Using AI and Big Data in the HealthCare Sector to help build a Smarter and more Intelligent HealthCare System

Sanjeev Kumar Marimekala
STSM and Thought Leader
IBM
svmarime@us.ibm.com

John Lamb, PhD,
Adjunct Faculty, Mathematics,
Pace University, Pleasantville, NY, USA
jlamb@pace.edu

Robert Epstein
Leader, Hybrid Cloud Distributed Automation
IBM
epsteinr@us.ibm.com

Vasundhara Bhupathi
DT Portfolio Manager-Applications
DXC Technology
vbhupathi@dxc.com

2024 IEEE World AI IoT Congress (AIIoT) | 979-8-3503-8780-3/24/$31.00 ©2024 IEEE | DOI: 10.1109/AIIoT61789.2024.10578989

Abstract -**The purpose of this paper is to demonstrate the use of AI and Big Data in HealthCare Sector to help build a smarter and more intelligent HealthCare System. Researchers in HealthCare Sectors are relying heavily on big data and compute power to build correlations by using statistical methods and artificial intelligence (AI) models. These models enable Healthcare Sector participants to manage HealthCare for a core set of the population. They also help providers to analyze the impact of decisions on their most vulnerable patients. There are many factors that are considered in performing big data analysis, some of them are: the patient's medical history, genetic information, eating habits and fitness regimen. The data that is analyzed includes several key decision-making processes. Some of the challenges with the data used include data quality, data validation, data knowledge, domain expertise, and data integration challenges with various end points. While performing data analysis, the HealthCare Sectors must take security and data governance (HIPPA regulations etc.) into consideration. Big data analysis follows the (4P) approach[1], preference, prediction, personalization, and promotion. The question that arises most often is the type of data that is the most reliable for analysis in the HealthCare Sector. Most HealthCare organizations use demographic information, diagnosis, treatment, prescription drugs, laboratory tests, physiologic monitoring data, hospitalization, and patient insurance for their analysis. Since the data comes from multiple sources[2], there is a big challenge to perform data integration, extraction, and transformation as it consumes large amounts of resources and compute power, coupled with the additional challenges of data aggregation, data enrichment and format inconsistencies. To address this challenge and to analyze the process completely requires data scientists who have domain knowledge and expertise to extract, enrich and transform data. and group them into a meaningful format for proper data analysis and research. The HealthCare Sector faces this as a major challenge. Another key component in big data analysis is the lack of data visualization tools that can take structured and unstructured data and build customized dashboards for data correlation. There are also challenges with big data management in the HealthCare Sector. This data needs to be highly secured, with proper guardrails with tightened security measures and controls, and with proper data governance in place. There is also a need for a robust infrastructure that can handle large amounts of medical data that can allow researchers to analyze, build data correlation models and generate meaningful insights by using AI models through prompt engineering techniques[3] to build data correlation. In our paper, our focus is to understand, through several use cases the key challenges in gathering and grouping HealthCare data (both structured and unstructured data) from various sources. We also focus on understanding the impact of technical advancements in emerging AI technologies and how it plays a vital role in defining and deriving meaningful data insights for research and for learning HealthCare data patterns with a primary focus on data authenticity, ethics, privacy, governance, integrity, and security.**

Index Terms – Generative AI, ChatGPT, Guardrails, Big Data, AI Models, Large Language Models, ML, Prompt Engineering and Prompt Tuning.

INTRODUCTION

AI and Big Data are being used significantly in HealthCare. Especially during the Covid 19 pandemic the AI played a vital role in finding the vaccine. This paper expands on that concept that AI can help solve the problem in HealthCare Sector.

Usage of Generative AI such as ChatGPT tool in HealthCare poses some challenging questions in our mind; A few of such examples were.

Does ChatGPT help in diagnosis of diseases or solve HealthCare Problems?
How accurately can the disease be diagnosed?
Is the Big data sufficient for the Generative AI?
Can ChatGPT help us to conduct research in HealthCare Sector?
What techniques and methods should we adopt while using ChatGPT to solve the problems in HealthCare?
Can we conclude on the results from ChatGPT, or should conduct additional tests with Realtime datasets?

