University for the Creative Arts

BERLIN SCHOOL OF BUSINESS & INNOVATION

Dissertation Title: Predicting disease outbreaks using big data analytics

Master title: MSc Data Analytics

Name: Rutwik Findoliya

Year: FEB 2024

ABSTRACT

The study examines the ability of big data analytics to help identify and predict outbreaks early, so this information can guide the development of an overall system for public health monitoring. With a positivist, deductive process, the team conducted a cross-sectional survey with 60 professionals from public health, epidemiology and healthcare analytics in the United Kingdom. The research highlights that people are moderately familiar with what big data can do, yet the use of these advanced technologies is less common than expected. Using data from electronic health records and climates is seen as safer than setting data collected from people's daily movements and online accounts. Privacy, data bias and algorithm transparency were clear ethical worries that prevented progress. Even though infrastructure and skill-related problems were mentioned, most respondents acknowledged that big data can greatly improve the practice of predicting epidemics in the future. This research adds to the knowledge base by examining what practitioners think, pointing out obstacles in operation and giving solutions for ethical leadership and inclusive design. It suggests that Government should provide tailored training, put more money into data systems and improve teamwork across sectors to encourage adoption. It needs models that are ethical, clear and able to be used widely so that new technology can be applied in predictive epidemiology.

CONTENTS

Contents

ACKNOWLEDGEMENTS

Statement of compliance with academic ethics and the avoidance of plagiarism

I honestly declare that this dissertation is entirely my own work and none of its part has been copied from printed or electronic sources, translated from foreign sources and reproduced from essays of other researchers or students. Wherever I have been based on ideas or other people texts I clearly declare it through the good use of references following academic ethics.

(In the case that it is proven that part of the essay does not constitute an original work, but a copy of an already published essay or from another source, the student will be expelled permanently from the postgraduate program.)

Name and Surname (Capital letters):

RUTWIK FINDOLIYA

....

Date: ....07 ..../......06..../...2024......

DISSERTATION THESIS

(Leave this page empty.)

INTRODUCTION

Introduction

Big data analysis is used in this study to help detect early signs of epidemics and boost both prediction of diseases and health monitoring. It explores where epidemic data comes from, how forecasting methods work and the ethical issues involved, to make sure epidemic forecasting helps keep the world healthy.

The identification and control of epidemic diseases have always been an important goal of health-care institutions, governments and probably the entire society. Disease surveillance and prediction have experienced a shift in the current society through technological advancements, especially the application of big data analytics (Ahmed et al., 2021). This particular field of study aims at identifying the population of diseases to investigate the potential of large datasets to help predict the development and projected growth of epidemics, as well as assist in prevention, resource optimisation, and the formation of policies (Sarumi, 2021).

Brief Analysis of the Research Topic

Endemics, including Coronavirus Disease 2019 (COVID-19), Ebola, SARS or Zika, have given a clear picture of the societal impact of delayed response as well as inadequate surveillance systems. The use of conventional epidemiological data through record retrieval, cross-checking the hospital records and laboratory confirmation is usually slow in identifying new disease outbreaks. This can lead to massive casualties, disruption of economic activities, as well as heightened panic in the society (Ahmed et al., 2021).

On the other hand, big data analytics offers the opportunity to increase the accuracy of forecasts by using new information from various sources such as social media, diseases, weather conditions, movement of the population, and searches for certain terms. Increased speed, volume, and type of data allow us to find early warning signs when traditional health systems have not yet noticed them (Souza, Leung and Cuzzocrea, 2020). For instance, HealthMap and Google Flu Trends have shown previously that the use of Internet-based systems has the potential to augment traditional surveillance. Therefore, the role of big data in epidemiology is more important than before due to the current globalised and urbanised society. It is important to note that computers' integration with epidemiological information produces less reactive and more preventive approaches to threats to people's well-being (Sarumi, 2021).

Trigger and Rationale

The motive for undertaking this research lies in two factors: COVID-19 and the digital transformation of healthcare data. Due to the highly disruptive nature of COVID-19, the importance of early detection that only higher-end disease surveillance systems could provide has been highlighted (Ahmed et al., 2021). Those countries that employed data science strategies were less flexible in stopping the spread compared to those in the group that used only public health measures.

The growing emergence of diseases like COVID-19 has therefore immensely driven difficulties in the traditional methods of disease monitoring and prediction. These techniques which rely on historical data sources are always very slow in identifying these occurrences and subsequently resulting in outbreak of diseases (Souza, Leung and Cuzzocrea, 2020). These epidemics are allowed to fester and spread due to the lack of early identification and containment measures taken hence contributing to high incidences of morbidity and mortality, economic impact and disruptions, and the ever increasing public fear. By incorporating real time data from such sources as that of social media, travel and environmental conditions the application of big data analytics in the improvement of systems of early warning presents immense possibility. It is important to realize that there is much that present schemes of surveillance can accomplish and it is to strive to decrease the consequences of further emergent epidemics by overcoming the current limitations (Sarumi, 2021). Consequently, this research seeks to establish various ways in which big data can be used to improve the functionality of disease outbreak prediction models and help in developing more effective global health security.

Aims and Objectives

Aim

The aim of the study is to assess the capability of big data analytics in the early diagnosis of an Epidemic and designing a framework that incorporates big data analytics into the health monitoring system.

Objectives

1. To explore the role of big data analytics in forecasting disease outbreaks and epidemic trends.
2. To identify key data sources and methodologies used in epidemic forecasting.
3. To evaluate the effectiveness of big data-driven forecasting models in predicting epidemic outbreaks.
4. To investigate the ethical and practical challenges of integrating big data analytics into disease forecasting and management.

Significance of the Study

In this study, the following significant contributions will be made:

Contribution to Knowledge: It will provide the knowledge of how big data is implemented in disease prediction and can contribute to the discovery of new approaches and methods for the improvement of epidemic prediction models. Through examining the utilization of data from various sources for purpose of forecasting, this study will therefore extend knowledge in public health informatics (Grybauskas, Pilinkienė and Stundžienė, 2021).

Contribution to Clinical Practice: Proper forecasting of diseases will enable adequate preparations, resource deployment, and care management within clinical environments. Finally, the findings of this study will enable health care workers to effectively plan for disease outbreaks in order not to overload the system.

Contribution to Research: This study would advance and enrich the current literature concerning the dynamism of the ethicality of big data in healthcare. Dispelling issues like privacy, correctness of data and involvement of active stakeholders in the process, the study will serve as a base for further research on ethical application of big data in epidemiological and public health fields.

Methodology

The method also includes the use of primary data collection, and this is through administering questionnaires on an online platform. The target respondents are public health practitioners, epidemiologists, data analysts and even healthcare policymakers who have prior knowledge about analytics in healthcare (Corsi et al., 2020). The participants would be recruited from with culture and job diversity in Germany only. Descriptive analysis and frequency analysis, as well as simple qualitative analysis in the form of thematic content analysis, shall be used in an attempt to analyse the data obtained. These tools were considered fit for use since they helped in managing survey responses and analysing them within the given time and research scope.

Some limitations of the study are in the reporting of the results, which may lean towards a societal bias, lack of organisational validity due to small sample size, and accuracy of the responses received by the research, since the information collected is reliable on the participants' responses. Besides, the study aims to reveal the perception of big data use rather than implementing the models by themselves, and this can be discussed in further research (Corsi et al., 2020).

Synopsis of the Chapters

Introduction

This chapter is devoted to the development of the research topic, including the explanation of its significance, purpose, research goals, and objectives. It also describes why this topic qualifies as an appropriate area to study and gives an overview of the design and layout of the research.

Chapter 1: Literature Review 1

The first literature review chapter investigates the theoretical foundations of big data analytics and its relevance in the healthcare domain. It outlines ideas of big data, its origins and the history of big data technologies within the sphere of public health.

Chapter 2: Literature Review Part 2

The second literature review chapter, therefore, narrows down its discussion on the topic of case studies of big data in disease outbreak prediction. It evaluates the model applied in previous outbreaks, examines the associated success factors, and explores other issues in the literature.

Chapter 3: Methodology

This chapter describes the research design, the particular data collection tool to be used, the sampling technique adopted, the data analysis technique, the consideration of ethical issues in the study and the research limitations. It explains why the survey method will be adopted in collecting data and how the data collected will be handled.

Chapter 4: Results and Analysis

This chapter focuses on reporting the findings of the survey in an organised manner in the form of tables, figures and narratives. As with statistical summaries and general patterns, specific value relationships are also emphasized.

Chapter 5: Discussion

The discussion chapter relates the research discoveries from the survey to the literature. It discusses the findings in light of public health and data science, concerning the consequences, drawbacks, and opportunities of the research.

Summary

This chapter, therefore, synthesises the study by restating the findings made with regard to the research questions. They provide guidance on the method implementation to develop predictive epidemiology to public health authorities and data scientists and highlight future research opportunities for the proposed approach.

CHAPTER ONE – LITERATURE REVIEW I

Introduction

This section looks at outbreaks of diseases, common surveillance systems, data from large datasets and the part played by AI and machine learning in epidemiology. It describes how outbreak detection has improved over time, what limitations traditional techniques face and how new technology supports more accurate disease prediction and faster public health reactions.

1.1 Understanding Disease Outbreaks

According to Ristaino et al. (2021), Epidemics have influenced the course of history, people's health care, and the field of medicine throughout history. Familiarising oneself with these terms and their background is crucial to correctly understand and approach elements of modern theory and practice, for instance, big data analytics of these events. In contrast, Zaçe et al. (2021) stated that the outbreak is usually characterised by many cases of disease that occur above a normal number in a particular population, geographical location or time. Such a disease, when it infects many people in different areas or states and within a short duration, may be termed as an epidemic or pandemic, according to the situation.

According to Ristaino et al. (2021), Epidemics may be described as the outbreak of diseases or an occurrence of cases of diseases in a given population or geographical area that exceeds the normal level expected in that area. This is relative since it is hard to define what is considered epidemic in a certain region and not in another. For instance, increased cases of malaria within sub-Saharan Africa in the rainy season are normal, while increased instances of the same in the Northern Europe region will cause an epidemic alarm. In contrast, Zaçe et al. (2021) stated that Epidemics are localised diseases which spread within a country or region, but pandemics are widespread diseases that extend over international borders and impact a large population. According to WHO, the pandemic means the distribution in the world and also the transmission from man to man and the involvement of society. The last pandemic was COVID-19, which began in late 2019 and rapidly affected almost all countries, resulting in unprecedented economic, social and health disruptions.

According to Zaçe et al. (2021), throughout the years, disease outbreaks have brought significant change or otherwise affected societies in various ways. The post is covering such topics as historical facts, diseases, and pandemics: the Black Death (bubonic plague), one of the most catastrophic pestilences that took place in the 14th century, claimed the lives of 75–200 million people in Eurasia. It was transmitted through trade caravans, and due to bad personal hygiene, this shows some of the first observed behavioural means that which people passed diseases. In contrast, according to Ristaino et al. (2021) Spanish Flu of 1918, which affected around one-third of the global population and resulted in more than fifty million deaths. Here, it was most severe in healthy persons in their productive prime—a departure from the usual flu epidemiology. The flu pandemics that occurred after World War can be attributed to the mobility of the troops and the existing and inadequate medical facilities. The given historical examples serve to illustrate how mobility, population density, and medical facilities determine the dynamics of an outbreak.

Outbreaks have become frequent and complicated in the 21st century by globalisation factors, climate change, urbanisation and antibiotic resistance. The Stringer is entirely correct in stating that the SARS outbreak of 2002-2003 was the first pandemic of the Information Age. It started in China, and it affected more than thirty countries before it was controlled. SARS also emphasised the need for international cooperation or synchronisation, and monitoring in real-time, all of which remained fairly weak at that time. In contrast, Zaçe et al. (2021) stated that in early 2009, the H1N1 influenza became a pandemic but proved not as deadly as expected; this underlines the specifics of emerging diseases. With the continuation of the death and chaos, the outbreak in West Africa of 2014–2016 exposed the fragility of public health systems, particularly in regions that lack adequate resources and the impact of false narratives on the response to contain the disease.

According to Zaçe et al. (2021), A more recent and contemporary example of the effectiveness as well as the weakness of contemporary epidemiological surveillance was reflected in the COVID-19 outbreak. The issue remained the same, and early responses were slow due to politics, miscommunication, and a lack of proper assimilation of information, even with the emergence of newer knowledge. However, the pandemic also demonstrated how valuable tools such as digital technologies and big data are in monitoring the progression. The use of real-time dashboards, availability of mobility data from smartphones, and social media feeds turned into a pivotal force in decision-making at the policy maker and individual levels.

Knowledge of these historical as well as conceptual trends is important for today and tomorrow's disease forecasting. They state that although pathogens differ, the dynamics of epidemic diffusion are similar and pertain to human interactions, the environment or the state of preparedness of health facilities. According to Ristaino et al. (2021), the use of big data analytics is a good chance to discover essential signals that might be unnoticed with more frequent and profound analysis of the past outbreaks.

1.2 Traditional Disease Surveillance Systems

Morris et al. (2021) in their publication defined Conventional systems for the identification of diseases to be the bedrock of approaches towards disease tracking and response. These are commonly developed by the world health organisation, the centres for disease control and prevention and various health departments around the world, which may entails the systematic collection and evaluation of health information. In contrast, MacAulay et al. (2022), they are used to diagnose the outbreak, study its spread, evaluate the containment and intervention measures, and to guide the policymakers. These systems have applicability to many contexts, but they also include deep flaws that are evident in new threats to global health.

As MacAulay et al. (2022) posit, other conventional surveys rely on health care facilities, laboratories, or departments of public health. These actors submit report of notifiable diseases through official channel frequently, usually weekly or monthly depending on diagnosis or clinical observations. For instance, the WHO has an implemented Integrated Disease Surveillance and Response (IDSR) in African nations while the CDC of the United States of America is called the National Notifiable Diseases Surveillance System (NNDSS). Morris et al. (2021) establish that these systems provide harmonized and stable data that eases the government from spending much time and inclined on data collection besides coming up with trends, stating disease burden as well as funds allocation.

As stated by Morris et al. (2021), one of the key strengths of a traditional surveillance system is that the process is strictly scientific. The data collected is usually standardized and is likely to pass through a health expert and derived from various settings including hospitals or laboratories and mortality registers. This makes gathered data more trustworthy for long-term researches and in order to create the policies. On the other hand, MacAulay et al. (2022) noted that these systems have legal-institutional legitimacy they must consequently follow it obliging the entities to report in those systems to be core accountability. They are integrated within the national health systems and were used in health crises to produce an effective response, in cases when signals are raised, demanding actions such as isolation, vaccine or quarantine.

According to MacAulay et al. (2022), this is due to the high degree of integration of these systems in the global space. Coordinates of surveillance networks such as the WHO's Global Outbreak Alert and Response Network (GOARN) exist through which countries can exchange knowledge and resources in instances of cross-border health threats. For instance, during the H1N1 influenza in 2009, information from national systems provided the WHO with the information it required to declare a public health emergency and release recommendations. Likewise, routine influenza surveillance contributes to the development of seasonal flu vaccines used around the world (Morris et al., 2021).

According to MacAulay et al. (2022), much the same, traditional identification systems also offer their disadvantages, particularly the speed, coverage, and flexibility. A limitation that is equally fatal in every aspect is the delays that characterise disease occurrence and data reporting. As opposed to Morris et al. (2021), due to the reliance on confirmed clinical diagnoses and lab results of these systems, there might be a significant delay in data ranging from days to weeks. This complicates the early diagnosis of the disease, especially when dealing with a new material pathogen or when the disease being transmitted has hardly noticeable symptoms.

Another limitation is underreporting. Low capacity in reporting; in low and middle-income countries, there are poor facility, shortcomings in laboratory facilities, and inadequate facilities, which leads to delayed reporting or even incomplete results. MacAulay et al. (2022) argued that only some of these cases are reported because of large administrative load, patient anonymity, or variation in reporting between countries with high average income. This may lead to an underestimation of diseases, especially regarding prevalence, and therefore inadequate or misplaced efforts in public health.

According to Morris et al. (2021), Traditional systems are also not designed to analyse new forms of non-traditional data that might serve as early warning indicators. For example, keywords, postings on social media and people's movements are likely to show emerging health issues before they reach the hospital. Nonetheless, such types of data are not suitable for incorporation in legacy systems since they are not supported by most of the systems. These different systems reflect rigidity due to bureaucratic and paper-based reporting that slows responsiveness in new and emerging health threats.

In addition, the lack of integration of a centralised database between agencies or across regions can slow the assessment of an outbreak's kinetic pattern. This problem arises due to decentralised health governance, giving rise to scattered databases, variant report structures, and weak links between offices. This was seen when various countries and even states lacked the data sharing which is vital in the fight against a pandemic for proper coordination, as was the case early in the COVID-19 pandemic (Morris et al., 2021).

