Montiel-Romero *et al. BMC Infectious Diseases* (2025) 25:38
https://doi.org/10.1186/s12879-024-10426-9

BMC Infectious Diseases

RESEARCH

Open Access

Recommended antibiotic treatment agreement between infectious diseases specialists and ChatGPT®

Santiago Montiel-Romero[1], Sandra Rajme-López[1], Carla Marina Román-Montes[1,2], Alvaro López-Iñiguez[1], Héctor Orlando Rivera-Villegas[1,2], Eric Ochoa-Hein[1,3], María Fernanda González-Lara[1,2], Alfredo Ponce-de-León[1], Karla María Tamez-Torres[1,2*†] and Bernardo Alfonso Martinez-Guerra[1,2*†]

Abstract

Background Antimicrobial resistance is a global threat to public health. Chat Generative Pre-trained Transformer (ChatGPT®) is a language model tool based on artificial intelligence. ChatGPT® could analyze data from antimicrobial susceptibility tests in real time, especially in places where infectious diseases (ID) specialists are not available. We aimed to evaluate the agreement between ChatGPT® and ID specialists regarding appropriate antibiotic prescription in simulated cases.

Methods Using data from microbiological isolates recovered in our center, we fabricated 100 cases of patients with different infections. Each case included age, infectious syndrome, isolated organism and complete antibiogram. Considering a precise set of instructions, the cases were introduced into ChatGPT® and presented to five ID specialists. For each case, we asked, (1) "What is the most appropriate antibiotic that should be prescribed to the patient in the clinical case?" and (2) "According to the interpretation of the antibiogram, what is the most probable mechanism of resistance?". We then calculated the agreement between ID specialists and ChatGPT®, as well as Cohen's kappa coefficient.

Results Regarding the recommended antibiotic prescription, agreement between ID specialists and ChatGPT® was observed in 51/100 cases. The calculated kappa coefficient was 0.48. Agreement on antimicrobial resistance mechanisms was observed in 42/100 cases. The calculated kappa coefficient was 0.39. In a subanalysis according to infectious syndromes and microorganisms, Agreement (range 25 – 80%) and kappa coefficients (range 0.21–0.79) varied.

Conclusion We found poor agreement between ID specialists and ChatGPT® regarding the recommended antibiotic management in simulated clinical cases.

[†]Karla María Tamez-Torres and Bernardo Alfonso Martinez-Guerra contributed equally to this manuscript.

*Correspondence:
Karla María Tamez-Torres
karla.tamezt@incmnsz.mx; karla.tamez@gmail.com
Bernardo Alfonso Martinez-Guerra
bernardo.martinezg@incmnsz.mx; beramg@gmail.com

Full list of author information is available at the end of the article

Key points

AI in ID remains under investigation. We evaluated the agreement between ChatGPT® and ID specialists regarding antimicrobial prescriptions. Agreement between ID specialists and ChatGPT® was observed in 51% of the cases. AI must not substitute ID consults.

Keywords Artificial intelligence, Machine learning, Infectious diseases, Antimicrobial resistance

Introduction

Antimicrobial resistance (AMR) is a survival mechanism in bacteria that can be expressed through selective pressure [1] due to excessive and inappropriate use of antibiotics resulting in the emergence of multidrug-resistant (MDR) pathogens [2]. The World Health Organization (WHO) has declared AMR as one of the top 10 major public health concerns and a major threat to human health [3]. In 2019, it was estimated that 4.95 million deaths were associated with AMR, being *Escherichia coli*, *Staphylococcus aureus*, *Klebsiella pneumoniae*, *Streptococcus pneumoniae*, *Acinetobacter baumannii*, and *Pseudomonas aeruginosa*, as well as lower respiratory tract, bloodstream, and intra-abdominal infections the most lethal organisms and infectious syndromes, respectively [4]. Appropriate use of antimicrobials is key for combating AMR [5].

