Human-AI Interaction in the Wild: A Case Study of How People Use a Health Chatbot

Xiangmin Fan
Chinese Academy of Sciences
Beijing, China
xiangmin@iscas.ac.cn

Dakuo Wang
IBM Research
Cambridge, MA, United States
dakuo.wang@ibm.com

Daren Chao
University of Toronto
Toronto, Canada
drchao@cs.toronto.edu

Feng Tian
Chinese Academy of Sciences
Beijing, China
tf@iel.iscas.ac.cn

Zhan Zhang
Pace University
New York, NY, United States
zzhang@pace.edu

CHI'20,, April 25–30, 2020, Honolulu, HI, USA
ACM 978-1-4503-6819-3/20/04.
https://doi.org/10.1145/3334480.XXXXXXX

Abstract
Chatbots are increasingly adopted in many use cases, yet there is little research documenting how people use and interact with them in real world scenarios. In this study, we focus on a self-diagnosis chatbot and analyze its usage data collected over a half-year period. Through both quantitative and qualitative analysis, we synthesized the users' behaviors while they were interacting with this healthcare chatbot. We observed two interesting user behaviors-"dropping" and "gaming"-and analyzed the types of factors led to these behaviors. Using these findings, we discuss the design implications for improving users' satisfaction and adoption of health chatbots.

Author Keywords
Self-diagnosis, Chatbot, Human-AI collaboration

Introduction
Patients and caregivers always face challenges obtaining timely medical advice and information from their healthcare providers due to the long wait time for an appointment [3]. With the advances of artificial intelligence (AI) technology in recent years, there is an opportunity to tackle the challenges and barriers faced by patients in seeking health information [4, 5]. AI-empowered intelligent systems, such as health chatbots, allow patients to share their symptoms and interactively ask questions, and provide live feedback

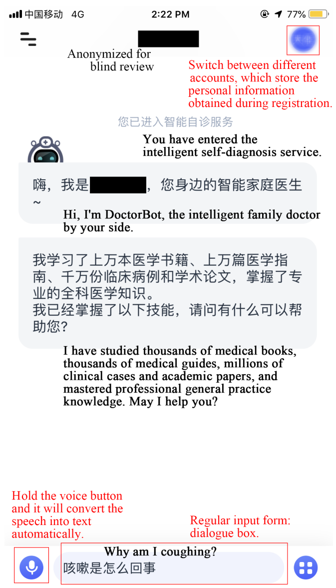

Figure 1: The starting point of the consultation. The user types in Chinese: "Why am I coughing?"

Figure 2: DoctorBot leads the conversation and prompts the user to provide more information.

to help patients triage and manage their conditions themselves [6, 2]. While this technology is still in its developmental phase, health chatbots could potentially alter the landscape of healthcare by increasing access to healthcare services and reducing unnecessary clinical visits [11, 13].

As these health chatbots are capable of enhancing patient experiences with healthcare, and potentially influencing health behaviors, research is needed to understand how to design such chatbots to enable patient-centered care and increase user adoption. Prior work has primarily focused on developing novel algorithms to improve the accuracy and effectiveness of chatbots' diagnoses (e.g., [14, 8, 10]). However, there is little HCI research on the real use of health chatbots [12]. More specifically, little is known about how people interact with health chatbots in the real world, what kinds of user behaviors occur during chatbot use, and what types of barriers exist in using this novel technology.

To bridge this knowledge gap, we examined one of the widely used chatbots in China, DoctorBot[1], which has tens of thousands of users. The large-scale deployment of DoctorBot provides us with an unique opportunity to gain an in-depth, empirical understanding of how people interact with and use these chatbots in the real world and what barriers hinder the delivery of these novel services. These insights can be used to guide the future design of health chatbots to better suit users' needs.

