Author/s:
Si, Y;Meng, Y;Chen, X;An, R;Mao, L;Li, B;Bateman, H;Zhang, H;Fan, H;Zu, J;Gong, S;Zhou, Z;Miao, Y;Fan, X;Chen, G

Title:
Quality safety and disparity of an AI chatbot in managing chronic diseases: simulated patient experiments

Date:
2025-12-01

Citation:
Si, Y., Meng, Y., Chen, X., An, R., Mao, L., Li, B., Bateman, H., Zhang, H., Fan, H., Zu, J., Gong, S., Zhou, Z., Miao, Y., Fan, X. & Chen, G. (2025). Quality safety and disparity of an AI chatbot in managing chronic diseases: simulated patient experiments. Npj Digital Medicine, 8 (1), pp.574-. https://doi.org/10.1038/s41746-025-01956-w.

Persistent Link:
https://hdl.handle.net/11343/363406

npj | digital medicine

Article

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-025-01956-w

Quality safety and disparity of an AI chatbot in managing chronic diseases: simulated patient experiments

Yafei Si[1], Yurun Meng[2], Xi Chen[3,4], Ruopeng An[5], Limin Mao[6], Bingqin Li[7], Hazel Bateman[8], Han Zhang[2], Hongbin Fan[2], Jiaqi Zu[2], Shaoqing Gong[9] ✉, Zhongliang Zhou[2], Yudong Miao[10], Xiaojing Fan[2] ✉ & Gang Chen[1]

The rapid development of AI solutions reveals opportunities to address the underdiagnosis and poor management of chronic conditions in developing settings. Using the method of simulated patients and experimental designs, we evaluate the quality, safety, and disparity of medical consultation with ERNIE Bot in China among 384 patient-AI trials. ERNIE Bot reached a diagnostic accuracy of 77.3%, correct drug prescriptions of 94.3%, but prescribed high rates of unnecessary medical tests (91.9%) and unnecessary medications (57.8%). Disparities were observed based on patient age and household economic status, with older and wealthier patients receiving more intensive care. Under standardized conditions, ERNIE Bot, ChatGPT, and DeepSeek demonstrated higher diagnostic accuracy but a greater tendency toward overprescription than human physicians. The results suggest the great potential of ERNIE Bot in empowering quality, accessibility, and affordability of healthcare provision in developing contexts, but also highlight critical risks related to safety and amplification of sociodemographic disparities.

The rapid development of AI solutions presents new opportunities to address urgent challenges in the prevention and management of non-communicable chronic diseases (NCDs) in low- and middle-income countries (LMICs). NCDs are now leading causes of mortality and morbidity worldwide, and are responsible for around 41 million (equivalent to 74%) global deaths in 2019[1]. Among these, cardiovascular diseases (CVDs) account for the largest proportion of deaths of 17.9 million people[2], followed by chronic respiratory diseases (4.1 million)[3]. Of all deaths, 77% are in LMICs[4]. The prevalence of chronic diseases is projected to continually rise in LMICs due to rapid population aging and lifestyle changes[5]. However, despite advancements in healthcare, many chronic conditions remain underdiagnosed and poorly managed[6,7], leading to excessive avoidable deaths. The heightened likelihood of underdiagnoses and poor chronic disease control is particularly concerning, especially among the more disadvantaged sub-populations living outside urban areas[8].

One of the most important contributors to the disproportionately high rates of underdiagnoses and poor management of NCDs in LMICs is the fact that a substantial proportion of primary care providers are less accessible, affordable, and qualified[6,9]. In rural areas of India, three out of every four people who seek primary healthcare use informal providers rather than licensed doctors or formal clinics[10]. Studies in China have found that only around one-quarter of NCD diagnoses and about one-third of medication prescriptions by primary care practitioners were deemed accurate and appropriate according to the standard clinical guidelines[11–15]. Developing countries like Ghana, Kenya, and Vietnam face the same challenge[16]. It is well documented that primary care practitioners outside the metropolis often lack the necessary resources, training, and support to diagnose and manage NCDs properly. This is further compounded by the severe shortage of essential healthcare facilities and personnel in more rural areas[17].

The emergence of generative AI presents new opportunities to improve healthcare accessibility. Unlike traditional clinical decision-support systems, many generative AI tools are freely available to the public and can provide health-related information across geographic and institutional boundaries. A growing body of research has demonstrated the effectiveness of generative

[1]Melbourne School of Population and Global Health, The University of Melbourne, Parkville, VIC, Australia. [2]School of Public Policy and Administration, Xi'an Jiaotong University, Xi'an, Shaanxi, China. [3]Department of Health Policy and Management, Yale School of Public Health, New Haven, CT, USA. [4]Department of Economics, Yale University, New Haven, CT, USA. [5]Silver School of Social Work, New York University, New York, NY, USA. [6]Center for Social Research in Health, UNSW Sydney, Kensington, NSW, Australia. [7]Social Policy Research Center, UNSW Sydney, Kensington, NSW, Australia. [8]UNSW Business School, UNSW Sydney, Kensington, NSW, Australia. [9]Luohe Medical College, Luohe, Henan, China. [10]College of Public Health, Zhengzhou University, Zhengzhou, Henan, China. ✉e-mail: gongshaoqingmd@163.com; emirada@163.com

AI in some CVDs[18] and orthopaedic diseases[19]. Patients, the general public, and health providers also favour the introduction of AI-powered healthcare services[20,21]. However, to our knowledge, the performance of AI tools in diagnosing and managing common NCDs in primary care settings remains scarce, especially in LMICs. In the absence of robust legal regulations and professional safeguards, there are rising concerns about the safety and ethical conduct of generative AI in patient interactions and clinical management processes[22]. This is particularly crucial in medical practices where patient-centred decision-making is based on providing accurate, reliable, and ethically safe information. To fill these gaps, this study evaluates the quality, safety, and disparity of two common NCD consultations provided by one of the most popular Chinese AI chatbots.

We used ERNIE (Enhanced Representation through kNowledge IntEgration) Bot, officially released in August 2023, one of the most popular AI chatbots in China (equivalent to ChatGPT used internationally), developed by Baidu. By April 2024, ERNIE Bot has recorded over 200 million active users in China, ranking first in comprehensive capabilities among China's large language models (LLMs), significantly outperforming the international average[23]. Compared to models like ChatGPT, ERNIE Bot has been uniquely developed and optimized for the Chinese language and cultural context. It is trained on a large-scale corpus that includes Chinese medical literature, regulatory documents, and clinical guidelines, making it particularly relevant for application in mainland China. ERNIE Bot is regarded as surpassing the general chat functions of an ordinary bot as a comprehensive platform to enable industry-quality AI-driven applications, including in healthcare, where ERNIE Bot has demonstrated its competency by passing the standard Chinese National Medical Licensing Examination[24]. However, its conversational capacity for medical consultations remains unclear. These features position ERNIE Bot as a locally optimized but globally significant case study for understanding AI in healthcare delivery within LMICs.