Artificial intelligence (AI) involves the development of computer systems capable of performing tasks that would otherwise require human intelligence [6]. The application of AI in medicine is a rapidly evolving field. It is being increasingly used in many specialties, including infectious diseases (ID) [7–9]. Chat Generative Pretrained Transformer (ChatGPT®) is an AI chatbot (i.e., an extended language model intended to simulate conversations with humans) relying on deep learning techniques to generate coherent responses resembling human conversations [10, 11]. The potential use of ChatGPT® in ID has been scarcely reported [10]. Although AI has been used in recent years to face AMR [12] and even pandemics (COVID-19) [13], its usefulness remains to be studied, particularly in developing countries.

ID specialists play a major role in various aspects of patient care and public health. Because ID specialists may not be widely available in all settings, tools to ensure appropriate antibiotic use are necessary. ChatGPT® capabilities in natural language processing, information synthesis, and resistance pattern recognition could result in a useful tool for analyzing antimicrobial susceptibility patterns and providing management suggestions. In this study, we aimed to evaluate the agreement between ChatGPT® and ID specialists regarding the appropriate antibiotic prescription for simulated cases using data from clinical isolates and their antibiograms.

Methods

We conducted a cross-sectional study using data from consecutive bacterial isolates obtained between January 1 2022, and June 30 2023, at our center's clinical microbiology laboratory. A total of 100 organisms (10 carbapenem-resistant *K. pneumoniae*, 10 carbapenem-resistant *E. coli*, 20 carbapenem-resistant *P. aeruginosa*, 10 carbapenem-resistant *A. baumanii*, 10 methicillin-resistant *S. aureus*, 10 vancomycin-resistant *Enterococcus faecium*, 10 *Klebsiella aerogenes*, 10 *Citrobacter freundii*, and 10 *Enterobacter cloacae* isolates) with their respective antimicrobial susceptibility tests results were used to fabricate clinical scenarios. Isolate identification and antimicrobial susceptibilities were obtained using the VITEK-2 system (Biomérieux®, Marcy-L'Etoile, FR). The simulated cases were constructed by 3 study investigators (S. M-R, K.M. T-T. and B.A. M-G.) and included age, infectious syndrome (e.g., bloodstream infection, pneumonia, intraabdominal infection, and pyelonephritis), the isolated organism and its complete antibiogram. Along with a precise set of instructions (Fig. 1), the cases were entered into ChatGPT® (GPT-4.0 series model) and presented to five board-certified ID specialists. For each case, we asked, (1) "What is the most appropriate antibiotic that should be prescribed to the patient in the clinical case?" and (2) "According to the interpretation of the antibiogram, what is the most probable mechanism of resistance?". ID specialists were required to submit their answers within five days and could use whatever resources they deemed necessary to provide their answers, as they could do when caring for any given patient.

No patient identifying data was used. A signed informed consent form was obtained from the participating ID specialists. Our study was approved by the Institutional Review Board (reference INF-4700-23-23-1), was conducted in accordance with the principles of the Declaration of Helsinki and complied with the Good Clinical Practice Guidelines.

Statistical analysis

Considering a power of 80%, a minimum acceptable Cohen's kappa coefficient (κ) of 0.6, and an expected kappa of 0.85, a minimum sample size of 90 simulated cases was estimated. The total sample was at 100 cases (complete cases can be found in the Supplementary appendix). Ten clinical per microorganism were

This is a case of a 62-year-old patient with pneumonia caused by Klebsiella pneumoniae. In vitro, the isolate is resistant to the antibiotics Ciprofloxacin, Nitrofurantoin, Trimethoprim-sulfamethoxazole, Ampicillin, Cephalothin, Cefuroxime, Cefotaxime, Ceftazidime, Ceftriaxone, Cefepime, Ertapenem, Meropenem and Ampicillin-sulbactam. In vitro, the isolate is susceptible to the antibiotics Amikacin and Gentamicin.

Question 1:
According to your interpretation of the antibiogram, what is the most appropriate antibiotic that should be prescribed to the patient in the clinical case?

____

To answer, consider the following:
Not all existing antibiotics that may be used for infections due to multidrug resistant organisms were tested. You may select an antibiotic that was not tested.
Please select only one antibiotic. If you believe combination therapy is warranted, state the combination that you consider best.
Only state the antibiotic, do not justify your decision.

Question 2:
According to your interpretation of the antibiogram, what is the most probable enzymatic mechanism of resistance to meropenem?

