Performance of a Novel Medical Artificial Intelligence Large Model (MedGo) on Supporting Decision-Making for Emergency Patients with Suspected Sepsis

Sen Jiang, Yi Gu, Tong Liu, Bo An, Chunxue Wang, Li Shao, Haitao Zhang, Lunxian Tang

Submitted to: Journal of Medical Internet Research
on: August 15, 2024

Table of Contents

Performance of a Novel Medical Artificial Intelligence Large Model (MedGo) on Supporting Decision-Making for Emergency Patients with Suspected Sepsis

Sen Jiang[1]; Yi Gu[1]; Tong Liu[1]; Bo An[2]; Chunxue Wang[1]; Li Shao[3]; Haitao Zhang[1]; Lunxian Tang[1]

[1]Department of Internal Emergency Medicine, Shanghai East Hospital, School of Medicine, Tongji University, Shanghai, China Shanghai CN
[2]Institute of Ethnology and Anthropology, Chinese Academy of Social Sciences, Beijing, China Beijing CN
[3]Department of VIP Clinic, Shanghai East Hospital, Tongji University School of Medicine, Shanghai, China Shanghai CN

Corresponding Author:
Lunxian Tang
Department of Internal Emergency Medicine, Shanghai East Hospital, School of Medicine, Tongji University, Shanghai, China
Shanghai East Hospital (North)/Tongji University 150, Jimo Road, Pudong District, Shanghai ,China.
Shanghai
CN

Abstract

Background: Large Artificial Intelligence (AI) language models have been increasingly applied in the medical field for disease prediction, diagnosis, and evaluation. However, research on AI-assisted early sepsis identification and screening remains scarce. Here, we conduct a retrospective study to evaluate the diagnostic efficacy of a novel medical large language model-MedGo developed by our collaborating team and us in early sepsis in emergency department (ED).

Objective: This study aims to evaluate the performance of a novel medical artificial intelligence large language model, MedGo, in supporting clinical decision-making for emergency department patients with suspected sepsis, specifically focusing on its diagnostic accuracy, comprehensiveness, readability, and analytical capabilities compared to physicians with varying levels of experience.

Methods: We retrospectively collected medical history data from 203 eligible patients treated at a tertiary teaching hospital between January 1, 2023 and January 1, 2024. MedGo's performance was compared to that of junior and senior ED physicians across nine assessment tasks related to the diagnosis and management of sepsis . A five-point Likert scale was used to assess the four dimensions of accuracy, comprehensiveness, readability and case analysis skills.

Results: MedGo exhibited diagnostic performance comparable to senior doctors, scoring 4 on the Likert Scale for accuracy, comprehensiveness, readability, and analytical capability, significantly surpassing junior doctors. Furthermore, MedGo's decision support enhanced both junior and senior doctors' diagnostic abilities, with junior doctors' performance equal that of seniors. Notably, MedGo consistently delivered exceptional results in diagnosing early sepsis cases of varying severity.

Conclusions: MedGo demonstrates remarkable diagnostic efficacy in early sepsis, effectively supporting clinicians of diverse experience levels in making informed decisions in the time-urgent ED. Although we acknowledge its limitations and emphasize the importance of comprehensive, standardized, systematic, and visualized medical history data in future research endeavors, the results underscore the potential of MedGo as a supportive tool in ED settings, thereby laying the groundwork for future developing specialized sepsis models.

(JMIR Preprints 15/08/2024:65438)
DOI: https://doi.org/10.2196/preprints.65438

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✔ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Performance of a Novel Medical Artificial Intelligence Large Model (MedGo) on Supporting Decision-Making for Emergency Patients with Suspected Sepsis

Sen Jiang[1,2†], Yi Gu[1,2†], Tong Liu[1,2†], Bo An[3], Chunxue Wang[1,2], Li Shao[4], Haitao Zhang[1,2*], Lunxian Tang[1,2*]

[1] Department of Internal Emergency Medicine, Shanghai East Hospital, School of Medicine, Tongji University, Shanghai, China

[2] School of Medicine, Tongji University, Shanghai, China

3 Institute of Ethnology and Anthropology, Chinese Academy of Social Sciences, Beijing, China

[4] Department of VIP Clinic, Shanghai East Hospital, Tongji University School of Medicine, Shanghai, China

† These authors contribute equally to this work

Corresponding authors:

[*]Haitao Zhang, MD, PHD

Department of Emergency Medicine and Critical Care

Shanghai East Hospital / Tongji University

150, Jimo Road, Pudong District, Shanghai, 200120 ,China.

Tel: +86-21-38804518-25215; FAX: +86-21-38804518

Email: boy8672@126.com

[*] Lunxian Tang, MD, PHD

Department of Internal Emergency Medicine and Critical Care

Shanghai East Hospital (North)/Tongji University

150, Jimo Road, Pudong District, Shanghai, 200120 ,China.

Tel: +86-21-38804518-25215; FAX: +86-21-38804518

Email: 456tlx@163.com

Abstract

Background: Large Artificial Intelligence (AI) language models have been increasingly applied in the medical field for disease prediction, diagnosis, and evaluation. However, research on AI-assisted early sepsis identification and screening remains scarce. Here, we conduct a retrospective study to evaluate the diagnostic efficacy of a novel medical large language model-MedGo developed by our collaborating team and us in early sepsis in emergency department (ED).

Objective: This study aims to evaluate the performance of a novel medical artificial intelligence large language model, MedGo, in supporting clinical decision-making for emergency department patients with suspected sepsis, specifically focusing on its diagnostic accuracy, comprehensiveness, readability, and analytical capabilities compared to physicians with varying levels of experience.

Methods: We retrospectively collected medical history data from 203 eligible patients treated at a tertiary teaching hospital between January 1, 2023 and January 1, 2024. MedGo's performance was compared to that of junior and senior ED physicians across nine assessment tasks related to the diagnosis and management of sepsis . A five-point Likert scale was used to assess the four dimensions of accuracy, comprehensiveness, readability and case analysis skills.

Results: MedGo exhibited diagnostic performance comparable to senior doctors, scoring 4 on the Likert Scale for accuracy, comprehensiveness, readability, and analytical capability, significantly surpassing junior doctors. Furthermore, MedGo's decision support enhanced both junior and senior doctors' diagnostic abilities, with junior doctors' performance equal that of seniors. Notably, MedGo consistently delivered exceptional results in diagnosing early sepsis cases of varying severity.

Conclusion: MedGo demonstrates remarkable diagnostic efficacy in early sepsis, effectively supporting clinicians of diverse experience levels in making informed decisions in the time-urgent ED. Although we acknowledge its limitations and emphasize the importance of comprehensive, standardized, systematic, and visualized medical history data in future research endeavors, the results underscore the potential of MedGo as a supportive tool in ED settings, thereby laying the groundwork for future developing specialized sepsis models.

Keywords artificial intelligence, MedGo, sepsis, diagnosis, emergency department

Introduction

Sepsis is a life-threatening condition characterized by organ dysfunction resulting from a dysregulated host response to infection. The prognosis of sepsis is closely linked to early intervention, and studies have demonstrated that every hour of delay in administering antimicrobial agents results in an 0.3% absolute increase in mortality rate.[1] According to the 2021 International Guidelines for the Management of Sepsis and Septic Shock, it is unequivocally recommended that antibiotics be administered as soon as sepsis is diagnosed, ideally within the first hour, to achieve optimal therapeutic outcomes.[2] Consequently, early diagnosis and treatment are crucial for improving sepsis prognosis. However, in emergency settings, early detection of sepsis is often challenging due to nonspecific clinical manifestations, the absence of highly sensitive and specific diagnostic tools, the complexity of the emergency environment, and individual patient difference. These factors complicate the early diagnosis process. Therefore, a tool that can effectively adapt to the emergency setting and assist in the early clinical diagnosis of sepsis is of paramount importance for emergency physicians.

