Original Research Article

DIGITAL HEALTH

DIGITAL HEALTH
Volume 10: 1–16
© The Author(s) 2024
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/20552076241260504
journals.sagepub.com/home/dhj
Sage

Assessing usability of intelligent guidance chatbots in Chinese hospitals: Cross-sectional study

Yanni Yang[1], Siyang Liu[1], Ping Lei[2], Zhengwei Huang[3], Lu Liu[4] and Yiting Tan[1]

Abstract

Objective: This study aimed to assessing usability of intelligent guidance chatbots (IGCs) in Chinese hospitals.

Methods: A cross-sectional study based on expert survey was conducted between August to December 2023. The survey assessed the usability of chatbots in 590 Chinese hospitals. One-way ANOVA was used to analyze the impact of the number of functions, human-like characteristics, number of outpatients, and staff size on the usability of the IGCs.

Results: The results indicate that there are 273 (46.27%) hospitals scoring above 45 points. In terms of function development, 581(98.47%) hospitals have set the number of functions between 1 and 5. Besides, 350 hospitals have excellent function implementation, accounting for 59.32%. In terms of the IGC's human-like characteristic, 220 hospitals have both an avatar and a nickname. Results of One-way ANOVA show that, the number of functions($F = 202.667$, $P < 0.001$), human-like characteristics($F = 372.29$, $P < 0.001$), staff size($F = 9.846$, $P < 0.001$), and the number of outpatients($F = 5.709$, $P = 0.004$) have significant impact on the usability of hospital IGCs.

Conclusions: This study found that the differences in the usability of hospital IGCs at various levels of the number of functions, human-like characteristics, number of outpatients, and staff size. These findings provide insights for deploying hospital IGCs and can inform improvements in patient's experience and adoption of chatbots.

Keywords

Usability, intelligent guidance chatbots, medical guidance system, outpatient service, medical care

Submission date: 18 January 2024; Acceptance date: 23 May 2024

Introduction

The practice of medical care exhibits significant variation across different systems and cultural contexts.[1] In China, for example, medical care system that is heavily specialized, patients access medical care services without a referral from a family doctor and do not know what specialty or provider they need to visit before arriving in the hospital's lobby.[1] In China, patients seeking medical care have several options for scheduling appointments with doctors, including in-person visits to hospitals, phone calls, or booking online. As information technology advances, a growing number of Chinese patients are turning to online platforms

[1]School of Literature and Media, China Three Gorges University, Yichang, Hubei, China
[2]Department of Orthopedics, Zhijiang Hospital of Traditional Chinese Medicine, Zhijiang, Hubei, China
[3]College of Economics & Management, China Three Gorges University, Yichang, Hubei, China
[4]College of Electrical Engineering & New Energy, China Three Gorges University, Yichang, Hubei, China

Corresponding author:
Ping Lei, Department of Orthopedics, Zhijiang Hospital of Traditional Chinese Medicine, Tuanjie Road No. 18, Zhijiang, Hubei 443200, China.
Email: leiping1988@foxmail.com

for appointments. Data from China's National Health Commission indicates that over 50% of appointments in Chinese tertiary hospitals are now made online.[2] With this shift trend, an issue arises that patients frequently make an appointment with incorrect departments or doctors online, necessitating a department or doctor change when they arrive at hospital's lobby. To enhance outpatient service quality and facilitate patient access to medical care, numerous hospitals have initiated the implement of intelligent medical guidance systems. This initiative forms a crucial part of the hospitals' endeavor towards smart healthcare infrastructure. In this era of smart healthcare services, a range of intelligent applications are bridging the gap between patients and medical staff.[3,4]

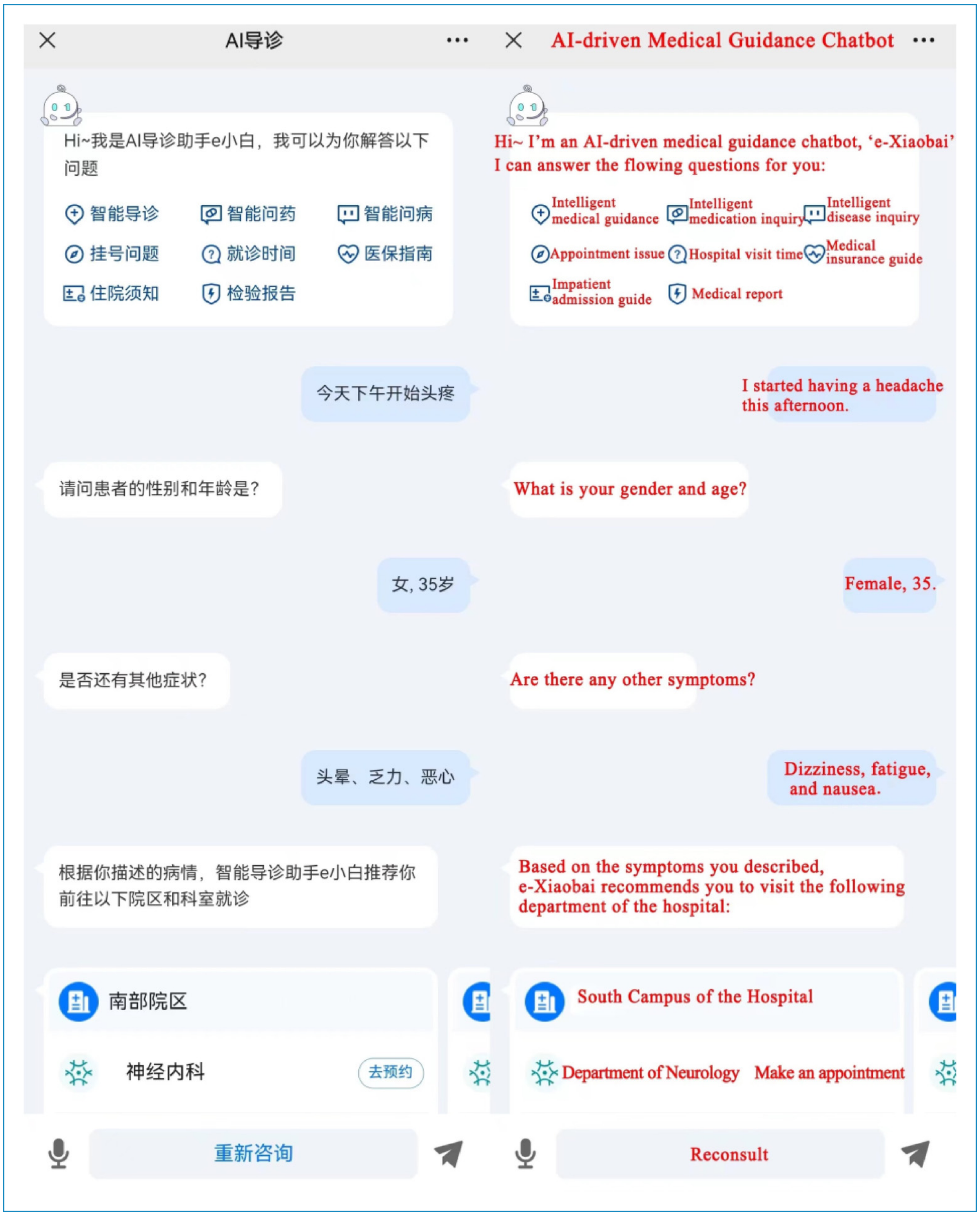
Artificial intelligence (AI)-driven conversational agents, offering medical guidance, are increasingly becoming pivotal in the facilitation and access to medical services.[5,6] These medical guidance agents, driven by conversational AI, leverage technologies such as natural language processing and machine learning. Such integration allows these systems to interpret and process human language, thereby enabling effective communication with patients and accurately discerning their needs to execute relevant tasks or provide appropriate responses.[7] In this research, we define these AI-driven conversational agents for medical guidance as intelligent guidance chatbots (IGCs; refer to Figure 1). Patients engage with these IGCs to acquire pre-visit information. During their conversation, patients can articulate symptoms to identify the most suitable medical department and arrange consultations with specialized physicians. These IGCs are instrumental in aiding patients to find the correct medical department or doctors and better comprehend patients' health conditions prior to their doctor visits.[8] Additionally, these IGCs can address user inquiries regarding medication, diseases, and other medical information, as well as provide guidance on healthcare insurance policies, hospitalization instructions, and more. Moreover, if the hospital offers specialized services, such as scheduling appointments with renowned experts, users can also obtain relevant information by engaging in dialogue with the IGCs.

