European Journal of Clinical and Experimental Medicine

ONLINE FIRST

Authors: Ankita Jawanpuria, Aruna Rani Behera, Chinmaya Dash, Mohammad Hifz Ur Rahman

Article type: Original Article

Received: 13 December 2023

Accepted: 5 February 2024

Published online: 18 February 2024

eISSN: 2544-1361

Eur J Clin Exp Med

doi: 10.15584/ejcem.2024.2.19

ChatGPT in hospital infection prevention and control – assessing knowledge of an AI model based on a validated questionnaire

Ankita Jawanpuria [1], Aruna Rani Behera [1], Chinmaya Dash [1], Mohammad Hifz Ur Rahman [2]

[1] Department of Microbiology, Manipal Tata Medical College, Manipal Academy of Higher Education, Manipal, India

[2] Department of Community Medicine, Manipal Tata Medical College, Manipal Academy of Higher Education, Manipal, India

Corresponding author: Aruna Rani Behera, e-mail: arunaapril11@gmail.com

ORCID

AJ: https://orcid.org/0009-0005-1994-6554

ARB: https://orcid.org/0000-0002-2451-8766

CD: https://orcid.org/0000-0001-5263-6973

MHUR: https://orcid.org/0000-0002-0039-5837

ABSTRACT

Introduction and aim. An AI model like ChatGPT is a good source of knowledge. We can explore the potential of AI models to complement the expertise of healthcare professionals by providing real-time, evidence-based information in infection prevention and control (IPC).

Material and methods. This study involved 110 queries related to IPC, validated by subject experts in IPC. The responses from ChatGPT were evaluated using Bloom's taxonomy by experienced microbiologists. The scores were divided as <3 as being a poor response, 3–4 as an average response, and >4 as a good response. Statistical analysis was done by correlation coefficient and Cohen's Kappa.

Results. The overall score was 4.33 (95% CI, q1 3.65–q3 4.64) indicating ChatGPT's substantial IPC knowledge. A good response (i.e.>4 score) was found in 70 (63.6%) questions, while in 10 (9%) questions, it showed a poor response. The poor response was seen in needle stick injury and personal protective equipment (PPE) doffing-related questions. The overall correlations were found to be significant. Cohen's Kappa confirmed moderate to substantial agreement between evaluators.

Conclusion. ChatGPT demonstrated a commendable understanding of IPC principles in various domains and the study identifies specific instances where the model may require further refinement especially in critical scenarios such as needlestick injuries and PPE doffing.

Keywords. artificial intelligence, ChatGPT, infection control, large language model, medical education

Introduction

According to a WHO 2022 report, out of every 100 patients in acute-care hospitals, seven patients in high-income countries and 15 patients in low- and middle-income countries acquire at least one healthcare-associated infection (HAI) during their hospital stay. On average, 1 in every 10 affected patients die from their HAI. Infection prevention and control measures play a vital role in maintaining patient safety within healthcare settings. Various studies have found knowledge gaps in healthcare workers regarding IPC.[1]

Many artificial intelligence (AI) models have come out as good knowledge models. AI models like ChatGPT, an advanced language model trained on a vast array of data, are capable of generating human-like responses to a wide range of queries. Several studies have been done to check knowledge of these AI models in various healthcare fields.[2-4]

Aim

This study aimed to assess the knowledge of the AI model ChatGPT using a validated questionnaire in the context of hospital infection prevention and control. By evaluating its understanding and ability to provide accurate information related to IPC, we can explore the potential of AI models to complement the expertise of healthcare professionals by providing real-time, evidence-based information in infection prevention and control.

Material and methods

This cross-sectional study was conducted using the ChatGPT AI model (generation 3.5). A series of 110 higher-order reasoning queries were posed to ChatGPT, covering various learning objectives in IPC, such as hand hygiene, HAI prevention, injection safety, antimicrobial resistance, biomedical waste management, environmental cleaning, and disinfection, employee immunization status, high-risk areas, standard precautions, bundle care approach, and central sterile supply department (CSSD).

To ensure the validity and bias of the questionnaire, it was rigorously validated by four subject experts with extensive experience in IPC. The first response generated by ChatGPT for each question was collected and stored in an MS Word file for further analysis.