In this study, we took a data-driven case study approach to analyze user behaviors on DoctorBot. By analyzing a total of 47,684 human-chatbot consultation cases, we found that users would consult a wider range of topics, including stigmatic diseases and others that would be out of the scope of conventional health services. Our results also revealed two interesting user scenarios: "Dropping"–a user drops the consultation process at a certain point, and "Gaming"–a user only explores the chatbot with no intention for medical consultation. We also analyzed the feedback provided by users to identify usability issues that would impact the user's satisfaction and engagement. A more comprehensive analysis on the interactions between users and AI conversational agent is underway. We hope these analyses not only instruct future design of healthcare chatbots to improve its usability and adoption rate, but also shed lights on the future directions of studying and designing to interact with human (e.g., patients and doctors) in the healthcare settings and beyond.

DoctorBot

DoctorBot is an AI-driven medical consultation platform, which utilizes large datasets, including numerous medical literature and clinical cases, to produce specific outputs according to users' enquiries. The interaction between DoctorBot and users is text-based. Users can use this application to consult medical diagnoses, disease information, drug-use instructions, and many other topics. Among those, self-diagnosis is the most popular and demanding service, and as such, becomes our primary focus of this study.

A complete self-diagnosis process can be revealed from Figure 1, 2 and 3. After receiving "why am I coughing?" in the dialogue box, DoctorBot will threw out a series of questions for more detailed information about the symptom or health concern that the user expressed. After filling in the questions, it finally give a diagnostic report including tentative medical diagnosis, prediction confidence level, treatment options and other relevant recommendations. The DoctorBot claims that the medical information, diagnoses and prescriptions can only be used for reference only.

[1] https://www.zuoshouyisheng.com/

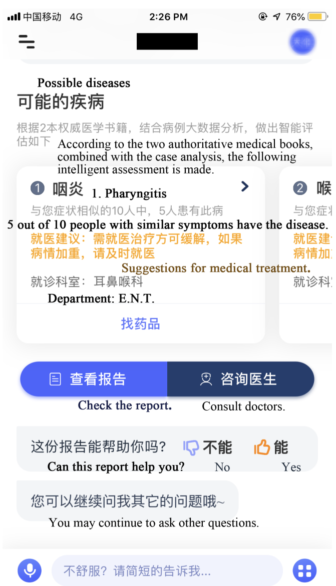

Figure 3: A report including possible disease diagnoses, medical advice and categories of diseases is generated by DoctorBot.

Methodology

To understand how people interact with self-diagnosis chatbot in the real world, we took a data-driven case study approach to analyze the system log generated between September 2018 and March 2019. The system log consists of 47,684 consultation cases initiated by 16,519 users. All identifiable information (e.g., phone number) was removed from the dataset to protect user privacy. Moreover, the users have consented at the point of registration that researchers are allowed to analyze their usage data for non-commercial purposes. The study procedure was approved by our University Institutional Review Board.

We first used statistical analysis to analyze the entire dataset, focusing on users' demographic information and general features of the consultations. By dosing so, we could gain an overview of the characteristics of the users and general patterns of application use. Given the fact that 30,710 out of 47,684 consultation cases were completed (e.g., a diagnostic report was given by the chatbot towards the end of consultation), whereas the rest cases were terminated halfway, we decided to perform in-depth analysis on both complete and incomplete cases using a variety of methods, including statistical analysis and content analysis. In particular, we focused on understanding the types of user behaviors occurred during chatbot use and the barriers that may hinder the effective use of the DoctorBot. We also analyzed the feedback provided by users to identify usability issues that would impact user experience.

Preliminary Findings

Finding 1: DoctorBot provides convenience for people and helps them overcome embarrassment for asking private and sensitive questions.
In total, there are 30,710 completed consultations which usually return a diagnostic report to the user. We analyzed the generated diagnostic reports to understand what types of disease or health concern people usually consult. We extracted and categorized the consulting topics using disease categories that are adopted by DoctorBot[2] and compared them with the typical categorization and usage of healthcare services in three primary hospitals in China [7] (Figure 4). The results showed that those diseases with mild symptoms such as gastroenterology and dermatology appeared in a lot of consultations with chatbot, the proportion of which is significantly higher than in the primary hospitals. One possible explanation is that people with mild symptoms would prefer consulting the necessity of clinical visits first, i.e., using self-diagnosis applications, rather than going to hospitals directly. In contrast, using the chatbot to consult emergency care issues is not very common. We also noticed the use of DoctorBot to seek medical help on stigmatic conditions, such as sexual dysfunction and vulvodynia. This observation highlights that people tended to use health chatbots to deal with medical conditions that often entail considerable privacy and social stigma issues.