Although IGCs are still in the exploratory phase, this technology has improved the accessibility of medical services for patients, reduced their reliance on manual guidance during hospital visits, and optimized the medical care process.[9] IGCs support patients in seeking medical advice anytime and anywhere,[10] which provide real-time feedback, assisting patients in understanding their symptoms, learning about their health conditions, and offering diagnostic suggestions.[11] Such chatbots as virtual conversational agents, mimicking human interaction, and directly provide medical advice to patients in a timely and cost-effective manner. For instance, IGCs facilitate appointment scheduling and online consultations, guide patients in answering a series of questions about their symptoms for disease diagnosis, and help confirm the severity of a disease.[12] In the realm of medical queries, employing IGCs as tools for medical consultation provides relevant information and decision support, streamlining the consultation process by which medical staff consult with patients, and answering patient queries and resolving doubts.[13]

Despite these potential benefits, similar to many other mobile health (mHealth) applications, IGCs have not yet been widely adopted by those who could benefit the most from this new technology.[14] Previous work has primarily focused on developing advanced algorithms to enhance the accuracy and effectiveness of the diagnostic capabilities of conversational agents.[15,16] However, it is necessary to improve navigational aids, such as intelligent guidance chatbots (IGCs), to streamline information flow and enhance patient outcomes. There has been limited research on the real-world usage of IGCs.[17] Most of the previous studies focus on the use of medical conversational bots in controlled environments and have limited understanding of the deployment of IGCs in various medical institutions, and the potential barriers and challenges associated with patients' use.[18]

Health information technology has been widely applied in the field of medical care, with a rich variety of dimensions and tools available for its evaluation. However, currently, there is no mature evaluation framework to support the assessment of IGCs. Previous evaluations of conversational agents have provided theoretical references, and scholars have proposed numerous methods for assessing medical conversational robots. These include the heuristic evaluation method proposed by Nielsen,[19] the international standard of ISO/IEC 25010 and ISO 9241-11,[20–23] and the use of questionnaires like SUS, UEQ, and Chatbottest.[13,24] Based on research characteristics, methods such as constructing an evaluation system that involves metrics like system accuracy, similarity to human conversation, user's satisfaction and perceived of usefulness, are used to assess the usability of conversational agents.[25–27] While previous research on health information technology evaluation is extensive in terms of dimensions and tools, an assessment framework specifically for IGCs has yet to be proposed.

To fill the aforementioned gap, we investigated the implement of IGCs in 590 tertiary grade-A hospitals in China, which is beneficial for understanding the usability of IGC in real-world settings, as well as the variations in usability under different technological conditions and organizational sizes. This study represents the first large-scale, real-world evaluation to investigate these issues. More specifically, we proposed the evaluation method to measure the usability of IGCs in these 590 tertiary hospitals. In addition, we highlighted the variations in usability under different technological and organizational conditions and the reasons for these variations, offering



Hi~我是AI导诊助手e小白，我可以为你解答以下问题

今天下午开始头疼

请问患者的性别和年龄是?

女，35岁

是否还有其他症状?

头晕、乏力、恶心

根据你描述的病情，智能导诊助手e小白推荐你前往以下院区和科室就诊

Hi~ I'm an AI-driven medical guidance chatbot, 'e-Xiaobai' I can answer the flowing questions for you:

I started having a headache this afternoon.

What is your gender and age?

Female, 35.

Are there any other symptoms?

Dizziness, fatigue, and nausea.

Based on the symptoms you described, e-Xiaobai recommends you to visit the following department of the hospital:

Figure 1. The interface of a medical guidance chatbot.

insights for optimizing medical guidance services and enhancing user experience.

Methods

Data sources and variables

An institution-based cross-sectional quantitative study was conducted at the China Three Gorges University, from August 31/2023 to December 01/2023. This study conducted a cross-sectional survey on whether the IGCs had been implemented by all 1651 tertiary grade-A hospitals in China (In Chinese medical system, hospitals are divided into multiple levels, among which the tertiary grade-A hospitals are the top medical institutions).[28] In this study, we initially compiled a list of 1651 hospitals. Subsequently, our team, comprising the three members who are also authors of this paper, conducted a preliminary assessment by systematically testing the IGCs in these hospitals. Through this process, we identified 603 hospitals that had implemented IGCs. However, during the subsequent data collection phase, we encountered usage issues with 13 of these hospitals, rendering it impossible to retrieve data from their IGCs. Therefore, we excluded these 13 hospitals from our final sample.

It was found that 590 hospitals had implemented intelligent guidance services and had complete data. Therefore, this study took the IGCs of these 590 tertiary grade-A hospitals in China as the research object. The required data was obtained from various sources such as the hospitals' intelligent guidance systems, official government websites, and relevant news reports about the hospitals, with the data collection period ending on August 31, 2023. We confirm that ethical approval is not needed as we did not conduct research on human subjects. Informed consent was not necessary for all of the analysis data to be publicly available from the website. Thus, the Ethics Committee of China Three Gorges University waived the need for informed consent. The details of the sample and data acquisition were mainly divided into the following two stages.

In the first phase, following Nielsen's research which suggests that five evaluators are optimal for finding usability problem,[29] we invited five experts with interdisciplinary backgrounds in health and information technology who had undergone training in evaluation methods. Two of them hold PhD in medical informatics and are actively involved in research pertaining to medical information management, while three are IT technicians engaged in the development of medical information systems. These experts were tasked with assessing the usability of the IGCs of these hospitals. Each expert was responsible for evaluating approximately 120 IGCs, ensuring a manageable workload. Consequently, each IGC was rated by one expert rater.

Furthermore, to ensure the reliability of the ratings, we conducted a preliminary assessment of inter- and intra-rater agreements. Prior to the formal evaluation, we tested the consistency of the experts using a small sample. Specifically, after providing training to the experts, we randomly selected 20 IGCs from different hospitals. Each expert assessed these 20 IGCs twice, with a one-month interval between assessments. The results of our analysis demonstrated strong inter-rater reliability, with an intraclass correlation coefficient (ICC) of 0.91 ($P < 0.001$) for the first assessment and 0.94 ($P < 0.001$) for the second assessment. These findings indicate high consistency among the expert raters in their evaluations. Additionally, the intra-rater reliability was also robust, with ICC coefficients exceeding 0.75 ($P < 0.001$) for both the first and second assessments, indicating consistency in evaluations over time.

