http://pubsonline.informs.org/journal/isre

INFORMATION SYSTEMS RESEARCH
Articles in Advance, pp. 1–23
ISSN 1047-7047 (print), ISSN 1526-5536 (online)

A Bitter Pill to Swallow? The Consequences of Patient Evaluation in Online Health Question-and-Answer Platforms

Chen Chen,[a] Dylan Walker[b,*]

[a] School of Management and Economics, The Chinese University of Hong Kong, Shenzhen 518172, China; [b] George L. Argyros School of Business and Economics, Chapman University, Orange, California 92866
*Corresponding author

Contact: chenchen2020@cuhk.edu.cn, https://orcid.org/0000-0001-6068-2806 (CC); dylan@chapman.edu, https://orcid.org/0000-0002-1354-4969 (DW)

Received: March 27, 2019
Revised: October 1, 2020; April 1, 2022
Accepted: July 9, 2022
Published Online in Articles in Advance: ■■■■ ■■, 2022

https://doi.org/10.1287/isre.2022.1158

Abstract. Online health question-and-answer (Q&A) platforms (OHQPs), where patients post health-related questions, evaluate advice from multiple doctors, and direct a bounty (monetary reward) to their most preferred answer, have become a prominent channel for patients to receive medical advice in China. To explore the quality of medical advice on these platforms, we analyzed data on patients' evaluation of ~497,000 answers to ~114,000 questions on one of the most popular OHQPs, 120ask.com, over a three-month period. We assembled a panel of independent physicians and instructed them to evaluate the quality of ~13,000 answers. We found that the quality of medical advice offered on the platform was high on average, and that low-quality answers were rare (6%). However, our results also indicate that patients lacked the ability to discriminate advice quality. They were as likely to choose the best answer as the worst. The medical accuracy of patient evaluation was worse in critical categories (cancer, internal medicine) and for vulnerable subpopulations (pediatrics). Given that millions of patients seek medical advice from OHQPs in China annually, the social and economic implications of this finding are troubling. To understand how patients evaluate advice, we trained deep neural networks to think like patients, allowing us to identify patients' positive and negative responses to different heurist cues. Although our results indicate that OHQPs perform well, we identified several concerns that should be addressed through platform design and policy changes. Because the Q&A process lacks peer review mechanisms, signals of advice quality are not conveyed to patients, forcing them to rely on heuristic cues, which cannot effectively guide them toward the best advice. We also found that the platform reputation metric was not correlated with the quality of the advice giver's advice, may effectively encourage patients to select lesser quality medical advice, and increased the risk of moral hazard for malicious players to intentionally provide less accurate but more agreeable advice for personal gain. Our analysis revealed bad actors on the platform, including drug promoters and spammers. Finally, we found that OHQPs exacerbated *care avoidance*. We discuss several potential policy changes to address these shortcomings.

History: Param Singh, Senior Editor; Idris Adjerid, Associate Editor.
Supplemental Material: The online appendix is available at https://doi.org/10.1287/isre.2022.1158.

Keywords: online healthcare • patient evaluation • care avoidance • deep learning • online health consulting • peer evaluation

Introduction

The digital platform revolution has fundamentally changed how people seek and obtain information online. In the domain of health, online health question-and-answer (Q&A) platforms (OHQPs) are a new type of popular platform that has emerged to connect patients seeking medical advice to physicians in an online setting. These platforms are impossible to ignore for two reasons. First, their societal and economic impacts are without question. Since 2006, OHQPs have grown into a multibillion-dollar business, attracting over 6% of all registered doctors in China. Their success can be attributed to a variety of factors, including supplementing and/or complementing traditional points of care; providing cheap, convenient, and rapid access to physician advice; and redistributing the means of access between patients and physicians. As such, OHQPs have the potential to alleviate inefficiencies and constraints in modern healthcare systems. Yet, little research has been done to evaluate the real-world performance of these platforms. It is unclear whether OHQPs succeed in dispensing quality medical advice to patients and how they complement traditional points of care. For example, do OHQPs promote offline follow-up when it is needed or instead enable or exacerbate *care avoidance*? These are important questions, as the social and economic consequences of supplying bad medical advice or promoting care avoidance could be catastrophic. Because these platforms provide an inexpensive alternative to offline medical advice seeking, it is likely that they have a disproportionate impact on economically disadvantaged populations.

Existing research cannot explain whether or to what extent advice seekers can discern the quality of the advice they receive nor how they select advice from the available alternatives on OHQPs. Although there is substantial research on patients' reactions to advice in offline settings, electronic health advice and on information seeking on related platforms, such as social Q&A platforms (SQPs), OHQPs are distinct in the type of information being sought, the scope of information asymmetry, the role that cognitive biases play, and the platform constraints and features that determine how information is conveyed and evaluated. Presently, we lack empirical estimates of the quality of advice accepted on OHQPs and a robust understanding of patient information evaluation and choice making in the presence of large information asymmetry. Put simply, we still know very little about OHQPs, and given their prevalence, potential, and risks, we cannot afford to ignore them.

We conducted a large-scale study of one of the most popular OHQPs. Our research questions (RQs) are as follows:

RQ1. What is the quality of advice dispensed through OHQPs?

RQ2. (a) How do patients select advice from answers provided by physicians on OHQPs? (b) And how often do patients select the best answers on OHQPs?

RQ3. What are the implications of patient/physician dynamics on OHQPs?

To answer these questions, we collected and analyzed hundreds of thousands of answers to questions posted on 120ask.com. Using a panel of experienced physicians, we obtained empirical estimates of the quality of advice that is offered and accepted across a wide range of medical topics. We leverage natural language processing techniques to capture a rich set of features of platform cues and advice conveyed to patients, including physicians' profile information and platform reputation, aspects of prognosis, suggestions for care, and the psychometric aspects of communication. To understand how patients evaluate and ultimately accept or reject medical advice, we employed deep learning neural network (NN) models trained to simulate patients' actual decision making. Using a novel bootstrapped feature perturbation protocol, we developed which is inspired by LIME (Ribeiro et al. 2016), we estimate the impact of these features on the outcome probability of patients accepting advice. To our knowledge, this is the first study that provides large-scale empirical evidence of the quality of advice and how patients respond to advice on OHQPs. Our results reveal that the average quality of advice dispensed on OHQPs is high, and poor-quality answers were relatively rare (6%). Yet, there remain some troubling aspects of how OHQPs function in the real world that have immediate implications for platform providers, physicians, and public health. In the next section, we describe the social and economic impact of OHQPs, describe how they operate, and introduce five important streams of research literature and relate them to our three research questions.

Context, Theory, and Related Work

The Social and Economic Impact of OHQPs

OHQPs have grown into a multibillion-dollar industry since 2006, filling a demand for access to physician consultation, and complementing brick-and-mortar point-of-care visits. Physicians are a scarce resource in China–only 3 million doctors are available to attend to 0.2 billion patients each year (Sohu 2017). Physicians in China are typically overworked, with more than 50% of them working in excess of 60 hours per week throughout the year (Changyexinxi 2017). Moreover, the allocation of care resources to patients is vastly skewed. Level 3A hospitals (the highest-ranked hospitals according to the Ministry of Health of the People's Republic of China) comprise only 7% of all hospitals in China, but undertake 49% of all clinical care and 43.1% of all hospitalization.[1] OHQPs have the potential to alleviate these supply-side disparities by matching patients with physicians from smaller or lesser-ranked hospitals, overcoming constraints of proximity or geography. OHQPs also have intrinsic advantages in facilitating convenient, economical, and timely interaction between patients and doctors. For example, one of the top OHQPs, 120ask.com (the platform we studied), has attracted more than 100,000 providers of medical advice who have collectively answered approximately 360 million questions (Baidu Baike 2019b). Recent events emphasize the important role that OHQPs (and online healthcare in general) play in meeting patients' needs and complementing or even supplementing offline healthcare. For example, during the COVID-19 pandemic that ravaged China in February and March of 2020, we observed a significant increase in newly registered patients and the number of questions asked on OHQPs, which nearly tripled during the nationwide lockdown. This is likely because OHQPs provided an alternative to in-person care at hospitals, which many patients avoided because of high concentrations of patients infected with COVID-19.

How OHQPs Operate

OHQPs provide online medical consultation by connecting patients with medical questions to physicians and other medical practitioners through a Q&A interface. Most OHQPs offer hierarchical consulting, with baseline medical consultation available through a very low-cost or free (for patients) Q&A interface, and more extensive higher-cost consultation with practitioners on the platform through online messaging (e.g., WeChat) or phone conversation. Although we study the OHQP 120ask.com, there are several other large OHQPs with similar features and design choices (see the online appendix for a detailed table of OHQPs).

In a typical Q&A interaction on a OHQP, a patient posts her question and one or more physicians or medical professionals[2] post their responses. After evaluating the available answers, the patient can select one as the winner, who will receive the bounty (monetary reward), and may elect (at additional cost) to ask more questions or schedule a follow-up consultation with the advice giver who supplied the winning answer. Advice givers may elect to be verified by the platform, in which case their personal and professional information (name, picture, hospital affiliation, qualifications, past activity on the platform) is either disclosed directly on the Q&A page or available through a hyperlink. In OHQPs, askers are patients who typically lack medical expertise, and information asymmetry between askers and advice givers (e.g., physicians) is high. Importantly, doctors are often monetarily incentivized to have their answers chosen, either directly through question bounties (that are rewarded based on patient choice), through higher-cost (follow-up) consultations, or reputation increases, which improve the chance for other patients to initiate higher-cost consultations.

Relevant Literature

Over the past decade, a wealth of research has examined how information systems interact with healthcare in terms of electronic health records or health information (Agarwal et al. 2010; Angst et al. 2010; Mishra et al. 2012; Yaraghi et al. 2015; Angst and Agarwal 2017; Atasoy et al. 2018a, b), online physician reviews (Gao et al. 2015, Lu and Rui 2015, Hao et al. 2017), online communities, social support, social media (Lapointe et al. 2014, Yan and Tan 2014, Guo et al. 2017, Bavafa et al. 2018), and m-health (Ghose et al. 2021). However, little to no research has been done on OHQPs, despite their vast popularity in China and potential to emerge globally as a channel for patient–doctor interaction. Five streams of research are relevant to our research questions on OHQPs. Research on SQPs are a good starting point, as OHQPs may be viewed as a specific subtype of SQP. However, some critical differences in the mechanism design common to OHQPs and the knowledge gap between (and role of) askers and advice givers may impact the quality of advice (RQ1) and how patients respond to it (RQ2). Research on patient response to healthcare advice in offline settings can potentially inform their response in OHQPs (RQ2). Theory and evidence of information processing behaviors when information asymmetry is high provides a good foundation to tackle which aspects of information patients in OHQPs can reasonably leverage for making decisions (RQ2). Furthermore, multiple forms of documented cognitive bias in patients responding to medical advice can guide our expectations on how patients might be driven to make suboptimal decisions (RQ2). Finally, research on postdecision behavior can inform some of the consequences and implications of OHQPs (RQ3). We discuss each of these at length below.

The most similar platforms to OHQPs are SQPs such as Yahoo! Answers and Stack Overflow, which allow askers to ask new questions and search answers to previous questions from archival records. Given the growing popularity of these platforms and their relevance to connected populations, it is unsurprising that they have received substantial attention from academic researchers, particularly in the field of information systems (Kim et al. 2007, Harper et al. 2009, Morris et al. 2010, Zhang 2010, Oh et al. 2012, Song et al. 2019). However, OHQPs differ from SQPs in some critically important ways. They have extremely high information asymmetry between askers and advice givers and a single evaluator for each question, they lack peer rating mechanisms, and they typically have poorly implemented search functionality[3] (Yang et al. 2008; Liu et al. 2014, 2017; Nie et al. 2014). In contrast, SQPs rely heavily on their peer-review and multiple-evaluator rating systems to promote high-quality answers and convenient search functionality to ensure good availability of information (Shah et al. 2009). The healthcare context is also an important distinction. The stakes in OHQPs are personal, most relevant to the asker, and can be significantly higher, given that poor medical advice may have detrimental and even catastrophic consequences (RQ3). There is some research on healthcare on social media and SQPs (Jin et al. 2016, Bae and Yi 2017, Yi 2018). For example, Jin et al. (2016) examine how patients evaluate advice from solvers in the online health SQP Baidu Knows. They found that patient responses depend on emotional support in advice, source credibility of advice, and competition between advice givers. However, the Baidu Knows SQP differs from OHQPs in many ways, as highlighted above. The most important distinctions are that advice givers get no bounty for answering questions, are from the general population and therefore typically lack medical expertise. SQPs also have peer review systems and much stronger search functionality. The role of information asymmetry, source credibility, competition, and other factors researchers of SQPs consider are clearly different when the asker and advice giver share similar expertise or are both patients, compared with when they are patient and physician (as is in OHQPs). Overall, findings from research on SQPs, even in the health context, are unlikely to carry over to OHQPs.