The collected responses were quantitatively evaluated by three authors of this study. Authors of the paper are experts in the field and intimately familiar with the study's objectives, methodologies, and context. This familiarity can contribute to a nuanced evaluation. Also, the authors have a deep understanding of the intricacies of the study, allowing for a more contextually informed evaluation. The authors were blind to each other’s evaluation and to decrease further bias, correlation coefficient analysis was performed. The evaluators gave zero to five marks to each question. The scores obtained were stored in an MS Excel sheet for analysis.

In this study, the answers generated by ChatGPT were categorized into specific levels of Bloom's taxonomy. Bloom's taxonomy is a hierarchical model that classifies educational learning objectives based on complexity and specificity. Taxonomy comprises six domains ranging from lower to higher levels of cognitive processes: knowledge, comprehension, application, analysis, synthesis, and evaluation. The assessment determined the specific Bloom taxonomy group, to which the answers from ChatGPT belonged.

Statistical analysis

The data was analyzed using software Jamovi version 2.4.5, including numbers, means, medians, standard deviations, and quartiles. Correlation coefficient between different evaluators was measured. Cohen's Kappa was employed to validate the reliability of the evaluators' domain categorizations. $p<0.05$ was taken as significant.

Results

A total of 110 higher-order reasoning queries were posed to the ChatGPT model, all the questions were given marks on a scale from zero to five.

A section-wise score is given in Table 1.

Table 1. Section wise score

S. No.	Section name	Median score	Lower CI*	Upper CI
1	Hand hygiene	4.42	3.48	4.79
2	Standard and transmission-based precautions	4.58	3.88	4.78
3	Personal protective equipment (PPE)	4.17	3.52	4.41
4	Hospital acquired infection	4.25	3.37	4.79
5	CSSD	4.67	3.95	4.85
6	Injection safety needle stick management	4.33	3.05	4.52
7	Antimicrobial stewardship program	4.42	3.97	4.67
8	High risk areas	4.33	4	4.64
9	Infection control policy	4.17	3.68	4.45
10	Disinfection and environmental cleaning	4.08	3.65	4.59

11	Bundle care approach	4.25	3.64	4.39

*CI= Confidence interval

The overall score was found to be 4.33 (95% CI, q1 3.65–q3 4.64) out of five.

We divided the scores as <3 as poor a response, 3-4 as an average response and >4 as a good response. The result of 110 questions was:

Response	Number of questions
Poor	10
Average	30
Good	70

Table 2. Some questions asked to ChatGPT

Questions	**Average marks**	**Answer domain**
If a nurse gets a needle stick injury, what first aid she should take?	2.3	Knowledge
What types of tests are available to detect blood-borne viruses, list all?	4.5	Knowledge
Which parameters determine hepatitis B virus (HBV) infectivity?	3	Knowledge
How we improve hand hygiene compliance among nurses?	4.8	Knowledge
What are the infection control requirements to establish a 4-bed medical Intensive care unit (ICU)?	4.3	Application
How to prevent bacterial contamination of red blood cells (RBC) units in blood banks?	4.6	Application

In hand hygiene questions, such as "How can we improve hand hygiene compliance among doctors?" the provided answer included points such as education and training, leadership commitment, awareness campaigns, easy access to hand hygiene facilities, peer support and accountability, continuous quality improvement, celebrating success, and patient and family involvement. The score given to this answer is 4.8 (Table 2). In questions related to standard and transmission-based precautions, like "What standards should be maintained in a negative pressure room?" the answer encompassed points such as airflow and

pressure, per hour air exchange, filtration, room integrity, directional airflow, and monitoring and maintenance, earning a score of 4.6.

In the HAI section, the question asked was, "What is the infectivity period of measles?" The answer provided was, "The infectivity period of measles refers to the timeframe during which an individual with measles can transmit the virus to others. The infectivity period typically lasts for approximately 4 days before the rash develops until 4 days after the rash appears," resulting in a score of 4.2.

In the CSSD section, the question was, "How can we maintain the air quality of CSSD?" The answer outlined points such as the Heating, Ventilation, and Air Conditioning (HVAC) system, positive pressure, air filtration, segregated airflow, ventilation and exhaust, regular maintenance, monitoring and testing, staff training, and practices, receiving a score of 4.7.

In the injection safety section, a question was asked, "If a nurse experiences a needlestick injury, what first aid should she take?" The answer included points such as first aid, follow-up, and counseling, although the suggestion was given to squeeze the finger (score 2.3).

In the PPE section, a question was asked, "If PPE is visibly soiled while on duty, what should be done before doffing?" The answer suggested all points like hand hygiene, proper doffing, and disposing of contaminated PPE, although there was no mention of removing soiling with an alcohol swab before removing PPE (score 3).