Based on these strengths and weaknesses, it can be concluded that, as crucial as traditional surveillance systems are to public health, they represent a starting point that is insufficient at present. Increased vulnerability to rapidly spreading transnational infectious diseases requires real-time, open-source, and predictive solutions. In contrast, according to Morris et al. (2021), the use of big data analytics when incorporated in traditional frameworks is a way forward for them. For such integration to take place, it needs to be grounded on the best practices as offered by system-based approaches.

1.3 Big Data Sources for Disease Prediction

Azmi et al. (2022) stated that Epidemiology is an academic discipline that has been characterised by the sudden arrival of big data as a relatively recent phenomenon that has revolutionised this field and provided new ways of analysing disease trends. Traditional data sources have some benefits, but most of the time, they are restricted by their small coverage, latency and rigid features. In addition, big data stores a lot of real-time information that may indicate a possible disease outbreak. Electronic Health Records (EHR), different types of social media, transport and location information, data on climate and weather conditions and Apple HealthKit data are among the most used sources for predicting diseases. All these information sources assist in enhancing situational awareness and expected cognition in public health (Ahmed et al., 2021).

EHRs are one of the main structures used in big data healthcare studies, according to Ahmed et al. (2021). They have data on patients' socio-demographic characteristics, diagnosis, laboratory data, prescriptions and clinical interactions. When combined across multiple healthcare organisations, EHRS may help in detecting symptom anomalies, higher admission rates or new types of diseases. For example, increased cases of respiratory illnesses in a particular community could indicate the start of flu season or a rise in COVID-19 cases. Azmi et al. (2022) stated that, in addition, EHRs help track patients longitudinally, which is necessary for assessment of disease and treatment patterns. However, Big Data problems like Data exchange or integration between multiple systems, privacy laws, and delayed input still limit their effectiveness in real-time outbreak prediction.

Ahmed et al. (2021) stated that, Social Media Data is wider from the population point of view and more real-time as compared to traditional surveys. Social media such as Twitter and Facebook are traditionally used to predict diseases or as barometers that can pick up early signs of disease through users' posts on their state of health, health complications or new disease regions. It is also important to use text mining and sentiment analysis to uncover clusters of symptoms that associate with emergent diseases or shifts in behaviours. According to Azmi et al. (2022), Social media participation increased in certain regions before cases were observed to escalate in the early phases of COVID-19. This was seen as an indie example of how digital epidemiology can be used alongside traditional surveillance. However, the type of data is often raw, scattered, and can contain misinformation, thus posing a need for filtering and ethical use of public data.

Mobility data drawn from phone, transportation and GPS-based applications guides understanding of human mobility, which is a core characteristic that drives the spread of diseases. Application-generated mobility reports of Google and Apple in the backdrop of the COVID-19 outbreak informed about the trends of movement restriction due to lockdown. According to Ahmed et al. (2021), these forms of movement patterns can help public health authorities to conduct an assessment of how infectious diseases could spread across regions, and the likelihood of importation into the new zones. Mobility information is quite convenient for incorporating into vectors such as malaria or dengue, to explore increased exposure in high-risk areas. Nonetheless, three key concerns that include privacy, representation, and ownership of data should be well handled.

Climate and Environmental Data also apply in the projections of the Occurrence of diseases, such as diseases that are sensitive to the climate or other ecological concerns. Temperature, rainfall, humidity and air quality determine the survival and dispersion of several parasite or disease-carrying organisms. For example, diseases like cholera, Zika, and dengue are known to be precipitated by conditions like rainfall and temperature changes. Satellite images and data from remote sensing help in real-time tracking of these conditions, allowing outbreak modelling (Ahmed et al., 2021). Incorporating climate data into surveillance systems helps in identifying early warning systems for diseases sensitive to climate change, but due to the intricate structure of ecosystems, usually advanced computational models are required for effective interpretation.

Azmi et al. (2022) stated that, Smartwatches and fitness trackers are some of the Wearable devices that are currently used to capture health data in the form of Heart rates, Body temperatures, oxygen levels, and Sleep patterns. When pooled and depersonalised, physiological data can indicate trends deviant from the population's health standards, and therefore can mark the onset of an infection before actual symptoms occur. It is then contradicted by Ahmed et al. (2021) that newer smart wearable device producers, such as Fitbit, have been able to notice changes in resting pulse during flu seasons about flu activity in the zones. While being a valuable resource, this data source is not harnessed to its full potential in the field of public health, as certain preconceptions, such as access, standardisation, and user consent, remain a concern.

1.4 Role of Artificial Intelligence (AI) and Machine Learning (ML) in Epidemiology

According to Hamilton et al. (2021), Artificial Intelligence (AI) and Machine Learning (ML) *h*ave introduced intelligent and powerful computational tools for data analysis in the field of epidemiology to identify trends and forecasts for diseases and outbreaks. The core difference between them is that AI and ML do not need prior assumptions and utilise linear models for pattern recognition. Chis would present more precise, immediate, and effective equipment to monitor as well as control the future spread of epidemics.

Chen et al. (2024) stated that, outbreak prediction is one of the primary use cases of AI in epidemiology. Some of the most popular algorithms are the decision tree, support vector machines, neural networks, and ensemble methods, which can be trained on historical cases of an outbreak in combination with other features like population density, mobility, climate, and social media. Such models can detect connections and the influence of certain variables that cannot be seen if using a rather straightforward analysis. For instance, systems such as BlueDot and HealthMap predicted the signals of COVID-19 even before an official alert using news semantic analytics, ticket bookings of airlines, and announcements in various languages. A system of early detection is instrumental in the activation of a fast response from the side of public health side.

AI is also used for the categorisation and grouping of diseases and their cases. It presents another capacity of algorithms to distinguish different diseases depending on symptoms, laboratory coefficients, and spread information. It is contradicted by according to Hamilton et al. (2021), it is particularly concerning if the symptoms of different diseases partially overlap, such as influenza, COVID-19, and other respiratory diseases. Classification models play a critical role in diagnosing patients and prioritising them for resource allocation based on this risk factor. NLP, which is a part of artificial intelligence, helps to identify insights from unstructured clinical notes and reports to intensify the surveillance (Chen et al., 2024).

According to Hamilton et al. (2021), one of the most notable applications of machine learning is in the modelling of diseases and their likely progression patterns. Recurrent neural networks and time-series models help predict new cases, hospitalisations, or death rates based on current trends. Chen et al. (2024) contradicted that these models assist in healthcare planning to estimate the requirements, such as the number of ICU beds, ventilators, and staff, among others. Machine learning has also been applied to estimate the effects of various principles of the population's activities, more related to abstinence or related to vaccination programs, regarding changes in the dynamics of the outbreak.

According to Chen et al. (2024), AI helps in performing syndromic surveillance, which entails identifying early illness signs in the population through data such as sales of over-the-counter drugs, emergency room visits, and web searches. These systems do not require a clinically evaluated diagnosis, and they can act as predictors in cases when the ability to perform a lab workup is not available or is performed later. In contrast, according to Hamilton et al. (2021), despite these advances, challenges remain. AI can then be hindered by the quality of data, biased data and also a lack of clear understanding of how decisions were made. Another problem is known as the black box problem. Privacy issues and specifically potential violation of patient data or real-time location data must also be safeguarded with proper ethical structures.

Conclusion

Big data analytics and AI help make epidemic prediction and control much more effective. Connecting these tools to standard systems improves early discovery, better reactions and smarter decisions on resources, while keeping ethics and quality data top priorities is necessary for successful public health.

CHAPTER TWO – LITERATURE REVIEW II

Introduction

It analyzes different predictive modelling approaches, challenges linked to ethics, difficulties in using big data and what stakeholders think about health surveillance. The article points out the need to improve how different types of data are handled and to focus on trust, transparency and cooperation to help in predicting outbreaks and managing public health.

2.1 Predictive Modelling Techniques in Epidemiological Forecasting

According to Martin-Moreno et al. (2022), the process of Predictive modelling is an essential element of contemporary epidemiological prediction, which helps authorities in charge of public health to estimate further developments of disease spread, to mobilize resources and provide necessary measures. With the availability of large scale availability of health data, many methods of prediction have been used starting from statistical methods to modern machine learning algorithms. Several approaches exist in the task of modelling and they include regression analysis, time series analysis, neural networks and hybrid systems. All the techniques has certain advantages and is used depending on the data characteristics and the goals of the planned public health interventions.

In their study, Gupta, Pandey, and Pal (2021) described regression models as some of the most popular analytical tools used in epidemiological research due to their effectiveness and easy interpretation. Linear regression models can make predictions about a certain outcome variable or dependent variable (for example number of disease circumstances) depending on one or several predictor variables or independent variables (for instance temperature, population density, or mobility). In contrast, Martin-Moreno et al. (2022) explain that Logistic regression, which is often used for binary outcome analysis, is sometimes used to determine the likelihood of disease occurrence based on factors such as age, vaccination timetable, and presence or history of underlying medical conditions. For instance, during the phases of influenza cycles, it is possible to employ logistic regression to define which groups are more endangered by infections and hospitalization. Nonetheless, it is worth mentioning the applicability of regression models is restrained by the assumptions like linear relationship and absence of correlation between variables, which is not always true in the case of disease transmission and especially in its dynamic progression.

This is especially useful where modelling of an event or an occurrence over time is necessary, time-series analysis is very essential in epidemiological forecasting. In using temporal data, the auto regressive integrated moving average (ARIMA), and Seasonal ARIMA (SARIMA) models can be used to model the past data and make prediction for future data. These models are useful in determining the periodicity and seasonality of disease like flu or dengue fever in which the disease reoccur on cyclic basis. As pointed out by Gupta, Pandey and Pal (2021), Time-series models are also used to evaluate the effect of control measures such as lockdowns, vaccination drives by comparing predicted values to actual number post the intervention. However, such models can be subjected to a noisy process and may encounter challenges when applied to a new structure such as a virus.

Neural network especially deep learning model have gained traction in the recent past due to their capability to learn even nonlinear transformations of the data without prior assumptions about the distribution of the data. RNNs and LSTMs are especially applicable in sequence analysis, therefore it is possible to apply them in predicting epidemic curves in time. For instance, in the period of COVID-19, LSTM models were used for day-by-day case projection, admission rate, and mortality rate predictions. These models operate by identifying patterns from the recorded data and then use those patterns to suggest the behaviour that should be expected in the future, and in general, the more data provided the better these models will be. Nevertheless, they mainly worked in an ‘opaque’ manner, making only a few interpretable, which poses a problem regarding their adoption particularly in influential health policy domains (Gupta, Pandey and Pal, 2021).

Hybrid systems interconnect one or more models so as to improve precision and flexibility. These models can for example include the combination of time series forecasting and a machine learning algorithm or the combination of outputs from a regression analysis and a neural network classification. Martin-Moreno et al. (2022) contradicted that, Hybrid models can be beneficial since it makes of the high points of each constituent while covering their weaknesses. For instance, it is possible to combine an autoregressive integrated moving average model for the temporal dependencies in the series while incorporating an artificial neural network for nonlinear dependency, like behaviour shift or climate change. It is also the practice of applying these approaches, which can lead to more efficient and flexible forecasting systems suitable to the new trends of the outbreaks.

However, these models are also judged on their level of sensitivity, specificity and accuracy. Another way in which the reliability of the model can be checked is by using cross-validation and back-tested to outbreak data. In addition to demonstrating when and to what extent the disease is likely to surge, they also allow for creating specific plans on different routes of development depending on the measures that are taken by the authorities (Gupta, Pandey and Pal, 2021).

According to Martin-Moreno et al. (2022), it is crucial in using predictive modelling strategies for improving the preparedness and preparedness of various health systems. On the one hand, known techniques like regression and time-series analysis are more transparent and easier to implement, while the more modern solutions, like neuron networks and combined ones, enable richer ways of computations and are less sensitive to noises. As pointed out by Gupta, Pandey, and Pal (2021) while advancing that the choice of model depends on data characteristics, the disease under consideration and the purpose of the study. With the developing technology and increasing volumes of data, the use of the predictive modelling will be even more important in addressing future public health crises.

2.2 Ethical, Legal, and Data Privacy Concerns in Health Data Analytics

According to Rajasegar, Gouthaman and Ponnusamy (2024), the integration of big data analytics systems in health and especially epidemiology discloses the great probability of the occurrence of diseases in the population and helps in providing an effective solution towards a healthier population. Althobaiti (2021) argued that it also gives rise to several ethical, legal and data privacy issues, most keenly accentuated since health-related information is sensitive and personal. The matter of data protection, consent and possible abuse if certain rights need to be respected, and sustainable when people are to be impacted by certain technological advancements.

According to Althobaiti (2021), the right to privacy is one of the major ethical issues when it comes to data collection and sharing. Health data can be described as private since it usually pertains to personal information linked with geographic location and behaviour. Rajasegar, Gouthaman and Ponnusamy (2024) argued that Electronic health records, mobility tracking, wearables, and social media – all include personal data to predict disease outcomes, and personal data, when misused, can hurt. The risks of such data include theft, being accessed by people who should not, or using the data for commercial or discriminatory purposes. For example, if a person has been evaluated to be at high risk of getting an infectious disease, such information, if disclosed to the public or used improperly, may lead to prejudice, discrimination, or even higher insurance costs.

The other important ethical consideration is informed consent. The values of many big data sources, including GPS or posts on social media, do not come from a healthcare setting. They may, therefore, not know that their data can be used for health surveillance needs. This former type of leakage poses a challenge to anonymity since even if data is anonymised, number matching might match the numbers in other datasets with the data collected. According to Rajasegar, Gouthaman and Ponnusamy (2024), awareness of the data gathered requires that people should understand the collected data, the use of the data, the availability of the data to other people and the consequences of using the data. There is often pressure to infringe on consent whenever there is an outbreak of diseases; nonetheless, this can open the floodgates for future encroachments on people's rights (Althobaiti, 2021).

According to Rajasegar, Gouthaman and Ponnusamy (2024), from a legal perspective, the GDPR in the EU and the HIPAA in the United States provide specific measures on the collection, processing, and sharing of health information. For instance, GDPR has provisions including data minimisation, limitation of purpose, and rights given to people to access and erase their data. High-risk data processing activities also must be carried out by the data controllers after the organisation has made an assessment. Althobaiti (2021) argued that, these laws are designed to open possibilities while protecting individuals, but doing so in the field of novel, complex, often algorithmically based on big decentralised data assets, like artificial intelligence and machine learning, is still problematic.

It emerged that legal and ethical risks require data governance to be managed well. It involves defining ownership, storage, sharing, and accountability for data that is collected and stored locally or by third parties. It is thus imperative for public health authorities, research institutions, and technology companies to clarify how these properties and ownerships are divided to design a clear, fair data-sharing agreement that protects people’s rights (Althobaiti, 2021). In this case, ethical review boards and data protection officers have a responsibility to monitor compliance with legal and ethical standards. However, there is emerging awareness of the concept of “data justice”, which encompasses fair usage of data that would not hurt the vulnerable groups.

According to Rajasegar, Gouthaman and Ponnusamy (2024), one key issue to address is algorithm bias and its relationship with fairness. AI learned from such data may be predisposed to discriminating between the healthy and the unhealthy, or the sick and the cured, thus perpetuating health disparities. For example, if an outbreak prediction model is dependent on information from large towns, it may offer scant representation for rural or minority groups, hence allocating resources to the wrong quarter and responding too late in areas with few or no representatives. Conversely Althobaiti (2021) stated that this means transparency in the creation of such algorithms and in making sure that they would not worsen inequalities when it comes to public health.

According to Althobaiti (2021), as big data analytics is deemed to revolutionise disease mapping and subsequent decision-making within the public health domain, such integration needs to face serious ethical and legal considerations. Privacy, accountability, adherence to the data protection laws, and fairness in the use of algorithms are not just standard requirements but are cornerstones upon which technology is developed so as not to undermine the rights and privacy of the individuals seeking the protection of the technology. With the increased use of health data in healthcare, there is a need to develop a sound ethical and legal foundation for the use of the data.

2.3 Challenges in Implementing Big Data for Health Surveillance

According to Chowdhury et al. (2024), despite the potential benefits, big data analytics of health data for surveillance and early warning of disease outbreaks is not without several challenges.

These include the technical, institutional, and infrastructural, which the study shall discuss below as the main barriers to a widespread and efficient implementation of big data tools in the field of public health. Mitigating these challenges is important to realise the benefits of real-time and predictive technologies in various areas of health, especially during outbreaks and epidemics (Rehman, Naz and Razzak, 2021).

