We are delighted to be considered for either a full-length or shorter presentation.

AI, Authority, and Expert–Client Interaction: Experimental Evidence from Healthcare*

Yuyu Chen† Hongbin Li‡ Lingsheng Meng§ Xinyao Qiu¶ Qingxu Yang‖

March 26, 2026

Abstract

Generative artificial intelligence (AI) may reshape expert–client decision-making by reducing informational asymmetries that underlie expert authority. We study this in healthcare, a canonical credence-good market characterized by substantial information asymmetry between physicians and patients. In a preregistered field experiment involving more than 10,000 outpatient visits in China, we randomly assign patients access to a generative AI chatbot before their consultation. Access to the chatbot changes clinical practice: prescription rates decline, particularly for Traditional Chinese Medicine and antibiotics, while diagnostic testing increases and short-term revisit rates fall. Evidence suggests that improved patient information changes treatment requests and constrains physician discretion. These shifts point to a broader phenomenon: the defensively cautious, liability-driven guardrails encoded in AI training propagate into real clinical decisions at scale. Survey evidence reveals a divergence in perceptions. Patients with chatbot access report lower satisfaction and intended adherence, whereas physicians perceive clearer communication and better patient preparation. These findings show that generative AI, and the institutional logic embedded in its training, reshapes clinical decision-making and the dynamics of the patient–physician interaction.

Keywords: Generative AI, Information Asymmetry, Healthcare Field Experiment.

JEL Codes: I11, I18, O33, D82.

*This experiment was preregistered at AEA RCT Registry (AEARCTR-0015851), and approved by Institutional Review Board of Guanghua School of Management, Peking University (#2025-14), and Human Research Ethics Committee (HREC) of University of Hong Kong (EA250830).

†Peking University.

‡Stanford University.

§Stanford University.

¶University of Hong Kong.

‖University of Hong Kong.

1 Introduction

Generative artificial intelligence (AI) is rapidly transforming economic activity. Much of the existing research studies AI as a production technology that augments worker productivity or substitutes for labor in specific tasks. Yet many economically important distortions arise not from production inefficiency, but from information asymmetries in markets where decisions are delegated to experts (such as physicians, lawyers, or financial advisors) whose judgments consumers cannot easily evaluate. In such settings, expertise confers not only knowledge but authority over decisions. The implications of generative AI may therefore extend beyond productivity: by reducing information asymmetries, it may also reshape the allocation of decision authority between experts and their clients.

We study this question in healthcare, a canonical credence-good market characterized by severe informational asymmetries and substantial expert discretion (Arrow, 1963; Darby and Karni, 1973; McGuire, 2000). Patients typically lack the knowledge required to assess treatment necessity or evaluate alternative options. As a result, clinical decisions are largely determined by physician judgment, shaped by practice style, professional norms, and institutional incentives (Currie et al., 2026). In such settings, informational asymmetries and expert discretion can generate overtreatment or other forms of inefficient care. At the same time, medical expertise is essential for appropriate care, and weakening deference to expertise may carry risks for adherence and treatment effectiveness.

Generative AI introduces a new force into this equilibrium. Large language models (LLMs) dramatically reduce the cost of accessing and processing medical information for patients.[1] In doing so, generative AI functions as a consumer-side information infrastructure. Unlike traditional reforms that target physician incentives or oversight, this technology changes the informational environment in which clinical decisions are made. Better-informed patients may both change their treatment requests and defer less to physician recommendations, potentially shifting clinical decisions and the balance of authority in the consultation. We further argue that generative AI operates through an additional channel: the information it provides may not be neutral, but reflects the defensive, liability-driven guardrails encoded in AI training, which could introduce a systematically cautious decision logic into expert–client interactions. The key question we ask is whether improving patient

[1] Health-related queries are already among the most common use cases of ChatGPT (https://openai.com/index/introducing-gpt-5/).

information via generative AI—without changing prices, contracts, or physician compensation—alters clinical decisions and the balance of authority in such high-stakes expert markets.

To empirically answer this question, we conducted a preregistered month-long field experiment that provides the first randomized evidence on how patients' access to generative AI tools reshapes clinical practice, healthcare delivery, and interactions between patients and physicians. Our randomized controlled trial took place in a large public hospital in China and involved over 10,000 outpatient visits. Patients were randomly assigned to receive, or not receive, access to an LLM-based chatbot prior to their appointments. The chatbot invited patients to describe their symptoms and health concerns in detail before the visit and provided free access to AI-generated, personalized guidance and general medical information.

We implement a two-layer randomization design. We first randomize physicians into an *Exposed* group or an *Unexposed* group. Among patients scheduling appointments with *Exposed* physicians, half are randomly assigned to receive access to the chatbot (*Treatment*) and half to receive no access (*Control*). Patients scheduling appointments with *Unexposed* physicians never receive chatbot access. This design allows us to compare clinical outcomes between patients with and without access to the chatbot who see the same physician, while also testing for within-physician spillovers in clinical practice affecting patients without access. Together, these comparisons shed light on both the direct effects of AI-assisted patient preparation and the ways in which generative AI tools may reshape physician behavior and care delivery.

Offering patients access to the chatbot leads to sizable changes in clinical practice and healthcare utilization. Within-physician comparison suggests that chatbot access reduces the likelihood of receiving any prescription by about 4.5 percentage points relative to a baseline rate of 87%, driven primarily by declines in Traditional Chinese Medicine and antibiotic prescribing. At the same time, chatbot access increases the probability that a visit includes diagnostic testing by roughly 2.7 percentage points from a baseline of 23%. This pattern of reduced prescribing and increased diagnostic testing may partly reflect the defensive, liability-driven guardrails encoded in AI training, which tend to caution against medication use and encourage diagnostic clarification. These shifts in clinical decision-making translate into lower medication expenditures, while diagnostic spending rises modestly, resulting in no statistically significant change in total healthcare expenditures. We find no evidence that chatbot access worsens short-term health outcomes. If anything, treated patients

are 1.1 percentage points less likely to revisit the same hospital within two weeks, although this reduction is only marginally significant and may reflect changes in care-seeking behavior rather than improvements in underlying health.

Our main results report intent-to-treat effects based on randomized access to the chatbot, regardless of actual usage. Linking chatbot usage logs to hospital administrative data shows that take-up is approximately 17 percent, with higher usage among younger, male, employed patients and first-time visitors. We therefore estimate treatment-on-the-treated effects by instrumenting actual chatbot usage with randomized assignment. Chatbot usage is associated with a 26 percentage point reduction in the likelihood of receiving any prescription, a 16 percentage point increase in diagnostic testing, and a 6.6 percentage point reduction in two-week revisits. These effects are roughly five times larger than the corresponding intent-to-treat estimates, reflecting partial take-up, and should be interpreted as local average treatment effects for patients who choose to use the chatbot when access is provided.

Next, we examine whether within-physician spillovers arise from physician learning or behavioral adjustment following repeated interactions with AI-prepared patients. Such spillovers would imply that patients who do not directly use the chatbot, such as older or less educated patients, could nevertheless benefit indirectly from generative AI. However, we find no evidence of such spillovers. Among patients without chatbot access, clinical outcomes do not differ between those who see *Exposed* physicians and those who see *Unexposed* physicians. We also find no persistence of effects: differences in clinical practice across physician groups are entirely driven by visits involving chatbot-access patients and disappear immediately after the experiment ends. This null result is economically informative. It suggests that generative AI does not automatically reshape expert norms or clinical decision-making, at least at the exposure rate in our experiment. Instead, the effects operate primarily through direct patient use. From a policy perspective, this implies that unequal access to AI tools may translate directly into unequal benefits, since improvements do not diffuse passively through provider learning.

To understand how patient access to the chatbot affects patient behavior and the physician–patient relationship, we combine administrative medical records with survey data from patients and physicians. We find that access to the chatbot does not materially change observable patient behavior, including appointment attendance, medication purchases, or completion of diagnostic

tests. However, survey evidence points to a divergence in perceptions of the consultation experience. Patients with chatbot access report lower overall satisfaction, perceive communication with physicians as less smooth, and express lower intended compliance with medical advice. Physicians who interact with AI-prepared patients also report lower patient compliance, but at the same time perceive these patients as easier to communicate with, with clearer symptom descriptions and a better understanding of their condition. The reduction in both intended and observed compliance is consistent with Finkelstein et al. (2022) and suggests that generative AI may shift informational authority within clinical encounters, reshaping how patients view physician expertise and potentially weakening adherence.

Finally, we explore two mechanisms that may explain these shifts in clinical practice. First, the intervention may operate by changing patients' treatment requests. By improving patient information, the chatbot may reduce their demand for unnecessary medications while increasing interest in diagnostic testing. Indeed, analysis of chatbot conversations shows that the model frequently advises symptom monitoring and cautions against Traditional Chinese Medicine and antibiotic use, while often discussing diagnostic testing. Moreover, the treatment effects are concentrated among physicians who report being receptive to patient input and largely absent among physicians who emphasize strict authority. These patterns suggest that better-informed patients are less likely to request unnecessary medications, and such requests translate into prescriptions primarily when physicians are receptive to patient input.

Second, the intervention may shift the balance of authority within the consultation. Better-informed patients may be less deferential to physician recommendations and more willing to question intensive treatment. Consistent with this hypothesis, we find that reductions in medication use are largest among physicians with higher baseline prescribing intensity, particularly for Traditional Chinese Medicine and antibiotics. In addition, as shown earlier, treated patients report lower intended compliance, suggesting that AI-assisted patients rely less on physician recommendations when making treatment decisions. Taken together, these patterns suggest that improved patient information limits physician discretion in treatment decisions.

This study makes three primary contributions to the economics literature. First, we provide the first large-scale causal evidence on the impact of generative AI as an information intermediary in a credence-good market. Existing studies typically examine information asymmetries in healthcare by

exploiting whether patients are physicians themselves (Grytten et al., 2011; Johnson and Rehavi, 2016; Frakes et al., 2021; Chen et al., 2025) or have family members in the medical profession (Artmann et al., 2022; Chen et al., 2022; Finkelstein et al., 2022) as sources of variation in patients' information, but the evidence from this literature is mixed. Some studies find that better-informed patients experience less overtreatment (Johnson and Rehavi, 2016), while others report only limited effects (Frakes et al., 2021; Artmann et al., 2022). By leveraging LLM-based technology, our study contributes to this body of literature by randomly providing a subset of patients with access to personalized, easy-to-understand information on their health conditions.[2] More broadly, the project also relates to studies of health-related information and education programs in developing country settings, as surveyed in Dupas (2011b).

Going beyond clinical decisions, we examine how access to information reshapes patient–physician interactions. Similar to Currie et al. (2011), who show that knowledgeable simulated patients constrain physicians' propensity to overprescribe antibiotics but experience less courteous interactions, we find that patients' access to LLM tools reduces prescribing while worsening self-reported experiences. At the same time, patients with access to LLM tools report lower compliance with physician recommendations, consistent with observational evidence that patients with family members in the medical profession exhibit reduced adherence (Finkelstein et al., 2022). These findings highlight the importance of looking beyond treatment and testing decisions to understand how information technologies reshape relational dynamics in clinical care.

Beyond outpatient healthcare, our study contributes to a broad literature on credence goods markets. These markets are economically significant and pervasive. From real estate (Levitt and Syverson, 2008) and financial advice (Mullainathan et al., 2012) to pharmacy (Bronnenberg et al., 2015), automobile and computer repair (Schneider, 2012; Busse et al., 2017; Rasch and Waibel, 2018; Kerschbamer et al., 2023), and taxi services (Balafoutas et al., 2013, 2017), consumers routinely delegate decisions to better-informed experts whose recommendations they cannot easily verify. A common thread across these settings is that informational asymmetries between experts and clients create scope for overtreatment, fraud, or misallocation. Consistent with this, Balafoutas et al. (2013),

[2]Related literatures study other types of health information, including second opinions (Mimra et al., 2016), patient decision aids that provide standardized materials about specific treatment options (surveyed by Stacey et al., 2017, 2024), provider report cards and online ratings that disclose information about healthcare provider quality (e.g., Dranove et al., 2003; Kolstad, 2013; Angerer et al., 2026), and educational campaigns on health risks and the importance of diagnostic testing (e.g., Dupas, 2011a; Lopez et al., 2024).

Busse et al. (2017), and Kerschbamer et al. (2023) show that reducing informational asymmetries can discipline expert behavior. Our study extends this literature by showing that generative AI can serve as a scalable consumer-side information technology, disciplining expert behavior in one of the most consequential credence goods markets – outpatient healthcare.

Finally, we contribute to the rapidly growing economics of AI literature. Much of the existing work studies generative AI as a production technology that augments worker productivity or substitutes for labor in specific tasks (e.g., Dell'Acqua et al., 2023; Peng et al., 2023; Noy and Zhang, 2023; Cui et al., 2024; Brynjolfsson et al., 2025; Dillon et al., 2025). Viewed through the lens of expert–client relationships, a smaller stream examines AI tools that assist experts themselves, including physicians (Abaluck et al., 2026), financial analysts (Bertomeu et al., 2025), and legal workers (Choi et al., 2024). Much less is known about generative AI as a tool for consumers who face informational disadvantages relative to the experts they rely on. The limited existing work includes Spatharioti et al. (2023), which uses online experiments to examine how LLM-based tools help consumers search for and compare products. By providing patients with access to an LLM-based chatbot prior to outpatient visits, our study offers the first large-scale randomized evidence on how generative AI can empower informationally disadvantaged clients by reducing their reliance on expert authority and reshaping the dynamics of the expert–client relationship.[3]

Our study also speaks to a broader question for the economics of AI that has received little attention: how do the norms and constraints encoded in AI training propagate into real-world decisions at scale? LLMs exhibit systematic behavioral patterns shaped by their design, including tendencies that vary with model architecture, training data, and scale (e.g., Filippas et al., 2024; Bowen et al., 2025; Garg and Fetzer, 2025; Bini et al., 2025). In high-stakes domains such as healthcare, these patterns are not incidental. Major AI developers have explicitly encoded defensive, liability-driven guardrails into their models, instructing them to avoid recommending specific treatments and to defer to licensed professionals. When such models are deployed as consumer-facing information tools, these encoded constraints enter real-life decisions at scale. Our findings suggest this propagation

[3]Recent randomized trials in clinical settings have tested integrating LLM tools into hospital workflows. Wan et al. (2024) find that an LLM chatbot assisting nurses at outpatient reception improves patient satisfaction and communication efficiency, and Tao et al. (2026) show that a pre-consultation chatbot performing clinical history-taking, preliminary diagnoses, and test ordering reduces specialist consultation time and improves coordination. Both studies deploy AI to augment provider functions and find improved patient satisfaction. In contrast, our study deploys AI as a patient-side information tool and examines its effects on both clinical decisions and the broader dynamics of the physician–patient relationship.

mechanism may be quantitatively important, with the potential to shift aggregate outcomes such as population-level healthcare utilization.

