When Peers Matter More:

The Dominance of Social Influence Over Algorithm Accuracy in Patient Decision Making

Abstract

This study explores the impact of technical and social information on human acceptance of algorithmic recommendations in a healthcare context. Specifically, we examine interactions between patients and an artificial intelligence (AI) gatekeeping app that recommends specialty care based on symptoms provided by patients. In a field experiment conducted at a large hospital, patients were randomly assigned to four groups, each receiving AI recommendations in one of four information disclosure scenarios: (1) recommendation only, (2) recommendation with the technical information of algorithm accuracy, (3) recommendation with the social information of peer acceptance, and (4) recommendation with both algorithm accuracy and peer acceptance. Our analysis shows that patient acceptance of AI recommendations increases in algorithm accuracy or peer acceptance when each information is presented independently. When presented together, algorithm accuracy and peer acceptance are complementary in influencing patient acceptance of AI recommendations, demonstrating a positive interaction effect. However, their marginal effects are asymmetric, with peer acceptance dominating algorithm accuracy. Furthermore, the impact of information disclosure depends on patient expertise level: patients with less capability of articulating symptoms or new patients without previous hospital experience are more likely to be influenced by the disclosed technical and social information.

Keywords: human-AI interaction, algorithm aversion, healthcare, gatekeeping, information disclosure, herding

1. Introduction

Artificial intelligence (AI) has increasingly influenced various aspects of our lives, particularly in shaping our decision-making processes through AI-based recommendations (Gomez-Uribe and Hunt 2016, Linden et al. 2003). However, despite the growing attention to human-AI interactions across diverse decision-making contexts, there is a notable scarcity of causal studies that explore how humans respond to AI recommendations (Brenner and Meyll 2020, Sun et al. 2022, Lu 2023, Kim et al. 2024). This gap is especially significant in the healthcare domain, where it is crucial to understand the attitudes of patients who lack sufficient medical knowledge and must bear the consequences of medical uncertainties when considering AI-generated advice. Pioneering evidence suggests that patients often perceive recommendations from human doctors differently than those from AI systems. For instance, Promberger and Baron (2006) suggested that 80% of patients are likely to accept advice from a human doctor, citing greater trust in the physician's judgment. By contrast, a survey by Robertson et al. (2023) revealed a more divided opinion among U.S. respondents, with 47.1% open to AI-driven diagnosis and treatment clinics. While these contrasting anecdotes may arise from variations in research methodologies and contexts, they underscore a fundamental yet unresolved question: what truly drives the acceptance of AI advice in patient decision-making?

The literature on human attitudes toward algorithmic recommendations in broader contexts also presents inconclusive findings. Some studies reveal AI aversion, suggesting that humans naturally distrust AI due to its perceived lack of moral authenticity, sincerity, or affective feelings (Dawes et al. 1989, Dietvorst et al. 2014, Castelo et al. 2019, Grove and Meehl 1996). Conversely, others indicate AI appreciation driven by the belief that AI is more intelligent than human beings (Logg et al. 2019). In this study, we aim to analyze key drivers that influence human acceptance of AI recommendations and reconcile these two seemingly contrasting perspectives on human attitudes toward AI. Specifically, we aim to investigate how information disclosure affects patients' acceptance of AI recommendations.

We focus on the impact of information disclosure on human-AI interactions because economists have long demonstrated its role in reducing information asymmetry and addressing related challenges in various decision-making contexts (Lu 2016, Jin et al. 2021). In healthcare, information is even more critical, as patients often lack the essential knowledge that healthcare providers possess, leaving them uncertain about the appropriate care needed. Consequently, patients rely on the expertise and guidance of healthcare professionals to direct them to the right care, with primary care physicians

(PCPs) acting as gatekeepers in this process (Grumbach et al. 1999, Schoen et al. 2006, Sripa et al. 2019). However, the gatekeeping role of PCPs varies globally, and research shows several challenges associated with PCP referrals, with some countries lacking a formal referral system and others struggling with insufficient staffing to adequately support the referral process (Osborn et al. 2015).

To tackle the practical challenges of healthcare gatekeeping, we collaborated with a leading hospital in China that deployed an AI-enabled gatekeeping app in January 2023. The app's gatekeeping algorithm, trained on millions of medical records and refined through extensive feedback from healthcare experts, achieves diagnostic accuracy comparable to that of an average primary care physician. The app allows patients to bypass traditional pre-consultations with physicians. Instead, patients can input their symptoms and biometric data into a mobile app, which subsequently recommends the appropriate specialty for the patients to make appointments with.[1] Using this app, we conducted a large-scale experiment at the hospital from March to August 2023. We focus on two important types of information about the algorithm: *technical* information regarding its accuracy, which indicates the performance of the recommendation, and *social* information in the form of peer acceptance rates, which reflect collective judgment of prior patients with similar symptoms who received the same specialty recommendation. We randomly assigned patients to one of four groups with varying information disclosure scenarios: (1) recommendation only, (2) recommendation with *algorithm accuracy*, (3) recommendation with *peer acceptance*, and (4) recommendation with both *algorithm accuracy* and *peer acceptance*. We will answer the following research questions: 1) How do the two types of information independently and jointly affect patient acceptance of AI recommendations? 2) How do the characteristics of patients and their interactions with AI moderate the effect of the disclosed information?

We adopt the following strategies when conducting empirical analyses. In our first set of analyses, we compare group 1 (no disclosure) with either group 2 (disclosure of algorithm accuracy) or group 3 (disclosure of peer acceptance). This allows us to examine the *independent* effect of each type of information on patient acceptance of AI recommendations. Our findings indicate that the likelihood of patients accepting AI recommendations increases in algorithm accuracy or peer acceptance. In other words, patients are more likely to accept AI recommendations when the algorithm demonstrates higher accuracy or when the recommendations are supported by peer endorsement.

[1] The recommendation is at the specialty or department level (e.g., neurology as opposed to psychology) rather than individual physicians.

In our second set of analyses, we compare group 1 (no disclosure) with group 4 (disclosure of algorithm accuracy and peer acceptance) to examine the *combined* effect of simultaneously presenting both information. This analysis revealed an intriguing herding effect: when both types of information are disclosed, the influence of algorithm accuracy diminishes, and patients rely more heavily on peer information. Specifically, when peer acceptance is low, patients are reluctant to accept AI recommendations, regardless of the algorithm's accuracy. Conversely, when peer acceptance is high, patients consistently lean towards accepting AI recommendations, irrespective of algorithm accuracy. Nonetheless, algorithm accuracy and peer acceptance complement each other's impact, as evidenced by the significantly positive interaction effect of these two factors. Overall, these findings suggest that the social context in which AI recommendations are presented significantly shapes patients' perceptions of AI. Social influence can override the perceived technical quality of the AI, yet both types of information complement each other, enhancing their overall impact.

To further explore heterogeneity at the individual level, we investigate how patient characteristics moderate the effect of disclosed information on their decision-making. We find that patients lacking prior experience with the intended healthcare services or those who struggle to articulate detailed health symptoms to the AI app are more heavily influenced by the disclosed information. Because these patients are more uncertain about the appropriate healthcare services to seek, the information provided to them is thus more influential in their decision-making. These results further validate the effectiveness of our information disclosure design in nudging patients' acceptance of AI recommendations, particularly in scenarios characterized by information asymmetry.

Our work advances the human-AI interaction literature by offering important insights into how individuals interact with AI in choosing healthcare services, which differs significantly from traditional algorithmic recommendations in e-commerce settings (Gomez-Uribe and Hunt 2016, Brenner and Meyll 2020). Unlike e-commerce, where recommendations are typically tailored to consumer preferences and involve primarily monetary risks, healthcare decisions entail significantly higher uncertainty in matching and far greater nonmonetary stakes, such as poor health outcomes and high time costs. We demonstrate that AI aversion is not unchangeable; rather, it can be gradually mitigated as the algorithm matures and peer acceptance grows. More importantly, we show that the herding effect plays a pivotal role in encouraging broader acceptance and even appreciation of AI recommendations in decision-making, even when the technical performance of the algorithm itself is not perfect. As AI continues to integrate into high-stakes environments like healthcare, understanding

these dynamics will be crucial for designing systems that effectively support and enhance human decision-making.

The significance of our research is also underscored by its potential to improve physician resource utilization and patient care in the healthcare sector. Various challenges, such as PCP shortages (Aluttis et al. 2014, Green et al. 2013) and human gatekeeper referral errors (Freeman et al. 2017, 2021), have led to limited healthcare access and delays in receiving care, ultimately compromising patient outcomes and reducing the efficiency of the healthcare system. Our study offers a cost-effective solution by strategically disclosing key information to nudge patients toward accepting AI recommendations for appropriate healthcare services. This approach can potentially relieve pressure on overburdened PCPs, speed up access to specialty care, and reduce unnecessary healthcare expenditures by bypassing consultation PCP visits.

The rest of the paper is organized as follows: Section 2 provides a literature review. Section 3 develops the hypotheses. Section 4 describes the research context and the experiment design. Sections 5 and 6 discuss the main results and moderating effects. Section 7 concludes the paper and proposes future research directions.

2. Related Literature

This study connects two streams of literature on human-AI interactions: algorithm aversion and algorithm appreciation. In this section, we first review these two streams of literature and then summarize the contribution of our paper.