The most pressing technical issues are, first, the challenge of data integration. Health information can be sourced or acquired in several ways, including EHRS, laboratories, wearable devices, social media, climate sensors, and mobility apps. While these sources may use different formats, standards, and terminologies for presenting the data, they cannot be merged directly into an analytical framework. It is contradicted by Chowdhury et al. (2024) that there is no ability to easily integrate data systems, and this, coupled with the absence of standard metadata, makes it difficult to consolidate and analyse data. For instance, different databases in a hospital may use a particular coding system for diagnosing conditions that is different from that used in the national surveillance databases, making the interpretations either disparate or delayed (Rehman, Naz and Razzak, 2021).

Also crucial is the issue of data quality and data reliability. Big data usually encompasses large sets of unstructured data collected in real-time, and, therefore, the information might be misleading or contain errors. Information such as customer data collected through social media networks can be easily obtained, yet it usually contains a lot of unwanted noise and cannot be considered scientifically proven. According to Chowdhury et al. (2024), the predictions influenced by erroneous information in the database rollover can contribute to false recommendations that can mislead activities aimed at public health improvement. Managing and preparing a vast data source for epidemiological use can be such a cumbersome task and is often time-consuming, especially due to the use of various tools and expertise that may at times not be easily accessible, especially in low and middle-income countries (Rehman, Naz and Razzak, 2021).

According to Rehman, Naz and Razzak (2021), there is another technical factor, computational capacity, which can also be seen or described as a technical challenge. High volume processing of large data is made possible with the help of computing infrastructure like efficient threads of servers, clouds and web-based storage systems. Most of the expanding public health agencies, especially those from developing nations, do not have access to such technologies. This makes it even more expensive and complicated to implement and maintain such systems, hence making the adoption of big data tools to routine health surveillance activities difficult (Chowdhury et al., 2024).

From an institutional outlook, one of the major barriers noted is organisational resistance to change. Many public health facilities are still operating with a traditional informal system of disease reporting and surveillance. Organisational change management involves the cultural change, training of personnel and leadership to embrace the new change and adopt the data approach (Rehman, Naz and Razzak, 2021). Health professionals may also be ignorant of completely depending on the algorithms or automated systems, this is because such systems mostly work in rather confined ways and are rather secretive. Lack of control over work, employees' privacy, and legal issues are all reasons that cause resistance to big data solutions.

According to Rehman, Naz and Razzak (2021), some institutional barriers affect the usage of social media, and these involve data governance and legal frameworks. One of the most significant challenges that can be identified is concerns related to data ownership and data sharing, and protection when it comes to the usage of Big Data for public health. Conversely, according to Chowdhury et al. (2024), National laws and regulations that protect the rights of the people to privacy, for instance, the general data protection regulation in Europe, may hinder the flow of data across borders that can be pivotal in disease surveillance. Closely related to this, issues of consent, anonymity, and data fairness need to be properly addressed, not to violate the rights of particular subjects while promoting public health interests.

According to Chowdhury et al. (2024), these challenges are further aggravated by infrastructural factors, especially in settings that are considered bearers of poor resources. Many regions still have inadequate healthcare facilities; limited access to the internet, outdated equipment and a dearth of professional experts in health informatics or data science. Conversely, Rehman, Naz and Razzak (2021) stated that, while digital tools are present, it means that intermittent electricity, a lack of technical support, and unreliable Internet connections severely challenge the continuity of big data projects. This is because the gaps in connectivity result in differences in surveillance strategies as well as hamper international cooperation in identifying and managing epidemics.

According to Rehman, Naz and Razzak (2021), the opportunities offered by big data for health surveillance are significant; however, it is hampered by a tangled array of technical, institutional, and infrastructural issues. For these challenges to be effectively addressed, it is going to take more than just investing in the digital systems, standardisation of data, training and educating the health stakeholders and formulation of legal and ethical frameworks. It is through such combined approaches that big data analytics can be incorporated into the global health systems to improve capitalisation on epidemics.

2.4 Stakeholder Perceptions and Public Trust in Data-Driven Health Monitoring

According to Kerasidou and Kerasidou (2023), the adoption and adoption of data-driven health monitoring systems depends not only on the technological factor but also on the readiness of the key stakeholders that including the healthcare professionals, public health organisations, the government, and the public. Conversely, Walvoord et al. (2022), it is crucial to comprehend these perceptions since big data analytics depends on trust, openness, and cooperation in analysing and forecasting diseases. If stakeholders do not trust the technology or consider it invasive, it has reduced value in the management of public health.

According to Walvoord et al. (2022), healthcare workers themselves are directly implicated in the creation and usage of data management systems. It has a considerable impact on the use of big data tools in clinical and epidemiologic research. In general, the clinician and the public health workers know that big data can help in disease diagnosis, resource allocation and real-time decisions. For instance, some of the enabled applications are predictive models for admissions and automated alert systems when there are unusual variations in patients' symptoms. Conversely, Kerasidou and Kerasidou (2023) stated that there are questions about the accuracy of these algorithms, the opacity of data handling, and how they might compromise clinical reasoning. While it is understood that some are reluctant to employ 'black box' like tools and models without explanation of results. It is crucial to train healthcare professionals and ensure that the development progress of such systems is made transparent to foster confidence.

Among the policy and public health authorities, data surveillance is commonly perceived through the prism of its effectiveness, scalability and capability to intervene at an early stage.

The analysis of data dashboards and mobility reports, especially, were instrumental in supporting the decisions of lockdowns, the organisation of the vaccination processes and the distribution of medical equipment and resources during the COVID-19 crisis. But these authorities are not oblivious of the political, ethical and legal ramifications of health data, especially in a democratic society. This raises fears of encroachment and privacy invasions by governments or private companies, which mars people's confidence when legal frameworks and accountability bodies are not well established and functional (Walvoord et al., 2022).

According to Kerasidou and Kerasidou (2023), the general public's trust in health data systems varies and is affected by the culture and experience with the healthcare system, and overall knowledge of the issue of privacy in the digital age. While some people are convinced that data-driven measures like contact tracing apps or apps for reporting symptoms are useful for public health, there are concerns about surveillance, data leaks, and loss of privacy. Walvoord et al. (2022) argued that this leads to suspicion because there is no clear statement on how data is gathered, anonymised, and used subsequently. For example, contact tracing apps in some states were lower as people are worried about being surveilled, or they do not trust data security.

Transparency, communication and consent must be maintained to establish public confidence. Trust in data sharing can hinge on open volition, perceived utility, and perceived control. Using campaign approaches, ethical committees, and the implementation of privacy by design will go a long way in fixing the trust dilemma between the public and the implementing institutions.

2.5 Research Gap

The need to improve disease outbreak predictions has been particularly enhanced by innovative big data analytics because research gaps still exist on the best approach for how this technology can be incorporated into public health surveillance. While previous research emphasises technological features of prognostic models or particular instances, there is little attention paid to systematic, ethical, or organisational constraints that limit the use, particularly in low-resource environments. Furthermore, research on stakeholders' perspectives is also scarce, especially concerning how trust and acceptance of implementation by healthcare professionals and the public impact it. There is also a lack of research that assesses the efficiency of combining multiple inputs which including social, mobility, and climate data, among others, into a cohesive real-time monitoring model. In addition, the ethical issues related to consent, data privacy, and bias in algorithms are usually recognised but not analysed thoroughly in practical applications. This research aims to fill these gaps by providing an intuitive understanding of the issues, perceptions from different stakeholders and governance issues associated with the use of big data in disease prediction.

Conclusion

Resolving issues related to technology, ethics, and organisations is necessary for developing big data analytics in epidemiology. Ensuring stakeholders trust the company and that their information is safe is important for the success of adoption. This research seeks to reduce existing gaps by looking at full models and governance structures to improve real-time and equal responses to diseases.

CHAPTER THREE – METHODOLOGY

The chapter outlined the methodology that underpinned the study investigating predicting disease outbreaks using big data analytics. The chapter explained the philosophical foundations, the approach and the methods used to analyse how advances in big data analytics can improve the accuracy and reliability of disease outbreak predictions. Data found from surveys and an analysis based on quantitative methods is used in this research. It gives a structured and fair perspective on how technology unfair public health prediction.

3.1 Positivism Research Philosophy

The research used a positivism research methodology. According to positivism, reality can be ascertained through the collection and interpretation of measurable, observable data. The use of a positivist method was pertinent as the aim of the study was to measure and analyse the links between variables such as the adoption of big data technologies and the effectiveness of disease outbreak forecasting (Ali, 2024).

Data garnered from health informatics and analytics professionals formed the basis for the results. Using a positivist methodology resulted in more reliable and definitive findings that were open to wider evaluation and application (Ali, 2024). It allowed the collection of reliable data through the use of surveys, analysis and modelling. Using positivist philosophy, researchers made their work more reliable and the findings more credible.

3.2 Deductive Research Approach

A deductive approach was adopted to examine whether applied theories held for real-world situations. Researchers initiated this study by examining relevant scholarship in the areas of big data analytics, epidemiology and health informatics to develop guiding principles and hypotheses to be tested (Hall, Hall and Shaw, 2022). These included assumptions such as: Employing real-time data analytics shortens the time it takes to respond to disease outbreaks, and gathering greater varieties and amounts of information results in improved predictions of disease spread.

In light of these theoretical ideas, specific hypotheses were derived for analysis using survey data. The researchers were then able to translate theoretical ideas into clearly stated hypotheses capable of empirical testing. For example, the study examined how the adoption of AI-powered predictive models in big data health analytics has influenced the speed with which outbreaks like dengue, malaria or COVID-19 are identified. Using the deductive method, the investigation was able to validate (or disprove) such assumptions by rigorously analysing survey data gathered from firsthand sources (Hall, Hall and Shaw, 2022).

3.3 Explanatory Research Design

An explanatory research design was nominated for this study to inspect how and why using big data analytics led to advances in disease outbreak forecast. Explanatory design was the most appropriate approach since the study aimed to found both the degree to which big data enhanced prediction accurateness and the underlying mechanisms involved in producing this improved performance (Bentouhami, Casas and Weyler, 2021).

With the explanatory design, the study was able to evaluate factors such as the makeup and size of datasets, the technologies utilised, such as Apache Hadoop and Spark and how often models were efficient. This study investigated the role of big data in predicting diseases (Bentouhami, Casas and Weyler, 2021). Experts analysed if using these devices could result in quicker identification of outbreaks and official responses from health experts.

Here, the study analysed the ways different elements interacted rather than just focusing on one single factor. It was studied how the progress of machine learning experts relates to the access to technology and how it changes the accuracy of predictions made with big data in public health. Because of these factors, the study can see clearly how different aspects contribute to the use of big data analytics in public health (Bentouhami, Casas and Weyler, 2021).

3.4 Sampling

People were chosen for the study using a purposive sampling technique. Participants in this non-probability sample were selected based on their expertise and work experience in big data applied to healthcare or public health (Omair, 2020).

Respondents were included only if they had at least one year of work experience with disease prediction or healthcare analytics tools. Participants should have had a role that allowed them to be actively engaged with statistics-driven applications (Omair, 2020). These experts came from different areas such as public health departments, global health groups, hospitals and companies providing health data services.

Of the 60 professionals who were reached, provided us with their responses. Because a lot of people took part in the survey, the study could be confident that our numbers were correct, and the careful selection ensured our data was informative. The study included 60 people purposively sampled and had at least one year of working in healthcare analytics. The review concentrated on the details and relevance because the experts were made to share their knowledge, ensuring that experts shared their knowledge and thus not to be audience wide. Although the sample was not random and so generalisability was limited, the research opted to obtain further information from a small group of very knowledgeable individuals rather than obtaining data from many.

3.5 Data Collection

The data analytics and public health analysis were done with data that was sourced from United Kingdom area. The event brought together people from organizations including public health departments, research groups, hospitals and companies involved in healthcare data. Since they possessed the training on forecasting and data technology, these professionals were in the best position to study outbreak prediction technologies in the UK.

Data Analytics and Public Health Specialists were contacted through a survey online for this study. In an attempt to collect people's data on the interface between data analytics and public health, the survey was conducted. The survey responses were collected from professionals who had experience related to data science, applying data analytics in epidemiology, data analytics in the healthcare IT and healthcare policy (Mazhar et al., 2021).

In the online survey, participants were asked to reflect upon several dissimilar things. Contributors in the survey were asked to share facts about themselves, their use of data, technology predilections, ability to predict and what they supposed were their main difficulties (Mazhar et al., 2021). The participants exposed their age, years working, job duties and the type of organisation they were part of. People shared the types of data they observed the most and how steadily they used precise datasets.

Study participants reported the data tools and platforms they depended on, together with details about their perceived trustworthiness. The study took into account how quickly responses were given, how correct predictions were, and the extent to which AI was being used by companies. The study explored the broad range of problems identified by respondents, which included concerns about data accuracy, compatibility and costs (Mazhar et al., 2021).

The study posted the survey on LinkedIn, notified members of related mailing lists and sent it directly to healthcare organisations by email. The study refined the survey by incorporating recommendations received during the trial period. The study received 60 completed surveys for response rate.

3.6 Validity and Reliability of Data

It was very important to verify that the data used was valid and reliable during the research process. Validity shows if the survey measured the right topics, and reliability checks if the answers are similar when the survey is taken more than once (Coleman, 2022).

Content validity was improved by creating the survey based on literature and getting expert advice from academics in the field of public health analytics. The questions used included messages that matched the research objectives (Coleman, 2022). Questions were grouped logically and made sure they measured what they were supposed to, such as types of data, tools used for analysis or expected times for predictions.

To measure internal consistency, Cronbach's Alpha was applied to multiple items in the same construct. A coefficient of 0.83 was calculated, suggesting that the findings were reliable (Coleman, 2022). Although the current study gives a Cronbach's Alpha value of 0.83, every multi-item component—for example, data usage, prediction accuracy and tool reliability—needs to be checked for reliability. The values of Cronbach's Alpha ought to be mentioned separately for every scale which makes the analysis more reliable. The survey process was stabilised and made reliable by maintaining equal conditions throughout. All individuals taking part in the study were provided with the same digital survey through a secure link and were told when to submit it. Participation was made easy because the survey could be accessed from multiple devices.

3.7 Ethical Considerations

Ensuring ethical compliance was a main principle in conducting this research. Because the study included humans, every effort was made to ensure research was handled appropriately at all times. They were designed according to the standard procedures and ethical rules of the British Educational Research Association (BERA) (Hasan, Rana and Chowdhury, 2025).

It was up to them to choose whether to take part in the survey. All individuals taking part in the survey had the chance to read the informed consent form explaining the aims, their rights and the questions in the study. No one else could participate unless an agreement was made ahead of time. Anonymity and confidentiality were guaranteed. When filling out the survey, there were no questions about people's names, emails or their job positions (Hasan, Rana and Chowdhury, 2025). Specially protected files were created so that the principal researcher could have access, but others could not. The data for reporting was combined to make sure nobody's identity was revealed.

When working with data, the study strictly followed regulations such as the General Data Protection Regulation (GDPR). In supporting non-maleficence, the research refrained from questioning anybody who might be pleased or anything that could include the institution. Any part of the questionnaire that seemed to be too personal or questionable was uninvolved. The study subjects were guaranteed that the information gathered would not be shared with industries and would be used only for scientific study. At all times, they made sure the participants' rights, security and dignity were preserved.

3.8 Data Analysis

Analysing the survey data was done mostly with descriptive statistics, relying on frequency distributions and graphs. These approaches helped present the main ideas of the participants' responses and show the key trends in the data (Bianconi, 2024).

The data was organised by frequency to see which types of responses were the most common. The percentage of respondents who retrieved information from structured and unstructured information sources was determined. This allowed us to draw conclusions about most popular big data systems and key challenges in forecasting of novel outbreaks. The researchers asked closed ended or multiple choice questions about types of data, predictive tools, response speed and how accurately people judge the results to study the key variables. Individual variables were given numerical values by using likert scales or counts applied to each.

Bar charts, pie charts and histograms were made for the purposes of making the findings easier to understand. Percentage of users choosing real time dashboards versus traditional batch tools was graphed by bar charts. Pie charts displayed how many respondents were from each type of health organisation, for example, public hospitals, research institutes or government departments (Bianconi, 2024). For numeric responses like the number of data streams used weekly, histograms displayed the data's distribution.

These methods helped prevent errors, made the results reproducible and kept the data statistically significant. The data revealed that there is a connection between the amount of data and how efficiently the prediction system functions (Bianconi, 2024).

3.9 Time Horizon

The study is using cross-sectional data, which refers to data gathered all at once rather than over many periods. Since the study aims to rate the effectiveness of strategies used to manage risk in healthcare project management, using this strategy is reasonable. Through this type of research, the study can gather up-to-date information about what healthcare project managers deal with in their current period. Using this tool, you can also conduct surveys that help you collect consistent answers from a range of participants in a short time.