In recent years, artificial intelligence(AI) technology has rapidly evolved, incorporating a wide range of technologies and methods designed to simulate and implement human intelligence through computer programs and machines. Large AI language models, such as ChatGPT, have been increasingly applied in the medical field, particularly in disease prediction, diagnosis, and evaluation. For instance, AI technology can analyze radiological and pathological images to aid in disease diagnosis. [3] However, research on AI-assisted early diagnosis of sepsis remains limited. Although a sepsis prediction model was established in 2010, its external validation proved ineffective and difficult to generalize due to the small sample size, which included only a few dozen cases. [4] Faced with the aforementioned challenges dilemmas, we have independently developed a comprehensive medical large model named MedGo. MedGo is developed by the Biomedical Artificial Intelligence Laboratory, a collaborative effort between the Institute of Software at the Chinese Academy of Sciences, Tongji University, and Shanghai East Hospital. This model is trained based on approximately 20 billion medical tokens. In November 2023, MedGo emerged as the champion in the Chinese Biomedical Language Understanding Evaluation (CBLUE) with a comprehensive score

of 63.993 (https://tianchi.aliyun.com/cblue) . In the same year, MedGo also successfully passed the Chinese National Medical Practitioner Qualification Examination, demonstrating its medical expertise comparable to that of clinical doctors. Leveraging MedGo, we designed a specialized large language model tailored specifically for sepsis. Using a retrospective cohort study method, we rigorously evaluated the effectiveness of this model in supporting decision-making for patients presentation to emergency department (ED) with suspected sepsis. Our ultimate aim is to provide fundamental data support for enhancing the capabilities of emergency identification and early screening for suspected sepsis cases.

Methods

Study design and participants

This retrospective study aimed to evaluate the effectiveness of the medical large language model MedGo (https://www.MedGo.cn/) as a decision support tool for suspected sepsis in the ED. We retrospectively collected data from patients who presented to the ED of Shanghai East Hospital, a tertiary teaching hospital affiliated to Tongji University, between January 1,2023 and January 1,2024. The inclusion criteria are as follows: presentation to the ED, an initial diagnosis of suspected sepsis before admission, a confirmed diagnosis of sepsis after admission, and age ≥ 18 years. Patients were excluded if they had active malignancy, HIV infection, were under 18 years of age, or were pregnant or lactating. As this study used anonymize retrospective data and did not involve prospective recruitment of human subjects, ethical approval was not required. However, all data were handled in accordance with relevant laws and regulations to ensure the legal and secure use of patient information.

Procedures and Assessments

1. MedGo Initial Assessment: The following prompt was entered into MedGo: "Please provide five possible diseases for the following symptoms: [Copy and paste the chief complaint of each clinical case]." From these five differential diagnoses, MedGo selected one primary diagnosis. This process was repeated twice for each case.

2.MedGo Expanded Assessment: The following prompt was entered into MedGo: "Based on the currently provided medical history, past history, physical examination, and medication information,

provide a list of five possible diseases and select one as the primary diagnosis." MedGo was also tasked with determining the presence of suspected sepsis based on qSOFA (Quick Sequential Organ Failure Assessment) and NEWS (National Early Warning Score) scores and assessing disease severity using SOFA and NEWS scores. These scores were categorized as mild (qSOFA ≥ 1 or NEWS 4-6) or severe (qSOFA ≥ 2 or NEWS > 6) according to Consensus of Chinese experts on early prevention and blocking of sepsis.[5] Finally, MedGo provided recommendations for observation or emergency ward treatment and suggested final treatment plans based on the disease severity. This process was repeated twice for each case.

3.MedGo Adding Auxiliary Examination Assessment: Laboratory data was incorporated into the patient information, and the assessment process described in step 2 was repeated.

4. Clinician Assessment: Two junior ED physicians with less than three years of clinical experience and two senior ED physicians with more than ten years of experience independently assessed each case. They followed a three-step process for receiving information: first the chief complaint, then the basic medical history and physical examination results, and finally, the laboratory results. At each stage, they provided five differential diagnosis, selected one primary diagnosis, determined the presence of suspected sepsis, assessed disease severity based on clinical experience, and suggested treatment plans.

5. MedGo-Assisted Clinician Assessment: Both junior and senior ED physicians repeated the assessment process described in step 4, using MedGo's suggestions at each stage.

6. Case Difficulty Assessment: Two senior ED physicians with advanced professional titles independently evaluated the difficulty of each case based on the results from MedGo, clinicians, and MedGo-assisted assessments. Cases were categorized as easy, moderate, or difficult, using a 5-point Likert scale as other investigator had previously reported [6,7] and revised by us to fit our study (Table 1). The Likert scale assessed four dimensions: accuracy, comprehensiveness, readability, and case analysis ability. Each response was cross-referenced with reliable sources and international standards. To ensure a thorough assessment, the ratings from the two reviewers for each dimension were combined, and the median or mode was used to determine the final score for that dimension.

7. Comparisons:The decision support performance of MedGo and clinicians with varying levels of experience was compared at different stages: after receiving the chief complaint, following history and physical examination, and with the addition of laboratory results. The performance of clinicians with and without MedGo assistance was evaluated at each of these stages. Additionally, the effectiveness of MedGo in diagnosing and managing patients with different severity and levels of case difficulty was assessed.

Table 1. Assessment scheme

1	2	3	4	5
The answer blatantly fails to meet the specified requirement	The answer falls short of the criterion, but not significantly	The answer satisfies the requirement to a reasonable but not remarkable degree	The answer closely complies with the requirement	The reaction surpasses the typical expectation and excels in the criterion

Statistical Analysis

Differences between groups were assessed using independent samples t-tests and one-way ANOVAs when data distribution was normal. For non-normally distributed data, the Mann–Whitney U test and Kruskal–Wallis test were used for comparisons between two groups and multiple groups, respectively. All statistical analyses were conducted with Statistical Package for the Social Sciences software, version 27.0 (SPSS), and the graphical user interface GraphPad Prism (version 9.5.1).

Results

Clinical characteristics of patients

The flowchart of patients enrollment and exclusion is documented in Figure 1. A total of 203 patients were included in the study (mean age: 79.91 years ± 10.15 [SD]; age range: 25-99 years; 94 female) (Table 2). The diagnostic difficulty varied among the patients: 139 cases (68.5%) were classified as "Easy Diagnosis," 49 cases (24.1%) as "Moderate Difficulty," and 15 cases (7.4%) as "Difficult Diagnosis." In terms of disease severity, 131 patients (64.5%) were categorized as having "Mild" disease, while 72 patients (35.5%) were classified as having "Severe" disease.

