Digital Health and Informatics Innovations for Sustainable Health Care Systems
J. Mantas et al. (Eds.)

doi:10.3233/SHTI240469

Evaluating ChatGPT 4.0's User Satisfaction Among Doctors Across Different Medical Departments

Loukas TRIANTAFYLLOPOULOS[a], Georgios FERETZAKIS[a,1], Lazaros TZELVES[b], Aikaterini SAKAGIANNI[c], Vassilios S. VERYKIOS[a] and Dimitris KALLES[a]
[a] *School of Science and Technology, Hellenic Open University, Patras, Greece*
[b] *Second Department of Urology, National and Kapodistrian University of Athens, Sismanogleio General Hospital, Athens, Greece*
[c] *Sismanogleio General Hospital, Intensive Care Unit, Marousi, Greece*

Abstract. In an era increasingly focused on integrating Artificial Intelligence (AI) into healthcare, the utility and user satisfaction of AI applications like ChatGPT have become pivotal research areas. This study, conducted in Greece, engaged 193 doctors from various medical departments who interacted with ChatGPT 4.0 through a custom web application. The participants, representing a diverse range of medical specialties, received responses from the specific chatbot tailored to their specific departmental inquiries. Their satisfaction was gauged using a validated form featuring a 1-to-5 rating scale. The results highlighted a possible correlation between the doctors' medical departments and their satisfaction levels with ChatGPT 4.0. Significantly, doctors from certain departments (like General Surgery and Cardiology) reported lower satisfaction scores, ranging from 2.73 to 2.80 out of 5, in contrast to their colleagues from departments like Biopathology and Orthopedics, who scored between 4.00 and 4.46 out of 5. This variation in satisfaction levels underscores the diverse needs within different medical specialties and illuminates both the potential of ChatGPT and the areas needing improvement, especially in delivering department-specific medical information. Despite its limitations, ChatGPT version 4.0 is emerging as a valuable tool in the medical community, indicating potential future advancements and more extensive integration into healthcare practices. The study's findings are crucial in understanding the distinct preferences and requirements of healthcare professionals across various medical departments, thereby guiding the future development of AI tools in healthcare.

Keywords. Artificial intelligence, ChatGPT, Doctor-Chatbot Interaction, Satisfaction, Medical Departments, Large language models, Greece

1. Introduction and Background

ChatGPT has made a significant impact in the healthcare sector [1]. Although it hasn't been trained using a specialized medical database [2], many researchers find its performance in delivering medical information to various nursing departments satisfactory [3-7]. However, a commonly noted limitation of ChatGPT is its performance in languages other than English, which forms the basis of its training [8-10].

[1] Corresponding Author: Georgios FERETZAKIS, Ph.D., School of Science and Technology, Hellenic Open University, Patra 263 35, Greece; E-mail: georgios.feretzakis@ac.eap.gr

In our study, we aim to assess the quality of information provided by ChatGPT on topics relevant to various nursing departments, specifically in Greek, with the assistance of doctors. We use a validated evaluation tool to analyze the satisfaction levels of several doctors across Greece regarding ChatGPT's accuracy, clarity of information, response time, and overall interaction. Additionally, the research seeks to determine whether there is a correlation between the doctors' overall satisfaction and the specific nursing department related to the evaluated information.

2. Materials and Methods

To evaluate physicians' satisfaction with ChatGPT across various medical specialties, a case study method was employed within the Greek healthcare context. A custom web application integrating HTML, JavaScript, PHP, CSS, and a secure MySQL database was developed for data collection and participant confidentiality. Doctors submitted their professional details and interacted with the virtual assistant under evaluation, rating their satisfaction on a scale from 1 to 5 based on criteria such as accuracy, clarity, relevance, response time, and overall satisfaction.

2.1. Participant selection

To ensure broad geographic and specialty representation, our study targeted a diverse sample of certified medical professionals by reaching out to all 63 medical associations in Greece. With no specific exclusion criteria applied, the aim was to include a wide spectrum of experiences and insights from the medical profession. Conducted from November to December 2023, the study encompassed 193 doctors from 27 medical departments across 35 Greek prefectures.