2 Institutional Background

2.1 Generative AI in Healthcare: Adoption and Guardrails

Generative AI has rapidly become a major channel through which people access health information. Health-related queries are among the most common use cases of large language models: OpenAI reports that approximately 230 million people discuss their health with ChatGPT each week,[4] and surveys indicate that roughly one in six American adults consults an AI chatbot for health information at least once a month (Kaiser Family Foundation, 2024). This growth has occurred largely ahead of formal regulatory frameworks, creating a setting in which AI systems interact with patients at massive scale without the oversight structures that govern licensed medical professionals.[5]

In response to safety concerns and legal exposure, major AI developers have explicitly encoded defensive guardrails into their models for health-related interactions. OpenAI's published model specification instructs its systems to avoid providing tailored medical advice, to refrain from recommending specific treatments or medications, and to direct users to licensed professionals instead.[6] Anthropic has similarly stated that Claude is designed to include contextual disclaimers, acknowledge uncertainty, and direct users to healthcare professionals for personalized guidance.[7] These guardrails are not merely stylistic: they are the product of deliberate training choices that systematically shape model outputs toward caution, diagnostic clarification, and deference to professional judgment.

[4] https://openai.com/index/introducing-chatgpt-health/

[5] The regulatory debate over AI's role in healthcare is actively evolving. California's Assembly Bill 489, signed into law in October 2025, directly addresses the boundary between AI and licensed medical practice: it prohibits developers and deployers of AI systems from using any terms, phrases, or design elements that imply the system is providing care from a licensed healthcare professional, and authorizes state licensing boards to investigate and penalize violations.https://leginfo.legislature.ca.gov/faces/billNavClient.xhtml?bill_id=202520260AB489.

[6] https://model-spec.openai.com/2025-04-11.html

[7] https://www.anthropic.com/news/healthcare-life-sciences

2.2 Healthcare Market and Institutional Context in China

Healthcare expenditures have risen rapidly worldwide. Economic research points to both demand-side and supply-side mechanisms as drivers of high healthcare spending. On the demand side, beyond the classic moral hazard problem,[8] patients often lack the medical knowledge needed to evaluate the necessity of care and may request treatments that are not clinically indicated. Physicians may comply with such requests, contributing to overutilization (e.g., Kravitz et al., 2005; Kotwani et al., 2010; Linder et al., 2014; Finkelstein et al., 2016; Lopez et al., 2022). On the supply side, physician characteristics and financial incentives directly shape treatment decisions, generating systematic variation in medical expenditures (e.g., Iizuka, 2012; Das et al., 2016; Finkelstein et al., 2016; Cutler et al., 2019; Badinski et al., 2023; Currie et al., 2026).[9]

Overutilization, particularly in the form of overprescription, is well documented in China (e.g., Li et al., 2012; Lu, 2014; Fang et al., 2021). The healthcare system in China is predominantly state-run, with public hospitals delivering the vast majority of services. Despite ongoing payment reforms, hospitals continue to operate largely under a fee-for-service payment model and receive limited direct government subsidies. As a result, hospital revenue remains closely tied to service volume. Pharmaceutical sales occur primarily within hospitals rather than through independent retail pharmacies. Between 2014 and 2019, drug expenditures represented more than 60 percent of inpatient spending and roughly 70 percent of outpatient spending in public hospitals.[10] These shares are substantially higher than those observed in many high-income health systems, where prescription drugs account for a considerably smaller fraction of total health expenditures.[11]

Motivated by this institutional setting, we examine whether providing patients with easy-to-understand, personalized medical information reduces prescription intensity in China. On the demand side, improved information may reduce patients' request for clinically unnecessary treatments. On the supply side, better-informed patients may constrain physicians' discretion to prescribe treatments

[8] China has near-universal health insurance coverage through two main public schemes: the Urban Employee Basic Medical Insurance (UEBMI), which covers formal-sector workers, and the Urban and Rural Resident Basic Medical Insurance (URRBMI), which covers rural residents and non-working urban populations.

[9] Moreover, Gottschalk et al. (2020) document that patients with higher socioeconomic status are significantly less likely to receive unnecessary treatment recommendations, indicating that physicians' propensity to overtreat depends on patient characteristics.

[10] China Health Statistical Yearbook, various years.

[11] Across OECD countries, pharmaceutical spending typically accounts for roughly 15–20% of total health expenditures. https://www.oecd.org/en/data/indicators/pharmaceutical-spending.html

that exceed clinical necessity when financial incentives favor higher prescribing.

Beyond overall prescribing, we focus on three drug categories: Traditional Chinese Medicine (TCM), antibiotics, and opioids. TCM is widely prescribed in Chinese public hospitals but is characterized by less standardized evidence on effectiveness and side effects,[12] and often higher or more variable pricing. Antibiotic overprescription has been well documented in China (Currie et al., 2013, 2014) and in the United States (Fleming-Dutra et al., 2016), raising concerns about antimicrobial resistance. Opioids, while less prevalent in China than in the United States (Maclean et al., 2020), have drawn substantial international attention due to concerns about misuse and addiction. We focus on these three categories because they are highly policy relevant and have been central to regulatory efforts and public health debates.

A further institutional feature of the Chinese outpatient system amplifies the informational asymmetries faced by patients. Unlike systems organized around general practitioners or family medicine, China has no meaningful primary care gatekeeping layer: patients schedule appointments directly with specialists without prior referral or triage by a generalist physician. As a result, patients arrive at outpatient specialist consultations without access to a trusted intermediary who might explain their condition in plain terms or provide continuity of medical guidance over time. The informational disadvantage patients face is therefore considerably more severe than in settings where a family doctor serves as an ongoing informational resource. Compounding this, physician–patient interactions in the Chinese outpatient system are largely episodic rather than ongoing: the same patient rarely sees the same specialist repeatedly. In credence-good markets, repeated interaction is one of the few mechanisms that can discipline expert behavior even absent formal oversight (Dulleck and Kerschbamer, 2006; Fong et al., 2022). Its near-absence in Chinese outpatient care leaves patients especially vulnerable to physician discretion and makes the potential value of patient-side information particularly large.

We conducted our experiment within this context by partnering with a large public county-level hospital in Sichuan Province. As the largest medical facility in its county, it employs 200 outpatient physicians who handle approximately 900 visits per day for a population of 600,000. As a county-level tertiary hospital, it is representative of the mid-tier public institutions that serve roughly 40% of

[12] Under new regulations effective July 1, 2026, TCMs that do not clearly disclose contraindications, adverse reactions, or precautions will not be approved upon re-registration. Reports indicate that over 70 percent of currently approved products may be affected. https://www.news.cn/politics/20260129/853dc80d18fb43d2b1d61676c46b5675/c.html

China's population and account for 35% of nationwide outpatient visits, enhancing the external validity of our findings.

This institutional setting provides several advantages for implementing and evaluating our intervention. First, the hospital's integrated electronic medical record (EMR) system allows for comprehensive outcome measurement. Second, its online appointment platform enables seamless randomized treatment assignment. Third, the predominantly rural patient population has limited exposure to digital technologies. This ensures a low baseline of AI adoption, minimizing the risk that background usage dilutes the treatment effect. Finally, like most public hospitals in China, our partnering hospital operates under severe capacity constraints. Outpatient consultations are brief, averaging about nine minutes per visit. Limited consultation time restricts explanation, counseling, and patient engagement, making patients' advance preparation and clear communication particularly valuable. In this setting, additional health information and support from generative AI tools may be especially relevant.

3 Experimental Design and Data

3.1 Intervention

Our intervention is a generative AI chatbot powered by a large language model built on a widely used general-purpose architecture, adapted for healthcare-related interactions, and integrated into the hospital's online appointment system. Immediately after a patient completes an online appointment booking, which occurs on average one day before the visit, those randomly assigned to the treatment group receive a personalized link to the chatbot. Simultaneously, the hospital sends a separate message encouraging patients to use the tool prior to their visit. Panels (a) and (b) of Figure 1 present the chatbot's initial landing screen and consultation interface.

The chatbot supports pre-visit patient preparation by eliciting symptom descriptions, structuring patient narratives, and providing general medical information. Throughout all interactions, the interface prominently reminds users that the chatbot does not replace professional medical advice and that clinical decisions should follow physicians' recommendations.

The system automatically logs all user interactions, including timestamps and message content. These records provide an objective measure of engagement with the intervention, which we link

to administrative and clinical outcomes using unique patient identifiers. This structure enables estimation of both intent-to-treat and treatment-on-the-treated effects.

The intervention is designed to reduce informational asymmetries that characterize outpatient care and, in doing so, may affect both communication and clinical decision-making. First, by prompting patients to organize and describe symptoms and to review relevant health information before the visit, the chatbot may improve pre-visit preparation and facilitate clearer communication during the consultation. This is particularly important in our setting, where the average outpatient visit lasts approximately nine minutes, making efficient and precise communication essential. Beyond communication, access to structured medical information may improve patients' understanding of their health conditions, typical diagnostic pathways, and the role of medications and tests. Better informed patients may therefore be less likely to request unnecessary treatments or prescriptions, reducing demand-driven overuse. Meanwhile, by narrowing informational gaps, the chatbot may shift the balance of authority within the consultation. Patients who arrive with greater knowledge and more structured questions may be better positioned to question recommendations and constrain physician incentives to over-prescribe or over-treat. Together, our intervention can reshape the clinical interaction between patient and physicians through multiple channels.

3.2 Two-Layer Randomization

We implement a two-layer randomized experimental design to identify the effects of providing patients with access to the chatbot while creating variation to detect within-physician spillovers. Physicians are first randomized at the physician level into an *Exposed* group (two-thirds) or an *Unexposed* group (one-third), stratified by seniority, education, and specialty. Physicians in the *Unexposed* group never interact with AI-assisted patients. Among patients scheduling appointments with *Exposed* physicians, half are randomly assigned to receive access to the chatbot (*Treatment*) and half to receive no access (*Control*). Patients scheduling appointments with *Unexposed* physicians never receive chatbot access.

Panel (c) of Figure 1 summarizes the two-layer randomization. This design generates three analytically distinct groups. Comparisons between the *Treatment* and *Control* groups identify the direct effect of patient access to the chatbot within the same physician panel. Comparisons involving the *Unexposed* group allow us to detect within-physician spillovers arising from physician learning or

behavioral adjustment following repeated interactions with AI-prepared patients.

The three analytical groups are formally labeled as:

1. **Treatment (T):** Patients with chatbot access, seeing *Exposed* physicians.
2. **Control (C):** Patients without chatbot access, seeing *Exposed* physicians.
3. **Unexposed (U):** Patients without chatbot access, seeing *Unexposed* physicians.

3.3 Data and Sample

The experiment was conducted between June 2 and July 3, 2025, at a partnering hospital. Because the intervention is delivered through the online appointment system, we restrict the sample to patients who scheduled outpatient visits online during this period. Patients who walked in without an online appointment, approximately one third of total outpatient visits, are excluded from both the experiment and the analysis sample.[13]

During the experimental period, 179 physicians provided outpatient care. Of these, 117 (two thirds) were randomized to the *Exposed* group and 62 (one third) to the *Unexposed* group. Panel A of Table 1 reports balance tests at the physician level and shows that the two groups are well balanced across observed characteristics, including age, gender, seniority, education, and specialty, with no statistically significant differences.

A total of 11,666 patients scheduled visits with *Exposed* physicians through the online appointment system and were randomized with equal probability into the *Treatment* or *Control* group. In addition, 6,010 patients scheduled visits with *Unexposed* physicians. Panel B of Table 1 reports balance tests at the patient level, comparing the *Treatment* and *Control* groups as well as the *Treatment* and *Unexposed* groups. Across these comparisons, we find no statistically significant differences in age, gender, occupation, first-time visit status, or advance booking time. The average patient is approximately 45 years old, and a large share report working as farmers (44%) or being unemployed or retired (24%). This demographic profile suggests that baseline familiarity with generative AI tools is likely to be limited in this population. Appointments are booked, on average, approximately

[13]We additionally exclude pediatric patients from the experiment. Approximately 0.6% of patients who booked an appointment but did not complete payment are also excluded from the analysis sample.

25 hours in advance, giving treated patients adequate opportunity to use the chatbot if they choose to do so.

Medical Records. We observe administrative outpatient medical records for all visits during the study period. These records include patient and physician identifiers, visit timestamps, diagnoses, detailed drug prescriptions, and orders for diagnostic examinations. For each prescribed drug and diagnostic test, we observe the listed price, which captures medical expenditures prior to insurance reimbursement. These data allow us to construct visit-level measures of clinical decision-making, treatment intensity, and spending outcomes.

Patient Survey. We invited patients to complete a brief post-consultation survey on the day of the visit or the following day, offering a monetary incentive of 2 to 10 RMB. The survey elicits patients' satisfaction with the visit, perceived quality of communication, trust in the physician's diagnosis, and intended adherence to medical advice. It also collects baseline characteristics, including education, general trust in medical providers, and attitudes toward traditional Chinese medicine. For patients in the *Treatment* group, the survey includes additional questions on chatbot usage and perceived usefulness. In total, we collected 2,247 patient responses, corresponding to a 13 percent response rate. The sample includes 640 respondents from the *Treatment* group, 789 from the *Control* group, and 818 from the *Unexposed* group.

Physician Survey. Following the conclusion of the one-month experiment, we surveyed all physicians who conducted outpatient consultations during the study period, achieving a response rate of 95.5 percent, with 171 of 179 physicians completing the questionnaire. The survey elicits physicians' assessments of their patients along four dimensions, clarity in symptom descriptions, patients' understanding of their conditions, ease of communication during the visit, and patients' potential adherence to medical advice. Both *Exposed* and *Unexposed* physicians were also asked to estimate how frequently their patients used AI tools prior to visits. Taken together, the physician and patient surveys complement the administrative data by validating treatment exposure and providing insight into how AI-assisted health information affects consultation dynamics and patient welfare.

4 Main Results

4.1 Graphical Evidence Across Experimental Groups

We graphically present mean outcomes for the *Treatment*, *Control*, and *Unexposed* groups over the experimental period, beginning with prescribing behavior. Panel (a) of Figure 2 shows that access to the chatbot is associated with a substantial reduction in the likelihood that a visit results in any prescription. In the *Treatment* group, the share of visits with at least one prescribed medication is 4 percentage points lower than in the *Control* group, where the mean prescribing rate is 88%, a notably high level. This difference is statistically significant at the 5% level. The prescribing rate in the *Treatment* group is also significantly lower than in the *Unexposed* group, while prescribing rates in the *Control* and *Unexposed* groups are statistically indistinguishable.

