical variable. The following analysis is a scatterplot of the survival status for Covid-19 patients with various critical underlying conditions in the dataset.

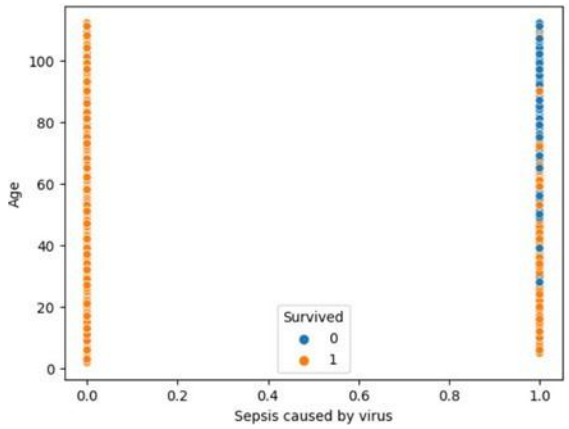

Figure 5: Sepsis versus survival status by age

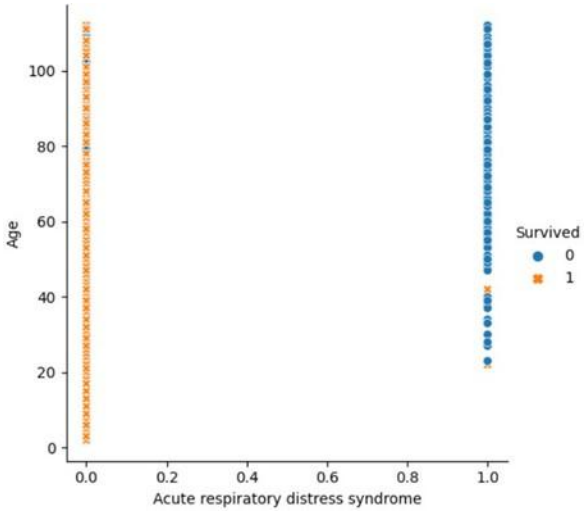

Figure 6: Acute Respiratory Distress Syndrome Vs Survival Status According to Age

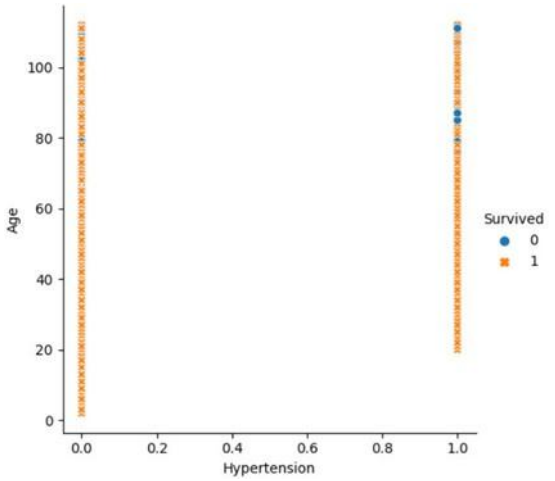

Figure 7: Hypertension Vs Survival Status According to Age

Correlation

Correlation plots determine which variables are connected and how strong their relationship is (Gu et al., 2016). The variables are transformed into numerical variables, and a heatmap is plotted to illustrate the data using different colours ranging from 0 to 1, with 1 being the most correlated and 0 representing the least correlated.