To create a realistic testing environment like daily medical consultations[25], the Simulated Patients (SPs) method was applied to the freely accessible version of ERNIE Bot 3.5. SPs are healthy individuals trained to systematically represent patients' key sociodemographic characteristics, medical histories, and biomedical status to facilitate patient-doctor communication during typical medical consultations. The SP method has been widely recognized as a "gold standard" for quality evaluation, particularly for primary healthcare in LMICs[16,26]. In this study, ERNIE Bot was designated as a doctor to offer medical consultation to the trained SPs from our previous studies[11,27,28]. Each SP presented a primary complaint (e.g., recent chest pain or shortness of breath) followed by predefined, consistent responses to all subsequent inquiries posed by ERNIE Bot. One complete SP-AI interaction was recorded as a trial. Two SPs were trained to use predefined response scripts[11,15] to ensure standardization across trials. The scripts were created based on clinical guidelines and validated by senior clinicians to ensure medical accuracy and completeness (see Supplementary Note 1). To ensure standardization, all SPs underwent structured training sessions, including script memorization, supervised rehearsals, and pilot trials.

Health disparities embedded in health systems, but learnt by AI chatbots, are also of interest in the study. First, gender, older age and socioeconomic status are the most common sources of health disparities. Second, like many LMICs, the urban-rural divide is a prominent feature in China. Especially, people with urban Hukou have better access to healthcare, education, housing, and employment opportunities than their rural counterparts[29]. Further, the Urban Employee Medical Insurance (UEMI) scheme caters to current or retired employees of government agencies, public or private enterprises, and institutions[30,31]. Compared with the Urban and Rural Resident Medical Insurance (URRMI) scheme catering to unemployed residents, the UEMI coverage is more comprehensive. Although the Hukou system and health insurance schemes are specific to China's policy context, they exemplify broader patterns of institutional inequality found in many LMICs—particularly those with segmented access to public services and unequal healthcare entitlements. The Hukou system captures institutionalized rural–urban disparities, which parallel urban–rural divides in access to health and social services in countries such as India, Indonesia, and Vietnam. Similarly, differential insurance coverage reflects disparities in financial protection and healthcare entitlements, an issue common in segmented or tiered health systems across LMICs.

Based on literature related to health disparities[27,29,32,33], six patient-level binary factors were used to assess variations in the AI-generated medical consultations, including i) gender (women vs. men), ii) age (65 years vs. 55 years old), iii) registered Hukou category (urban vs. non-urban), iv) permanent residence (urban vs. rural), v) household economic status (poor vs. rich), and vi) health insurance coverage (UEMI vs. URRMI). Following common physician practice, SPs revealed their gender and age information at the beginning of a consultation, the Hukou and residence information during a consultation, and the household economic status and health insurance coverage before the prescription of medications. We acknowledge that only the most apparent patient traits and levels are included in the study to simplify the experiments, although other factors are also important. These six patient-level factors were randomly assigned to SPs, resulting in 64 ($=2^6$) artificially manipulated scenarios.

The study is innovative in assessing the quality, safety, and disparity of medical consultations provided by AI chatbots and offers several methodological advantages. First, SPs allow for the creation of standardized scenarios in which disease conditions and relevant optimal care could be predefined, enabling subsequent direct comparison of AI-generated medical consultations against clinical guidelines. Second, SPs help ensure consistency in symptom presentation by reducing unobservable variations during doctor-patient communication. Third, SPs can record the entire consultation process and outcomes in detail, minimizing the recall bias inherent in traditional patient self-completed surveys. Fourth, because the background medical history and SP responses are standardized, except for deliberately varied patient traits, differences in outcomes should be attributable to the AI model rather than patient preferences or demands. Finally, the SP method avoids exposing real patients to potential harm during the evaluation of AI-generated consultations. In the study, we trained two SPs to present the common diseases, i.e., unstable angina and asthma. These conditions were selected due to their high burden among older adults and their prior use in existing literature[11,14,34].

Results

Data were collected from the beginning of December 2023 to April 2024 (see SP-AI interaction examples in Supplementary Note 3). Each disease condition was presented to ERNIE Bot three times to increase the trial's robustness. New chats were created to ensure the AI did not carry over its understanding from one trial to another. Out of the 64 independent SP scenarios, a final sample of 384 trials ($=2^6*2*3$) was generated, half ($n = 192$) for unstable angina and the other half for asthma. All six traits were orthogonally presented, with each trait level having 96 counts for each disease. All SP-AI trials successfully generated experimental data for the analysis. The Bot's responses were cross-validated with the most recent standard clinical guidelines to create four care quality and safety indicators.

Quality and safety indicators

Based on the 384 independent trials, overall ERNIE Bot completed 14.5% (95% CI: 13.8–15.3%) of the standard full checklist items and 20.3% (95% CI: 18.4–22.1%) of the standard essential checklist items. ERNIE Bot performed better for unstable angina (full-17.6%, 95% CI: 16.6–18.6%; essential- 35.4%, 95% CI: 33.6–37.2%) than for asthma (full-11.5%, 95% CI: 10.6–12.3%; essential-5.1%, 95% CI: 4.1–6.1%). The detailed checklist items for the two diseases are reported in Supplementary Note 4.

Despite such low-to-moderate levels of adherence to standard checklists, ERNIE Bot performed much more satisfactorily in the last two quality indicators overall, where correct diagnosis rates (77.3%, 95% CI: 73.1–81.5%) and correct medication prescription rates (94.3%, 95% CI: 91.9–96.6%) reached medium high-to-high. Here, ERNIE Bot performed equally well for unstable angina (correct diagnosis-76.6%, 95% CI:

Table 1 | Quality and safety performance by ERNIE (*N* = 384)

	Unstable angina (*n* = 192)		Asthma (*n* = 192)		Total (*n* = 384)	
	Mean	95% CI	Mean	95% CI	Mean	95% CI
Quality indicators						
% completion of the full checklists (0-1)	0.176	0.166, 0.186	0.115	0.106, 0.123	0.145	0.138, 0.153
% completion of the essential checklists (core, subset, 0-1)	0.354	0.336, 0.372	0.051	0.041, 0.061	0.203	0.184, 0.221
% Correct diagnosis (0-1)	0.766	0.705, 0.826	0.781	0.722, 0.840	0.773	0.731, 0.815
% Correct medication (0-1)	0.948	0.916, 0.980	0.938	0.903, 0.972	0.943	0.919, 0.966
Safety						
No. of tests requested	3.09	2.91, 3.27	3.10	2.90, 3.30	3.09	2.96, 3.23
% Unnecessary test requested (0-1)	0.969	0.944, 0.994	0.870	0.822, 0.918	0.919	0.892, 0.947
No. of medication prescribed	3.97	3.69, 4.26	4.21	3.91, 4.51	4.09	3.89, 4.30
% Inappropriate medication prescribed (0-1)	0.526	0.455, 0.597	0.630	0.561, 0.699	0.578	0.529, 0.628

Note. Means and 95% confidence intervals (CIs) for binary and continuous variables.