In the second phase, we analyzed the function development and the human-like characteristics of IGCs of these 590 hospitals. In terms of function development, the Chinese content, implementation and the number of functions were analyzed. The content of function denotes the collection of specific functions included in an intelligent guidance chatbot, such as department recommendation, medical queries, or vaccine information. The implementation of function refers to the extent to which a series of functions of the intelligent guidance chatbot can be realized. Whether the chatbot can navigate patients to the appropriate department based on their described symptoms, or answer questions about medications based on user queries. For human-like characteristics, we primarily focused on whether the sample IGCs had a nickname or an avatar. Additionally, we gathered data on the staff size and the number of outpatient of the sample hospitals, which had been published on their official websites. To simplify the data analysis process and enhance the effectiveness and reliability of data analysis, we conducted a group analysis of the following variables. The grouping details are as presented in Table 1. The grouping of the staff size and the number of outpatients are based on the distribution characteristics of the data sample in this study.

These 590 hospitals are distributed across 31 provinces and municipalities in China (data from Hong Kong, Macau, and Taiwan not included). The diversity in geographical distribution and technological application among these hospitals makes them well-suited to meet the research needs.

Usability evaluation for IGCs

Usability is a key indicator of the quality of user experience. Usability assessments enable developers of medical information systems to promptly identify issues that arise during human-computer interactions. As a classic method for assessing usability, Nielsen's ten principles evaluate the usability of products or services from the user's perspective. It is important to note that Nielsen's principles represent a general evaluation framework. For this study, we have developed a usability evaluation scale for IGC of hospital based on Nielsen's usability heuristics, adapted to the domain-specific features and interface characteristics of these chatbots, as shown in Table 2. To quantitatively evaluate our study sample, each indicator is described and assigned a value, with the cumulative total of these values across the 10 indicators constituting the measure of usability.

Statistical analysis methods

To investigate the differences in the usability of intelligent guiding chatbots from technical and organizational aspects, we analyzed the impact of four variables: the number of functions, human-like characteristics, staff size, and the numbers of outpatient. First, we grouped each variable, then used SPSS 25.0 software to conduct tests for normality and homogeneity of variance test. And then, we performed analysis of variance (ANOVA) and used the Least Significant Difference (LSD) method to compare the significance of differences between sample groups. Modern statistical research and practice suggest that factor analysis of variance can be effectively performed when the sample size is sufficient and the data distribution is approximately normal. Therefore, in the process of performing normality

Table 1. Variable definition and grouping.

	Variable definition	Grouping of variables
Number of functions	It indicates the number of specific functions developed in an intelligent guidance chatbot.	Group 1: the number of functions is between 1 and 2. Group 2: the number of functions is 3 or more.
Human-like characteristics	It represents the assignment of a nickname or an avatar to the intelligent guidance chatbot.	Group 1: An IGC has no nickname or avatar. Group 2: An IGC has either a nickname or an avatar.
Staff size	It refers to the latest number of employed staff as published by each case sample hospital.	Group 1: A hospital has a staff size of fewer than 1500 people. Group 2: A hospital's staff size ranges between 1500 and 3000 people. Group 3: A hospital has a staff size of more than 3000 people.
Number of outpatients	It indicates the latest yearly number of outpatient visits as published by each case sample hospital.	Group 1: A hospital has fewer than 500 thousands outpatient visits. Group 2: The number of outpatients of a hospital ranges between 500 thousands and 1.5 million. Group 3: A hospital has more than 1.5 million outpatient visits.

tests, we conducted observation analysis by drawing Q-Q plots. Additionally, given the wide applicability of the Levene test, this study used the Levene test to examine the homogeneity of variances of the data, determining whether the variances of each sample group are equal to the variables.

Results

Descriptive statistical analysis

Result of usability evaluation. According to the usability evaluation method for IGCs proposed in this study, the statistical results indicate that there are 273 (46.27%) hospitals scoring above 45 points (50 points is the maximum score), 307 (52.03%) hospitals scoring between 35 to 44 points, and 10 (1.7%) hospitals scoring below 35 points. Among hospitals that have implemented IGCs, the usability evaluation results show that some hospitals have developed their intelligent guidance chatbots well, the majority still need to strengthen their efforts. The horizontal comparison was made of the scores obtained by various hospitals across 10 indicators, as shown in Figure 2.

The results show that in the three indicators of aesthetics and minimalist design (589/590, number of hospitals with a score of 5/total sample size), flexibility and efficiency of use (570/590), and recognition rather than recall (584/590), the majority of these hospitals' IGCs scored 5 points. Regarding the two indicators of consistency and standards (474/590) and help users recognize, diagnose, and recover from errors (447/590), about 80% of hospitals have achieved good performance. For the three indicators of visibility of system status (350/590), user control and freedom (377/590), and error prevention (387/590), about 60% of hospitals have achieved good performance. In addition, for indicators consistency and standards (293/590), help and documentation (232/590), around 50% of the hospitals' chatbots have suboptimal scores.

Results of function development and human-like characteristics. Function development and human-like characteristics technically reflect the implement status of IGCs in hospitals. Sample analysis results indicate that the main functions of IGCs are appointment for hospital visits, medical enquiries, medical insurance guidelines, medical test result inquiries, inpatient guidelines, and vaccine searches. However, many hospitals have devised many functions in IGCs, these functions may not be implemented effectively. A statistical analysis was conducted on the number of functions and the implementation status of these functions of the 590 hospitals' IGCs, with results presented in Figure 3. The implementation status of the IGCs' functions was divided into five levels; the number of functions was counted based on what is displayed on the interface of the hospital's IGC. Overall, there are 76 hospitals with one function setting, 78 hospitals with two function settings, 30 hospitals with three function settings, 103 hospitals with four function settings, and 57 hospitals with five function settings, 98.47% (581/590) of hospitals have set the number of functions between 1 and 5. Besides, 350 hospitals have excellent function implementation, accounting for 59.32% (350/590).

Table 2. Usability evaluation scale for IGCs.

Nielsen usability heuristic	Description of usability indicators	Indicator scoring
Visibility of system status	The chatbot displays clear navigation steps and current status, informing users about what is happening and their position in the navigation process.	Scoring range: 1 to 5 points • Barely functional with access to the system dialogue interface and no status prompts, scores 1 point; • Partially functional system navigation with basic status prompts that mostly meet the needs, scores 3 points; • Fully functional system navigation with very clear, timely, and helpful status prompts, scores 5 points; • 2 and 4 points are awarded for performance levels that fall between 1 to 3 and 3 to 5, respectively.
Match between system and the real world	The chatbot's system dialogue is designed in alignment with the users' real-world habits, utilizing easily understandable and familiar medical terminology or colloquial expressions, while avoiding overly complex or specialized vocabulary.	Scoring range: 1 to 5 points • The dialogue design is obscure and difficult to understand, scoring 1 point; • The dialogue design generally aligns with user habits, with occasional uncommon professional terms, scoring 3 points; • The dialogue design fully conforms to user habits and is easily understandable by users, scoring 5 points; • 2 and 4 points are awarded for performance levels that fall between 1 to 3 and 3 to 5, respectively.
User control and freedom	If a user mistakenly selects an option or navigates to the wrong page, they can easily revert to the previous state without losing any information.	Scoring range: 1 to 5 points • Users have no option for undoing or redoing actions, with no freedom in navigation and forced to follow a fixed path, scoring 1 point; • Undo and redo options are available but need improvement for better usability, offering some degree of navigational freedom but with limitations, scoring 3 points; • Undo and redo options are clearly user-friendly, allowing for unrestricted navigation and transitions, scoring 5 points; • 2 and 4 points are given for performance levels that fall between 1 to 3 and 3 to 5, respectively.
Consistency and standards	The chatbot system maintains consistency in design and operation across different pages and sections.	Scoring range: 1 to 5 points • Interface elements are inconsistent, using a variety of terms or languages to describe the same operation or function, and the workflow for similar tasks is completely different, scoring 1 point; • Most interface elements are consistent, with terms or languages used for the same operation or function being consistent in most cases, and the workflow for similar tasks is largely consistent, scoring 3 points; • Interface is completely consistent throughout the system, using uniform terms and language, and the workflow for all similar tasks is entirely consistent, scoring 5 points; • 2 and 4 points are awarded for performance levels that fall between 1 to 3 and 3 to 5, respectively.