____

To answer, consider the following:
Select only one mechanism, do not justify your decision.

Fig. 1 Instructions set example

Table 1 Level of agreement in different scenarios

Level of agreement	Answer from ChatGPT®	Answer from ID specialist
Complete agreement	Recommend antibiotic A	Recommend antibiotic A
Complete agreement	Recommend antibiotics A and B	Recommend antibiotics A and B
Partial agreement	Recommend antibiotics A and B	Recommend antibiotic A
Partial agreement	Recommend antibiotics A and B	Recommend antibiotics A and C
Complete agreement	Existence of AMR mechanism A	Existence of AMR mechanism A
Complete agreement	Existence of AMR mechanisms A and B	Existence of AMR mechanisms A and B
Partial agreement	Existence of AMR mechanism A	Existence of AMR mechanisms A and B
Partial agreement	Existence of AMR mechanisms A and B	Existence of AMR mechanisms A and C

AMR antimicrobial resistance, *ID* infectious diseases

fabricated (except for carbapenem-resistant *P. aeruginosa*, for which 20 cases were built). Twenty cases were randomly assigned to each ID specialist. Complete agreement was considered when the answers between ID specialists and ChatGPT® were identical. As an exploratory analysis, partial agreement was calculated. Partial agreement was considered when two or more answers were given, and any level of agreement, according to the authors, was observed. Scenarios for complete and partial agreement are described in Table 1.

For the calculation of κ, a random agreement probability of 5% was assumed. The interpretation of the kappa coefficient was as follows: <0: No agreement, 0.00-0.20: Slight agreement, 0.21–0.40: Fair agreement, 0.41–0.60: Moderate agreement, 0.61–0.80: Substantial agreement, 0.81–0.99: Near perfect agreement, 1: Perfect agreement. An independent researcher, blinded to the source of the answers, performed the analysis.

Table 2 Results. Complete agreement

	Agreement (%)	κ
Recommended antibiotic	51/100 cases (51)	0.48 (moderate)
Antimicrobial resistance mechanism	42/100 cases (42)	0.39 (fair)
Recommended antibiotic in cases of BSI	18/34 cases (53)	0.51 (moderate)
Antimicrobial resistance mechanism in cases of BSI	18/34 cases (53)	0.51 (moderate)
Recommended antibiotic in cases of pneumonia	8/20 cases (40)	0.37 (fair)
Antimicrobial resistance mechanism in cases of pneumonia	8/20 cases (40)	0.37 (fair)
Recommended antibiotic in cases of pyelonephritis	7/18 cases (39)	0.36 (fair)
Antimicrobial resistance mechanism in cases of pyelonephritis	6/18 cases (33)	0.29 (fair)
Recommended antibiotic in cases of IAI	18/28 cases (64)	0.62 (substantial)
Antimicrobial resistance mechanism in cases of IAI	10/28 cases (36)	0.33 (fair)
Recommended antibiotic in cases of infection by GPC	14/20 cases (70)	0.68 (substantial)
Antimicrobial resistance mechanism in cases of infection by GPC	14/20 cases (70)	0.68 (substantial)
Recommended antibiotic in cases of infection by GNB	37/80 cases (46)	0.43 (moderate)
Antimicrobial resistance mechanism in cases of infection by GNB	28/80 cases (35)	0.32 (fair)

BSI bloodstream infection, *GNB* Gram-negative bacilli, *GPC* Gram-positive cocci, *IAI* intraabdominal infection, κ kappa coefficient

Results

As for the recommended antibiotic, there was agreement between ID specialists and ChatGPT® in 51/100 cases (51%), with a κ of 0.48. Agreement on the mechanism of antimicrobial resistance was observed in 42 cases (42%) with a calculated κ of 0.39. Regarding the recommended antibiotic in the cases of bloodstream infection, agreement was observed in 18/34 cases (53%, κ 0.51). As for the recommended antimicrobial for pneumonia, agreement was observed in 8/20 cases (40%, κ 0.37). Agreement in the recommended antibiotic was observed in 7/18 (39%, κ 0.36) and in 18/28 (64%, κ 0.62) cases of pyelonephritis and intraabdominal infection, respectively. Regarding the recommended antibiotic for infections caused by Gram-positive cocci, agreement was observed in 14/20 cases (70%, κ 0.68). There was agreement in 37/80 (46%, κ 0.43) cases of infections due to Gram-negative bacilli. Table 2 summarizes these results.