The second stream of relevant research focuses on how patients respond to *professional* healthcare advice and is most relevant to patient response in OHQPs (RQ2). Extensive research has been conducted on patient response to healthcare advice in more traditional offline, in-person care settings (Kaplan et al. 1989, Haskard Zolnierek and Dimatteo 2009, Francis et al. 2010), but little research exists on online healthcare consulting environments. Cao et al. (2017) studied how patients select a physician to consult online, but not their response to the advice they receive through consultation. Wu (2018) studied online healthcare communities (which often coexist in parallel with Q&A sections of OHQPs) and found that perceived usefulness and patient satisfaction explain most of the variance in patient's continued use of these communities. Some work has shown that patients behave irrationally in the context of in-person medical care (Case et al. 2005, Rodoletz et al. 2005, Bass et al. 2006, Reach 2015), but the online setting of OHQPs differs in several important ways. First, in terms of commitment, patients spend much more money in offline settings (hundreds or thousands of Chinese yuan renminbi) as opposed to OHQPs (which may be free or require only small micropayments of less than CNY 1) and time or effort (physical presence versus posting online). The increase in commitment in offline consulting likely indicates that patients suspect their concern is serious enough to warrant an office visit. This suggests that in OHQPs, patients may be less predisposed to believe a concern is serious and might prefer advice that downplays severity (RQ2). Second, in offline consulting, there is no explicit competition between physicians offering advice, as patients typically have a one-to-one relationship with the consulting physician and must expend additional resources to attain a second opinion. The third difference is in terms of accountability and is most relevant to the quality of advice dispensed on OHQPs (RQ1). In offline consulting, the dispensation of poor medical advice is more visible and easily traceable (through official records that are historically maintained and subject to review), whereas on OHQPs, physicians bear little consequence for giving poor advice, as the platform does not review the quality of advice nor make prior answers easily searchable. Although OHQPs do provide reputation metrics for physicians, based on activity on the platform and the ratio of accepted answers, it is not clear whether such signals promote doctors who give more accurate advice (RQ3). Finally, the face-to-face nature of offline consulting allows physicians to gauge and respond to patient's attitudes when they confer advice and places the physician in a more authoritative position. In contrast, physicians in OHQPs cannot gauge or respond to patients' reactions, and OHQPs make patients the explicit authority by allowing them to designate the correct answer and assign the benefit (bounty/reputation increase) to the physician who provided their preferred answer. These distinctions make OHQPs a unique context where both patient and doctor behaviors may significantly deviate from those in offline settings, which has direct implications for our research questions (RQ2 and RQ3). Yet, the prevalence and influence of OHQPs makes it absolutely crucial that we understand these behaviors and their impacts on advice quality and patient evaluation, which have serious implications for public health.

The third stream of research relates to how individuals process information and is directly relevant to how patients choose advice on OHQPs (RQ2). The most relevant theoretical framework for information processing to OHQPs is dual-process theory. Variants of dual-process theories have evolved in different disciplines, primarily psychology, economics, and marketing (Chaiken 1987, Chaiken and Trope 1999, Wei and Watts 2008, Glöckner and Witteman 2010). Despite their differences, these theories consistently stipulate that humans engage in two different types of processing when they encounter new information: people may logically analyze new information content or instead may attend to heuristic cues associated with that content. However, when a person lacks the necessary expertise to analyze information logically, he must rely entirely on heuristic cues to decide whether to adopt the information (Trumbo 1999), which has been specifically shown to hold in online environments (Meservy et al. 2014). In OHQPs, the platform determines which heuristic cues are visible for users to leverage in their decision-making process. When the expertise gap (i.e., information asymmetry) is high, as in healthcare settings, patients must often rely solely upon heuristic processing (Jin et al. 2016). In such cases, the platform constraints become critically important to the overall performance of the system. It is therefore important to understand whether, how, and to what extent patients leverage heuristic cues in evaluating healthcare advice and whether and to what extent their evaluations are biased (RQ2).

The fourth stream of research relates to bias in information processing, which is also relevant to how patients respond to and choose advice on OHQPs (RQ2). A wealth of research suggests that people are subject to bias when processing information in a variety of forms, and underlying mechanisms include information selection and avoidance due to confirmation bias (Trope and Bassok 1982, Nickerson and Bias 1998, Pohl 2004); cognitive dissonance (Hyman and Sheatsley 1947, Festinger 1962, Trope 1979); the desire to avoid information that evokes anxiety, discomfort, and other negative feelings (Case et al. 2005, Rodoletz et al. 2005, Sweeny et al. 2010); and the need for validation or defensive reasoning (Kunda 1990, Jain and Maheswaran 2002, Hart et al. 2009). Substantial research has shown that patients are prone to avoid, willingly ignore, or reject bad news, even to their own detriment (Case et al. 2005, Rodoletz et al. 2005). For example, potential HIV carriers often intentionally avoid seeking test results or even blatantly reject them, out of denial (Sweeny et al. 2010). Taken together, prior research suggests that patients in OHQPs will be particularly prone to bias, but lacking expertise, must rely on heuristic cues provided by the platform and embedded in physicians' communications (RQ2).

The last relevant stream of research relates to postchoice decision making, which is a rich topic of study in psychology and economics. This stream of research directly relates to our question on the consequences of OHQPs (RQ3). A well-known classical work in psychology studied how choice might impact postdecision evaluation and found an increase in preference for chosen goods and a decrease in preference for unchosen goods (Gerard and White 1983). A more recent study has confirmed that this phenomenon is present not only in subjects' self-report measures but directly in measures of brain activation (Izuma et al. 2010). In addition, the differentiation and consolidation theory of decision making predicts consolidation processes that work in favor of the chosen alternative (Svenson 1992). Taken together, this research suggests that a patient's act of choosing an answer to their question from the set of alternative answers will increase their tendency to act on the chosen advice and influence their downstream behaviors (RQ3).

In the remainder of this paper, we describe our data collection and processing methods, analyze the quality of advice on a large OHQP, and estimate the impact of heuristic cue features of medical advice on patients' tendency to select advice.

Methodology

Data Collection and Preprocessing

To answer our research questions, we scraped data from a Chinese OHQP, 120ask.com, spanning a three-month period in 2015. The platform 120ask is an online health Q&A platform where hundreds of thousands of registered doctors, medical practitioners (e.g., registered nurses), and other advice givers provide medical advice for an exceptionally low cost. The platform has ~320,000 visitors daily and has been searched on Baidu about 100 million times since its inception. Patient fees and advice-giver payment for Q&A consultation differ across different OHQPs and over time, but on 120ask in 2015, nearly all questions were free for patients. Advice givers were provided with a small monetary incentive (CNY 0.10) to answer each question. If their answer was accepted by the patient, they received a bounty of CNY 0.15. Both of these payments to advice givers were subsidized by the platform. It was possible for patients to add an extra bounty (CNY 1–CNY 100) out of their own pocket to increase attention to their question, though this rarely ever happened (<0.5% of questions). Patients on 120ask can also opt for further consulting through WeChat applications or over the phone. Such consulting is typically prorated by the minute or conversation at a price set by the doctor and listed on his or her profile page. We chose 120ask because (1) it is one of the largest OHQP sites, with more than ~100,000 officially registered doctors participating and thousands of questions answered on a daily basis; (2) data on all questions and answers and doctors' information can be scraped via conventional methods; (3) by design, the full text of questions and answers on the site are highly formatted (i.e., almost all answers contain two distinct parts: diagnosis and suggestions), which facilitates natural language processing; and (4) for each question, a patient is allowed to select only one correct answer, which rules out the possibility of multiple selections, which can significantly complicate analysis. Because we are focused on patient evaluation among a set of multiple alternatives, we excluded from our analysis questions that were not evaluated or were answered by fewer than two doctors. The whole data set consists of 114,037 questions and 496,842 answers. To ensure time invariance of our findings, we also scraped data from the same site spanning a three-month period in 2019 and performed the same analysis on it. We observed largely the same results (see the online appendix).

Evaluating the Accuracy of Online Advice

To objectively evaluate the accuracy of doctors' diagnosis and suggestions, we conducted a survey among eight experienced physicians (referred to as *evaluators*) in China who had not participated in any OHQP at the time of the survey. We selected evaluators with strong qualifications and experience. Our expert evaluators were trained in China in a reputable medical school, received broad and systematic training across all fields prior to specialization, and had approximately 8–30 years of experience practicing medicine in China. They were recruited through an informal social network forum where physicians exchange medical advice and network professionally. Each was paid CNY 2,000 (approximately USD 300) for their effort, which is approximately one-half to one-third of their monthly salary. They spent, on average, 20 to 30 hours over the course of three weeks to complete their evaluations.

We randomly sampled 3,000 questions (from a pool of ~114,000) that had multiple answers. This sample generated 12,767 different question/answer pairs (survey units) with two items: "How much do you agree with this doctor's diagnosis?" (from 1 = totally disagree to 5 = totally agree), and "How confident are you in this judgement?" (from 1 = totally unconfident to 5 = very confident). To remove the potential for bias associated with doctors' identity or status, we provided questions and answers only, with no supplemental information on the doctors who provided the answers. Each survey unit (question/answer pair) was evaluated by three independent evaluators.

Evaluators were instructed to evaluate the entirety of advice within an answer (diagnosis, prognosis, and suggestions for care) on a five-point Likert scale. They were instructed to rate confidence at one if they felt that the question fell outside of their expertise and were explicitly encouraged to say "I don't know" when not sure, to improve the credibility of the evaluation. Evaluators were also encouraged to use any outside resources they felt were necessary to aid in assessing the advice given. In a follow-up survey that we conducted, three of the four physicians who responded indicated that they consulted outside resources. To ensure that the physicians in our panel of respondents were reading the questions and answers carefully, we incorporated an attention check randomly into our survey. The attention check instructed them to report specific answers, which, if not followed, would invalidate the respondent's evaluations entirely. None of our expert evaluators failed the attention check.

We scored each answer by aggregating evaluator ratings according to the following formula:

$$S_i = \sum_{j \in (c_{i,j} > 2)} \frac{(c_{i,j})^2}{\sum_{j \in (c_{i,j} > 2)} (c_{i,j})^2} r_{i,j},$$

where S_i is the rating score of the answer i, $c_{i,j}$ is the confidence of the evaluation of answer i by evaluator j (restricted to the same question), $r_{i,j}$ is the rating of answer i by evaluator j (restricted to the same question), and the sums are taken over all evaluators j that evaluated answers to the question, provided that their confidence exceeded two. In other words, the score of each answer was constructed as the confidence-squared weighted average of all scores of confidence level 3 or higher (scores from evaluations with confidence level 2 or lower were omitted to mitigate professional uncertainty). The motivation for creating an aggregated score weighted by the square of the confidence derives from treating the confidence as an implicit reciprocal of the standard deviation, which codifies the panel physician's uncertainty in evaluating the question. This interpretation has the benefit of discounting low confidence ratings much more substantially in their contribution to the overall score. Our results are robust to different choices for weighting confidence in aggregating scores (see the online appendix).

Table 1. Summary Statistics on Evaluator Ratings

Statistics	Weighted score	Average confidence level
Mean	4.28	4.14
Standard deviation	0.67	0.55
Min	1	1
25%	4.0	3.67
50%	4.53	4.33
75%	4.76	4.67
Max	5	5
N	12,762	12,762

For each question, the answer with highest overall aggregated score was designated as the correct answer. When two or more answers to the same question were tied by aggregated score, we used the highest confidence level from any of the evaluators as the tiebreaker. In ties where the highest evaluator confidence scores were equal (8% of total answers), we allowed for multiple answers to be designated as correct. Summary statistics on evaluator ratings are provided in Table 1.

Overall, the ratings were consistent across different evaluators. The average deviation of each physician's rating from the weighted score was only 0.57, much less than the mean of weighted score of all answers (~4.3). In addition, evaluators agreed approximately 80% of the time to within one rating score, and more than 80% of time at least two out of three evaluators reached consensus. Although the most common ways of measuring interrater reliability, such as Cohen's kappa, intraclass correlation (ICC), and Krippendorff's alpha, do not incorporate rater confidence, which is unique and important in our setting, we nonetheless calculated interrater reliability in the following ways. For all ratings, ICC(3, k) was 0.47, indicating fair agreement, and for ratings with the highest confidence, it was 0.65, indicating good agreement. In addition, we calculated a customized version of Krippendorff's alpha that incorporated rater confidence (see the online appendix for details), yielding $\alpha = 0.7637$ (95% confidence intervals (CIs): 0.7633, 0.7642), indicating reasonable agreement. We believe that these results likely reflect some medical uncertainty in evaluating advice quality that is natural in this setting (rather than unreliability in raters). That reliability increases with confidence justifies our choice of a scoring mechanism that confers more weight to more confident ratings.

Regarding our research question on the quality of advice dispensed on OHQPs (RQ1), from the above analysis, we conclude that good healthcare advice is represented among the answers to most questions. For each question, at least one of the answers had an average rating above three, with a stated confidence of three or higher by at least one evaluator, implying that patients had at least one reasonable option to choose from.

To understand why some answers were poorly rated (score of <3), we randomly selected 70 such answers (half of which were selected by patients) along with their questions and recruited four evaluators from our original panel to reevaluate the answers and provide more details. We first asked whether they agreed with the original evaluations given by their peers (93% of the time, they did). We then asked them to write a brief paragraph on why they agreed or disagreed. Each answer was evaluated by two evaluators, and the results were manually coded.