70.5% - 82.6%; correct prescription - 94.8%, 95% CI: 91.6–98.0%) and asthma (correct diagnosis - 78.1%, 95% CI: 72.2–84.0%; correct prescription - 93.8%, 95% CI: 90.3–97.2%).

Regarding safety, on average, ERNIE Bot had requested 3.09 (95% CI: 2.96–3.23; range 0–7) lab tests and prescribed 4.09 (95% CI: 3.89–4.30; range 0–14) medications. Among the 384 trials, ERNIE Bot reached alarmingly high rates of requesting unnecessary lab tests (91.9%, 95% CI: 89.2–94.7%) and prescribing inappropriate or even potentially harmful medications (57.8%, 95% CI: 52.9–62.8%). For both disease conditions, ERNIE Bot performed equally poorly. For unstable angina, ERNIE requested 3.09 (95% CI: 2.91–3.27; range 0–7) lab tests and prescribed 3.97 (95% CI: 3.69–4.26; range 0–12) medications. Among the 192 trials, 96.9% (95% CI: 94.4–99.4%) included unnecessary lab tests, and 52.6% (95% CI: 45.5–59.7%) included inappropriate medications. For asthma, ERNIE Bot requested 3.10 (95% CI: 2.90–3.30; range 0–6) lab tests and prescribed 4.21 (95% CI: 3.91–4.51; range 0–14) medications. Among the 192 trials, 87.0% (95% CI: 82.2–91.8%) included unnecessary lab tests, and 63.0% (95% CI: 56.1–69.9%) included inappropriate medications. The results are presented in Table 1.

Influences of the six patient-level factors: bivariable associations

As shown in Fig. 1, compared with SPs aged 55 years, for those aged 65 years ERNIE Bot achieved a relatively higher correct diagnosis rate (82.3% vs. 72.4%; $P = 0.021$) and prescribed marginally more medications (4.26 vs. 3.92; $P = 0.052$); compared with poorer patients, for wealthier ones, ERNIE Bot requested substantially more lab tests (3.26 vs. 2.93; $P = 0.009$) as well as prescribed more medications (4.45 vs. 3.73; $P < 0.001$); in Supplementary Fig. 4, compared to patients with URRMI health insurance coverage, for those covered by UEMI, ERNIE Bot prescribed relatively more medications (4.28 vs. 3.90; $P = 0.030$).

However, neither gender, residential Hukou registration, nor permanent residence of the SP patients had any differential influence over the eight performance indicators (Supplementary Figs. 1–3).

Influences of six patient-level factors: multivariable regression model estimation

As shown in Table 2, ERNIE Bot performed better in achieving higher correct diagnosis rates for the older SPs (aged 65 vs 55 years - 9.8%, 95% CI: 1.7% to 18.0%; $P < 0.05$). There was also a slightly increased possibility for ERNIE Bot to request more lab tests 0.323, 95% CI: 0.059 to 0.587; $P < 0.05$) and a substantially increased likelihood of prescribing more medications (0.724, 95% CI: 0.327 to 1.121; $P < 0.001$) for the wealthier SPs. Again, no performance variations were identified regarding SPs' gender, residential hukou registration, or permanent residential locations.

Further, compared with unstable angina, asthma was associated with significantly lower adherence to the checklist (complete- −6.2%, 95% CI: −7.5% to -4.9%; $P < 0.001$; essential- −32.1%, 95% CI: −34.5% to −29.6%; $P < 0.001$). Interestingly, asthma was linked with a reduced likelihood of unnecessary lab test requests (−10.2%, 95% CI: -16.2% to −4.2%; $P < 0.001$) on the one hand but an increased possibility of inappropriately prescribed medications (10.4%, 95% CI: 0.8% to 19.9%; $P < 0.05$) on the other.

Comparison of ERNIE Bot with China's primary care providers, ChatGPT, and DeepSeek

We conducted additional SP trials using the same clinical scenarios to benchmark ERNIE Bot's performance against healthcare providers and other popular LLMs. These included consultations with primary care providers in Luohe, China, and two advanced LLMs: ChatGPT-4o and DeepSeek R1 (Table 3). We deliberately set up eight SPs in February 2025, collecting 40 independent trials (20 for unstable angina and 20 for asthma) for each comparator under the same case scenarios and standardized protocols. This design allows for a controlled, internally valid comparison across human and AI-based care.

Primary care providers completed 26.1% (95% CI: 22.1–30.1%) of the full checklist items and 37.1% (95% CI: 27.9–46.4%) of the essential checklist items. They achieved relatively low rates of correct diagnosis (25.0%, 95% CI: 11.0–39.0%) and correct medication prescription (10.0%, 95% CI: 0.3–19.7%). In contrast, ChatGPT-4o completed 41.3% (95% CI: 39.3–43.4%) of the complete checklist and 53.3% (95% CI: 45.8–60.7%) of the essential checklist, achieving high diagnostic accuracy (92.5%, 95% CI: 80.1–97.4%) and perfect prescription accuracy (100.0%, 95% CI: 100.0–100.0%). DeepSeek R1 performed similarly, completing 47.8% (95% CI: 44.8–50.8%) of the complete checklist and 64.6% (95% CI: 56.1–73.1%) of the essential checklist, with perfect scores in both diagnosis and medication prescription (100.0%, 95% CI: 100.0–100.0%).

Regarding safety indicators, primary care providers requested 2.78 (95% CI: 2.31–3.24) laboratory tests and prescribed 0.65 (95% CI: 0.22–1.08) medications on average. Unnecessary test orders were recorded in 35.0% (95% CI: 19.6–50.4%) of cases, and inappropriate or potentially harmful medications were prescribed in 20.0% (95% CI: 7.0–33.0%) of cases. In comparison, ChatGPT-4o requested 3.65 (95% CI: 3.22–4.08) lab tests and prescribed 5.50 (95% CI: 5.05–5.95) medications, with substantially higher rates of unnecessary tests (92.5%, 95% CI: 80.1–97.4%) and inappropriate prescriptions (67.5%, 95% CI: 52.3–82.7%). DeepSeek R1 exhibited similar patterns, requesting 4.93 (95% CI: 4.41–5.44) lab tests and prescribing 5.92 (95% CI: 5.53–6.32) medications, with rates of unnecessary tests and inappropriate medications reaching 100.0% (95% CI: 100.0–100.0%) and 60.0% (95% CI: 44.1–75.9%), respectively.