We next examine prescribing patterns across major medication categories, including Traditional Chinese Medicine (TCM), antibiotics, and opioids. As shown in Panels (b) and (c) of Figure 2, prescribing of both TCM and antibiotics declines sharply in the *Treatment* group, with statistically significant differences relative to both comparison groups. Opioid prescribing is also lower among treated patients, but the estimated differences relative to the *Control* and *Unexposed* groups are not statistically significant. This is expected given the low baseline rate of opioid prescribing in China, at below 3% in both the *Control* and *Unexposed* groups, as shown in Panel (d).

At the same time, physicians seeing treated patients are more likely to order diagnostic tests. Panel (e) shows that in the *Control* and *Unexposed* groups, approximately 21–22% of visits involve at least one diagnostic test, compared with roughly 25% in the *Treatment* group. The increase in testing is statistically significant relative to both comparison groups. This pattern of reduced prescribing and increased diagnostic testing may partly reflect the defensive, liability-driven guardrails encoded in AI training, which tend to caution against medication use and encourage diagnostic clarification.

We next assess whether these changes in clinical management are associated with differences in patient outcomes. One outcome observed in the administrative data is whether a patient revisits the same hospital within two weeks of the initial consultation, coded as a binary indicator. Panel (f) of Figure 2 shows that patients in the *Treatment* group are 1.4 percentage points less likely to revisit within two weeks than patients in the *Control* group, corresponding to an 8.8% reduction. This difference is statistically significant. Because the *Unexposed* group exhibits a similar two-week

revisit rate to the *Control* group, the same conclusion holds when comparing treated patients to unexposed patients.

The reduction in short-term revisits is consistent with improvements in the effectiveness of initial consultations, but it may also reflect other mechanisms. For example, better-prepared patients may feel less need to seek follow-up care, even in the absence of changes in clinical quality, or the intervention may affect patients' expectations about the necessity of additional visits. Because the revisit measure captures subsequent healthcare utilization rather than health status directly, these results should be interpreted as changes in follow-up behavior rather than definitive evidence of improved clinical outcomes. In this rural setting, where patients often face substantial travel costs, reductions in revisits nonetheless have potential welfare implications, although the underlying channels remain open to interpretation.

4.2 Within-Physician Intent-to-Treat Effects

A key pattern in the graphical results is the similarity of outcomes between the *Control* and *Unexposed* groups, suggesting limited within-physician spillovers. We formally assess the presence of such spillovers in subsequent analyses. Motivated by this pattern, our main regression analyses restrict the sample to visits with *Exposed* physicians. Within this sample, we estimate the causal effects of offering access to the LLM-based chatbot on clinical outcomes using the following specification:

$$Y_{ip} = \alpha + \beta \text{Treated}_{ip} + \gamma X_{ip} + \delta_p + \varepsilon_{ip} \tag{1}$$

where the unit of observation is the outpatient visit for patient i seeing physician p. The outcome variable Y_{ip} represents various measures of the consultation, including prescription patterns, diagnostic procedures, and healthcare expenses. The key explanatory variable, Treated_{ip}, is a binary indicator equal to one if patient i was randomly assigned access to the chatbot prior to the visit. We include a vector of patient characteristics X_{ip} (age, gender, occupation, and first-time visit status) and physician fixed effects δ_p to control for unobserved physician heterogeneity that might influence treatment outcomes. The error term ε_{ip} captures unobserved factors affecting individual visits.

The coefficient of interest, β, identifies the *Intent-to-Treat* (ITT) effect, capturing the differential outcomes when patients are assigned access to the chatbot, regardless of actual usage. This approach

preserves the integrity of the randomization and provides policy-relevant estimates of the impact of *offering* AI tools in a real-world clinical setting. It is important to note that if within-physician spillovers occur, for example, if physicians adjust their general practice patterns for all patients after interacting with AI-informed ones, the control group within this sample may also be indirectly affected. Consequently, our estimate of β should be interpreted as a conservative lower bound of the true impact relative to a counterfactual of zero exposure.

Table 2 reports within-physician intent-to-treat estimates of the effects of offering chatbot access on clinical practice patterns among visits with *Exposed* physicians. Consistent with the graphical evidence, the regression results show that access to the chatbot significantly reduces the likelihood that a visit results in any prescription by about 4.5 percentage points relative to a sample mean of 87%. This decline is driven primarily by reductions in prescribing of Traditional Chinese Medicine and antibiotics, which fall by 2.2 and 1.5 percentage points, respectively, while opioid prescribing also declines but is not statistically significant. At the same time, chatbot access increases the probability that a visit involves any diagnostic test by approximately 2.7 percentage points relative to a sample mean of 23%. Finally, treated patients are about 1.1 percentage points less likely to revisit the same hospital within two weeks; the effect is marginally statistically significant. Estimates are nearly identical with and without controls for patient characteristics, indicating that the results are not driven by observable differences in patient composition.

Table 3 reports intent-to-treat effects on healthcare expenditures, measured as the sum of listed prices for prescribed drugs and diagnostic tests prior to insurance reimbursement. Consistent with the changes in prescribing behavior documented in Table 2, access to the chatbot is associated with reductions in medication expenditures, significant at the 10% level. In the specification that controls for patient characteristics, TCM expenditures decline by approximately 1.7 RMB, corresponding to a 13% reduction relative to the sample mean, while Western Medicine expenditures decline by about 5.3 RMB relative to a sample mean of 64.8 RMB. At the same time, diagnostic expenditures increase by roughly 6.5 RMB, although this estimate is not statistically significant. As a result of these offsetting changes across expenditure components, we find no statistically significant effect of chatbot access on total healthcare expenditures.

4.3 Patient Take-up and Treatment-on-the-Treated Effects

The results above report intent-to-treat effects based on randomized access to the chatbot, regardless of whether patients actively used the tool. We now examine patient take-up of the chatbot and estimate treatment-on-the-treated effects that scale the intent-to-treat estimates by actual usage.

Linking chatbot usage logs to the hospital administrative system allows us to observe which patients actually used the chatbot when access was assigned. Among patients in the *Treatment* group, approximately 17 percent used the chatbot. By design, patients in the *Control* group were not given access and therefore could not use the tool. Appendix Table A1 reports first-stage estimates of the relationship between treatment assignment and chatbot usage. Assignment to the *Treatment* group increases the probability that a patient uses the chatbot by 17 percentage points. Usage is self-selected. Appendix Table A2 shows that, conditional on being assigned access to the chatbot, younger patients, men, employed patients, and first-time visitors are more likely to use it.

We next estimate treatment-on-the-treated effects, which recover the average effect of chatbot usage among patients who complied with the assignment. We instrument actual chatbot usage with randomized assignment to the *Treatment* group, using the same within-physician specification as in the intent-to-treat analysis. Table 4 reports the effects of chatbot usage on clinical practice patterns. Consistent with the intent-to-treat results, chatbot usage is associated with a 26 percentage point reduction in the probability that a visit results in any prescription, mostly driven by declines of about 13 percentage points in Traditional Chinese Medicine prescribing and 9 percentage points in antibiotic prescribing. At the same time, chatbot usage increases the likelihood of diagnostic testing by roughly 16 percentage points and reduces the probability of a two-week revisit by about 6.6 percentage points. These magnitudes are approximately five times larger than the corresponding intent-to-treat estimates due to partial take-up of the chatbot, which is approximately 17 percent in the treatment group. Accordingly, the instrumental variable estimates should be interpreted as local average treatment effects for patients who choose to use the chatbot when access is provided.

Table 5 reports the local average treatment effects on healthcare expenditures. Chatbot usage is associated with sizable reductions in medication expenditures, with TCM expenditures declining by about 10 RMB and Western Medicine expenditures declining by roughly 31 RMB, both statistically significant at the 10 percent level. These reductions are offset by an increase in diagnostic expenditures

of approximately 38 RMB, although this estimate is imprecisely estimated. For both Tables 4 and 5, first-stage F-statistics exceed 1,100 across specifications, indicating a strong first stage.

5 Dynamic and Spillover Effects

5.1 Dynamic Effects

Having established the average effects of chatbot access and usage, we next examine how these effects evolve over time and whether they generate spillovers. To study the temporal dynamics of the intervention, we implement an event-study framework. This approach serves three purposes. First, it provides a diagnostic check on the physician-level randomization by testing for differential pre-trends between *Exposed* and *Unexposed* physicians. Second, it characterizes how outcomes for *Exposed* physicians evolve over the course of the intervention. Third, it allows us to assess whether any effects persist after the formal intervention period ends.

We aggregate the data to the physician-week level and estimate the following specification:

$$Y_{pt} = \alpha + \sum_{k \neq -1} \beta_k \cdot \left(\text{Exposed}_p \times \mathbb{1}\{\text{Week}_t = k\}\right) + \delta_p + \tau_t + \varepsilon_{pt}, \tag{2}$$

where Y_{pt} denotes the outcome for physician p in week t, such as the rate of Traditional Chinese Medicine prescribing or diagnostic testing. The indicator $\mathbb{1}\{\text{Week}_t = k\}$ indexes weeks relative to the intervention start date of June 2, 2025. The coefficient for $k = -1$, corresponding to the week from May 26 to June 1, 2025, is normalized to zero. The variable Exposed_p equals one for physicians randomized to the *Exposed* group.

The coefficients of interest are $\{\beta_k\}_{k=-12}^{12}$. For $k < -1$, the estimated coefficients capture differences in outcomes between *Exposed* and *Unexposed* physicians prior to the intervention and provide a test for pre-existing differential trends. For $k \geq 0$, the coefficients trace the dynamic response of outcomes for *Exposed* physicians relative to *Unexposed* physicians. The intervention period extends through day 5 of week $k = 4$; coefficients for subsequent weeks therefore capture post-intervention dynamics. Weeks earlier than $k = -12$ and later than $k = +12$ are binned at the respective endpoints. All specifications include physician fixed effects δ_p, which absorb time-invariant differences across physicians, and week fixed effects τ_t, which control for common temporal shocks.

Standard errors are clustered at the physician level to account for serial correlation within physicians over time. Physician-week observations are weighted by the number of patient visits.

Figure 3 presents the results for medication prescribing rates (overall, Traditional Chinese Medicine, antibiotics, and opioids), diagnostic testing, and revisit rates. For all outcomes, we observe no evidence of pre-trends prior to the experiment, suggesting balance between *Exposed* and *Unexposed* physicians.

Weeks 0 through 3 correspond to the four full weeks of the intervention period, so the coefficients β_0–β_3 capture the treatment effect and its evolution during the experiment. Because the analysis compares outcomes between *Exposed* and *Unexposed* physicians, the estimated effects are mechanically attenuated. Only half of the patients seen by *Exposed* physicians are assigned access to the chatbot, and take-up among eligible patients is incomplete (approximately 17%). As a result, the magnitudes of the estimates are expected to be small and can be interpreted as roughly one-half of the corresponding intention-to-treat effects and approximately one-tenth of the average treatment-on-the-treated effects.

As shown in Figure 3, we observe clear reductions in overall medication prescribing and Traditional Chinese Medicine prescribing beginning in week 1, while the estimated effects on antibiotic and opioid prescribing are substantially noisier. Starting in week 1, we also observe a clear increase in diagnostic testing rates, along with a noisy and statistically insignificant decrease in revisit rates. While point estimates vary across weeks, we find no statistically significant evidence that the treatment effects change between week 1 and week 3.

The experiment concludes in the middle of week 4. Beginning in week 4, we find no differences in any of these outcomes between *Exposed* and *Unexposed* physicians. This pattern indicates no persistence in physicians' clinical practices following the intervention. Once the experiment ends, outcomes for *Exposed* physicians revert to their pre-intervention levels.

5.2 Within-Physician Spillovers

Having established that there are no spillovers to patients who see *Exposed* physicians after the experiment ends, we next test for within physician spillovers during the intervention period. Specifically, we examine whether interactions with patients who have access to the chatbot affect outcomes for patients without access who see the same physician. To do so, we adopt the event study framework in equation (2) with one modification. We exclude all visits by patients assigned

access to the chatbot. This restriction allows us to compare outcomes for patients without chatbot access who see *Exposed* physicians (*Control* group) with outcomes for patients without chatbot access who see *Unexposed* physicians (*Unexposed* group). If within physician spillovers or rebound effects are present, we would expect to observe differences in outcomes between these two groups of patients, even though neither group has access to the chatbot.

Appendix Figure A1 reports the results for the same set of clinical outcomes. Across all outcomes, we find no statistically significant differences between patients randomly assigned to the *Control* group and those assigned to the *Unexposed* group. This finding is consistent with the patterns observed in Figure 2. Combined with the absence of persistent effects documented in Figure 3, these results indicate that interactions with AI-prepared patients do not generate spillovers to other patients. In particular, physicians do not appear to update their clinical practices in ways that affect patients who do not have access to the chatbot, either contemporaneously or after the intervention ends. This pattern suggests that the effects of the chatbot operate primarily through direct patient use rather than through lasting changes in physician behavior.

From a policy perspective, these results raise concerns about unequal access and uptake. Because chatbot adopters are disproportionately younger, male, and employed, policies aimed at expanding access or improving usability may be necessary to ensure that generative AI tools benefit a broader patient population.

One potential concern is that exposure to AI-prepared patients may generate spillovers across physicians, for example through informal communication or peer learning. Such spillovers would bias our estimates of persistence and within-physician spillovers toward zero by contaminating the comparison group if *Unexposed* physicians indirectly update their practice over time. Several features of our setting mitigate this concern. First, physicians conduct consultations independently in separate rooms, limiting opportunities to observe or discuss specific patient interactions. Second, prescribing patterns among *Unexposed* physicians who never interact with treated patients remain stable over time. Third, for across-physician spillovers to drive the within-physician estimates to zero, the magnitude of across-physician learning would need to be comparable to the magnitude of within-physician learning generated by direct exposure to treated patients. We consider across-physician learning of that scale unlikely in our setting. Taken together, these patterns indicate that across-physician spillovers are unlikely to change the qualitative interpretation of our results.

6 Patient Behavior and Physician–Patient Relationship

Figure 4 summarizes changes in patient behavior and perceptions by treatment status. Panels (a) through (c) draw on administrative medical records. Panel (a) reports patient attrition, measured as whether patients who scheduled an appointment ultimately attended their consultation. We find no evidence that access to the chatbot discourages patients from seeing physicians or substitutes for in-person care, as attendance rates are similar across the *Treatment*, *Control*, and *Unexposed* groups. Panels (b) and (c) capture realized post-consultation behavior, measuring whether patients purchased prescribed medications and whether they completed prescribed diagnostic tests following the visit. For both outcomes, we find no statistically meaningful differences across the *Treatment*, *Control*, and *Unexposed* groups. This suggests that access to the chatbot does not materially affect patients' decisions to follow through on physicians' recommendations along these observable dimensions.