Analysis

Patient Evaluation Performance

To assess the accuracy of patient evaluation, we compared answers accepted by patients to those designated as correct by our expert evaluators. The results are striking. On average, patients selected the best answer (as designated by our evaluators) in only 31% of all cases (922 out of 2,968). Furthermore, in approximately 29% of the cases (857 out of 2,968), patients selected the worst answer (as designated by our evaluators). Approximately one-fifth of all answers that were designated as "entirely incorrect" (score of <2) by our evaluators were chosen by patients (113 out of 519), indicating that extremely bad advice, although less common, was likely to be chosen when given. In addition, patients are almost as likely to choose answers within the top 25% by evaluation score (chosen 24.5% of the time) as they are to choose answers in the bottom 25% (chosen 21.7% of the time), answering our second research question (RQ2(b)). Figure 1(a) shows the distribution of answer scores chosen by patients and those that are the best answers.[4] Figure 1(b) shows the distribution of the difference between the score of the best and the patient chosen answer.

It could be that patients seem to perform poorly not because they are poor evaluators, but because for some questions, the quality of answers may be low and may not vary substantially—that is, perhaps many questions have answers that are of similar low quality and the patient must choose among these. To ascertain whether this is true, we plotted the distribution of

Figure 1. Distribution of Evaluation Score

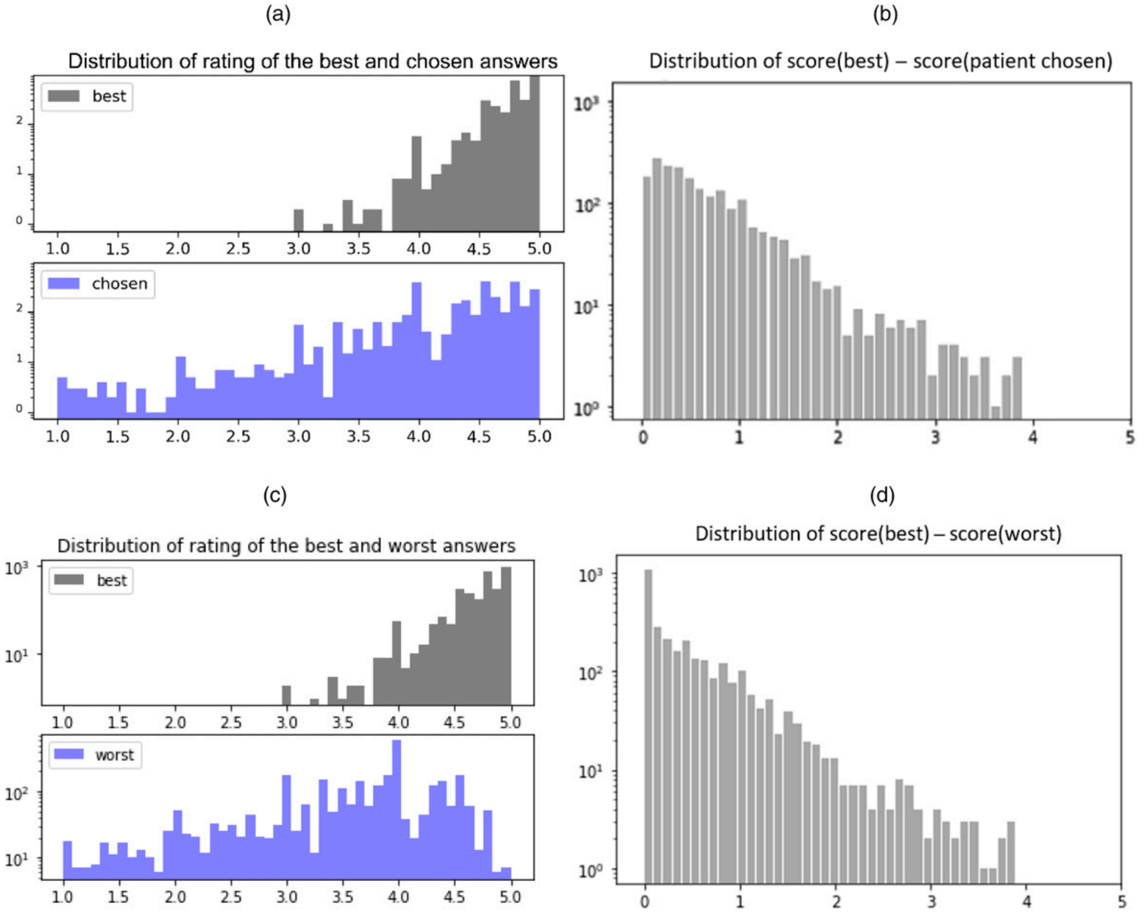


Table 2. Summary Statistics of Patients' Evaluation Accuracy by Disease Category

Category	Description	Num of answers in each category	Mean score according to evaluators	Min score	Standard deviation	Percent patients chose best answer (%)
cat_recreational	Healthy life style	84	4.151	1.333	0.793	65.2
cat_sex	STD	207	4.351	1.300	0.678	42.0
cat_pifu	Dermatology	497	4.255	1.000	0.705	39.8
cat_chuanran	Infectious disease	273	4.349	1.000	0.673	36.4
cat_pingxing	Skin-related condition	587	4.359	1.364	0.559	34.1
cat_waike	Surgical department	2,077	4.289	0.000	0.646	34.0
cat_wuguan	Ear, nose, and throat	841	4.197	0.000	0.740	31.7
cat_other	Other	137	4.364	0.000	0.732	31.4
cat_zhongyi	Chinese medicine	323	4.227	1.000	0.724	30.4
cat_fuchan	Obstetrics and gynecology	2,618	4.358	1.000	0.627	30.2
cat_zhongliu	Tumor	347	4.090	1.455	0.761	29.3
cat_xinli	Psychiatrics	306	4.325	1.800	0.601	28.6
cat_neke	Internal medicine	2,630	4.241	0.000	0.701	27.7
cat_meirong	Cosmetics	206	4.272	1.158	0.614	27.7
cat_zhengxing	Plastic surgery	891	4.288	1.000	0.600	26.9
cat_erke	Pediatrics	728	4.229	1.000	0.726	26.3

scores of top-rated and bottom-rated answers to each question in Figure 1(c). The distribution of the difference between the scores of the best and worst answers are displayed in Figure 1(d). Clearly, the best and worst answers are distributed differently. However, to determine whether patient choose poor answers because they lack good alternatives, we repeated our estimation of how often patients chose incorrect answers by only looking at questions that had a clear winner, as defined by questions where the quality score of the best answer was at least one full point above that of the second best answer. In this case, patients chose the best answer only 40% of the time and the worst answer 44% of the time. For questions with a clear winner, on average, patients were better at choosing the best answer (40% for questions with a clear winner vs. 31% for all questions), but much more likely to choose the worst answer (44% for questions with a clear winner vs. 29% for all questions). Overall, patients chose an answer that was at least one full point lower than the best answer 50.7% of the time. Thus, patients do not select poor advice because they lack good alternatives.

Patient Performance Across Disease Categories

While choosing low-quality medical advice can lead to harmful health consequences, not all categories of advice seeking are equally consequential. It is likely that the consequence of misjudgment varies over different types of conditions or disease categories. For example, in categories related to cancers/tumors, pediatrics (a vulnerable group where even a mild condition could be life-threatening if not properly attended), and internal medicine (mostly internal organ disease affecting the heart and liver), the damage caused by following bad advice could be much more severe. We analyzed patients' performance of evaluation by condition or disease category. The results are displayed in Table 2 (we omit the max score as it is identically five across all categories).

Interestingly, patients assessing answers to questions in higher-vulnerability categories are even more prone to poor judgment, as accuracy from all three categories ranked in the bottom 6 of 16. The average difference between the score of the answer the patient chose and the score of the best answer has similar ranking across disease categories (see Table A2 in the online appendix). Patients did particularly poor in pediatrics (ranked last), where most of the askers are parents of babies or toddlers. It is unclear why this is the case. In general, we speculate that reasons for poor performance in any category could include the following: (1) diseases from that category are more complicated; (2) external sources of information for that category are more varied in quality, including sources of misinformation; (3) patients may be more prone to bias when affected by such conditions, leading to irrational judgment. Indeed, cursory analysis of questions from the worse-performing categories reveals a higher level of anxiety (higher counts of words that match the Chinese Linguistic Inquiry and Word Count (LIWC) dimension of "anxiety").

When Patients Choose Poor Advice

Although we found that patients select suboptimal advice, this may not be particularly problematic if the quality of the advice they select is still high. On the other hand, when patients select advice that is poor (quality score of <3) and considerably worse that the available alternatives, this can lead to serious adverse health outcomes. Even though we found that answers of poor quality were rare (~6%) on the platform overall, because patients lack the ability to discriminate advice based on its quality, they still choose poor-quality advice a substantial percentage of the time. Figure 2 displays the average percentage of the time that patients choose poor-quality advice compared with the quality of the best alternative answer.

Even when the best answer available was of relatively high quality (score of >4), patients chose poor-quality advice 2%–15% of the time. Given the large scale of the platform, the tendency for patients to select poor advice when it is offered can lead to thousands of patients acting on bad advice.

Here we provide a rough estimate of the annual proliferation of poor medical advice through our OHQP, if the platform were to continue unchecked. On average, patients selected answers to approximately ~440,000 questions each year (extrapolating from the ~110,000 questions where patients had more than one choice and selected an answer from our three-month data period). Assuming the proportions of poor medical advice given and selected by patients to be constant, we can expect ~124,000 poor answers (score of <3) from all categories, of which ~23,000 will be chosen by patients. In the three vulnerable categories we identified (pediatrics, cancer/tumors and internal medicine), we can expect ~32,000 low-quality answers given each year, of which ~5,700 answers will be chosen by patients. Even if patients act on only a fraction of selected advice, the consequences for public health are serious. Moreover, as we will show, several mechanisms on the platform incentivize patients to do so, including the platform reputation system, the lack of searchable records of patient–physician interaction, and lack of physician accountability. It is reasonable to wonder why poor-quality answers are given on the platform and why an answer would be rated poorly by an experienced physician. We explore bad answers and bad actors on the platform in the next section.

Exploring the Bad Answers and Bad Actors on the Platform

To understand why our panel of physician evaluators evaluated the quality of an answer as low, we conducted a follow-up survey with some of the physician evaluators from our original panel. We randomly sampled 70 answers that were poorly rated (approximately one-third of answers rated 2 or below) and asked four evaluators from our original panel to indicate whether they agreed with the evaluation and to explain why or why not. In 93% of all cases, our follow-up evaluators agreed with the original assessment. We then manually coded their explanations, yielding the categories shown in Table 3 (in descending order, according to the number of answers assigned to each category).

The second and third categories are somewhat generic for poor-quality answers, involving advice that is not medically correct or is incomplete. However, other categories reveal more interesting facets of bad advice.

Figure 2. Patient Tendency to Select Poor-Quality Advice

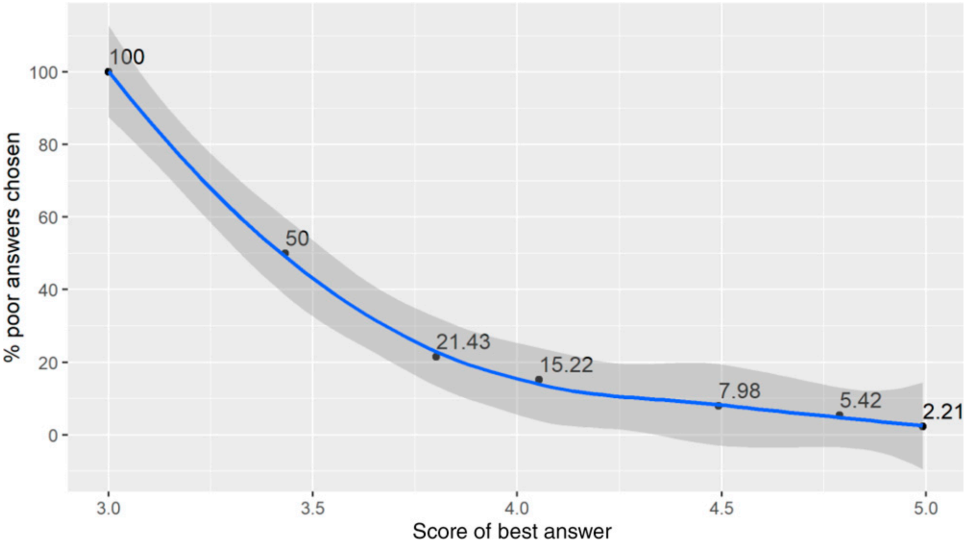


Table 3. Categories of Bad Advice

Category	Description of bad advice
1	The answer does not match the question
2	The advice giver provided a partial but not complete answer
3	The answer is medically incorrect
4	The advice giver appears to be promoting a specific drug with little concern for the patient's issue
5	The answer downplays the severity of the patient's problem
6	The answer is not concise, often containing too much information that is not related to the question

The first and sixth categories seem to pertain to advice givers on the platform that answer a large volume of questions, often with less care. Some advice givers seem to have adopted the strategy of answering questions by copying and pasting the same answer to many questions, possibly by copying the information from other resources. Our hypothesis is that some advice givers have adopted a volume strategy by answering many similar questions with less tailored answers, to obtain bounties.