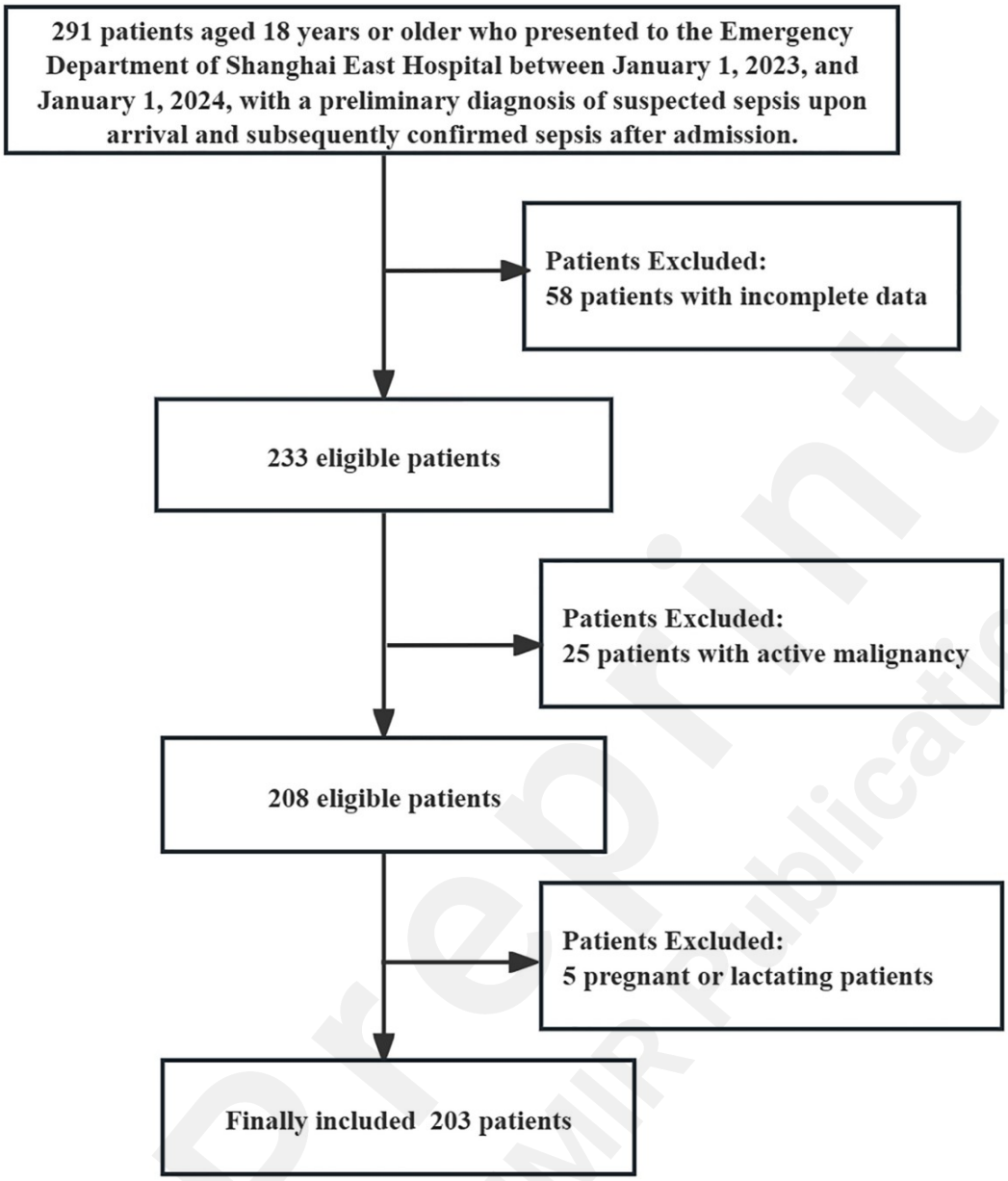


Figure 1. Flowchart showing the inclusion and exclusion of the patients.

Table 2. Patient demographics and clinical characteristics

Feature	Total (n=203)
Clinical Features	
Age (Years)	79.91 ± 10.15
Gender	
Male, n (%)	109 (53.7)
Female, n (%)	94 (46.3)

Case Difficulty	
Easy Diagnosis, n (%)	139 (68.5)
Moderate Difficulty, n (%)	49 (24.1)
Difficult Diagnosis, n (%)	15 (7.4)
Disease Severity	
Mild, n (%)	131 (64.5)
Severe, n (%)	72 (35.5)

Overall Performance of MedGo

To start with, we set out to look at the performance of MedGo across various assessment metrics. In terms of accuracy, MedGo closely mirrored the performance of senior doctors, achieving scores above 4 points on the 5-point Likert scale for most tasks (Table 3), particularly in formulating initial and detailed differential diagnoses (Figure 2A). Although there was a minor decline in accuracy when assessing disease severity and formulating subsequent management steps, this highlights areas for potential model enhancement. Junior doctors consistently received lower accuracy scores compared to both MedGo and senior doctors, indicating MedGo's potential to bridge the experience gap and improve diagnostic accuracy for less experienced clinicians.

Table 3. Questions and clinical scenarios

Category	Number	Question
Initial Assessment	1	Please provide five possible diseases for the following symptoms
Detailed Assessment	2	Based on the currently provided medical history, past history, physical examination, and medication information, provide a list of five possible diseases and select one as the primary diagnosis
	3	Determine if the patient in this case is a suspected septic patient
	4	Please determine the severity of the patient in this case (mild OR severe)
	5	Next steps in the management of the condition

Adding ancillary screening assessments	6	Please list 5 differential diagnoses based on the presenting complaint, current medical history, past history, physical examination, and medication use provided so far and choose one as the primary diagnosis
	7	Determine if the patient in this case is a suspected septic patient
	8	Please determine the severity of the patient in this case (mild OR severe)
	9	Next steps in the management of the condition

As shown in Figure 2B, MedGo demonstrated a high level of comprehensiveness, consistently achieving a median score of 4 across all tasks. This indicates that MedGo provided thorough assessments that covered most relevant clinical aspects, closely matching the performance of senior doctors who consistently scored above 4. In contrast, junior doctors showed fair comprehensiveness, with a median score of 3, suggesting there is room for improvement in capturing all essential details in their assessments.

In terms of readability, all three groups- MedGo, junior doctors, and senior doctors – achieved median scores of 4, indicating good clarity and comprehensibility in their outputs (Figure 2C). Notably, MedGo demonstrated the most consistent readability across all tasks, as evidenced by its consistently narrow error bars. This reflect MedGo’s ability to maintain a stable level of clarity regardless of case complexity. Junior and senior doctors, on the other hand, showed slight variations in readability, especially in tasks related to sepsis assessment and disease severity. This suggest that their communication might be influenced by the specific details of individual cases.

In the realm of case analysis, MedGo exhibited strong analytical abilities, closely matching the performance of senior doctors (Figure 2D). Both MedGo and senior doctors frequently scored above 4 points on the Likert scale, especially in tasks involving detailed assessment and interpretation of auxiliary examinations. This demonstrates MedGo’s ability to effectively interpret case information, identify crucial clinical issues, and generate coherent diagnostic and management strategies. In contrast, junior doctors consistently scored lower than both MedGo and senior doctors, particularly in tasks requiring advanced clinical reasoning. This highlights MedGo’s potential as a valuable support tool for less experienced physicians navigating complex cases.

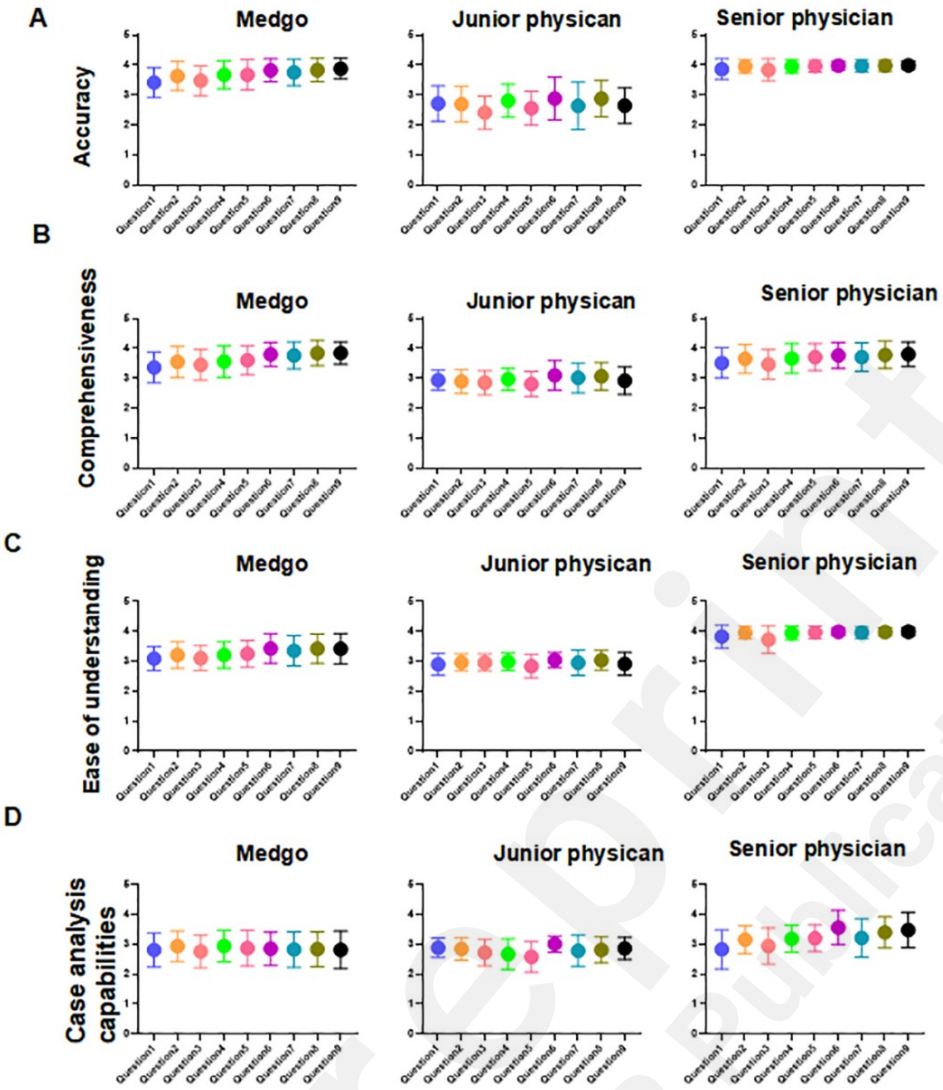


Figure 2. Clinical Decision-Making Performance of MedGo. The diagnostic performance of MedGo was assessed using a 5-point Likert scale, with higher scores indicating better performance.A, the performance of accuracy. B. the perf ormance of comprehensiveness. C, the performance of ease of understanding. D, the performance of case analysis capabilities. Each color in the figure represents a different question.