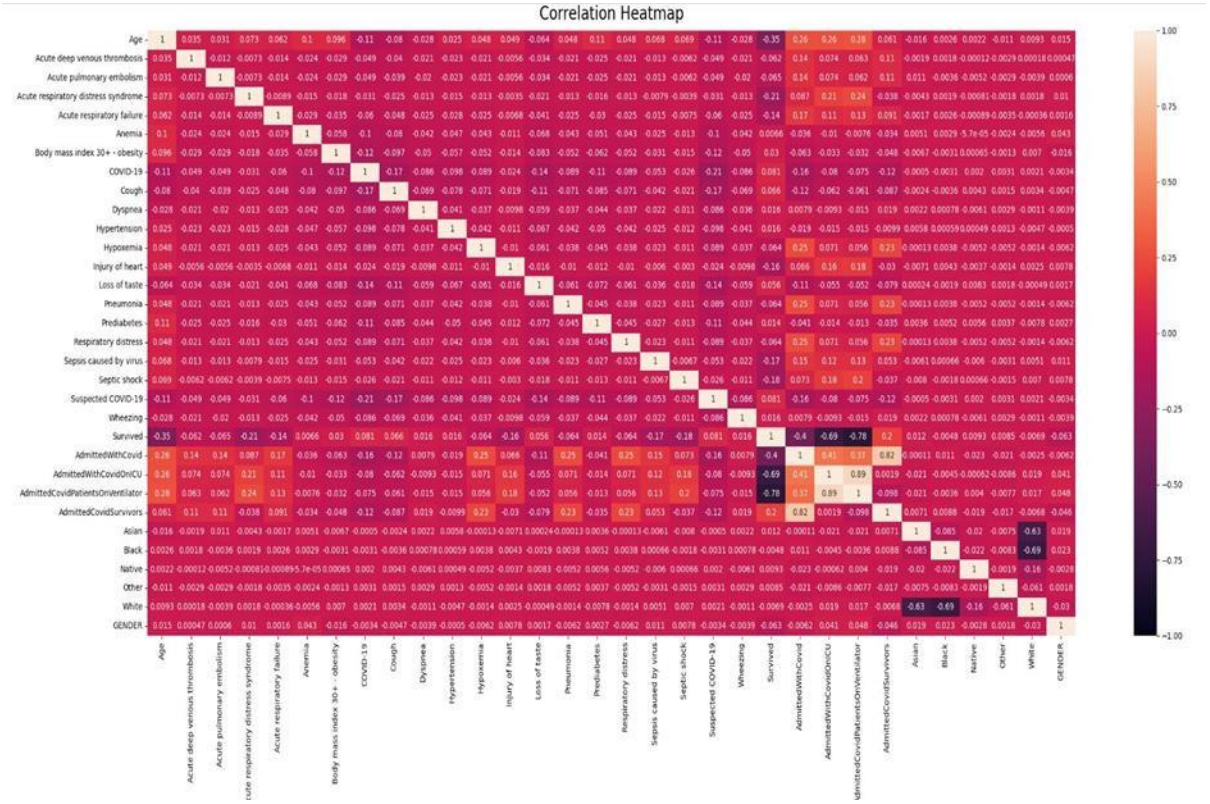

Figure 8: Heat-map of the Numerical Independent and Dependent Feature

Modeling

Here, the three algorithms utilised are decision tree, random forest and logistics regression. Furthermore, the dataset is imbalanced. Based on the dataset knowledge, a confusion matrix and evaluation metrics are introduced to understand the model's performance.

Decision Tree

The Decision Tree algorithm was trained and tested on the dataset in building the Data Tree model. The below result shows the model performance of the algorithm.

```
181…  print(classification_report(y_test, y_pred))
```

	precision	recall	f1-score	support
0	1.00	1.00	1.00	1354
1	1.00	1.00	1.00	14022
accuracy			1.00	15376
macro avg	1.00	1.00	1.00	15376
weighted avg	1.00	1.00	1.00	15376

Figure 9: Decision Tree Model Performance

Qualitative Results and Analysis

The study involved four respondents whereby all were key informants. All the respondents were healthcare workers based in hospitals. They represented different sectors of the health systems. The study sought their views on various population health segmentation, data-driven approaches, data analytics, the role of effectiveness on service delivery, the benefits, barriers and enablers to development and implementation, and

strategies to develop and improve data-driven population health segmentation approaches. Several themes were explored, including the general population health segmentation, where respondents talked about the understanding of data-driven approaches, if there are factors affecting the population health management strategies and if there are roles they play in developing and implementing data-driven approaches in population health segmentation strategies. The respondents also discussed issues concerning development, frameworks defining them, data analytics in creating an intelligence-led approach, benefits, barriers, and strategies.