2.1 Algorithm Aversion

The academic literature has long documented humans' general resistance to algorithms, a phenomenon referred to as algorithm aversion. This literature can be traced back to Meehl's seminal work on clinical versus statistical prediction (Meehl 1954). Subsequent studies have provided supporting evidence of human resistance to using actuarial methods (i.e., statistical algorithms) in decision-making despite their superior performance (Dawes 1979, Dawes et al. 1989).

Early works have identified various reasons why algorithm aversion occurs, primarily based on anecdotal evidence. Dawes et al. (1989) suggested that professionals, particularly in the field of clinical psychology, resist statistical methods due to a lack of familiarity with these methods, overconfidence in their clinical judgment, and a misconception that group statistics do not account for individual heterogeneity. Similarly, Grove and Meehl (1996) demonstrated that mechanical

methods, which involve algorithmic and objective procedures (e.g., equation), are typically equal to or superior to clinical methods in predictive tasks. However, these methods face widespread resistance. Besides the perception of human superiority and concerns about the applicability of generic algorithms to specific populations, Grove and Meehl (1996) propose several additional reasons why professionals (i.e., physicians in their context) are skeptical of algorithmic methods. For example, algorithms are considered inhumane and deprive individuals of dignity; algorithmic models are seen as assuming an orderly, rigid, and deterministic world, which is perceived as neither able to handle uncertainty nor capable of learning and improving over time.

Later works began to explore the causes of algorithmic aversion using empirical evidence across various contexts. Through two survey studies, Promberger and Baron (2006) found that individuals, when faced with medical decisions, preferred recommendations from physicians over computer programs. Such preference was attributed to two reasons: First, when patients decide whether or not to have medical procedures, they would not like to carry the responsibility for that decision but rather shift it to someone else, and physicians are believed to be capable of taking the responsibility than computers; Second, patients place greater trust in physicians than computer programs in making a good recommendation. Dietvorst et al. (2014) demonstrated through lab experiments that individuals favored humans over algorithms in forecast-based settings, such as predicting the potential of MBA candidates and making offers. Specifically, individuals were significantly more likely to use humans that produced 13-97% more errors than algorithms after seeing them perform. This was because of a quicker loss of confidence in algorithms than in humans after witnessing the same errors. Castelo et al. (2019) discovered a lower trust in algorithms for subjective (vs. objective) tasks because consumers mistakenly believe that algorithms lack the necessary capabilities to perform subjective tasks. Longoni and Cian (2022) proposed the "word of machine" phenomenon, where the trade-off between utilitarian and hedonic attributes influences the preference for or resistance to AI-based recommendations. The word-of-mouth effect arises from a lay belief that AI recommenders are more competent than human recommenders in the utilitarian realm and less so in the hedonic realm. As a consequence, individuals are more (less) likely to choose AI recommenders when utilitarian (hedonic) attributes are salient. Sun et al. (2022) found that packing workers at the warehouses of the Alibaba Group, one of the world's largest e-commerce companies, deviate from algorithmic prescriptions of the box size for packages, and such deviations are due to both workers having more information and

better solutions than algorithms, and their aversion, inability, or discretion to follow algorithmic prescriptions.

2.2 Algorithm Appreciation

While much of the existing literature suggests aversion towards algorithms, recent studies indicate that preferences for algorithms are possible. Logg et al. (2019) define "algorithm appreciation" as the phenomenon where laypeople are more likely to follow advice believed to be from an algorithm rather than a person. They found consistent evidence of algorithm appreciation across multiple experiments. However, the authors noted that such appreciation diminished when individuals chose between an algorithm's estimate and their own (vs. an external advisor's) or when they had expertise in the task at hand. Further research has explored how algorithm design impacts human attitudes towards algorithms. For instance, Dietvorst et al. (2018) found that individuals prefer algorithms they can modify, Castelo et al. (2019) showed that enhancing algorithms' perceived affective human-likeness helps increase the use of algorithms for subjective tasks, and Clegg et al. (2023) demonstrated that consumers generally favor products with high-adaptivity algorithms because they perceive these algorithms as more creative than low-adaptivity ones. Recent work by Bai et al. (2022) showed that workers perceive the algorithmic (versus human-based) assignment process as fairer even though the underlying rule to assign work is the same, suggesting human preference for the algorithm in their research context.

2.3 Our Contributions

Our paper makes several contributions to the literature. First, the existing literature on human-AI interactions predominantly focuses on the commercial marketplace, where recommendations are typically transactional and driven by consumer preferences. However, there is a noticeable gap in research within the healthcare context, where decisions carry higher stakes and involve greater uncertainty (Longoni et al. 2019). The lack of research on patient-AI interactions is attributed to various factors, such as limited deployment of patient-facing AI applications, challenges in collecting healthcare data, and difficulties in conducting field experiments in real-world clinical settings. Our study addresses this gap by conducting a large-scale field experiment in a hospital system and using the practical gatekeeping context as the test field.

Second, our work bridges the algorithm aversion and algorithm appreciation literature. While variations in research methodologies and contexts contribute to the mixed evidence on human

attitudes towards AI, the fundamental issue lies in understanding what drives AI acceptance. From the information disclosure perspective, we examine the independent and combined effects of algorithmic technical quality and peer-based social opinions, demonstrating that AI aversion can shift toward AI appreciation as algorithm accuracy improves and peer acceptance increases.

Third, our paper contributes to the healthcare gatekeeping literature and practice (Freeman et al. 2017, 2021). By offering insights for the effective deployment of AI-enabled gatekeeping in the healthcare system, our work highlights a cost-effective solution to increase the acceptance rate of AI recommendations and efficiently direct patients to appropriate care based on their symptoms. This approach shows promise in alleviating the burden on PCPs, reducing patient wait times, and lowering unnecessary healthcare expenditures by allowing patients to bypass initial PCP consultations.

3. Hypothesis Development

Our paper examines how the design of information disclosure policies impacts individuals' acceptance of algorithmic recommendations. As we have discussed in the introduction, information asymmetry is especially crucial within the healthcare context because patients frequently lack the expertise to identify the appropriate healthcare services for their symptoms, particularly when specialist care is required (Lu 2012). Consequently, providing information about AI recommendations can potentially influence patients' acceptance decisions. In this section, we develop theoretical predictions regarding the independent and combined effects of *algorithm accuracy* and *peer acceptance* on patients' acceptance of algorithmic recommendations.

3.1. Independent Effect of Technical Information: Algorithm Accuracy

In our setting, algorithm accuracy measures how well the AI-recommended specialty care matches a patient's symptoms, which signals the technical quality of the algorithm. Disclosing algorithm accuracy can affect patient acceptance of AI recommendations through several channels.

First, previous research indicates that the level of machine intelligence, or its capabilities, significantly influences individuals' cognitive and emotional trust in AI (Glikson and Woolley, 2020). Since algorithm accuracy indicates how reliable the gatekeeping algorithm is, higher algorithm accuracy will enhance patients' trust in its recommendations. Moreover, trust is a critical factor in people's interactions with AI-infused systems, with higher trust increasing users' belief that AI will help them achieve their desired goal, thereby making them more likely to accept AI (Kelly et al.,

2023). Thus, with higher algorithm accuracy, we expect patients will be more likely to accept AI recommendations.

Second, according to the technology acceptance model, perceived usefulness significantly contributes to individuals' intentions to adopt technology, ultimately driving their actual use of the technology (Davis 1989). Since algorithm accuracy informs patients of the extent of the match between their symptoms and the AI-recommended specialty care, it will influence the perceptions of the usefulness of the algorithm output and, therefore, affect patients' acceptance of algorithmic recommendations. When the algorithm accuracy level increases, patients will perceive AI recommendations as valid and useful in guiding them to proper care, which will encourage them to adhere to AI recommendations by scheduling appointments with the recommended specialty. Based on these arguments, we propose the following hypothesis for empirical testing:

Hypothesis 1 *Patient acceptance of AI recommendations increases in algorithm accuracy.*

3.2. Independent Effect of Social Information: Peer Acceptance

In our setting, peer acceptance measures the acceptance likelihood of peer patients who received the same specialty recommendation, thereby providing relevant social information. Peer decision is likely an important factor in affecting patient acceptance of AI recommendations due to individuals' general tendency to follow crowd behavior. The seminal work of Banerjee (1992) presents important results on herding in observational learning, where individuals in a sequential decision model, disregard their own information and simply follow the herd. A main cause of herding is that individuals assume other decision-makers possess some information that is important for their decision-making. Prior works have demonstrated herding behavior in various contexts, such as demand herding (Hu et al. 2016, Xu et al. 2021), online review ratings (Lee et al. 2015), product sales (Liu 2006), online auctions (Simonsohn and Ariely 2008), and analysts' earnings forecasts (Clement and Tse 2005).

In the healthcare context, the provision of social information might lead to herding behavior for two main reasons. First, the literature shows that individuals follow others to reduce uncertainty in situations where they lack sufficient information. For instance, Davis et al. (2021) showed that when faced with uncertain service quality from firms, consumers use social information to guide their choices. In the healthcare context, due to the nature of credence services and the lack of medical knowledge, patients may find it challenging to make informed decisions based on their private information. Therefore, disclosing the peer acceptance rate can effectively signal the quality of AI

recommendations, influencing patients' perceptions. Second, Raafat et al. (2009) show that an underlying psychological and neuroscientific basis for herding behavior is risk aversion. Because healthcare decision-making carries higher stakes compared to decisions due to the potential for significant health consequences, patients may be willing to follow the herd as a risk-averse strategy. When the peer acceptance rate is high, individual patients will be more likely to accept AI recommendations, believing that the specialty advice from AI is safe as many people are making the same decision. Based on these considerations, we propose the following hypothesis:

Hypothesis 2 *Patient acceptance of AI recommendations increases in peer acceptance.*

3.3. Combined Effect of Algorithm Accuracy and Peer Acceptance

While both algorithm accuracy and peer acceptance can influence patient decisions, what would happen if they were disclosed together? How will patients weigh the two types of information when deciding whether or not to accept AI recommendations? We develop hypotheses from two aspects, corresponding to the marginal and interaction effects.