In our paper, before we try to address these problems, let us look at how the patient data is integrated in HealthCare Sector. The diagram below shows the typical data transformation [4]at a high level in HealthCare Sector.

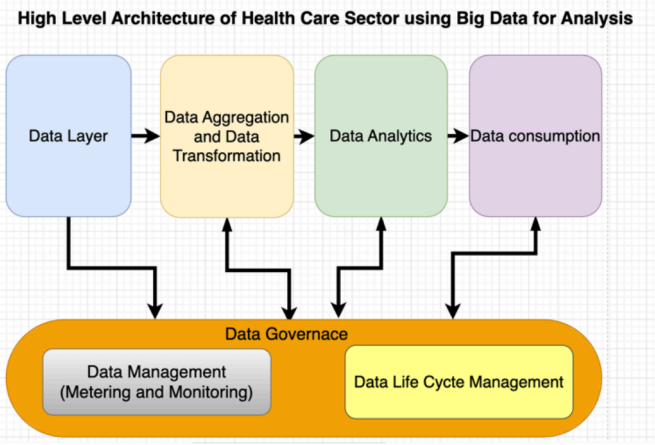


Figure 1

As we can see from Figure 1, that data is captured from different sources, and it goes through data transformation using techniques such as Extract Transform Load (ETL) or Extract Load Transform (ELT). The data transformation increases efficiency for analytics and enables data driven decisions. The Large Language Models are trained using these enriched datasets to solve the specific problems for better accuracy. In HealthCare the data confidentiality and privacy is the key concern, so any Large Language Model especially Opensource ones such as Transformer Model (ChatGPT) cannot be used in HealthCare Sector. So, there is a need for closed or proprietary models that can be trained using these enriched datasets to serve the HealthCare Sectors. Again, these proprietary models need to comply with the industry standards and adhere to laws and regulations set by the local government and establish guardrails to address bias, privacy concerns and ethics. Big data fosters innovation in Generative AI. It also enables models to explore new patterns and ideas, that can lead to unexpected and groundbreaking creations.

HYPOTHESIS

In our hypothesis, we consider that Big Data, data quality, GPU's and closed LLM's[5], Cloud Services and diagnostic data sources are crucial for Generative AI to solve HealthCare problems and to conduct research activities. In HealthCare the accuracy is very critical, and the models need to generate accuracy up to Eight 9's (99.999999 %)[6] when diagnosing and predicting the diseases. In HealthCare the multimodal[7] approach could be better alternate to yield high accuracy as the results from one Large Language model can be verified by using another Large Language Model for similar datasets to compare and validate so that there is less ambiguity in assessing the results. Our assumption is that ChatGPT does not have the access to the real HealthCare data. It can only generate random data sets and by using prompt Engineering and Prompt tuning techniques, we can provide input in textual format as a query to ChatGPT to solve our problem conceptually and by using Generative AI such as ChatGPT it helps us in researching and make logical reasoning.

KEY FACTORS OR MATERIALS

- Big Data
- Data Quality
- Proprietary LLM
- Data Governance
- Compute Power GPU's
- Multimodal
- Validation of results generated by (LLM's)
- Cloud Services
- Prompt Techniques
- Prompt Tuning
- Diagnostic Data
- HealthCare Regulations and Compliance Standards[8]

PROCEDURE

We have used ChatGPT 3.5 to study and research our use case. We have designed the use case scenario using Intelligent patient Bed Placement in a small, medium and large Hospitals that will assist patient to get recommendation from Generative AI providing the information about optimum bed alignment, space utilization, patient flow and acuity, workflow efficiency, patient satisfaction, infection control and emergency preparedness for different seasons such as summer, spring, and fall. Our approach in deriving results from ChatGPT 3.5 was using prompt engineering technique.

Case Study: Intelligent Patient Bed Placements in Hospitals using Al and Big Data.