A question regarding infection control policy was asked, "Where should a doctor sit in a respiratory outdoor department (OPD)?" The answer provided points such as physical distancing, proper ventilation, hand hygiene, and the availability of PPE facilities, receiving a score of 4.

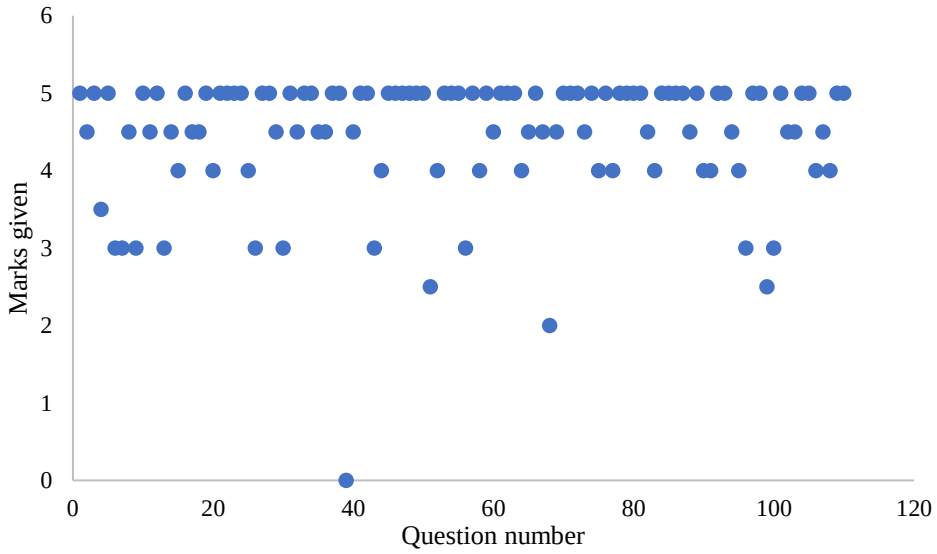

Fig. 1. Score distribution by evaluator 1

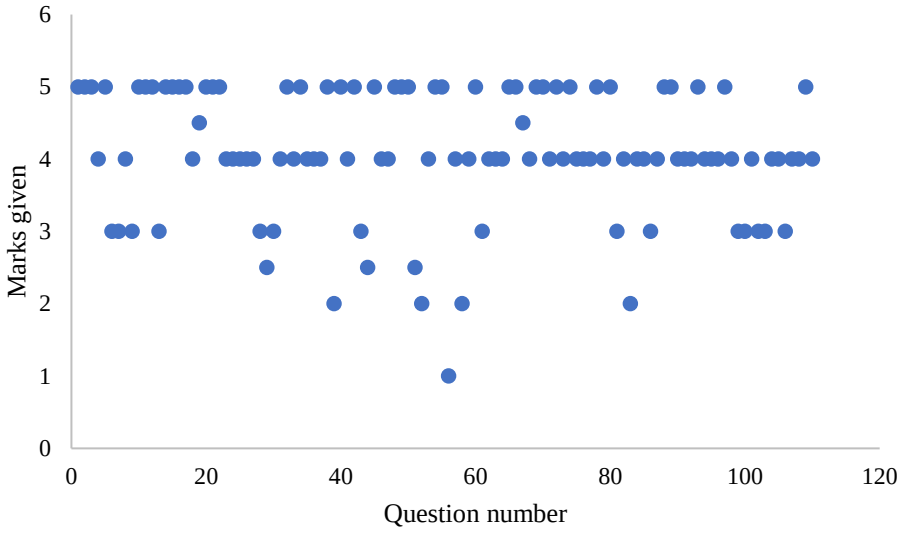

Fig. 2. Score distribution by evaluator 2

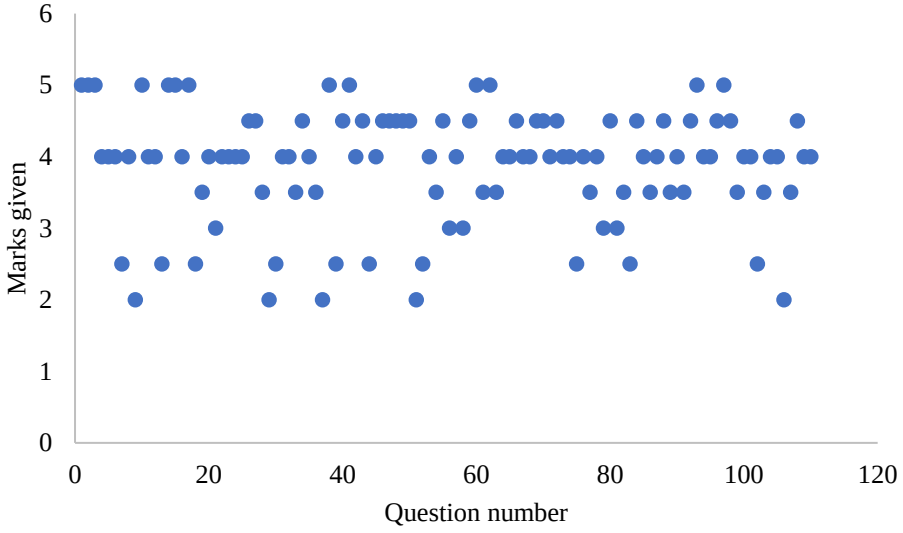

Fig. 3. Score distribution by evaluator 3

In Figure 1, 2 and 3, score distributions by evaluators are shown where, x axis shows question number and the y axis denotes marks given.