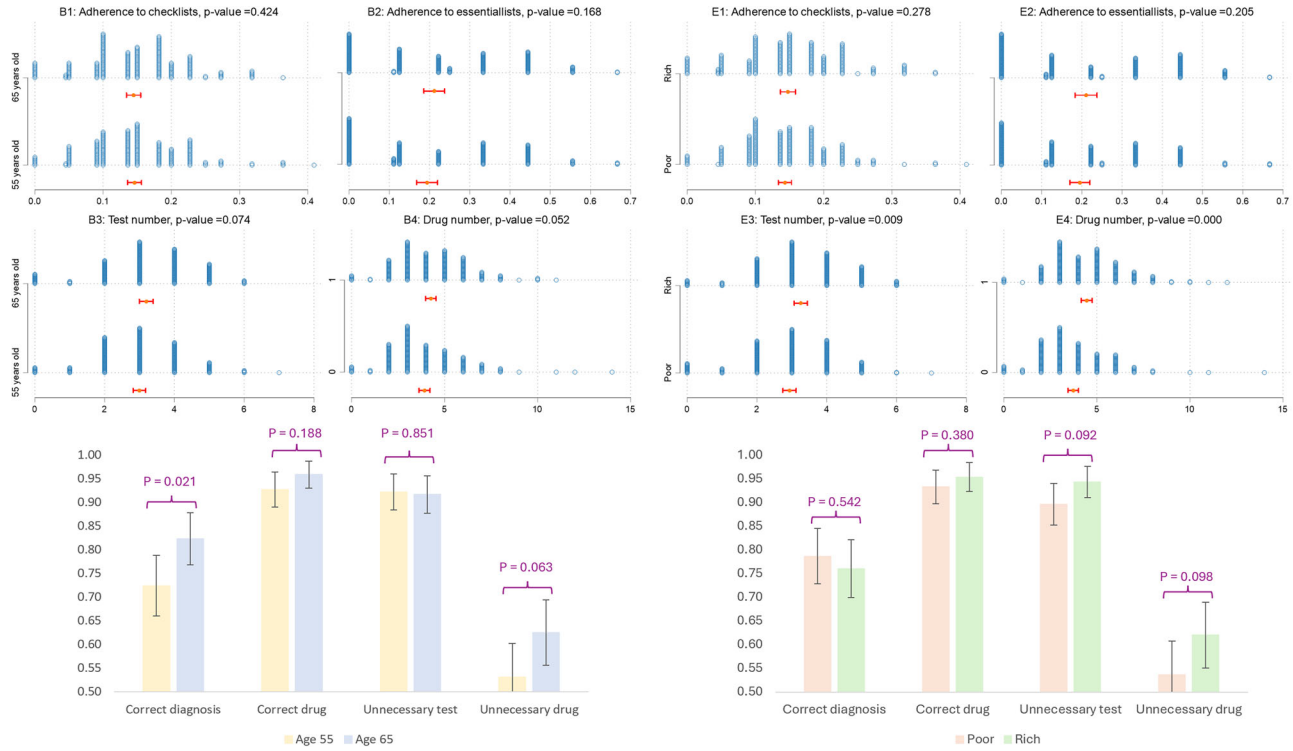

Fig. 1 | The quality and safety indicators of AI consultations by patient age and household economic status. Note: Means and 95% confidence intervals (CIs) are presented in red, including the distribution of all observations in blue dots; chi-square tests were performed on binary and analysis of variance (ANOVA) for continuous variables.

Table 2 | Influences of six patient-level factors on the quality and safety performance of the AI consultations

Quality indicators	Full checklists		Essential checklists		Correct diagnosis rate		Correct medication rate	
	dy/dx	95% CI	dy/dx	95% CI	dy/dx	95% CI	dy/dx	95% CI
Asthma (ref: Unstable Angina)	−0.062***	−0.075, −0.049	−0.321***	−0.345, −0.296	0.011	−0.071, 0.093	−0.010	−0.054, 0.034
Male (ref: female)	−0.007	−0.020, 0.006	−0.005	−0.025, 0.015	−0.005	−0.088, 0.078	0.014	−0.030, 0.058
65 years old (ref: 55 years old)	−0.001	−0.014, 0.012	0.018*	−0.002, 0.038	0.098**	0.017, 0.180	0.032	−0.014, 0.078
Urban registration (ref: non-urban)	0.009	−0.004, 0.022	0.005	−0.015, 0.025	0.046	−0.036, 0.129	0.001	−0.045, 0.046
Urban residence (ref: rural)	0.009	−0.004, 0.022	0.011	−0.009, 0.032	0.013	−0.069, 0.096	0.010	−0.036, 0.055
Wealthier (ref: poorer) household economic status	0.004	−0.009, 0.017	0.015	−0.005, 0.036	−0.023	−0.105, 0.060	0.020	−0.026, 0.065
UEMI (ref: URRMI)	0.008	−0.005, 0.021	−0.000	−0.021, 0.020	−0.057	−0.139, 0.026	−0.010	−0.056, 0.035
Safety indicators	No. of lab tests requested		Unnecessary requested lab tests rate		No. of medications prescribed		Inappropriate prescribed medications rate	
	dy/dx	95% CI	dy/dx	95% CI	dy/dx	95% CI	dy/dx	95% CI
Asthma (ref: Unstable Angina)	0.010	−0.254, 0.274	−0.102***	−0.162, −0.042	0.234	−0.162, 0.631	0.104**	0.008, 0.199
Male (ref: female)	0.094	−0.170, 0.358	0.024	−0.028, 0.076	−0.214	-0.610, 0.183	−0.063	−0.159, 0.034
65 years old (ref: 55 years old)	0.198	−0.066, 0.462	−0.003	−0.056, 0.050	0.339*	−0.058, 0.735	0.093*	−0.002, 0.189
Urban registration (ref: non-urban)	0.156	−0.108, 0.420	−0.001	−0.053, 0.051	0.089	−0.308, 0.485	−0.010	−0.108, 0.087
Urban residence (ref: rural)	0.156	−0.108, 0.420	0.004	−0.048, 0.056	−0.036	−0.433, 0.360	−0.033	−0.130, 0.064
Wealthier (ref: poorer) household economic status	0.323**	0.059, 0.587	0.044	−0.009, 0.097	0.724***	0.327, 1.121	0.083*	−0.013, 0.180
UEMI (ref: URRMI)	0.073	−0.191, 0.337	−0.017	−0.068, 0.035	0.391*	−0.006, 0.787	0.021	−0.076, 0.118

*$p < 0.1$, **$p < 0.05$, ***$p < 0.01$; Coefficients (dy/dx, all in absolute decimal points) and the 95% confidence intervals (CIs) estimated by the multivariable General Linear or Poisson Regression models. *UEMI* Urban Employee Medical Insurance, *URRMI* Urban and Rural Resident Medical Insurance.