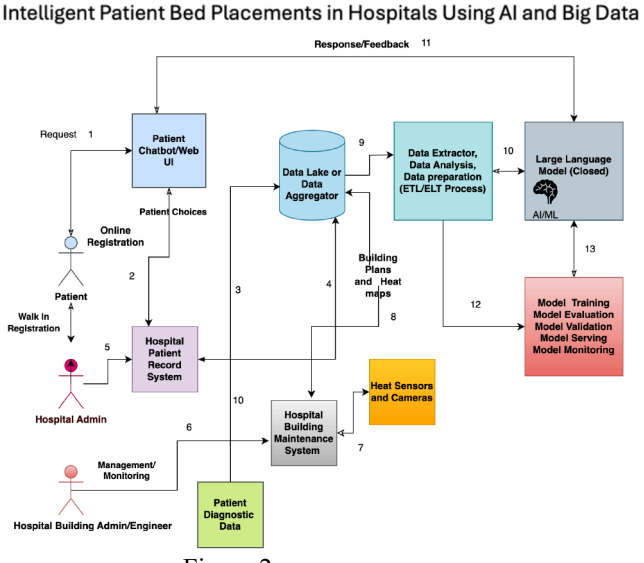


Figure 2

Detailed description of the above Figure 2.

1) Patient uses chatbot or UI (Mobile or Desktop App) to online Hospital registration and selects the bed preference.
2) The request goes through the Hospital patient record system.
3) The upstream information from patient diagnostic data source is sent to the Hospital Data Lake or Data Warehouse.
4) The Hospital building maintenance system sends the heating, cooling, and energy consumption, bed position to the Data Lake.
5) If the patient registers at the hospital, the hospital admin will input the patient preferences and this information is recorded in the Hospital patient record system.
6) The Hospital building Admin or Engineer monitors the Hospital Building Maintenance System.
7) The Hospital Building Maintenance System receives the images of heatmaps (hot and cold spots) from Heat Sensors and Cameras.
8) Hospital Building Maintenance System sends the information about heatmap and images of hot and cold spots of patient room to the Data Lake.
9) The data from Data Lake is Extracted, Transformed and Loaded (ETL/ELT) and then Labeled/Grouped into specific datasets.
10)The Large Language Model (LLM) uses the datasets from step 9.
11) The response is sent to the patient in real time, the experiences would be based on the choice of bed preferences they made using the LLM simulation.
12) If there is a feedback from patient or any changes to the Hospital bed preferences then the model will adjust and learn based on the feedback from patient and the input parameters provided by the patient.
13) At this step, the model training, model evaluation, model validation, model serving and model monitoring is done to make further improvements in model performance[9] and accuracy.

DATA

The data shown below is collected using ChatGPT 3.5 using prompt engineering technique. The ChatGPT generated the randomized data that can be tailored to the specific needs based on the use case requirements.

Data categories:

- Seasons
- Optimum Bed Alignment
- Space Utilization
- Patient Flow and Acuity
- Workflow Efficiency
- Patient Satisfaction
- Infection Control
- Emergency Preparedness

Data Groups:

Seasons: Summer, Spring, Fall
Hospital Size: Small, Medium, Large
Data Category: Small, Medium, Large

Data Category	Small	Medium	Large
Seasons	Summer	Spring	Fall
Optimum Bed Alignment	Align beds to maximize airflow and natural ventilation. Consider placing beds away from direct sunlight.	Align beds to ensure adequate airflow and natural ventilation. Consider positioning beds to receive morning sunlight.	Align beds to optimize heating distribution. Position beds away from cold drafts.
Space Utilization	Average room size: 250 sq. ft. Recommended bed spacing: 3 ft.	Average room size: 400 sq. ft. Recommended bed spacing: 4 ft.	Average room size: 600 sq. ft. Recommended bed spacing: 5 ft.
Patient Flow and Acuity	Average bed occupancy rate: 80% Peak occupancy period: 10:00 AM - 2:00 PM	Average bed occupancy rate: 85% Peak occupancy period: 9:00 AM - 3:00 PM	Average bed occupancy rate: 90% Peak occupancy period: 8:00 AM - 4:00 PM
Workflow Efficiency	Average time-motion study result: 12 minutes/hour. Staff feedback survey rating: 4.5/5	Average time-motion study result: 10 minutes/hour. Staff feedback survey rating: 4.7/5	Average time-motion study result: 8 minutes/hour. Staff feedback survey rating: 4.8/5
Patient Satisfaction	Average patient satisfaction score: 9/10. Top patient complaint: Lack of privacy	Average patient satisfaction score: 8.5/10. Top patient complaint: Noise	Average patient satisfaction score: 8/10. Top patient complaint: Temperature
Infection Control	HealthCare-associated infection (HAI) rate: 2% Correlation between bed spacing and HAI rate: No significant correlation found	HealthCare-associated infection (HAI) rate: 1.5% Correlation between bed spacing and HAI rate: No significant correlation found	HealthCare-associated infection (HAI) rate: 1% Correlation between bed spacing and HAI rate: No significant correlation found
Emergency Preparedness	Number of emergency evacuation routes: 3	Number of emergency evacuation routes: 4	Number of emergency evacuation routes: 5