(continued)

Table 2. Continued.

Nielsen usability heuristic	Description of usability indicators	Indicator scoring
Error prevention	The chatbot incorporates measures to prevent user errors.	Scoring range: 1 to 5 points • There are no confirmation steps for important or irreversible operations, the interface has few or unclear guides and prompts that can lead to misunderstandings, error messages are vague or offer no solutions, scoring 1 point; • Some key operations include confirmation steps, the interface provides basic guidance and prompts, error messages are generally clear and offer some solutions, scoring 3 points; • All significant and irreversible operations require user confirmation, the interface provides clear guidance and prompts, almost eliminating the possibility of errors, error messages are very clear and provide comprehensive solutions, scoring 5 points; • 2 and 4 points are assigned for performance levels that fall between 1 to 3 and 3 to 5, respectively.
Recognition rather than recall	Users can easily find various commands for the chatbot service without the need to deliberately memorize how to operate it.	Scoring range: 1 to 5 points • Almost no available options are clearly visible, there is almost no step indication in multi-step tasks, it fails to remember any history or commonly used options, and there are almost no prompts or explanations before critical tasks, scoring 1 point; • Some available options are clearly visible, step indications are mostly clear, it can remember some history and commonly used options, and most critical tasks have some prompts or explanations, scoring 3 points; • All available options are clearly visible, every multi-step task has clear step-by-step instructions, it provides convenient access to history and commonly used options, and every critical task is preceded by clear prompts or explanations, scoring 5 points; • 2 and 4 points are allocated for performance levels that fall between 1 to 3 and 3 to 5, respectively.
Flexibility and efficiency of use	The chatbot is efficient and flexible, suitable for both novices and experts, and allows for frequent user interactions.	Scoring range: 1 to 5 points • All users are forced to use complex operation processes with almost no real-time feedback, scoring 1 point; • Different levels of operational complexity are provided, and most important actions have some form of immediate feedback, scoring 3 points; • Highly flexible and personalized operation processes are available for users of different experience levels, and all operations come with timely and clear real-time feedback, scoring 5 points; • 2 and 4 points are given for performance levels that fall between 1 to 3 and 3 to 5, respectively.

(continued)

Table 2. Continued.

Nielsen usability heuristic	Description of usability indicators	Indicator scoring
Aesthetic and minimalist design	The chatbot features a simple and aesthetically pleasing design, free from irrelevant content and superfluous information.	Scoring range: 1 to 5 points • The interface is overly complex, with many unnecessary and redundant elements, and overly dense information, scoring 1 point; • Interface elements are generally reasonable, but there is still room for improvement, with a moderate information density, scoring 3 points; • Every interface element is necessary, with no redundant elements, and the information layout is logical and well-organized, scoring 5 points; • 2 and 4 points are awarded for performance levels that fall between 1 to 3 and 3 to 5, respectively.
Help users recognize, diagnose, and recover from errors	The chatbot accurately identifies issues that arise during the service and provides effective solutions.	Scoring range: 1 to 5 points • Error messages are vague or missing, errors are difficult to fix, and there is no way to report errors or get support, scoring 1 point; • Error messages are basically clear, offering some solutions, with basic error reporting and support available, scoring 3 points; • Error messages are very clear, providing explicit steps for resolution, clear methods for error reporting, and immediate and efficient support, scoring 5 points; • 2 and 4 points are awarded for performance levels that fall between 1 to 3 and 3 to 5, respectively.
Help and documentation	The chatbot includes features akin to help documents, offering prompts and assistance.	Scoring range: 1 to 5 points • Help documentation is very hard to find, with no practical examples or FAQs provided, scoring 1 point; • Help documentation is relatively conspicuous but still needs improvement, offering some practical examples or FAQs, scoring 3 points; • Help and documentation are very easy to find, providing a comprehensive and practical set of examples or FAQs, scoring 5 points; • 2 and 4 points are given for performance levels that fall between 1 to 3 and 3 to 5, respectively.

In terms of the IGC's human-like characteristic, 220 hospitals have both an avatar and a nickname, 103 have either an avatar or a nickname, and 267 have neither. We found that the more detailed the human-like characteristics of IGC, such as having a nickname or an avatar, the more advanced the system development tends to be. A well-crafted human-like characteristics of IGC enhance the user experience, making human-chatbot interaction more natural and appealing. Furthermore, it was found that nicknames often have characteristics like homogeneity, cuteness, and user-friendliness, with examples like 'Xiao Mi', 'Xiao Rui', 'Xiao Die', 'Doctor Xiao Ci', etc. The avatar, on the other hand, are mostly represented by cartoon robots, cartoon nurses, cartoon doctors, and other cartoon figures. The following are the nicknames and avatars of some of the IGCs, as shown in Figure 4.

Results of number of outpatients and staff size of hospitals. There were approximately 8.04 billion outpatient visits nationwide in China, with around 3.88 billion of these occurring in hospitals across various regions. The statistics indicate that in the most recent year, the total number of outpatient visits in these 590 hospitals amounted to about 705 million. On average, these hospitals had about 1.1949 million outpatient visits per year, and the average staff size was approximately 2126 employees. Overall, the vast

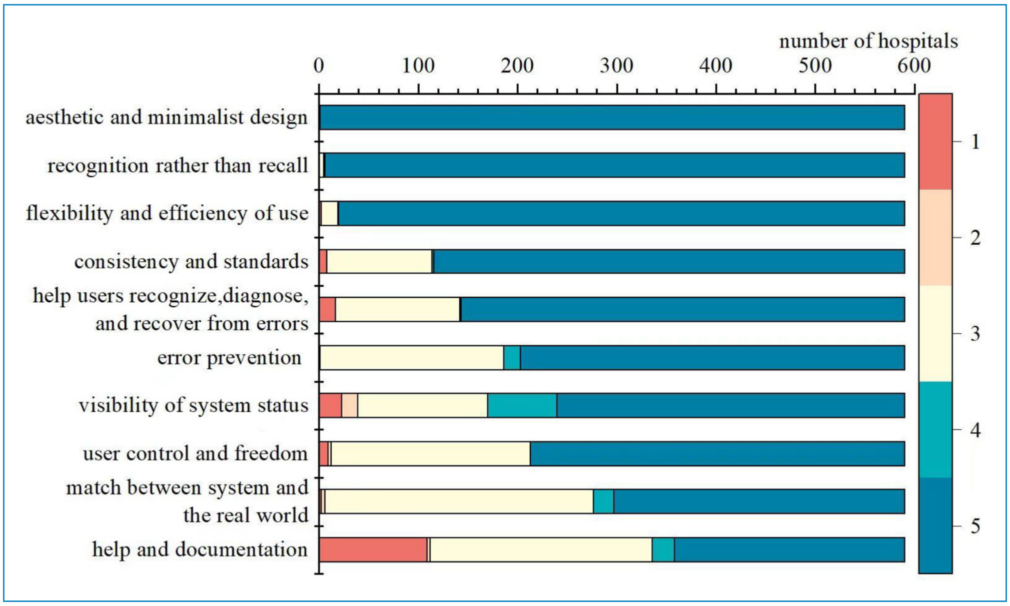


Figure 2. Comparison of the scores across 10 indicators of usability evaluation.