Figure 10: Word cloud displaying frequency and keywords used

Population Health Segmentation

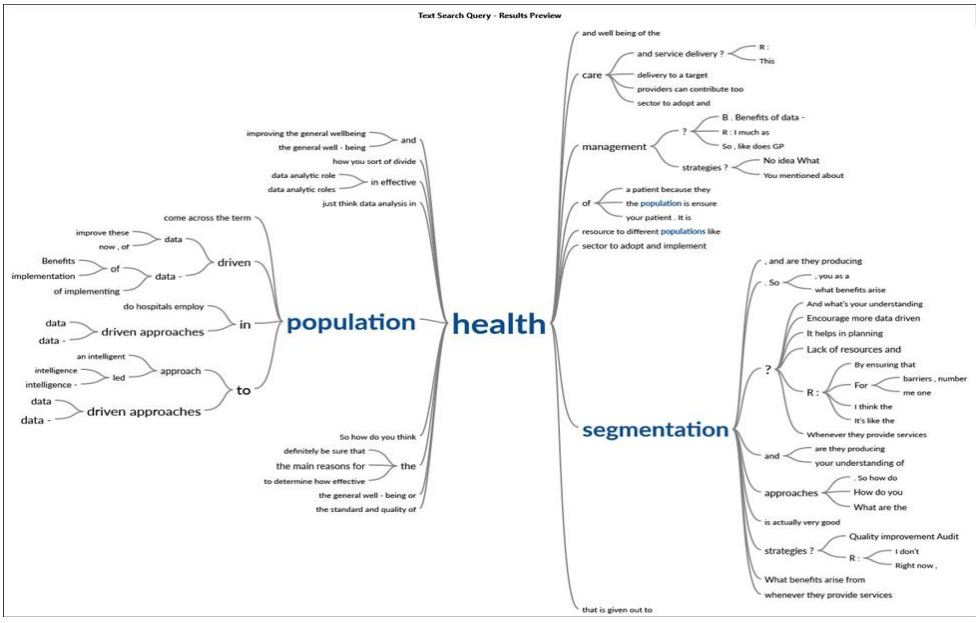

Figure 11: Word tree

Data-driven Approaches

This theme highlighted some of the aspects or approaches hospitals use to employ in population health management strategies. Under this theme, two sub-categories were developed: Frameworks to develop and implement data-driven approaches, and role in developing and implementing data-driven approaches.

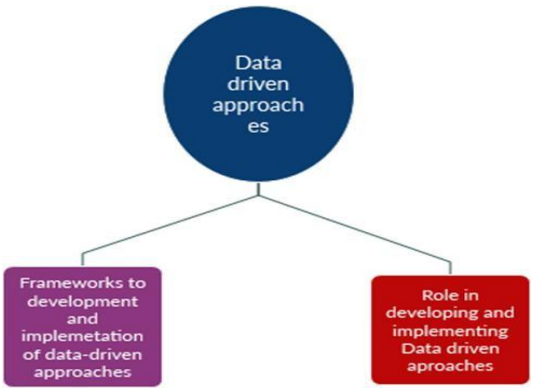

Figure 12: Mind map

Benefits of Data-driven Population

Health Segmentation Under this theme, the study sought to understand the benefits of data analytics in population health segmentation and its roles in effective healthcare and service delivery. This is mostly useful regarding health systems' existing data analytics structure to enable data-driven population health segmentation. Two sub-categories arise from this theme: population health segmentation, the Roles in effective healthcare and service delivery.

Roles in Effective Healthcare and Service Delivery.

This theme highlights the impact of data-driven population health segmentation on data analytics roles in effective healthcare and service delivery. The respondents, if not all, agreed and contributed that data analytics play a big role in service delivery and healthcare:

"I think without data analysis the healthcare would not function at all because there is data analysis even done from things like what medication we use and where we get it from, even the software, the systems that we use that use data analysis. So, I don't think healthcare would function at all without data analysis." [[Transcripts\\Recording 1>]

"They help in early detection of serious illness like cancer. They also help in early detection of deteriorating patients. A patient that would deteriorate will probably die. The data analytics like the tool-the news score, it helps us in the ward to know that," Okay this patient is deteriorating, if something is not done quickly the patient will most likely die." So, it helps in early detection of deteriorating patients. "[[Transcripts\\Recording 2>]

CONCLUSION

Population health management systems can use data analytics to generate intelligence-led population health segmentation measurements, trends, reports, infographics, or task lists Considering the secondary dataset we utilised for this study, analysis to covid-19 patients. While we covered what medications and procedures were used for covid-19 patients, there was no extensive research on which procedures and medications were more effective. The analysis can be extended to include what medication and procedures are more effective and check the impact of its supplies on the mortality rate. The study found a gap in knowledge of data driven approaches to population health segmentation from the primary data. Therefore, National Health Service (NHS) must channel more resources to increase awareness of using data to allocate health resources to the communities and decide the healthcare service patients need in the hospital.