We conjecture that peer acceptance has a greater impact on patient acceptance of AI recommendations than algorithm accuracy when both types of information are presented. This is derived from two rationales. First, prior literature shows that human decision-makers often struggle to utilize all available information effectively due to limited attention (Kahneman 1973). Additionally, researchers have demonstrated that more information does not necessarily enhance decision-making due to information overload (Davis et al. 2021, Edmunds and Morris 2000). As such, when disclosing two pieces of information, patients might only be capable of effectively interpreting the information that is easier to digest. During human-AI interactions within the healthcare setting, patients are likely to prioritize social information over the technical accuracy of the algorithm because peer acceptance is more relatable and easier to understand than the technical information about a black-box algorithm. Second, healthcare decisions inherently involve high levels of uncertainty and risk, which may lead individuals to rely more on peer opinions, as they might feel their own knowledge is insufficient for making the right decision. Hence, following the crowd wisdom might be a safer and less risky option for patients. A similar phenomenon regarding patients' use of technical and social information is documented by Li et al. (2023), who showed that patients place greater trust in Yelp ratings on social media compared to government ratings that assess the technical quality of nursing home care. Accordingly, we formulate the following hypothesis:

Hypothesis 3 *When presented together, peer acceptance dominates algorithm accuracy in influencing patient acceptance of AI recommendations.*

We expect that the technical and social information will complement each other in influencing patients' acceptance of AI recommendations, for the following reasons. Suppose algorithm accuracy and peer acceptance rate are consistent and at reasonably high levels, the decision process for the patient is straightforward, and the interaction effect will be positive. Following Hypotheses 1 and 2, when algorithm accuracy or peer acceptance increases, patients will be more likely to trust and adopt AI recommendations due to the perceptions of higher reliability and usefulness. Hence, conditional on a given level of algorithm accuracy, when patients know that many of their peers are accepting gatekeeping decisions from their peers, they will be more willing to adopt AI recommendations because peer acceptance reinforces that the recommendation is likely valid. In the same vein, conditional on a given level of peer acceptance, when patients know that the algorithm predictions have high precision, patients will be more likely to accept AI recommendations because they have greater confidence that the recommended specialty care is likely correct. Based on these considerations, we propose the following hypothesis for an empirical test:

Hypothesis 4 *When presented together, algorithm accuracy and peer acceptance are complementary in influencing patient acceptance of AI recommendations, demonstrating a positive interaction effect.*

4. Research Context and Data

We collaborate with a large teaching hospital in China. As of 2023, the hospital provides pediatric care, with 1,000 inpatient beds and an annual outpatient volume of about seven hundred thousand. The hospital implemented an AI-enabled gatekeeping app, aiming to help patients locate appropriate care more efficiently and to alleviate the burden on frontline healthcare workers responsible for referring patients to various specialties. The AI gatekeeping app guides patients in scheduling medical appointments with appropriate types of specialty, such as dermatology, cardiology, and psychiatry. In particular, a patient can conveniently initiate a conversation with the AI app on a computer or a mobile phone, where patients can provide clinical and demographic information at their fingertips,

and the app generates recommendations for specialty care based on patient input. More information about the app's user interface can be found in Figure 1.[2]

4.1 Experiment Design

Human-AI interactions can be complex and opaque, particularly due to black-box algorithms. Such phenomenon is particularly salient in the healthcare industry due to the nature of credence service, characterized by asymmetric information between experts (i.e., physicians) and consumers (patients) regarding the nature and quality of care. As a result, providing additional information is critical in shaping patients' perceptions of AI and trust in AI-enabled recommendations.

We designed two informational nudges and examined how they influence patients' decisions to accept algorithmic recommendations. The first nudge provides information about *algorithm accuracy*, which measures the algorithmic performance for a specific user session and serves as an indicator of its technical quality. The algorithm accuracy is determined by the AI gatekeeping algorithm and the user input, which is a similarity-based measure that reflects the extent of matching between patient symptoms and the common symptoms of the recommended diagnosis (i.e., specialty). The second nudge provides information about *peer acceptance*, which measures the average acceptance rate among previous patients who received the same specialty recommendation before our experiment, representing group responses to algorithmic recommendations. As depicted in Figure 2, the experiment comprises four treatment arms: patients in Group 1 view recommendations only; patients in Group 2 view recommendations along with algorithm accuracy; patients in Group 3 view recommendations along with peer acceptance rate; and patients in Group 4 view recommendations along with both algorithm accuracy and peer acceptance rate. Examples of the gatekeeping sessions can be found in Figures 3 and 4.

We ran the experiment from March to August 2023, after the app's post-adoption usage had stabilized. By this time, both hospitals and patients had become familiar with the system, and frequent program updates were no longer necessary, allowing us to avoid any contamination from early adoption effects in our experiment. Patients were randomly assigned to one and only one treatment arm, receiving a unique presentation of the recommendation based on patient input and the algorithm prediction. Since each patient receives a random patient identifier (ID) upon registration with the

[2] *Note.* In this study, we do not distinguish between patients and guardians of patients because the latter interact with the gatekeeping app and make all the healthcare decisions on behalf of the former. Therefore, we will refer to users of the app as patients for the rest of the paper.

hospital, which remains unchanged over time, we use the patient ID to facilitate the random assignment of individuals to different treatment arms.[3] The size of patients belonging to different groups is fairly balanced, with 1325 individuals in Group 1, 1316 in Group 2, 1341 in Group 3, and 1372 in Group 4.

4.2 Variable Construction

The outcome variable of interest is a patient's decision, *acceptance*, a binary variable that equals one if a patient follows the recommendation by scheduling an appointment with any physicians in the recommended specialty (i.e., department) and zero otherwise. The key independent variables consist of three measures. The first is a binary variable, *treated*, which equals one for observations in the treated conditions with disclosed information (i.e., Groups 2, 3, or 4) and zero for the baseline condition with recommendation only (i.e., Group 1). The second variable is *Accuracy*, a continuous measure of the algorithm performance in the recommendation, where a higher value suggests higher recommendation accuracy of a gatekeeping session. The third variable is *PeerRate*, a continuous measure of the acceptance of prior patients who received the same recommendation (i.e., the same recommended specialty department).

We also collect various characteristics at the patient and the user-AI session level for robustness checks and heterogeneity analyses, including social demographic information (i.e., patient age, gender, new vs. established patient) and user-AI interactions (i.e., the number of checked symptoms, and whether a patient provides additional information besides the check boxes while interacting with the algorithm). Table 1 presents the summary statistics and definitions of key variables.

4.3 Identification

Our causal identification strategy hinges on the random assignment of patients to different treatment arms, a well-established method in experimental design that mitigates concerns about selection bias. By randomizing patients, we ensure that both observed and unobserved factors that could potentially influence treatment outcomes are distributed evenly across treatment groups. Randomization effectively controls for confounders that may not be directly measurable, such as underlying health

[3] Specifically, the assignment is based on the last two digits of a patient ID: patients whose ID end in two odd digits are assigned to Group 1, patients whose ID end in an odd digit followed by an even digit are assigned to Group 2, patients whose ID end in two even digits are assigned to Group 3, and patients whose IDs end in an even digit followed by an odd digit are assigned to Group 4.

conditions, patient preferences, or other contextual factors that could bias the causal identification of the treatment effect.

In an ideal research design, we would also randomize algorithm accuracy and peer acceptance. However, in a real-world clinical setting, randomizing these two measures among patients is infeasible due to ethical concerns. It is important to note that algorithm accuracy and peer acceptance are inherently influenced by patient symptoms. While the random assignment of patients to treatment arms helps minimize biases that might arise from unmeasured confounders, we also checked the distribution of algorithm accuracy and peer acceptance across the control and treatment groups to address any concerns. As shown in Table 2, both variables are well-balanced across the groups, confirming that the randomization effectively distributed patients with varying conditions into different groups. Therefore, any differences in outcomes are likely due to the treatment effects rather than unobserved imbalances related to these measures or patient conditions.

Besides the algorithm accuracy and peer acceptance measures, we also observe balanced patient characteristics, including patient demographics, past experience with the hospital, and characteristics of user-AI sessions, between the treatment groups (i.e., groups 2, 3, and 4) and the control group (i.e., group 1), as evidenced by the fact that all between-group standardized differences are below the threshold of 0.10, which is recommended by the literature for evaluating between-group covariate balance (Austin 2009) (see Panel A of Table 2). In addition, we carry out a between-group *t*-test for differences in means and report the result in Panel B of Table 2 and find that all p-values are above 0.1. Therefore, both balance checks suggest the proper random assignment process in the experiment and lay the foundation for the causal inference in our analyses.