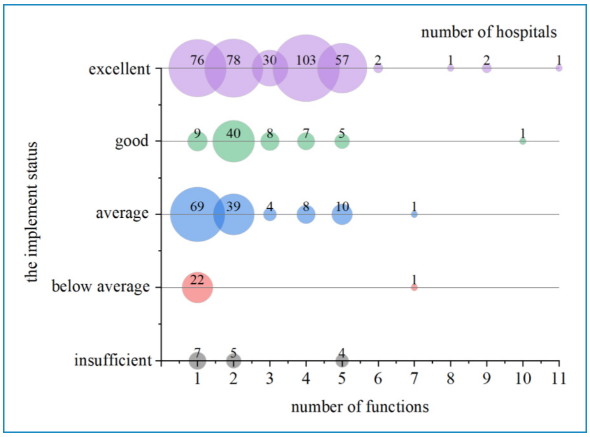


Figure 3. Results of the implementation status and number of functions.

majority of hospitals had a large number of outpatient visits, resulting in high demand for outpatient consultation and a heavy workload of outpatient tasks.

Results of analysis of variance (ANOVA)

Normality test. Q-Q plots were generated using SPSS 25.0 to illustrate the usability of IGCs at various levels of four variables: the number of functions, human-like characteristics, staff size, and number of outpatients. The results are shown in Appendix 1 to 4. In these plots, the diagonal line represents the Q-Q line that would be expected if the usability of IGCs followed a normal distribution, and the blue dots represent the actual scatter plot of the usability of IGCs. From these graphs, it can be observed that the usability of IGCs with these four variables at different levels is concentrated near the normal Q-Q line, indicating that overall, they approximately follow a normal distribution.

Homogeneity of variance test. Before conducting a one-way analysis of variance (ANOVA), a test for homogeneity of variance was performed on the data. This test was applied to the four variables: number of functions, human-like characteristics, staff size, and number of outpatients. The test results, as shown in Table 3, indicate that all four variables had significance results greater than 0.05, demonstrating homogeneity of variances. Combined with the results of the normality test, these four variables meet the prerequisite conditions for conducting an ANOVA.

Results of one-way analysis of variance. A one-way ANOVA was conducted on the four variables respectively, resulting in descriptions of samples with different distributions (see Table 4) and the results of ANOVA (see Table 5). Take the variable of number of functions as an example, it can be observed that the mean value of usability for Group 1 is 42.29, while for Group 2 it is 46.89, indicating a significant difference between the two groups. Further, as shown in Table 5, the P-value of the number of functions is less than the significance level α=0.05, and the F-value is also relatively high. This suggests that there is a significant difference in the overall mean values of the usability at different levels of the number of functions. The results of Tables 4 and 5 show that the four variables both have significant impact on the usability of IGCs.

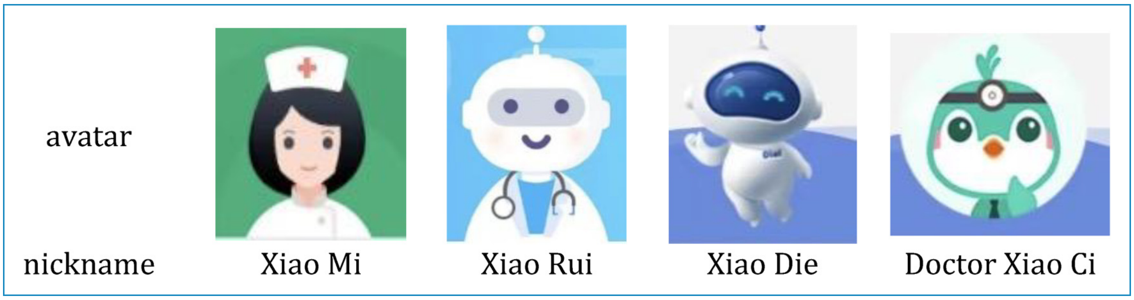


Figure 4. Nicknames and avatars of some of the IGCs.

Table 3. Levene test of homogeneity of variance.

	Levene statistic	Df1	Df2	Significance (P value)
Number of functions	0.214	1	588	0.644
Human-like characteristics	2.251	1	588	0.134
Staff size	1.663	2	587	0.191
Number of outpatients	2.149	2	587	0.117

Table 4. Sample descriptions for different distributions of the four variables.

						95% confidence interval of the mean	
Variables	Grouping	N	Mean	Standard deviation	Standard error	Lower bound	Upper bound
Number of functions	Group 1 (1~2)	345	42.29	3.774	0.203	41.89	42.69
	Group 2 (>=3)	245	46.89	3.982	0.254	46.38	47.39
Human-like characteristics	Group 1(0)	267	41.14	3.234	0.198	40.75	41.53
	Group 2(1~2)	323	46.73	3.712	0.207	46.32	47.14
Staff size	Group 1(<1500)	261	43.32	4.674	.289	42.75	43.89
	Group 2(1500~3000)	201	44.69	4.270	.301	44.10	45.29
	Group 3(>3000)	128	45.22	4.049	.358	44.51	45.93
Number of outpatients	Group 1 (<50)	196	43.34	4.571	.326	42.70	43.99
	Group 2 (50~150)	248	44.51	4.584	.291	43.94	45.09
	Group 3 (>150)	146	44.82	3.984	.330	44.17	45.47
	Total	590	44.20	4.474	.184	43.84	44.56

Table 5. Results of the variance analysis for different distributions of the four variables.

Variables	Source of variance	Sum of squares (SS)	Degree of freedom (Df)	Mean square (MS)	F value	P value
Number of functions	Between Groups	3022.168	1	3022.168	202.667	<0.001
	Within Groups	8768.232	588	14.912		
Human-like characteristics	Between Groups	4570.961	1	4570.961	372.290	<0.001
	Within Groups	7219.439	588	12.278		
Staff size	Between Groups	382.684	2	191.342	9.846	<0.001
	Within Groups	11407.716	587	19.434		
Number of outpatients	Between Groups	224.969	2	112.485	5.709	.004
	Within Groups	11565.431	587	19.703		
	Total	11790.400	589			

Moreover, given the significant impact of staff size on the usability of IGCs, it is necessary to perform post-hoc analyses to identify which specific groups exhibit significant differences. For this purpose, this paper employs the LSD test for post-hoc multiple comparison analysis. The test results, as shown in Table 6, indicate that there are significant differences in the impact on thstem and the real e usability of IGCs between Group 1 (small hospitals) and both Group 2 (medium-sized hospitals) and Group 3 (large hospitals). However, there is no significant difference between Group 2 and Group 3. Besides, we conducted the same test for the variable of the number of outpatients, as shown in Table 6. The results indicate significant differences in the impact on the usability of IGCs between Group 1 (<50) and both Group 2 (50∼150) and Group 3 (>150), while no significant difference exists between Group 2 and 3.