The fourth category, which relates to potential drug promotion, is interesting because it reveals the unexpected incentives of some advice givers on the platform. We note that the platform does not enforce that all advice givers to be licensed to practice medicine, and although they encourage advice givers to provide their license information and become verified, it is not required. The opportunity for drug promotion in this unregulated channel is troubling.

The fifth category may indicate that advice givers could be swayed to appease patients, which may be driven by (unconscious or conscious) tendencies to comfort the patients by downplaying the severity of their issues.

These findings suggest that although most advice givers provided high-quality advice and seemed to have good intentions, there are clear signals of bad actors on the platform, including spammers and drug promoters. Traditionally, reputation systems are one solution to fend off bad actors on a platform. However, because the reputation metric for advice givers on 120ask is the ratio of accepted answers, it depends entirely on patient choices. As a result, because reputation does not involve any sort of peer review, and because patients do not perform well as evaluators, the current reputation system on 120ask does not identify good actors and distinguish them from bad ones.

We found further evidence of "spammy" advice givers. Operationally, we define a spammer as an advice giver who gave the same answer to at least five different questions. Using this criterion, we identified 383 spammers out of the 16,828 advice givers (2.3%) in our data. These answers sometimes matched answers from other advice givers who were not identified as spammers, suggesting that they may have been copied from other advice givers. For example, one physician gave three variants of nearly identical answers to 1,479 questions in the area of gynecopathy that promoted the synthesized medicine 育宫培麟丸 ("YuGong PeiLin Pills"). Another advice giver (who was not a licensed practitioner) focused on late-stage cancer related questions, and for eight questions on multiple types of cancer (liver, stomach, and lung), gave exactly the same answer that suggested generically using traditional herbal medicine.

Beyond spammers, we also looked for specific evidence of active drug promotion. In the follow-up survey that we conducted to understand bad answers, our evaluators identified three suspicious promoters (whose answers were designated as category 4—the advice giver appears to be promoting a specific drug with little concern for the patient's issue), of whom two were confirmed as dedicated drug promoters by looking at their entire answer history. The two specific drugs were Lukfey and 微络康洗胰清糖素("WeiLuoKang" or "Pancreas Washing Sugar Cleansing Element"). Lukfey is suspicious, as it claims to be a world-renowned western medicine but is in fact a synthetic herbal drug that was invented and manufactured by a domestic Chinese firm. It was advertised on Baidu, but we could find no mention of it outside of China. One physician dedicated approximately 500 answers to promote Lukfey. The other drug, 微络康洗胰清糖素("WeiLuoKang" or "Pancreas Washing Sugar Cleansing Element"), is a healthcare supplement product that is marketed for reducing blood sugar. We consulted with three physicians from our panel who claimed to have never heard of the drug, believed it to be suspicious, and cautioned that it should not be used as a replacement for blood sugar–reducing drugs. One chief physician dedicated approximately 300 answers to promoting this drug as an elixir for diabetes.

We can understand why some advice givers might be financially incentivized to adopt a spam- or volume-based approach to answering questions—they benefit directly from micropayments. We estimate that such advice givers, who may answer thousands of questions each month, can earn up to CNY 350 (see the online appendix for details of the estimation).

Overall, this evidence suggests that bad actors do exist on the platform, some of whom are licensed to practice medicine and verified by the platform. Platform peer review mechanisms could help to limit bad actors. We discuss this idea further in the discussion section.

Modeling Patient Evaluation with Neural Networks

To understand why patients are poor evaluators, and, more specifically, how they use available information to evaluate answers (RQ2(a)), we estimate the impact of cues from content, phrase and psychometric language in answers, and the contextual information provided by the platform (such as doctor credentials, hospital ranks, etc.) on the likelihood of a patient designating an answer as correct. Conventionally, this could be accomplished through discrete choice modeling (implemented through, e.g., conditional logit (clogit) regression). Although this approach provides clean interpretability of coefficient estimates, it also requires several assumptions, including the specification itself and the independence of irrelevant alternatives assumption. Moreover, it is not guaranteed to achieve maximal predictive performance. To overcome these issues, we adopted a deep NN modeling approach, a commonly used technique in machine learning. Neural networks can accept the entire feature set of all answers presented to the patient as an input, and combine these features parametrically though hidden layers, which allows for arbitrarily complex mathematical dependence (contingent on neural network depth, width, and nonlinear activation). These aspects of neural network models allow us to relax assumptions about specification and independence of irrelevant alternatives, and to more closely mirror the decision problem faced by actual patients.

The universal approximation theorem states that a feedforward network with a single hidden layer containing a finite number of neurons can approximate continuous functions on compact subsets of $\mathbb{R}^n$, under mild assumptions on the activation function (Hornik et al. 1989). This theorem proves that neural networks can achieve mathematical equivalency with conditional logit regression (if such a specification is indeed appropriate), provided that the neural network is complex enough and well parameterized. Indeed, a comparison of estimates from a discrete choice clogit model and the neural network model yield consistent results. However, we found that the neural network model outperformed the clogit model significantly in terms of predictive power, achieving accuracy of 58.09% (95% CIs: 58.06%, 58.13%), compared with 54.01% (95% CIs: 53.28%, 54.74%) for the clogit model, which is approximately a 10% relative increase. The results of the clogit regression can be found in Table A6 of the online appendix.

We built and trained two neural networks: a *real patient NN*, to simulate the decision making of actual patients,[5] and an *ideal patient NN*, to simulate the decision making of a hypothetical ideal patient who leverages heuristic cues to make the best possible decision.[6] Both are supervised learning processes trained on the same heuristic features of the data (for fair comparison), with only one major difference. For the real patient model, the predicted outcome is encoded by a one-hot vector, where one corresponds to the answer chosen by patients and zero otherwise. For the ideal patient model, the predicted outcome is encoded by a one-hot vector where one corresponds to the answer designated as the best answer. In other words, the ideal patient model would try to make the most correct choice, after seeing the same input as the real patient model. The ideal patient NN allows us to compare the performance of actual patients to a hypothetical "best-performing patient evaluator" to determine the relative gap in performance. We can leverage the real patient NN to estimate the impact of features on patient evaluation and further compare with the ideal patient NN to understand how a hypothetical ideal patient would weigh heuristic features of answers differently from real patients. To accomplish this, we need an approach to derive explanations from the predictions of real and ideal NN models. We describe such an approach below, but first we turn to a description of the heuristic features of answers and how they were encoded.

Encoding the Features of Answers

Each answer on the platform is highly structured and contains cues provided by both the platform (context) and the answer itself (content). Context cues include credentials (a physician's occupational rank, affiliated hospital rank, listed expertise, and activity on the platform) and effort (the number of words in an answer, whether the doctor asks detailed questions). A doctor's occupational rank (e.g., chief or associate chief physician) measures their social status in the field of medicine and can signal their experience, credentials, achievement, and reputation. The rank of a hospital where a physician works is also a signal. In China, hospitals are officially classified into three levels by the Ministry of Health according to metrics such as size, endowment, number of visiting patients, yearly performance, and number of highly reputed experts (Baidu Baike 2019a). The highest rank is level 3A. The level of a hospital likely serves as an indirect reflection of a doctor's social status. Being affiliated with a more prestigious hospital positively signals a doctors' status. OHQPs also provide reputation information, including the number of questions physicians have answered and the ratio of accepted answers. Other context cues include the number of informative words in an answer and whether physicians habitually include postface text (such as a disclaimer, impersonal coda, or outro or closing statement) that is common across all their answers.

Content cues include diagnosis and prognosis (the section of an answer that classifies a patient's ailment and likelihood of future outcomes given that classification), suggestions (an indication of what a patient should do, typically in plain language), and other communication language (which may include words of comfort or encouragement). Diagnoses and prognoses typically contain medical terminology or jargon that can be difficult for a patient to understand, and that is therefore less likely to significantly impact the patients' evaluation (in accordance with dual-process theory). However, suggestions and other communications contain a great deal of heuristic cues—plain, nonprofessional phrases that are comprehensible to laymen and intended to communicate, explain, or instruct—that a patient can readily use to evaluate an answer. To encode these cues into numeric features, we turn to natural language processing methods applied to Chinese text.

Word choice can provide rich information about beliefs, fears, thinking patterns, social relationships, and personalities. It is therefore interesting to explore the psychometric dimensions of doctors' answers and examine whether and to what extent they affect patients' evaluations. To capture psychometric features of answers, we leverage the Chinese LIWC dictionary, a scientific lexicon that captures the tones of Chinese phrases or words by attributing to each one or more psychometric dimensions (Huang et al. 2012). For example, the dimension "anxiety" includes words such as "焦虑" ("anxious"), "不知所措" ("unsettled"), "漫无目的" ("aimless"), and "危机" ("danger"). The Chinese LIWC dictionary is very inclusive but can be prone to mismatched characterizations, because many Chinese words and characters are polysemantic. To avoid erroneous characterizations, we excluded dimensions that were mismatched more than 50% of the time from this study. This procedure resulted in 18 LIWC dimensions, including *discrepancy*, *affiliation*, *male*, *female*, *sexual*, *differentiation*, *certain*, *feel*, *cause*, *achieve*, *death*, *family*, *eating*, *anxiety*, *risk*, *reward*, *tentative*, and *friend*.

In addition to issues of mismatched characterization, we also found that LIWC alone failed to capture a variety of common phrases used by doctors to express their attitudes toward patients' situations (e.g., expressions of empathy, attempts to soothe, declaration of warnings, downplaying the severity of a condition). To account for this, we built a heuristic dictionary of common phrases to compliment Chinese LIWC features. To ensure that our heuristic dictionary was meaningful and appropriate, we consulted several experienced Chinese physicians. The initial heuristic dictionary was constructed by manually inspecting the content of 3,000 answers. This dictionary was then sent to a panel of experienced Chinese physicians who were instructed to examine whether words or phrases were correctly assigned to each dimension, whether such dimensions were sufficiently inclusive, and to provide corrections when appropriate. Finally, the content of an additional 1,000 answers was manually examined using the dictionary to ensure that each dimension did not exclude or miss pertinent words or phrases. This resulted in the addition of three prominent dimensions: *optimism*, *comforting*, and *frankness*. Importantly, we allowed for soft matching of phrases with interstitial words (in the form of wild cards). For example, "没有*问题" (where the asterisk indicates a wildcard) matched both "没有问题" ("not an issue") and "没有大问题" ("not a serious issue").

Beyond the psychometric dimensions associated with word choice, patient evaluation likely depended upon the types of suggestions they received from doctors. To capture this, we constructed a dictionary of phrases to match common types of suggestions given by doctors, using a procedure similar to that described above in consultation with a panel of Chinese doctors. This resulted in five types of general suggestions: suggestions involving diet, in-person checkup, in-person treatment, getting rest, and exercise. We excluded suggestions to take medicine, as almost all doctors' answers contained such suggestions, and there was significant heterogeneity in medicine type. Soft matching of phrases with interstitial words was performed for suggestions in the same manner as described above.

To map doctors' answers onto psychometric and suggestion features, we used regular expression matching. The powerful regular expression system allows for accommodation of flexible terms, negation, conjunction/disjunction of different phrases with the matching terms, and positive/negative look-back. For example, it captures "你最好赶紧去医院" ("you need to go the hospital immediately"), "建议立即手术" ("I suggest immediate surgery"), and "去正规医生检查" ("should visit a doctor to get examined"), while negating "别去医院" ("do not go to the hospital"). Phrases with the same or similar meanings (such as "去医院" and "到医院," which both suggest going to the hospital) were all captured by introducing disjunction in matching terms. To ensure the equivalency between regular expression terms and phrases in our constructed dictionaries, we performed extensive tests using the manual inspection of content from a randomly selected sample of 1,000 answers to ensure that regular expressions matched all desired phrases and terms inclusively. Complete lists of all answer features categorized by context cues, phrases and suggestions, and LIWC dimensions are provided in Tables A3–A5 of the online appendix.

Neural Network Structure and Training Procedure

For both real and ideal patient neural network models, we prepared the input data corresponding to each question and set of answers in the following way. We extracted all possible heuristic features (a total of 34) that represented the content or context information of each answer. Next, we converted each feature to a vector of binary indicator variables (through binarization or discretization) and concatenated all feature vectors for each answer, yielding a 7 × 147 input vector for each question.[7] For features that are naturally binary (e.g., an answer either includes the suggestion to go to the hospital or it does not), we coded inclusion as one (zero otherwise). For features that represented count variables or log-transformed count variables, we discretized values and introduced dummies for low, medium, and high values, corresponding to values that fell into the bottom 25%, middle 50%, and top 25% quantiles, respectively. Binarizing features yields a standardized representation across different variable types and allows us to meaningfully compare the content that patients were exposed to when they evaluated answers to their questions. By including all answers to a given question as an input, the neural network is able to capture the actual evaluation task that patients faced.