Panels (d) through (f) are based on patient survey responses and therefore reflect a more selected sample. While these survey-based measures should be interpreted with caution, they provide insight into how access to the chatbot may have affected the consultation experience. Treated patients report lower overall satisfaction, with 83.5% of respondents reporting being very satisfied, compared with approximately 86% in both the *Control* and *Unexposed* groups. Treated patients also perceive communication with physicians as less smooth, as reflected in a 2 percentage point reduction in the likelihood of rating the interaction as very smooth. Finally, intended compliance with physicians' recommendations declines substantially among treated patients, with the share reporting an intention to fully comply falling from around 90% in the *Control* and *Unexposed* groups to about 84% in the *Treatment* group.

Taken together, these patterns suggest that while the chatbot does not meaningfully alter observable patient behavior, it may negatively affect patients' perceptions of the clinical encounter and their interactions with physicians. One possibility is that when patients act more knowledgeable, they receive less polite responses from physicians, a phenomenon documented by Currie et al. (2011). Another possibility is that access to additional health information prior to the visit changes patients' expectations about the consultation, including which questions should be addressed or how much explanation they anticipate from the physician. When these expectations are not fully met, patients may report lower satisfaction or perceive communication as less smooth, even if clinical decisions and

follow-through remain largely unchanged. This interpretation is consistent with evidence from Ayers et al. (2023), which shows that in response to patient questions posted on social media, AI chatbot replies are rated more favorably than physician responses overall and on measures of empathy and clarity.

While the patient survey provides insight into the consultation experience from the patient's perspective, it reflects a selected sample of respondents. We therefore complement this evidence with a physician survey that captures the consultation from the physicians' perspective. The physician survey achieves a high completion rate and thus does not suffer from the same selection concerns, although inference is limited by the smaller sample size of physicians rather than visits. Accordingly, we interpret the physician survey results as descriptive and suggestive evidence on potential mechanisms.

Figure 5 summarizes physicians' perceptions of patient behavior along four dimensions, comparing responses from physicians in the *Exposed* and *Unexposed* groups.[14] In Panel (a), physicians in the *Exposed* group are more likely to report that patients describe their symptoms clearly, with 71.0% of physicians reporting that patients clearly describe their symptoms, compared with 65.1% among *Unexposed* physicians. Panel (b) shows a similar pattern for perceived patient understanding of their condition, with 42.2% of *Exposed* physicians reporting good understanding, compared with 31.2% in the *Unexposed* group. Panel (c) indicates smoother perceived physician–patient communication in the *Exposed* group, with 66.1% reporting fairly smooth communication, relative to 58.7% among *Unexposed* physicians. Taken together, the patient and physician surveys reveal an important divergence in perspectives. From the patient's perspective, access to the chatbot is associated with worse perceived communication and lower satisfaction with the visit, whereas from the physician's perspective, patients appear better prepared and communicate more clearly.

Despite these opposing assessments of communication quality, both patients and physicians report lower adherence or intended compliance with medical advice among treated patients. Panel (d) of Figure 5 shows a 5 percentage point reduction in physician-reported patient adherence in the *Exposed* group. This pattern mirrors the patient survey findings and highlights a potential tension between improved patient preparedness and reduced adherence to physicians' recommendations. This is

[14] As a sanity check, we elicit physicians' beliefs about the share of their patients who used generative AI tools prior to the consultation for their health condition during the experimental month. The median perceived share is 20% among physicians in the *Exposed* group and 10% among those in the *Unexposed* group.

consistent with prior evidence showing that patients with access to expertise exhibit lower adherence (Finkelstein et al., 2022). From a policy perspective, this pattern underscores a potential trade-off: while AI tools may enhance patient understanding and improve aspects of the consultation process, they may also alter authority dynamics in ways that affect adherence, with potential implications for both care delivery and health outcomes.

7 Potential Mechanisms

The intervention may affect clinical decisions through multiple channels. In particular, improving patient access to medical information may change patients' treatment requests and shift the balance of authority within the consultation. We examine evidence consistent with each mechanism below.

7.1 Changes in Patient Treatment Requests

One potential mechanism is that the chatbot changes patients' treatment requests. In outpatient care, overtreatment can arise when imperfectly informed patients request unnecessary treatment and physicians comply to avoid conflict (Lopez et al., 2022). By providing patients with structured symptom guidance and general medical information, the chatbot may reduce requests for unnecessary medications while increasing interest in diagnostic clarification.

Analysis of the conversation logs between patients and the generative AI model provides suggestive evidence consistent with this mechanism. In 81.73% of conversations, the chatbot advises patients to monitor their symptoms or notes that their conditions may resolve on their own without medication, which may reduce patients' requests for drug treatment. Traditional Chinese Medicine is mentioned in 11.89% of conversations, yet the chatbot rarely endorses its use. Antibiotics are discussed in 14.99% of conversations, and in 85.6% of those cases the chatbot explicitly discourages their use. By contrast, 34.43% of conversations include discussion of specific diagnostic tests. Taken together, these patterns suggest that the intervention may shift patients' treatment requests away from immediate medication use, particularly treatments with limited clinical support, and toward diagnostic clarification and symptom monitoring.

This mechanism generates a testable prediction. If the reduction in prescribing and increase in testing are driven by changes in patients' treatment requests, the effects should be more pronounced

among physicians who are responsive to patient input. When physicians place greater emphasis on their own authority and assign less weight to patient requests, improvements in patient information are less likely to affect prescribing and testing decisions. In contrast, when physicians are more receptive to patient input, shifts in patients' treatment requests should translate more directly into changes in clinical decisions.

To examine this prediction, we conduct a heterogeneity analysis using two measures from the physician survey. We asked physicians whether they view medical knowledge acquired by patients from external sources as helpful and whether patients should strictly follow physicians' recommendations. These measures capture variation in physician openness and authority orientation. The results are consistent with this mechanism. Panel A of Table 6 shows that the reduction in medication prescribing and increase in diagnostic testing are substantially larger among physicians who report being highly open to patient-acquired health knowledge. In contrast, Panel B shows that the treatment effects are close to zero among physicians who place strong emphasis on physician authority. These patterns suggest that better-informed patients are less likely to request unnecessary medications, and such requests translate into prescriptions primarily when physicians are receptive to patient input.

7.2 Shifts in Authority within the Consultation

The intervention may also affect the balance of authority within the consultation. By narrowing informational asymmetries, the chatbot may reduce patients' reliance on physician expertise and make them less deferential to physicians' recommendations (Currie et al., 2011). Survey responses from both patients and physicians provide suggestive evidence consistent with this interpretation. Treated patients report lower intended compliance with medical advice, and physicians interacting with treated patients also perceive lower patient adherence. These patterns suggest that AI-assisted patients rely less on physician recommendations when making treatment decisions.

Such changes in the consultation dynamic may limit physicians' discretion in treatment decisions. If improved patient information reduces deference to physician authority, the effects of the intervention should be strongest when patients interact with physicians who tend to practice more intensively.[15]

[15]One potential explanation for heterogeneity in prescribing intensity is financially motivated physician-induced demand. However, such incentives are limited in our institutional setting. Physicians' salaries and bonuses are not tied to drug sales, and hospitals operate under a zero-markup policy for medications (Yi et al., 2015; Yip et al., 2019; Fang

In such cases, better-informed patients may be more willing to question aggressive prescribing and steer treatment decisions toward more conservative care.

We first test this prediction using physicians' self-reported treatment attitudes. Table 7 examines whether the treatment effects are larger among physicians who express more favorable views toward prescribing specific medications. In the physician survey, we elicit attitudes toward the use of Traditional Chinese Medicine, antibiotics, and opioids, and classify physicians according to whether they report more favorable or more conservative views toward each category.

The results are consistent with this interpretation. In column (1) of Table 7, the interaction between treatment assignment and a more pro-TCM attitude is negative and statistically significant, indicating that the decline in TCM prescribing is substantially larger among physicians who are more inclined to prescribe TCM. A similar, though imprecisely estimated, pattern appears for antibiotics in column (2). We find no differential effects for opioids, consistent with their low baseline prescribing rates in China.

We next examine heterogeneity using physicians' revealed prescribing behavior prior to the experiment. Specifically, we classify physicians according to whether their prescribing rates for TCM, antibiotics, and opioids between January and May 2025, before the start of the experiment, exceed the median within their department. This measure captures baseline treatment intensity based on actual prescribing patterns rather than self-reported attitudes. Table 8 reports the results. The reduction in medication use is significantly larger among physicians with higher baseline prescribing intensity. In particular, the decline in TCM prescribing is concentrated among physicians whose pre-experiment TCM prescribing rates were above the departmental median. A similar pattern appears for antibiotics, while we find no differential effects for opioids, again consistent with their low baseline prescribing rates in China.

Taken together, these patterns suggest that when patients interact with physicians who tend to prescribe more intensively, improved patient information allows them to question such recommendations and constrain aggressive prescribing. This evidence is consistent with improved patient

et al., 2021). By contrast, physicians' bonuses may be linked to revenues from testing and procedures. If prescribing were primarily driven by financial incentives (e.g., Gruber and Owings, 1996; Liu et al., 2009; Iizuka, 2012; Lu, 2014; Fang et al., 2021), one would expect excess testing rather than excess prescribing in the absence of the intervention, and a reduction in testing rather than the increase we observe following the intervention. These institutional features suggest that variation in prescribing intensity is more likely to reflect differences in physicians' practice styles rather than direct financial incentives.

information shifting authority within the consultation and limiting physicians' discretion in treatment decisions.

8 Conclusion

This paper provides the first large-scale randomized evidence on how patient access to generative AI affects healthcare delivery in a real-world clinical setting. We show that providing patients with access to a generative AI chatbot prior to outpatient visits meaningfully alters clinical practice. Medication prescribing declines, particularly for Traditional Chinese Medicine and antibiotics, while diagnostic testing increases, without raising total medical expenditures. Short-term revisit rates decline marginally, suggesting no deterioration in care quality. At the same time, we find no evidence of spillovers in physicians' behavior toward patients without chatbot access, indicating that the effects operate primarily through direct patient use rather than persistent changes in physician practice.

Our mechanism analyses suggest that these changes arise through two related channels. First, improved patient information changes patients' treatment requests, reducing requests for unnecessary medications while increasing interest in diagnostic clarification. Second, improved information appears to shift the balance of authority within the consultation, limiting physicians' discretion in treatment decisions, particularly among physicians with higher baseline prescribing intensity. We further argue that the specific pattern of clinical change we document—reduced prescribing and increased diagnostic testing—is consistent with the defensive, liability-driven guardrails encoded in AI training by developers, which systematically caution against medication use and encourage diagnostic clarification. This points to a broader phenomenon: when AI systems are deployed as consumer-facing information tools at scale, the norms encoded in their training propagate into real clinical decisions, with the potential to shift aggregate outcomes such as population-level healthcare utilization.

At the same time, our survey results reveal a tension between the use of generative AI tools and the physician–patient relationship. While physicians report clearer communication and better patient preparation when interacting with AI-assisted patients, patients themselves report lower satisfaction and reduced intended adherence to medical advice. These patterns suggest that generative AI may

reshape expectations about clinical encounters and alter the dynamics of the physician–patient interaction in ways that affect compliance.

Overall, our findings highlight both the promise and the limits of generative AI as a tool for reducing information asymmetries in healthcare. Generative AI can improve patients' access to medical information and empower them to constrain physicians' discretion in treatment decisions. However, these shifts in authority within the clinical encounter may also introduce new frictions in physician–patient relationships. More broadly, our study underscores the importance of understanding how information technology such as generative AI, together with the norms encoded in its training, reshape interactions between experts and the clients they serve.

References

Abaluck, Jason, Robert Pless, Nirmal Ravi, Anja Sautmann, and Aaron Schwartz, "Does LLM Assistance Improve Healthcare Delivery? An Evaluation Using On-site Physicians and Laboratory Tests," *National Bureau of Economic Research Working Paper*, 2026.

Angerer, Silvia, Daniela Glätzle-Rützler, Wanda Mimra, Thomas Rittmannsberger, and Christian Waibel, "The Value of Rating Systems in Credence Goods Markets," *The Economic Journal*, 2026, p. ueag011.

Arrow, Kenneth J, "Uncertainty and the Welfare Economics of Medical Care," *The American Economic Review*, 1963, *53* (5), 941–973.

Artmann, Elisabeth, Hessel Oosterbeek, and Bas Van Der Klaauw, "Do Doctors Improve the Health Care of Their Parents? Evidence from Admission Lotteries," *American Economic Journal: Applied Economics*, 2022, *14* (3), 164–184.

Ayers, John W, Adam Poliak, Mark Dredze, Eric C Leas, Zechariah Zhu, Jessica B Kelley, Dennis J Faix, Aaron M Goodman, Christopher A Longhurst, Michael Hogarth et al., "Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum," *JAMA internal medicine*, 2023, *183* (6), 589–596.

Badinski, Ivan, Amy Finkelstein, Matthew Gentzkow, and Peter Hull, "Geographic Variation in Healthcare Utilization: The Role of Physicians," *National Bureau of Economic Research Working Paper*, 2023.

Balafoutas, Loukas, Adrian Beck, Rudolf Kerschbamer, and Matthias Sutter, "What Drives Taxi Drivers? A Field Experiment on Fraud in a Market for Credence Goods," *Review of Economic Studies*, 2013, *80* (3), 876–891.

_ , Rudolf Kerschbamer, and Matthias Sutter, "Second-Degree Moral Hazard in a Real-World Credence Goods Market," *The Economic Journal*, 2017, *127* (599), 1–18.

Bertomeu, Jeremy, Yupeng Lin, Yibin Liu, and Zhenghui Ni, "The Impact of Generative AI on Information Processing: Evidence from the Ban of ChatGPT in Italy," *Journal of Accounting and Economics*, 2025, p. 101782.

Bini, Pietro, Lin William Cong, Xing Huang, and Lawrence J Jin, "Behavioral Economics of AI: LLM Biases and Corrections," *Available at SSRN 5213130*, 2025.

Bowen, Donald E, Luke CD Stein, S McKay Price, and Ke Yang, "Measuring and Mitigating Racial Disparities in LLMs: Evidence from a Mortgage Underwriting Experiment," *Working Paper*, 2025.

Bronnenberg, Bart J, Jean-Pierre Dubé, Matthew Gentzkow, and Jesse M Shapiro, "Do Pharmacists Buy Bayer? Informed Shoppers and the Brand Premium," *The Quarterly Journal of Economics*, 2015, *130* (4), 1669–1726.

Brynjolfsson, Erik, Danielle Li, and Lindsey Raymond, "Generative AI at Work," *The Quarterly Journal of Economics*, 2025, *140* (2), 889–942.

Busse, Meghan R, Ayelet Israeli, and Florian Zettelmeyer, "Repairing the Damage: The Effect of Price Knowledge and Gender on Auto Repair Price Quotes," *Journal of Marketing Research*, 2017, *54* (1), 75–95.