RESULTS

Our case study and analysis has revealed important findings. We have used ChatGPT to simulate a sample size of 5000 patients with the randomized data. We have used Python's matplotlib to plot the results. We have plotted the same by different age group vs hospital sizes in a graph format, bed positioning, infection rate, energy saving by season, satisfaction, accessibility, patient flow.

Patient Bed Placement Analysis:

Hospital Bed Positioning Strategies - Pediatric (0-12)

Parameters	Pediatric (0-12)	Pediatric (0-12)
Hospital Size	Medium	Small
Room Size	Pediatric	Pediatric
Bed Positioning Strategy	Install adjustable bed rails for safety	Use colorful and engaging decoration to create a welcoming environment
Heating	Use space heater to maintain warmth in pediatric rooms	Use radiant floor heating for even warmth
Cooling	Ensure adequate ventilation to prevent overheating	Provide oscillating fans or cooling as needed
Infection Rate	Moderate	Low

Hospital Bed Positioning Strategies - Teen (13-18)

Parameters	Teen (13-18)
Hospital Size	Medium
Room Size	Standard
Bed Positioning Strategy	Provide Ergonomic Furniture for Studying and relaxation
Heating	Ensure adequate heating for comfortable living
Cooling	Maintain Comfortable room temperature during summer
Infection Rate	Low

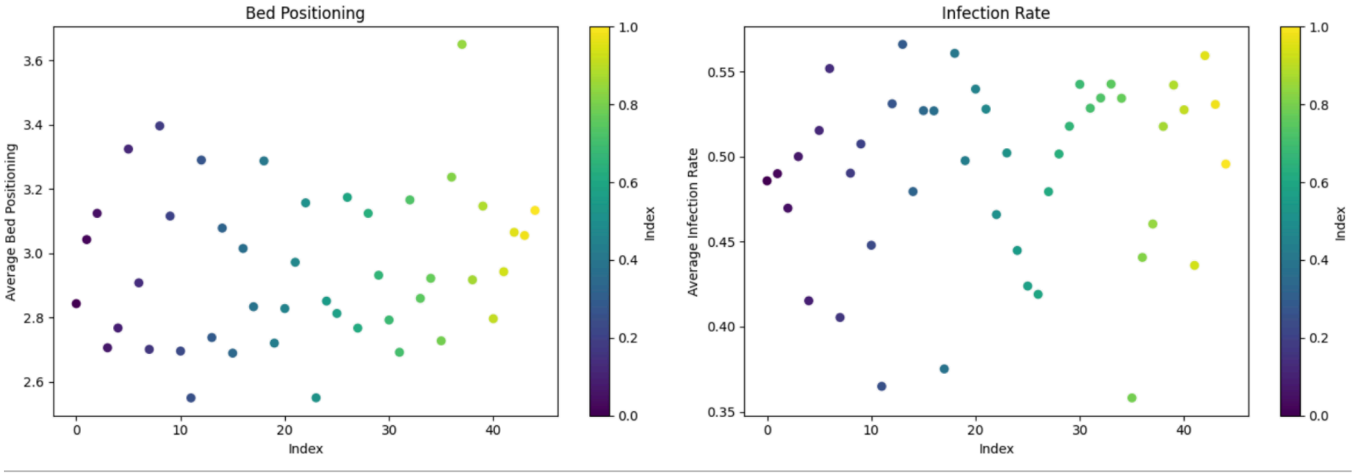
Hospital Bed Positioning Strategies - Adult (18-65)