The structure of the neural network is displayed in Figure 3. It is composed of a 7 × 147 input layer (where 147 is the length of feature vector of each answer correspond to a branch), two intermediate layers (of sizes 34 and 11) for each branch, and a scalar output layer for each branch, corresponding to whether the answer was chosen. The resulting output is a vector of seven scalars. For training purposes, a softmax activation layer was added after the original output layer, designated as the true output (y), to permit training with a cross-entropy loss function. The true output represents the probability for each answer of being chosen. As switching the order of inputs (e.g., switching the order of the first and second answer) should in theory not impact a patient's choice, the neural network was constrained to be symmetric with regard to the seven branches, such that all branches share the same weight matrix.

To train and evaluate the predictive performance of the real and ideal patient neural network models, we followed the standard approach and divided the data into training (80%) and test (20%) sets.[8] For the loss function, we used the cross-entropy between training output and actual output (y). Weights were updated during training through RMSPROP,[9] an improved version of stochastic gradient descend. Each model was trained for up to 20 epochs with early stopping to prevent overfitting.

We assessed the robustness of the neural network structure by repeating our analysis with an additional intermediate layer and by changing the number of nodes in each layer (see the online appendix for details). For each alteration, we resampled training/test sets and reran the training procedure. Overall, we observe little to no variation in training accuracy (<1.5%), affirming the robustness of our method. The real patient NN was trained and tested on an 80% training split of the data set of ~110,000 questions evaluated by real patients, minimizing the cross-entropy between the neural network's prediction of the patient's selected answer and the actual answer selected by the patient. Overall, the real patient NN mimicked the patients' choice with an accuracy of 58.0 ± 0.1%. To determine whether this performance was meaningful, we estimated a baseline performance by training the real patient model on the same input, but with the output randomized. This randomized baseline model only selected patients' choices about 25.0 ± 0.0% of time, suggesting that our real patient NN is a significant improvement and a reasonable approximation of

Figure 3. (Color online) The Structure of the Neural Net with Shared Weights Between Branches

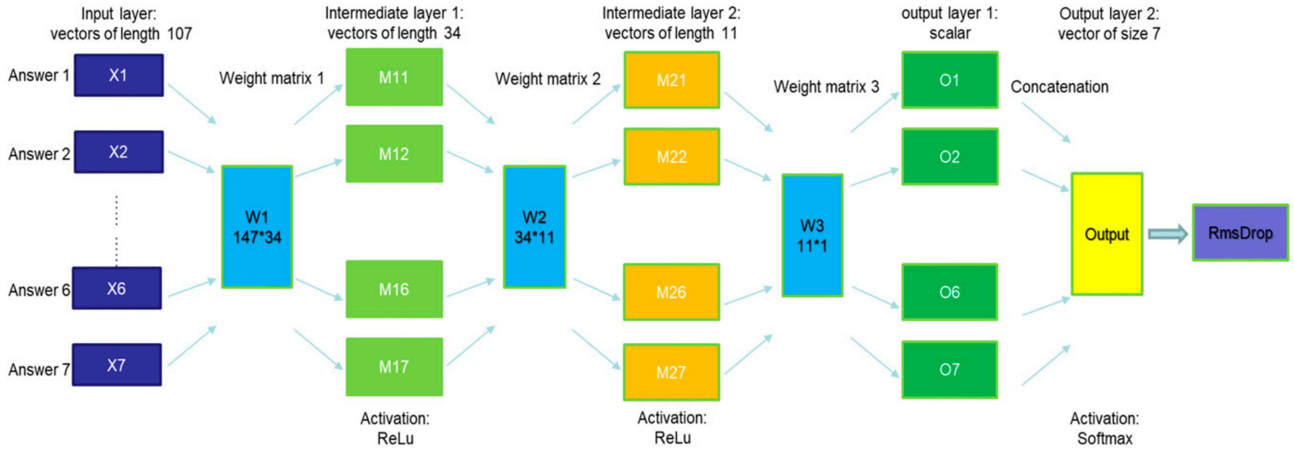
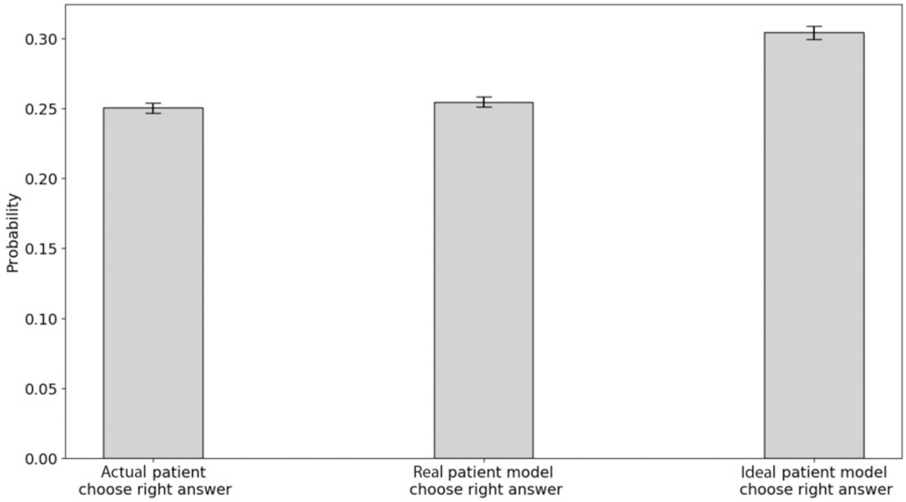


Figure 4. Comparison of Actual Patients to Real and Ideal Patient NN Models



an actual patient. The ideal patient NN was trained on an 80% training split (~2,500) of the ~3,000 questions and answers that were evaluated by our panel of experienced physicians in China, minimizing the cross-entropy between the model's chosen answer and the actual best answer chosen by experts. Our results are displayed in Figure 4.

In the data set of ~3,000 questions evaluated by our panel, the actual patients selected the (evaluator-determined) best advice only 25.4% ± 0.4% of time.

To fairly compare how the real and ideal patient NN models performed in selecting the best advice, we assessed them on the 20% test split of the ~3,000 questions evaluated by our panel (which was not used to train either model). The real patient NN selected the best advice with similar performance as actual patients (25.0 ± 0.4%). This is to be expected,[10] as the real patient NN was trained to mimic the actual patients. However, we should expect the ideal patient NN to pick the best advice more often because it is trained to make the best choice. As shown in Figure 4, the ideal patient NN (30.4 ± 0.4%) outperformed the real patient NN (25.4 ± 0.5%) on choosing the best advice by 6.0% (t-statistic = −60.06, $p < 0.001$). This hypothetical ideal patient model allows us to estimate the best-case scenario for patients' evaluation, using only heuristic processing. The gap in performance between real and ideal patients can be attributed to a combination of subjective bias in patient evaluation and a tendency for physicians to incorporate heuristic features that evoke a negative response in good advice. However, the overall performance is still poor. That is, even an ideal patient is limited in her ability to make the best choice because of her lack of medical expertise and reliance on only heuristic cues. This implies that the platform must reduce the need for patients to rely on heuristic cues to evaluate advice. Incorporating peer review of answers would convey signals of professional consensus to the patient. Permitting advice givers to comment on or up- or down-vote answers could accomplish this, and such mechanisms are commonplace in many SQPs (e.g., Stack Overflow). Given that medical practitioners are willing to answer questions on the platform in exchange for micropayments and platform reputation, similar incentives to evaluate and respond to advice on the platform would likely be viable.

Interpreting Neural Networks via the Perturbation Protocol

One of the challenges for neural networks is interpretation, as weights in the network are not directly associated with contributions of features. Multiple methods have been developed to explain neural network predictions: Garson's algorithm interprets the relative importance of predictors in connection with predicted outcomes by analyzing model weights. The Lek profile method explores the relationship of the outcome and a predictor by holding other predictors at constant values. Partial dependence plots visualize the relationship between an outcome and one or two predictors (reviewed by Zhang et al. 2018). However, a widely accepted and popular modern approach to achieve interpretability is a technique known as LIME (locally interpretable model-agnostic explanations). LIME is a model-agnostic method that provides local interpretability by perturbing the input of individual data samples to understand how the predictions change for those samples (Ribeiro et al. 2016). Our approach is most similar to LIME in that it also uses perturbations to understand predictions. However, we are interested in understanding the impact of a feature on predicted outcomes globally, across a large set of predictions made by the model.

Extending from the LIME technique of connecting local perturbations with predicted outcomes, we developed a similar perturbation procedure that allows us to make global explanations of input features' contributions to predictions. Using this protocol, we can identify how the presence or absence of heuristic features in an answer relates to the likelihood of a patient selecting it. Recall that all feature variables are binary indicators that encode inclusion (for naturally binary features) or discretized level (for count or log-transformed count variables). We define a *tune-up perturbation* of a feature as a change from absence to presence (for naturally binary features, such as existence of a preface or disclaimer in an answer) or from one discretized level to the next higher level (e.g., from low to medium, for count or log-transformed count features, such as the number of words that express comfort). We define *tune-down perturbations* similarly. Tune-up or tune-down perturbations of a particular feature are not operationally possible for every answer to a question. For example, answers that express the highest/lowest value of a discretized feature cannot be further increased/decreased. We define an answer as *eligible* for tune-up/tune-down perturbations for feature X if it is operationally possible.

Our perturbation analysis procedure is performed as follows. For each feature X, let U_X (D_x) define the set of all answers that are eligible for tune-up (tune-down) perturbations. We first check whether the size of these sets ($|U_X|$, $|D_X|$) are sufficiently balanced (and stop if they are not). We bootstrap sample $|U_X|$ ($|D_X|$) times with replacement from these sets to obtain the sets U_X^S, D_X^S (where s indexes the sampled set). We tune-up perturb feature X in each answer in the set U_X^S, leaving the features of all other answers to the same question unchanged, and feed the entire feature vector for all answers into the neural network model. We define $\Delta P_{U^s,x}$ as the fraction of tune-up perturbed answers that the neural network model chose. We repeat the procedure with tune-down perturbation, yielding $\Delta P_{D^s,x}$. If the effect of tune-up and tune-down perturbations on the predicted outcome is consistent, then we expect $\Delta P_{D^s,x}$ to be similar to $\Delta P_{U^s,x}$ in magnitude but opposite in sign. We repeat this procedure 100 times and pool the results together to form the set $\{\Delta P_{S,x}\} = \{\Delta P_{U^s,x}\} \cup \{-\Delta P_{D^s,x}\}$ of size 200. Finally, we estimate the mean (ΔP_X) and lower and upper 70% and 90% confidence intervals from this set. If tune-up and tune-down perturbations have inconsistent effects on predicted outcomes, then this measure will vary substantially across samples, ultimately yielding large confidence intervals.

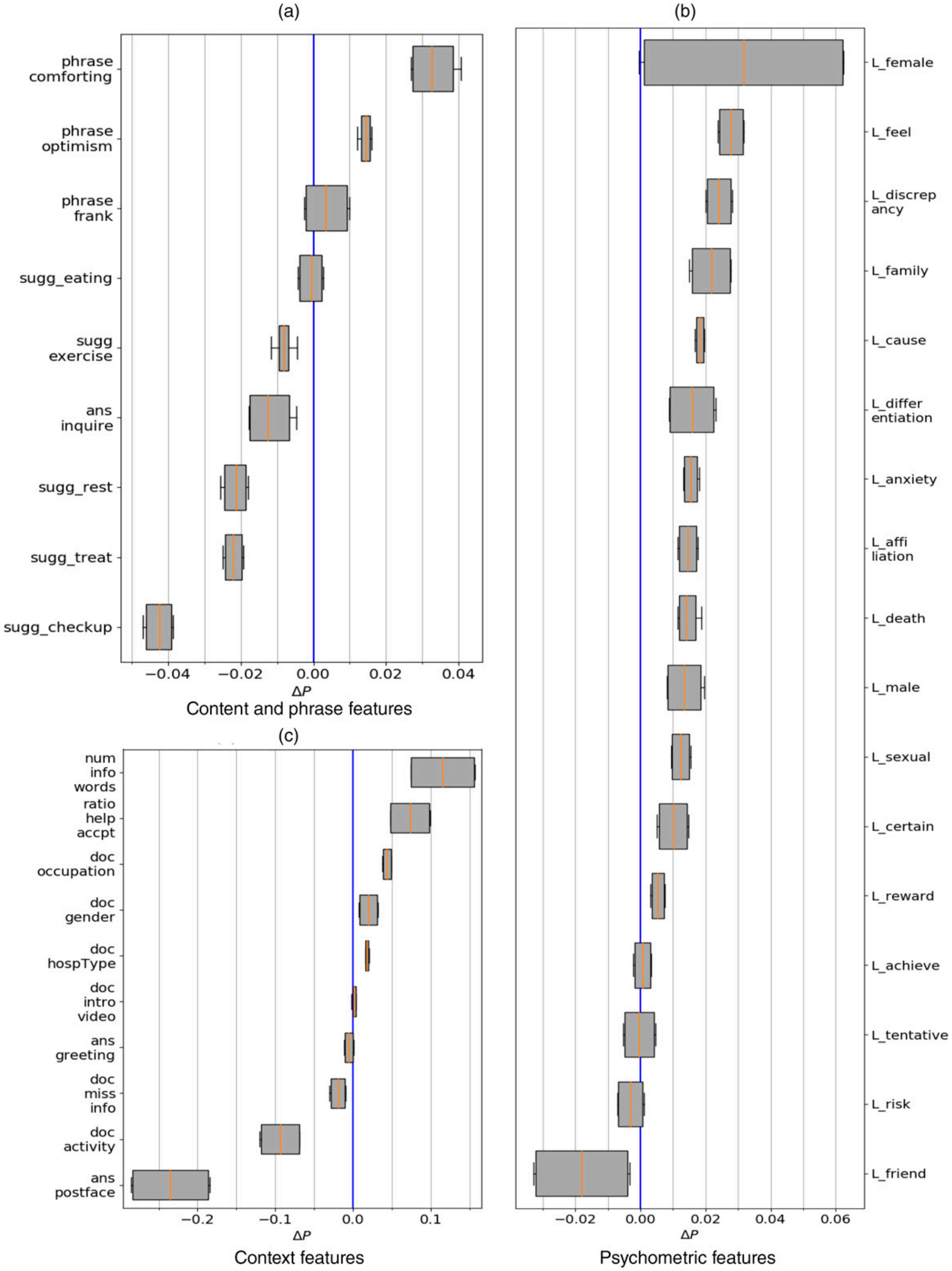
To determine which heuristic features explain patient choice, we performed the above perturbation protocol on the real patient model by selecting and perturbing questions from the entire data set of ~110,000 questions evaluated by patients. We are also interested in how features increase the likelihood of patients selecting *the most correct answer*. To determine this, we repeated the perturbation procedure on the real patient model by selecting and perturbing the features of the *best answer* of the subset of questions in the evaluated data set where actual patients had selected a suboptimal answer (this occurred in ~2,200 questions). However, we found that the estimates were nearly identical, though with wider confidence intervals due to the limited size of the data (see the online appendix). This indicates that *features have the same impact on patient evaluation regardless of the extent to which the answer is correct*. In other words, real patients (and the real patient model) respond to features of any answer in the same way, regardless of its quality, as they cannot discriminate the quality of answers. It also implies that any potential correlation between features and the tendency for an answer to be correct are not driving our findings. The impact of all feature perturbations on the probability for a patient to accept an answer are displayed in Figure 5. Mean estimates ΔP_X and lower and upper 90% confidence intervals are displayed in the first three columns of Table 4.