Declaration by Authors

Acknowledgement: None

Source of Funding: None

Conflict of Interest: The authors declare no conflict of interest.

REFERENCES

1. Mishra A. Creating Machine Learning Models with Scikit- learn [Internet]. 2019. p. 79–114. Available from: https://www.researchgate.net/publication/335606294_Creating_Machine_Learning_Models_with_Scikit-learn
2. Hosmer DW, Lemeshow S. Applied Logistic Regression. Hoboken, NJ, USA: John Wiley & Sons, Inc.; 2000.
3. Annane D, Bellissant E, Cavaillon JM. Septic shock. The Lancet [Internet]. 2005 Jan 1;365(9453):63–78. Available from: https://www.sciencedirect.com/science/article/pii/S0140673604176678

4. Antognini D, Musat C, Faltings B. Multi-Dimensional Explanation of Target Variables from Documents. Proceedings of the AAAI Conference on Artificial Intelligence. 2021 May 18;35(14):12507–15.
5. Atasoy H, Greenwood BN, McCullough JS. The Digitization of Patient Care: A Review of the Effects of Electronic Health Records on Health Care Quality and Utilization. Annual Review of Public Health [Internet]. 2019 Apr;40(1):487–500. Available from: https://www.annualreviews.org/doi/10.1146/annurev-publhealth-040218-044206
6. Bardsley M, Steventon A. Untapped Potential:Investing n health and care data analytics. [Internet]. 2019. Available from: https://www.researchgate.net/publication/337022547_Untapped_PotentialInvesting_n_health_and_care_data_analytics
7. Batko K, Ślęzak A. The Use of Big Data Analytics in Healthcare. Journal of Big Data [Internet]. 2022 Jan 6;9(3). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8733917/
8. King EN, Ryan TP. A Preliminary Investigation of Maximum Likelihood Logistic Regression versus Exact Logistic Regression. The American Statistician. 2002 Aug;56(3):163–70.
9. Marija Burinskiene, Vitalija Rudzkiene. Application of logit regression models for the identification of market segments. Journal of Business Economics and Management. 2007 Dec 31;8(4):253–8.
10. Caelen O. A Bayesian interpretation of the confusion matrix. Annals of Mathematics and Artificial Intelligence. 2017 Sep 11;81(3-4):429–50.
11. Caldararo KL, Nash DB. Population Health Research: Early Description of the Organizational Shift Toward Population Health Management and Defining a Vision for Leadership. Population Health Management. 2017 Oct;20(5):368–73.
12. Caruana R, Lou Y, Gehrke J, Koch P, Sturm M, Elhadad N. Intelligible Models for HealthCare. Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '15. 2015;
13. Census 2021 [Internet]. Census 2021. 2021. Available from: https://census.gov.uk/
14. Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics [Internet]. 2020 Jan 2;21(1). Available from: https://link.springer.com/article/10.1186/s12864-019-6413-7
15. Chojnicki M, Neumann-Podczaska A, Seostianin M, Tomczak Z, Tariq H, Chudek J, et al. Long-Term Survival of Older Patients Hospitalized for COVID-19. Do Clinical Characteristics upon Admission Matter? International Journal of Environmental Research and Public Health [Internet]. 2021 Oct 12 [cited 2022 Apr 10];18(20):10671. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8535841/
16. Chong JL, Lim KK, Matchar DB. Population segmentation based on healthcare needs: a systematic review. Systematic Reviews [Internet]. 2019 Aug 13;8(1). Available from: https://systematicreviewsjournal.biomedcentral.com/articles/10.1186/s13643-019-1105-6
17. Ömay Çokluk Bökeoğlu. Logistic Regression: Concept and Application. Educational Sciences Theory & Practice [Internet]. 2010 Jun [cited 2025 May 23];10(3):1397–407. Available from: https://www.researchgate.net/publication/289651384_Logistic_Regression_Concept_and_Application
18. Cozzoli N, Salvatore FP, Faccilongo N, Milone M. How can big data analytics be used for healthcare organization management? Literary framework and future research from a systematic review. BMC Health Services Research [Internet]. 2022 Jun 22;22(1). Available from: https://bmchealthservres.biomedcentral.com/articles/10.1186/s12913-022-08167-z
19. Creswell J. RESEARCH DESIGN Qualitative, Quantitative. and Mixed Methods Approaches SECOND EDITION SAGE Publications International Educational and Professional Publisher Thousand Oaks London New Delhi [Internet]. 2003. Available from: https://cumming.ucalgary.ca/sites/default/files/teams/82/communications/Creswell%202003%20-%20Research%20Design%20-%20Qualitative%2C%20Quantitative%20and%20Mixed%20Methods.pdf
20. Dash S, Shakyawar SK, Sharma M, Kaushik S. Big data in healthcare: Management,

