BMJ Open

Effects of interacting with a large language model compared with a human coach on the clinical diagnostic process and outcomes among fourth-year medical students: study protocol for a prospective, randomised experiment using patient vignettes

Juliane E Kämmer ,[1] Wolf E Hautz ,[1] Gert Krummrey ,[2] Thomas C Sauter ,[1] Dorothea Penders ,[3,4] Tanja Birrenbach ,[1] Nadine Bienefeld [5]

To cite: Kämmer JE, Hautz WE, Krummrey G, *et al*. Effects of interacting with a large language model compared with a human coach on the clinical diagnostic process and outcomes among fourth-year medical students: study protocol for a prospective, randomised experiment using patient vignettes. *BMJ Open* 2024;**14**:e087469. doi:10.1136/bmjopen-2024-087469

► Prepublication history and additional supplemental material for this paper are available online. To view these files, please visit the journal online (https://doi.org/10.1136/bmjopen-2024-087469).

Received 10 April 2024
Accepted 02 July 2024

For numbered affiliations see end of article.

Correspondence to
Dr Juliane E Kämmer;
juliane.kaemmer@unibe.ch

ABSTRACT

Introduction Versatile large language models (LLMs) have the potential to augment diagnostic decision-making by assisting diagnosticians, thanks to their ability to engage in open-ended, natural conversations and their comprehensive knowledge access. Yet the novelty of LLMs in diagnostic decision-making introduces uncertainties regarding their impact. Clinicians unfamiliar with the use of LLMs in their professional context may rely on general attitudes towards LLMs more broadly, potentially hindering thoughtful use and critical evaluation of their input, leading to either over-reliance and lack of critical thinking or an unwillingness to use LLMs as diagnostic aids. To address these concerns, this study examines the influence on the diagnostic process and outcomes of interacting with an LLM compared with a human coach, and of prior training vs no training for interacting with either of these 'coaches'. Our findings aim to illuminate the potential benefits and risks of employing artificial intelligence (AI) in diagnostic decision-making.

Methods and analysis We are conducting a prospective, randomised experiment with N=158 fourth-year medical students from Charité Medical School, Berlin, Germany. Participants are asked to diagnose patient vignettes after being assigned to either a human coach or ChatGPT and after either training or no training (both between-subject factors). We are specifically collecting data on the effects of using either of these 'coaches' and of additional training on information search, number of hypotheses entertained, diagnostic accuracy and confidence. Statistical methods will include linear mixed effects models. Exploratory analyses of the interaction patterns and attitudes towards AI will also generate more generalisable knowledge about the role of AI in medicine.

Ethics and dissemination The Bern Cantonal Ethics Committee considered the study exempt from full ethical review (BASEC No: Req-2023-01396). All methods will be conducted in accordance with relevant guidelines and regulations. Participation is voluntary and informed consent will be obtained. Results will be published in peer-reviewed scientific medical journals. Authorship will be determined according to the International Committee of Medical Journal Editors guidelines.

STRENGTHS AND LIMITATIONS OF THIS STUDY

- ⇒ The study is a prospective randomised controlled study of advanced medical students diagnosing complex patient cases.
- ⇒ The study includes a comparison of consultations with either a large language model or a human coach, enhancing the clinical validity of the study.
- ⇒ The detailed analysis of both the diagnostic process and its outcomes adds depth to the research findings.
- ⇒ Only advanced medical students are included in the study, potentially constraining the generalisability of the results to broader medical student populations.

INTRODUCTION

Medical diagnostic errors, defined as wrong, delayed or missed diagnoses, pose a serious threat to quality of care and patient safety, affecting 5%–15% of the patients who present to healthcare systems.[1–3] In the 2015 landmark report 'Improving Diagnosis in Healthcare', the US National Academy of Medicine warned that 'most people will experience a diagnostic error throughout their lifetime, sometimes with devastating consequences'.[4] Importantly, among harmful diagnostic errors, 84% are preventable but at the same time have higher rates of mortality than other types of error (29% vs 7%).[5 6] In a systematic review of malpractice claims worldwide, diagnostic errors were the most common and most expensive type of claim, reflecting 26%–63% of all cases.[7] Consequently, there is an urgent need for improving diagnostic decision-making in healthcare.

In recent years, specialised computerised diagnostic decision support systems such as differential diagnosis generators have been developed, showing the potential to improve the quality of diagnoses.[8] Additionally, since large language models (LLMs) based on generative pre-trained transformer (GPT) methodology have been widely disseminated, applications such as ChatGPT (Open AI) have raised hopes that such tools will become a valuable asset for (medical) education,[9 10] as well as for consultation and clinical decision support.[11–15] Recently, researchers have endeavoured to explore ChatGPT's potential and limitations in the healthcare domain, testing its medical proficiency. Across countries, they have demonstrated its ability to successfully pass medical licensing exams,[9 10 16 17] which may render ChatGPT-based chatbots a particularly useful resource for junior physicians. Thus, by leveraging their broad medical knowledge base, their capacity to engage in open-ended, natural conversations and their ability to process complex (patient) data, ChatGPT-based chatbots have the potential to augment diagnostic decision-making processes[18] and assist learners in medical education settings.[10]

However, the novelty of LLMs in diagnostic decision-making introduces uncertainties regarding their impact. Clinicians unfamiliar with using LLMs in their professional context may rely on general positive or negative attitudes towards artificial intelligence (AI), potentially hindering thoughtful use and critical evaluation of their input, leading to either over-reliance and lack of critical thinking or the neglect of AI's potential.[19–23] It is, therefore, imperative to comprehensively explore the extent, application and constraints of LLMs in clinical decision support to guarantee their conscientious and efficient implementation in practice.[12 18 24 25] To address these concerns, this prospective, randomised controlled clinical vignette study examines the influence of decision support using an LLM (ChatGPT) on the diagnostic process and outcomes compared with that of a human coach. This will advance the understanding of how human–AI collaboration can be leveraged to enhance diagnostic decision-making.