Table 3 | Comparing ERNIE Bot with China's Primary Care Physicians, ChatGPT 4o, and DeepSeek R1

	ERNIE Bot (*n*=384)		Physicians[#] (*n*=40)		ChatGPT 4o (*n*=40)		DeepSeek R1 (*n*=40)	
	Mean	95% CI	Mean	95% CI	Mean	95% CI	Mean	95% CI
Quality indicators								
% completion of the full checklists (0-1)	0.145	0.138, 0.153	0.261	0.221, 0.301	0.413	0.393, 0.434	0.478	0.448, 0.508
% completion of the essential checklists (core, subset, 0-1)	0.203	0.184, 0.221	0.371	0.279, 0.464	0.533	0.458, 0.607	0.646	0.561, 0.731
% Correct diagnosis (0-1)	0.773	0.731, 0.815	0.250	0.110, 0.390	0.925	0.801, 0.974[$]	1.000	1.000, 1.000
% Correct medication (0-1)	0.943	0.919, 0.966	0.100	0.003, 0.197	1.000	1.000, 1.000	1.000	1.000, 1.000
Safety								
No. of tests requested	3.09	2.96, 3.23	2.78	2.31, 3.24	3.65	3.22, 4.08	4.93	4.41, 5.44
% Unnecessary test requested (0-1)	0.919	0.892, 0.947	0.350	0.196, 0.504	0.925	0.801, 0.974[$]	1.000	1.000, 1.000
No. of medication prescribed	4.09	3.89, 4.30	0.65	0.22, 1.08	5.50	5.05, 5.95	5.93	5.53, 6.32
% Inappropriate medication prescribed (0-1)	0.578	0.529, 0.628	0.200	0.070, 0.330	0.675	0.523, 0.827	0.600	0.441, 0.759

[#]The SP-physician data was collected in Luohe, China, in 2025. The 40 SP visits in Luohe City were randomly sampled through a multistage random cluster sampling strategy. Asthma and unstable angina were equally stratified among visits. Luohe is a prefecture-level city in central Henan Province, China, with a population of approximately 2.37 million and a well-developed primary healthcare system. As a mid-sized city with a mix of urban and peri-urban communities, Luohe reflects many structural and resource characteristics typical of primary care delivery in China's low- and middle-income regions. It is also situated along major transport corridors, making it logistically accessible for SP research. Luohe's healthcare infrastructure includes a broad network of community health service centres and township clinics operating under national essential public health programs. These features make Luohe a representative setting for evaluating the quality and safety of routine outpatient care delivered by human primary care providers and for benchmarking AI performance in a real-world, yet generalizable, LMIC context. Means and 95% confidence intervals (CIs) for binary and continuous variables. [$]Wilson CIs were presented.

Sensitivity analysis

Although physicians and AI chatbots were compared using the same quality metrics, AI chatbots may be advantageous by providing more diagnoses and drug prescriptions. To fully address this concern, we benchmark physicians by using only the first diagnosis and drug prescription from AI chatbots. We find the rates of correct first diagnosis and appropriate medication dropped only slightly but still substantially outperformed those of physicians in both dimensions (Supplementary Table 1). In addition, despite the higher prevalence of unnecessary prescriptions in AI chatbots, we find their proportion of unnecessary lab tests and medications was very comparable to physicians (Supplementary Table 1).

Discussion

The rise of generative AI, exemplified by LLMs like ChatGPT and ERNIE Bot, is transforming healthcare landscapes, especially in LMICs. These regions, aided by the growing internet and smartphone access[35], are increasingly using AI chatbots for medical consultation. This study provides one of the first empirical evaluations of a widely accessible generative AI chatbot, ERNIE Bot 3.5, for chronic disease management in a low-resource setting, benchmarking its performance against human clinicians and other LLMs under standardized conditions, providing critical insights into the care quality, safety, and disparity.

ERNIE Bot achieved a relatively high diagnostic accuracy (77.3%) and correct drug prescription (94.3%). The performance remained high even when using the first diagnosis (55.5%) and first drug prescription (86.2%). The results are consistent with a pilot study using ChatGPT 3.5 covering 9 chronic and infectious diseases[28], although ChatGPT performed better in managing NCDs than infectious diseases. Consistent with national and international efforts to improve data surveillance[36], studies using similar SP methods suggested that primary care providers in LMICs like China, India, and Kenya can reach correct diagnoses in 12–52% of visits[11,16]. These results indicate that ERNIE Bot has the potential to address significant gaps in healthcare delivery by empowering less qualified healthcare providers in LMICs settings and addressing the underdiagnosis and poor management of NCDs.

Another notable finding is that ERNIE Bot completed a relatively small proportion of the standard checklist items, and primary care physicians completed a similar proportion. While the ability of LLMs to achieve high diagnostic accuracy with minimal checklist adherence highlights their powerful pattern recognition capabilities, it also raises concerns about transparency, reproducibility, and medico-legal accountability. In traditional clinical encounters, checklist adherence is a proxy for thorough history-taking and contributes to medical accountability[37]. Incomplete documentation or reasoning trails could hinder clinician oversight, auditability, and patient trust. In AI-driven interactions, low process completeness could lead to missed comorbidities or contradictions that are not explicitly prompted. Future development should prioritize explainability and interactive probing capabilities to ensure that AI tools do not sacrifice safety for efficiency.

However, one of the most concerning findings is the high rate of unnecessary lab tests requested (91.9%) and medications prescribed (57.8%) by ERNIE Bot. The pattern is consistent with our previous findings using ChatGPT 3.5[28]. Earlier studies suggested that primary care doctors offered low-value care in 28-64% of SP visits in LIMCs[11,16], which is mainly driven by finance and organization, thinking frameworks[38], and patient-physician relationships[39]. In contrast, the observed tendency of AI toward overprescription and over-requesting pathology tests may reflect both the lack of real-world accountability mechanisms[28] and potential biases in training data that reward exhaustive workups[40]. Without external constraints, generative AI models may prioritize comprehensiveness over clinical appropriateness. In resource-constrained settings, such overprescription drives up unnecessary costs and increases patient exposure to potential harm, offsetting the intended benefits of AI-driven efficiency.

While this study focuses on ERNIE Bot, it is essential to situate its performance within the broader ecosystem of generative AI models and human physicians. We find that primary care providers in Luohe, China can only reach very low accuracy in correct diagnosis (25.0%) and correct drug prescriptions (10%); compared with ERNIE Bot 3.5, ChatGPT-4o and DeepSeek-R1 which have reported similar or even higher diagnostic accuracy and prescribing reliability in clinical simulations, since they are regarded as more advanced but paid AI models[41,42]. Although direct head-to-head trials between physicians and LLMs are still limited, these models appear to share strengths, such as high recall for diagnostic hypotheses, and limitations in tendencies toward overprescription[28]. The results indicate that this common feature of LLMs and that professional oversight is necessary for the automated decision-making process in AI medical consultation. Together, these results reinforce the importance of rigorous, context-specific evaluation of AI tools before large-scale deployment. Future studies should

extend benchmarking to a broader range of acute and chronic conditions, explore dynamic interactions with real patients, and conduct prospective head-to-head comparisons between AI chatbots and human clinicians across diverse LMIC contexts.