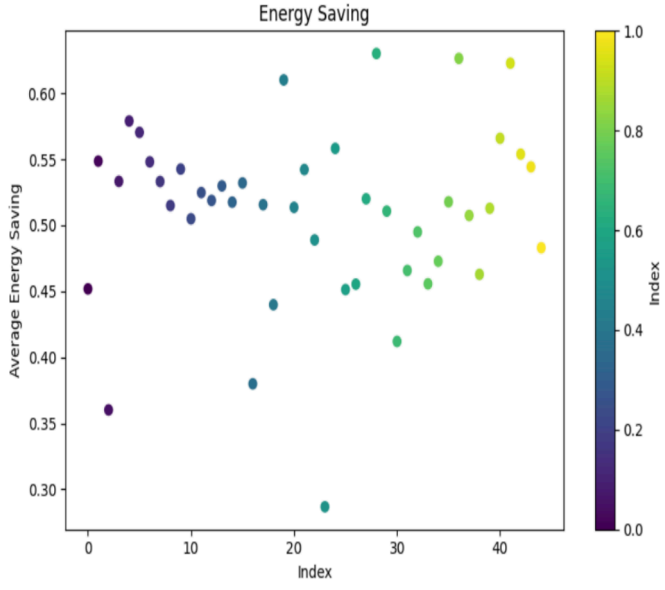
Parameters	Adult (18-65)	Adult (18-65)
Hospital Size	Small	Small
Room Size	Standard	Standard
Bed Positioning Strategy	Please bed near window for better ventilation and natural light	Avoid placing bed near high-traffic area to minimize noise
Heating	Use central heating system to maintain optimum temperature	Ensure proper insulation to refrain heat in colder months
Cooling	Install ceiling fan for additional air circulation	Provide portable air conditioner for cooling in warmer months
Infection Rate	Low	Low

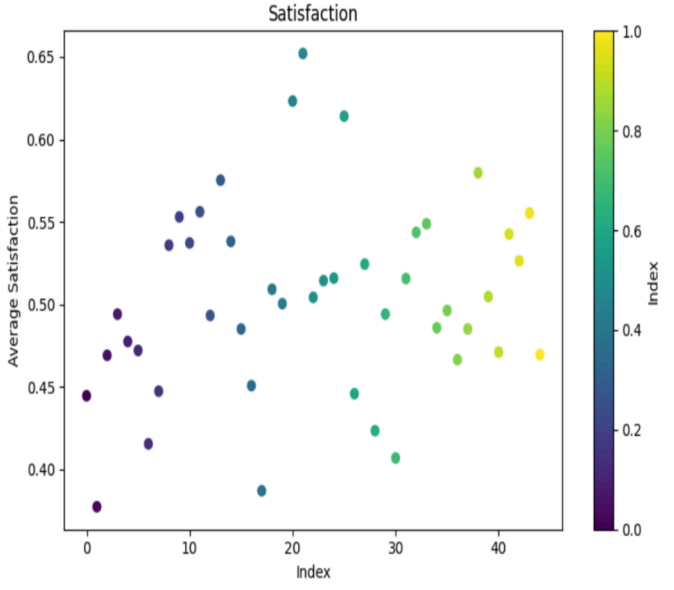
Hospital Bed Positioning Strategies - Elderly (65+)

