International Journal of Latest Engineering and Management Research (IJLEMR)
ISSN: 2455-4847
www.ijlemr.com || Volume 07 - Issue 11 || November 2022 || PP. 01-05

Impact of Data Science on Health Genomic

Mehmood Ali Mohammed[1]
[1]*University of the Cumberlands, Williamsburg, KY USA*

Abstract: Data science is analyzing data and extracting critical knowledge. Its application in a technical field is usually within medicine, where wisdom encompasses integrating principles and techniques that enable extensive data analysis. On the other hand, genomics is a multi-disciplinary approach focusing on understanding the biological structure, functional components, and evolution of genetic bodies. The concept of genomics covers the scientific sequencing of genomes, analysis of the transcripts, and determining the overall set of genes. This research paper sought to investigate the potential impact of data science in advancing genomics in healthcare. Also, the study sought to discover the potential challenges and future possibilities of data science in genomic medicine. The theories applied were the Data Science Theory, the Theory of Information Systems, the Kinship Theory, and the Neo-selectionist Theory. The researcher utilized the mixed research design by deploying the quantitative and the qualitative research design to address the study objectives. The results indicated the existence of a positive connotation between the variables of data science and genomics. Data analytics, data strategy, and operationalization positively influenced the genomics study in healthcare. The findings revealed that data science, along with its highlighted variables, had a positive impact on the practice of genomics in medicine and healthcare. The association was significant and thus provided a basis to encourage using the concepts in combination within healthcare.

Keywords: Data Science, Health Genomics, Information System, Health care system

I. INTRODUCTION

Data science is a relatively new paradigm considered one of the most significant and advanced discoveries in the 21st century. According to [1], data science has been deployed successfully to accelerate the unearthing of probability outcomes within several domains. Therefore, data science can be perceived as a practice that is concerned with the analysis of data and extraction of critical knowledge. The concept involves analyzing Big Data, thus obtaining correlations between variables while estimating the probable outcomes and errors. As a result, the application of data science in a technical field such as medicine encompasses integrating principles and techniques that enable the analysis of extensive data in terms of scale, velocity, and volume to accelerate the study of phenomena that have been represented by the collected data [2]. The specific objectives studied in this article include the following:

- Understanding the concept of data science and genomics.
- Elaborating on the application of data science in health genomics.
- Finding out the pros and cons of data science in genomics.
- Determining the implications of data science in genomics.

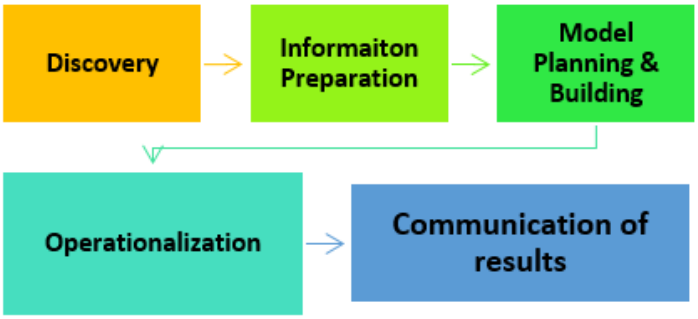

Fig 1. Data Science Process

Genomics is a multi-disciplinary and scientific approach that seeks to understand or comprehend the biological structure, functional components, and evolution of genetic bodies as well as their interactions, molecular elements, and the cloning of genomes to understand organismic phenomics [3], genomics covers the scientific sequencing of genomes, analysis of the transcripts and determining the overall set of genes. We should

note that genomics is divided into functional, structural, and comparative, which deals with constructing high-resolution genetic maps, characterizing genes, and comparing these aspects across genes. There have been suggestions regarding the interplay between data science and genomics. The former focuses on extracting knowledgeable insights from data at scale and how the same can be applied in comprehending genomics among organisms [4-5]. This paper sought to investigate the potential impact of data science in advancing genomics in healthcare.

II. BACKGROUND

Navarro et al. [6] conducted a study on the application of data science in the field of genomics-based on the umbrella approach. The research contextualized data science as an umbrella-like term comprising different subdomains. Therefore, the study focused on integrating genomics as a specified subdomain known as 3V data and the 4M model and analyzed cultural and technical exports and imports between the subdomains of data science and genomics. Furthermore, the authors discussed how data privacy, ownership, and value continue to be issues that affect the application of data science in the field of genomics. Royd et al. [7] conducted a study on the generation of health data concerning genomics in the United Kingdom. The study explored the current models used in generating genomic health data and the underlying legal, ethical, and policy issues associated with analyzing and interpreting genomic data, especially in the case of using DNA.