5 Main Results

5.1 Independent Effects of Algorithm Accuracy and Peer Acceptance

To estimate the treatment effects of algorithm accuracy and peer effect on individual patients' decisions, we apply the standard linear probability model. Our unit of observation is at the user session level, and the main estimation equations are specified as follows:

$$Y_i = \alpha + \beta_1 treated_i + \beta_2 treated_i * Accuracy_i + \gamma X_i + \varepsilon_i \,, \qquad (1)$$

$$Y_i = \alpha + \beta_1 treated_i + \beta_2 treated_i * PeerRate_i + \gamma X_i + \varepsilon_i \,, \qquad (2)$$

where Y_i represents a patient's acceptance of the recommendation in session i, which is a binary variable that equals one if the patient schedules an appointment with a doctor in the specialty

department recommended by the gatekeeping app and zero otherwise; $treated_i$ is a binary variable which equals one for observations in treatment groups 2 or 3, and zero for observations in group 1; $Accuracy_i$ and $PeerRate_i$ are continuous measures representing algorithm performance and peer acceptance, respectively; and X_i include various controls for patients and user sessions. The sum of β_1 and $\beta_2 * Accuracy_i$ (or $\beta_2 * PeerRate_i$) captures the treatment effect of the information disclosure, conditional on a given level of algorithm accuracy (or peer acceptance rate).

We estimate equation (1) using observations from Groups 1 and 2, where patients in Group 1 served as the control group, receiving only AI recommendations without any additional information. In contrast, patients in Group 2 received recommendations but were also informed about the algorithm accuracy for the specific recommendation. By analyzing the responses from both groups, we aimed to gain insights into whether disclosing information about algorithm accuracy influenced patients' willingness to accept AI recommendations. Similarly, we estimate equation (2) using observations from Groups 1 and 3, which allows us to examine the effect of incorporating information about peer acceptance on patients' decisions.

We report the regression results in columns (1) and (2) of Table 3. We find significantly negative coefficient estimates for the variable *treated* and significantly positive estimates for the interaction terms *treated* * *Accuracy* and *treated* * *PeerRate*. These results suggest that when algorithm accuracy or peer acceptance is low, patients tend to reject AI recommendations. On the other hand, when *Accuracy* or *PeerRate* increases, patients' likelihood to accept AI recommendations increases. The findings align with the literature. For instance, Dietvorst et al. (2018) suggest that algorithm aversion arises after people learn that algorithms are imperfect. With relatively low algorithm accuracy and peer acceptance, information about the imperfection of the algorithms raises patients' awareness of the algorithm's limitations, which leads to disappointment or skepticism and reduces their willingness to accept AI advice.

To further enhance the precision of the estimate, we additionally control for patient and session-level characteristics and report the results in columns (3) and (4) of Table 3. We find robust estimates similar to the baseline model, which further demonstrates the success of the random assignment and the causal identification of the treatment effects. For easier interpretations of the economic magnitude at different levels of *Accuracy* and *PeerRate*, we visualize the treatment effects. In Figure 5, the *x*-axis represents the accuracy level of the AI recommendation, ranging from 0% to 100%. The *y*-axis denotes the acceptance rate of the AI recommendation by patients at a given level of *Accuracy*,

calculated as the *treated* + *treated* * *Accuracy* from the coefficient estimates in column (3) of Table 3. As shown from the graph, after disclosing the prediction accuracy, there was a general tendency among patients to reject AI recommendations, indicating a lack of trust in AI. However, as depicted by the upward slope in the graph, the acceptance of AI recommendations increases as the accuracy level rises. Specifically, positive treatment effects began to appear when the accuracy level surpassed 77.1%. When the accuracy reaches 85%, the treatment effect becomes significantly positive, with the treatment effect being 0.040 at the 5% significance level.

In Figure 6, we plot the peer acceptance rate from the range of 0% to 100% on the *x*-axis and the corresponding acceptance rate of AI recommendations by patients on the *y*-axis (i.e., *treated* + *treated* * *PeerRate*). Similar to the effect of disclosing algorithm accuracy, we observe an upward trend as the peer acceptance rate increases. The reflection point for *PeerRate* is 44.6%, below which the overall treatment effect is negative. When *PeerRate* reaches 49%, the treatment effect becomes positively significant, with the treatment effect being 0.038 at the 5% significance level.

In summary, our results demonstrate the effectiveness of information disclosure proposed in the experiment. Our findings suggest that higher algorithm accuracy and peer acceptance increase a focal patient's likelihood of adopting algorithmic recommendations. Intuitively, accuracy signals the algorithm's technical quality, and the peer acceptance rate reflects the social judgment of the AI app. Hence, when algorithm accuracy and peer acceptance are higher, patients are more likely to perceive AI recommendations as reliable, leading to increased acceptance.

5.2 Combined Effect of Algorithm Accuracy and Peer Acceptance

We are interested in patients' responses when presented with both technical and social information. Specifically, we aim to understand how patients process these two types of information. To achieve this, we adopt the following specification and conduct analyses using observations from Groups 1 and 4:

$$Y_i = \alpha + \beta_1 treated_i + \beta_2 treated_i * Accuracy_i + \beta_3 treated_i * PeerRate_i + \beta_4 treated_i * Accuracy_i * PeerRate_i + \gamma X_i + \varepsilon_i , \qquad (3)$$

where $treated_i$ is a binary variable which equals one for observations in group 4 and zero for observations in group 1; the other variables remain the same as in equations (1) and (2). In this new specification, β_2 and β_3 capture the *marginal effect* of algorithm accuracy and peer acceptance on patient acceptance of AI recommendations. The coefficient β_4 is the cross-partial derivative of Y with respect to *Accuracy* and *PeerRate*, which represents the *interaction effect* between the two

independent variables. Following the common notion of complements and substitutes in Economics, the interaction effect is considered complementary (or substitutive) when $\beta_4 > 0$ (or $\beta_4 < 0$).

We estimate equation (3) and report the results in Table 4. As shown in column (1), the coefficient of *treated* is consistent with the previous section, where the negatively significant coefficient suggests that when $Accuracy$ and $PeerRate$ levels are minimal (i.e., zero), disclosing information will lead to rejections of AI recommendations. Very interestingly, we find that the presence of peer acceptance diminishes the marginal effect of algorithm accuracy. As evidenced by the coefficient of $treated * Accuracy$, while the coefficient remains positive, the magnitude is small and its effect becomes insignificant. By contrast, the baseline effect of peer acceptance remains positive and significant, as suggested by the positive and significant coefficient of $treated * PeerRate$. Based on the F-test of the marginal effects of these two factors, we find a significant difference in magnitude, suggesting that social information from peers indeed overweighs the effect of the algorithm performance when both are presented to patients. These findings thus support Hypothesis 3.

The coefficient of $treated * Accuracy * PeerRate$ is significantly positive, suggesting a positive interaction effect. In other words, when $PeerRate$ increases, the partial effect of $Accuracy$ on AI acceptance becomes stronger, or vice versa. The results thus support Hypothesis 4, implying that the social and technical information complements each other in influencing patients' acceptance of AI recommendations. The findings make sense because an increase in one measure boosts the effect of the other by enhancing individuals' trust and perceived quality of the AI recommendation. When additional controls for user sessions are included, we find consistent results, as shown in column (2) of Table 4, providing further support to Hypotheses 3 and 4.

Given the Table 4 result, we can calculate the treatment effect with any combination of algorithm accuracy and peer acceptance. For illustration purposes, we discuss two examples using coefficient estimates in column (2) of Table 4. When both signals are at low levels (i.e., $Accuracy = 20\%$ and $PeerRate = 20\%$), the overall treatment effect is -0.353 (= -0.507 + 0.071*20% + 0.585*20% + 0.558*20%*20%) and significant at 1%, suggesting a significantly lower likelihood of algorithm acceptance. Following the same calculation, when we increase accuracy to 80% and keep the peer acceptance rate at 20%, the overall treatment effect is still significantly negative, with the treatment effect being -0.244 ($p < 0.01$). The results reinforce the dominating effect of social influence in the presence of both information in Hypothesis 3, suggesting that individuals are more subject to the influence of peer patients than algorithm accuracy, especially when viewing misaligned signals.

5.3 Alternative Specification: Categorical Independent Variables

To effectively demonstrate the combined treatment effect when presenting both information to patients, we next conduct the following analysis, in which we assign each observation into a specific category by binning the continuous measures of algorithm accuracy and peer acceptance into three groups. This exercise allows us to examine the overall treatment effect for different combinations of the two independent variables.

$$Y_i = \alpha_0 + \sum_{j\in\{Low,Medium,High\},k\in\{Low,Medium,High\}} \beta_{jk} Grp_{i,\{jk\}} + \gamma X_i + \varepsilon_i \,, \qquad (4)$$

where Y_i represents a patient's acceptance of an AI recommendation in a user session i and X_i $includes$ various controls for patients and user sessions. For each user session, we discretize the continuous measures, $PeerRate$ and $Accuracy$, into low, medium, or high group based on the cutoff value at 30% and 70% percentiles of the corresponding variable. As such, for a specific user session i, depending on the algorithm accuracy and peer acceptance levels, a session uniquely belongs to a group $Grp_{i,\{jk\}}$, and the corresponding coefficient β_{jk} captures the combined treatment effect for the corresponding levels of accuracy and peer acceptance rate. For instance, the coefficient $\beta_{j=Low\ PeerRate,k=High\ Accuracy}$ indicates the treatment effect when a user session is associated with low peer acceptance and high algorithm accuracy.