Finding 2: Some users "drop" conversations halfway.
Given the large number of incomplete consultations (16,974, 35.6% of total cases), we performed an in-depth analysis on those conversations. Our analysis revealed that users would withdraw at a certain point without completing the consulting process. This may indicate that users encountered barriers in using DoctorBot. We name this phenomena *dropping* behavior.

We analyzed the correlation between consultation rounds and the dropping behavior. The statistical evidences show that users most probably drop the consultation within the

[2] The coding is done by the authors, referenced ICD-11, International Classification of Diseases 11th Revision, which is used by most clinical systems. https://icd.who.int/

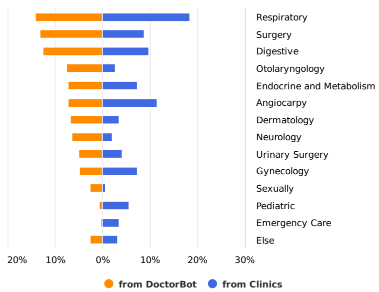

Figure 4: Distribution of Diagnosed Result Departments. The orange bars represent the distribution of all diagnoses by DoctorBot over the typical categorization of healthcare services, while the blue bars represent the clinics' distribution.

Figure 5: DoctorBot: comment box and pre-defined reasons.

first five rounds of conversation. We also examined the factors that could potentially cause people to terminate their consultation session, such as the complexity of questions. More than half of the incomplete conversations occurred on symptom checking questions. The reason is that these symptom-related questions are usually extremely complex and hard to describe, and can easily overwhelm the users.

Finding 3: Some users "game" the platform.
In addition to dropping, we also observed another interesting phenomenon–gaming behavior–that emerged based on our analysis of completed consultation cases. Through analysis of 3,000 consultation cases using a content analysis approach [9], we noticed that a group of users pretended to be a patient and used the chatbot for non-therapeutic purposes. Gaming behavior has been well studied in intelligent tutoring systems [1] but rarely reported in health technology research. Studying and detecting gaming behavior is very valuable because it can 1) inform the design of health chatbots to provide timely interventions; and 2) increase the accuracy of chatbots' diagnoses by removing the noisy information from the training model.

The analysis of 3,000 completed consultation cases allowed us to identify 241 gaming conversations and 5 patterns of gaming behavior: containing nonsense words, inconsistency between main query and subsequent statements, continuously answering "unclear" or "nothing" or using the same words for all kinds of questions, answering each question quickly or quitting the conversations quickly, spending around the same amount of time on each question.

Finding 4: Disease type is highly related to users' experience and their satisfaction level.
Towards the end of a consultation, DoctorBot prompts the user to rate the experience as either positive or negative, as shown in the bottom part of Figure 3. If a negative rating is chosen, the system asks the user to provide further explanations by using a comment box and selecting from a set of pre-defined reasons (e.g., "The result is inaccurate", "Too much content", etc.) as shown in Figure 5.

In total, we collected 3,832 pieces of feedback, with 2,185 positive ratings and 1,646 negative ratings. We first performed statistical analysis to identify the factors that lead to negative ratings, such as conversation duration and disease types. We found that the number of conversation round and the total time spent on the conversation have significant impact on user experience. That is, if a conversation lasts less than 5 rounds, the chatbot is likely to receive more negative ratings. In contrast, too many questions asked by the chatbot could overwhelm the user, causing negative feedback. Interestingly, the disease type is also highly related to users' experience and their satisfaction level. For example, medical advice about commonly seen diseases such as stomachache and respiratory issues usually receive positive ratings. However, those diseases with unknown or complex causal mechanisms such as dermatosis and genecopathy are challenging for the chatbot to provide meaningful diagnosis; therefore, negative rating against DoctorBot-generated outputs is fairly common.