2.2. Evaluation Tool and Data Analysis

The evaluation criteria for the study were developed following an extensive literature review on information accuracy, response time in healthcare, and user satisfaction across various medical fields, such as surgery [10], dermatology [11], and emergency medicine [12,13]. To ensure comprehensive content validity, our criteria were meticulously designed to encompass all relevant dimensions of user satisfaction, drawing upon existing frameworks and expert consultations. Additionally, the reliability and validity of the assessment tool were rigorously tested. Internal consistency was confirmed using Cronbach's Alpha, and inter-item relationships were examined through a correlation matrix. Construct and criterion-related validity were further established through detailed factor analysis and correlation tests, verifying the tool's capability to accurately measure user satisfaction.

Data analysis was conducted using Python 3.7, utilizing libraries such as Pandas, Matplotlib, Seaborn, numpy, and FactorAnalyzer to facilitate the processing and visualization of data. Pandas was employed to compute detailed descriptive statistics, providing insights into the distribution and central tendencies of the data. Visualization tools in Matplotlib and Seaborn were used to depict trends and correlations, enhancing the interpretability of the data. The numpy library was essential for numerical computations, while FactorAnalyzer was used for conducting factor analysis to explore underlying variables.

To ensure the integrity of our analysis, 13 of the 193 initial evaluations were excluded due to incomplete data, resulting in a dataset of 180 comprehensive responses. Given the variability in sample sizes across different medical specialties, particularly in subgroups with fewer than ten participants, we used non-parametric statistical methods suited for these conditions. The Kruskal-Wallis test was employed to assess differences in satisfaction ratings across departments. For pairwise comparisons among departments with sufficient data, we conducted Mann-Whitney U tests and applied the Bonferroni correction to adjust the significance level, controlling the risk of type I errors.

3. Results

Table 1 initially presents the results of the validity and reliability analysis tests for the assessment tool used in the research. The factor analysis confirmed the tool's construct validity by revealing two significant factors that demonstrate distinct dimensions of satisfaction. The criterion-related validity was robust, as evidenced by strong correlations between satisfaction and related ratings. Additionally, the internal consistency of the tool, measured by Cronbach's Alpha, was high (>0.7), indicating reliable measurements across different items.

Table 1. Summary of Validity and Reliability Results

Statistical Measure	Value	Description
Bartlett's Test p-value	1.305 x 10^-93	Suitable data for factor analysis.
KMO Measure	0.776	Adequate sampling adequacy for factor analysis.
Factor Eigenvalues	2.858, 0.752	
Cronbach's Alpha	0.855	Good reliability.
Correlation: Satisfaction-Accuracy	r=0.762, p<0.001	Strong positive correlation
Correlation: Satisfaction - Clarity	r=0.790, p<0.001	Strong positive correlation
Correlation: Satisfaction - Response Time	r=0.482, p<0.001	Moderate positive correlation

Table 2 presents the average evaluation scores for ChatGPT across different medical departments, with a minimum sample size of seven participants per department.

Table 2. Descriptive Statistics of Evaluation Scores by Medical Department ($n \geq 7$)

Medical Department	Sample	Accuracy Mean *(SD)*	Response Time Mean *(SD)*	Clarity Mean *(SD)*	Total Satisfaction Mean *(SD)*
Biopathology	8	3,75 (*1,28*)	4,50 (*1,07*)	3,63 (*1,19*)	4,00 (*0,93*)
Cardiology	10	3,30 (*1,25*)	3,80 (*1,23*)	3,00 (*1,56*)	2,80 (*1,48*)
Gastroenterology	7	3,00 (*1,63*)	3,43 (*1,81*)	3,00 (*1,91*)	3,29 (*1,70*)
Internists	29	3,90 (*1,01*)	3,86 (*1,13*)	4,07 (*1,00*)	3,79 (*0,98*)
Neurology	7	3,57 (*1,40*)	3,86 (*1,35*)	3,29 (*1,38*)	3,43 (*1,40*)
Pediatric Pathology	7	3,71 (*0,76*)	3,57 (*0,79*)	3,00 (*1,41*)	3,00 (*1,00*)
General Surgery	15	2,93 (*1,39*)	3,40 (*1,40*)	2,67 (*1,29*)	2,73 (*1,03*)
Gynaecology	7	3,43 (*1,13*)	4,00 (*1,41*)	3,29 (*1,11*)	3,57 (*0,98*)
Orthopedics	13	4,23 (*0,60*)	4,54 (*0,78*)	4,23 (*0,83*)	4,46 (*0,66*)
Urology	28	3,29 (*1,41*)	3,71 (*1,15*)	3,11 (*1,45*)	3,21 (*1,32*)