Chen, Stacey H, Jennjou Chen, Hongwei Chuang, and Tzu-Hsin Lin, "Physicians Treating Physicians: Relational and Informational Advantages in Treatment and Survival," *Journal of Labor Economics*, 2025, *43* (1), 15–46.

Chen, Yiqun, Petra Persson, and Maria Polyakova, "The Roots of Health Inequality and the Value of Intrafamily Expertise," *American Economic Journal: Applied Economics*, 2022, *14* (3), 185–223.

Choi, Jonathan H, Amy B Monahan, and Daniel Schwarcz, "Lawyering in the Age of Artificial Intelligence," *Minn. L. Rev.*, 2024, *109*, 147.

Cui, Zheyuan Kevin, Mert Demirer, Sonia Jaffe, Leon Musolff, Sida Peng, and Tobias Salz, "The Effects of Generative AI on High Skilled Work: Evidence from Three Field Experiments with Software Developers," *Available at SSRN 4945566*, 2024.

Currie, Janet, W Bentley MacLeod, and Kate Musen, "Doctor Decision-Making and Patient Outcomes," *Journal of Economic Literature*, 2026, *64* (1), 141–194.

_ , Wanchuan Lin, and Juanjuan Meng, "Social Networks and Externalities from Gift Exchange: Evidence from a Field Experiment," *Journal of Public Economics*, 2013, *107*, 19–30.

_ , _ , and _ , "Addressing Antibiotic Abuse in China: An Experimental Audit Study," *Journal of Development Economics*, 2014, *110*, 39–51.

_ , _ , and Wei Zhang, "Patient Knowledge and Antibiotic Abuse: Evidence from an Audit Study in China," *Journal of Health Economics*, 2011, *30* (5), 933–949.

Cutler, David, Jonathan S Skinner, Ariel Dora Stern, and David Wennberg, "Physician Beliefs and Patient Preferences: A New Look at Regional Variation in Health Care Spending," *American Economic Journal: Economic Policy*, 2019, *11* (1), 192–221.

Darby, Michael R and Edi Karni, "Free Competition and the Optimal Amount of Fraud," *The Journal of Law and Economics*, 1973, *16* (1), 67–88.

Das, Jishnu, Alaka Holla, Aakash Mohpal, and Karthik Muralidharan, "Quality and Accountability in Health Care Delivery: Audit-Study Evidence from Primary Care in India," *American Economic Review*, 2016, *106* (12), 3765–3799.

Dell'Acqua, Fabrizio, Edward McFowland III, Ethan R Mollick, Hila Lifshitz-Assaf, Katherine Kellogg, Saran Rajendran, Lisa Krayer, François Candelon, and Karim R Lakhani, "Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality," *Harvard Business School Technology & Operations Mgt. Unit Working Paper*, 2023, (24-013).

Dillon, Eleanor W, Sonia Jaffe, Nicole Immorlica, and Christopher T Stanton, "Shifting Work Patterns with Generative AI," *National Bureau of Economic Research Working Paper*, 2025.

Dranove, David, Daniel Kessler, Mark McClellan, and Mark Satterthwaite, "Is More Information Better? The Effects of "Report Cards" on Health Care Providers," *Journal of Political Economy*, 2003, *111* (3), 555–588.

Dulleck, Uwe and Rudolf Kerschbamer, "On Doctors, Mechanics, and Computer Specialists: The Economics of Credence Goods," *Journal of Economic literature*, 2006, *44* (1), 5–42.

Dupas, Pascaline, "Do Teenagers Respond to HIV Risk Information? Evidence from a Field Experiment in Kenya," *American Economic Journal: Applied Economics*, 2011, *3* (1), 1–34.

_ , "Health Behavior in Developing Countries," *Annual Review of Economics*, 2011, *3* (1), 425–449.

fai Fong, Yuk, Ting Liu, and Xiaoxuan Meng, "Trust Building in Credence Goods Markets," *American Economic Journal: Microeconomics*, 2022, *14* (1), 490–528.

Fang, Hanming, Xiaoyan Lei, Julie Shi, and Xuejie Yi, "Physician-Induced Demand: Evidence from China's Drug Price Zero-markup Policy," *National Bureau of Economic Research Working Paper*, 2021.

Filippas, Apostolos, John J Horton, and Benjamin S Manning, "Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus?," in "Proceedings of the 25th ACM Conference on Economics and Computation" 2024, pp. 614–615.

Finkelstein, Amy, Matthew Gentzkow, and Heidi Williams, "Sources of Geographic Variation in Health Care: Evidence from Patient Migration," *The Quarterly Journal of Economics*, 2016, *131* (4), 1681–1726.

_ , Petra Persson, Maria Polyakova, and Jesse M Shapiro, "A Taste of Their Own Medicine: Guideline Adherence and Access to Expertise," *American Economic Review: Insights*, 2022, *4* (4), 507–526.

Fleming-Dutra, Katherine E, Adam L Hersh, Daniel J Shapiro, Monina Bartoces, Eva A Enns, Thomas M File Jr, Jonathan A Finkelstein, Jeffrey S Gerber, David Y Hyun, Jeffrey A Linder et al., "Prevalence of Inappropriate Antibiotic Prescriptions among US Ambulatory Care Visits, 2010-2011," *JAMA*, 2016, *315* (17), 1864–1873.

Frakes, Michael, Jonathan Gruber, and Anupam Jena, "Is Great Information Good Enough? Evidence from Physicians as Patients," *Journal of Health Economics*, 2021, *75*, 102406.

Garg, Prashant and Thiemo Fetzer, "Artificial Intelligence Health Advice Accuracy Varies across Languages and Contexts," *arXiv preprint arXiv:2504.18310*, 2025.

Gottschalk, Felix, Wanda Mimra, and Christian Waibel, "Health Services as Credence Goods: A Field Experiment," *The Economic Journal*, 2020, *130* (629), 1346–1383.

Gruber, Jonathan and Maria Owings, "Physician Financial Incentives and Cesarean Section Delivery," *RAND Journal of Economics*, 1996, *27* (1).

Grytten, Jostein, Irene Skau, and Rune Sørensen, "Do Expert Patients Get Better Treatment than Others? Agency Discrimination and Statistical Discrimination in Obstetrics," *Journal of Health Economics*, 2011, *30* (1), 163–180.

Iizuka, Toshiaki, "Physician Agency and Adoption of Generic Pharmaceuticals," *American Economic Review*, 2012, *102* (6), 2826–2858.

Johnson, Erin M and M Marit Rehavi, "Physicians Treating Physicians: Information and Incentives in Childbirth," *American Economic Journal: Economic Policy*, 2016, *8* (1), 115–141.

Kaiser Family Foundation, "Poll: Most Who Use Artificial Intelligence Doubt AI Chatbots Provide Accurate Health Information," Technical Report, Kaiser Family Foundation 2024.

Kerschbamer, Rudolf, Daniel Neururer, and Matthias Sutter, "Credence Goods Markets, Online Information and Repair Prices: A Natural Field Experiment," *Journal of Public Economics*, 2023, *222*, 104891.

Kolstad, Jonathan T, "Information and Quality when Motivation Is Intrinsic: Evidence from Surgeon Report Cards," *American Economic Review*, 2013, *103* (7), 2875–2910.

Kotwani, Anita, Chand Wattal, Shashi Katewa, PC Joshi, and Kathleen Holloway, "Factors Influencing Primary Care Physicians to Prescribe Antibiotics in Delhi India," *Family Practice*, 2010, *27* (6), 684–690.

Kravitz, Richard L, Ronald M Epstein, Mitchell D Feldman, Carol E Franz, Rahman Azari, Michael S Wilkes, Ladson Hinton, and Peter Franks, "Influence of Patients' Requests for Direct-to-Consumer Advertised Antidepressants: A Randomized Controlled Trial," *JAMA*, 2005, *293* (16), 1995–2002.

Levitt, Steven D and Chad Syverson, "Market Distortions when Agents are Better Informed: The Value of Information in Real Estate Transactions," *The Review of Economics and Statistics*, 2008, *90* (4), 599–611.

Li, Yongbin, Jing Xu, Fang Wang, Bin Wang, Liqun Liu, Wanli Hou, Hong Fan, Yeqing Tong, Juan Zhang, and Zuxun Lu, "Overprescribing in China, Driven by Financial Incentives, Results in Very High Use of Antibiotics, Injections, and Corticosteroids," *Health affairs*, 2012, *31* (5), 1075–1082.

Linder, Jeffrey A, Jason N Doctor, Mark W Friedberg, Harry Reyes Nieva, Caroline Birks, Daniella Meeker, and Craig R Fox, "Time of Day and the Decision to Prescribe Antibiotics," *JAMA Internal Medicine*, 2014, *174* (12), 2029–2031.

Liu, Ya-Ming, Yea-Huei Kao Yang, and Chee-Ruey Hsieh, "Financial Incentives and Physicians' Prescription Decisions on the Choice Between Brand-Name and Generic Drugs: Evidence from Taiwan," *Journal of Health Economics*, 2009, *28* (2), 341–349.

Lopez, Carolina, Anja Sautmann, and Simone G Schaner, "Do Patients Value High-Quality Medical Care? Experimental Evidence from Malaria Diagnosis and Treatment," *National Bureau of Economic Research Working Paper*, 2024.

_ , _ , and Simone Schaner, "Does Patient Demand Contribute to the Overuse of Prescription Drugs?," *American Economic Journal: Applied Economics*, 2022, *14* (1), 225–260.

Lu, Fangwen, "Insurance Coverage and Agency Problems in Doctor Prescriptions: Evidence from a Field Experiment in China," *Journal of Development Economics*, 2014, *106*, 156–167.

Maclean, Johanna Catherine, Justine Mallatt, Christopher J Ruhm, and Kosali Simon, "Economic Studies on the Opioid Crisis: A Review," *National Bureau of Economic Research Working Paper*, 2020.

McGuire, Thomas G, "Physician Agency," *Handbook of Health Economics*, 2000, *1*, 461–536.

Mimra, Wanda, Alexander Rasch, and Christian Waibel, "Second Opinions in Markets for Expert Services: Experimental Evidence," *Journal of Economic Behavior & Organization*, 2016, *131*, 106–125.

Mullainathan, Sendhil, Markus Noeth, and Antoinette Schoar, "The Market for Financial Advice: An Audit Study," *National Bureau of Economic Research Working Paper*, 2012.

Noy, Shakked and Whitney Zhang, "Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence," *Science*, 2023, *381* (6654), 187–192.

Peng, Sida, Eirini Kalliamvakou, Peter Cihon, and Mert Demirer, "The Impact of AI on Developer Productivity: Evidence from GitHub Copilot," *arXiv preprint arXiv:2302.06590*, 2023.

Rasch, Alexander and Christian Waibel, "What Drives Fraud in a Credence Goods Market?–Evidence from a Field Study," *Oxford Bulletin of Economics and Statistics*, 2018, *80* (3), 605–624.

Schneider, Henry S, "Agency Problems and Reputation in Expert Services: Evidence from Auto Repair," *The Journal of Industrial Economics*, 2012, *60* (3), 406–433.

Spatharioti, Sofia Eleni, David M Rothschild, Daniel G Goldstein, and Jake M Hofman, "Comparing Traditional and LLM-based Search for Consumer Choice: A Randomized Experiment," *arXiv preprint arXiv:2307.03744*, 2023.

Stacey, Dawn, France Légaré, Krystina Lewis, Michael J Barry, Carol L Bennett, Karen B Eden, Margaret Holmes-Rovner, Hilary Llewellyn-Thomas, Anne Lyddiatt, Richard Thomson et al., "Decision Aids for People Facing Health Treatment or Screening Decisions," *Cochrane Database of Systematic Reviews*, 2017, (4).

_ , Krystina B. Lewis, Maureen Smith, Meg Carley, Robert Volk, Elisa E. Douglas, Lissa Pacheco-Brousseau, Jeanette Finderup, Janet Gunderson, Michael J. Barry, Carol L. Bennett, Paulina Bravo, Karina Steffensen, Amédé Gogovor, Ian D. Graham, Shannon E. Kelly, France Légaré, Henning Sondergaard, Richard Thomson, Logan Trenaman, and Lyndal Trevena,

"Decision aids for people facing health treatment or screening decisions," *Cochrane Database of Systematic Reviews*, 2024, *2024* (1), CD001431. Art. No.: CD001431. Update of: Cochrane Database Syst Rev. 2017;4:CD001431.

Tao, Xinge, Shuya Zhou, Kai Ding, Sairan Li, Yanzeng Li, Boyou Wu, Qirui Huang, Wangyue Chen, Muzi Shen, En Meng et al., "An LLM Chatbot to Facilitate Primary-to-Specialist Care Transitions: A Randomized Controlled Trial," *Nature Medicine*, 2026, pp. 1–9.

Wan, Peixing, Zigeng Huang, Wenjun Tang, Yulan Nie, Dajun Pei, Shaofen Deng, Jing Chen, Yizhi Zhou, Hongru Duan, Qingyu Chen et al., "Outpatient Reception via Collaboration between Nurses and a Large Language Model: a Randomized Controlled Trial," *Nature Medicine*, 2024, *30* (10), 2878–2885.

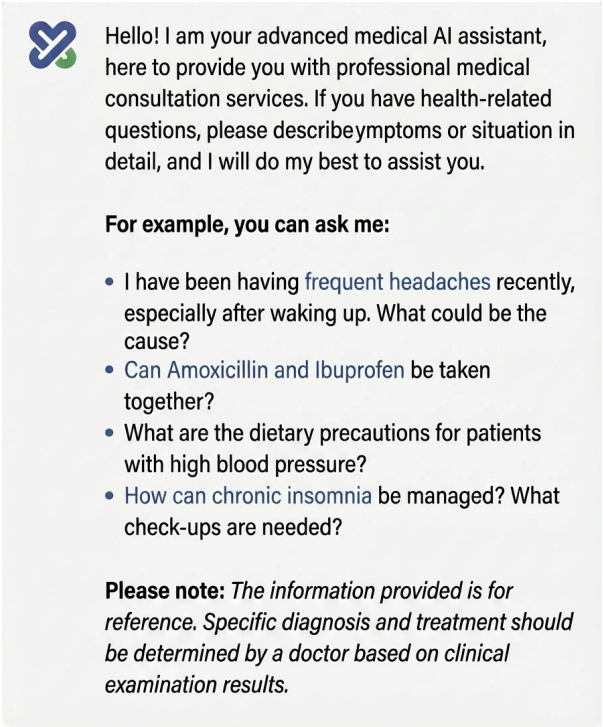
Yi, Hongmei, Grant Miller, Linxiu Zhang, Shaoping Li, and Scott Rozelle, "Intended and Unintended Consequences of China's Zero Markup Drug Policy," *Health Affairs*, 2015, *34* (8), 1391–1398.