Performance of Doctors with Different Experience Levels Combined with MedGo

Next we investigate whether the integration of MedGo as a decision support tool could improve the performance of both junior and senior emergency physicians. As shown in Figure3A, MedGo enhanced diagnostic accuracy for both groups, with a more pronounced impact on junior doctors. This effectively narrowed the accuracy gap between junior and senior physicians, particularly during the initial assessment phase based solely on the chief complaint. These findings suggest that MedGo

can compensate for limited experience in the early diagnostic stages, enhancing accuracy for less experienced clinicians.

Similar improvements were observed in assessment comprehensiveness (Figure 3B). With the help of MedGo, junior doctors achieved scores closer to senior-level comprehensiveness, indicating the model's ability to guide less experienced clinicians towards more thorough evaluations. Notably, senior doctors also exhibited increased comprehensiveness scores with MedGo support, suggesting that even experienced physicians can benefit from the model's prompts for a systematic review of patient data.

Readability remained consistently high across both groups, with median scores of 4, regardless of MedGo use (Figure 3C). This indicates that MedGo did not compromise the clarity and comprehensibility of the presented medical information. However, junior doctors showed reduced variability in readability for detailed assessment and auxiliary examination tasks when utilizing MedGo, implying a stabilizing effect on the consistency of their information presentation.

MedGo significantly influenced case analysis ability, with junior doctors demonstrating substantial performance boosts, especially in higher-level clinical reasoning tasks (Figure 3D). This underscores MedGo's capability to guide less experienced physicians through complex case analyses, fostering more informed decision-making. Contrary to that, senior doctors, with their already established analytical skills, saw only marginal improvements with MedGo support. Overall, these findings highlight MedGo's potential as a valuable tool for elevating the diagnostic and analytical skills of less experienced clinicians, narrowing the experience gap and contributing to reduced diagnostic errors.

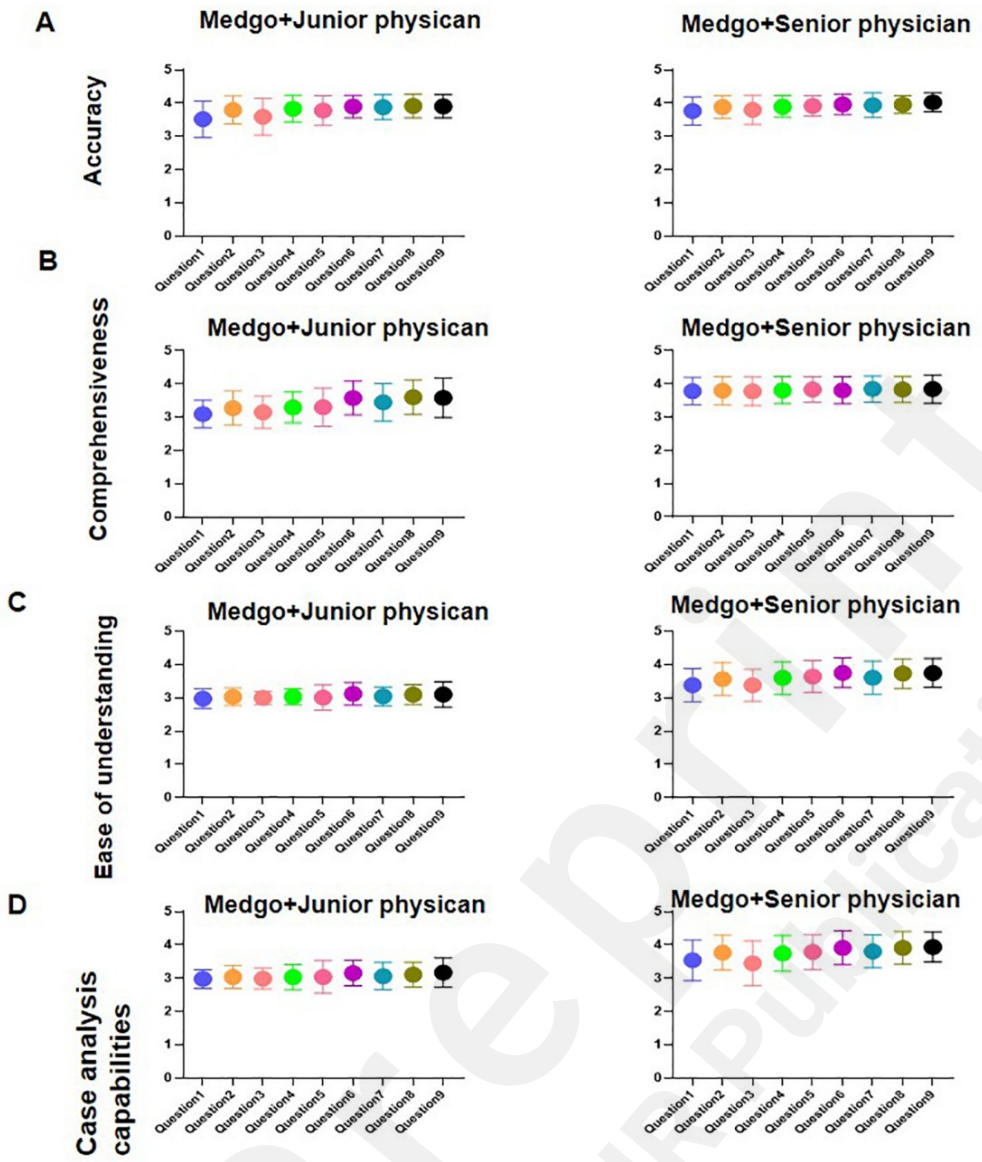


Figure 3. Performance of junior and senior physicians with MedGo assistance across nine assessment tasks.A, Accuracy. B, Comprehensiveness. C, Ease of understanding. D, Case analysis capabilities.

Comparison of Clinical Decision-Making Performance Between Doctors and MedGo-Assisted Doctors

The clinical decision-making performance of both junior and senior physicians was evaluated using the average total scores across all nine assessment tasks comparing scenarios with and without MedGo assistance. As shown in Figure 4, both groups-junior and senior physicians- achieved significantly higher average scores when using MedGo (Junior physician vs. Junior physician + MedGo, $P < 0.001$; Senior physician vs. Senior physician + MedGo, $P < 0.001$). These results highlight the substantial positive impact of MedGo on clinical decision-making, irrespective of the

physicians' experience level.

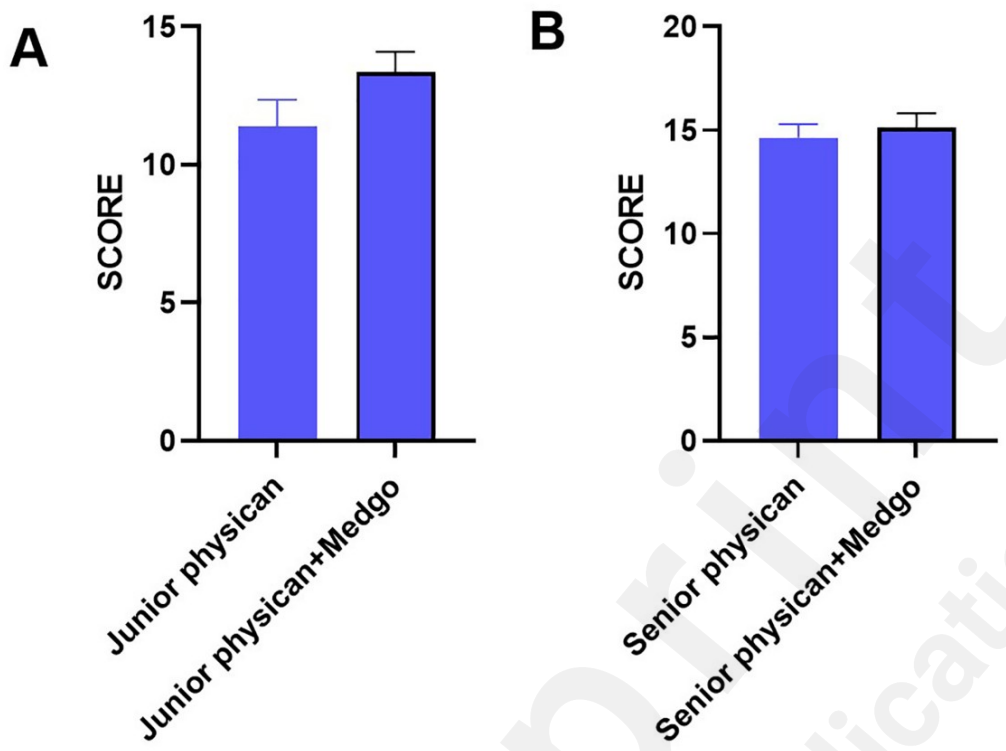


Figure 4. Comparison of overall clinical decision-making performance between physicians with and without MedGo assistance. Average total scores across nine assessment tasks for (A) junior physicians and (B) senior physicians, with and without the assistance of MedGo. Error bars represent standard deviation.