The point of the study is to discover practical insights that can benefit ongoing and future medical projects, considering the recent pandemic and other modern difficulties. While doing a long-term study could reveal more changes in risk mitigation strategies, the time and resources available in our study required using a cross-sectional design. It facilitates efficient analysis of many variables, helping advise on better practices for managing project risks within the healthcare sector.

CHAPTER FOUR – FINDINGS / ANALYSIS / DISCUSSION

4.1 FINDINGS

4.1.1 Introduction and Participant Profile

It gives the results from a survey sent to 60 professionals from fields like public health, healthcare analytics, epidemiology and related subjects in the UK. There were 24 close-ended questions in the survey meant to explore the core topics on how big data analytics can predict the likelihood of disease outbreaks.

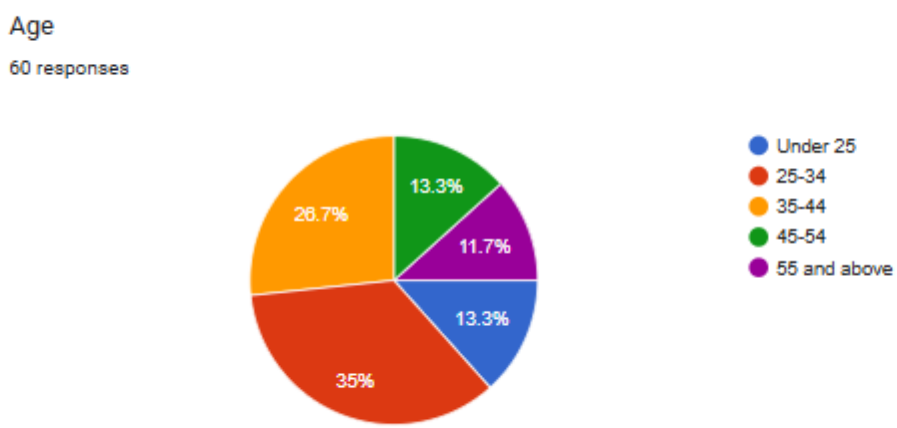

Figure 1: Age

Information on who was interviewed adds important background to the study. About one-third of people (30) in the sample were under the age of 25, nearly half (21) fell between 25 and 34, 30% (16) were 35–44, 14% (8) were 45–54 and 7% (7) were older than 55.

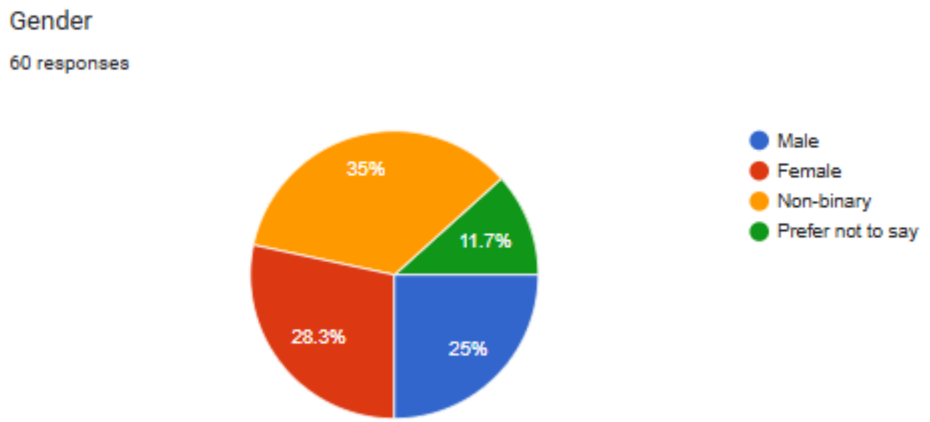

Figure 2: Gender

The survey included 15 males, 17 females, 21 non-binary individuals and 7 people who chose not to reveal their gender.

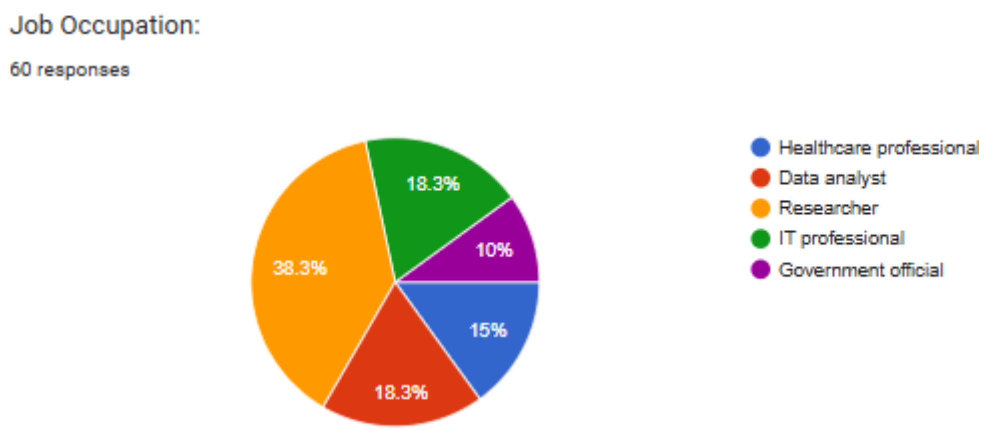

Figure 3: Job Occupation

Out of the 50 participants, the domains included researchers, data analysts, IT professionals, healthcare professionals and government officials.

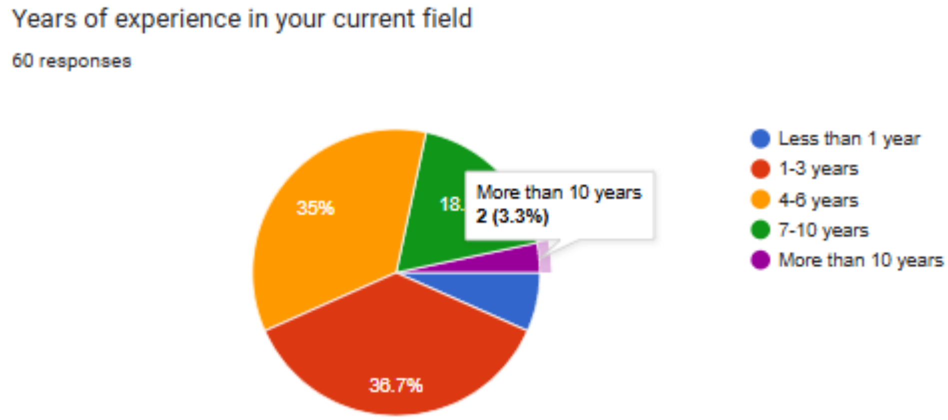

Figure 4: Professional Experience

In terms of how long they had spent in the IT sector, most gathered 1–3 years (22) or 4–6 years (21) of experience; there were fewer respondents with less than 1 year (4), 7–10 years (11) or more than 10 years (2). Because of this diversity, the outcomes from brainstorming are more helpful and helpful to many readers. This analysis is grouped into five important themes as guided by the objectives: (1) How aware and important big data analytics is seen, (2) The data and tools being used, (3) The success and problems related to forecasting epidemics, (4) Practical and ethical questions that come up during integration and (5) Everyone's general thoughts on how big data affects epidemic prediction.

4.1.2 Awareness and Perceived Importance of Big Data Analytics in Disease Outbreak Forecasting

The theme begins with looking at people's understanding of and reactions to using big data analytics in disease outbreak forecasting, aiming to achieve the first objective.

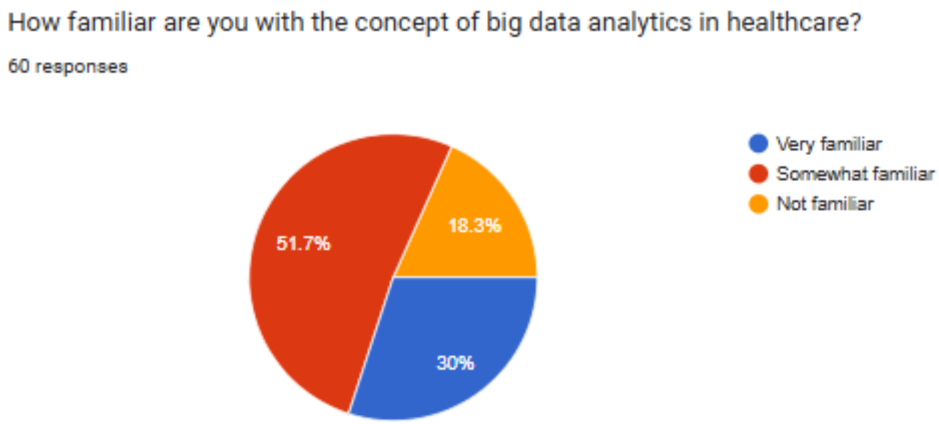

Figure 5: Familiarity with Big Data Analytics

Some participants mentioned that they were very familiar with big data in healthcare (18), some indicated they had some exposure (31) and a few were not very familiar (11). Overall, most people seem aware, but some lack specific details, perhaps because they did not have enough opportunities to learn.

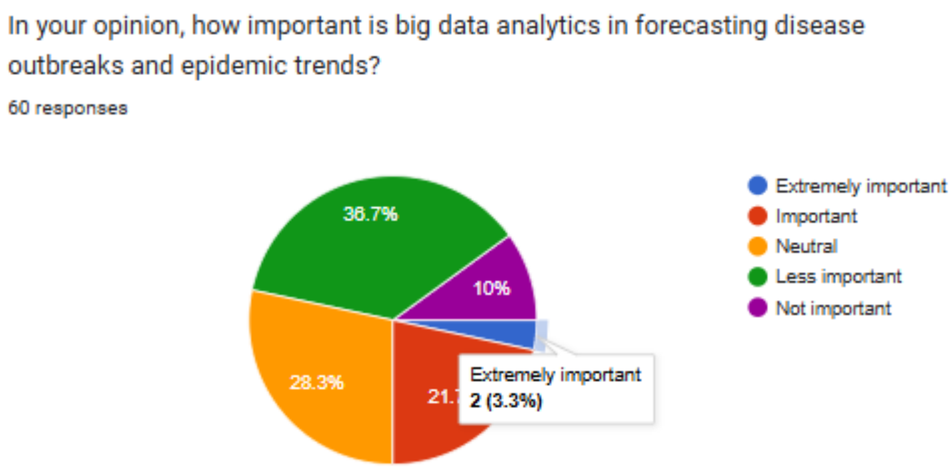

Figure 6: Perceived Importance in Forecasting

It is important to remember how important perception can be in forecasting. When it comes to the role of big data analytics in epidemic forecasting, only two people thought it was extremely important, thirteen found it important, twenty-two considered it less significant, six said it was not important and seventeen were neutral. The large group suggesting it is less crucial may be showing that they do not find it useful these days or prefer the typical ways.

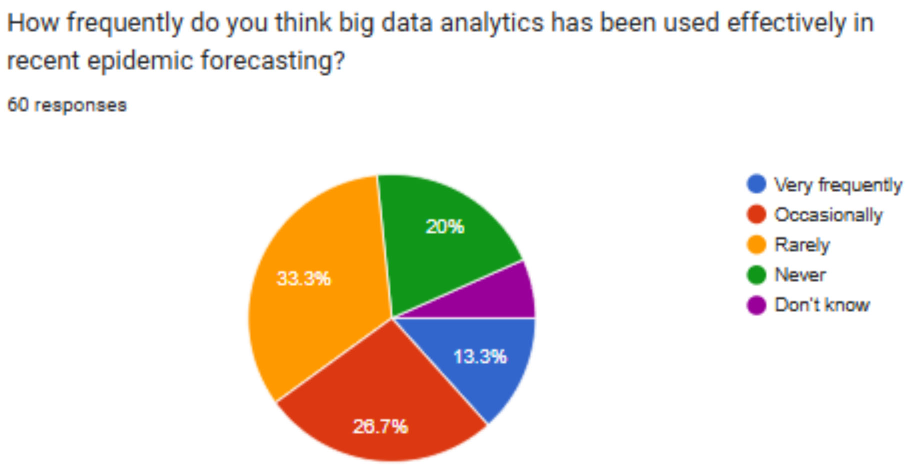

Figure 7Frequency of Effective Use

The majority of respondents thought big data had been effectively applied only occasionally. Four people did not know how to answer the question. This demonstrates that there seems to be a difference between the idea of big data and its actual adoption in public health surveillance..

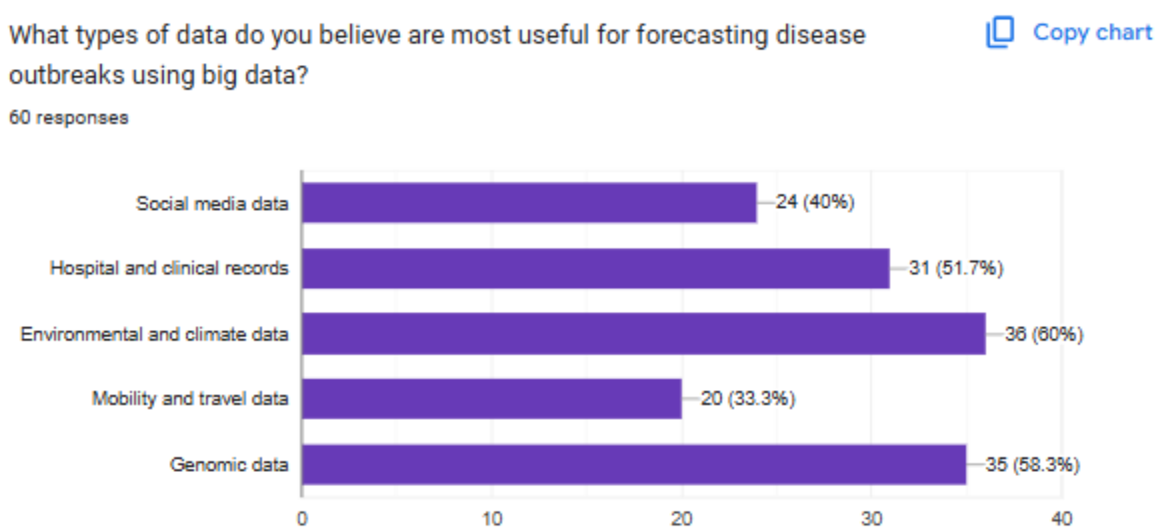

Figure 8Most Useful Data Types for Forecasting

Taking part in the research, experts concluded that environmental and climate data (36), genomic data (35) and hospital and clinical records (31) were particularly useful for predicting outbreaks. While social media and mobility/travel data were valuable, less importance was given to them.

Here, the use of established and validated data is still being prioritised, although it is now also clear that new digital types of data are becoming more important.

4.1.3 Key Data Sources and Analytical Methodologies for Epidemic Forecasting

This theme is under the second objective and focuses on the sources of information and the best methods to analyse them.

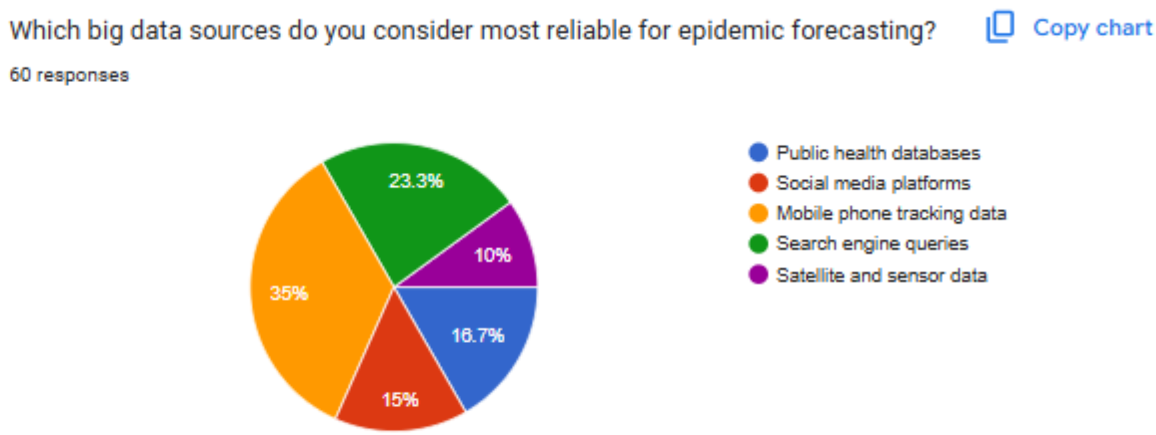

Figure 9Reliability of Big Data Sources

People said that tracking data from mobile phones was considered the most reliable big data source, with search engine queries in second place (21 and 14 responses). According to the survey, public health databases (10), social media platforms (9) and satellite and sensor data (6) were chosen less often by participants. It demonstrates a belief that mobility and searching data can detect epidemiological changes soon after they happen.