Columns (1) and (2) of Table 5 report the estimates of β_{jk} following equation (4). The coefficient estimates of the first three rows are all significantly negative, suggesting that individuals resist accepting AI recommendations when the peer acceptance rate is low (i.e., Low *PeerRate* group), regardless of the algorithm's accuracy level. This underscores the significance of peer effect in shaping patient decisions, wherein a low peer acceptance rate diminishes acceptance of AI recommendations. By contrast, when peer acceptance is high (i.e., High *PeerRate* group), all treatment estimates are significantly positive. The results indicate that patients consistently lean towards accepting AI recommendations, validating the dominant impact of peer influence.

Nonetheless, algorithm accuracy still matters, as we can see a slightly increasing magnitude of the treatment effect when the accuracy level increases. This echoes the notion that the technical quality signal (i.e., algorithm accuracy), enhances acceptance of AI recommendations, particularly when aligned with high peer acceptance rates. More interestingly, when we move to the case when signals from peer patients are ambiguous (i.e., Medium *PeerRate* group), individuals exhibit a reliance on the technical information of the algorithm. Only recommendations with high accuracy led

to acceptance, highlighting the pivotal role of algorithmic precision in guiding decision-making when social cues are ambiguous.

To check the robustness of our results to different cutoff values, we then discretize the continuous measures, $Accuracy_i$ and $PeerRate_k$, into low, medium, or high group based on the cutoff value at 20% and 80% percentiles for the continuous variables $Accuracy$ and $PeerRate$. Such grouping leads to smaller sizes for the low and high groups and more observations in the medium group. Nonetheless, as shown in columns (3) and (4) of Table 5, we find consistent and robust results.

6 Analyses of the Moderating Effects

In this section, we examine how individual heterogeneity influences responses to AI recommendations, which helps to understand the mechanisms underlying the effect of information disclosure about the AI gatekeeping app. Given that patients often lack priors about the care they need, we disclose information that is intended to update patients' beliefs and help them make informed decisions. If this mechanism indeed exists, we will anticipate a larger treatment effect on patients who lack priors (i.e., are more uncertain) about the type of services they need, because these patients will rely more on the additional information before making decisions. To test this mechanism, we develop two empirical tests that leverage variations in patients' understanding of a specific medical process.

6.1 The Moderating Effect of Patient Health Literacy

In the first test, we categorize patients into two groups based on whether they provide customized input when communicating their symptoms to the AI app by leveraging a nuanced feature in our setting. Specifically, patients were provided with two options for self-description when interacting with the AI app during a user session: (1) check boxes for patients to indicate symptoms preprogrammed by the app and (2) a text box for patients to provide additional health information. The former is mandatory, while the latter is optional. We hypothesize that patients who can articulate their health conditions using the text box have better health literacy than those who only use the check box to indicate preprogrammed symptoms. Consequently, they may have a better idea of the type of care they need and may be less influenced by the information disclosed about the AI app. Following this logic, we perform the empirical test using subgroups of patients *with* and *without* customized input via the text box.

The results are reported in Table 6, with columns (1) and (3) corresponding to patients *without* additional input via the text box, and columns (2) and (4) corresponding to patients *with* additional

input via the text box (i.e., with better health literacy). From columns (1) and (2), we can see that the effect of algorithm accuracy for patients lacking customized input almost doubles the effect of patients with customized input. From columns (3) and (4), we observe that the effect of peer acceptance is also larger in magnitude for patients without customized input. In terms of magnitude, a 10% increase in algorithm accuracy led to an 8% increase in recommendation acceptance for patients *without* customized inputs, whereas it led to only a 4% increase for patients *with* customized inputs. Similarly, a 10% increase in peer acceptance resulted in an 11% increase in recommendation acceptance for patients *without* customized inputs, compared to an 8% increase for patients *with* customized inputs. Overall, the empirical results are consistent with our intuition—patients who did not articulate their health conditions using the text box exhibited a greater reliance on AI accuracy and peer acceptance information than those who could describe their symptoms using the text box, likely due to a lack of health literacy.

6.2 The Moderating Effect of Patient Experience

In the second test, we differentiate between new and existing patients.[4] Our rationale is that existing patients, being more familiar with their health conditions and the hospital, may have better prior knowledge about the care they need than new patients. Accordingly, new patients will be more influenced by the information disclosure about the gatekeeping app than existing patients. Similar to the first test, we performed subgroup analysis using the sample of new and existing patients.

We report the results in Table 7. As shown in columns (1) and (2), the treatment effect of algorithm accuracy is much larger in magnitude on new patients than on existing patients, with a 10% increase in algorithm accuracy leading to a 6% increase for new patients but only a 3% increase for existing patients. This suggests that new patients more heavily rely on algorithmic performance information to make decisions. This can be attributed to the limited information that new patients possess regarding the hospital's healthcare delivery processes and the quality of services provided. In the absence of prior experience with the hospital, new patients place greater importance on technical quality indicators such as algorithm accuracy when making healthcare decisions. Conversely, existing patients who are already familiar with the hospital system rely less on algorithm accuracy.

Interestingly, despite the drastic heterogeneity in patients' reliance on accuracy level, both new and existing patients demonstrate similar herding effects (see columns 3 and 4 of Table 7), suggesting

[4] We define new patient as someone who had not visited the hospital within a year before the start of our sample window (i.e., between March 2022 and February 2023).

that peer opinions exert a comparable influence on both groups. This finding echoes the result from the combined effect, where peer acceptance is a powerful social signal that affects patient decision-making more than algorithm accuracy.

7 Conclusion

Researchers have investigated how technology has revolutionized various aspects of healthcare delivery, such as improving access to care (Delana et al. 2023), enhancing delivery efficiency (Sun et al. 2020), and improving service quality (Li et al. 2022). With the fast-evolving AI-based technology applications, our work is among the first to investigate patient-facing AI applications in real-world clinical settings, offering valuable managerial insights for their implementation.

In this study, we conducted a large-scale field experiment in a hospital to examine factors that shape patient acceptance of algorithmic recommendations. Our findings reveal that individuals' acceptance of AI recommendations increases with both the accuracy of the algorithm and the level of peer acceptance. Notably, we identify an interesting pattern in which patients rely more on peer choices than on algorithm accuracy in the presence of both information. A strong social signal from peer acceptance often dictates a patient's decision to accept AI recommendations, regardless of the technical performance of the algorithm. We further empirically demonstrate the moderating role of patient characteristics. Specifically, the treatment effect is more pronounced on patients with limited prior knowledge about healthcare services—such as those who struggle to articulate their symptoms or new patients of the hospital.

Our paper provides timely insights into human-AI interactions, bridging the gap between algorithm aversion and algorithm appreciation in the existing literature. We show that AI aversion and AI appreciation are not fixed opposing perspectives, but rather stages in an evolving process that can be influenced and reconciled through changes in algorithm accuracy and peer acceptance. Traditional views in the literature suggest a sharp divide between those who distrust AI and those who embrace it. However, our study suggests these attitudes are more fluid and can shift over time. Our research shows that human trust in AI is not solely based on an algorithm's performance but is heavily shaped by social validation. As individuals observe more people around them accepting and benefiting from AI recommendations, they become more likely to follow suit, even if they initially had doubts. Moreover, as individuals become more familiar with AI applications and experience positive outcomes, their confidence in AI increases. This growing familiarity reduces uncertainty and doubt, gradually shifting their perspectives from aversion to appreciation. Peer influence can amplify

this process by reinforcing positive experiences and normalizing AI usage in decision-making contexts. Ultimately, as AI technologies continue to improve and people become more accustomed to relying on them, the collective social acceptance of AI will likely grow. The transition from AI aversion to AI appreciation is thus not just a technological matter but a social one, driven by both the evolving capabilities of AI and the influence of peer behaviors. Over time, this shift may become inevitable, as AI becomes more deeply integrated into everyday decision-making processes.

Our study further offers practical insights for healthcare decision-makers and the broader audience. We highlight AI gatekeeping as a promising application to improve referral efficiency, allowing patients to receive appropriate specialty care more quickly than through traditional PCPs. When algorithms have superior performance and high acceptance of AI recommendations is desired, strategic information disclosure will increase human acceptance of AI. Additionally, practitioners should account for user heterogeneity when designing AI applications. More broadly, organizations deploying AI should recognize the value of disclosing both technical and social information to support informed decision-making, especially in settings where information asymmetry is high. This asymmetry often stems from the black-box nature of algorithms and the public's lack of domain-specific knowledge in settings like healthcare.

Reference

Aluttis, C., Bishaw, T., & Frank, M. W. (2014). The workforce for health in a globalized context – global shortages and international migration. *Global Health Action*, *7*(1), 23611.

Austin, P. C. (2009). Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Statistics in medicine*, 28(25), 3083-3107.

Bai, B., Dai, H., Zhang, D. J., Zhang, F., & Hu, H. (2022). The impacts of algorithmic work assignment on fairness perceptions and productivity: evidence from field experiments. *Manufacturing & Service Operations Management*, *24*(6), 3060–3078.

Banerjee, A. V. (1992). A simple model of herd behavior. *The Quarterly Journal of Economics*, *107*(3), 797–817.

Brenner, L., & Meyll, T. (2020). Robo-advisors: A substitute for human financial advice? *Journal of Behavioral and Experimental Finance*, *25*, 100275.