The marks given by the three evaluators were different, so to decrease bias and subjectivity, correlation coefficient was performed. The overall correlations between Evaluator 1 and Evaluator 2, Evaluator 2 and Evaluator 3, and Evaluator 3 and Evaluator 1 were found to be 0.55, 0.67, and 0.39, respectively (Fig. 4). These values provide an overview of the general agreement between each pair of evaluators across the entire set of questions.

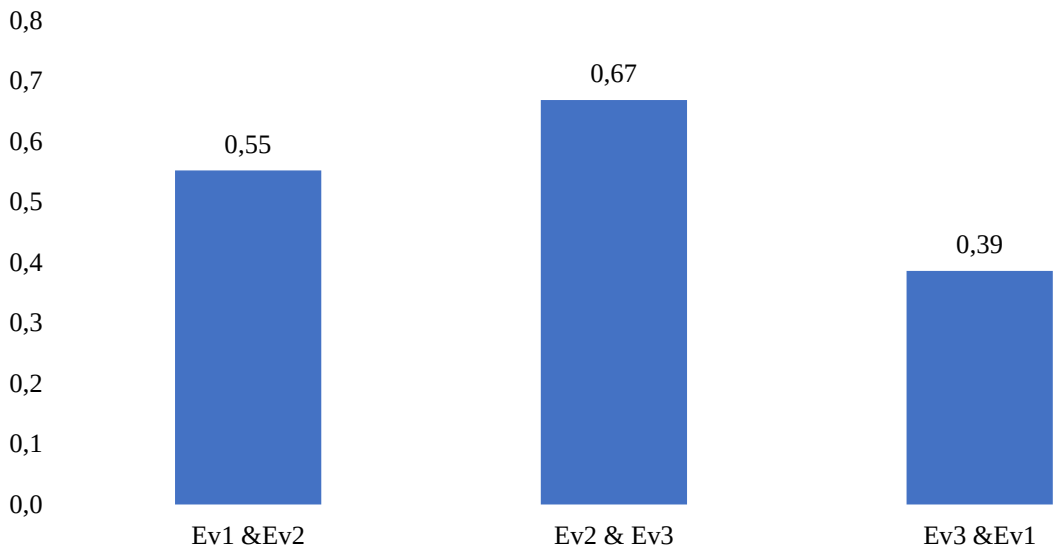

Fig. 4. Correlation coefficient between different evaluators (Ev)

To assess the consistency between different pairs of evaluators, correlations were calculated for all 11 sections of the questions (Fig. 5).