Orthopedics consistently ranks highest in all categories, with average scores above 4.20, while Biopathology performs well, particularly in response time and total satisfaction, both averaging over 4.0. Internists also show strong results, notably in

clarity. In contrast, General Surgery scores the lowest across all criteria, with total satisfaction averaging just 2.73. Cardiology, despite a decent sample size, registers lower overall satisfaction at 2.80. Other departments, including Gastroenterology, Neurology, Pediatric Pathology, Gynaecology, and Urology, display moderate scores; both Gastroenterology and Pediatric Pathology average above 3.0 in total satisfaction.

In the analysis of overall satisfaction ratings across 27 different medical departments, the ANOVA test yielded a statistic of 1.6291, indicating a moderate degree of variance between groups. With a p-value of 0.0492, slightly below the conventional threshold of 0.05, there appears to be a statistically significant difference in the satisfaction ratings among the departments.

4. Discussion

The academic community frequently assesses the reliability of information provided by ChatGPT, a Large Language Model (LLM), particularly in relation to medical issues across various departments, as referenced in the literature [5-9]. This study evaluated doctor satisfaction with ChatGPT's information quality across 27 different medical departments using defined criteria. Generally, the results suggest that ChatGPT provides satisfactory information, consistent with findings from other researchers like Nielsen et al. [14] and Kaare et al. [7]. For instance, Nielsen et al. reported satisfactory accuracy in an otolaryngology ward, with overall scores of 3.51 and category-specific scores of 3.41 out of 5. Kaare et al. evaluated the accuracy in an orthopedic department, finding scores around 4 out of 5 or 1.6 out of 2. Further, ANOVA analysis indicated that satisfaction levels significantly vary by medical department. This variability highlights the uneven extent of ChatGPT's training across different specialties, with some departments receiving more comprehensive and accurate information than others. This disparity underscores the need for AI systems like ChatGPT to continually adapt and evolve to meet the specific challenges and requirements of different medical fields.

However, our study is not without limitations. The focus solely on Greek data may lead to inaccuracies in ChatGPT's responses, a concern echoed in studies involving non-English languages [8,9]. Additionally, the subjective nature of the non-standardized five-point rating scale used for evaluating various parameters could introduce inconsistencies. This issue is also reflected in related research [14-15]. Moreover, participants' awareness that the responses were generated by ChatGPT might have influenced their evaluations, a challenge noted in Kaare et al.'s study [7]. To gain more comprehensive insights, it would be beneficial to expand this study to include a broader medical community and to incorporate newer models of ChatGPT or other LLMs, despite the challenges posed by these limitations.

5. Conclusions

Using a validated assessment tool, we evaluated the satisfaction levels of Greek physicians from 27 different nursing departments during their interactions with ChatGPT 4.0. This evaluation focused on four distinct criteria, examining both average satisfaction levels and variance through ANOVA analysis. Our findings indicate a marginally significant relationship between satisfaction and departmental affiliations. Although most departments described the chatbot's information quality as moderate, the

orthopedics department reported the highest satisfaction, in stark contrast to the lower scores from cardiology and general surgery. The thorough validation of the assessment tool underscores its reliability and effectiveness in measuring physician satisfaction with AI tools in healthcare, suggesting its potential for broader application.

References

[1] Venkataswamy R, Janamala V, Cherukuri RC. Realization of Humanoid Doctor and Real-Time Diagnostics of Disease Using Internet of Things, Edge Impulse Platform, and ChatGPT. Ann Biomed Eng. 2023;1-3. doi: 10.1007/s10439-023-03316-9.