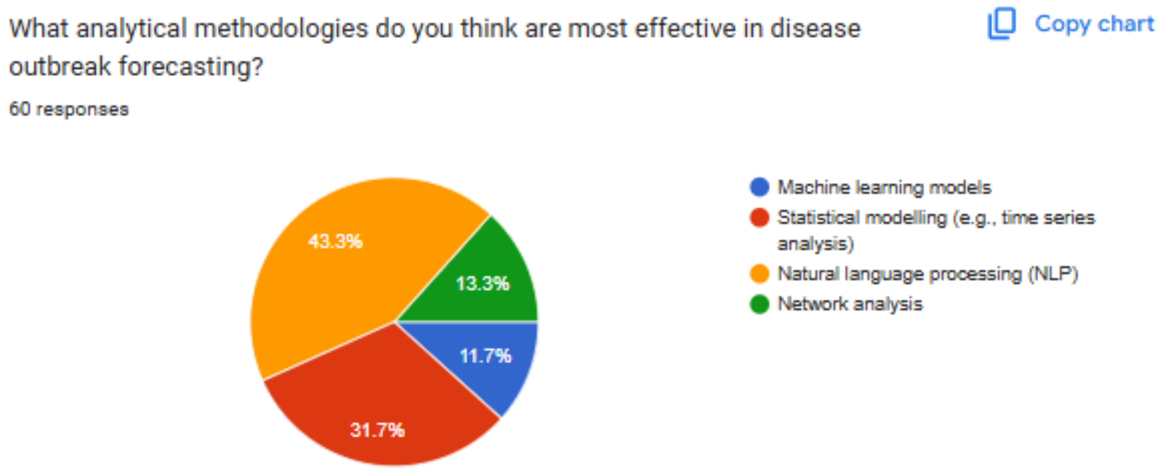

Figure 10Preferred Analytical Methodologies

Natural language processing (NLP) was cited the most (26), and statistical models based on time series analysis came next (19). The group gave less support to machine learning models and network analysis (6) than to other methods. The data show that although advanced technologies are widely recognised, classical statistics remain widely used.

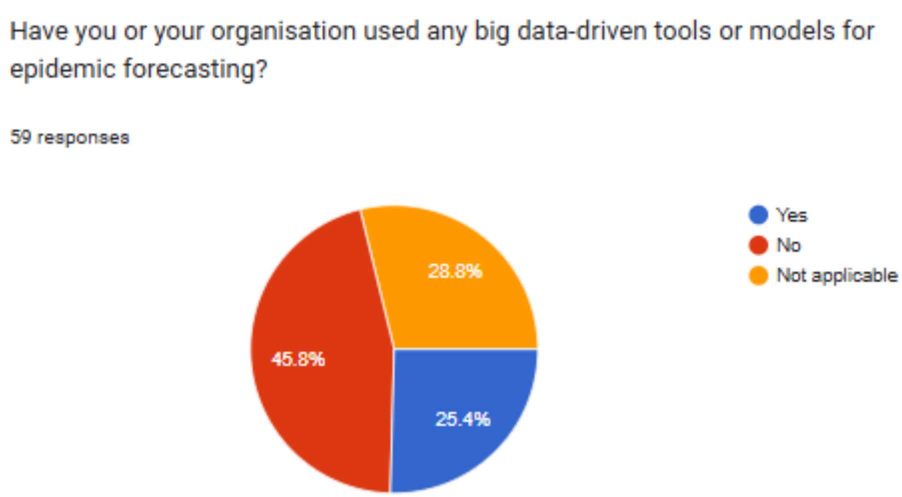

Figure 11Use of Big Data-Driven Tools

When asked, 15 respondents said they or their organisations depended on big data-based prediction instruments, but 27 chose no and 17 indicated the subject did not apply to them. It indicates that only a few are utilising big data tools right now.

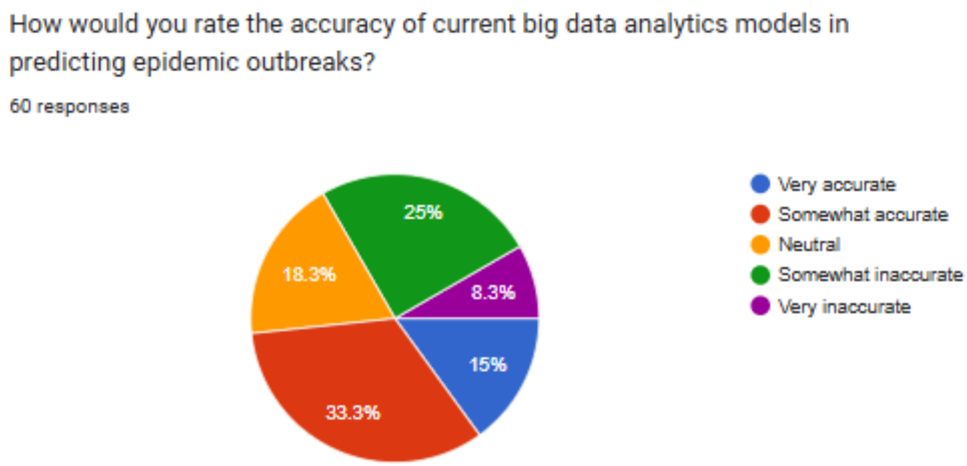

Figure 12Accuracy of Models

Out of 55 participants, 9 rated the accuracy of models as "very high," 20 as "somewhat high," 11 were neutral, and 20 rated the models as having either "somewhat" or "very low" accuracy. The varied opinions reveal that some people are still not convinced about the accuracy of predictive systems.

4.1.4 Effectiveness and Challenges of Big Data-Driven Forecasting Models

This concerns the third goal, analysing if and what stands in the way of effective big data forecasting.

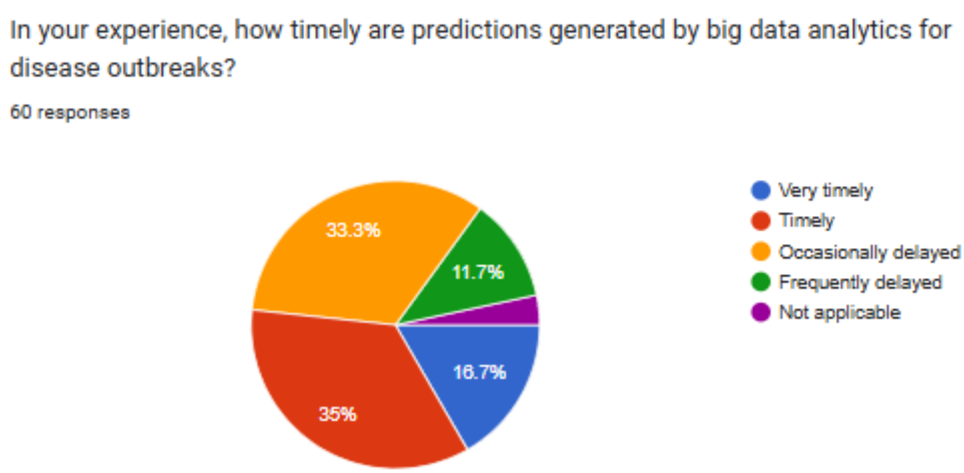

Figure 13Timeliness of Predictions

Most (31) stated that predictions were prepared in a timely way. However, a review of 27 judgment reports revealed that occasionally the judges took too long to make their pronouncements..

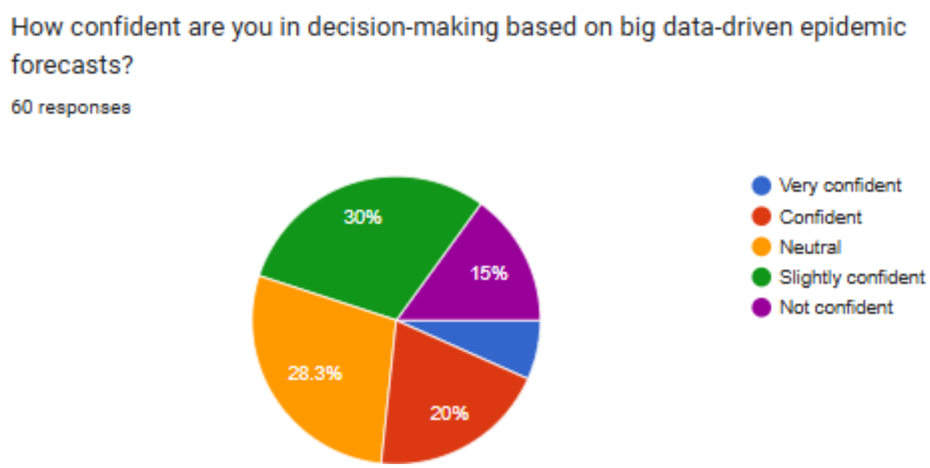

Figure 14Confidence in Decision-Making

Experts were quite unsure of their decisions based on big data forecasts: 16 were “very confident” or “confident,” but 27 felt neutral or “slightly confident,” and another 9 stated they were “not confident” with such forecasts. It means that using predictive analytics for important medical choices should be done carefully.

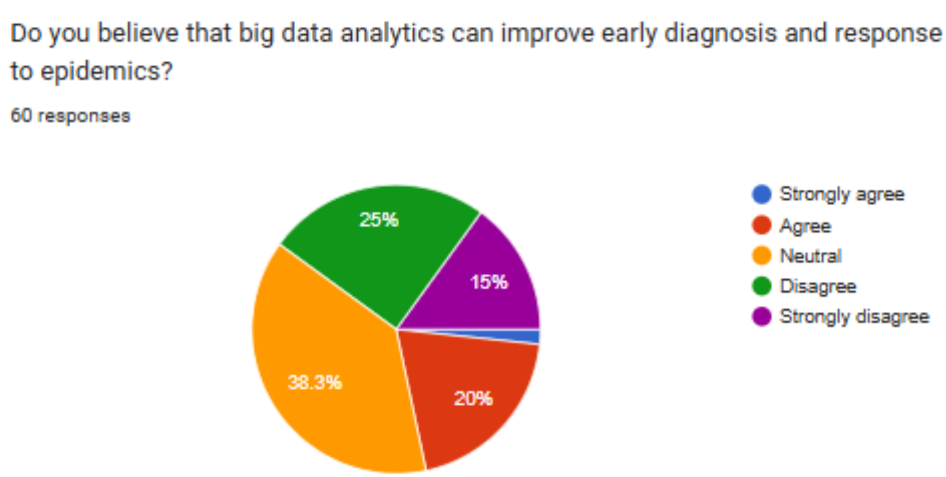

Figure 15Belief in Big Data’s Role in Early Diagnosis

Although only 13 participants considered big data analytics helpful for the early discovery of new diseases and quicker handling of epidemics, 24 gave a negative opinion. This suggests that not all public health officials are sure about how much big data helps manage outbreaks.

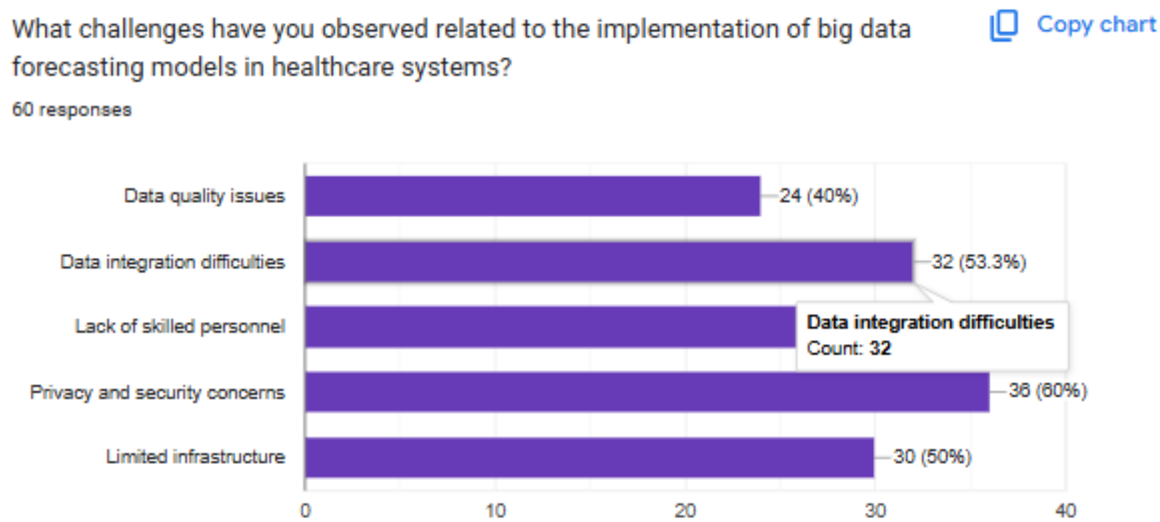

Figure 16Challenges in Implementation

Respondents listed several major issues: privacy and security (36), troubles with joining data (32), lacking enough infrastructure (30), having an insufficient number of skilled workers (27) and worries about data quality (24). They indicate the different obstacles that organisations encounter when using big data solutions.

4.1.5 Ethical and Practical Challenges of Integrating Big Data Analytics

This is the theme that addresses the fourth research objective by investigating both ethical challenges and whether big data can be easily applied.

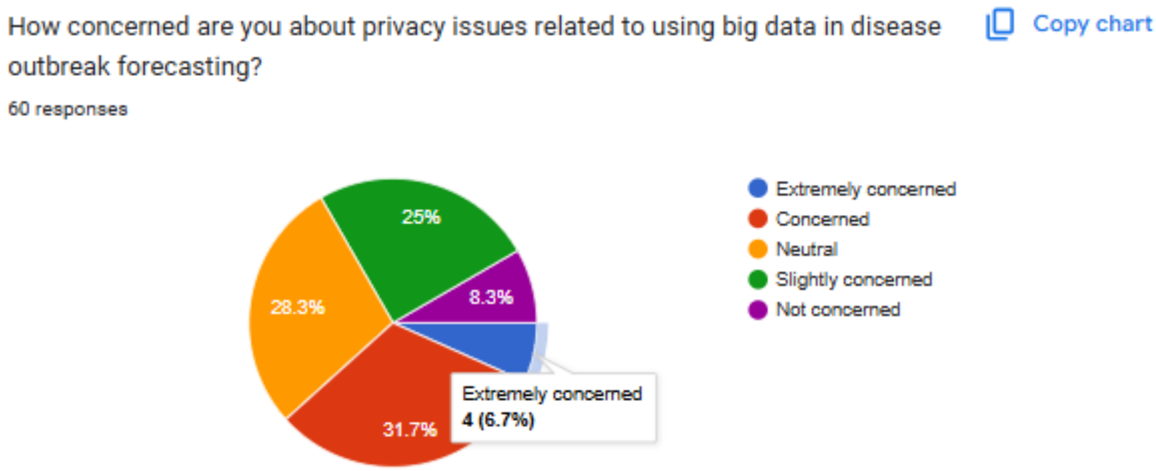

Figure 17: Privacy Concerns

Most of the respondents mentioned worries about privacy, as 23 were "extremely" or "concerned," 17 were in the middle, and 20 were either "slightly" or "not concerned." Privacy issues are still among the main concerns in the field of big data in health analytics.

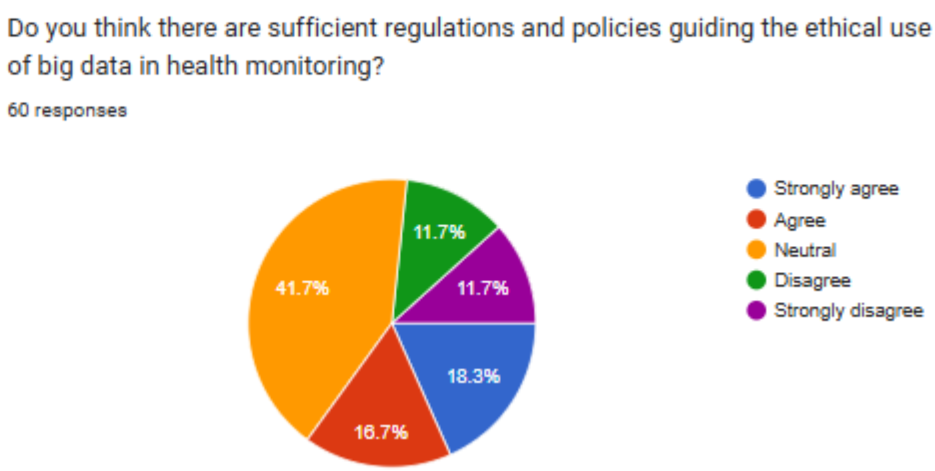

Figure 18Sufficiency of Regulations

Not everyone agreed on whether the policies are sufficient. Twenty-one people agreed or strongly agreed, fourteen disagreed or strongly disagreed, and twenty-five were neutral. This shows that some people are unsure about how well legal rules for big data are executed and regulated.