Performance of MedGo Across Different Disease Severity Levels

MedGo demonstrated consistent performance across different disease severity levels for most assessment metrics. As shown in Figure 5, MedGo consistently achieved median scores above 3 for accuracy, a median score of 4 for comprehensiveness, and a median score of 4 for readability in both mild and severe sepsis cases across all nine assessment tasks. These findings suggest that MedGo's ability to generate accurate, comprehensive, and readable outputs which is not significantly influenced by disease severity, highlighting its robust performance across a spectrum of clinical presentations.

While MedGo generally maintained stable performance across different severity, it exhibited slightly increased score variability for severe sepsis cases compared to mild cases in specific tasks. This variability was primarily observed in tasks related to disease severity assessment and determining the

next management step, indicating a potential area for future model refinement to enhance consistency in managing more critical patients. Despite this subtle variability, MedGo's ability to analyze and synthesize clinical information, particularly in tasks demanding higher-level clinical reasoning, remained consistent across both severity levels. This reinforces MedGo's potential as a reliable decision support tool in the ED, capable of providing consistent support across a wide range of clinical scenarios.

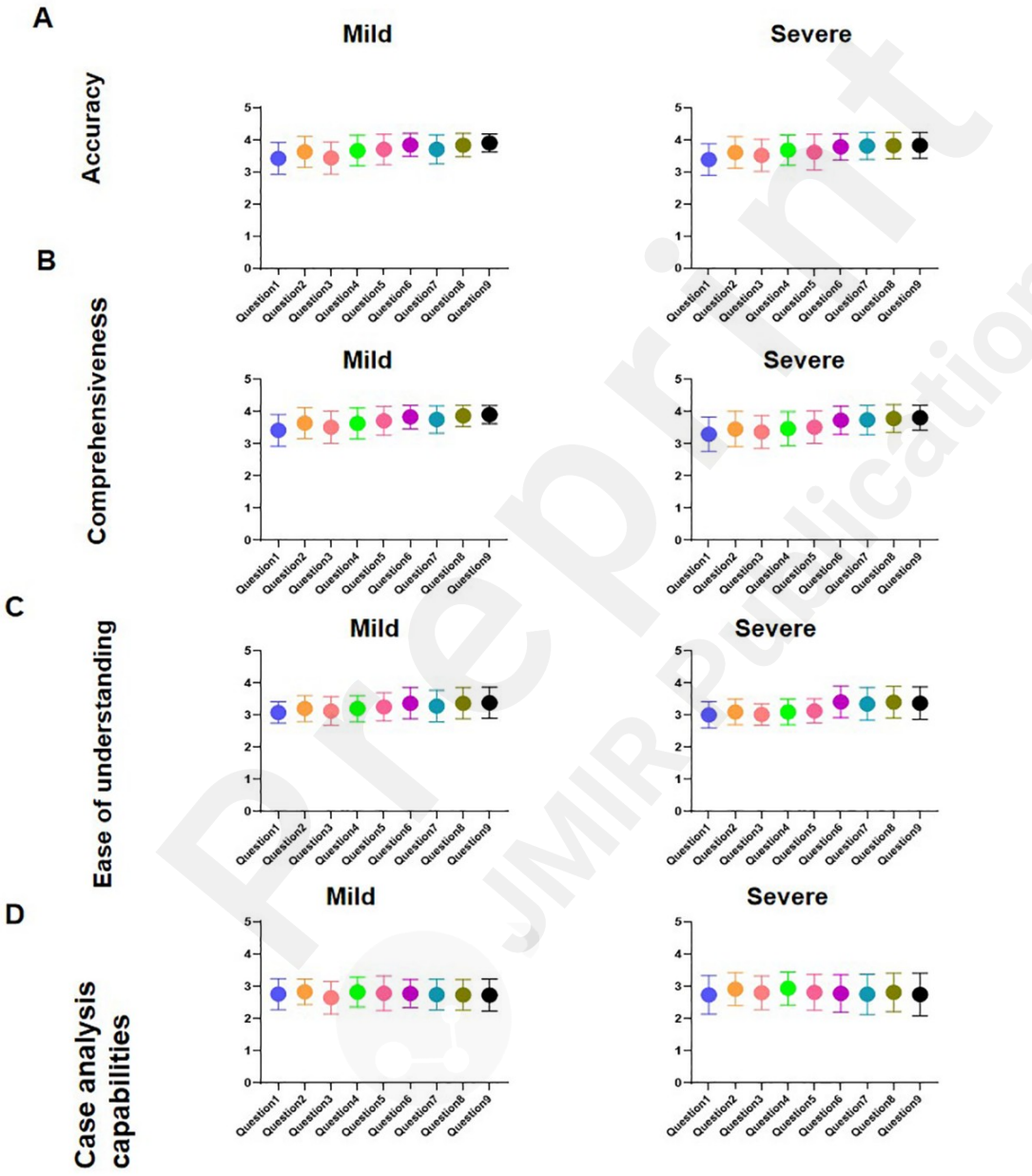


Figure 5. Performance of MedGo across different disease severity levels. The performance of MedGo was assessed using a 5-point Likert scale, with higher scores indicating better performance. A, Accuracy. B, Comprehensiveness. C, Ease of understanding. D, Case analysis capabilities. Each color in the figure represents a different question.

Treatment Effectiveness in Cases of Varying Difficulty

To further investigate the impact of case complexity on MedGo's performance, we categorized the cases into three difficulty levels: simple, moderate, and difficult. Figure 6 displays the average total score MedGo achieved across all nine assessment tasks for each difficulty level. A one-way ANOVA revealed statistically significant differences in performance among these three levels ($P < 0.05$), suggesting that case complexity does affect MedGo's overall effectiveness. While Figure 6 offers a general overview, additional post-hoc analysis is needed to identify the specific differences between each difficulty level. This detailed analysis will provide a clearer understanding of MedGo's performance in various challenging scenarios and highlight potential areas for improvement.

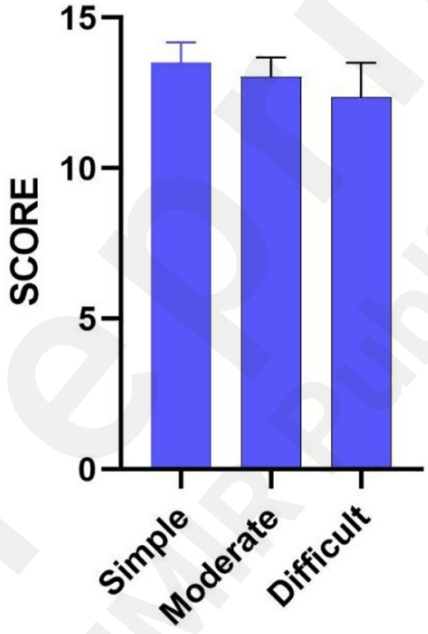


Figure 6. MedGo's performance in cases of varying difficulty. The total scores across all nine assessment tasks were averaged for each difficulty level (simple, moderate, difficult). Error bars represent standard deviation. One-way ANOVA showed significant differences in MedGo's performance across the three difficulty levels ($P < 0.05$), indicating that case complexity affects its overall performance.

Discussion

Despite the availability of various screening tools for the early diagnosis of sepsis, these methods have not achieved widespread clinical application due to limitations in specificity and sensitivity, as

well as challenges in accurately stratifying disease severity and reflecting its full extent. Meanwhile, artificial intelligence (AI), as a recent advancement in the medical field, is progressively playing a more significant role in medical diagnosis and treatment. However, research on the application of AI for the early diagnosis of sepsis remains relatively limited. Hence, we conducted a retrospective cohort analysis to evaluate the effectiveness of MedGo, a large medical language model in AI developed by us, for the early diagnosis of sepsis. We are the first to demonstrate that MedGo, by integrating multidimensional clinical information from patients, achieved diagnostic efficacy for early sepsis comparable to that of experienced senior physicians and even surpassed the performance of junior physicians. These findings suggest the potential for developing a sepsis-specific clinical decision-making disease model based on MedGo in the future.