Discussion

Principal findings

Nearly half of the sample hospitals have suboptimal scores for five indicators of IGC's usability. Analysis of the usability of IGCs in 590 hospitals shows that many hospitals have suboptimal performance in terms of visibility of system status, user control and freedom, error prevention, match between the system and the real world, and help and documentation. Approximately 40% of the hospitals' chatbots exhibited average performance in the indicators of visibility of system status, user control and freedom, and error prevention. These indicators reflect issues in the system design of the IGCs. The average performance in visibility of system status can be attributed to inadequate interface interaction design and a lack of effective feedback mechanisms, making it difficult for users to perceive the current system status of the chatbot. The lower scores in user control and freedom are primarily due to a lack of flexible navigation options in the system, such as 'undo', 'redo', or 'back' steps, which can make users feel constrained during operation and unable to freely explore or correct errors. The lower scores in error prevention involve not fully considering possible user errors or lacking effective mechanisms to prevent errors.

Therefore, in the system design of IGCs, firstly, interfaces of chatbots should be easy to understand and use, offering clear status indicators such as progress bars and confirmation messages to enhance the visibility of the system's status. Secondly, it is crucial to ensure that users have sufficient control during their interactions with the chatbots, such as allowing them to interrupt operations at any time or providing a variety of choices for user-driven decision-making. Thirdly, error-tolerant interaction processes need to be designed, offering clear error messages and quick recovery options, enabling users to easily correct mistakes and return to the correct path.

Additionally, for the indicators of match between the system and the real world, and help and documentation, around 50% of the hospitals' chatbots have suboptimal scores. The issue arises from the chatbots employing language that is too technical or specialized, or their manner of conversation not matching the linguistic habits and cultural background of the target user group, resulting in difficulties for users in understanding and communication. The low scores for the help and documentation are due to insufficient help resources, difficulty in accessing or using the help function, and delay in updating help documentation.

Table 6. Multiple comparison results for different staff size distributions.

						95% confidence interval	
variables	Grouping	Comparative group	Mean difference	Standard error	P value	Lower bound	Upper bound
Staff size	Group 1 (<1500)	Group 2	−1.370[a]	.414	.001	−2.18	-.56
		Group 3	−1.897[a]	.476	<0.001	−2.83	-.96
	Group 2 (1500∼3000)	Group 1	1.370[a]	.414	.001	.56	2.18
		Group 3	-.527	.499	.291	−1.51	.45
	Group 3 (>3000)	Group 1	1.897[a]	.476	<0.001	.96	2.83
		Group 2	.527	.499	.291	-.45	1.51
Number of outpatients	Group 1 (<50)	Group 2	−1.170[a]	.424	.006	−2.00	-.34
		Group 3	−1.480[a]	.485	.002	−2.43	-.53
	Group 2 (50∼150)	Group 1	1.170[a]	.424	.006	.34	2.00
		Group 3	-.310	.463	.504	−1.22	.60
	Group 3 (>150)	Group 1	1.480[a]	.485	.002	.53	2.43
		Group 2	.310	.463	.504	-.60	1.22

[a] $P < 0.05$.

Studies have pointed out that there are differences in the reading level of different users,[30,31] and it is recommended to use simpler words, shorter sentences, and remove complex medical terms in patient educational materials.[32] Therefore, in order to improve the efficiency of human-computer interaction, in the interaction interface with the chatbots, commands should be consistent with user language habits, avoiding technical jargon or complex commands. The help and documentation indicator is often overlooked. However, to enhance user understanding and empower users with self-service capabilities, it is crucial to strengthen the development of help systems and provide real-time, contextually relevant help information.

Function development significantly affects the usability of IGCs. In terms of function development, survey results indicate that the primary functions of IGCs in most hospital include department registration, medical inquiries, medical insurance guides, test result inquiries, inpatient guidelines, and vaccine searches. More than half of the hospitals have set 1 or 2 functions, and more than half have IGCs whose functions are well implemented. However, many hospitals have developed some functions, but these functions didn't really work out. Looking at the current well-established IGCs, such as those at Union Hospital, Tongji Medical College, Huazhong University of Science and Technology, or Shanghai Ninth People's Hospital, their advantages primarily manifest in the mature development of various functions, all capable of providing medical guidance to assist patients in problem-solving or offering help. For example, if a patient's primary task is to find a specific department in hospital, the chatbot should offer a straightforward and direct appointment function for department registration. To enhance the function development level of IGCs and meet patients' needs, it is crucial to conduct a detailed assessment of existing functions to identify issues and shortcomings, and to establish effective user feedback mechanisms to collect user opinions promptly. Furthermore, variance analysis results show that the number of functions significantly impacts the usability of IGCs. Although having a greater number of functions is not always better, the number setting of functions reflect the importance hospitals place on implementing IGCs.

Optimizing human-like characteristic design is beneficial for enhancing the usability of IGCs. In the aspect of human-like characteristic, among the 590 hospitals' IGCs, nearly half lack both nickname and avatar. Even for those that nickname or avatar are set, there is a high degree of homogeneity in the nicknames and avatars of chatbots across different hospitals. Moreover, variance analysis results show that human-like characteristic significantly impacts the usability of IGCs. The presentation of human-like characteristic is conducive to shortening the psychological distance with patients and increasing patients' trust and favorability in the hospital.[33] To highlight the human-like characteristics of chatbots, design elements could incorporate the hospital's history and the cultural features of its location. Alternatively, hospital staff, patients, and their families can be invited to participate in the design process of the chatbot's nickname and avatar through online surveys or social media interactions. Such user-involved design not only enhances the acceptance of the chatbots but also makes them more aligned with users' preferences and needs.

Hospitals of vary sizes have different strategies in the implement of IGCs. From the perspective of hospital's staff size, variance analysis results demonstrate that staff size has a significant impact on the usability of IGCs. There are notable differences in the usability of IGCs between small hospitals and medium-to-large-sized hospitals. Overall, hospitals with larger staff sizes tend to have higher usability scores for their chatbots. This indicates that deployment strategies for chatbots should differ according to the scale of the hospital. Large hospitals, with more resources, can invest in more advanced technologies and higher quality IGCs. For smaller hospitals, where resources may be more limited, the design of chatbot systems should avoid overly complex functions. Instead, the focus should be on core functionalities that enhance efficiency and patient satisfaction, such as basic medical guidance and information inquiry, ensuring that the chatbot's functions match the hospital's scale and needs.

The larger the amount of outpatient consultation the more necessary it is to enhancing the implement of IGCs. Regarding the number of outpatients, many hospitals in various regions have large numbers of outpatients, facing significant pressure in medical guidance. The variance analysis from this study shows that the number of outpatients significantly influences the usability of IGCs. There is a significant difference in usability between hospitals with fewer patients compared to those with more. Overall, hospitals with more outpatients tend to have higher usability scores for their chatbots. In hospitals with more outpatients, deploying IGCs is crucial to alleviate the pressure on medical consultation staff and improve service efficiency. On the one hand, technology is key to the effective functioning of chatbot systems, and continuous technological improvements and upgrades are essential to optimize outpatient medical consultation process and effectively handle the complex outpatient environment. On the other hand, increasing the acceptance of chatbots among hospital staff and patients is vital, which can be achieved through training or promotion, thereby enhancing users' confidence and ability to use chatbots.