Summary

In this paper, we conducted quantitative and qualitative analyses on system logs of a self-diagnosis chatbot, namely DoctorBot, which is a mobile application capable of providing text-based medical advice to users' enquiries. We have shown results from four aspects of understanding users' needs and behaviors. Based on the results, we plan to conduct mixed method research on different groups of people who have used the software to form a more comprehensive picture of users' attitudes and experiences in the future.

REFERENCES

[1] Ryan Baker, Jason Walonoski, Neil Heffernan, Ido Roll, Albert Corbett, and Kenneth Koedinger. 2008. Why students engage in "gaming the system" behavior in interactive learning environments. *Journal of Interactive Learning Research* 19, 2 (2008), 185–224.

[2] Ahmed Fadhil and Silvia Gabrielli. 2017. Addressing challenges in promoting healthy lifestyles: the al-chatbot approach. In *Proceedings of the 11th EAI international conference on pervasive computing Technologies for Healthcare*. ACM, 261–265.

[3] Xinning Gui, Yu Chen, Yubo Kou, Katie Pine, and Yunan Chen. 2017. Investigating Support Seeking from Peers for Pregnancy in Online Health Communities. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (2017), 50.

[4] Jianxing He, Sally L Baxter, Jie Xu, Jiming Xu, Xingtao Zhou, and Kang Zhang. 2019. The practical implementation of artificial intelligence technologies in medicine. *Nature medicine* 25, 1 (2019), 30.

[5] Gareth Iacobucci. 2019. NHS long term plan: Care to be shifted away from hospitals in “21st century” service model. *BMJ* 364 (2019). DOI: http://dx.doi.org/10.1136/bmj.l85

[6] Mirjana Ivanovic and Marija Semnic. 2018. The Role of Agent Technologies in Personalized Medicine. In *2018 5th International Conference on Systems and Informatics (ICSAI)*. IEEE, 299–304.

[7] Aihua Lin and Qing Liu. 2000. an analysis of disease classification of patients in different levels of hospitals in Guangdong. *Chinese Journal of Hospital Statistics* 7, 3 (2000), 138–140.

[8] Divya Madhu, CJ Neeraj Jain, Elmy Sebastain, Shinoy Shaji, and Anandhu Ajayakumar. 2017. A novel approach for medical assistance using trained chatbot. In *2017 International Conference on Inventive Communication and Computational Technologies (ICICCT)*. IEEE, 243–246.

[9] Philipp Mayring. 2000. Qualitative Content Analysis. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research* 1, 2 (2000). DOI: http://dx.doi.org/10.17169/fqs-1.2.1089

[10] Saurav Kumar Mishra, Dhirendra Bharti, and Nidhi Mishra. 2018. Dr. Vdoc: A Medical Chatbot that Acts as a Virtual Doctor. *Research & Reviews: Journal of Medical Science and Technology* 6, 3 (2018), 16–20.

[11] Joao Luis Zeni Montenegro, Cristiano André da Costa, and Rodrigo da Rosa Righi. 2019. Survey of conversational agents in health. *Expert Systems with Applications* (2019).

[12] Tom Nadarzynski, Oliver Miles, Aimee Cowie, and Damien Ridge. 2019. Acceptability of artificial intelligence (AI)-led chatbot services in healthcare: A mixed-methods study. *Digital health* 5 (2019), 2055207619871808.

[13] Amiya Kumar Tripathy, Rebeck Carvalho, Keshav Pawaskar, Suraj Yadav, and Vijay Yadav. 2015. Mobile based healthcare management using artificial intelligence. In *2015 International Conference on Technologies for Sustainable Development (ICTSD)*. IEEE, 1–6.

[14] Joseph Weizenbaum. 1966. ELIZA—a computer program for the study of natural language communication between man and machine. *Commun. ACM* 9, 1 (1966), 36–45.