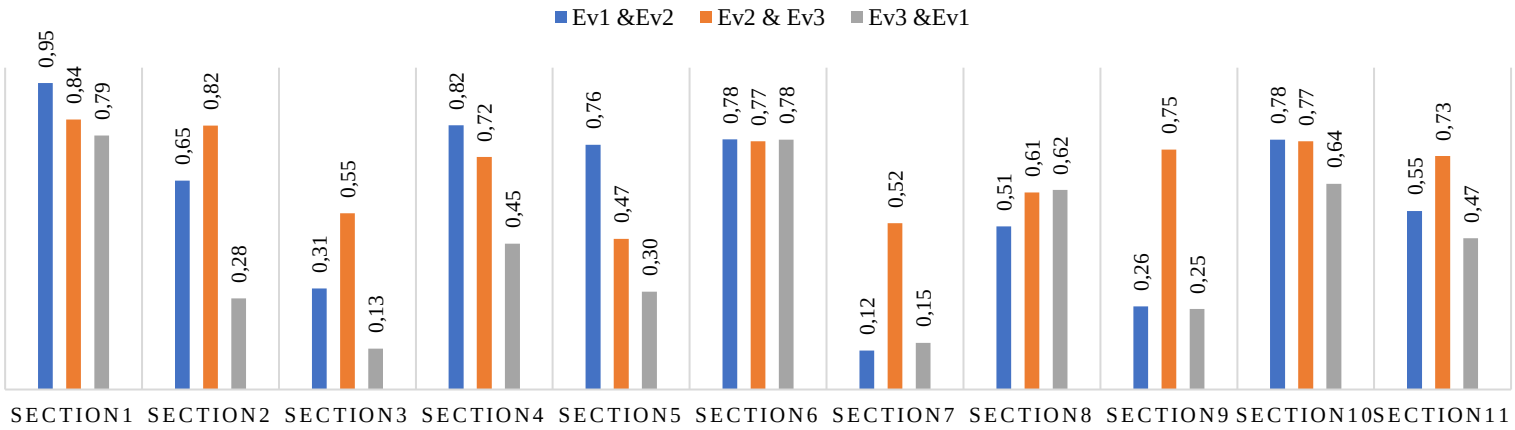

Fig. 5. Correlation coefficient between different evaluators for different sections (Ev – evaluator)

Overall, the highest correlation was observed between Evaluator 2 and Evaluator 3, indicating the closest agreement between these two evaluators. In contrast, the correlation between Evaluator 3 and Evaluator 1 was relatively lower. While some sections demonstrated strong consensus and agreement among all evaluators, others like the antimicrobial stewardship program (section 7) and infection control policy (section 9) showed mixed results with moderate to low agreement.

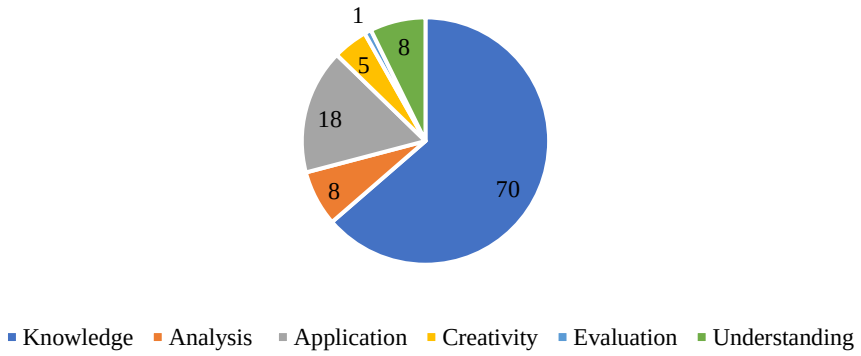

Fig. 6. Categorical analysis of answers

As shown in figure 6, out of 110, 70 answers were in the knowledge domain and 18 in the application domain. Rest questions showed creativity, understanding, analysis, and evaluation-based answers.

Table 3. Cohen's Kappa, to validate the reliability of the evaluators' domain categorizations*

Qualitative criteria	Kappa agreement				
Ev 1 & Ev 2					
Agreement	Expected agreement	Kappa	Std. Err.	Z	p>Z
67.27%	44.14%	0.4141	0.0549	7.55	0
Ev2 & Ev3					
Agreement	Expected agreement	Kappa	Std. Err.	Z	p>Z
76.36%	46.36%	0.5594	0.0563	9,93	0
Ev3 & Ev1					
Agreement	Expected agreement	Kappa	Std. Err.	Z	p>Z
59.09%	45.80%	0.2452	0.056	4.38	0

* Ev – evaluator, Std. Err. – standard error, Z – Z score

To validate the reliability of the evaluators' domain categorizations, Cohen's Kappa was employed. Cohen's kappa is a quantitative measure of reliability for two raters who are rating the same thing, correcting for how often the raters may agree by chance. These insights have practical implications for improving the validity of the evaluation process and enhancing the overall quality of the study's findings. For evaluator 1 and evaluator 2, the Kappa value of 0.41 indicated a moderate level of agreement beyond chance, with 67% of their domain categorizations matching. Similarly, for evaluator 2 and evaluator 3, the Kappa value of 0.55 signified substantial agreement beyond chance, with 76% of their domain categorizations matching. For evaluator 3 and evaluator 1, the Kappa value of 0.24 indicated a fair level of agreement beyond chance, with 59% of their domain categorizations matching (Table 3).