Yip, Winnie, Hongqiao Fu, Angela T Chen, Tiemin Zhai, Weiyan Jian, Roman Xu, Jay Pan, Min Hu, Zhongliang Zhou, Qiulin Chen et al., "10 Years of Health-Care Reform in China: Progress and Gaps in Universal Health Coverage," *The Lancet*, 2019, *394* (10204), 1192–1204.

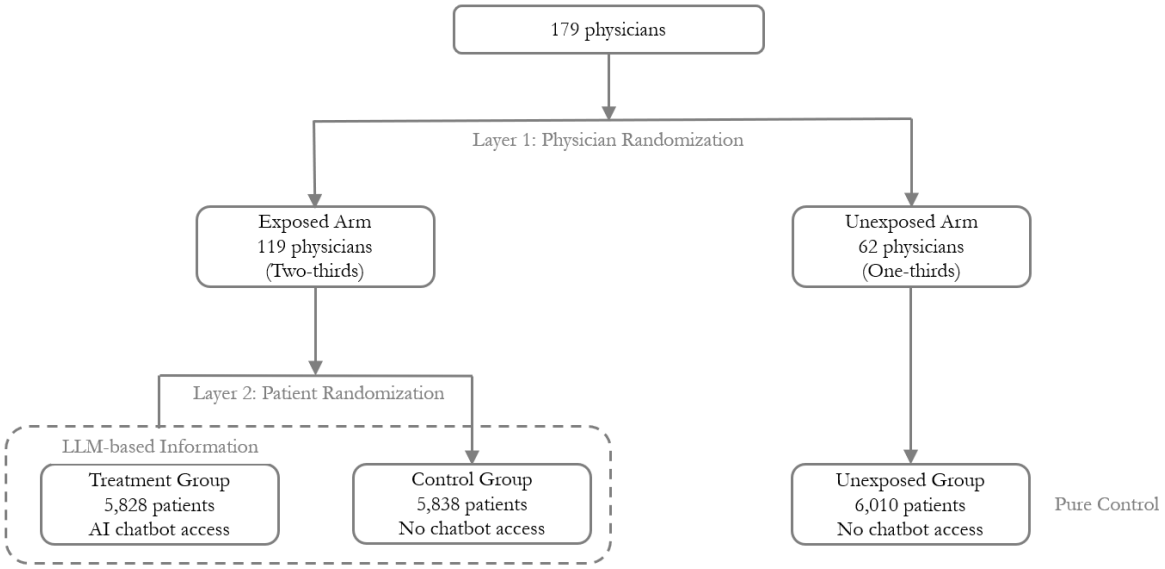
Figures and Tables



(a) Initial Interface

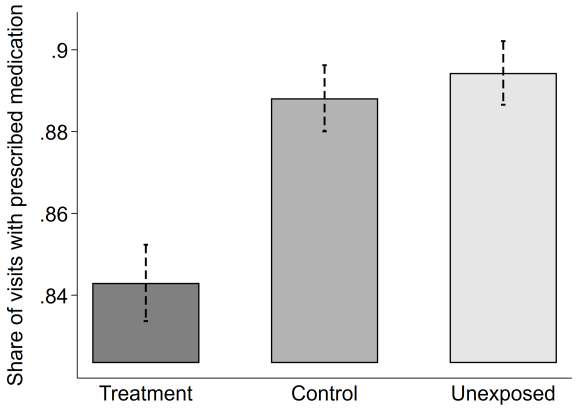


(b) Consultation Interface

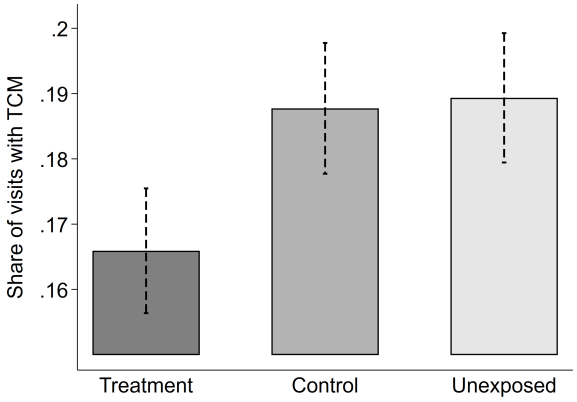


(c) Experimental Design

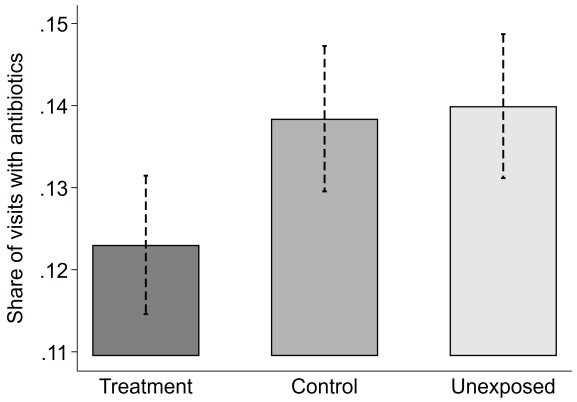
Figure 1. Medical AI Chatbot User Interfaces and Experimental Design Flowchart



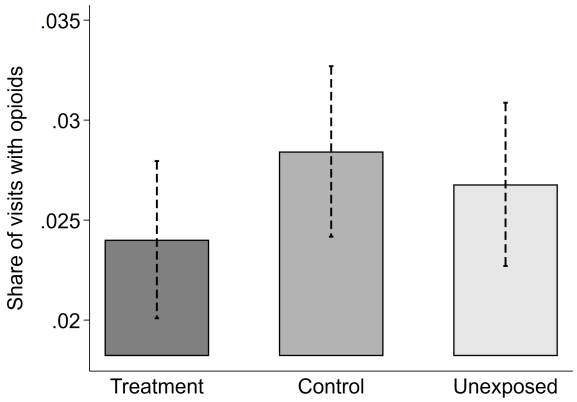
(a) Medication Prescription



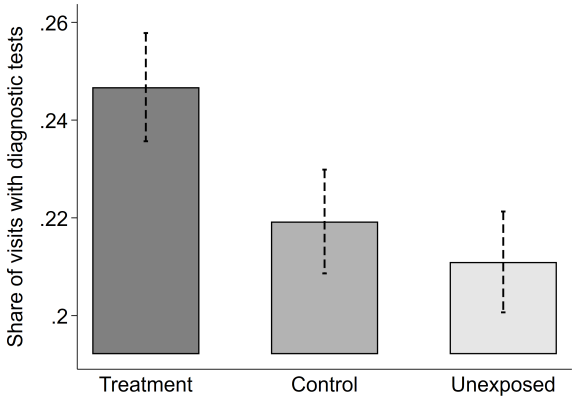
(b) TCM



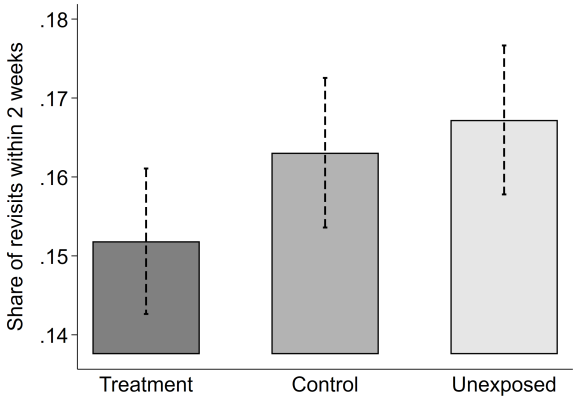
(c) Antibiotics



(d) Opioids



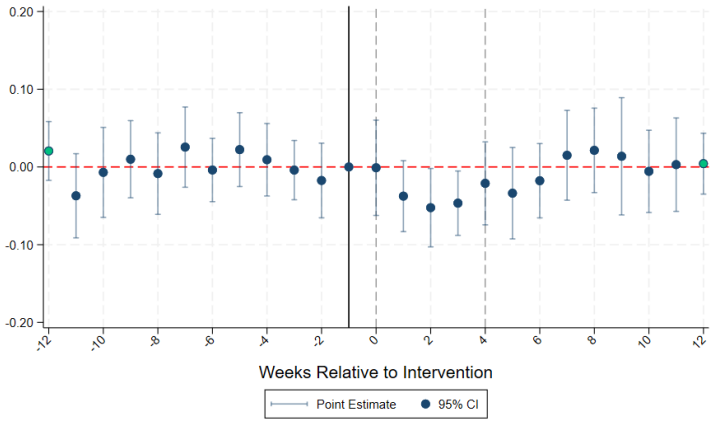
(e) Diagnostic tests



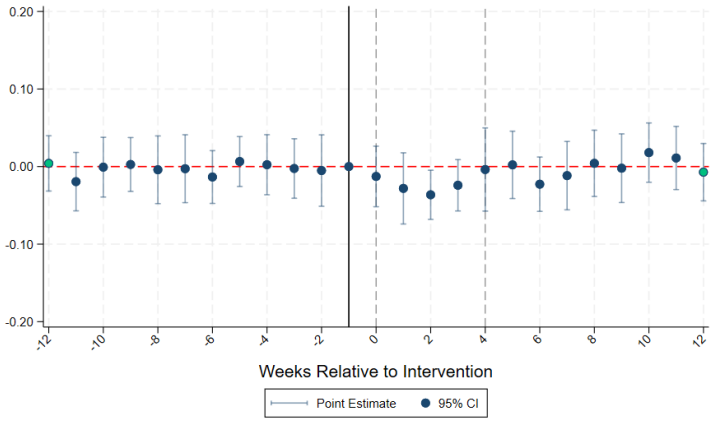
(f) Revisit within 2 weeks

Figure 2. Prescriptions and Clinical Practices by Treatment Status

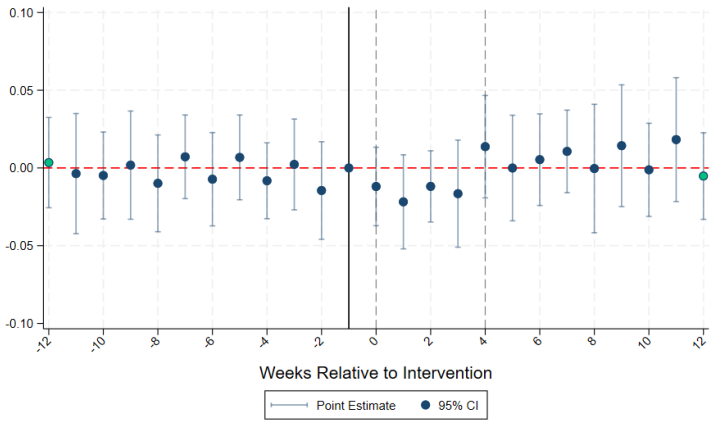
Notes: Bars represent group means with 95% confidence intervals (dashed lines). Treatment group received access to AI chatbot before consultations. Control group did not receive access but were treated by physicians exposed to AI-informed patients (exposed arm). Unexposed group did not receive access and were treated by physicians with no AI-exposed patients (unexposed arm). TCM = Traditional Chinese Medicine.



(a) Medication Prescription

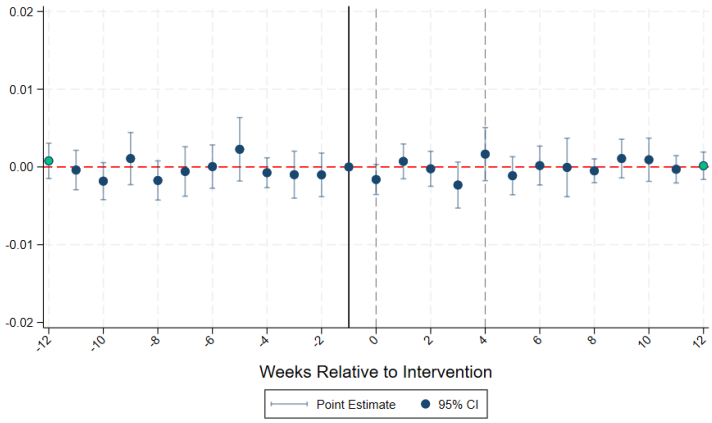


(b) TCM

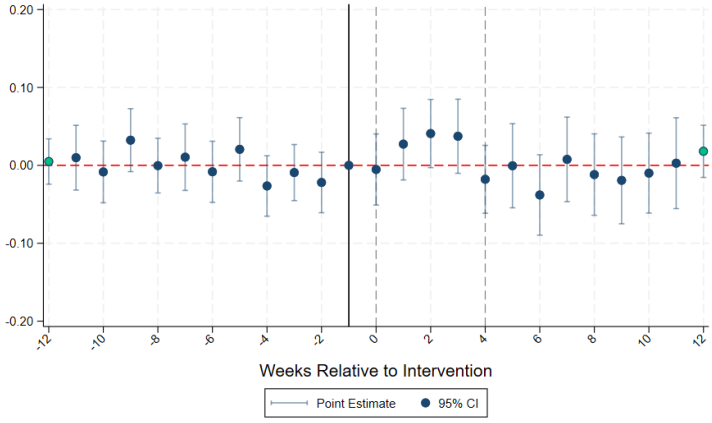


(c) Antibiotics

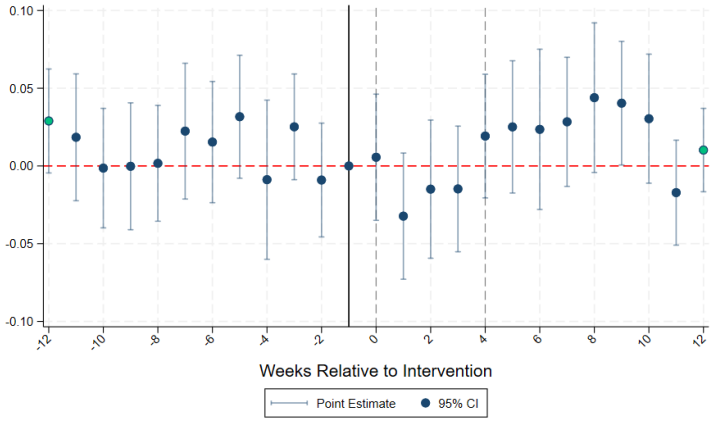
Figure 3. Dynamic Effects of Exposure to Treated Patients on Prescriptions and Clinical Practices