[2] Sarraju A, Bruemmer D, Van Iterson E, Cho L, Rodriguez F, Laffin L. Appropriateness of cardiovascular disease prevention recommendations obtained from a popular online chat-based artificial intelligence model. JAMA. 2023;329(10):842-844. doi: 10.1001/jama.2023.1044.

[3] Manolitsis I, Feretzakis G, Tzelves L, Kalles D, Katsimperis S, Angelopoulos P, et al. Training ChatGPT Models in Assisting Urologists in Daily Practice. Stud Health Technol Inform. 2023 Jun 29;305:576-579. doi: 10.3233/SHTI230562.

[4] Talyshinskii A, Naik N, Hameed BMZ, Zhanbyrbekuly U, Khairli G, Guliev B, et al. Expanding horizons and navigating challenges for enhanced clinical workflows: ChatGPT in urology. Front Surg. 2023 Sep 7;10:1257191. doi: 10.3389/fsurg.2023.1257191.

[5] Alessandri Bonetti M, Giorgino R, Gallo Afflitto G, De Lorenzi F, Egro FM. How does ChatGPT perform on the italian residency admission national exam compared to 15,869 medical graduates?. Ann Biomed Eng. 2023;1-5. doi: 10.1007/s10439-023-03318-7.

[6] Borchert RJ, Hickman CR, Pepys J, Sadler TJ. Performance of ChatGPT on the Situational Judgement Test—A Professional Dilemmas–Based Examination for Doctors in the United Kingdom. JMIR Med Educ. 2023;9(1):e48978. doi: 10.2196/48978.

[7] Kaarre J, Feldt R, Keeling LE, Dadoo S, Zsidai B, Hughes JD, et al. Exploring the potential of ChatGPT as a supplementary tool for providing orthopaedic information. Knee Surg Sports Traumatol Arthrosc. 2023;31(11):5190-5198. doi: 10.1007/s00167-023-07529-2.

[8] Oztermeli AD, Oztermeli A. ChatGPT performance in the medical specialty exam: An observational study. Medicine (Baltimore). 2023;102(32):e34673. doi: 10.1097/MD.0000000000034673.

[9] Use of ChatGPT on Taiwan's Examination for Medical Doctors. Ann Biomed Eng. 2023;1-3. doi: 10.1007/s10439-023-03308-9.

[10] Yeo YH, Samaan JS, Ng WH, Ma X, Ting PS, Kwak MS, et al. GPT-4 outperforms ChatGPT in answering non-English questions related to cirrhosis. medRxiv. 2023-05. doi: 10.1101/2023.05.04.23289482.

[11] Samaan S, Yeo YH, Rajeev N, Hawley L, Abel S, Ng WH, et al. Assessing the accuracy of responses by the language model ChatGPT to questions regarding bariatric surgery. Obes Surg. 2023;1-7. doi: 10.1007/s11695-023-06603-5.

[12] Mondal H, Mondal S, Podder I. Using ChatGPT for writing articles for patients' education for dermatological diseases: A pilot study. Indian Dermatol Online J. 2023;14(4):482-486. doi: 10.4103/idoj.idoj_72_23.

[13] El Dahdah J, Kassab J, El Helou MC, Gaballa A, Sayles III S, Phelan MP. ChatGPT: A Valuable Tool for Emergency Medical Assistance. Ann Emerg Med. 2023. doi: 10.1016/j.annemergmed.2023.04.027.

[14] Nielsen JP, von Buchwald C, Grønhøj C. Validity of the large language model ChatGPT (GPT4) as a patient information source in otolaryngology by a variety of doctors in a tertiary otorhinolaryngology department. Acta Otolaryngol. 2023;143(9):779-782. doi: 10.1080/00016489.2023.2254809.

[15] Anastasio AT, Mills IV FB, Karavan Jr MP, Adams Jr SB. Evaluating the Quality and Usability of Artificial Intelligence–Generated Responses to Common Patient Questions in Foot and Ankle Surgery. Foot Ankle Orthop. 2023;8(4):24730114231209919. doi: 10.1177/24730114231209919.