The significant (low p>Z) values of 0.00001 for all three pairs of evaluators underscored the statistical significance of the observed agreements, reinforcing the reliability of their domain categorizations.

Discussion

The results of this research study provide valuable insights into the knowledge base of ChatGPT in IPC, the overall score was found to be 4.33 (95% CI, q1 3.65–q3 4.64). The median score was almost similar in all the sections. A good response (i.e. >4 score) was seen in 70 (63.6%) questions while in 10 (9%) questions, it showed poor response.

ChatGPT consistently provided commendable responses to inquiries, with exemplary instances attached as annexures to this paper. Noteworthy is its comprehensive guidance when prompted for infection control measures to establish a 4-bedded medical ICU, where it emphasized adherence to "Universal precautions." These responses showcase the model's adeptness in offering practical and informed recommendations for healthcare scenarios. The model provided insightful and well-structured answers to questions related to hand hygiene, standard and transmission-based precautions, HAI, CSSD, injection safety, PPE, and infection control policies.

In the context of hand hygiene, ChatGPT outlined a comprehensive strategy to improve compliance among doctors, covering crucial aspects such as education, training, leadership commitment, awareness campaigns, facility accessibility, peer support, accountability, continuous quality improvement, and patient and family involvement. In the section on HAI, ChatGPT accurately conveyed information regarding the infectivity period of measles, underlining the critical timeframe during which an individual with measles can transmit the virus.

However, the study identified a notable discrepancy in the injection safety section, where ChatGPT suggested squeezing the finger as part of first aid for a nurse experiencing a needlestick injury, contrary to established protocols. Similarly, in the PPE section, ChatGPT omitted a crucial step related to decontamination with an alcohol swab before removing visibly soiled PPE. This indicates a nuanced gap in the model's understanding of detailed protocols.

In studies done by Sinha et al. and Ghosh et al. in pathology and biochemistry respectively, the score was 4.08 and 4.0, which is similar to our score.[3,7] Studies have been done to solve image-based queries in pathology, ophthalmology, and dermatology using deep learning and convolutional neural networks (CNN).[8–10] With the help of technical expertise, we can also make deep learning networks to solve complex IPC problems.[11-13]

The overall correlations between different pairs of evaluators for the 110 questions were calculated, indicating agreement among the evaluators was on the positive side. While some sections demonstrated strong consensus and agreement among all evaluators, others like the antimicrobial stewardship program and infection control policy showed mixed results with moderate to low agreement. Although Sinha et al. and Ghosh et al. in their studies showed good inter-rater comparability which can be due to objective answer type questions.[3,7] In our study, the questions were based on routine healthcare activities, as well as real life scenarios, and evaluators were blind towards each other's evaluation.

While doing categorical analysis, most of the answers are in the knowledge domain followed by the application domain. This correlates with the fact that this is mainly a knowledge model which can show human-like responses.

To validate the reliability of the evaluators' domain categorizations, Cohen's Kappa was employed, yielding Kappa values of 0.41, 0.55, and 0.24 for the different pairs of evaluators. These Kappa values indicated moderate to substantial agreement beyond chance, with statistically significant results.

Study limitations

The study also highlights some limitations and areas for improvement. It was a questionnaire-based study which encompassed inquiries related to diverse facets of infection control. It's important to note that the nature of the queries could vary depending on the specific context of hospital settings. ChatGPT belongs to the category of large language models (LLMs), characterized by their capacity to update their knowledge base consistently. There were some other limitations like lengthy answers given for all types of questions and ChatGPT was unable to provide references, and guidelines for its source of information. Apart from that, we couldn't compare this model with other AI models, which can be done in further studies in this area.

Conclusion

AI models like ChatGPT are a good source of knowledge. Overall, ChatGPT demonstrated a commendable understanding of IPC principles in various domains like hand hygiene, standard precautions, hospital acquired infections, infection control policy, etc. The study identifies specific instances where the model may require further refinement to align consistently with established protocols, especially in critical scenarios such as needlestick injuries and PPE doffing. The findings underscore the importance of ongoing

model training and validation to enhance its reliability in providing accurate and contextually appropriate information in healthcare settings. The integration of AI models like ChatGPT in IPC practices could enhance patient safety and overall outcomes by providing healthcare professionals with reliable and up-to-date information. It can complement the expertise of healthcare professionals and support decision-making processes in real time.