Castelo, N., Bos, M. W., & Lehmann, D. R. (2019). Task-dependent algorithm aversion. *Journal of Marketing Research*, *56*(5), 809–825.

Clegg, M., Hofstetter, R., de Bellis, E., & Schmitt, B. H. (2023). Unveiling the mind of the machine. *Journal of Consumer Research*, ucad075.

Clement, M. B., & Tse, S. Y. (2005). Financial analyst characteristics and herding behavior in forecasting. *The Journal of Finance*, *60*(1), 307–341.

Davis, A. M., Gaur, V., & Kim, D. (2021). Consumer learning from own experience and social information: an experimental study. *Management Science*, *67*(5), 2924–2943.
Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, 319–340.
Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. In *Rationality and Social Responsibility* (pp. 571–582). Psychology Press.
Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science, New Series*, *243*(4899), 1668–1674.
Delana, K., Deo, S., Ramdas, K., Subburaman, G.-B. B., & Ravilla, T. (2023). Multichannel delivery in healthcare: the impact of telemedicine centers in southern India. *Management Science*, *69*(5), 2568–2586.
Dietvorst, B. J., Simmons, J. P., & Massey, C. (2014). Algorithm aversion: people erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology*, *144*(1).
Dietvorst, B. J., Simmons, J. P., & Massey, C. (2018). Overcoming algorithm aversion: people will use imperfect algorithms if they can (even slightly) modify them. *Management Science*, *64*(3), 1155–1170.
Edmunds, A., & Morris, A. (2000). The problem of information overload in business organisations: A review of the literature. *International Journal of Information Management*, *20*(1), 17–28.
Freeman, M., Robinson, S., & Scholtes, S. (2021). Gatekeeping, fast and slow: an empirical study of referral errors in the emergency department. *Management Science*, *67*(7), 4209–4232.
Freeman, M., Savva, N., & Scholtes, S. (2017). Gatekeepers at work: an empirical analysis of a maternity unit. *Management Science*, *63*(10), 3147–3167.
Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: review of empirical research. *Academy of Management Annals*, *14*(2), 627–660.
Gomez-Uribe, C. A., & Hunt, N. (2016). The netflix recommender system: algorithms, business value, and innovation. *ACM Transactions on Management Information Systems*, *6*(4), 1–19.
Green, L. V., Savin, S., & Lu, Y. (2013). Green, L. V., Savin, S., & Lu, Y. (2013). Primary care physician shortages could be eliminated through use of teams, nonphysicians, and electronic communication. *Health Affairs*, *32*(1), 11–19.
Grove, W. M., & Meehl, P. E. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical–statistical controversy. *Psychology, Public Policy, and Law*, *2*(2), 293.
Grumbach, K., Selby, J. V., Damberg, C., Bindman, A. B., Quesenberry Jr, C., Truman, A., & Uratsu, C. (1999). Resolving the gatekeeper conundrum: What patients value in primary care and referrals to specialists. *JAMA*, *282*(3), 261–266.
Hu, M., Milner, J., & Wu, J. (2016). Liking and following and the newsvendor: Operations and marketing policies under social influence. *Management Science*, *62*(3), 867–879.
Jin, G. Z., Luca, M., & Martin, D. (2021). Is no news (perceived as) bad news? An experimental investigation of information disclosure. *American Economic Journal: Microeconomics*, *13*(2), 141–173.
Kahneman, D. (1973). *Attention and effort* (Vol. 1063).
Kelly, S., Kaye, S.-A., & Oviedo-Trespalacios, O. (2023). What factors contribute to the acceptance of artificial intelligence? A systematic review. *Telematics and Informatics*, *77*, 101925.
Kim, Y., Mitrofanov, D., & Xu, Y., AI and worker learning: evidence from a large-scale field experiment (September 19, 2024). SSRN: https://papers.ssrn.com/abstract_id=4962212

Lee, Y.-J., Hosanagar, K., & Tan, Y. (2015). Do I Follow Do I follow my friends or the crowd? Information cascades in online movie ratings. *Management Science*, *61*(9), 2241–2258.

Li, Y., Lu, L. X., & Lu, S. F. (2023). Does social media dominate government report cards in influencing nursing home demand?. *Manufacturing & Service Operations Management*, 25(6), 2085-2105.

Li, Y., Lu, L. X., Lu, S. F., & Chen, J. (2022). The value of health information technology interoperability: Evidence from interhospital transfer of heart attack patients. *Manufacturing & Service Operations Management*, *24*(2), 827–845.

Linden, G., Smith, B., & York, J. (2003). Amazon. com recommendations: Item-to-item collaborative filtering. *IEEE Internet Computing*, *7*(1), 76–80.

Liu, Y. (2006). Word of Mouth for Movies: Its dynamics and impact on box office revenue. *Journal of Marketing*, *70*(3), 74–89.

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, *151*, 90–103.

Longoni, C., Bonezzi, A., & Morewedge, C. K. (2019). Resistance to medical artificial intelligence. *Journal of Consumer Research*, *46*(4), 629–650.

Longoni, C., & Cian, L. (2022). Artificial intelligence in utilitarian vs. hedonic contexts: The "word-of-machine" effect. *Journal of Marketing*, *86*(1), 91–108.

Lu, S.F. (2012). Multitasking, information disclosure, and product quality: Evidence from nursing homes. *Journal of Economics & Management Strategy*, 21(3), 673-705.

Lu, S.F. (2016). The role of donations in quality disclosure: evidence from nonprofit nursing homes. *American Journal of Health Economics*, 2(4), 431-462.

Lu, S.F. (2023). Comment on Chapters 1 and 2: building physician trust in artificial intelligence. *In The Economics of Artificial Intelligence: Health Care Challenges*, 139-142. National Bureau of Economic Research, Inc.

Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence.*

Osborn, R., Moulds, D., Schneider, E. C., Doty, M. M., Squires, D., & Sarnak, D. O. (2015). Primary care physicians in ten countries report challenges caring for patients with complex health needs. *Health Affairs*, *34*(12), 2104–2112.

Promberger, M., & Baron, J. (2006). Do patients trust computers? *Journal of Behavioral Decision Making*, *19*(5), 455–468.

Raafat, R. M., Chater, N., & Frith, C. (2009). Herding in humans. *Trends in Cognitive Sciences*, *13*(10), 420–428.

Robertson, C., Woods, A., Bergstrand, K., Findley, J., Balser, C., & Slepian, M. J. (2023). Diverse patients' attitudes towards Artificial Intelligence (AI) in diagnosis. PLOS Digital Health, 2(5), e0000237.

Schoen, C., Osborn, R., Huynh, P. T., Doty, M., Peugh, J., & Zapert, K. (2006). On The Front Lines Of Care: Primary Care Doctors' Office Systems, Experiences, And Views In Seven Countries: Country variations in primary care practices indicate opportunities to learn to improve outcomes and efficiency. *Health Affairs*, *25*(Suppl1), W555–W571.

Simonsohn, U., & Ariely, D. (2008). When rational sellers face nonrational buyers: evidence from herding on eBay. *Management Science*, *54*(9), 1624–1637.

Sripa, P., Hayhoe, B., Garg, P., Majeed, A., & Greenfield, G. (2019). Impact of GP gatekeeping on quality of care, and health outcomes, use, and expenditure: A systematic review. *British Journal of General Practice*, *69*(682), e294–e303.

Sun, J., Zhang, D. J., Hu, H., & Van Mieghem, J. A. (2022). Predicting human discretion to adjust algorithmic prescription: A large-scale field experiment in warehouse operations. *Management Science*, *68*(2), 846–865.

Sun, S., Lu, S. F., & Rui, H. (2020). Does telemedicine reduce emergency room congestion? Evidence from New York state. *Information Systems Research*, *31*(3), 972–986.

Xu, Y., Armony, M., & Ghose, A. (2021). The interplay between online reviews and physician demand: An empirical investigation. *Management Science*, *67*(12), 7344–7361.

Table 1. Summary Statistics

Variable	Mean	S.D.	Definition
Outcome Variable			
Acceptance	0.473	0.499	Binary variable indicating whether a patient accepts AI recommendation
Treatment Variable			
Accuracy	0.761	0.193	Continuous measure of the algorithm performance in a recommendation
PeerRate	0.438	0.233	Continuous measure of the average acceptance rete by prior patients who received the same recommendation
Key Controls			
patAge	65.207	47.005	Age of the pediatric patient involved (measured in months)
patGender	0.410	0.492	Binary variable indicating whether a patient is female
isNewPatient	0.641	0.480	Binary variable indicating whether a patient is a new patient
isPatientInput	0.715	0.451	Binary variable indicating whether a patient has additional input via the chat box during a gatekeeping session
numCheckbox	2.516	1.722	Continuous measure of the number of checkboxes selected by a patient during a gatekeeping session

Note. This table reports the definitions and summary statistics for the key variables in the paper. S.D. stands for standard deviation.