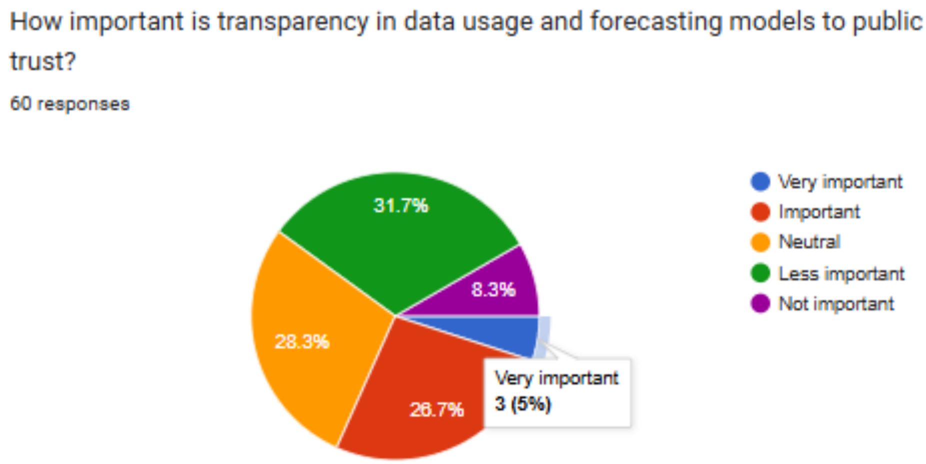

Figure 19Importance of Transparency

For 19 persons, being transparent about data usage and forecasting models was very important, for 17 it was important, and for 24 persons it was less or not important. Many believe transparency is vital for gaining public trust, but it is not always seen as the main concern.

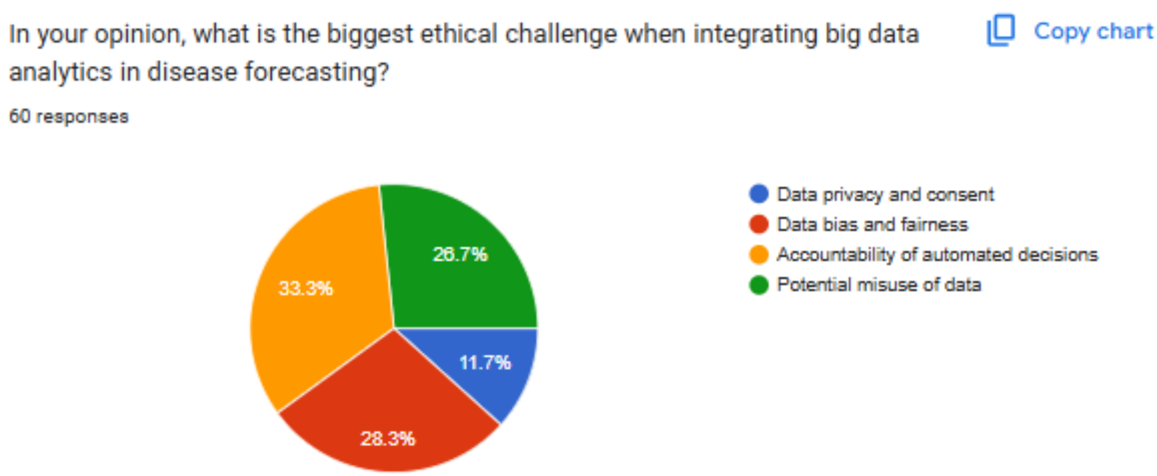

Figure 20: Biggest Ethical Challenges

They picked accountability in automation (20), data bias and fairness (17), the potential for data misuse (16) and data privacy and consent (7) as the major ethical issues. This demonstrates the level of complexity involved when it comes to ethical issues in using big data.

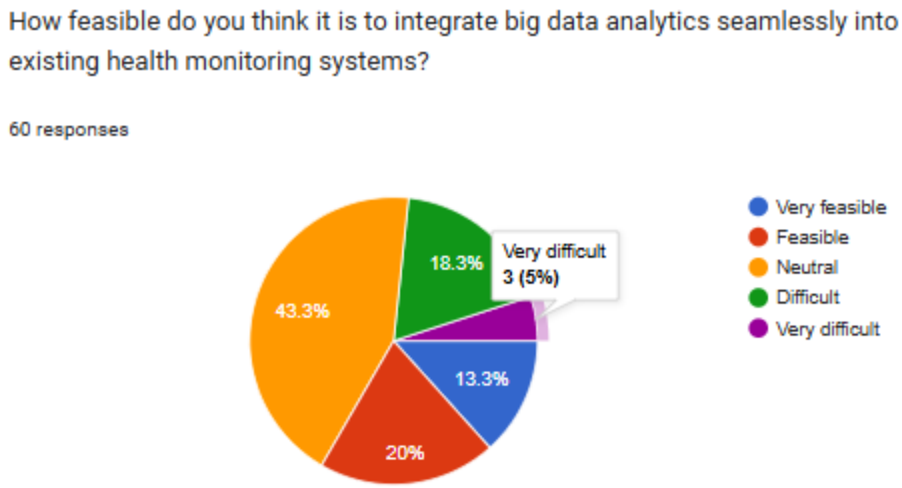

Figure 21Feasibility of Integration

Out of all the participants, only 20 stated that the integration of big data analytics into current health monitoring was a possible or highly possible option. Twenty-six did not have any issues and 14 reported that the process was not easy at all, indicating hard systemic and technical barriers.

Figure 22: Needed Support and Resources

Funding, investments, better data infrastructure, legal rules, public awareness and education and training and capacity building are important for coping with the challenges. This underscores how meaningful having the right amount of support programs is.

4.1.6 General Perceptions and Future Directions

The last theme includes the wider thoughts from respondents on the expected future use of big data analytics in predicting epidemics.

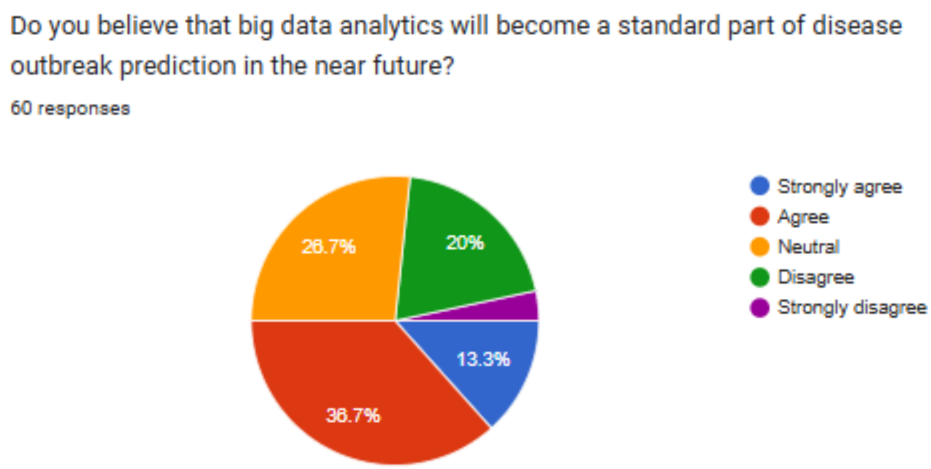

Figure 23Adoption as a Standard Tool

Thirty respondents (a majority) thought big data analytics would eventually become the norm for predicting disease outbreaks, but 14 did not agree with this and 16 were unsure. There is cautious hope that future integration will occur.

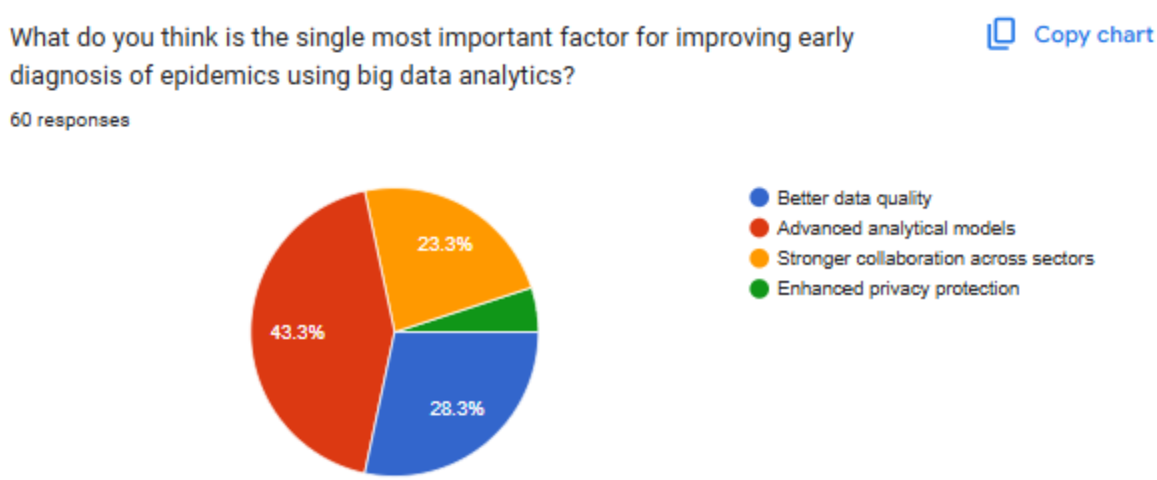

Figure 24Key Factors for Improving Early Diagnosis

Most respondents pointed out that advanced analytics and better quality of data are the biggest factors for helping early diagnosis. While there was a mention of boosting collaboration between sectors (14) and offering more secure privacy rights (3), these points were believed to be less important.

4.1.7 Summary of Findings

Survey respondents proved to be moderately aware of the subject, yet remained cautiously hopeful. Although many people are aware of what big data analytics can do for epidemic forecasting, not many believe that today's models are very accurate or easy to use in practice. Many people still trust traditional data sources the most, though newly available data from social media and mobility are gaining importance. It is best to mix classical statistical methods with the latest AI technologies, according to analytical approaches. There are many things that slow down the use of big data, including privacy fears, merging different data sources, deficient infrastructure and employees lacking certain skills. Matters such as how algorithms are used, evidence of data unconscious or conscious bias and their misuse are as important to consider as how vague the rules are. Though there are various obstacles, most believe that effective use of big data in epidemic prediction will become standard only when data is improved, models are more advanced, people cooperate, and frameworks are improved.

4.2 ANALYSIS

This chapter examines the results from the survey presented in the previous chapter based on how they relate to the research objectives. The analysis uses both numbers and personal observations to go over the data, identify relationships between important variables and study how big data analytics is useful for predicting disease outbreaks. The study results are discussed as a whole, showing their role in revealing awareness, preferences on data, tools, challenges in practice, ethics and the views on future advancements in epidemic forecasting with big data.

Awareness and Perceptions of Big Data Analytics in Epidemic Forecasting

Most of the respondents reported having reasonable knowledge of big data analytics in healthcare. Being aware of these aspects is necessary since it guides people's views on the technology's use in predicting epidemics. Yet, it appears that a significant number of participants do not understand cyber security well enough, which may limit the acceptance and use of it. To reduce this gap, people need more education and hands-on experience, which would help them integrate big data with public health efforts.

Nevertheless, people's opinions on the value of big data in predicting disease outbreaks were not all the same. Although a small group said it was extremely important, a large group had neutral or sceptical feelings about it. This may suggest that although there are many hopes for big data in public health, its practical use has not fully emerged. The result also fits with what people believe about the rare success of big data in recent efforts to predict epidemics, which points out a gap between possibilities and actions.

People participating in the survey considered environmental and climate facts, genomic information and average clinical records to be the most important for predicting disease outbreaks. It points out that public health decisions are still mainly based on legitimate, orderly data sets that are trusted. The clear picture is that social media and mobility data are valued in data-driven businesses, even though people are adopting them hesitantly. Because of these issues, researchers are still challenged by the mixture of unstructured and complex types of data in their models.

Data Sources and Analytical Methodologies: Preferences and Reliability

The assessment of different data sources indicated that mobility data from mobile phones and search engines was preferred. This choice is explained by the belief that this data can offer very accurate and up-to-date measures of population activity, which play a key role in predicting the spread of diseases. On the other hand, social media and satellite information were considered less reliable, which may have resulted from noise in the data, their representativeness and differences in quality.

The survey points out that NLP is favoured by most, just like classical statistical modelling and time series analysis. Since there is interest in using unstructured tools, statistical techniques are also important because of their transparent nature and proven effectiveness. It seems that machine learning models and network analysis were preferred less because people had concerns about how well those approaches work in vital public health cases and about understanding what they do.

It appears that using tools based on big data in epidemic forecasting is underdeveloped, showing a big practical deficit as compared to theoretical results. This situation may be caused by issues with infrastructure, organisation or limited resources, or it could come from doubts about the current model's accuracy. The concerns about the reliability of predictive models lead to uncertainty, which shows there should be continuous model validation, and clarity about any uncertainties should be provided to those involved.

Effectiveness and Operational Challenges in Big Data Forecasting

More than half of the respondents believe forecasts arrive in time, though a noticeable minority says that some forecasts arrive too late to act quickly. Such inconsistency puts the important advantage of early warning in big data analytics at risk.

Most of the respondents reported that they did not really trust using big data to make decisions. If scientists are sceptical about predictions, it might become challenging to put forecasts into practice, so it is important to increase the reliability and clarity of such models. People's uncertainty about big data analytics likely helps explain the division of views on whether it helps epidemics be detect and handle epidemics faster.

Significant problems often hinder the implementation of big data forecasting tools. Because many people worry about their health information, privacy and security are taken very seriously by related organisations. Issues with combining data and the lack of good quality data make it harder to create clear models from various resources. A lack of infrastructure facilities and specialists points to significant barriers in the area, especially where resources are very limited. It is important to remember that technical progress can help little without enhancing skills, policies and infrastructure at the same time.

Ethical and Practical Challenges of Integrating Big Data Analytics

Integrating big data analytics in public health is mainly about addressing ethical concerns. Most participants answered that they are greatly concerned about privacy when using technology. Additionally, folks observe that existing rules do not adequately handle the challenges of handling big data in healthcare systems.

Better trust and acceptance with the public is often achieved through transparency in getting, handling and using information. There are some people who think openness should be valued most, while others question how important it is, maybe due to the contrast between transparency and privacy. There are ethical issues that include privacy, accountability in algorithm use, bias in data, fairness and misleading use of information. As big data analytics brings up several ethical matters, there is a strong need for systems that ensure users apply it responsibly and fairly.

Describing big data tools as compatible with existing monitoring systems provokes scepticism, since only half of the respondents are sure they could be adopted seamlessly. Such doubts point out that historical software, difficulty in combining different systems and routine workflow setbacks all play a part. They stress that having more funds, modern data systems, clear regulations, helpful information for everyone, and additional training for professionals is important to handle these obstacles.

General Perceptions and Future Outlook

Even with these concerns present, it seems there is hope that big data analytics will improve epidemic forecasting. The majority of respondents agree that big data will soon form a common part of handling disease outbreaks as issues are tackled.

When focusing on ways to improve early diagnosis with big data, respondents give top importance to processed data and the accuracy of calculations. Keeping both elements important is in line with the idea that solid methodology needs good, representative data to offer strong results. Even though it is not frequently mentioned, teamwork among various sectors and strong privacy rules are important for furthering epidemic forecasting.

Integration and Relationships in Findings

It turns out that awareness, chosen sources of data, used methods, real-world operations and ethics all interact in a complicated way. People who are familiar with big data appear to trust its usefulness and accuracy more, suggesting that learning more is key to a good opinion of it.

Privacy, data quality, infrastructure, and workforce challenges are all related, indicating that big data projects depend on handling all of them together. Trust and transparency are crucial for making the public and professionals more accepting, which then decides whether AI can be adopted and used well.

Summary

While big data analytics appears to boost disease outbreak prediction, it still faces many serious qualitative and quantitative issues. Though some are familiar with alternative approaches and data, traditional ways of looking at information are still mainly used. When considering adopting new technology, challenges in operations and ethics must be dealt with by looking at technical, organisational and legal issues.

The results support the goals of the research by displaying people’s current viewpoints, listing what helps or hinders progress and setting important things to focus on in the future. Such insights show us how to advance big data applications in epidemiology and make sure that predictive tools are not only advanced technically but also ethical, easy to operate and are accepted by many.

4.3 DISCUSSION

This chapter compares the results from the survey study in Chapter Four with the relevant information found in Chapters One and Two. This is to ensure the new data is located within the main academic discussions on big data analytics for outbreak forecasting, paying attention to where they agree, disagree and provide new ideas. This thorough review points out that the present findings may agree with or differ from old studies and offer new ideas. The discussion first examines where our findings align with the literature, then where they differ, and finally what we discovered in the survey that may help explain technology use and ethics in public health informatics, all seen through the lens of existing theories.