(d) Opioids



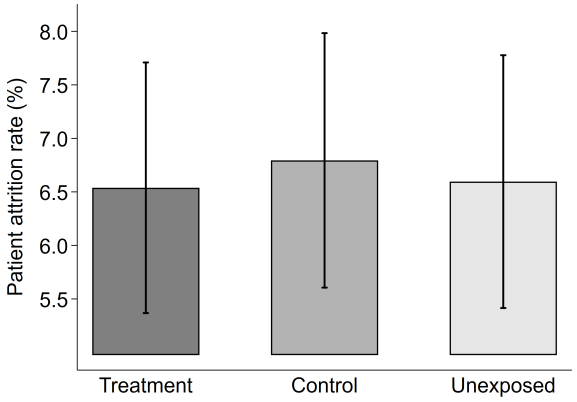
(e) Diagnostic tests



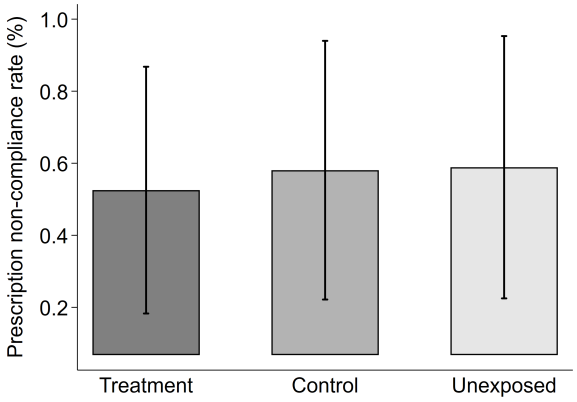
(f) Revisit within 2 weeks

Figure 3. Dynamic Effects of Exposure to Treated Patients on Prescriptions and Clinical Practices (Continued)

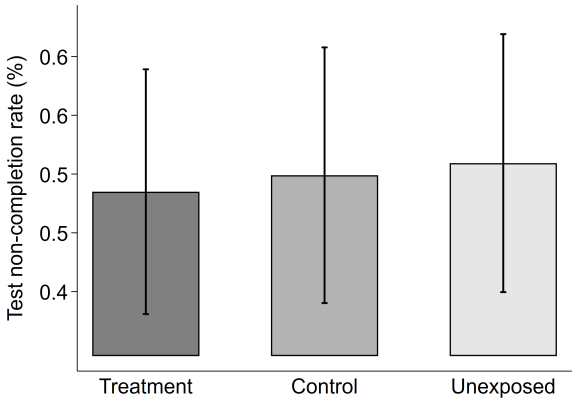
Notes: This figure presents the point estimates with 95% confidence intervals from equation (2). Data are aggregated to the physician-week level using all patient visits from 2025, weighted by the number of visits per physician per week. Relative weeks beyond −12 and +12 are grouped at the endpoints. The baseline period is Week −1 (May 26–June 1, 2025). Standard errors are clustered at the physician level.



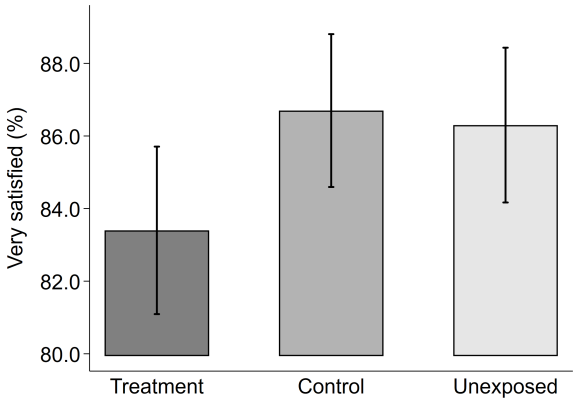
(a) Patient Attrition



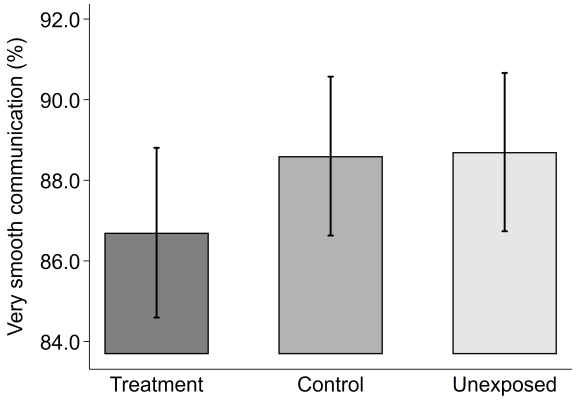
(b) Medication Non-Purchase



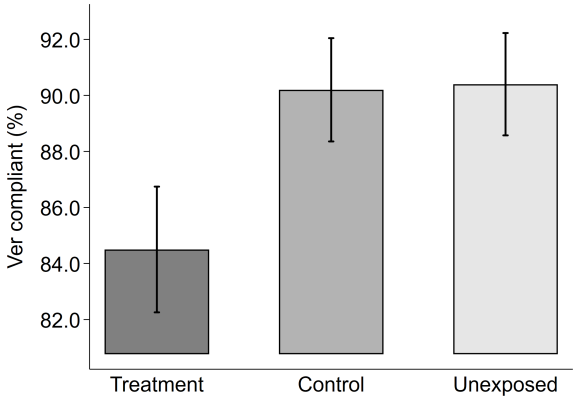
(c) Diagnostic Test Non-Completion



(d) Overall Satisfaction



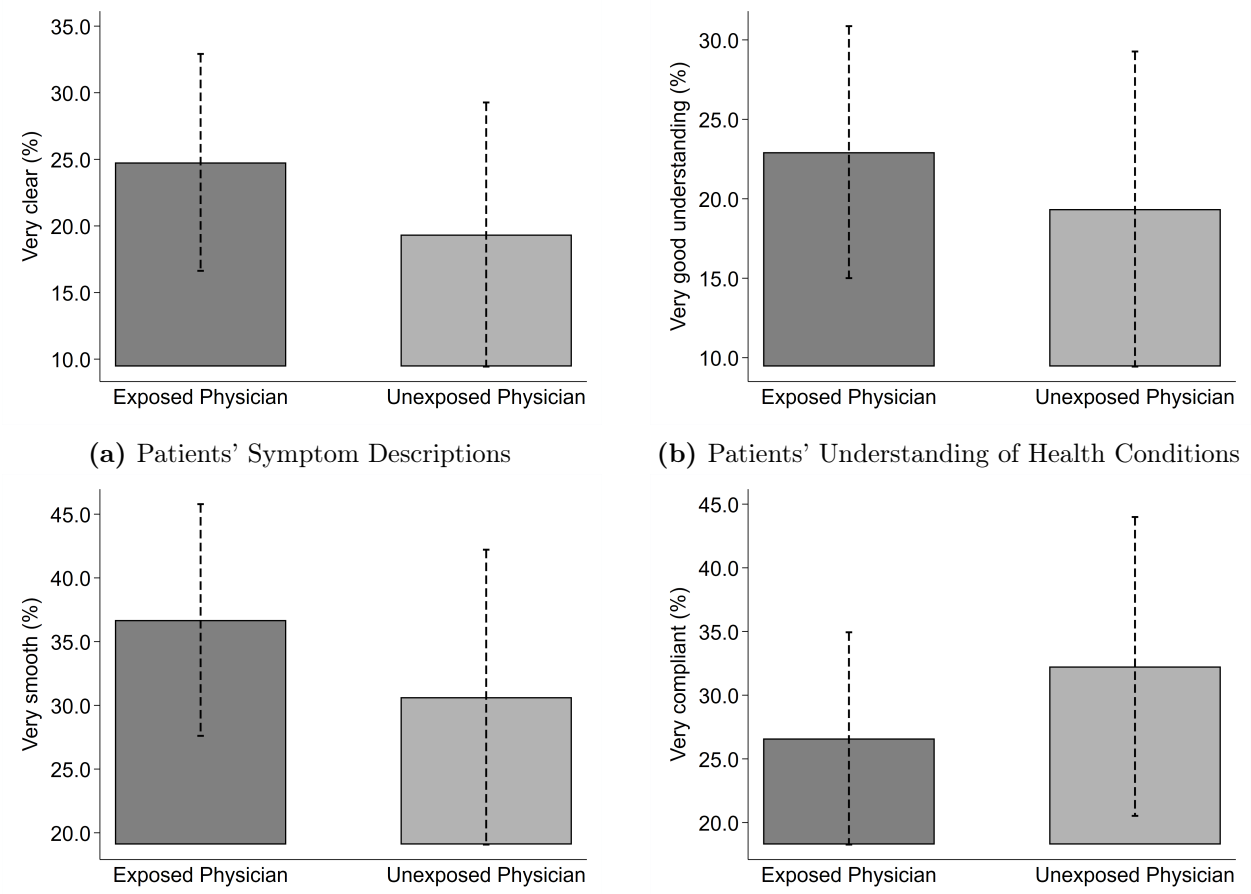
(e) Physician–Patient Communication



(f) Intended Compliance

Figure 4. Patient Behavior and Perceptions by Treatment Status

Notes: Panels (a)–(c) report outcomes constructed from administrative medical records and capture patient attendance and realized post-consultation behavior. Panels (d)–(f) report outcomes from a post-consultation patient survey, comparing respondents in the *Treatment*, *Control*, and *Unexposed* groups. Percentages denote the share of patients with the indicated outcome or selecting the indicated response category, with 95% confidence intervals (dashed lines).



(a) Patients' Symptom Descriptions

(b) Patients' Understanding of Health Conditions

(c) Physician–Patient Communication

(d) Patient Adherence to Medical Advice

Figure 5. Physicians' Perceptions of Patient Behaviors during Consultations

Notes: This figure reports physicians' survey responses collected after the conclusion of the experiment, comparing physicians randomized to the *Exposed* and *Unexposed* groups. Outcomes reflect physicians' perceptions of patient behavior during consultations. Percentages denote the share of physicians selecting the indicated response category, with 95% confidence intervals (dashed lines).

Table 1. Balance Test: Physician and Patient Characteristics

Panel A: Physician Characteristics

Variables	Exposed Arm		Unexposed Arm		Diff	p-value
	Mean	S.D.	Mean	S.D.		
Age	37.94	8.52	39.48	9.61	−1.54	0.29
Male	0.61	0.49	0.69	0.46	−0.09	0.25
Senior physician	0.30	0.46	0.29	0.46	0.01	0.90
College degree or above	0.91	0.29	0.92	0.27	−0.01	0.76
Specialty: Internal Medicine	0.34	0.48	0.34	0.48	−0.00	0.97
Specialty: Surgery	0.26	0.44	0.23	0.42	0.04	0.56
Specialty: Psychiatry	0.24	0.43	0.27	0.45	−0.04	0.62
Specialty: Emergency	0.05	0.22	0.08	0.27	−0.03	0.47
Specialty: Other Dept.	0.12	0.33	0.11	0.32	0.01	0.89
Physicians	117		62			

Panel B: Patient Characteristics

	Treatment		Control		Unexposed		Treat vs Ctrl		Treat vs Unexp	
	Mean	S.D.	Mean	S.D.	Mean	S.D.	Diff	p-value	Diff	p-value
Age	45.36	16.92	45.33	16.93	45.49	17.24	0.02	0.95	−0.15	0.63
Male	0.36	0.48	0.36	0.48	0.36	0.48	−0.00	0.96	0.00	0.93
Occupation: Farmer	0.44	0.50	0.44	0.50	0.44	0.50	0.00	0.98	−0.00	0.97
Occupation: Employed	0.24	0.43	0.23	0.42	0.24	0.43	0.01	0.20	−0.01	0.18
Occupation: Student	0.09	0.29	0.09	0.29	0.09	0.29	−0.00	0.92	0.00	0.89
Unemployed or retired	0.23	0.42	0.24	0.43	0.23	0.42	−0.01	0.21	0.01	0.19
First-time Visit	0.70	0.46	0.70	0.46	0.70	0.46	−0.00	0.95	0.00	0.95
Advance Booking Time (hours)	24.63	27.08	24.90	28.02	24.96	27.59	-0.28	0.55	-0.32	0.48
Patients	5,828		5,838		6,010					

Table 2. Within-Physician Intent-to-Treat Effects on Clinical Practice Patterns

	Dependent Variables					
	Prescribed medication	TCM	Antibiotics	Opioids	Diagnostic tests	Revisit
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: No Controls						
Treatment Group	-0.045^{***}	-0.022^{***}	-0.015^{**}	-0.004	0.027^{***}	-0.011^{*}
	(0.006)	(0.007)	(0.006)	(0.003)	(0.008)	(0.007)
Mean	0.87	0.18	0.13	0.03	0.23	0.16
Observations	11,666	11,666	11,666	11,666	11,666	11,666
Panel B: Patient Characteristics						
Treatment Group	-0.045^{***}	-0.022^{***}	-0.016^{**}	-0.004	0.027^{***}	-0.011^{*}
	(0.006)	(0.007)	(0.006)	(0.003)	(0.008)	(0.007)
Patient Char.	Y	Y	Y	Y	Y	Y
Mean	0.87	0.18	0.13	0.03	0.23	0.16
Observations	11,666	11,666	11,666	11,666	11,666	11,666

Notes: Patient characteristics include age, gender, occupation, and dummy for first visit. TCM = Traditional Chinese Medicine; WM = Western Medicine. Standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 3. Within-Physician Intent-to-Treat Effects on Healthcare Costs

	Dependent Variables				
	TCM costs (RMB) (1)	WM costs (RMB) (2)	Diagnostic costs (RMB) (3)	Other costs (RMB) (4)	Total costs (RMB) (5)
Panel A: No Controls					
Treatment Group	−1.686*	−5.276*	6.542	0.174	−0.245
	(0.944)	(2.939)	(5.661)	(3.293)	(7.263)
Mean	13.10	64.81	92.69	35.88	206.48
Observations	11,666	11,666	11,666	11,666	11,666
Panel B: Patient Characteristics					
Treatment Group	−1.712*	−5.265*	6.532	0.171	−0.274
	(0.944)	(2.940)	(5.661)	(3.294)	(7.266)
Patient Char.	Y	Y	Y	Y	Y
Mean	13.10	64.81	92.69	35.88	206.48
Observations	11,666	11,666	11,666	11,666	11,666

Notes: Patient characteristics include age, gender, occupation, and dummy for first visit. TCM = Traditional Chinese Medicine; WM = Western Medicine. Standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 4. Instrumental Variable Estimates: Treatment Effects on Clinical Practice Patterns

	Dependent Variables					
	Prescribed medication	TCM	Antibiotics	Opioids	Diagnostic tests	Revisit
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: No Controls						
Chatbot Usage	-0.265^{***}	-0.128^{***}	-0.090^{**}	-0.026	0.161^{***}	-0.066^{*}
	(0.037)	(0.041)	(0.037)	(0.017)	(0.046)	(0.039)
Mean	0.87	0.18	0.13	0.03	0.23	0.16
Observations	11,666	11,666	11,666	11,666	11,666	11,666
First Stage F-statistics			1200.244			
Panel B: Patient Characteristics						
Chatbot Usage	-0.264^{***}	-0.128^{***}	-0.091^{**}	-0.026	0.161^{***}	-0.066^{*}
	(0.037)	(0.041)	(0.037)	(0.017)	(0.046)	(0.039)
Patient Char.	Y	Y	Y	Y	Y	Y
Mean	0.87	0.18	0.13	0.03	0.23	0.16
Observations	11,666	11,666	11,666	11,666	11,666	11,666
First Stage F-statistics			1198.717			