Khoury et al. [8] investigated the intersection of genomics health and big data, a component of data science, and the operationalization of public health. With precision public health (PPH) emerging as a significant response to the rapidly increasing availability of biobanks, genomics, and other primary sources of bug data in health care, understanding data science was vital. On the other hand, He et al. [9] conducted a study on the application of data analytics, a branch of data science in genomic medicine. The study reviewed the challenges of manipulating large-scale sequencing data and the variable clinical data retrieved from the EHRs system for medical genomics. The authors introduced potential solutions to various difficulties in manipulating, managing, and analyzing clinical and genomic data. However, there was limited focus on the broader aspect of data science and how it affects the practice of genomics.

III. RESEARCH OBJECTIVE

This study relied on a general objective and a list of specific objectives that provided a structure for conducting the research. The researcher developed objectives from the study topic and the overall research question, which entailed investigating the impact of data science on the practice of genomics in health.

The specific objectives included:

- To understand the concept of data science and genomics.
- To understand the application of data science in health genomics.
- To determine the pros and cons of data science in genomics.
- To determine the implications of data science in genomics.
- To discover data science's potential challenges and future possibilities in genomic medicine.

IV. LITERATURE REVIEW

The researcher utilized several theories that provided a basis for understanding the potential relationship between data science and health genomics. The study employed theories related to data science and its elements and the components of genomics that interact within the scope of healthcare. These theories included the Data Science Theory, the Theory of Information Systems, the Kinship Theory, and the Neo-selectionist Theory. The Data Science Theory is a form of a theoretical approach to dealing with data and the models associated with understanding knowledge discovery and solving a scientific problem [10-11]. The theory-based data science model seeks to represent the processes that are well-understood in terms of scientific principles. The theory provides the concept of data science with a more scientific perspective and basis for understanding the various forms of data and their respective interpretations.

The Theory of Information Systems is a critical theoretical domain for understanding the potential application of data science in medicine, particularly genomics. The theory is usually described in terms of the underlying constitution, the representation, and the intended achievements [12-13]. The thesis comprises a set of statements that are bound in terms of language, tend to capture specific items, and propose the nature of the relationship between the underlying concepts. Sometimes, the statements can be presented or complemented using diagrams, graphs, and tables. Regarding data science, the theory helps capture the complex world and then contribute to the existing pool of knowledge. Therefore, data science can be viewed as a concept and practice that informs the selection of accepted constructs, uses data to enhance comprehension, and generates meaning or interpretation from theory [14].

On the other hand, the Kinship Theory is regarded as a kin selection theoretical model where the male and female alleles tend to undergo different patterns within the social environment. The theory focuses on the genes that have an expression level governing the extent to which behavioral and psychological interactions occur within and outside individuals [15]. However, the theoretical kinship framework has undergone a paradigm shift concerning genealogy. Considering this, there has been a move from the traditional paradigm of kinship concerning genomics as studies have adopted the genealogical paradigm. This relatively new paradigm focuses on reproduction as a method for gendering a relationship formally represented by selected kin [16]. This theoretical perspective provides a much-needed basis for understanding the concept and practice of genomics in healthcare.

V. FRAMEWORK

This section focused on the methodology the researcher adopted to achieve the study objectives and answer the research questions. First, the team described the research design, discussed the study population, and then explained the sample size and selection criteria. The other sub-sections included the process of collecting data and the method of data analysis. The researcher utilized the mixed research design to explain the facts and features highlighted within the topic area and the target population. According to Paul [17], the selected method was suitable for this study as it will assist in responding to the research problem by discovering the associations between the independent and dependent variables chosen to represent the study objectives. , the mixed research design enabled the research to provide accurate and valid representations of the research variables by deploying the quantitative and the qualitative research design to address the study objectives [18].

The study population is the group of units for which the collected data sets assist in making inferences. The target population describes the units against which the study findings will be generalized [19].In this case, the population to be investigated will include health practitioners, particularly medical researchers, and IT specialists in medical facilities across Kentucky. The study focused on this population based on their understanding and experience with the topic of interest. Sample design refers to the framework that serves as the foundation for selecting the study sample, thus affecting significant elements of the research process [17, 20]. In this context, the researcher defined the sampling frame to represent the population of interest from which the survey sample was drawn. An example was drawn from all the hospitals across Kentucky using the census technique that listed the specific characteristics and elements that constituted the proper samples.