In the exploratory analysis, partial agreement regarding the recommended antibiotic and mechanism of resistance was observed in 68/100 (68%, κ 0.66) and 46/100 (46%, κ 0.43) cases, respectively (Table S1).

Discussion

We aimed to evaluate the agreement between ID specialists and ChatGPT® V4.0 regarding the recommended antibiotic prescription in simulated clinical cases. Agreement was infrequent, and moderate, as assessed by the Kappa coefficient. These results might be due to the platform's lack of contextual awareness, where the diagnostic and therapeutic focus should consider local epidemiology, AMR patterns and the availability of antimicrobial susceptibility testing, among other factors. In the exploratory subgroup analysis, intriguing results were observed. For bacteremia and infections due to Gram-positive cocci, partial agreement on recommended antibiotic was substantial and near perfect, respectively. These perplexing results could be explained by differences in the quantity of internet available information (e.g. in June 3rd, 2024, a PubMed® search for the MeSH terms "Gram-Positive bacteria" and "Gram-Negative bacteria" returned 561,445 and 888,627 results, respectively. For the MeSH terms "Bacteremia" and "Pneumonia, the search returned 33,615 and 36,7499 results, respectively). Importantly, these results must be cautiously interpreted as our definitions for partial agreement do not necessarily reflect the best clinical practice.

Our study statistically evaluates the agreement between ChatGPT® and ID specialists. A previous study evaluated the performance of ChatGPT® Version 3.5 in 40 medical offices in the Netherlands. The diagnostic and therapeutic advice recommended by ChatGPT® was compared to that given by an ID and clinical microbiology specialists. Results were classified on a scale from 1 (bad, incorrect advice) to 5 (excellent, corresponding to the advice given by infectious disease and clinical microbiology specialists). The advice given by ChatGPT® was of moderate quality with a median score of 2,8 [10]. It is known that these AI language models work through automatic learning algorithms trained to recognize and predict linguistic patterns by extracting information from websites, social media forums, open-access books, and articles, among others, which may not always include the most recent or accurate information. Of note, the repeating of patterns and questions could further affect the AI's response to similar problems. Additional issues with AI platforms with respect to ID (as well as with many other specialties) is that knowledge is constantly being developed and updated. Moreover, a recent article highlighted the complexity of the ID clinical field when compared to other medical specialties by using a range of metrics such as UpToDate® articles by specialty, number of recent evidence-based recommendations, and new FDA-approved molecules from 1985 to 2022 [14]. Regardless, large language models are evolving to render positive results with a quick and continuous improvement in the quality of answers. Recent research evaluated the clinical use of Med-PaLM (Medicine-Pathways Language Model)®, a large language model created by Google to answer medical questions. The authors reported that the AI model did not reach enough depth and quality as the physicians who were asked the same questions [15]. Importantly, further research did report a superior performance of version 2 of Med-PaLM 2 when answering United States Medical Licensing Examination questions [16].