Notes: Instrumental variable estimates using random assignment as instrument. Patient characteristics include age, gender, occupation, and dummy for first visit. TCM = Traditional Chinese Medicine; WM = Western Medicine. Standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 5. Instrumental Variable Estimates: Treatment Effects on Healthcare Costs

	Dependent Variables				
	TCM costs (RMB) (1)	WM costs (RMB) (2)	Diagnostic costs (RMB) (3)	Other costs (RMB) (4)	Total costs (RMB) (5)
Panel A: No Controls					
Chatbot Usage	−9.884*	−30.933*	38.359	1.023	−1.436
	(5.542)	(17.228)	(33.186)	(19.306)	(42.583)
Mean	13.10	64.81	92.69	35.88	206.48
Observations	11,666	11,666	11,666	11,666	11,666
First Stage F-statistics			1200.244		
Panel B: Patient Characteristics					
Chatbot Usage	−10.044*	−30.878*	38.310	1.005	−1.607
	(5.542)	(17.234)	(33.192)	(19.311)	(42.602)
Patient Char.	Y	Y	Y	Y	Y
Mean	13.10	64.81	92.69	35.88	206.48
Observations	11,666	11,666	11,666	11,666	11,666
First Stage F-statistics			1198.717		

Notes: Instrumental variable estimates using random assignment as instrument. Patient characteristics include age, gender, occupation, and dummy for first visit. TCM = Traditional Chinese Medicine; WM = Western Medicine. Standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 6. Heterogeneous Treatment Effects by Physicians' Attitudes

	Dependent variables					
	Prescribed medication	TCM	Antibiotics	Opioids	Diagnostic tests	Revisit
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A. Physician Openness						
Treat × Openness	−0.069***	−0.037**	−0.017	−0.000	0.087***	−0.031
	(0.017)	(0.019)	(0.016)	(0.007)	(0.021)	(0.023)
Treat	−0.035***	−0.015	−0.011	0.000	0.007	0.004
	(0.007)	(0.008)	(0.007)	(0.003)	(0.009)	(0.009)
Patient Char.	Y	Y	Y	Y	Y	Y
Physician FE	Y	Y	Y	Y	Y	Y
Mean	0.87	0.18	0.13	0.03	0.23	0.16
Observations	11,666	11,666	11,666	11,666	11,666	11,666
Panel B. Physician Authority						
Treat × Authority	0.036**	0.021*	0.023**	0.004	−0.031	0.007
	(0.016)	(0.012)	(0.011)	(0.007)	(0.020)	(0.017)
Treat	−0.068***	−0.031***	−0.026***	−0.003	0.046***	−0.019
	(0.014)	(0.011)	(0.010)	(0.006)	(0.018)	(0.016)
Patient Char.	Y	Y	Y	Y	Y	Y
Physician FE	Y	Y	Y	Y	Y	Y
Mean	0.87	0.18	0.13	0.03	0.23	0.16
Observations	11,666	11,666	11,666	11,666	11,666	11,666

Notes: This table reports heterogeneous treatment effects by physicians' attitudes measured in the post-experiment survey. In Panel A, *Openness* equals one if the physician reported that medical knowledge acquired by patients from external sources is "very helpful" to physicians, and zero otherwise. In Panel B, *Authority* equals one if the physician strongly agreed with the statement "Patients should closely follow physicians' recommendations," and zero otherwise. All regressions control for patient characteristics (age, gender, occupation, and first-visit indicator) and include physician fixed effects. TCM denotes Traditional Chinese Medicine. Standard errors are reported in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table 7. Heterogeneous Treatment Effects by Physicians' Attitudes Toward Medications

	Dependent Variables		
	TCM	Antibiotics	Opioids
	(1)	(2)	(3)
Treat × Pro-TCM	−0.029**		
	(0.014)		
Treat × Pro-antibiotic		−0.017	
		(0.012)	
Treat × Pro-opioid			−0.001
			(0.007)
Treat	−0.013	−0.009	0.001
	(0.011)	(0.009)	(0.006)
Patient Char.	Y	Y	Y
Physician FE	Y	Y	Y
Mean	0.18	0.13	0.03
Observations	11,666	11,666	11,666

Notes: This table reports heterogeneous treatment effects by physicians' attitudes toward different medications. Pro-TCM equals one if the physician selected "Equal emphasis on TCM and Western medicine" when asked about the relationship between TCM and Western medicine, and zero if the physician selected "Completely Western medicine" or "Western medicine primary, TCM supplementary." Pro-antibiotic equals one if the physician selected "Use is generally appropriate" when asked about antibiotic use in China, and zero if the physician selected "Obviously overused" or "Slightly overused." Pro-opioid equals one if the physician selected "Use is generally appropriate" when asked about opioid analgesic use in China, and zero otherwise. All regressions control for patient characteristics (age, gender, occupation, and an indicator for first visit) and include physician fixed effects. TCM = Traditional Chinese Medicine. Standard errors are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

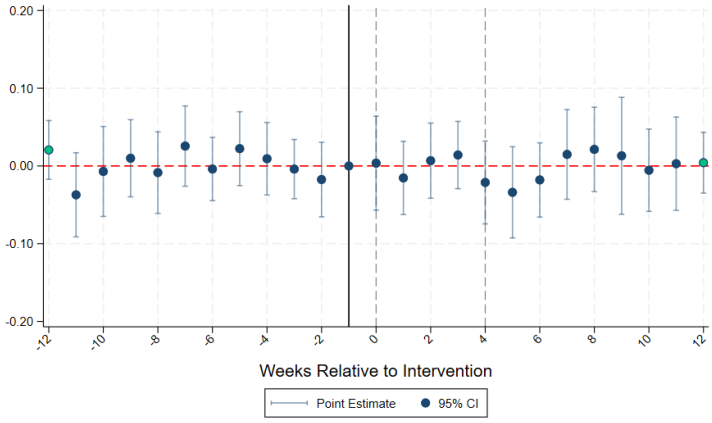
Table 8. Heterogeneous Treatment Effects by Physicians' Baseline Prescribing Intensity

	Dependent Variables		
	TCM	Antibiotics	Opioids
	(1)	(2)	(3)
Treat × High-TCM	−0.023**		
	(0.012)		
Treat × High-antibiotic		−0.019**	
		(0.009)	
Treat × High-opioid			−0.004
			(0.004)
Treat	−0.005	−0.003	0.004
	(0.009)	(0.008)	(0.005)
Patient Char.	Y	Y	Y
Physician FE	Y	Y	Y
Mean	0.18	0.13	0.03
Observations	11,666	11,666	11,666

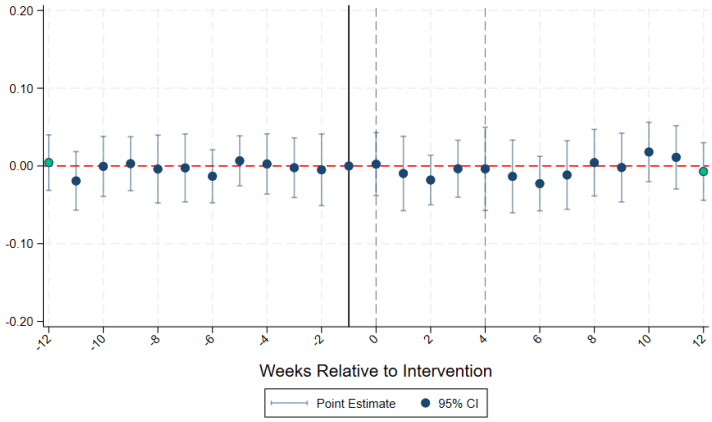
Notes: This table reports heterogeneous treatment effects based on whether a physician has above-average treatment intensity within their department prior to the experiment between January and May 2025. High TCM (High Antibiotic, High Opioid) is an indicator equal to one if the physician's pre-experiment mean prescribing rate for TCM (antibiotics, opioids) exceeds the department median, and zero otherwise. All regressions control for patient characteristics (age, gender, occupation, and an indicator for first visit) and include physician fixed effects. TCM = Traditional Chinese Medicine. Standard errors are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

ONLINE APPENDIX

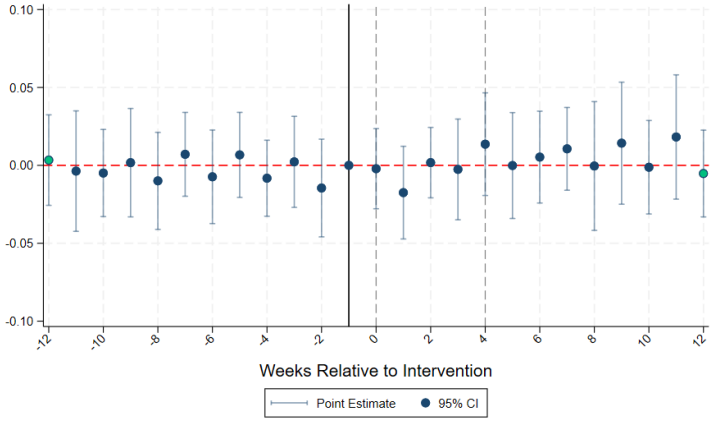
A Additional Figures and Tables



(a) Medication Prescription

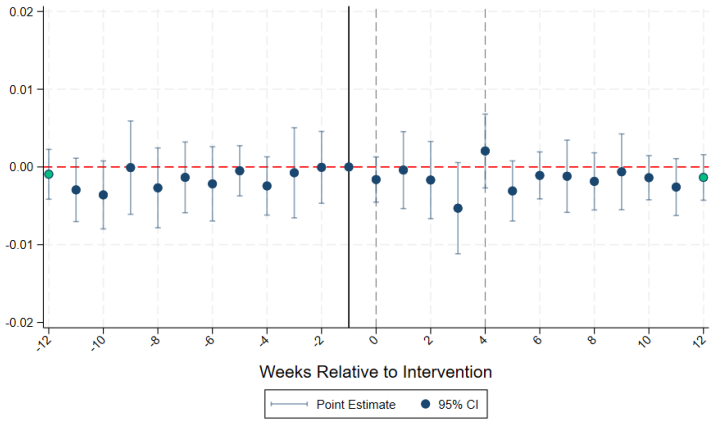


(b) TCM

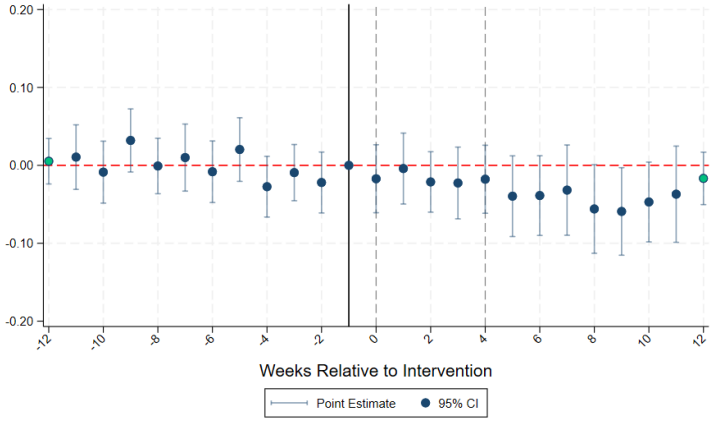


(c) Antibiotics

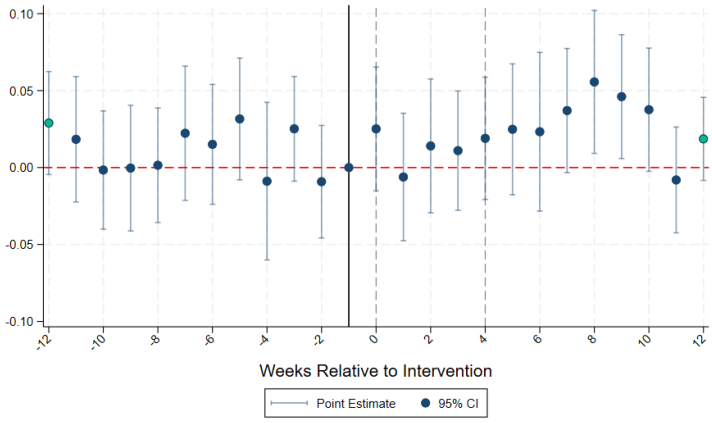
Figure A1. Dynamic Effects of Exposure to Treated Patients on Prescriptions and Clinical Practices, Excluding Treated Patients



(d) Opioids



(e) Diagnostic tests



(f) Revisit within 2 weeks

Figure A1. Dynamic Effects of Exposure to Treated Patients on Prescriptions and Clinical Practices, Excluding Treated Patients (Continued)

Notes: This figure presents the point estimates with 95% confidence intervals from equation (2). Data are aggregated to the physician-week level using patient visits from 2025, comparing those from control group and unexposed group, weighted by the number of visits per physician per week. Relative weeks beyond −12 and +12 are grouped at the endpoints. The baseline period is Week −1 (May 26–June 1, 2025). Standard errors are clustered at the physician level.

Table A1. Effect of Treatment on AI Adoption

	Whether Using AI Consulting Before Visit (1=Yes)	
	(1)	(2)
Treatment Group	0.171***	0.170***
	(0.005)	(0.005)
Patient Characteristics	No	Yes
Observations	11,666	11,666
R-squared	0.093	0.093

Notes: This table reports estimates from linear probability models of AI adoption on treatment assignment, using both the treatment group and control group subsamples. The dependent variable equals one if the patient used AI consulting before the visit, and zero otherwise. Column (1) includes only the treatment indicator, while Column (2) additionally controls for patient characteristics, including age, gender, occupation, and dummy for first visit. Standard errors are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A2. Patient Characteristics and AI Usage

	Whether Using AI Consulting Before Visit (1=Yes)			
	(1)	(2)	(3)	(4)
Age	−0.005***	−0.006***	−0.006***	−0.006***
	(0.000)	(0.000)	(0.000)	(0.000)
Male		0.075***	0.055***	0.055***
		(0.009)	(0.012)	(0.012)
Students			−0.001	−0.009
			(0.019)	(0.020)
Farmers			−0.010	−0.008
			(0.013)	(0.013)
Employed			0.067***	0.066***
			(0.013)	(0.014)
Unemployed/Retired			0.024	0.022
			(0.025)	(0.027)
First Visit				0.024**
				(0.010)
Observations	5,828	5,828	5,828	5,828

Notes: This table reports estimates from linear probability models examining the relationship between patient characteristics and AI usage. The dependent variable equals one if the patient used AI consulting before the visit, and zero otherwise. The sample is restricted to patients in the treatment group who were exposed to AI consulting. Columns (1)–(4) sequentially add patient characteristics, including age, gender, student status, occupation, employment status, and whether first visit. Standard errors are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.