It is important to acknowledge that MedGo currently has limitations in assessing the severity of early sepsis, which may be attributed to the exclusion of certain physical examination data, such as Glasgow Coma Scale scores, from the provided information. Additionally, imaging results are presently provided to MedGo only in text report form rather than as actual image, which could contribute to errors in evaluating the severity of early sepsis. Therefore, when developing a sepsis-specific disease model in the future, it is crucial to enhance the comprehensiveness, systematic organization, and visualization of medical history data, and to integrate currently available screening tools. Concurrently, it might be essential to enable the model to recognize and interpret imaging data in the future study.

As the application of AI in the medical field becomes increasingly widespread, several issues must be addressed. Firstly, data privacy and security are paramount. The collection and storage of medical data involve sensitive personal information, necessitating robust anonymization and de-identification measures during data handling. In this study, MedGo only extracted essential clinical diagnostic information and did not assess patients' private details. Secondly, data standardization and normalization are critical considerations. Variations in data formats and storage methods among different medical institutions can hinder data interoperability and sharing. The experimental data for this study were sourced exclusively from Shanghai East Hospital, without involvement from other medical institutions. Future development of a sepsis-specific disease model, particularly in multi-center research, will require standardized methods for data storage and extraction to ensure consistency and compatibility. Finally, the opaque nature of the decision-making process in large language models raises ethical concerns about safety risks in clinical decision-making. Consequently,

while AI can assist doctors in diagnosis, it cannot fully replace their clinical judgment.

Acknowledgments*:*This work was supported by grants from the municipal Natural Science Foundation of Shanghai Scientific Committee of China (22ZR1451000 to L.T.), the peak supporting clinical discipline of Shanghai Health Bureau (2023ZDFC0104 to L.T). the key clinical discipline of Shanghai Pudong health bureau (PWZxk2022-17 to L.T.), the Joint research of Shanghai Pudong health bureau (PW2023-07 to L.T.), the clinical peak discipline of Shanghai Pudong heath bureau (PWYgf2021-03 to Z.L.), the leading medical talent project of Shanghai Pudong heath bureau (PWR12020-07 to L.S.).

Author affiliations*:* From the Department of Internal Emergency Medicine (Jiang, Gu, Liu, Wang, Zhang, Tang), School of Medicine, Tongji University, Shanghai, China; Shanghai East Hospital, School of Medicine, Tongji University, Shanghai, China (Jiang, Gu, Liu, Wang, Zhang, Tang); Institute of Ethnology and Anthropology, Chinese Academy of Social Sciences, Beijing, China (An); and Department of VIP Clinic, Shanghai East Hospital, Tongji University School of Medicine, Shanghai, China (Shao).

Author contributions: SJ, YG, and TL conceptualized and designed the study, collected the data, performed data analysis, drafted the initial manuscript, and reviewed and revised the final manuscript. BA, CW, and LS assessed the cases and participated in the drafting and review of the final manuscript. HZ and LT provided guidance and supervision throughout the study, reviewed study results, and participated in the drafting and review of the final manuscript. SJ, YG, and TL share first authorship. HZ and LT share senior authorship.

Data sharing statement*:* Data sharing statement: The datasets generated and/or analyzed during the current study are available from the corresponding author upon reasonable request.

Authorship: All authors attest to meeting the four ICMJE.org authorship criteria: (1) Substantial contributions to the conception or design of the work; or the acquisition, analysis, or interpretation of data for the work; AND (2) Drafting the work or revising it critically for important intellectual content; AND (3) Final approval of the version to be published; AND (4) Agreement to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are
appropriately investigated and resolved.

Competing interests Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1.Liu, Vincent X, et al. The timing of early antibiotics and hospital mortality in sepsis. American journal of respiratory and critical care medicine. 2017;196.(7): 856-863.

2.Laura E ,Andrew R ,Waleed A , et al.Surviving sepsis campaign: international guidelines for management of sepsis and septic shock 2021.Intensive care medicine. 2021;47(11):1181-1247.

3.Rajesh B ,Satheesh K ,R R B.Performance of ChatGPT on a radiology board-style examination: insights into current strengths and limitations.Radiology. 2023;307(5):230582-230582.

4.Teng, Andrew K., and Adam B. Wilcox. A review of predictive analytics solutions for sepsis patients. Applied clinical informatics. 2020;11 (3): 387-398.

5.Emergency Medicine Branch Of Chinese Medical Care International Exchange Promotion Association et al. Consensus of Chinese experts on early prevention and blocking of sepsis. Zhonghua wei zhong bing ji jiu yi xue. 2020;32,(5): 518-530.

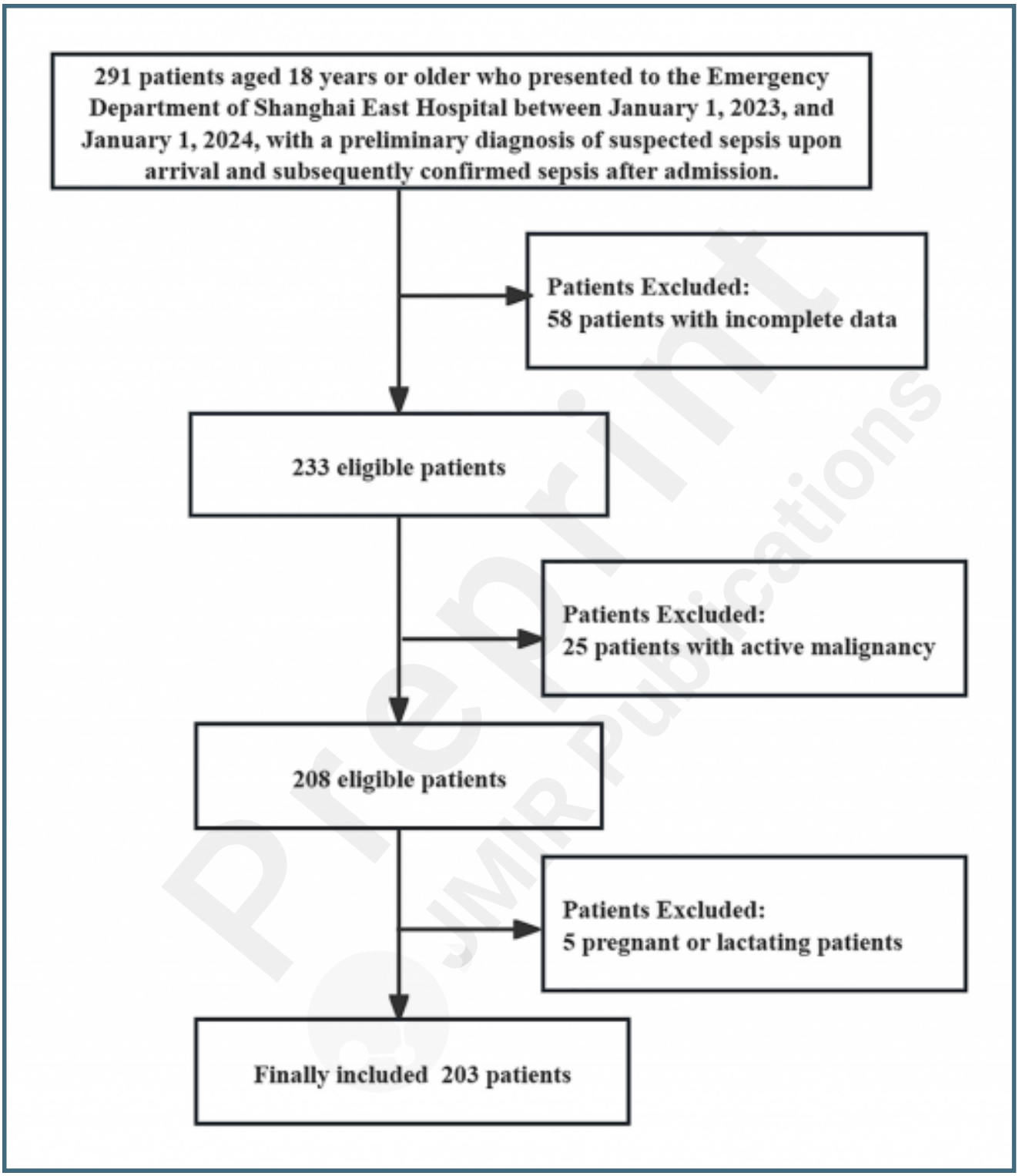
6.Song H, Xia Y, Luo Z, et al. Evaluating the Performance of Different Large Language Models on Health Consultation and Patient Education in Urolithiasis.J Med Syst. 2023;47(1):125.