Our study presents limitations that must be acknowledged. It was a cross-sectional study that did not evaluate the evolving nature of AI models. Additionally, the questions used in the cases, asked for the ideal antibiotic and not the best available drug. Regardless, local drug availability may have influenced the clinicians' answers. In our region, antibiotics recommended for the treatment of infections due to carbapenem resistant *Enterobacteriaceae*, such as aztreonam, imipenem-cilastatin-relebactam, meropenem-vaborbactam, cefiderocol and eravacycline are not commercially available. Additionally, given the limited availability of new antibiotics, AI could recommend unavailable drugs if no specific regional context is provided. We did not include complete contextual epidemiological and drug-availability information in the questions because we believe that not all non-ID clinicians may be aware of this data. Although interesting and approachable, we believe that including these contextual data sets in the questions could not fully replicate real-life use of Chat-GPT® and could guide the AI model into rendering different answers, modifying the end-results, and giving a false security regarding the clinical accuracy of Chat-GPT® in difficult medical scenarios. Because of adaptability, consistency of Chat-GPT®'s answers could vary across time. Although the consistency of the answers must be considered when studying AI models, determining the consistency of the answers across time was not included in the objectives of our study. The latter was considered because we aimed to simulate real-life clinical scenarios, in which the same questions are not commonly asked more than once, as clinical decisions must be undertaken in a timely manner. We considered that asking the same questions across time would not fully replicate real-life use of Chat-GPT®. In our study, we did not compare the answers to the recommendations provided in clinical guidelines because international guidelines recommend antibiotics that are not available in our region, which could mean that, when comparing Chat-GPT® to guidelines, the results could not be applicable in specific regional contexts. In addition to guidelines, ID specialists guide their evidence-based recommendations in regional, epidemiological, and specific clinical (e.g., chronic comorbidities, drug-drug interactions, etc.) scenarios. In our study we instructed the ID specialist and Chat-GPT® to not jusify their answer. The lack of justification was instructed to render concrete and specific answers by both Chat-GPT® and ID specialists. Strengths of our study include the blinded nature of the analysis and the fact the cases were built using real world data. The comprehensive set of instructions may further facilitate our study's reproducibility.

In the hands of inexperienced physicians, ChatGPT® might provide inaccurate information that could negatively impact patients' health. While AI has long been under development, our results, in accordance with previously published studies and current expert opinions, suggest that AI must not replace clinical specialists [17], but rather function as an aid in patient care. Moreover, there is solid evidence that ID consultations improve mortality rates and are associated with better outcomes in patients with different infectious syndromes [18–20].

Our research also suggests that, given the lack of consistency, clinical specialists must supervise the use of AI platforms. Additionally, when using large language models, human guided AI education should be sought. The current and future use of AI in the field of medicine and ID is undeniable; hence we must be familiar with AI platforms and find the best way to make the most of them. Further educational programs for AI human use in clinical medicine are needed. Likewise, it is necessary to foresee and recognize the ethical conflicts that could arise.

Conclusion

We found a moderate agreement between ID specialists and ChatGPT® regarding the recommended antibiotic management in simulated clinical cases. Our results do not support the use of ChatGPT® for ID-related decision-making. However, improvements in AI are expected; therefore, further and continuous research must be undertaken.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12879-024-10426-9.

Supplementary Material 1

Acknowledgements

We wish to acknowledge our Institution´s tireless personnel.

Author contributions

SM presented the idea of the manuscript. SM, BM and KT contributed to conceptualization, analysis and writing the original draft of this manuscript. SR, FG, CR, EO, AL and HR participated as the five ID specialists to answer the clinical cases. All authors reviewed the manuscript and contributed to method design, and writing, critically reviewing, and editing the final version of this manuscript. S.M.: Santiago MontielB.M.: Bernardo MartinezK.T.: Karla TamezS.R.: Sandra Rajme F.G.: Fernanda González C.R.: Carla Román E.O.: Eric OchoaA.L.: Álvaro LópezH.R.: Hector Rivera.

Funding
This research received no external funding.

Data availability
The datasets used and analyzed are available from the corresponding authors upon request.

Declarations

Ethics approval and consent to participate
No patient identifying data was used. A signed informed consent form was obtained the participating ID specialists. Our study was approved by the Institutional Review Board (Comité de Investigación y Comité de Ética en Investigación of the Instituto Nacional de Ciencias Médicas y Nutrición Salvador Zubirán, reference INF-4700-23-23-1), was conducted in accordance with the principles of the Declaration of Helsinki and complied with the Good Clinical Practice Guidelines.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Clinical trial
Not applicable.

Artificial intelligence software use
No artificial intelligence software was used to write this report.