analysis and future prospects. Journal of Big Data [Internet]. 2019;6(1):1–25. Available from: https://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0217-0

21. Davis J, Goadrich M. The relationship between Precision-Recall and ROC curves. Proceedings of the 23rd international conference on Machine learning - ICML '06 [Internet]. 2006; Available from: http://pages.cs.wisc.edu/~jdavis/davisgoadrichcamera2.pdf
22. Deng X, Liu Q, Deng Y, Mahadevan S. An improved method to construct basic probability assignment based on the confusion matrix for classification problem. Information Sciences. 2016 May;340-341:250–61.
23. Martino MD, Furfaro S, Mulas MF, Mataloni F, Santurri M, Paris A, et al. [Population segmentation as a tool for planning community healthcare networks: the key role of social and health information systems.]. PubMed. 2022 Feb 1;113(2):97–104.
24. Nascimento IJB do, Marcolino MS, Abdulazeem HM, Weerasekara I, Azzopardi-Muscat N, Gonçalves MA, et al. Impact of Big Data Analytics on People's Health: Overview of Systematic Reviews and Recommendations for Future Studies. Journal of Medical Internet Research [Internet]. 2021 Apr 13;23(4):e27275. Available from: https://www.jmir.org/2021/4/e27275/
25. El Naqa I, Murphy MJ. What Is Machine Learning? Machine Learning in Radiation Oncology [Internet]. 2015;1(1):3–11. Available from: https://link.springer.com/chapter/10.1007/978-3-319-18305-3_1
26. El Naqa I, Ruan D, Valdes G, Dekker A, McNutt T, Ge Y, et al. Machine learning and modeling: Data, validation, communication challenges. Medical Physics. 2018 Aug 24;45(10):e834–40.
27. 27.Emami A, Javanmardi F, Akbari A, Kojuri J, Bakhtiari H, Rezaei T, et al. Survival rate in hypertensive patients with COVID-19. Clinical and Experimental Hypertension. 2020 Aug 24;43(1):77–80.
28. Evans WD, Thomas CN, Favatas D, Smyser J, Briggs J. Digital Segmentation of Priority Populations in Public Health. Health Education & Behavior. 2019 Nov 19;46(2_suppl):81S89S.
29. Rinaldo A, Solimun, Nurjannah. Computational Statistics with Dummy Variables. IntechOpen eBooks. 2022 Apr 6;
30. Friedman J H. A Recursive Partitioning Decision Rule for Nonparametric Classification. IEEE Transactions on Computers. 1977 Apr; C-26(4):404–8.
31. Galetsi P, Katsaliaki K, Kumar S. Values, challenges and future directions of big data analytics in healthcare: A systematic review. Social Science & Medicine. 2019 Sep;241(241):112533.
32. Gallo M, Sá MPBO, Doulamis IP, Hussein N, Laforgia PL, Kampaktsis PN, et al. Transcatheter valve-in-valve implantation for degenerated bioprosthetic aortic and mitral valves – an update on indications, techniques, and clinical results. Expert Review of Medical Devices. 2021 Jun 15;18(7):597–608.
33. Gamache R, Kharrazi H, Weiner JP. Public and Population Health Informatics: The Bridging of Big Data to Benefit Communities. Yearbook of Medical Informatics [Internet]. 2018 Aug 1;27(1):199–206. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6115205/
34. Ghosh C. Data Pre-processing. Data Analysis with Machine Learning for Psychologists. 2022;55–85.
35. Goncalves A, Ray P, Soper B, Stevens J, Coyle L, Sales AP. Generation and evaluation of synthetic patient data. BMC Medical Research Methodology. 2020 May 7;20(1).
36. Gu Z, Eils R, Schlesner M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics (Oxford, England) [Internet]. 2016 Sep 15 [cited 2020 Mar 9];32(18):2847–9. Available from: https://www.ncbi.nlm.nih.gov/pubmed/27207943
37. Healy LM. Logistic Regression: An Overview [Internet]. 2006. Available from: https://www.researchgate.net/publication/255597415_Logistic_Regression_An_Overview
38. Jolly S, Gupta N. Understanding and Implementing Machine Learning Models with Dummy Variables with Low Variance. Springer eBooks. 2020 Aug 2;477–87.