We can also leverage our perturbation protocol to understand how an ideal patient might *weigh heuristic features differently*. To do so, we performed the above perturbation protocol on both the real and ideal patient models. For each feature, we calculated the difference $\Delta P_X^{ideal} - \Delta P_X^{real}$ and its lower and upper 90% confidence intervals. The results are displayed in last three columns of Table 4.

Figure 5 displays the impact of perturbation of a given feature on the change in probability for a patient to select an answer, for (a) content and phrase features, (b) context features, and (c) psychometric features. To facilitate comparison of effect sizes, features are displayed in rank-descending order. Lines denote mean values; boxes denote 70% confidence intervals; whiskers denote 90% confidence intervals. Note that the scale of the y-axis is different across subplots.

From these results we can draw several inferences on how patients respond to advice on OHQPs (RQ2). Below, we describe these inferences in terms of contextual cues, content and phrase cues, and the psychometric aspects of language used in advice. We include effect sizes as increases or decreases in a patient's probability to accept the answer when the feature in question is perturbed. Some of these findings have immediate policy implications: doctors can be encouraged to incorporate features that contribute positively to patients' decisions and discouraged from incorporating features that contribute negatively. It is important to note that any potential policy should refrain from

Figure 5. (Color online) The Impact of Features on Patient Selection



making recommendations that would alter actual diagnoses, prognoses, treatment, or care suggestions. Instead, policy recommendations could focus on the manner that advice is communicated in terms of the psychometric characteristics of language used and other aspects of advice (such as overall length or inclusion of copied postface text) or platform participation (such as doctor profile completeness).

Contextual Cues

The quality of a doctor's reputation and participation on the platform could send a strong signal to patients

Table 4. The Perturbation Impact of Features on Patients' Probability to Accept an Answer (ΔP)

	Real patient model (%)			Ideal minus real (%)		
Features	ΔP	LCI	UCI	Diff(ΔP)	LCI	UCI
ans_postface	−23.52	−28.51	−18.38	30.06	24.91	34.00
doc_activity	−9.39	−11.95	−6.86	20.44	16.36	24.66
sugg_check	−4.23	−4.71	−3.86	15.93	13.96	17.71
sugg_treat	−2.21	−2.50	−1.93	12.27	9.32	15.48
sugg_rest	−2.13	−2.57	−1.81	−26.46	−31.70	−23.46
doc_miss_info	−1.85	−3.04	−0.93	—	—	—
L_friend	−1.82	−3.28	−0.34	−8.12	−9.82	−6.84
ans_inquire	−1.26	−1.79	−0.46	—	—	—
sugg_exercise	−0.82	−1.16	−0.45	−8.54	−11.4	−4.39
ans_greeting	−0.53	−1.11	0.06	3.23	1.71	5.08
L_risk	−0.31	−0.71	0.12	3.62	2.24	4.94
sugg_eat	−0.06	−0.44	0.28	−7.57	−8.92	−6.11
L_tentative	−0.04	−0.54	0.45	−3.63	−4.99	−2.28
L_achieve	0.07	−0.22	0.33	−8.29	−9.67	−6.99
doc_intro_vid	0.15	−0.17	0.40	—	—	—
phrase_frank	0.34	−0.25	1.00	−9.21	−13.4	−6.32
L_reward	0.53	0.31	0.76	−1.01	−1.85	−0.04
L_certain	1.00	0.51	1.48	1.18	−0.25	2.75
L_sexual	1.23	0.94	1.54	0.83	−2.04	3.95
L_male	1.34	0.81	1.97	6.43	4.31	10.10
L_death	1.40	1.15	1.88	−17.82	−23.64	−10.08
phrase_optimism	1.45	1.2	1.6	4.05	−0.17	7.83
L_affiliation	1.45	1.13	1.76	−4.58	−5.54	−3.49
L_anxiety	1.53	1.32	1.81	−12.89	−14.38	−11.07
L_differentiation	1.58	0.87	2.30	2.01	−1.69	5.92
doc_hosp_type	1.82	1.60	2.08	—	—	—
L_cause	1.82	1.68	1.96	0.22	−0.60	1.16
doc_gender	1.96	0.75	3.22	—	—	—
L_family	2.18	1.50	2.78	−2.84	−4.2	−0.54
L_discrepancy	2.40	2.01	2.81	0.50	−0.87	1.65
L_feel	2.78	2.37	3.18	−2.29	−3.90	−0.59
L_female	3.17	−0.04	6.24	−7.98	−11.39	−3.61
phrase_comforting	3.26	2.69	4.06	−4.39	−7.39	−2.84
doc_occupation	4.40	3.82	4.97	−5.62	−9.37	−1.95
ratio_help_accept	7.36	4.79	9.93	−4.52	−10.17	1.15
num_info_words	11.6	7.44	15.7	−17.38	−33.14	−2.02

Note. LCI and UCI refer to lower and upper 90% confidence intervals of ΔP in columns 2 and 3 and lower and upper 90% confidence intervals of Diff(ΔP) in columns 5 and 6.

when evaluating advice. Our findings show that patients had preference for advice from doctors with larger ratios of accepted advice, with a 7.4% increase in probability to accept advice for each quantile increase in the ratio of accepted advice (e.g., from below 0.15% to below 5%). Notably, the ideal patient model weighed platform reputation less (though the difference is marginal). When coupled with our previous finding that patients on average accept the best advice in less than one-third of all cases, this is extremely troubling. The reputation metric (accepted advice ratio) provided on the platform *increases the tendency for patients to accept advice from physicians who are prone to give suboptimal advice*. Moreover, it increases the moral hazard for physicians to offer less accurate but more enticing advice for a better chance of having their advice selected. In other words, physicians who dispense less accurate advice that is more appealing to patients not only benefit from bounty, but also from increased platform reputation. Indeed, doctors who ranked in the top 25% in the ratio of accepted advice on average scored lower in their answers (4.17 versus 4.33) and were less likely to provide the best answer according to our evaluators (21.8% versus 25.3%), yet had answers that were more likely to be selected by patients (19.0% versus 12.0%) than those who ranked in the bottom 25%. The evidence points to a flawed reputation system that tends to misguide patients and increase moral hazard of advice givers. Patients also seemed to exhibit a decreased preference for advice from physicians who are highly active on the platform, with a 9% decrease in probability to accept advice for each quantile increase in number of questions answered (e.g., from below 100 to below 10,000). In contrast, the ideal patient model tended to prefer answers from doctors that were more active on the platform.

Patients also reacted to cues that reflect doctors' credibility and ability, such as doctors' qualifications and the ranks of their hospitals. Our findings show that patients were 4.4% more likely to accept advice from doctors with a higher qualification (from nondoctor medical practitioners[11] to nurses to residents to chief physicians). In contrast, the ideal patient model tended to discount the ranks of doctors. We found that higher-ranked doctors, on average, did not provide higher-quality advice. Patients were also more likely to accept answers from doctors coming from larger hospitals, with a 1.8% increase in probability. They seemed to dislike advice from doctors with incomplete profiles, with a 1.9% decrease in probability to accept advice from doctors when their profile was perturbed to include missing information. There is also evidence that patients exhibit a gender bias, preferring answers from male doctors, with a 2.0% increase in probability. This is consistent with studies that have shown patient preference for male doctors in offline settings (Schmittdiel et al. 2000); though, the online nature of OHQPs rules out some explanations for gender bias (such as comfort with physical examination) that pertain to offline settings.

Content and Phrases

Regarding the informational content of answers, patients appreciated longer answers (a possible indicator of physician effort), but only when the content of advice was personalized and informative. Increasing the number of informative words in an answer by one quantile (e.g., from below 100 to below 300) leads to a 11.6% increase in probability for patients to accept the answer. In contrast, many doctors active on the OHQP habitually concluded with postfaces—large sections of text in their answers that are identical across all advice they dispense. Patients strongly disliked the inclusion of such uninformative text in answers, with an 23.5% decrease in probability to select an answer when uninformative text was included, perhaps because it increases cognitive load. In contrast, the ideal patient model preferred answers with such postfaces, perhaps because they are correlated with advice quality.

Regarding diagnosis, prognosis, and suggestions, patients disliked the suggestion to go to a hospital for further examination (4.2% decrease in probability to accept advice when this suggestion was included) or treatment (2.2% decrease in probability to accept advice when this suggestion was included). In contrast, the ideal patient model strongly preferred such suggestions. Indeed, in 22% of all cases where patients opted for suboptimal advice, the answer they selected did not suggest in-person care or treatment, whereas the best answer (according to our panel of evaluators) did. This did not occur only for suboptimal answers that were close in quality to the best answer. There were 138 questions (out of ~3,000 questions that were evaluated by our panel of physicians) where the average quality of answers that recommend in-person care exceed that of those that did not by more than one rating point. Among these, patients selected the lower-rated answers (that did not suggest follow-up care) in 42% of all cases (52 out of 138 questions). Taken together, this is strong evidence that OHQPs enable or exacerbate care avoidance by providing patients with advice from a medical expert that sanctifies their preference to avoid in-person care (RQ3). Patients also disliked suggestions to exercise (0.8% decrease in probability if included) or rest (2.1% decrease in probability if included).

In terms of common phrases used in advice, patients showed preference for the use of comforting phrases (e.g., “it’s okay,” “don’t worry,” “not a serious issue”) by 3.3%. The ideal patient discounted answers with comforting phrases. Patients also seemed to prefer the use of optimistic phrases by 1.5% (e.g., “can be cured,” “not life-threatening,” “prognosis is good,” “low risk”). On the other hand, patients did not react consistently to frank language pertaining to prognosis (“consequences could be serious,” “prognosis is bad,” “difficult to fully cure”).

Psychometric Aspects of Communication

In terms of psychometric aspects of communication, patients preferred language that communicated understandable explanations and the limits of understanding in terms of discrepancy (2.4%; e.g., “可能” (“possible”), “可以” (“could,” “able to,” “may”)), differentiation (1.6%; e.g., “不能” (“cannot”), “可能” (“perhaps”), “而” (“rather”), “除了” (“except”), “比较” (“compare,” “contrast”), etc.), and cause (1.8%; e.g., “引起” (“cause”), “导致” (“lead to”), “基于” (“based on”), etc.). They also preferred language that acknowledged what they were feeling (2.9%; e.g., “痛” (“pain”), “伤害” (“hurt”), “严重” (“serious” [feeling], “heavy” [feeling]), or acknowledged anxiety (1.5%; e.g., “不适” (“uncomfortable”), “担心” (“worry”), “紧张” (“nervous”), etc.).[12]

Suggestions that incorporated language specific to female (e.g., “母” (“mom”), “妈” (“mother”)) or male (e.g., “他” (“he”), “男孩” (“boy”)) gender were preferred by patients, with 3.2% (though with a relatively wide confidence interval) and 1.3% increases in adoption probability, respectively. Language relating to the social involvement of others in patients’ lives such as affiliation (e.g., “帮助” (“help,” “assist”), “伴” (“companion”), “伙伴” (“partner”)), family (e.g., “宝宝” (“baby”), “怀孕” (“pregnant”)), and sexual (e.g., “孕” (“pregnant”), “性” (“sex”)) were preferred by patients (respectively associated with 1.45%, 2.2%, and 1.2% increases in probability of accepting advice), though this may not be relevant to all medical questions.