Limitations

This study primarily explored the application and deployment of intelligent guidance chatbots in hospitals from the perspective of usability. In addition to the analysis of function development, human-like characteristics, hospital staff size, and the number of outpatients, which mainly focuses on the technical and organizational aspects, other factors such as economic conditions, internet development in a region, and even hospitals' attitudes towards technology can also influence their decision-making and deployment of IGCs in hospitals. For instance, medical organizations in economically prosperous areas may have more resources and funds to invest in new technologies. The economic environment of a hospital's location influences its decision-making regarding intelligent guidance chatbot adoption. Furthermore, in regions with well-developed internet infrastructures, high-speed and stable network connections provide the necessary foundation for the real-time operation and data exchange of IGCs. Therefore, in future research, we aim to explore the adoption behavior of hospitals towards IGCs from multiple perspectives. Besides, in terms of research methodology, we consider the use of regression analysis methods in the future to more comprehensively analyze the influencing factors of the usability of IGCs.

Conclusions

In China, an increasing number of outpatients are opting for online appointment scheduling, and AI-driven intelligent guidance chatbots are being adopted by more and more hospitals for outpatient appointment services. Previous research on intelligent guidance chatbots is not extensive. This paper analyzed the usability of IGCs in 590 tertiary hospitals in China. It is the first study to use a large-scale heterogeneous dataset to investigate the real-world usage of IGCs in hospitals. We assessed the usability of IGCs, function development, human-like characteristics of these 590 hospitals' chatbots. Using variance analysis, we explored the differences in the usability of hospital IGCs at various levels of the number of functions, human-like characteristics, number of outpatients, and staff size. These findings provide insights for deploying hospital IGCs and can inform improvements in patient's experience and adoption of chatbots.

This study makes three significant contributions. Firstly, we provide a detailed analysis of the use of AI-driven IGCs in Chinese hospitals, which is a topic still underexplored in the health informatics community. Secondly, previous studies often focused on a single case, such as evaluating a specific chatbot or other health information technology. Building on this theoretical foundation, we proposed a method for assessing the usability of IGCs, aimed at evaluating the development status of hospital IGCs. Thirdly, We analyzed how the usability of IGCs varies at different levels of function development, human-like characteristics, staff size, and the number of outpatients. We identify challenges and barriers that hinder the adoption and usage of these chatbots and provide references for optimizing hospital outpatient appointment services, enhancing patient's experience with using chatbots, and alleviating pressure in outpatient triage.

Acknowledgements: The authors extend their gratitude to Lihua Chen, Siying He, and Yaping Ouyang for their invaluable assistance in data collection. We also express our sincere thanks to all the participants in this study.

Contributorship: YY, SL, and PL researched the literature and conceived the study. YY, SL LL and YT were involved in data analysis. YY, PL, and ZH wrote the first draft of the manuscript. All authors reviewed and edited the manuscript and approved the final version of the manuscript.

Data availability: The datasets used and analyzed during the study are available from the corresponding author on reasonable request.

Declaration of conflicting interests: The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Ethical approval: We confirm that such approval is not needed as we did not conduct research on human subjects. Informed consent was not necessary for all of the analysis data to be publicly available from the website.

Funding: The authors received no financial support for the research, authorship, and/or publication of this article.

Guarantor: YY and PL.

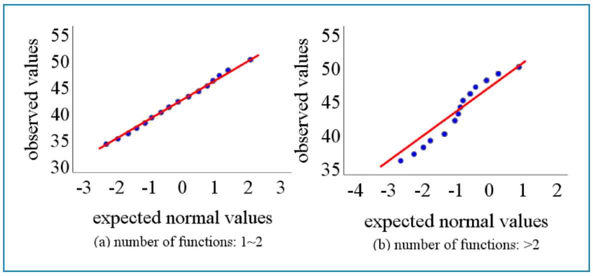
ORCID iD: Yanni Yang https://orcid.org/0000-0002-3878-9989

References

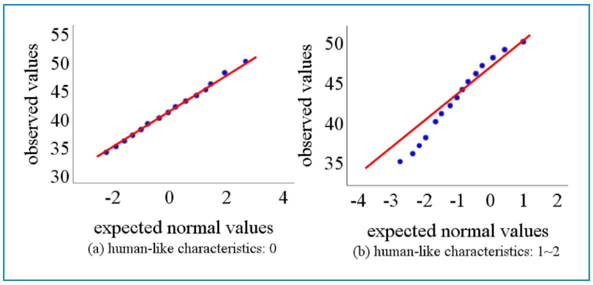
1. Ma AC, Meng Z and Ding X. Performance review of intelligent guidance robot at the outpatient clinic setting. *Cureus* 2021; 13: e16840.
2. National Health Commission of the People's Republic of China. Response to the "Proposal on Accelerating the Promotion of Mobile Medical Services". Available at: http://www.nhc.gov.cn/wjw/jiany/202111/c7ca63f972284bf2bf8bf3470e4ae1b6.shtml (2021, accessed 20 December 2023).
3. Palanica A, Flaschner P, Thommandram A, et al. Physicians' perceptions of chatbots in health care: cross-sectional web-based survey. *J Med Internet Res* 2019; 21: e12887.
4. He J, Baxter SL, Xu J, et al. The practical implementation of artificial intelligence technologies in medicine. *Nat Med* 2019; 25: 30–36.
5. Bibault JE, Chaix B, Nectoux P, et al. Healthcare ex Machina: are conversational agents ready for prime time in oncology? *Clinical and Translational Radiation Oncology* 2019; 16: 55–59.
6. Valtolina S, Barricelli BR and Di Gaetano S. Communicability of traditional interfaces VS chatbots in healthcare and smart home domains. *Behavior & Information Technology* 2020; 39: 108–132.
7. Hsu IC and Yu JD. A medical chatbot using machine learning and natural language understanding. *Multimed Tools Appl* 2022; 81: 23777–23799.
8. Deng Z, Tian Z, Xue J, et al. What predicts patients' satisfaction and continuous use of intelligent medical guidance? The moderating effect of consulting experience. *Behavior & Information Technology* 2023; 11: 1–18.
9. Montenegro JLZ, da Costa CA and da Rosa Righi R. Survey of conversational agents in health. *Expert Syst Appl* 2019; 129: 56–67.
10. Meyer AND, Giardina TD, Spitzmueller C, et al. Patient perspectives on the usefulness of an artificial intelligence–assisted symptom checker: cross-sectional survey study. *J Med Internet Res* 2020; 22: e14679.
11. Hamet P and Tremblay J. Artificial intelligence in medicine. *Metabolism* 2017; 69: 36–40.
12. Jovanović M, Baez M and Casati F. Chatbots as conversational healthcare services. *IEEE Internet Comput* 2020; 25: 44–51.
13. Minutolo A, Damiano E, De Pietro G, et al. A conversational agent for querying Italian patient information leaflets and improving health literacy. *Comput Biol Med* 2022; 141: 105004.
14. Laumer S, Maier CF and Gubler FT. Chatbot acceptance in healthcare: Explaining user adoption of conversational agents for disease diagnosis. In: Proceeding of the 27th European Conference on Information System, 2019.
15. Liu J, Li C, Huang Y, et al. An intelligent medical guidance and recommendation model driven by patient-physician communication data. *Front Public Health* 2023; 11: 1–19.
16. Zhou X, Li Y and Liang W. CNN-RNN based intelligent recommendation for online medical pre-diagnosis support. *IEEE/ACM Trans Comput Biol Bioinf* 2020; 18: 912–921.
17. Nadarzynski T, Miles O, Cowie A, et al. Acceptability of artificial intelligence (AI)-led chatbot services in healthcare: a mixed-methods study. *Digital Health* 2019; 5: 1–12.
18. Fan X, Chao D, Zhang Z, et al. Utilization of self-diagnosis health chatbots in real-world settings: case study. *J Med Internet Res* 2021; 23: e19928.
19. Nielsen J. Finding usability problems through heuristic evaluation. In: Proceedings of the SIGCHI conference on Human factors in computing systems, 1992.
20. Santa Barletta V, Caivano D, Colizzi L, et al. Clinical-chatbot AHP evaluation based on "quality in use" of ISO/IEC 25010. *Int J Med Inf* 2023; 170: 104951.