Table 2. Balance Check

Panel A Between-group Standard Difference

	Group 1		Group 2		Std. Diff.	Group 3		Std. Diff.	Group 4		Std. Diff.
	Mean	S.D.	Mean	S.D.		Mean	S.D.		Mean	S.D.	
Accuracy	0.760	0.193	0.754	0.193	0.033	0.761	0.194	-0.001	0.769	0.192	-0.046
PeerRate	0.436	0.233	0.440	0.234	-0.020	0.443	0.229	-0.030	0.432	0.235	0.015
patAge	65.550	45.187	65.800	47.457	-0.005	63.130	48.008	0.052	66.340	47.287	-0.017
patGender	0.407	0.491	0.409	0.492	-0.004	0.407	0.491	-0.001	0.418	0.493	-0.022
isNewPatient	0.648	0.478	0.643	0.479	0.011	0.635	0.482	0.027	0.639	0.480	0.019
isPatientInput	0.705	0.456	0.725	0.447	-0.044	0.706	0.456	-0.003	0.724	0.447	-0.042
numCheckbox	2.549	1.792	2.546	1.678	0.002	2.517	1.742	0.018	2.455	1.677	0.054
Sample Size (N)	**1325**		**1316**			**1341**			**1372**		

Note. In this table, we calculate between-group standardized differences in means to compare the balance of covariates between each treatment condition (i.e., Groups 2, 3, and 4) and the control condition (i.e., Group 1). S.D. stands for standard deviation. Std. Diff. stands for the standardized difference and a value greater than 0.1 (in absolute terms) is considered to indicate a meaningful imbalance.

Panel B Between-group T-test

	Group 1	Group 2	p-value	Group 3	t test	Group 4	p-value
Accuracy	0.760	0.754	0.396	0.760	0.978	0.769	0.231
	(0.193)	(0.193)		(0.194)		(0.192)	
PeerRate	0.436	0.440	0.616	0.443	0.431	0.432	0.697
	(0.233)	(0.234)		(0.229)		(0.235)	
patAge	65.551	65.803	0.889	63.126	0.180	66.338	0.659
	(45.187)	(47.457)		(48.008)		(47.287)	
patGender	0.407	0.409	0.916	0.407	0.985	0.418	0.568
	(0.491)	(0.492)		(0.491)		(0.493)	
isNewPatient	0.648	0.643	0.770	0.635	0.486	0.639	0.622
	(0.478)	(0.479)		(0.482)		(0.48)	
isPatientInput	0.705	0.725	0.255	0.706	0.942	0.724	0.279
	(0.456)	(0.447)		(0.456)		(0.447)	
numCheckbox	1.456	1.390	0.964	1.420	0.641	1.401	0.160
	(0.951)	(0.858)		(0.896)		(0.832)	

Note. In this table, we perform the pairwise two-sample t-test and report p-values to determine whether the difference between the means of each treatment condition (i.e., Groups 2, 3, and 4) are statistically significant from the control condition (i.e., Group 1). A p-value smaller than 0.1 is considered to indicate a meaningful imbalance.

Table 3. Independent Effect

Dependent Variables	Acceptance			
Samples	group 1&2 NoInfo & AccInfo	group 1&3 NoInfo & PeerInfo	group 1&2 NoInfo & AccInfo	group 1&3 NoInfo & PeerInfo
	(1)	(2)	(3)	(4)
treated	-0.403***	-0.398***	-0.388***	-0.386***
	(0.057)	(0.031)	(0.057)	(0.031)
treated*Accuracy	0.524***		0.503***	
	(0.071)		(0.071)	
treated*PeerRate		0.890***		0.866***
		(0.057)		(0.056)
Patient and Session Controls	N	N	Y	Y
R-squared	0.020	0.084	0.047	0.121
Observations	2641	2666	2641	2666

Note. *** p<0.01, ** p<0.05, * p<0.1. In the analyses, *treated* is defined as zero for observations in group 1(recommendation only) and one for observations in group 2 (recommendation plus information on algorighm accuracy) and group 3 (recommendation plus information on peer acceptance rate).

Table 4 Combined Effect

Dependent Variables	Acceptance	
Samples	group 1&4 NoInfo & BothInfo	
	(1)	(2)
treated	-0.485***	-0.507***
	(0.097)	(0.095)
treated*Accuracy	0.053	0.071
	(0.124)	(0.122)
treated*PeerRate	0.430**	0.585***
	(0.211)	(0.208)
treated*Accuracy*PeerRate	0.734***	0.558**
	(0.269)	(0.266)
Patient and Session Controls	N	Y
F-test of the coefficients for treated*Accuracy and treated*PeerRate		
F-value	8.10	14.97
R-squared	0.124	0.158
Observations	2697	2697

Note. *** p<0.01, ** p<0.05, * p<0.1. In the analyses, *treated* is defined as zero for observations in group 1(recommendation only) and one for observations in group 4 (recommendation plus the information on algorighm accuracy and peer acceptance rate).

Table 5. Combined Effect by Levels of Disclosed Information

Dependent Variables		Acceptance			
Samples		group 1&4 No Info & Both Info 30% & 70% cutoffs		group 1&4 No Info & Both Info 20% & 80% cutoffs	
		(1)	(2)	(3)	(4)
Low PeerRate	Low Accuracy	-0.361***	-0.369***	-0.360***	-0.386***
		(0.037)	(0.036)	(0.056)	(0.055)
	Medium Accuracy	-0.367***	-0.367***	-0.441***	-0.446***
		(0.041)	(0.040)	(0.043)	(0.043)
	High Accuracy	-0.297***	-0.307***	-0.133**	-0.141**
		(0.040)	(0.040)	(0.065)	(0.065)
Medium PeerRate	Low Accuracy	0.091	0.093	-0.062	-0.052
		(0.059)	(0.058)	(0.045)	(0.045)
	Medium Accuracy	-0.033	-0.017	-0.048*	-0.040
		(0.040)	(0.040)	(0.025)	(0.025)
	High Accuracy	0.292***	0.276***	0.251***	0.241***
		(0.042)	(0.042)	(0.037)	(0.037)
High PeerRate	Low Accuracy	0.126***	0.143***	0.183***	0.216***
		(0.040)	(0.040)	(0.062)	(0.061)
	Medium Accuracy	0.135***	0.134***	0.161***	0.153***
		(0.032)	(0.032)	(0.037)	(0.036)
	High Accuracy	0.329***	0.328***	0.355***	0.338***
		(0.039)	(0.038)	(0.066)	(0.065)
Patient and Session Controls		N	Y	N	Y
	R-squared	0.137	0.168	0.095	0.129
	Observations	2697	2697	2697	2697

Note. *** p<0.01, ** p<0.05, * p<0.1. In the analyses, *treated* is defined as zero for observations in group 1(recommendation only) and one for observations in group 4 (recommendation plus the information on algorighm accuracy and peer acceptance rate).

Table 6. Heterogeneity by Patient Health Literacy

Dependent Variables	Acceptance			
Samples	group 1&2 NoInfo & AccInfo		group 1&3 NoInfo & PeerInfo	
	without patient input	with patient input	without patient input	with patient input
	(1)	(2)	(3)	(4)
treated	-0.635***	-0.310***	-0.538***	-0.348***
	(0.116)	(0.066)	(0.067)	(0.035)
treated*Accuracy	0.806***	0.412***		
	(0.148)	(0.082)		
treated*PeerRate			1.076***	0.821***
			(0.119)	(0.064)
Patient and Session Controls	Y	Y	Y	Y
R-squared	0.069	0.036	0.141	0.110
Observations	753	1888	785	1881

Note. *** p<0.01, ** p<0.05, * p<0.1. In the analyses, treated is defined as zero for observations in group 1(recommendation only) and one for observations in group 2 (recommendation plus information on algorighm accuracy) and group 3 (recommendation plus information on peer acceptance rate).

Table 7. Heterogeneity by New vs. Existing Patient

Dependent Variables	Acceptance			
Samples	group 1&2 NoInfo & AccInfo		group 1&3 NoInfo & PeerInfo	
	New Patient	Existing Patient	New Patient	Existing Patient
	(1)	(2)	(3)	(4)
treated	-0.490***	-0.210**	-0.408***	-0.353***
	(0.071)	(0.096)	(0.038)	(0.053)
treated*Accuracy	0.610***	0.314***		
	(0.088)	(0.122)		
treated*PeerRate			0.887***	0.829***
			(0.069)	(0.096)
Patient and Session Controls	Y	Y	Y	Y
R-squared	0.060	0.033	0.129	0.120
Observations	1705	936	1711	955

Note. *** p<0.01, ** p<0.05, * p<0.1. In the analyses, treated is defined as zero for observations in group 1(recommendation only) and one for observations in group 2 (recommendation plus information on algorighm accuracy) and group 3 (recommendation plus information on peer acceptance rate).