1. Survey Findings Supporting Established Literature

The Recognition of Big Data's Transformative Potential

Many survey respondents agree with the key argument in the literature that big data analytics greatly improve epidemiological surveillance and prediction of disease outbreaks (Ahmed et al., 2021; Sarumi, 2021). Most people surveyed showed moderate to high understanding of big data, indicating that more professionals are realising and welcoming the rising significance of digital epidemiology. The link between being enthusiastic in theory and practical experience seems to show that healthcare and data experts have prepared work for adopting new theories (Atuahene, Kanjanabootra and Gajendran, 2022).

Likewise, those who participated in the survey preferred electronically kept health data, genetic information and data from the environment which the literature also views as key aspects of solid public health observation (Shukla, Dhyani and Pujari, 2022). This shows that applying scientifically solid and organised data is still vital for producing dependable and effective forecasts concerning outbreaks, as laid out in the first literature review chapter.

Methodological Concordance: NLP and Statistical Models

The report's choice of natural language processing (NLP) and classical statistical models, which involve time series analysis, as their preferred approaches to analysis matches what is found in current studies. Hamilton et al. (2021) and Chen et al. (2024) explain that NLP becomes more important as it can analyze unstructured clinical notes, text from social media and public health reports, all of which are relevant now due to the surge in data. Gupta, Pandey and Pal (2021) note that their usefulness and proven ability to represent changes in diseases over time make time series and regression models valuable. It shows that professionals value both new AI and time-proven techniques, as consensus in the literature is that effective forecasting often requires combining different methods (Riezler and Hagmann, 2024).

Ethical Concerns and Privacy Prioritisation

People who took the survey also emphasized the importance of protecting data privacy and security, much as in Chapter Two. The data supports what Ros et al. (2020) have argued about the vulnerability of health data and the possible problems that can occur if it is not used correctly. Just like that, there is disagreement amongst experts on if laws like GDPR and HIPAA are powerful enough to handle the difficulties caused by big data. This study further proves that ethics shapes adoption of big data in epidemic forecasting and can affect it either positively or negatively (Pina et al., 2024).

2. Contradictions and Divergences: Survey Findings Challenging Literature Perspectives

Perceived Importance and Practical Utilisation: A Reality Check

Big data analytics is widely reported to significantly help in surveillance of epidemics in the published material (Pillai and Kumar, 2021). However, the survey shows that professionals are less certain. Half of the respondents considered big data as either less important or not important for predicting outbreaks and the majority felt it was not used effectively in recent epidemics. From this, it is clear that academics tend to idealize certain things while practitioners encounter more realistic issues in the field. This situation undermines positive accounts by demonstrating a "implementation gap" that happens when organisations cannot fully act on technology due to being slow, lacking resources and being caught up in a complex system. It makes it clear that future research and policy should deal with these practical matters as well as new technology (Ahmed et al., 2021).

Limited Enthusiasm for Advanced Machine Learning

Although machine learning and hybrid AI are becoming more important in epidemiology according to recent research (Martin-Moreno et al., 2022; Chen et al., 2024), our survey found that few respondents were willing to use them. They selected easier-to-interpret models such as NLP and statistical approaches. This disapproval goes against what the literature suggests which is that machine learning is the best way to achieve predictions (Nti and Quarcoo, 2022).

It is clear from the findings that individuals are reluctant to use these tools because of their unclear models, a struggle to explain their findings and the "black box" aspect, as Hamilton et al. (2021) suggested. This situation makes people question the literature's blind faith in AI, calling for ways that both provide clear information and maintain strong prediction models.

3. Novel Contributions from the Survey: Fresh Perspectives on Big Data Analytics in Epidemic Forecasting

Differentiated Trust in Big Data Sources

People's opinions and beliefs about different big data sources have been explored more deeply through this survey. Usual health and genetic information are thought to be very trusted and now, practitioners view mobile phone activity and what people search for on the internet as better data sources than those from social media and satellites (Corsi et al., 2020). As a result, the information gathered can provide important details that show how things such as data quality, accessibility and relevance in a topic shape acceptance.

By studying this issue, socio-technical systems theory gets a more detailed idea of how trust and the usefulness of technology influence which data sources are used in public health systems (Zhao et al., 2024).

Detailed Profiling of Operational and Resource Barriers

The survey also measures the relative importance that physicians and other professionals give to implementation problems. Concerns about privacy, data mixing problems, lack of trained personnel, poor infrastructure and insufficient funding were repeatedly pointed out, helping to determine which issues needed action, unlike in most academic works that just list them (Chowdhury et al., 2024; Rehman, Naz and Razzak, 2021).

This analysis supports Rogers' diffusion of innovation theory (2003) by pointing out what hinders knowing and using the technologies and proposes interventions such as education and necessary government policies (Awotunde et al., 2021).

Mixed Attitudes Toward Transparency and Trust

A surprising result is that many people think both for and against transparency when it comes to using and creating data forecasts. Although transparency is highlighted by theory as vital for ethical compliance, many kindergarten teachers did not see it as a crucial factor (Taylor, 2017). It highlights problems that develop when entities want more privacy, faster operations and less openness. All this complexity requires us to develop new ideas for ethical guidelines that can handle the variety of situations and interests at play (An, Babanejaddehaki and Papagelis, 2024).

4. Theoretical Perspectives Illuminating Survey Findings

Socio-Technical Systems Theory as an Explanatory Lens

Focusing on the links between social, technical, ethical and organisational challenges emphasizes the importance of socio-technical systems theory according to Baxter and Sommerville (2011). It describes how successful use of big data analytics depends on the way technology, the skills of users, company culture and rules and regulations reinforce each other. The way various forms of data, challenges in operating and ethics were treated in the survey highlights that technology's operation is always affected by the environment (Münch et al., 2022).

Diffusion of Innovations Theory in Adoption Dynamics

The evidence matches well with Rogers' Diffusion of Innovations framework (2003) and explains why certain models and data sets (interpretable, traditional) spread faster than others (advanced machine learning, coming from social media). People's preferences for certain professions can be explained by complexity, compatibility with current practices and the ability to notice benefits from their use (Pinho, Franco and Mendes, 2020). Furthermore, the importance of training, funding and institution support is consistent with Rogers' belief that good communication and social systems help people embrace new ideas.

Data Justice Theory and Ethical Governance

The main issues discussed in the survey match the ideas of Data Justice Theory (Taylor, 2017) which highlights fairness, inclusivity and power more than standard data ethics do. The emphasis on algorithmic bias, misuse and concerns about privacy follows the framework's request for fairness among data practices for those who need protection. The detailed survey outcomes focusing on transparency and trust suggest there is a consistent effort to balance ethics and real-world factors, so data justice efforts must be flexible (VILJOEN, 2021).

It becomes clear that even though the key ideas behind big data are proven, its use in daily practice has many obstacles. The survey points out that having advanced technology is not enough for a successful outcome and impact. Rather, ethical governance, preparation within the organisation, effective workers and how resources are used are just as important. Bringing in the views of experts, this study gives a deeper meaning to theories by relating them to practical situations and issues (Menzli et al., 2022). Because of this integrated approach, research, policies and strategies for epidemic forecasting can be designed more effectively.

This chapter shows how the study findings back up some literature while outlining gaps in others on big data analytics used for epidemic forecasting. In line with what the literature says, the evidence states that big data can make a big difference, data sources need to be reliable and ethical matters are key (Dencik, Hintz and Redden, 2022). In addition, researchers found that practitioners are not very impressed by how big data is used now, feel little interest in advanced machine learning and have complex opinions about transparency and trust that do not fit the standard academic beliefs. New discoveries about different kinds of data source trust, the need to clarify priority resources and complicated ethical challenges offer support for theories and prove the significance of bringing together society and technology, supporting innovations and ensuring justice in data use for public health (Menzli et al., 2022).

These aspects strengthen the field by joining theories with what is done in practice, giving useful and workable suggestions for using big data analytics as an effective, ethical and generally accepted way to predict disease outbreaks.

CONCLUDING REMARKS

The last chapter gathers the key results, looks at the impact they have on current studies, lists the challenges faced and suggests ideas for further research and actions. The main goal is to summarise the study's impact on understanding disease outbreak prediction with data and to suggest improvements for this field.

Key Findings

The purpose of the research was to investigate how big data analytics can assist in early detection of epidemics and to create a plan for including big data in medical monitoring systems. The investigation was based on four objectives, and each objective was handled by processing the survey data from 60 healthcare and data experts. Details about the main findings related to the objectives are given below.

Objective 1: The Importance of Big Data Analytics in Predicting Outbreaks

While healthcare professionals are mostly aware of big data analytics, there are still many who do not know a lot about it. It means that even today, education and sharing of new digital health tools are still facing difficulties.

Most people noticed that big data could play an important role in future epidemic predictions, although they were unsure about its practical benefits now. It was often seen that in recent epidemic responses, organizations did not use big data as often as they might have. While the idea is accepted, applying it in practice is still limited, making technology's promise and practical use quite different, a reality known as the implementation gap.

Moreover, many people in the survey felt that established data sources, for example, patient charts and genomic information, were most helpful, and they were a bit more careful about accepting new data types like what's collected from social media and mobile apps.

Objective 2: Important Types of Information and Forecasting Methods in Epidemics

It was discovered that mobile phone tracking and search engine information are considered to be more reliable for large data sets than either social media or satellite data. It shows that experts view the topic carefully, thinking about new ideas as well as how accurately data reflects reality.

To analyse data, NLP and time series analysis were relied upon more than other methods. Machine learning and hybrid AI models, which got attention in the research, were ranked much lower, possibly due to issues related to transparency.

Although some participants had tried large data-driven forecasts, there was only moderate trust in the effectiveness of the forecasts, which means adoption is still limited.

Objective 3: Success rate of forecasting models that use Big Data

There was a mixed opinion about how well and how fast forecasts were being produced. Even though some believed the predictions were useful, several people found that the wait for the information broke down the early warning system.

Although some people showed confidence in making decisions with big data analytics, many others voiced their doubts. Also, less than 25 per cent of respondents strongly thought that big data brings much success in early detection and response.

The report stated that certain elements seriously affected eHealth operations, such as low data quality, challenges connecting diverse datasets, insufficient workforce skills, privacy and security risks and a lack of supportive infrastructure when resources are scarce.

Objective 4: Dealing with the Ethics and Practical Issues of Using Big Data

Privacy was the main issue, with lots of participants stating that they were either “extremely” or “somewhat” worried about their data being used properly and ethically. Many dealt with the question of whether current legislation was adequate, making it clear that new regulations are required.

People felt that open reporting of data use matters, yet there were mixed opinions on this, maybe because privacy is sometimes a concern.

Besides privacy, respondents pointed out that algorithmic accountability, data bias and possible misuse are important ethical issues in the field. They demonstrate that dealing with ethics in big data epidemiology is very complex.

There were doubts about smoothly mixing big data analytics into existing health systems, especially because it was suggested that more training, better infrastructure, clearer regulations, extra funds, and higher public awareness were all necessary.

Impact on the Literature

The study contributes a lot to the current literature on how big data can improve forecasting in epidemics.

1. The study backs up theories by previous researchers, proving that big data analytics is seen as a major change and that traditional data is still at the heart of it. It justifies concerns about ethics and points out socio-technical issues as the main obstacles.

2. Using the accounts of practitioners, the study explains that the role of big data in practice is less advanced than what the literature hopes for. This quality means that the works become more realistic by highlighting adoption's actual challenges.

3. This study adds new information about how people trust and incorporate different kinds of data sources, more detailed than what is commonly covered by other studies that mix different forms of data.

4. The research uncovered unique feelings about data transparency, suggesting the need for new ethics to handle conflicting goals and a mix of stakeholders. It widens the scope of ethics by noticing everyday problems that arise when doing what is right and wrong.

5. Using Socio-Technical Systems Theory, Diffusion of Innovations Theory and Data Justice Perspectives allows for an interpretive lens that strengthens both the theory and its usefulness in public health big data.

All in all, the study merges both theory and practical work, giving a detailed and flexible picture of big data analytics' issues and chances in this area.

Limitations of the Study

Even though it offers useful information, the study is also restricted by a number of issues.

1. The survey received answers from 60 participants, and while enough for exploratory analysis, it does not allow for generalisation. Participants who responded were intentionally chosen to know big data and public health, which may have biased the results.

2. The survey centred on individuals located in particular regions such as the UK and Europe. Thus, data and results may not represent the whole world's healthcare, mainly where regulation and infrastructure differ widely in low- and middle-income countries.

3. Using self-reports, the survey may be influenced in various ways by how participants answer. Because of this, how accurately people describe their familiarity, level of confidence, and challenges may suffer.

4. The design of the study meant it could not track changes in opinions and activities over extended periods, as fields in technology are growing quickly.

5. It was more important for the research to study perceptions than to evaluate actual outcomes or technical details. As a result, we cannot draw proper conclusions regarding how effective or important the model is in the health laboratory.

Future Recommendations

Building on the findings and limitations, several recommendations emerge to advance research, policy, and practice in big data-driven epidemic forecasting:

Promote Training: Programs aimed at public health experts and data scientists should be made to help them understand technical knowledge and what is right and wrong in using information.

Promote Multi-Sectoral Collaboration: Since combining information and methods needs "all hands on deck", information and analysis should be coordinated with experts in healthcare, technology, education and policy fields. Teamwork approaches make data sharing, standardisation and increasing trust easier (Majeed and Hwang, 2021).

Invest in Infrastructure and Capacity Building: Investing usually in information technology, fast processors and employees helps, particularly in short-staffed and underfunded places, to provide accurate predictions for outbreaks.

Enhance Transparency and Ethical Governance: Governments and businesses must develop clear, flexible policies and ethical rules that protect data while making algorithm development and data use open to oversight (Rodríguez et al., 2022).

Encourage Incremental Adoption of Advanced Analytics: Help design algorithms that are a bit easier to understand and use at first, and gradually increase their use once trust in their results is gained.

Conduct Longitudinal and Outcome-Based Research: It is suggested that future research use long-term studies and evaluation of main outcomes to investigate more meaningful aspects of forecasting (Chao et al., 2023).

Include Broader Geographic and Cultural Contexts: Looking at more parts of the world, such as low- and middle-income countries, may highlight certain health problems and solutions, contributing to globally fair health safety.

Foster Public Engagement and Awareness: Teaching the public about big data in health surveillance may result in more approval, better information sharing and cleaner data when involvement is valued (Sheng et al., 2020).

Summary

By linking the study's findings to existing studies, this chapter identified areas that fit well, revealing important disagreements and showing unique results. It points out that big data analytics is full of opportunities but still needs a good mix of technology, ethics and real investments. Even though some aspects aren't generalisable, the research supplies a strong basis and practical steps for improving predictive epidemiology. By increasing knowledge of how big data works in public health, this dissertation helps advance discussions on the topic and enables the use of digital innovations for ethical, prompt and equal predictions of outbreaks.

BIBLIOGRAPHY

Ahmed, I., Ahmad, M., Jeon, G. and Piccialli, F. (2021). A Framework for Pandemic Prediction Using Big Data Analytics. *Big Data Research*, 25, p.100190. doi:https://doi.org/10.1016/j.bdr.2021.100190.

Ali, I.M. (2024). A Guide for Positivist Research Paradigm: from Philosophy to Methodology. *Ideology Journal*, 9(2), pp.187–196. doi:https://doi.org/10.24191/idealogy.v9i2.596.

Althobaiti, K. (2021). Surveillance in Next-Generation Personalized Healthcare: Science and Ethics of Data Analytics in Healthcare. *The New Bioethics*, 27(4), pp.1–25. doi:https://doi.org/10.1080/20502877.2021.1993055.

An, A., Babanejaddehaki, G. and Papagelis, M. (2024). Disease Outbreak Detection and Forecasting: A Review of Methods and Data Sources. *ACM Transactions on Computing for Healthcare*, 6(2). doi:https://doi.org/10.1145/3708549.

Atuahene, B.T., Kanjanabootra, S. and Gajendran, T. (2022). Transformative role of big data through enabling capability recognition in construction. *Construction Management and Economics*, pp.1–24. doi:https://doi.org/10.1080/01446193.2022.2132523.

Awotunde, J.B., Jimoh, R.G., Oladipo, I.D., Abdulraheem, M., Jimoh, T.B. and Ajamu, G.J. (2021). Big Data and Data Analytics for an Enhanced COVID-19 Epidemic Management. *Artificial Intelligence for COVID-19*, pp.11–29. doi:https://doi.org/10.1007/978-3-030-69744-0_2.