21. Moldt JA, Festl-Wietek T, Mamlouk AM, et al. Assessing medical students' perceived stress levels by comparing a chatbot-based approach to the perceived stress questionnaire (PSQ20) in a mixed-methods study. *Digital Health* 2022; 8: 1–12.
22. Ghaleb M, Almurtadha Y, Algarni F, et al. Mining the chatbot brain to improve COVID-19 bot response accuracy. *Computers, Materials & Continua* 2022; 70: 2619–2638.
23. Cao H, Zhang Z, Evans RD, et al. Barriers and enablers to the implementation of intelligent guidance systems for patients in Chinese tertiary transfer hospitals: usability evaluation. *IEEE Trans Eng Manage* 2021; 70: 2634–2643.
24. Omoregbe NAI, Ndaman IO, Misra S, et al. Text messaging-based medical diagnosis using natural language processing and fuzzy logic. *J Healthc Eng* 2020; 2020: 1–14.
25. Chaix B, Bibault JE, Romain R, et al. Assessing the performances of a chatbot to collect real-life data of patients suffering from primary headache disorders. *Digital Health* 2022; 8: 20552076221097783.
26. Hong G, Smith M and Lin S. The AI will see you now: feasibility and acceptability of a conversational AI medical interviewing system. *JMIR Formative Research* 2022; 6: e37028.
27. Nam KH, Kim DH, Lee JH, et al. Conversational artificial intelligence for spinal pain questionnaire: validation and user satisfaction. *Neurospine* 2022; 19: 348–356.
28. National Health Commission of the People's Republic of China. China Health Statistical Yearbook 2022. Available at: http://www.nhc.gov.cn/mohwsbwstjxxzx/tjtjnj/202305/6ef68aac6bd14c1eb9375e01a0faa1fb/files/b05b3d958fc546d98261d165cea4adba.pdf (2023, accessed 20 December 2023).
29. Nielsen J. How to conduct a heuristic evaluation. Available at: https://www.ingenieriasimple.com/usabilidad/HeuristicEvaluation.pdf (1995, accessed 15 November 2023).
30. Prabhu AV, Gupta R, Kim C, et al. Patient education materials in dermatology: addressing the health literacy needs of patients. *JAMA Dermatol* 2016; 152: 946–947.
31. Gajjar AA, Patel S, Patel SV, et al. Readability of cerebrovascular diseases online educational material from major cerebrovascular organizations. *J Neurointerv Surg* 2024. Available at: https://jnis.bmj.com/content/early/2024/02/23/jnis-2023-021205 (2024, accessed 13 April 2024).
32. Kamath P, Zheng R, Narasimman M, et al. Evaluation of online patient education materials concerning skin cancers. *J Am Acad Dermatol* 2021; 84: 190–191.
33. Graham S, Depp C, Lee EE, et al. Artificial intelligence for mental health and mental illnesses: an overview. *Curr Psychiatry Rep* 2019; 21: 1–18.

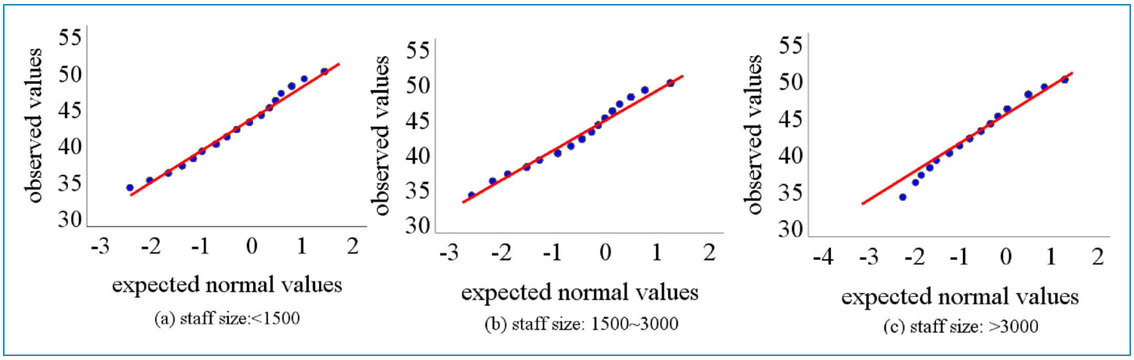
Appendices



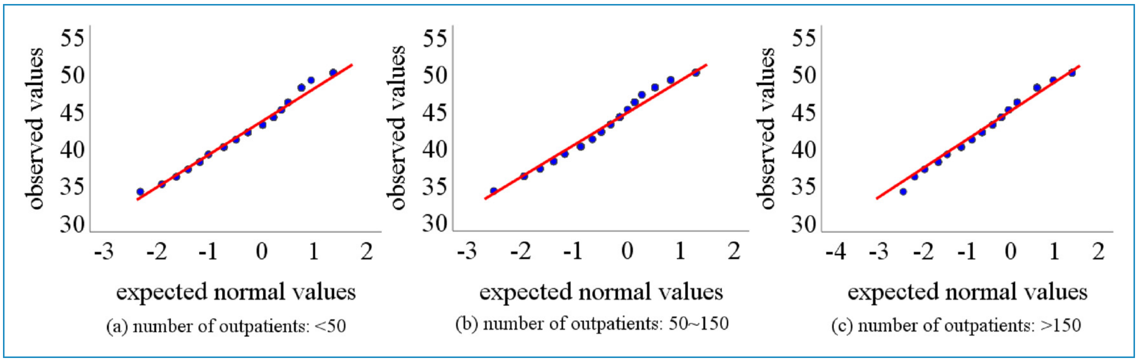
Appendix 1. Q-Q Plot of usability of IGC at different levels of the number of functions.



Appendix 2. Q-Q Plot of usability of IGC at different levels of the human-like characteristics.



Appendix 3. Q-Q Plot of usability of IGC at different levels of the staff size.



(a) number of outpatients: <50

(b) number of outpatients: 50~150

(c) number of outpatients: >150

Appendix 4. Q-Q Plot of usability of IGC at different levels of the number of outpatients.