Figure 1. User Interface of the APP

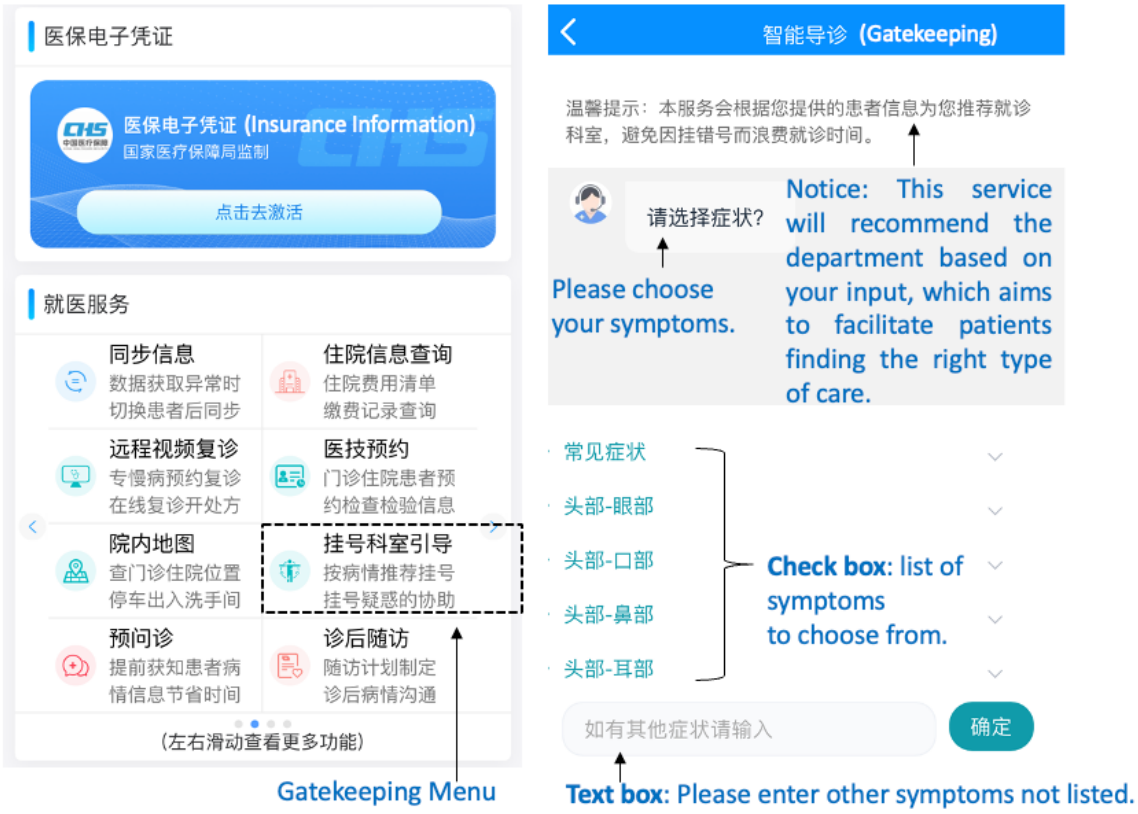


Figure 2. Illustration of the Experiment Design

Recommendation Based on your symptom, the AI expert suggests you visit the following department for a consultation. (Disclaimer: This is only advice, not equating a practitioner's diagnosis.)	**Group 1**
Recommendation [Algorithm Accuracy] Based on your symptom, the AI expert suggests you visit the following department for a consultation. (Disclaimer: This is only advice, not equating a practitioner's diagnosis.)	**Group 2**
Recommendation [Peer Acceptance] Based on your symptom, the AI expert suggests you visit the following department for a consultation. (Disclaimer: This is only advice, not equating a practitioner's diagnosis.)	**Group 3**
Recommendation [Algorithm Accuracy & Peer Acceptance] Based on your symptom, the AI expert suggests you visit the following department for a consultation. (Disclaimer: This is only advice, not equating a practitioner's diagnosis.)	**Group 4**

Figure 3. An Example of an AI Gatekeeping Session

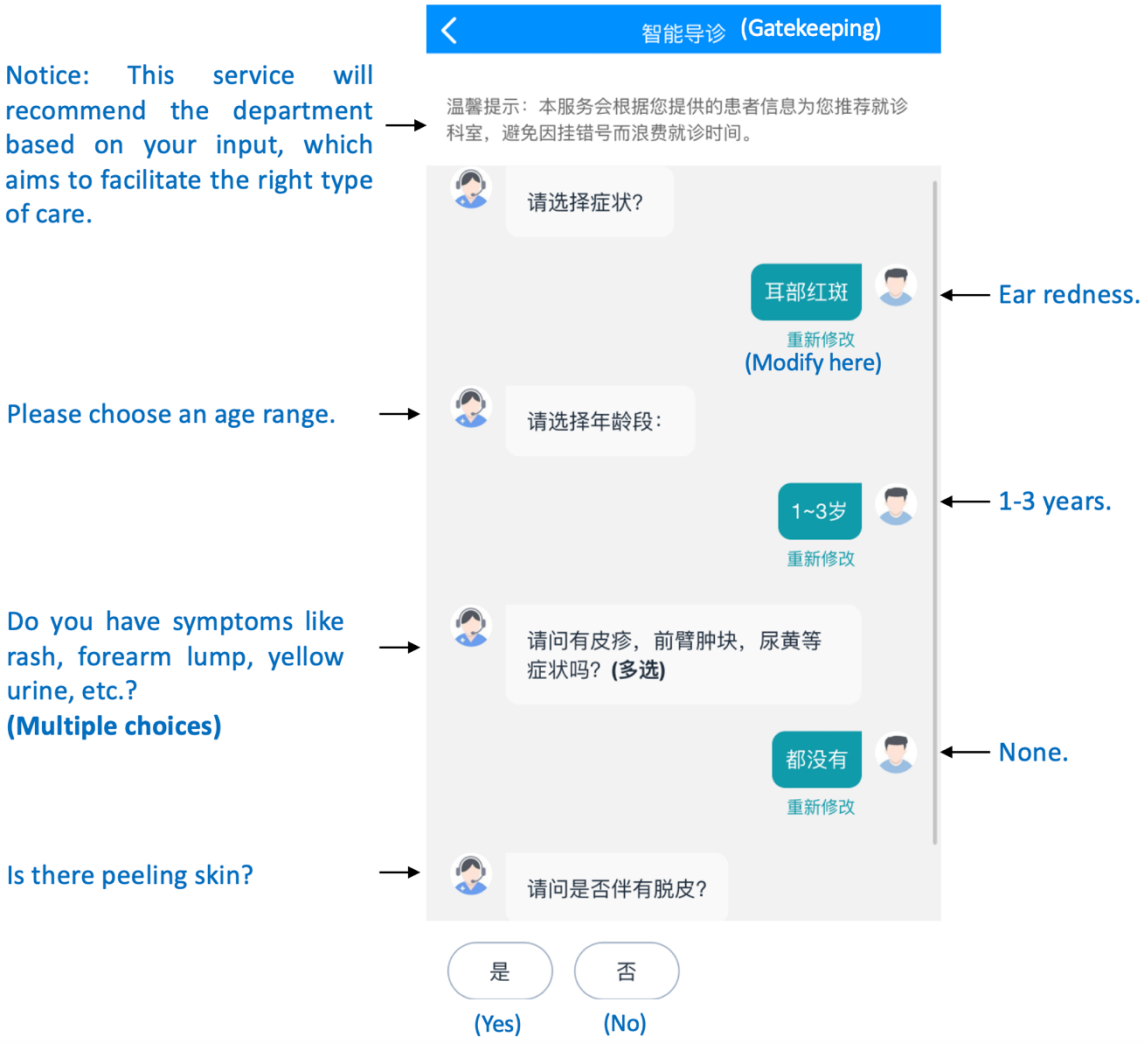


Figure 4. An Example of AI Gatekeeping Output



Figure 5 Visualization of the Independent Effect of Algorithm Accuracy

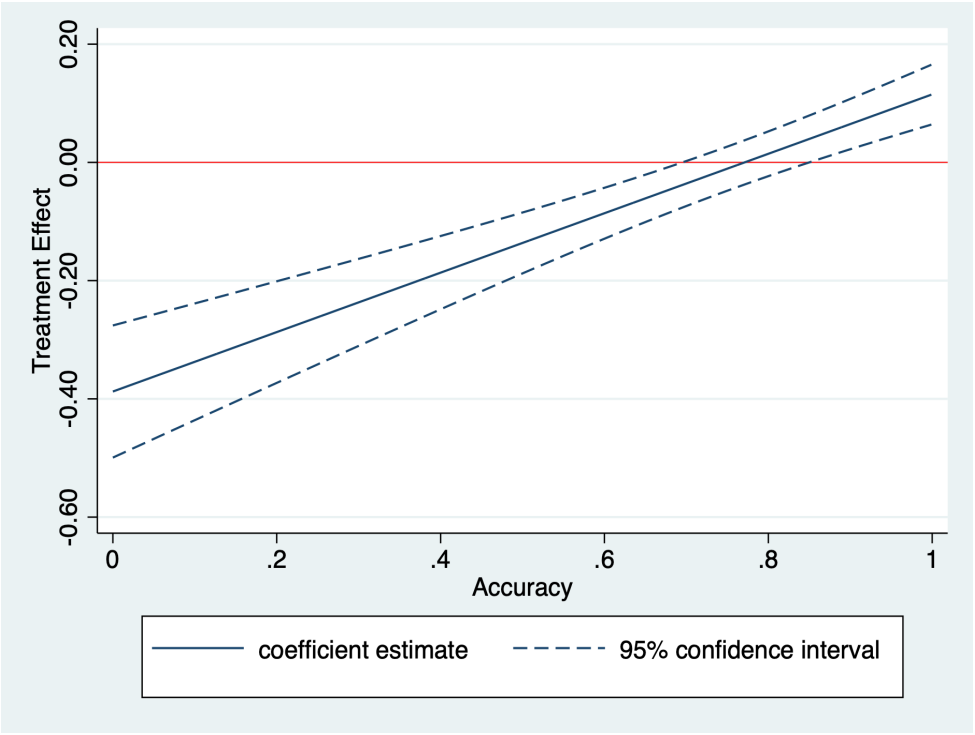


Figure 6 Visualization of the Independent Effect of Peer Acceptance