Azmi, J., Arif, M., Nafis, M.T., Alam, M.A., Tanweer, S. and Wang, G. (2022). A systematic review on machine learning approaches for cardiovascular disease prediction using medical big data. *Medical Engineering & Physics*, 105, p.103825. doi:https://doi.org/10.1016/j.medengphy.2022.103825.

Bentouhami, H., Casas, L. and Weyler, J. (2021). Reporting of 'Theoretical Design' in Explanatory Research: A Critical Appraisal of Research on Early Life Exposure to Antibiotics and the Occurrence of Asthma. *Clinical Epidemiology*, [online] Volume 13, pp.755–767. doi:https://doi.org/10.2147/clep.s318287.

Bianconi, F. (2024). Frequency Distributions. *Data and Process Visualisation for Graphic Communication*, pp.67–80. doi:https://doi.org/10.1007/978-3-031-57051-3_5.

Chao, K., Sarker, N.I., Ali, I., Radin, B., Azman, A. and Shaed, M.M. (2023). Big data-driven public health policy making: Potential for the healthcare industry. *Heliyon*, 9(9), pp.e19681–e19681. doi:https://doi.org/10.1016/j.heliyon.2023.e19681.

Chen, S., Yu, J., Chamouni, S., Wang, Y. and Li, Y. (2024). Integrating machine learning and artificial intelligence in life-course epidemiology: pathways to innovative public health solutions. *BMC Medicine*, 22(1). doi:https://doi.org/10.1186/s12916-024-03566-x.

Chowdhury, A.T., Saha, P., Majid, M.E., Mushtak, A. and Kabir, M.A. (2024). Application of Big Data in Infectious Disease Surveillance: Contemporary Challenges and Solutions. pp.51–71. doi:https://doi.org/10.1007/978-3-031-59967-5_3.

Coleman, P. (2022). Validity and Reliability within Qualitative Research for the Caring Sciences. *International Journal of Caring Sciences*, 14(3), pp.2041–2045. doi:https://oro.open.ac.uk/81588/.

Corsi, A., de Souza, F.F., Pagani, R.N. and Kovaleski, J.L. (2020). Big data analytics as a tool for fighting pandemics: a systematic review of literature. *Journal of Ambient Intelligence and Humanised Computing*, [online] 12(10), pp.1–18. doi:https://doi.org/10.1007/s12652-020-02617-4.

Corsi, A., de Souza, F.F., Pagani, R.N. and 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*, [online] 12(10), pp.1–18. doi:https://doi.org/10.1007/s12652-020-02617-4.

Dencik, L., Hintz, A. and Redden, J. (2022). *Data Justice*. https://www.torrossa.com/gs/resourceProxy?an=5409499&publisher=FZ7200.

Grybauskas, A., Pilinkienė, V. and Stundžienė, A. (2021). Predictive analytics using Big Data for the real estate market during the COVID-19 pandemic. *Journal of Big Data*, 8(1). doi:https://doi.org/10.1186/s40537-021-00476-0.

Gupta, R., Pandey, G. and Pal, S.K. (2021). Comparative analysis of epidemiological models for COVID-19 pandemic predictions. *Biostatistics & Epidemiology*, 5(1), pp.69–91. doi:https://doi.org/10.1080/24709360.2021.1913709.

Hall, J.R., Hall, S.S. and Shaw, E.H. (2022). A Deductive Approach to a Systematic Review of Entrepreneurship Literature. *Management Review Quarterly*, [online] 73(1), pp.1–30. doi:https://doi.org/10.1007/s11301-022-00266-9.

Hamilton, A.J., Strauss, A.T., Martinez, D.A., Hinson, J.S., Levin, S., Lin, G. and Klein, E.Y. (2021). Machine learning and artificial intelligence: applications in healthcare epidemiology. *Antimicrobial Stewardship & Healthcare Epidemiology*, [online] 1(1). doi:https://doi.org/10.1017/ash.2021.192.

Hasan, N., Rana, R. and Chowdhury, S. (2025). *Ethical considerations in research*. [online] Academia.edu. Available at: https://www.academia.edu/download/74657806/Ethical_Considerations_In_Research.pdf.

Kerasidou, A. and Kerasidou, C. (2023). Data-driven research and healthcare: public trust, data governance and the NHS. *BMC Medical Ethics*, 24(1). doi:https://doi.org/10.1186/s12910-023-00922-z.

MacAulay, S., Ellison, A.R., Kille, P. and Cable, J. (2022). Moving towards improved surveillance and earlier diagnosis of aquatic pathogens: From traditional methods to emerging technologies. *Reviews in Aquaculture*, 14(4), pp.1813–1829. doi:https://doi.org/10.1111/raq.12674.

Majeed, A. and Hwang, S.O. (2021). Data-Driven Analytics Leveraging Artificial Intelligence in the Era of COVID-19: An Insightful Review of Recent Developments. *Symmetry*, 14(1), p.16. doi:https://doi.org/10.3390/sym14010016.

Martin-Moreno, J.M., Alegre-Martinez, A., Martin-Gorgojo, V., Alfonso-Sanchez, J.L., Torres, F. and Pallares-Carratala, V. (2022). Predictive Models for Forecasting Public Health Scenarios: Practical Experiences Applied during the First Wave of the COVID-19 Pandemic. *International Journal of Environmental Research and Public Health*, 19(9), p.5546. doi:https://doi.org/10.3390/ijerph19095546.

Mazhar, S.A., Anjum, R., Anwar, A.I. and Khan, A.A. (2021). Methods of Data Collection: a Fundamental Tool of Research. *Journal of Integrated Community Health*, 10(1), pp.6–10. doi:https://doi.org/10.24321/2319.9113.202101.

Menzli, L.J., Smirani, L.K., Boulahia, J.A. and Hadjouni, M. (2022). Investigation of Open Educational Resources Adoption in Higher Education Using Rogers' Diffusion of Innovation Theory. *Heliyon*, 8(7), p.e09885. doi:https://doi.org/10.1016/j.heliyon.2022.e09885.

Morris, C.E., Géniaux, G., Nédellec, C., Sauvion, N. and Soubeyrand, S. (2021). One Health concepts and challenges for surveillance, forecasting, and mitigation of plant disease beyond the traditional scope of crop production. *Plant Pathology*, 71(1), pp.86–97. doi:https://doi.org/10.1111/ppa.13446.

Münch, C., Marx, E., Benz, L., Hartmann, E. and Matzner, M. (2022). Capabilities of digital servitization: Evidence from the socio-technical systems theory. *Technological Forecasting and Social Change*, 176, p.121361. doi:https://doi.org/10.1016/j.techfore.2021.121361.

Nti, I. and Quarcoo, J. (2022). *A mini-review of machine learning in big data analytics: Applications, challenges, and prospects | TUP Journals & Magazine | IEEE Xplore*. [online] ieeexplore.ieee.org. Available at: https://ieeexplore.ieee.org/abstract/document/9691296/.

Omair, A. (2020). Sample size estimation and sampling techniques for selecting a representative sample. *Journal of Health Specialties*, 2(4), pp.142–147. doi:http://www.thejhs.org/article_issn-1658-600X-year-2014-volume-2-issue-4-spage-142-epage-147-aulast-Omair.html.

Pillai, S.V. and Kumar, R.S. (2021). The role of data-driven artificial intelligence on COVID-19 disease management in public sphere: a review. *DECISION*. doi:https://doi.org/10.1007/s40622-021-00289-3.

Pina, E., Ramos, J., Jorge, H., Váz, P., Silva, J., Wanzeller, C., Abbasi, M. and Martins, P. (2024). Data Privacy and Ethical Considerations in Database Management. *Journal of Cybersecurity and Privacy*, [online] 4(3), pp.494–517. doi:https://doi.org/10.3390/jcp4030024.

Pinho, C., Franco, M. and Mendes, L. (2020). Application of innovation diffusion theory to the E-learning process: higher education context. *Education and Information Technologies*, 26(1), pp.421–440. doi:https://doi.org/10.1007/s10639-020-10269-2.

Rajasegar, R.S., Gouthaman, P. and Ponnusamy, V. (2024). Data Privacy and Ethics in Data Analytics. *Studies in big data*, pp.195–213. doi:https://doi.org/10.1007/978-981-97-0448-4_10.

Rehman, A., Naz, S. and Razzak, I. (2021). Leveraging big data analytics in healthcare enhancement: trends, challenges and opportunities. *Multimedia Systems*, [online] 28(4), pp.1339–1371. doi:https://doi.org/10.1007/s00530-020-00736-8.

Riezler, S. and Hagmann, M. (2024). *Validity, reliability, and significance: Empirical methods for NLP and data science*. [online] Google Books. Available at: https://books.google.com/books?hl=en&lr=&id=aX4NEQAAQBAJ&oi=fnd&pg=PP7&dq=Methodological+Concordance:+NLP+and+Statistical+Models&ots=nyIBSk_kVl&sig=08YO256zVogrqpN3o_jYuOuC-7w.

Ristaino, J.B., Anderson, P.K., Bebber, D.P., Brauman, K.A., Cunniffe, N.J., Fedoroff, N.V., Finegold, C., Garrett, K.A., Gilligan, C.A., Jones, C.M., Martin, M.D., MacDonald, G.K., Neenan, P., Records, A., Schmale, D.G., Tateosian, L. and Wei, Q. (2021). The Persistent Threat of Emerging Plant Disease Pandemics to Global Food Security. *Proceedings of the National Academy of Sciences*, [online] 118(23), p.e2022239118. doi:https://doi.org/10.1073/pnas.2022239118.

Rodríguez, A., Agarwal, P.C., Ho, J. and Patel, M.A. (2022). Data-Centric Epidemic Forecasting: A Survey. *arXiv (Cornell University)*. doi:https://doi.org/10.48550/arxiv.2207.09370.

Ros, F., Kush, R., Friedman, C., Gil Zorzo, E., Rivero Corte, P., Rubin, J.C., Sanchez, B., Stocco, P. and Van Houweling, D. (2020). Addressing the Covid-19 pandemic and future public health challenges through global collaboration and a data-driven systems approach. *Learning Health Systems*, 5(1). doi:https://doi.org/10.1002/lrh2.10253.

Sarumi, O.A. (2021). Machine Learning-Based Big Data Analytics Framework for Ebola Outbreak Surveillance. *Advances in Intelligent Systems and Computing*, pp.580–589. doi:https://doi.org/10.1007/978-3-030-71187-0_53.

Sheng, J., Amoah, J.A., Khan, Z. and Wang, X. (2020). COVID-19 Pandemic in the New Era of Big Data Analytics: Methodological Innovations and Future Research Directions. *British Journal of Management*, [online] 32(4). doi:https://doi.org/10.1111/1467-8551.12441.

Shukla, J., Dhyani, S. and Pujari, P. (2022). Big data analytics, a transformative approach to long-term restoration research for meeting global targets. *Restoration Ecology*. doi:https://doi.org/10.1111/rec.13814.

Souza, J., Leung, C.K. and Cuzzocrea, A. (2020). An Innovative Big Data Predictive Analytics Framework over Hybrid Big Data Sources with an Application for Disease Analytics. *Advanced Information Networking and Applications*, pp.669–680. doi:https://doi.org/10.1007/978-3-030-44041-1_59.

VILJOEN, S. (2021). A Relational Theory of Data Governance. *The Yale Law Journal*, [online] 131(2), pp.573–654. doi:https://doi.org/10.2307/45400961.

Walvoord, E.C., Howenstine, M.S., Allen, B.L., Ribera, A.K., Nabhan, Z.M., Tori, A.J., Eichholtz, R.D. and Dankoski, M.E. (2022). Engaging All Stakeholders to Create a Trusted, Data-Driven, Process Improvement Approach to Addressing Learner Mistreatment. *Teaching and Learning in Medicine*, 36(1), pp.61–71. doi:https://doi.org/10.1080/10401334.2022.2122979.

Zaçe, D., Hoxhaj, I., Orfino, A., Viteritti, A.M., Janiri, L. and Di Pietro, M.L. (2021). Interventions to address mental health issues in healthcare workers during infectious disease outbreaks: A systematic review. *Journal of Psychiatric Research*, 136, pp.319–333. doi:https://doi.org/10.1016/j.jpsychires.2021.02.019.

Zhao, A.P., Li, S., Cao, Z., Hu, P.J.-H., Wang, J., Xiang, Y., Xie, D. and Lu, X. (2024). AI for Science: Predicting Infectious Diseases. *Journal of safety science and resilience*, 5(2). doi:https://doi.org/10.1016/j.jnlssr.2024.02.002.

APPENDIX

Questionnaire: Predicting Disease Outbreaks Using Big Data Analytics

Demographic Information

1. Age:
   - Under 25 / 25-34 / 35-44 / 45-54 / 55 and above
2. Gender:
   - Male / Female / Non-binary / Prefer not to say
3. Job Occupation:
   - Healthcare professional / Data analyst / Researcher / IT professional / Government official / Other (please specify) _______
4. Years of experience in your current field:
   - Less than 1 year / 1-3 years / 4-6 years / 7-10 years / More than 10 years

Section 1: Role of Big Data Analytics in Forecasting Disease Outbreaks How familiar are you with the concept of big data analytics in healthcare?

   - Very familiar / Somewhat familiar / Not familiar

2. In your opinion, how important is big data analytics in forecasting disease outbreaks and epidemic trends?
   - Extremely important / Important / Neutral / Less important / Not important
3. How frequently do you think big data analytics has been used effectively in recent epidemic forecasting?
   - Very frequently / Occasionally / Rarely / Never / Don't know

4. What types of data do you believe are most useful for forecasting disease outbreaks using big data? (Select all that apply)
    - Social media data
    - Hospital and clinical records
    - Environmental and climate data
    - Mobility and travel data
    - Genomic data
    - Other (please specify) _______

Section 2: Key Data Sources and Methodologies in Epidemic Forecasting

5. Which big data sources do you consider most reliable for epidemic forecasting?
    - Public health databases
    - Social media platforms
    - Mobile phone tracking data
    - Search engine queries
    - Satellite and sensor data
6. What analytical methodologies do you think are most effective in disease outbreak forecasting?
    - Machine learning models
    - Statistical modelling (e.g., time series analysis)
    - Natural language processing (NLP)
    - Network analysis
    - Other (please specify) _______

7. Have you or your organisation used any big data-driven tools or models for epidemic forecasting?
    - Yes / No / Not applicable
8. How would you rate the accuracy of current big data analytics models in predicting epidemic outbreaks?
    - Very accurate / Somewhat accurate / Neutral / Somewhat inaccurate / Very inaccurate

Section 3: Effectiveness of Big Data-Driven Forecasting Models

9. In your experience, how timely are predictions generated by big data analytics for disease outbreaks?
    - Very timely / Timely / Occasionally delayed / Frequently delayed / Not applicable
10. How confident are you in decision-making based on big data-driven epidemic forecasts?
    - Very confident / Confident / Neutral / Slightly confident / Not confident
11. Do you believe that big data analytics can improve early diagnosis and response to epidemics?
    - Strongly agree / Agree / Neutral / Disagree / Strongly disagree
12. What challenges have you observed related to the implementation of big data forecasting models in healthcare systems? (Select all that apply)
    - Data quality issues
    - Data integration difficulties
    - Lack of skilled personnel
    - Privacy and security concerns
    - Limited infrastructure

- Other (please specify) _______

Section 4: Ethical and Practical Challenges of Integrating Big Data Analytics

13. How concerned are you about privacy issues related to using big data in disease outbreak forecasting?
    - Extremely concerned / Concerned / Neutral / Slightly concerned / Not concerned
14. Do you think there are sufficient regulations and policies guiding the ethical use of big data in health monitoring?
    - Strongly agree / Agree / Neutral / Disagree / Strongly disagree
15. How important is transparency in data usage and forecasting models to public trust?
    - Very important / Important / Neutral / Less important / Not important
16. In your opinion, what is the biggest ethical challenge when integrating big data analytics in disease forecasting?
    - Data privacy and consent
    - Data bias and fairness
    - Accountability of automated decisions
    - Potential misuse of data
    - Other (please specify) _______
17. How feasible do you think it is to integrate big data analytics seamlessly into existing health monitoring systems?
    - Very feasible / Feasible / Neutral / Difficult / Very difficult
18. What support or resources are needed most to overcome practical challenges in using big data for epidemic forecasting? (Select all that apply)
    - Training and capacity building

- Improved data infrastructure
- Regulatory frameworks
- Funding and investments
- Public awareness and education
- Other (please specify) _______

Section 5: General Perceptions

19. Do you believe that big data analytics will become a standard part of disease outbreak prediction in the near future?
    - Strongly agree / Agree / Neutral / Disagree / Strongly disagree
20. What do you think is the single most important factor for improving early diagnosis of epidemics using big data analytics?
    - Better data quality
    - Advanced analytical models
    - Stronger collaboration across sectors
    - Enhanced privacy protection
    - Other (please specify) _______