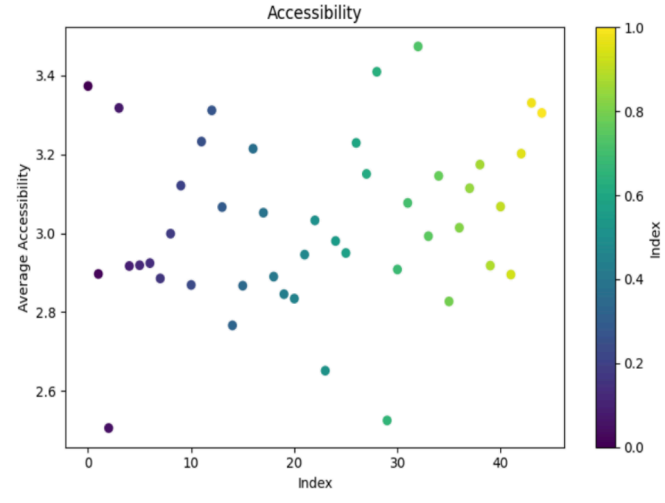
Parameters	Elderly 65+
Hospital Size	Large
Room Size	Standard
Bed Positioning Strategy	Ensure Easy Access to bathroom facilities
Heating	Install Thermostat-controlled heating for individual room comfort
Cooling	Implement cross ventilation to improve air quality
Infection rate	Moderate

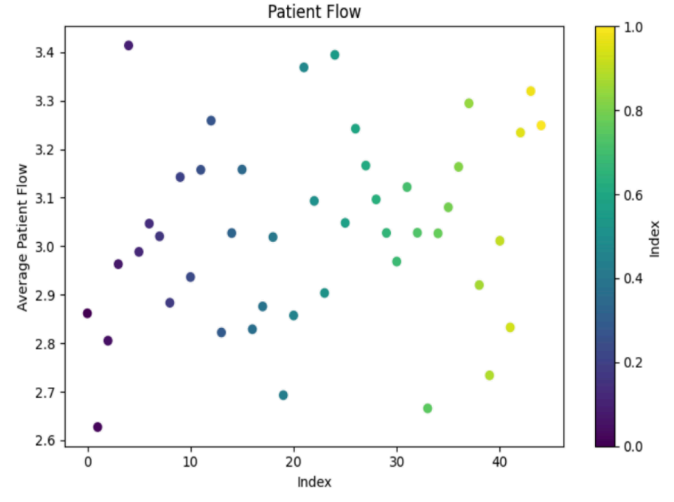
Scatter plots:



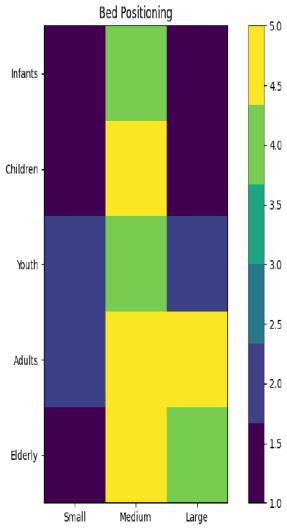


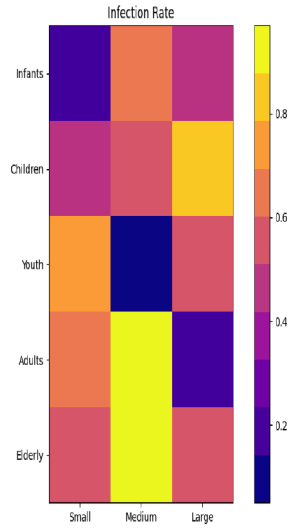


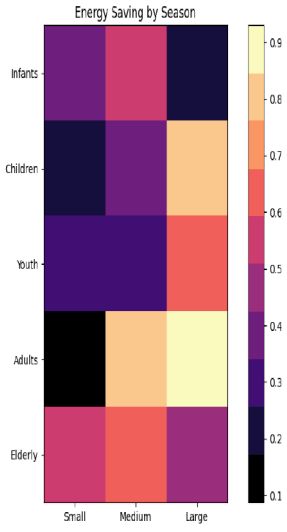


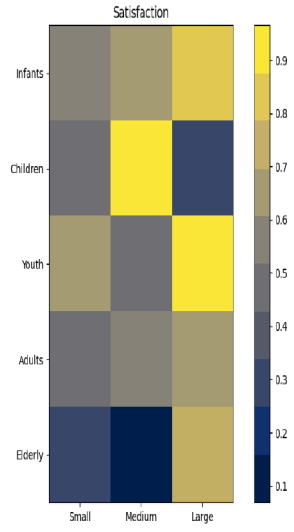


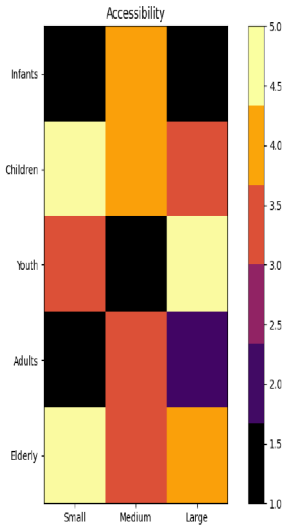
Heatmaps

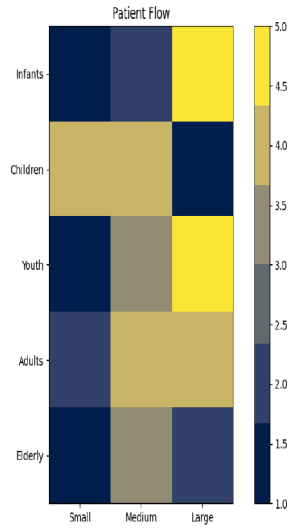












Interpretation: Darker colors usually indicate higher values or frequencies, while lighter colors indicate lower values or frequencies.

Patterns: Look for any patterns or trends in the heatmap. For example, do certain age groups consistently have higher satisfaction scores than others? Are there any abrupt changes or gradients in color intensity that might indicate differences between age groups or other variables?

Correlations: If you're comparing multiple variables in the heatmap, pay attention to how the colors change across different combinations of variables. Are there any areas of high correlation (e.g., darker regions) or low correlation (e.g., lighter regions) between variables?

Context: Finally, consider the context of our data.

Analysis and Findings:

Bed position: There is a high frequency for a small hospital across all age groups.

Infection Rate: The frequency is high for infants in small hospital, for children it is high in medium hospital and for adults it is high in large hospital.

Energy saving: The frequency is high for infants, children, and youth for small, medium, and large hospitals.

Satisfaction: For Children the frequency is higher in large hospital and for elderly it higher for small and medium hospital.

Accessibility: The frequency is higher for infants in small and large hospital, for youth it is higher in small and medium hospital whereas for adult is higher in small, medium, and large hospital

Patient Flow: The frequency is higher for infants, youth, adults, and elderly for a small hospital whereas it is higher for children in large hospital.

Our research is conducted using ChatGPT 3.5 to simulate the use case.

The data shown above is for demonstration purpose only. It does not represent real-world data from the internet or any specific dataset.

The purpose of the sample data was to showcase how to create a summary table and plot the data using Python, allowing users to visualize and analyze the relationships between different demographic and facility characteristics and various metrics such as bed positioning, infection rate, energy saving, satisfaction, accessibility, and patient flow.

If you have real-world data that you'd like to analyze or visualize using similar techniques, you can adapt the code to work with your dataset and explore the relationships between different variables in your specific context.

CONCLUSION

In this paper, we have researched and documented on how ChatGPT can assist in solving the problem in HealthCare Sectors with the assumption that real datasets are available for thorough analysis and problem determination for a real use case. We can also solve the problem with simulated data. Especially in HealthCare Sectors as the data is sensitive and cannot be shared. The ChatGPT assists in generating random data to provide the analysis that needs to be validated with real time data for a specific use case.

Finally, we would like to conclude that from our key findings through the use case that Generative AI (ChatGPT) can solve the problem conceptually when compared to traditional approach of solving the problem. In traditional research approach data gathering is the real issue, but ChatGPT makes it easier using randomized data. The techniques such as prompt engineering and prompt tuning helps us in providing key inputs to ChatGPT and to obtain desired output to solve a problem with our actual Data sets (Simulating Big Data Environment). The accuracy depends on how the Large Language Model is trained on and the types of data sets it uses. In our findings, we have found that ChatGPT 3.5 can generate random data and assist in solving a problem. We have also verified if the data provided by ChatGPT is from Internet by asking ChatGPT this question. And when we have asked the similar question with ChatGPT and the answer was no and it has confirmed that the data is randomly generated to simulate our use case.