This study also sheds light on the disparities perpetuated by the application of AI in healthcare[43–45]. ERNIE Bot mainly exhibited a significantly higher rate of achieving an accurate diagnosis for older adults than for younger ones. It is reasonable since chronological age is often considered a key contributor to the onset of chronic conditions[33]. However, it was unexpected that older adults received marginally more medications and had a higher chance of receiving unnecessary medications. Further, patients from better-off households received more lab tests and medicines than those from poorer ones. ERNIE Bot overserved wealthier patients, which inevitably leads to a higher chance of excessive pathology tests and inappropriate medication prescriptions. This is supported, to some extent, by real-world evidence where patients with more generous health insurance coverage or higher out-of-pocket affordability tend to obtain more medical resources[46,47]. In general, offering AI models a budget constraint in their decision-making has been understudied. Offer information on insurance type or socioeconomic status may not be as direct as a budget constraint. We will consider pursuing this as a future direction. Again, no performance variations were identified regarding SPs' gender, residential hukou registration, or permanent residential location.

Although ERNIE Bot is not integrated into health systems in China, its growing accessibility through commercial platforms raises the possibility of informal use in health decision-making. Potential integration pathways may include deployment as a triage tool for low-acuity conditions, a health literacy assistant for patients, or a decision-support tool for less-experienced clinicians in under-resourced settings. However, integrating AI tools like ERNIE Bot into healthcare systems presents both an opportunity and a challenge. Studies in China have shown that LLMs can improve primary diabetes care and outpatient reception[48,49], but equitably scaling the findings will require attention to rural, low-resource settings[50]. ERNIE Bot holds promise in alleviating the burden of NCDs by extending diagnostic and treatment capabilities in settings where resources are scarce. However, our findings also emphasize balancing AI's potential with necessary safeguards. Especially, 'do not harm' remains a foundational principle about using AI in health care. Such integration would require rigorous evaluation of safety, clinical validity, and system compatibility.

Future research should embed ethical, stakeholder-driven design principles from the outset to enhance the safety and equity of AI chatbots in healthcare. Rather than assessing AI safety and equity after deployment, proactive engagement with key stakeholders, including patients, health care providers, and policymakers, at an early stage is essential. This early engagement will capture diverse expectations, values, and concerns, particularly from underrepresented groups, thereby informing the ethical, cultural, and contextual alignment of AI chatbot systems. Second, future work should focus on operationalizing safety and responsibility through practical, empirically validated mechanisms. Building on stakeholder insights and empirical performance evaluations, the development of automated AI alignment solutions and best practice toolkits should be prioritised. Agent-based tools for pre- and post-processing AI-generated outputs and user guides should be co-designed and iteratively refined through stakeholder workshops and chatbot re-testing cycles. Third, future studies can explore collaborative decision-making models that integrate LLMs with human providers to evaluate their assistive potential in real-world clinical workflows. This translational approach may offer tangible, practice-ready solutions for policymakers, AI developers, and healthcare institutions.

This study acknowledges several limitations. First, our analysis focuses on two specific chronic conditions, which may limit the generalizability of the findings to other diseases or specialties. Unstable angina and asthma were selected due to their clinical significance in ageing populations, the availability of established national clinical guidelines, and their suitability for SP simulation. Importantly, the presenting symptoms of these conditions align with some of the most common complaints encountered in primary care, thereby enhancing our study's relevance and practical value for primary healthcare settings. Second, the SP method may not fully capture the complexity of real-world patient interactions. However, previous studies have shown that provider behaviour toward SPs closely mirrors their behaviour with actual patients[10,37]. Third, we did not account for emotional communications. Compared to factual information exchange, patient-centred communication is also essential, as it is perceived as trustworthy, accurate, reliable, and actionable[51]. Fourth, ERNIE Bot has been trained on data containing the Chinese language, limiting our results' broader applicability to other healthcare contexts. The evolving nature of generative AI models means that outputs may vary over time as models are updated, potentially affecting replicability.

Despite the limitations, we present one of the first empirical evaluations of a generative AI chatbot's diagnostic and prescribing performance against human clinicians and frontier LLMs under standardized, real-world simulated conditions in a low-resource setting. While ERNIE Bot demonstrated high diagnostic accuracy and medication appropriateness, critical challenges remain, including low adherence to standard clinical processes, high rates of unnecessary care, and amplification of socioeconomic disparities. These findings highlight AI chatbots' dual potential to expand healthcare access while introducing new risks if deployed without safeguards. Future development and integration of AI systems should prioritize equity-centred design, explainability, rigorous, context-specific validation, and continuous human oversight to ensure AI chatbots contribute safely and ethically to strengthening global health systems.

Methods

Ethical approvals were obtained from the relevant Chinese institutional review boards: the First Affiliated Hospital of Xi'an Jiaotong University (No: LLSBPJ-2024-WT-019) and Luohe Medical College (No: LYZLL-2024012).

We clarify that each SP was assigned one disease case (unstable angina or asthma). Each case was tested three times to evaluate repeatability through independently initiated sessions. A new, independent chat session was initiated for each trial to avoid memory retention effects. In addition, the AI chatbots' memory was cleared before a new chat. This ensured that AI chatbots treated each interaction as a separate, first-time consultation, maintaining consistency and real-world reliability of the outputs. All trials were completed in the same sitting to ensure consistency and minimize external variability. SPs did not take any diagnostic tests themselves for AI consultations. This is because SPs were trained to portray predefined clinical cases representing common diseases, where appropriate history-taking alone was sufficient to support an unambiguous and accurate diagnosis and treatment recommendation.

Mandarin was used to test the performance of ERNIE Bot, ChatGPT, and DeepSeek to be consistent with human physicians. Written consent forms were obtained from hospitals and physicians before the SPs' visits, but physicians were not aware of the diseases to be tested. The SP scripts have been translated into English and can be found in Supplementary Note 1. Physicians' and AI chatbots' responses were cross-validated with the most updated standard clinical guidelines, *Guidelines for the Prevention and Control of Bronchial Asthma (2020 Edition)* and *Guidelines for the Diagnosis and Treatment of Unstable Angina (2024 Edition)*, for the two selected NCDs, to assess the accuracy and appropriateness of its diagnoses and medication prescriptions. A panel of six senior doctors and pharmacists with over 15 years of clinical experience at tertiary hospitals independently reviewed and validated the scripts[11,52]. The details about the scripts and their associated checklists can be found in Supplementary Note 2.