Author details
[1]Department of Infectious Diseases, Instituto Nacional de Ciencias Médicas y Nutrición Salvador Zubirán, Mexico City, Mexico 15 Vasco de Quiroga, Belisario Domínguez Secc 16 Tlalpan, Mexico City 14080
[2]Clinical Microbiology Laboratory, Department of Infectious Diseases, Instituto Nacional de Ciencias Médicas y Nutrición Salvador Zubirán, Mexico City, Mexico15 Vasco de Quiroga, Belisario Domínguez Secc 16 Tlalpan, Mexico City 14080
[3]Hospital Epidemiology Department, Instituto Nacional de Ciencias Médicas y Nutrición Salvador Zubirán, Mexico City, Mexico 15 Vasco de Quiroga, Belisario Domínguez Secc 16 Tlalpan, Mexico City 14080

Received: 5 November 2024 / Accepted: 27 December 2024
Published online: 07 January 2025

References

1. Larsson DGJ, Flach CF. Antibiotic resistance in the environment. Nat Rev Microbiol. 2022;20:257–69.
2. Magiorakos AP, Srinivasan A, Carey RB, et al. Multidrug-resistant, extensively drug-resistant and pandrug-resistant bacteria: an international expert proposal for interim standard definitions for acquired resistance. Clin Microbiol Infect. 2012;18:268–81.
3. World Health Organization. Global research agenda for antimicrobial resistance in human health. 2023.
4. Murray CJ, Ikuta KS, Sharara F, et al. Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. Lancet. 2022;399:629–55.
5. Ilias I, Siempos E, Ioannidou ME. Falagas. The difference between adequate and appropriate Antimicrobial treatment. Clin Infect Dis. 2008;46:642.
6. Rajkomar A, Dean J, Kohane I. Machine learning in Medicine. N Engl J Med. 2019;380:1347–58.
7. Beam AL, Drazen JM, Kohane IS, Leong T-Y, Manrai AK, Rubin EJ. Artificial Intelligence in Medicine. N Engl J Med. 2023; 388:1220–1221. Available at: http://www.nejm.org/doi/https://doi.org/10.1056/NEJMe2206291
8. Brownstein JS, Rader B, Astley CM, Tian H. Advances in Artificial Intelligence for Infectious-Disease Surveillance. N Engl J Med. 2023;388:1597–607.
9. Cheng K, Li Z, He Y, et al. Potential use of Artificial Intelligence in Infectious Disease: take ChatGPT as an Example. Ann Biomed Eng. 2023;51:1130–5.
10. Sarink MJ, Bakker IL, Anas AA, Yusuf E. A study on the performance of ChatGPT in infectious diseases clinical consultation. Clin Microbiol Infect. 2023;29:1088–9.
11. Howard A, Hope W, Gerada A. ChatGPT and antimicrobial advice: the end of the consulting infection doctor? Lancet Infect Dis. 2023; 23:405–406. Available at: https://doi.org/10.1101/2023.01.23
12. Pascucci M, Royer G, Adamek J et al. AI-based mobile application to fight antibiotic resistance. Nat Commun 2021; 12.
13. Kamel Boulos MN, Geraghty EM. Geographical tracking and mapping of coronavirus disease COVID-19/severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) epidemic and associated events around the world: how 21st century GIS technologies are supporting the global fight against outbreaks and epidemics. Int J Health Geogr 2020; 19.
14. Grundy B, Houpt E. Complexity of infectious diseases compared with Other Medical subspecialties. Open Forum Infect Dis. 2023; 10.
15. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. Nature. 2023;620:172–80.
16. Singhal K, Tu T, Gottweis J et al. Towards Expert-Level Medical Question Answering with Large Language Models. 2023; Available at: http://arxiv.org/abs/2305.09617
17. Drazen JM, Kohane IS, Leong T-Y, Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI Chatbot for Medicine. N Engl J Med 2023; 388.
18. Bai AD, Showler A, Burry L, et al. Impact of infectious disease consultation on quality of care, mortality, and length of stay in staphylococcus aureus bacteremia: results from a large multicenter cohort study. Clin Infect Dis. 2015;60:1451–61.
19. Vogel M, Schmitz RPH, Hagel S, et al. Infectious disease consultation for Staphylococcus aureus bacteremia - A systematic review and meta-analysis. J Infect. 2016;72:19–28.
20. Shulder S, Tamma PD, Fiawoo S, et al. Infectious diseases Consultation Associated with reduced mortality in Gram-negative bacteremia. Clin Infect Dis. 2023;77:1234–7.

Publisher's note