As we conclude, the AI can assist humans in solving problems provided that we all as a good citizen take equal responsibilities, implement and adopt guardrails to protect data, maintain privacy, adhere to the local laws and regulations for adopting and building AI assisted ecosystem [10] by advancing future generations with more AI capabilities. As big data continues to grow, its role in advancing Generative AI will become increasingly significant, and it is opening new frontiers for AI capabilities [11] and applications.

REFERENCES

[1] J. Gu, “Research on Precision Marketing Strategy and Personalized Recommendation Method Based on Big Data Drive,” Wirel. Commun. Mob. Comput., vol. 2022, p. e6751413, Apr. 2022, doi: 10.1155/2022/6751413.

[2] M. Sterling, “Situated big data and big data analytics for healthcare,” in 2017 IEEE Global Humanitarian Technology Conference (GHTC), Oct. 2017, pp. 1–1. doi: 10.1109/GHTC.2017.8239322.

[3] D. Patel, S. Kadbhane, M. Sameed, A. Chandorkar, and A. Rumale, “Prompt Engineering Using Artificial Intelligence,” IJARCCE, vol. 12, Oct. 2023, doi: 10.17148/IJARCCE.2023.121018.

[4] H. Gui, R. Zheng, C. Ma, H. Fan, and L. Xu, “An Architecture for Healthcare Big Data Management and Analysis,” Nov. 2016, pp. 154–160. doi: 10.1007/978-3-319-48335-1_17.

[5] T. Ahmed, C. Bird, P. Devanbu, and S. Chakraborty, “Studying LLM Performance on Closed- and Open-source Data.” arXiv, Feb. 23, 2024. Accessed: Mar. 30, 2024. [Online]. Available: http://arxiv.org/abs/2402.15100

[6] sravan, “Improving Model Performance from 99.9% to 99.999999%,” Hyperspec AI. Accessed: Mar. 30, 2024. [Online]. Available: https://hyperspec.ai/improving-model-performance-from-99-9-to-99-999999/

[7] Q. Cai, H. Wang, Z. Li, and X. Liu, “A Survey on Multimodal Data-Driven Smart Healthcare Systems: Approaches and Applications,” *IEEE Access*, vol. 7, pp. 133583–133599, 2019, doi: 10.1109/ACCESS.2019.2941419.

[8] A. Baig, “Healthcare Compliance Laws and Regulations: Overview,” Securiti. Accessed: Mar. 30, 2024. [Online]. Available: https://securiti.ai/blog/healthcare-compliance-laws/

[9] A. Alem and S. Kumar, “Deep Learning Models Performance Evaluations for Remote Sensed Image Classification,” *IEEE Access*, vol. 10, pp. 111784–111793, 2022, doi: 10.1109/ACCESS.2022.3215264.

[10] P. Sarao, M. Milind, G. N. P. V. Babu, R. Rameshbhai Savaliya, M. Devi, and M. Tiwari, “Hybrid Artificial Ecosystem Optimization Algorithm based on Search Manager Framework for Big Data Environment,” in 2023 Third International Conference on Artificial Intelligence and Smart Energy (ICAIS), Feb. 2023, pp. 892–897. doi: 10.1109/ICAIS56108.2023.10073919.

[11] “Artificial Intelligence in Its Many Forms Will Be the Most Important Area of Technology in 2024, According to New IEEE Global Survey of CIOs, CTOs, and Technology Leaders.” Accessed: Mar. 30, 2024. [Online]. Available: https://www.ieee.org/about/news/2023/news-release-2023-survey-results.html

AUTHOR INFORMATION

Sanjeev Kumar Marimekala, STSM and Thought Leader, IBM, USA

Robert Epstein, Leader, Hybrid Cloud Distributed Automation Optimization & SRE, IBM, USA

John Lamb, PhD, Adjunct Faculty, Mathematics, Pace University, Pleasantville, NY, USA

Vasundhara Bhupathi, DT Portfolio Manager Applications, DXC Technology, USA