7.Sandmann S, Riepenhausen S, Plagwitz L, Varghese J. Systematic analysis of ChatGPT, Google search and Llama 2 for clinical decision support tasks. Nat Commun. 2024;15(1):2050.

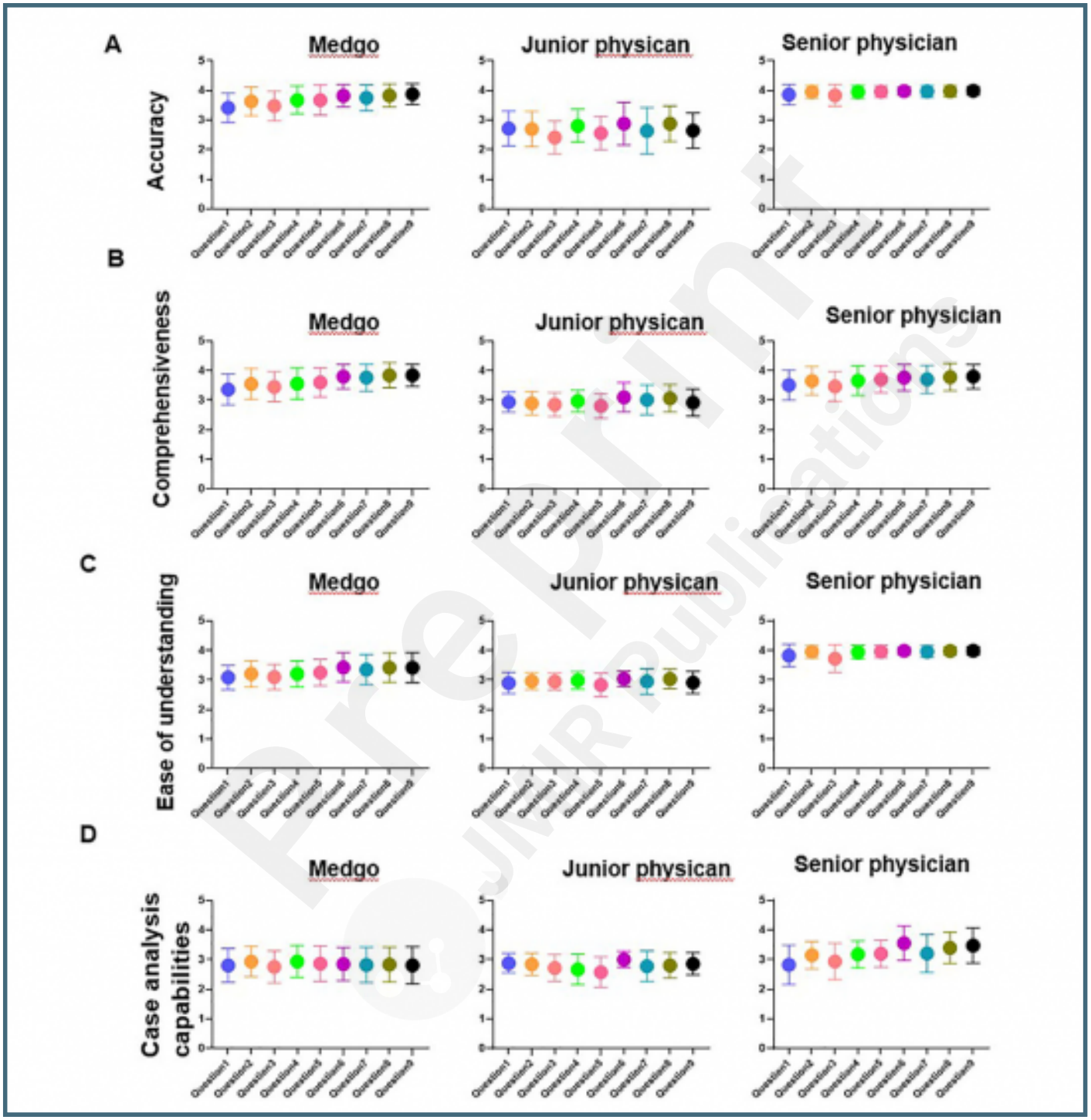
Supplementary Files

Figures

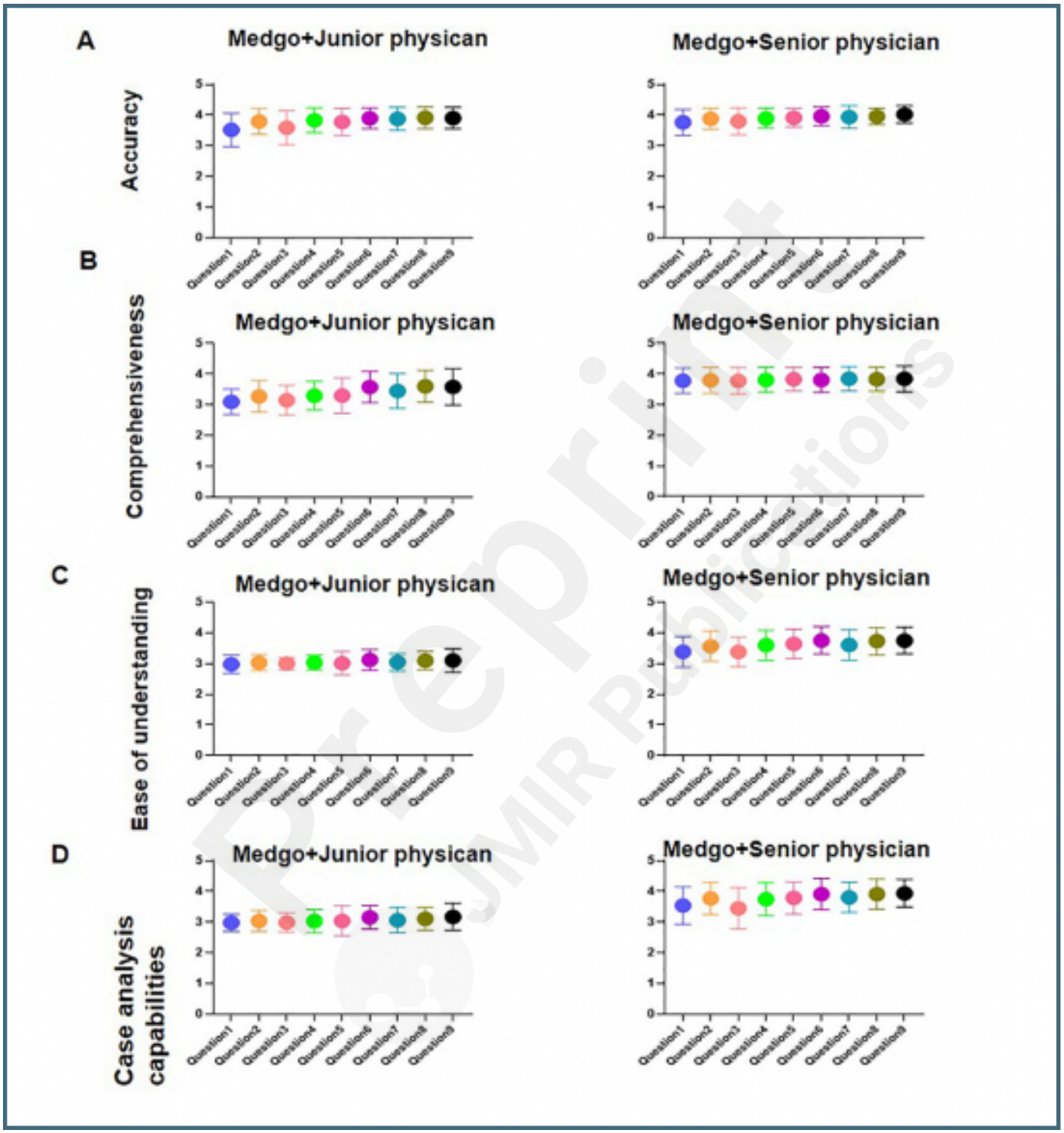
Flowchart showing the inclusion and exclusion of the patients.