Four quality indicators reflected the extent to which patients receive timely and accurate diagnoses and evidence-based treatment[53]. (1) Adherence to the standard complete checklist: including clinical inquiries and recommended laboratory-based pathology tests in agreement with the standard complete checklists[14,34]. This first indicator was coded as a continuous variable, ranging from 0 (nil agreement) to 1 (complete agreement). (2) Adherence to the standard essential checklist: including clinical inquiries and recommended laboratory tests in agreement with the standard

'essential' (core) checklists (a subset). This second indicator was also coded as a continuous variable, ranging from 0 (nil agreement) to 1 (complete agreement). (3) Correct diagnosis: At the end of each trial, the artificial SP directly requested that ERNIE Bot provide a diagnosis. This third binary indicator was assigned to either 1 (correct) if the AI-driven consultation trial produced at least one expected diagnosis according to the standard guidelines[37] or 0 (incorrect / misdiagnosis). (4) Correct medication prescription: Similarly, the fourth binary indicator was assigned to 1 (correct) if at least one guideline-recommended medication was prescribed at each AI trial. Otherwise, it was assigned to 0 (incorrect), denoting irrelevant, unnecessary, or even potentially harmful medication advice. We note that this is a commonly accepted rule when using SP to evaluate health care quality, although the standard is somewhat low in high-income countries.

An additional four safety indicators were included, focused on the AI-generated outcomes that were incongruent with the standard diagnostic and treatment guidelines: (i) the absolute number of irrelevant or unnecessary pathology tests requested (the 5th indicator, a numeric continuous variable), (ii) the presence of any of these test requests (the 6th binary indicator), (iii) the absolute number of inappropriate medications prescribed (the 7th indicator, a numeric continuous variable), and (iv) the presence of any of these medication prescriptions (the 8th binary indicator).

Descriptive analysis was conducted to summarize the four quality and four safety indicators of the total sample, respectively, for each disease condition. Apart from the absolute trial numbers, means and standard deviations (SDs) were used to report the continuous variable indicators, whereas proportions were used for the rest of the binary variable indicators.

Next, trial outcomes involving the six patient-level factors were examined. To assess the AI-generated outcome variations, chi-square tests were performed on binary variables and analysis of variance (ANOVA) for continuous variables.

Finally, to identify the associations of the six patient factors with the extent of variability of the quality and safety indicators, generalized linear models (GLM) were applied for continuous variable indicators and probit regressions for binary variable indicators. Average marginal effects were reported, and 95% confidence intervals (CIs) were presented. Statistical significance was set at $p < 0.05$. All analyses were performed in Stata 18.0 (Stata Corporation, College Station, TX).

Data availability

Data from this simulation study are available to anyone for any non-commercial purposes. The data from human physicians are not publicly available due to restrictions of the ethics approval for this study.

Code availability

The code scripts used in this analysis are available from the corresponding authors upon reasonable request.

Received: 8 January 2025; Accepted: 17 August 2025;
Published online: 25 September 2025

References

1. Ahmed, S. M. et al. Delivering non-communicable disease services through primary health care in selected South Asian countries: are health systems prepared?. *Lancet Glob. Health* **12**, e1706–e1719 (2024).
2. Roth, G. A. et al. Global Burden of Cardiovascular Diseases and Risk Factors, 1990–2019. *J. Am. Coll. Cardiol.* **76**, 2982–3021 (2020).
3. Momtazmanesh, S. et al. Global burden of chronic respiratory diseases and risk factors, 1990–2019: an update from the Global Burden of Disease Study 2019. *EClinicalMedicine* **59**, 101936 (2023).
4. Darmawan, E. S. et al. Beyond the Plate: Uncovering Inequalities in Fruit and Vegetable Intake across Indonesian Districts. *Nutrients* **15**, 2160 (2023).
5. Li, X. et al. The primary health-care system in China. *Lancet* **390**, 2584–2594 (2017).
6. Lu, J. et al. Prevalence, awareness, treatment, and control of hypertension in China: data from 1· 7 million adults in a population-based screening study (China PEACE Million Persons Project). *Lancet* **390**, 2549–2558 (2017).
7. Zhang, M. et al. Prevalence, awareness, treatment, and control of hypertension in China, 2004-18: findings from six rounds of a national survey. *BMJ*. **380**, e071952 (2023).
8. Wang, L. et al. Prevalence and ethnic pattern of diabetes and prediabetes in China in 2013. *Jama* **317**, 2515–2523 (2017).
9. Xiong, S. et al. Factors associated with the uptake of the national essential public health service package for hypertension and type-2 diabetes management in China's primary health care system: a mixed-methods study. *Lancet Reg. Heal. Pac*. **31**, 100664 (2023).
10. Das, J., Chowdhury, A., Hussam, R. & Banerjee, A. V. The impact of training informal health care providers in India: A randomized controlled trial. *Science* **354**, 6308 (2016).
11. Si, Y. et al. Quantifying the financial impact of overuse in primary care in China: A standardised patient study. *Soc. Sci. Med*. **320**, 115670 (2023).
12. Si, Y. et al. The quality of telemedicine consultations for sexually transmitted infections in China. *Health Policy Plan.* **39**, 307–317 (2024).
13. Su, M., Zhou, Z., Si, Y. & Fan, X. The Association Between Patient-Centered Communication and Primary Care Quality in Urban China: Evidence From a Standardized Patient Study. *Front. Public Health* **9**, 779293 (2022).
14. Sylvia, S. et al. Survey using incognito standardized patients shows poor quality care in China's rural clinics. *Health Policy Plan.* **30**, 322–333 (2015).
15. Si, Y., Chen, G., Zhou, Z., Yip, W. & Chen, X. The impact of physician-patient gender match on healthcare quality: An experiment in China. *Soc. Sci. Med*. **380**, 118166 (2025).
16. Kwan, A. et al. Use of standardised patients for healthcare quality research in low-and middle-income countries. *BMJ Glob. Health* **4**, e001669 (2019).
17. Li, D. et al. Unequal distribution of health human resource in mainland China: what are the determinants from a comprehensive perspective? *Int. J. Equity Health* **17**, 29 (2018).
18. Sarraju, A. et al. Appropriateness of cardiovascular disease prevention recommendations obtained from a popular online Chat-based Artificial Intelligence Model. *JAMA* **329**, 842–844 (2023).
19. Kuroiwa, T. et al. The potential of ChatGPT as a self-diagnostic tool in common orthopedic diseases: exploratory study. *J. Med. Internet Res.* **25**, e47621 (2023).
20. Vo, V. et al. Multi-stakeholder preferences for the use of artificial intelligence in healthcare: A systematic review and thematic analysis. *Soc. Sci. Med.* **338**, 116357 (2023).
21. Kurniawan, M. H., Handiyani, H., Nuraini, T., Hariyati, R. T. S. & Sutrisno, S. A systematic review of artificial intelligence-powered (AI-powered) chatbot intervention for managing chronic illness. *Ann. Med.* **56**, 2302980 (2024).
22. Wang, C. et al. Ethical considerations of using ChatGPT in health care. *J. Med. Internet Res.* **25**, e48009 (2023).
23. Frost & Sullivan. 2024 China Large Language Model Evaluatoin Analysis Result. https://www.frostchina.com/content/insight/detail/6600efdba2aa84f5d87e82df (2024).
24. Huang, L. et al. The performance evaluation of artificial intelligence ERNIE bot in Chinese National Medical Licensing Examination. *Postgrad. Med. J.* **100**, qgae062 (2024).
25. Johri, S. et al. An evaluation framework for clinical use of large language models in patient interaction tasks. *Nat. Med.* **31**, 1–10 (2025).
26. Wiseman, V. et al. Using Unannounced Standardised Patients to Obtain Data on Quality of Care in Low-Income and Middle-Income Countries: Key Challenges and Opportunities. (BMJ Specialist Journals, 2019).