39. Jordan MI, Mitchell TM. Machine learning: Trends, perspectives, and prospects. Science [Internet]. 2015 Jul 16;349(6245):255–60. Available from: https://www.science.org/doi/abs/10.1126/science.aaa8415
40. Kindig D, Stoddart G. What Is Population Health? American Journal of Public Health [Internet]. 2003 Mar;93(3):380–3. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1447747/
41. E. Blankmeyer, Ryan TP. King, E. N., and Ryan, T. P. (2002), A preliminary investigation of maximum likelihood logistic regression versus exact logistic regression, the American statistician, 56, 163-170: Comment by blankmeyer and reply [3] (multiple letters). 2003 Nov 1;57(4):320–1. Available from: https://www.researchgate.net/publication/296378977_King_E_N_and_Ryan_T_P_2002_A_preliminary_investigation_of_maximum_likelihood_logistic_regression_versus_exact_logistic_regression_the_American_statistician_56_163-170_Comment_by_blankmeyer_and_reply_3
42. Kumar S, Singh M. Big data analytics for healthcare industry: impact, applications, and tools. Big Data Mining and Analytics. 2019 Mar;2(1):48–57.
43. Lu W, Yu S, Liu H, Suo L, Tang K, Hu J, et al. Survival Analysis and Risk Factors in COVID-19 Patients. Disaster Medicine and Public Health Preparedness. 2021 Mar 25;1–6.
44. LYNN J, STRAUBE BM, BELL KM, JENCKS SF, KAMBIC RT. Using Population Segmentation to Provide Better Health Care for All: The "Bridges to Health" Model. The Milbank Quarterly [Internet]. 2007 Jun;85(2):185–208. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2690331/
45. McLeroy KR, Bibeau D, Steckler A, Glanz K. An Ecological Perspective on Health Promotion Programs. Health Education Quarterly [Internet]. 1988 Dec;15(4):351–77. Available from: https://pubmed.ncbi.nlm.nih.gov/3068205/
46. Mehta N, Pandit A, Kulkarni M. Elements of Healthcare Big Data Analytics. Studies in Big Data. 2019 Oct 2;23–43.
47. Mihajlovic S, A Kupusinac, Dragan Vojo Ivetic, Berković I. The Use of Python in the field of Artifical Intelligence [Internet]. ResearchGate. unknown; 2020. Available from: https://www.researchgate.net/publication/366578422_The_Use_of_Python_in_the_field_of_Artifical_Intelligence
48. Olegas Niaksu. CRISP Data Mining Methodology Extension for Medical Domain. 2015 Jan 1;3(2):92–109. Available from: https://www.researchgate.net/publication/277775478_CRISP_Data_Mining_Methodology_Extension_for_Medical_Domain
49. O'Brien JM, Ali NA, Aberegg SK, Abraham E. Sepsis. The American Journal of Medicine [Internet]. 2007 Dec;120(12):1012–22. Available from: https://www.sciencedirect.com/science/article/abs/pii/S0002934307005566
50. Raghupathi W, Raghupathi V. An empirical study of chronic diseases in the United States: A visual analytics approach to public health. International Journal of Environmental Research and Public Health [Internet]. 2019 Mar 1;15(3):1–24. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5876976/

******