Declarations

Funding

This research did not receive support from any funding agencies.

Author contributions

Conceptualization, A.J., A.R.B. and C.D ; Methodology, A.R.B. and A.J.; Software, M.H.U.R.; Validation, A.J., A.R.B. and C.D. ; Formal Analysis, A.J. and M.H.U.R.; Investigation, A.J. and A.R.B; Resources, A.J. and A.R.B; Data Curation, A.J. and A.R.B; Writing – Original Draft Preparation, A.J. and A.R.B., M.H.U.R.; Writing – Review & Editing, A.J., A.R.B. and M.H.U.R.; Visualization, A.J., A.R.B. and C.D Supervision, A.J. and A.R.B. Project Administration, A.J. and A.R.B.

Conflicts of interest

The authors declare no conflict of interest.

Data availability

The data sets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Ethics approval

The study was approved by Institutional Ethical Committee with No. MTMC/IEC/2023/11.

References

1. Alhumaid S, Al Mutair A, Al Alawi Z, et al. Knowledge of infection prevention and control among healthcare workers and factors influencing compliance: a systematic review. *Antimicrob Resist Infect Control.* 2021;10(1):86. doi: 10.1186/s13756-021-00957-0
2. Sallam M. ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. *Healthcare (Basel).* 2023;11(6):887. doi: 10.3390/healthcare11060887

3. Sinha RK, Deb Roy A, Kumar N, Mondal H. Applicability of ChatGPT in Assisting to Solve Higher Order Problems in Pathology. *Cureus*. 2023;15(2):35237. doi:10.7759/cureus.35237
4. Lahat A, Shachar E, Avidan B, Shatz Z, Glicksberg BS, Klang E. Evaluating the use of large language model in identifying top research questions in gastroenterology. *Sci Rep*. 2023;13 (1):4164. doi: 10.1038/s41598-023-31412-2
5. Ponsford MJ, Ward TJC, Stoneham SM, et al. A Systematic Review and Meta-Analysis of Inpatient Mortality Associated with Nosocomial and Community COVID-19 Exposes the Vulnerability of Immunosuppressed Adults. *Front Immunol*. 2021;12:744696. doi: 10.3389/fimmu.2021.744696
6. Fitzpatrick F, Doherty A, Lacey G. Using Artificial Intelligence in Infection Prevention. *Curr Treat Options Infect Dis*. 2020;12(2):135-144. doi: 10.1007/s40506-020-00216-7
7. Ghosh A, Bir A. Evaluating ChatGPT's Ability to Solve Higher-Order Questions on the Competency-Based Medical Education Curriculum in Medical Biochemistry. *Cureus*. 2023;15(4):37023. doi: 10.7759/cureus.37023
8. Khosravi P, Kazemi E, Imielinski M, Elemento O, Hajirasouliha I. Deep Convolutional Neural Networks Enable Discrimination of Heterogeneous Digital Pathology Images. *EBioMedicine*. 2018;27:317-328. doi: 10.1016/j.ebiom.2017.12.026
9. Abràmoff MD, Lou Y, Erginay A, et al. Improved Automated Detection of Diabetic Retinopathy on a Publicly Available Dataset Through Integration of Deep Learning. *Invest Ophthalmol Vis Sci*. 2016;57(13):5200-5206. doi: 10.1167/iovs.16-19964
10. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*. 2017;542(7639):115-118. doi: 10.1038/nature21056
11. Classen DC, Longhurst C, Thomas EJ. Bending the patient safety curve: how much can AI help? *NPJ Digit Med*. 2023;6(1):2. doi: 10.1038/s41746-022-00731-5
12. Reverberi C, Rigon T, Solari A, et al. Experimental evidence of effective human-AI collaboration in medical decision-making. *Sci Rep*. 2022;12(1):14952. doi: 10.1038/s41598-022-18751-2
13. Khanna NN, Maindarkar MA, Viswanathan V, et al. Economics of Artificial Intelligence in Healthcare: Diagnosis vs. Treatment. *Healthcare (Basel)*. 2022;10(12):2493. doi: 10.3390/healthcare10122493