27. Si, Y., Zhou, Z., Su, M. & Chen, X. Revisiting gender gap in quality of health care in urban China: a standardised patient audit study. *Lancet* **394**, S25 (2019).
28. Si, Y. et al. Quality and Accountability of ChatGPT in health care in low- and middle-income countries: simulated patient study. *J. Med. Internet Res.* **26**, e56121 (2024).
29. Zhu, Y. & Österle, A. Rural-urban disparities in unmet long-term care needs in China: The role of the hukou status. *Soc. Sci. Med.* **191**, 30–37 (2017).
30. Su, M. et al. Comparing the effects of China's three basic health insurance schemes on the equity of health-related quality of life: using the method of coarsened exact matching. *Health Qual. Life Outcomes* **16**, 41 (2018).
31. Xiong, S. et al. Using routinely collected data to determine care cascades of hypertension and type-2 diabetes management in China: a cross-sectional study. *Lancet Reg. Heal. Pac.* **45**, 101019 (2024).
32. Balafoutas, L., Kerschbamer, R. & Sutter, M. Second-degree moral hazard in a real-world credence goods market. *Econ. J.* **127**, 1–18 (2017).
33. São José, J. M. S., Amado, C. A. F., Ilinca, S., Buttigieg, S. C. & Taghizadeh Larsson, A. Ageism in health care: a systematic review of operational definitions and inductive conceptualizations. *Gerontologist* **59**, e98–e108 (2019).
34. Das, J. et al. In urban and rural India, a standardized patient study showed low levels of provider training and huge quality gaps. *Health Aff.* **31**, 2774–2784 (2012).
35. Howarth, J. How Many People Own Smartphones (2023-2028). *Exploding Topics* https://explodingtopics.com/blog/smartphone-stats (2023).
36. Xu, D.(R. oman) et al. Improving Data Surveillance Resilience Beyond COVID-19: Experiences of Primary heAlth Care quAlity Cohort In ChinA (ACACIA) Using Unannounced Standardized Patients. *Am. J. Public Health* **112**, 913–922 (2022).
37. Das, J., Holla, A., Mohpal, A. & Muralidharan, K. Quality and accountability in health care delivery: audit-study evidence from primary care in India. *Am. Econ. Rev.* **106**, 3765–3799 (2016).
38. Weeks, J. C. et al. Patients' expectations about effects of chemotherapy for advanced cancer. *N. Engl. J. Med.* **367**, 1616–1625 (2012).
39. He, A. J. The doctor–patient relationship, defensive medicine and overprescription in Chinese public hospitals: Evidence from a cross-sectional survey in Shenzhen city. *Soc. Sci. Med.* **123**, 64–71 (2014).
40. Sellamuthu, S. et al. AI-based recommendation model for effective decision to maximise ROI. *Soft Comput*. 1–10 (2023).
41. Zeng, D., Qin, Y., Sheng, B. & Wong, T. Y. DeepSeek's "Low-Cost" Adoption Across China's Hospital Systems: Too Fast, Too Soon?. *JAMA* **333**, 1866–1869 (2025).
42. Tordjman, M. et al. Comparative benchmarking of the DeepSeek large language model on medical tasks and clinical reasoning. *Nat. Med.* **31**, 2550–2555 (2025).
43. Zack, T. et al. Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study. *Lancet Digit. Health* **6**, e12–e22 (2024).
44. Omiye, J. A., Lester, J. C., Spichak, S., Rotemberg, V. & Daneshjou, R. Large language models propagate race-based medicine. *NPJ Digit. Med.* **6**, 195 (2023).
45. DeCamp, M. & Lindvall, C. Mitigating bias in AI at the point of care. *Science* **381**, 150–152 (2023).
46. Gottschalk, F., Mimra, W. & Waibel, C. Health services as credence goods: A field experiment. *Econ. J.* **130**, 1346–1383 (2020).
47. Kerschbamer, R., Neururer, D. & Sutter, M. Insurance coverage of customers induces dishonesty of sellers in markets for credence goods. *Proc. Natl. Acad. Sci.* **113**, 7454–7458 (2016).
48. Li, J. et al. Integrated image-based deep learning and language models for primary diabetes care. *Nat. Med.* **30**, 1–11 (2024).
49. Wan, P. et al. Outpatient reception via collaboration between nurses and a large language model: a randomized controlled trial. *Nat. Med.* **30**, 2878–2885 (2024).
50. Yip, W. Improving primary healthcare with generative AI. *Nat. Med*. **30**, 1–2 (2024).
51. Bertakis, K. D. & Azari, R. Patient-centered care is associated with decreased health care utilization. *J. Am. Board Fam. Med.* **24**, 229–239 (2011).
52. Su, M. et al. Comparing the Quality of Primary Care between Public and Private Providers in Urban China: A Standardized Patient Study. *Int. J. Environ. Res. Public. Health* **18**, 5060 (2021).
53. Das, J., Woskie, L., Rajbhandari, R., Abbasi, K. & Jha, A. Rethinking assumptions about delivery of healthcare: implications for universal health coverage. *BMJ*. **361**, k1716 (2018).

Acknowledgements

No funding was available to support this study. YS acknowledges the support from the National Social Science Foundation of China (no. 23AZD091) to conduct healthy ageing research. XC acknowledges financial support from the Drazen scholarship and the Aden scholarship, which are dedicated to research on Chinese healthcare systems. The authors acknowledge the helpful comments from Virginia Wiseman, Michael Kidd, and participants of the UNSW Ageing Futures Institute Annual Symposium 2024.

Author contributions

Y.S. drafted the main manuscript text; Y.S., Y.M. and X.F. prepared tables and figures. X.C., R.A., L.M., B.L., H.B., H.Z., H.F., J.Z., S.G., Z.Z., Y.M. and G.C. edited the manuscript. All authors reviewed and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41746-025-01956-w.

Correspondence and requests for materials should be addressed to Shaoqing Gong or Xiaojing Fan.