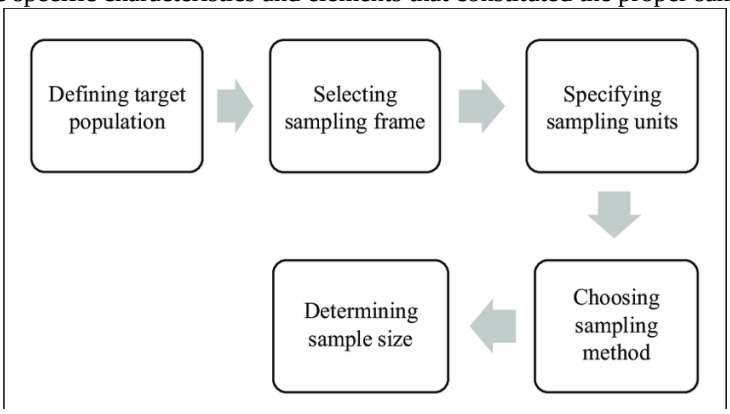

Fig 2: Sampling Process

The researcher employed both interviews and questionnaires to collect the required primary data. The questionnaires were used to collect data from the selected healthcare staff. We should note that questionnaires are regarded as appropriate for such studies since they enable the collection of valuable information that is not directly observable and accomplishments [21]. The questionnaire comprised both open and close-ended questions and helped obtain objective data from the participants since they could not be manipulated in any way by the study. According to [20], questionnaires provide an added advantage since they are less costly and require less development time. The data instrument addressed the research objective and was subdivided into two sub-sections. The first section enquired about the general information of the respondents, while the subsequent team focused on the goals. Qualitative data was obtained through key informant interviews with hospital heads of departments and their respective medical superintendent.

VI. REVIEW AND REGRESSION ANALYSIS

This section discusses the data analysis and the study results and then provides an interpretation. The researcher utilized primary data collected using questionnaires and interview schedules administered to selected

health practitioners in the health sector. The study sought to investigate the impact of data science on the practice of genomics in health. The specific objectives were to understand the concept of data science and genomics, to comprehend the application of data science in health genomics, and to find out the pros and cons of data science in genomics. Other objectives included determining the implications of data science in genomics and discovering the potential challenges and future possibilities of data science in genomic medicine. In descriptive statistics, the researcher used data science and genomics variables to investigate the association between the study variables. Below is a summary of the illustrative representations of the variables in terms of mean, standard deviation, minimum and maximum values:

Table 1. Descriptive Representation of Data

	N	**Maximum**	**Minimum**	**Mean**	**Std. Deviation**
Data Analytics	90	18.00	4.00	10.92	5.06
Big Data	90	11.00	8.00	10.83	2.10
Data Strategy	90	9.40	8.50	10.34	0.75
Data Operationalization	90	8.20	7.20	8.82	1.11
Genomics	90	7.95	0.423	3.67	4.143

Based on the findings above, the maximum value for data analytics as a variable for data science was 18.00, while the minimum value was 10.92. The mean was 10.92, and the standard deviation was 5.06. Big data had a maximum weight of 11.00, a minimum value of 8.00, and a mean of 10.83, while the standard deviation was 2.10. The maximum value for Data strategy was 9.40, the minimum value was 8.50, the mean was 10.34, and the standard deviation was 0.75. As for data operationalization as a component of data science, the maximum value was 8.20, while the minimum value was 7.20. The mean for the variable was 8.82, while the standard deviation was 1.11. Genomics had a minimum and maximum value of 0.423 and 7.95, respectively, and a mean of 3.67. The standard deviation for market value per share was 4.143.

Regression Analysis

Table 2. Model Summary

R	R Square	Adjusted R Square	Std. error of the Estimate	Durbin-Watson
0.381	0.145	0.105	3.921	0.537

The correlation coefficient for the model is 0.381, thus implying that there exists a connotation between the variables of data science, i.e., data analytics, data strategy, data operationalization, and big data. The adjusted R square was 0.105, with the value implying that the model elaborated 10.5% of the influence of the variables of data science. From the findings of analyzing the questionnaire, data analytics which was a variable of data science had a positive effect on genomics medicine. Data strategy was discovered to influence genomic practice at health facilities positively. In addition, big data was found to have a significant and positive impact on health genomics. This study focused on establishing the impact of data science on the practice of genomics in healthcare. The results indicated the existence of a positive connotation between the variables of data science and genomics. Also, big data, which involves the computation of large amounts of informational data, had a significant and positive influence on studying the biological structure, functions, and evolution of genomic elements. From these findings, it was evident that data science, along with its highlighted variables, had a positive impact on the practice of genomics in medicine and healthcare.