Our perturbation analysis reveals how features are correlated with the probability of answers being chosen by patients. High positive estimates for features suggested potential favorability, though we cannot conclude that such features *lead* to the selection of an answer. We also explored how the combination of features of an answer increased its overall agreeableness to patients. We define the overall agreeableness score of an answer by combining the contributions from the large set of heuristic weight estimates for features that are present in the answer, using the standard inverse variance weighting procedure. We defined more (less) agreeable answers as those with an agreeableness score within the top (bottom) 20%. We compared the quality scores of professionally evaluated answers to their agreeableness and found that more agreeable answers had slightly lower quality (4.33 for agreeable, 4.15 for less agreeable; p-value of difference of <0.001). However, the percentage of agreeable answers that were rated as poor quality (score of <3) was significantly larger for more agreeable answers (11% more agreeable versus 5% less agreeable; $p < 0.001$). This indicates that heuristic features that were favored by patients were higher in answers with lower quality.

Knowing how patients react to the language used by physicians on the platform could be leveraged to guide doctors in communicating advice to patients online (RQ3). For example, we could suggest physicians provide “a spoonful of sugar to help the medicine go down.” We can estimate an upper bound on the efficacy of such a policy by using a variation of our perturbation protocol. Starting with all questions where patients chose the wrong answer, we simultaneously perturb up (down) all features that evoke positive (negative) responses that could reasonably be changed by such a policy without altering diagnoses, prognoses, or suggestions for care.[13] Doing so leads to a 53% increase in the proportion of correctly chosen answers.

Discussion and Conclusion

Online health Q&A platforms have become a major business success likely because, compared with traditional brick-and-mortar hospital or office visits, they provide a channel for patients to reach out to doctors that is faster and more economical and convenient. To understand patient evaluation in OHQPs, we collected and analyzed data from one of the largest OHQPs. The data consist of 496,842 answers to 114,037 questions, and profile information from 16,828 doctors. To our knowledge, ours is the first large-scale study to use empirical data to analyze patients' decision-making processes and evaluation behavior and quality of medical advice on OHQPs. Our findings acknowledge the importance and contribution of OHQPs in providing online care to patients. OHQPs and other online healthcare consulting platforms seem to be a valuable complement to in-person care and, as such, could see substantial growth worldwide. More importantly, OHQPs may be vital in circumstances when access to traditional points of care is limited. For example, during the COVID-19 pandemic, particularly vulnerable or immunocompromised patients (such a cancer survivors), were advised to delay hospital visits because of the possibility of contracting the coronavirus. In such circumstances, OHQPs can play a vital role in providing medical advice to patients in need. Overall, although we find that the OHQP we studied on average promoted good medical advice, our results also reveal problems with patient evaluation. Although patients had a reasonable answer to choose for most questions, they chose suboptimal advice in over two-thirds of all cases and often chose poor advice when it was offered over high-quality alternatives. Accepting poor-quality advice is likely a precursor to bad (and potentially even catastrophic) healthcare outcomes. This is particularly the case for patients with more serious conditions (cancer and heart and liver disease) or vulnerable patient groups (pediatrics), where patients tend to perform even worse than average in selecting good answers.

To understand why patients perform poorly as evaluators, we estimated the impact of heuristic cues on patient evaluation behavior. Using extensive natural language processing of the full text content of answers, we codified a rich set of heuristic features on the content, phrases, context, and psychometric language in advice. We estimated the impacts of these features on the tendency for patients to select advice using state-of-the-art deep neural networks trained to mimic the decision making of patients. We found that heuristic processing can explain a substantial amount of the variation in patient decision making when selecting advice. Importantly, we identified the extent to which patients respond positively or negatively to different features. Our finding that patients have a strong negative reaction to suggestions to seek offline follow-up care or treatment suggests that OHQPs enable or exacerbate care avoidance. A recent study shows that about 48.9% of the Chinese population did not to seek in-person treatment when they were sick, and 29.6% chose not to be hospitalized when they should have been (National Health Commission of the People's Republic of China 2016). Care avoidance behavior is more prevalent in poorer, more rural areas, where OHQPs may be preferred over other points of consultation and care because of their lower cost. Our study suggests OHQPs exacerbate these tendencies as they yield professional advice that provides further justification for not seeking treatment. There have been multiple reports that patients delayed their treatment because of blind trust of online opinions that nearly lead to death (Sohu 2015, Wang et al. 2020). Practical efforts to address the problems with OHQPs may include policy interventions that affect patients, physicians, or alter other platform mechanisms.

Our findings indicate several aspects of the design of these platforms that seem to threaten patient reception and subsequent adoption of high-quality health advice. The lack of search functionality that is typical in most OHQPs limits patients' ability to assess existing advice on the platform that may be related to their condition. On the other hand, better search functionality could reduce the volume of new questions (and ultimately, platform revenue) by removing the need to ask new questions (as they can more easily find the best matched answers to their questions by searching). Lay evaluation of advice solely by the patient, who we found is unable to assess the accuracy of medical advice and susceptible to multiple forms of cognitive bias, seems clearly problematic. The lack of an expert peer review or rating mechanisms permits the few bad actors to provide poor advice off the official record, without negatively impacting their online reputation, and without conveying any kind of expert consensus to patients. Moreover, when combined with bounties and platform reputation, this can create a moral hazard for those bad actors who may be incentivized to tailor advice to meet patient preference at the expense of medical accuracy.

Our study also provides insights that apply beyond the immediate context of OHQPs. For example, it is important to understand how patients react to medical advice when it is delivered digitally, as this can affect patient follow-through and satisfaction. As remote, telemedicine, and digital consultation care provision grow, this will only become more important. Our study also brings up several issues surrounding regulation and oversight of digital healthcare provision. Although reputation is a common metric adopted by many platforms, we are not aware of any specific regulations regarding reputation metrics on health-related platforms. Yet, if it affects patient decisions, then care should be taken to ensure that reputation is correlated with quality of care. This holds not only for OHQPs but also for other sites and platforms in the digital healthcare ecosystem (such as patient review portals). We stipulate that post hoc evaluation of advice by peer experts should be an essential component of any online healthcare consulting platform, as it provide a confirmation of advice quality and an alternative (or supplemental) measure of reputation that is correlated with giving quality medical advice. For example, the platform askdr.co relies on offline reputation (such as providing and certifying doctors' credentials and displaying doctors' offline reputed performance) to ensure quality of performance online. However, in the OHQP we studied, we found that the quality of answers did not differ significantly between higher-ranked doctors (e.g., chief physicians) and lower-ranked doctors (e.g., residents). Although post hoc evaluation of advice is viable in OHQPs (where physicians seem willing to perform work in exchange for micropayments and reputation), it may be more difficult to implement in other online health consulting environments. Such post hoc evaluation (or peer review) would require extra work that may necessitate additional incentives, or that alternatively may need to be subsidized by platforms. By its nature, digital healthcare is more susceptible to abuse from bad actors than traditional care provision and regulation is also less established. We posit that in digital healthcare settings, expert peer review and/or auditing can increase quality of care, guard against bad actors, and deliver signals of professional consensus to patients to enable better choices.

Other policy changes include educating patients to be better evaluators (such as sending a tip to patients to encourage them to carefully evaluate advice that may be hard to hear) or educating physicians on "digital bedside manner" to better communicate advice on the platform analogous to training for bedside manner in offline interactions between patients and doctors (Rhee and Bird 1996, Anderson et al. 2007). Of the two, we believe the latter may be more effective. Our findings suggest that policies targeting patients and physicians on the platform could increase the probability of the patient selecting the best advice by up to 6% or 53%, respectively, which translates to selection of optimal advice in ~83,000 or ~731,000 more questions each year, under the best-case scenario. Real-world tests of the efficacy of physician or patient guidance policies warrant the rigor of randomized controlled experiments. More generally, a variety of interventions on both the supply and demand sides of OHQPs are possible, in terms of communication or platform mechanism changes, and could yield practical insights to deal with the poor patient evaluations in OHQPs.

Our study revealed the existence of bad actors on the platform, who, though in the minority, could benefit at the expense of harming patients. This is to be expected, as the platform did little to mitigate the risk of moral hazard. Though, our results showed that in general patients received good advice. We suspect this is due to the professionalism and conscientiousness of physicians on the platform. Still, OHQPs and other online healthcare platforms should take deliberate measures to weed out bad actors.

Beyond contextual insights, our study demonstrates the use of deep learning methods on large data on individual choices as a viable improvement over more conventional discrete choice modeling. The method is versatile enough to yield models that are mathematically equivalent to conventional discrete choice models without requiring specification assumptions that may be incompatible with the true data generation process. Indeed, we found that the deep learning method yielded superior predictive performance over discrete choice modeling. In terms of methodological contribution, the perturbation protocol that we describe is based on established methods in machine learning, is quite general, and allows for interpretable estimation of the impact of input features on the predicted outcome (output) of the neural network. The method can also be applied to other online expert consultation platforms with similar settings (such as legal advice).

Our approach is not without limitations. The OHQP we examined was one of the top three by usage, and we believe that our findings should generalize well to other OHQPs and even some other online healthcare environments, though there may be cases for which this does not hold. For example, the presentation of and response to context cues (such as doctor reputation, profile information, and so on) may vary across different platforms. Askers may also be sensitive to subtle changes in how platforms present information in general. We explored heuristic features that were understandable to the layperson and a good fit with our aim to understand the determinants of patient choice. However, our NLP approach would not be suitable for pulling out features of detailed medical diagnoses and prescriptive advice, given that the necessary medical lexicons are not well established (particularly in other languages). Furthermore, it may not work for cases where patients (or askers, in general) possess varying levels of professional knowledge.

Although neural net models are less susceptible to some kinds of bias because they function as universal approximators, they are still susceptible to omitted variable bias. Though, in our context, the threat of omitted variable bias is somewhat reduced because we have all the information that patients see and we have endeavored to encode a rich set of features that that are comprehensible to patients (who lack expertise to base their decisions on complex medical aspects of advice).

Overall, our study affirms the potential of OHQPs in delivering quality medical advice that is timely and economical, and can circumvent resource-based, geographic, or circumstantial barriers that limit access to care. We view our problematic findings of poor patient evaluation, a flawed reputation metric, exacerbation of care avoidance, and the existence of bad actors through the lens of potential platform design changes and policy solutions.

Acknowledgments

The authors acknowledge Lan Wang and Jayakanth Srinivasan for their helpful discussions and feedback.

Endnotes

[1] According to the 2017 official annual report release by the National Health Commission of China, http://www.nhc.gov.cn/wjw/ (accessed March 21, 2019).

[2] Most advice givers are officially registered medical workers, including doctors (majority), nurses, technicians, nutritionists, and psychologists. Less than 35% of them are unregistered or choose not to disclose their registered occupation. For the purposes of brevity, we use the terms doctors, physicians, and advice givers interchangeability to refer to all advice givers on the platforms.

[3] Most OHQPs (such as 120ask or Haodf) have very poor search function compared with SQPs such as Stack Overflow. The latter typically leverage both their own search engine and are indexed by external search engines (e.g., Google) to help askers identify relevant questions and answers. OHQPs, on the other hand, are often not indexed by external search engines (e.g., Baidu) and provide search results that are outdated, not meaningfully ranked, and observably poor matches to query terms.

[4] For a detailed summary of scores of all answers, see Table A1 in the online appendix.

[5] This neural network was trained on 80% of the ~110,000 questions whose answers were evaluated by an actual patient.

[6] This neural network was trained on 80% of the ~3,000 questions whose answers were evaluated by our panel of experts.

[7] Although almost all questions received at most seven answers, many received fewer (only 5% of questions received more than seven answers). We represented the feature vector of a missing answer as all zeros. This precludes the neural network from assigning a nonzero probability of selection for missing answers.

[8] Because of limitations of data size, particularly for the ideal patient model (for which we have only 3,000 professionally evaluated answers), we did not use formal cross-validation to explore the hyperparameters of the model (such as the number of intermediate layers and the number of nodes in intermediate layers). Instead, we explored variations to model structure separately, as tests of robustness, and determined that such variations did not yield substantial performance increases (see the online appendix for details).

[9] RMSProp is an unpublished algorithm first proposed in a Coursera course. For more information, see: https://optimization.cbe.cornell.edu/index.php?title=RMSProp.

[10] Data from these estimates exclude questions where best answers are tied and those with more than seven answers, yielding a slightly lower patient evaluation accuracy of 25%, relative to the accuracy of 31% for the entire data set.

[11] Nonregistered medical workers, referred to as "medical members" by the platform, are those who may know medicine but do not possess a license to practice medicine.

[12] For brevity, we include in parentheses the percentage changes in probability of a patient accepting an answer for the psychometric feature dimensions described in this paragraph. For the sake of interpretability, we include both original Chinese characters and phrase translations for typical examples of words or phrases within answers for each psychometric dimension from LIWC.