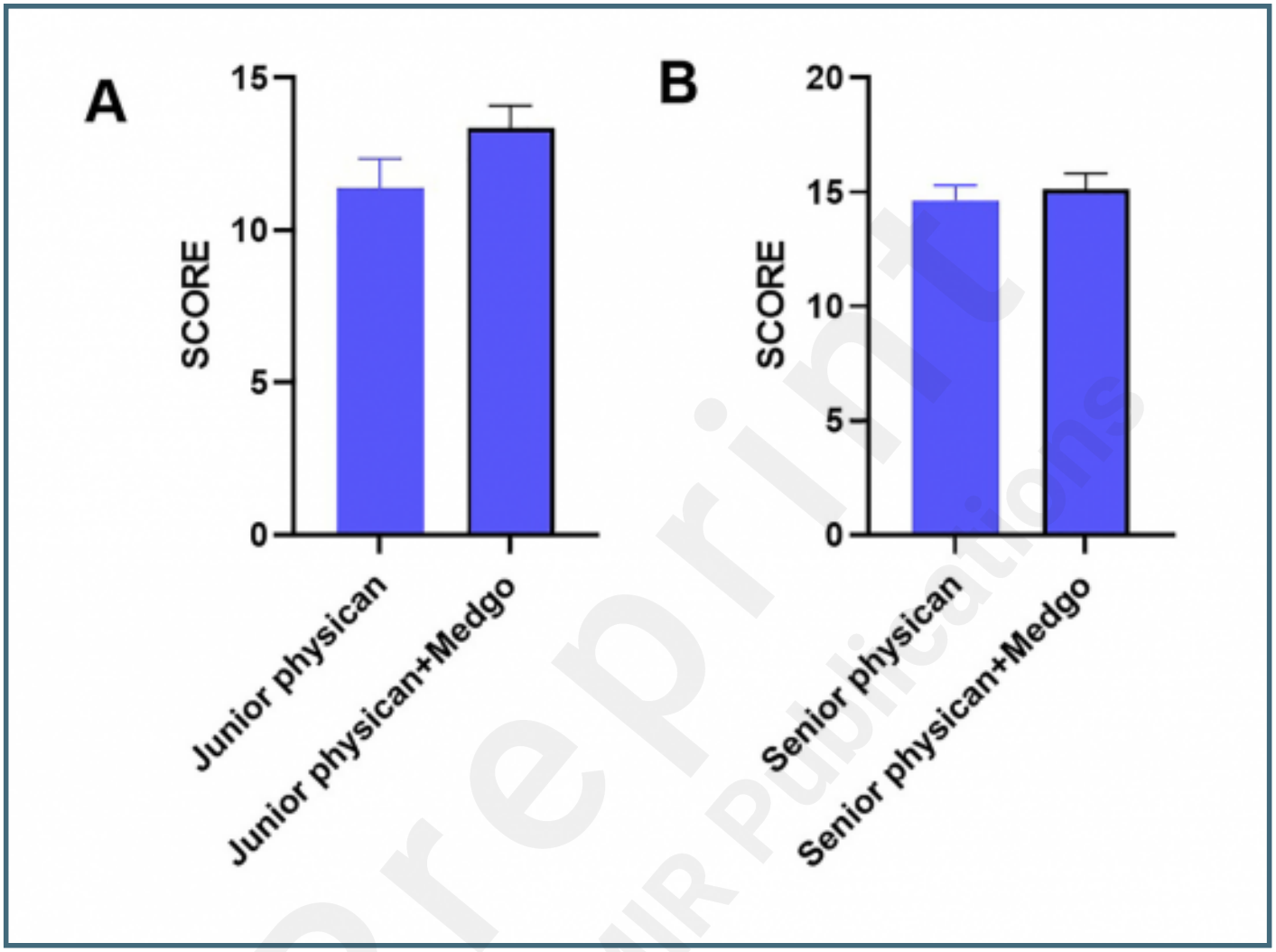
Clinical Decision-Making Performance of MedGo. The diagnostic performance of MedGo was assessed using a 5-point Likert scale, with higher scores indicating better performance.A, the performance of accuracy. B. the perf ormance of comprehensiveness. C, the performance of ease of understanding. D, the performance of case analysis capabilities. Each color in the figure represents a different question.



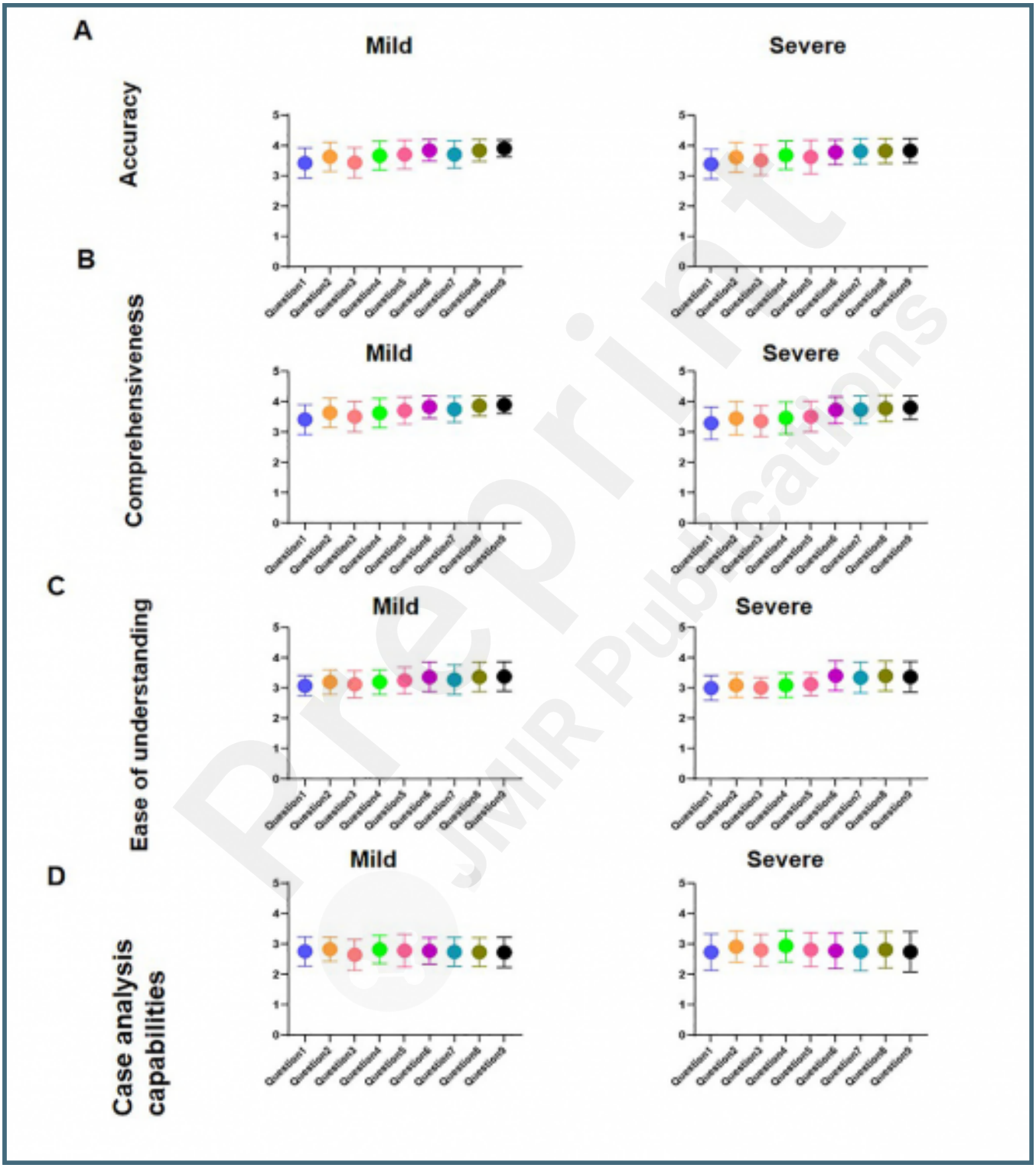
Performance of junior and senior physicians with MedGo assistance across nine assessment tasks.A, Accuracy. B, Comprehensiveness. C, Ease of understanding. D, Case analysis capabilities.



Comparison of overall clinical decision-making performance between physicians with and without MedGo assistance. Average total scores across nine assessment tasks for (A) junior physicians and (B) senior physicians, with and without the assistance of MedGo. Error bars represent standard deviation.



Performance of MedGo across different disease severity levels. The performance of MedGo was assessed using a 5-point Likert scale, with higher scores indicating better performance. A, Accuracy. B, Comprehensiveness. C, Ease of understanding. D, Case analysis capabilities. Each color in the figure represents a different question.



MedGo's performance in cases of varying difficulty. The total scores across all nine assessment tasks were averaged for each difficulty level (simple, moderate, difficult). Error bars represent standard deviation. One-way ANOVA showed significant differences in MedGo's performance across the three difficulty levels ($P < 0.05$), indicating that case complexity affects its overall performance.