VII. Conclusions

The association was significant and thus provided a basis to encourage data-driven decision-making using healthcare concepts in combination with healthcare. By implementing the variables of data science in the field of genomics, there is a considerable opportunity to improve the practice through better data management and reliable outcomes. Since the study focused on ascertaining the association between data science and genomics, the anticipated outcomes sought to elaborate on any positive or negative association between the variables. In this case, the study found that data science provides a wide range of opportunities for genomics,

hence the recommendation that the latter adopt the former within its practice. However, some cons were discovered in terms of the application of data science in genomics, including data privacy and protection issues. These issues tend to plague the unrestricted use of data, hence the limitations in terms of the extent to which human data, especially DNA, can be used for genomic purposes. Furthermore, there is a need for further research concerning the influence of data science on modern genomic practice in light of the adoption of advanced technological resources that have ethical and social considerations.

VIII. Acknowledgements

I am thankful to my manager in my workplace and my university professors for encouraging me to write this article about the impact of data science in healthcare and how it is shaping modern eHealth.

References

[1]. Brodie, M. L. (2019). What Is Data Science? Applied Data Science, pp.101-130.
[2]. Das, S. R. (2016). Data Science: Theories, Models, Algorithms, and Analytics. Apache Press.
[3]. Solanke, A., & Tribhuvan, K. (2015). Genomics: An Integrative Approach for Molecular Biology. In S. M. Khurana, & M. Singh, Biotechnology—Progress and Prospects (pp. 234-270). Studium press.
[4]. Navarro, F. C., Mohsen, H., Yan, C., Li, S., Gu, M., Meyerson, W., & Gerstein, M. (2019). Genomics and data science: an application within an umbrella. Genome Biology , 20(109), 45-58.
[5]. Sharma, P., Dash, B., & Ansari, M. F. (2022). Anti-phishing techniques – a review of Cyber Defense Mechanisms. IJARCCE, 11(7). https://doi.org/10.17148/ijarcce.2022.11728
[6]. Ormondroyd, E., Border, P., Hayward , J., &Papanikitas, A. (2022). Genomic health data generation in the UK: a 360 view. European Journal of Human Genetics volume, 30(1), 782-789.
[7]. Nassaji, H. (2015). Qualitative and descriptive research: Data type versus data analysis. Language Teaching Research, 1-45. doi:https://doi.org/10.1177/1362168815572747
[8]. Khoury, M. J., Armstrong, G., Bunnell, R. E., Cyril, J., &Iademarco, M. F. (2020). The intersection of genomics and big data with public health: Opportunities for precision public health. PLOS Medicine, 17(10), 67-76.
[9]. Sharif, Md Haris Uddin, and Mehmood Ali Mohammed. "A literature review of financial losses statistics for cyber security and future trend." World Journal of Advanced Research and Reviews 15.1 (2022): 138-156.
[10]. Karpatne, A., Atluri, G., Faghmous, J. H., Steinbach, M., Banerjee, A., Ganguly, A., . . . Kumar, V. (2019). Theory-guided Data Science: A New Paradigm for Scientific Discovery from Data. Data Science, 1-27.
[11]. Dash, B., & Ansari, M. F. (2022). Self-service analytics for data-driven decision making during COVID-19 pandemic: An organization's best defense. Academia Letters, 2.
[12]. Rizk, A., &Elragal , A. (2020). Data science: developing theoretical contributions in information systems via text analytics. Journal of Big Data volume, 7(7), 13-21.
[13]. Ansari, M. F., Sharma, P. K., & Dash, B. (2022). Prevention of phishing attacks using AI-based Cybersecurity Awareness Training. International Journal of Smart Sensor and Adhoc Network., 61–72. https://doi.org/10.47893/ijssan.2022.1221
[14]. Hjørland, B. (2018). Theoretical development of information science: A brief history. Journal of Information Science, 1-35.
[15]. Patten, M. M., Ross, L., Curley, J. P., Queller, D. C., Bonduriansky, R., & Wolf , J. B. (2014). The evolution of genomic imprinting: theories, predictions and empirical tests. Heredity, 113, 119–128.
[16]. Read, D. W. (2017). Kinship theory: A paradigm shift. Ethnology , 46(4):329-364.
[17]. Paul, L. J. (2017). Sample Design. Encyclopedia of Survey Research Methods, 1(2): 1-25.
[18]. Wyk, B. v. (2019). Research design and methods: Part 1. Post-Graduate Enrolment and Throughput , Univerisity of Western Cape.
[19]. Barnett, V. (2012). Sample Survey: Principles and Methods (3rd ed.). London: Arnold.
[20]. Dash, B. (2021). A hybrid solution for extracting information from unstructured data using optical character recognition (OCR) with natural language processing (NLP).
[21]. Thompson, C. B. (2017). Descriptive Data Analysis. Retrieved August 26, 2021, from https://www.airmedicaljournal.com/article/S1067-991X(08)00297-6/pdf