[13] Features that could reasonably be changed include all psychometric features, comforting phrases, missing information in the doctor's profile, use of postface text, and the number of informative words.

References

Agarwal R, Gao G, DesRoches C, Jha AK (2010) The digital transformation of healthcare: Current status and the road ahead. *Inform. Systems Res.* 21(4):796–809.

Anderson R, Barbara A, Feldman S (2007) What patients want: A content analysis of key qualities that influence patient satisfaction. *J. Medical Practice Management* 22(5):255–261.

Angst CM, Agarwal R (2017) Adoption of electronic health records in the presence of privacy concerns: The elaboration likelihood model and individual persuasion. *MIS Quart.* 33(2):339.

Angst CM, Agarwal R, Sambamurthy V, Kelley K (2010) Social contagion and information technology diffusion: The adoption of electronic medical records in U.S. hospitals. *Management Sci.* 56(8): 1219–1241.

Atasoy H, Chen Pu, Ganju K (2018a) The spillover effects of health IT investments on regional healthcare costs. *Management Sci.* 64(6):2515–2534.

Atasoy H, Demirezen EM, Chen PY (2018b) Impacts of patient characteristics and care fragmentation on the value of HIEs. Fox School of Business Research Paper No. 18-035. http://dx.doi.org/10.2139/ssrn.3191566.

Bae BJ, Yi YJ (2017) What answers do questioners want on social Q&A? User preferences of answers about STDs. *Internet Res.* 27(5):1104–1121.

Baidu Baike (2019a) Hospital classification standard. Retrieved March 21, https://baike.baidu.com/item/医院等级划分标准.

Baidu Baike (2019b) 120ask.com. Retrieved March 21, https://baike.baidu.com/item/有问必答网/4709400?fr=aladdin.

Bass SB, Ruzek SB, Gordon TF, Fleisher L, McKeown-Conn N, Moore D (2006) Relationship of Internet health information use with patient behavior and self-efficacy: Experiences of newly diagnosed cancer patients who contact the National Cancer Institute's Cancer Information Service. *J. Health Comm.* 11(2):219–236.

Bavafa H, Hitt LM, Terwiesch C (2018) The impact of e-visits on visit frequencies and patient health: Evidence from primary care. *Management Sci.* 64(12):5461–5480.

Cao X, Liu Y, Zhu Z, Hu J, Chen X (2017) Online selection of a physician by patients: Empirical study from elaboration likelihood perspective. *Comput. Human Behav.* 73:403–412.

Case DO, Andrews JE, Johnson JD, Allard SL (2005) Avoiding vs. seeking: The relationship of information seeking to avoidance, blunting, coping, dissonance, and related concepts. *J. Medical Library Assoc.* 93(3):353–362.

Chaiken S (1987) The heuristic model of persuasion. Zanna M, Olson J, Herman C, eds. *Social influence: The Ontario Symposium*, Vol. 5 (Lawrence Erlbaum, Hillsdale, NJ), 3–39.

Chaiken S, Trope Y, eds. (1999) *Dual-Process Theories in Social Psychology* (Guilford Press, New York).

Changyexinxi (2017) Analysis of the survival status of Chinese doctors, 2017. Retrieved March 21, 2019, http://www.chyxx.com/industry/201710/570290.html.

Festinger L (1962) Cognitive dissonance. *Sci. Amer.* 207(4):93–106.

Francis V, Korsch BM, Morris MJ (2010) Gaps in doctor-patient communication. *New England J. Med.* 280(10):535–540.

Gao G, Greenwood BN, Agarwal R, McCullough JS (2015) Vocal minority and silent majority: How do online ratings reflect population perceptions of quality? *MIS Quart.* 39(3):565–589.

Gerard HB, White GL (1983) Post-decisional reevaluation of choice alternatives. *Personality Social Psych. Bull.* 9(3):365–369.

Ghose A, Guo X, Li B, Dang Y (2021) Empowering patients using smart mobile health platforms: Evidence from a randomized field experiment. Preprint, submitted February 10, https://arxiv.org/abs/2102.05506.

Glöckner A, Witteman C (2010) Beyond dual-process models: A categorisation of processes underlying intuitive judgement and decision making. *Thinking Reasoning* 16(1):1–25.

Guo X, Dang Y, Gu B, He Y, Chen W (2017) The crowding out effect of monetary incentives—Evidence from a natural experiment on online healthcare platforms. *INFORMS 9th Conf. Inform. Systems*, Houston, TX.

Hao H, Zhang K, Wang W, Gao G (2017) A tale of two countries: International comparison of online doctor reviews between China and the United States. *Internat. J. Medical Inform.* 99: 37–44.

Harper FM, Moy D, Konstan JA (2009) Facts or friends? Distinguishing informational and conversational questions in social Q&A sites. *Proc. SIGCHI Conf. Human Factors Comput. Systems* (Association for Computing Machinery, New York), 759–768.

Hart W, Albarracín D, Eagly AH, Brechan I, Lindberg MJ, Merrill L (2009) Feeling validated vs. being correct: A meta-analysis of selective exposure to information. *Psych. Bull.* 135(4):555–588.

Haskard Zolnierek KB, Dimatteo MR (2009) Physician communication and patient adherence to treatment: A meta-analysis. *Medical Care* 47(8):826–834.

Hornik K, Stinchcombe M, White H (1989) Multilayer feedforward networks are universal approximators. *Neural Networks* 2(5):359–366.

Huang CL, Chung CK, Hui N, Lin YC, Seih YT, Lam BCP, Chen WC, Bond MH, Pennebaker JW (2012) The development of the Chinese Linguistic Inquiry and Word Count dictionary. *Chinese J. Psych.* 54(2):185–201.

Hyman HH, Sheatsley PB (1947) Some reasons why information campaigns fail. *Public Opinion Quart.* 11(3):412–423.

Izuma K, Matsumoto M, Murayama K, Samejima K, Sadato N, Matsumoto K (2010) Neural correlates of cognitive dissonance and choice-induced preference change. *Proc. Natl. Acad. Sci. USA* 107(51):22014–22019.

Jain SP, Maheswaran D (2002) Motivated reasoning: A depth-of-processing perspective. *J. Consumer Res.* 26(4):358–371.

Jin J, Yan X, Li Y, Li Y (2016) How users adopt healthcare information: An empirical study of an online Q&A community. *Internat. J. Medical Inform.* 86:91–103.

Kaplan SH, Greenfield S, Ware JE (1989) Assessing the effects of physician-patient interactions on the outcomes of chronic disease. *Medical Care* 27(3, Suppl)S110–S127.

Kim S, Oh JS, Oh S (2007) Best-answer selection criteria in a social Q&A site from the user-oriented relevance perspective. *Proc. Amer. Soc. Inform. Sci. Tech.* 44(1):1–15.

Kunda Z (1990) The case for motivated reasoning. *Psych. Bull.* 108(3):480–498.

Lapointe L, Ramaprasad J, Vedel I (2014) Creating health awareness: A social media enabled collaboration. *Health Tech.* 4(1):43–57.

Liu X, Yao Y, Deng Z (2017) An empirical study of customer satisfaction and loyalty on health websites. *WHICEB 2017 Proc.*, vol. 1 (Association for Information Systems, Atlanta), 621–630.

Liu TX, Yang J, Adamic LA, Chen Y (2014) Crowdsourcing with all-pay auctions: A field experiment on Taskcn. *Management Sci.* 60(8):2020–2037.

Lu SF, Rui H (2015) Can we trust online physician ratings? Evidence from cardiac surgeons in Florida. Proc. 48th Annual Hawaii Internat. Conf. Systems Sci. (Institute of Electrical and Electronics Engineers, Piscataway, NJ), 2876–2885.

Meservy TO, Jensen ML, Fadel KJ (2014) Evaluation of competing candidate solutions in electronic networks of practice. *Inform. Systems Res.* 25(1):15–34.

Mishra AN, Anderson C, Angst CM, Agarwal R (2012) Electronic health records assimilation and physician identity evolution: An identity theory perspective. *Inform. Systems Res.* 23(3 part 1): 738–760.

Morris MR, Teevan J, Panovich K (2010) What do people ask their social networks, and why? A survey study of status message Q&A behavior. *Proc. SIGCHI Conf. Human Factors Comput. Systems* (Association for Computing Machinery, New York), 1739–1748.

National Health Commission of the People's Republic of China (2016) The 3rd national health service investigation and analysis. Accessed August 10, 2022, http://www.nhc.gov.cn/wjw/index.shtml.

Nickerson RS, Bias C (1998) A ubiquitous phenomenon in many guises. *Rev. General Psych.* 2(2):175–220.

Nie L, Li T, Akbari M, Shen J, Chua TS (2014) WenZher: Comprehensive vertical search for healthcare domain. Proc. 37th Internat. ACM SIGIR Conf. Res. Development Inform. Retrieval (Association for Computing Machinery, New York), 1245–1246.

Oh S, Yi YJ, Worrall A (2012) Quality of health answers in social Q&A. *Proc. Amer. Soc. Inform. Sci. Tech.* 49(1):1–6.

Pohl RF (2004) *Cognitive Illusions: A Handbook on Fallacies and Biases in Thinking, Judgement and Memory* (Psychology Press, London).

Reach G (2015) Patients' nonadherence and doctors' clinical inertia: Two faces of medical irrationality. *Diabetes Management* 5(3):167–181.

Rhee KJ, Bird J (1996) Perceptions and satisfaction with emergency department care. *J. Emergency Med.* 14(6):679–683.

Ribeiro MT, Singh S, Guestrin C (2016) "Why should I trust you?" Explaining the predictions of any classifier. Proc. 22nd ACM SIGKDD Internat. Conf. Knowledge Discovery Data Mining (Association for Computing Machinery, New York), 1135–1144.

Rodoletz M, Mangan CE, Miller SM, Sedlacek TV, Schroeder CM (2005) Applications of the monitoring process model to coping with severe long-term medical threats. *Health Psych.* 15(3):216–225.

Schmittdiel J, Grumbach K, Selby JV, Quesenberry CP (2000) Effect of physician and patient gender concordance on patient satisfaction and preventive care practices. *J. General Internal Med.* 15(11):761–769.

Shah C, Oh S, Oh JS (2009) Research agenda for social Q&A. *Library Inform. Sci. Res.* 31(4):205–209.

Sohu (2015) Blind trust in online doctors almost delayed the treatment until it is too late. Accessed March 21, 2019, http://www.sohu.com/a/32011460_160929.

Sohu (2017) How many doctors and nurses are there? Here are the offical figures. Accessed March 21, 2019, https://www.sohu.com/a/192148447_758942.

Song Y, Sahoo N, Ofek E (2019) When and how to diversify—A multi-category utility model for personalized content recommendation. *Management Sci.* 65(8):3737–3757.

Svenson O (1992) Differentiation and consolidation theory of human decision making: A frame of reference for the study of pre- and post-decision processes. *Acta Psychologica* 80(1–3):143–168.

Sweeny K, Melnyk D, Miller W, Shepperd JA (2010) Information avoidance: Who, what, when, and why. *Rev. General Psych.* 14(4):340–353.

Trope Y (1979) Uncertainty-reducing properties of achievement tasks. *J. Personality Soc. Psych.* 37(9):1505–1518.

Trope Y, Bassok M (1982) Confirmatory and diagnosing strategies in social information gathering. *J. Personality Soc. Psych.* 43(1):22–34.

Trumbo CW (1999) Heuristic-systematic information processing and risk judgment. *Risk Anal.* 19(3):391–400.

Wang W, Wu Y, Liang C (2020) The responsibility and risk prevention of online healthcare. Accessed August 10, 2022, https://www.sohu.com/a/426068207_120051855.

Wei Z, Watts SA (2008) Capitalizing on content: Information adoption in two online communities. *J. Assoc. Inform. Systems* 9(2):72–93.

Wu B (2018) Patient continued use of online healthcare communities: Web mining of patient-doctor communication. *J. Med. Internet Res.* 20(4):e126.

Yan L, Tan Y (2014) Feeling blue? Go online: An empirical study of social support among patients. *Inform. Systems Res.* 25(4):690–709.

Yang J, Adamic LA, Ackerman MS (2008) Crowdsourcing and knowledge sharing. Proc. Ninth ACM Conf. Electronic Commerce (Association for Computing Machinery, New York), 246–255.

Yaraghi N, Du AY, Sharman R, Gopal RD, Ramesh R (2015) Health information exchange as a multisided platform: Adoption, usage, and practice involvement in service co-production. *Inform. Systems Res.* 26(1):1–18.

Yi YJ (2018) Sexual health information-seeking behavior on a social media site: Predictors of best answer selection. *Online Inform. Rev.* 42(6):880–897.

Zhang Y (2010) Contextualizing consumer health information searching. Proc. *First ACM Internat. Health Inform. Sympos.* (Association for Computing Machinery, New York), 210–219.

Zhang Z, Beck MW, Winkler DA, Huang B, Sibanda W, Goyal H (2018) Opening the black box of neural networks: methods for interpreting neural network models in clinical applications. *Ann*. Translational Medicine 6(11):216–216.