Leveraging AI for enhanced diagnostic decision-making

What makes an LLM such as ChatGPT a potentially useful coach during the diagnostic journey? In their review of recent literature on ChatGPT in clinical decision support, Ferdush *et al*[18] listed a number of relevant attributes: For example, (a) LLMs can analyse patient data and take into account relevant clinical guidelines, understand complex medical information and aid in data interpretation; using identified patterns in patient data, LLMs can propose relevant differential diagnoses of high accuracy,[26] potentially counteracting premature closure.[27] (b) Thanks to their vast knowledge base of similar cases reported in medical literature, LLMs can remind professionals of rare or complex diseases typically in danger of being overlooked. (c) LLMs possess pertinent knowledge spanning multiple medical specialties and healthcare settings, making them a useful resource in any specialty and allowing the integration of information from different medical domains. (d) With LLMs, healthcare professionals can access clinical guidelines and best practices in real time and from one source, which supports them in making informed decisions.[18] Last, (e) LLMs may take over the role of advisors,[28 29] and (peer) coaches or teachers[30 31] who guide learners through the diagnostic process by reminding them of important steps to take or differential diagnoses to consider.

There are also potential drawbacks to consider in the context of diagnostic decision-making: (a) LLMs have been observed to occasionally miss relevant patient information, exhibit hallucinations (ie, confident yet wrong responses), display biases stemming from biased training data (eg, due to under-representation of certain demographics) and show limited contextual understanding.[18] (b) Further, there is the fear that over-reliance on LLMs may lead to reduced learning opportunities[11] and deskilling and hence an increased risk of diagnostic errors in the long run. Last and contrary to this, (c) clinicians may refute insights provided by LLMs as they tend to overlook the support offered by computerised diagnostic decision support systems.[22]

Thus, given the novelty of LLMs and the lack of experience with using GPTs in the diagnostic process and for medical education, a deeper exploration of the benefits, limitations and possible applications of LLMs for medical diagnosis and education is warranted. Our study, therefore, aims to (a) investigate the effects of an LLM (ChatGPT) on the diagnostic process, accuracy, number of diagnostic hypotheses and user confidence and (b) explore how the LLM is used during diagnosis. As LLMs generate human-like text responses in conversational settings, we compare the use of ChatGPT assistance with that of assistance from a human coach with more experience, the usual resource for junior physicians in medical educational settings.[32]

The role of the hypothesis space for diagnostic error

Of the multiple reasons for diagnostic error (such as technical failures or poorly cooperating patients), cognitive factors such as faulty information synthesis most frequently contribute to diagnostic error.[6 33] To illustrate, 89% of diagnostic error malpractice claims involved failures in clinical reasoning, the largest study on such claims found.[34]

Decades of research into clinical reasoning, diagnostic decision-making, or one of its many synonyms provide some insights into possible causes and remedies of diagnostic error.[27] It is now well established that clinicians generate diagnostic hypotheses within minutes of an encounter with a patient,[35 36] sometimes even much faster.[37] These initial hypotheses are of paramount importance for the accuracy of the final diagnosis because clinicians hardly ever add other hypotheses to the diagnoses they consider later on.[35] This is an important point because—in contrast to the process of scientific inquiry—physicians tend to conduct diagnostic tests that confirm their initial hypothesis rather than potentially refuting it.[35 38] Furthermore, they distort incoming additional findings in favour of the initial idea.[39 40] What distinguishes expert diagnosticians from novices is neither faster nor more but just better initial hypotheses.[41 42] This understanding of the importance of the initial hypothesis for the accuracy of the final diagnosis aligns well with the observation that the most commonly observed biases in clinical reasoning—availability bias, confirmation bias, satisfaction of search and premature closure[27 43–47]—all relate to the space of initially considered differential diagnoses.

Given that broadening the differential diagnoses can mitigate diagnostic errors,[48–51] it appears imperative to raise awareness among diagnosticians about this possibility. Furthermore, the quality of LLM output and advice is sensitive to the formulation of inquiries.[52 53] Therefore, providing single training instructions that offer a rationale for expanding the hypothesis space in diagnostic decision-making, along with practical illustrations on how to effectively elicit information from their coaches (whether human or ChatGPT) will likely enhance the coaches' impact. This single training will improve participants' reasoning and ability to leverage the coach's assistance, leading to better diagnostic outcomes, such as an increased number and relevance of diagnostic hypotheses and greater accuracy in the final diagnosis. Consequently, we will examine the impact of instructional training (training vs no training) along with human versus AI assistance. We aim to provide insights that elucidate the necessary guidance for the effective use of LLMs in diagnostic decision-making.

METHODS AND ANALYSIS

This study seeks to elucidate the differential (or analogous) use patterns between users of ChatGPT and those using a human coach in the context of diagnostic decision-making, along with their respective impacts on the diagnostic process and outcomes as well as user confidence. There is also significant practical interest in examining whether ChatGPT exhibits a more pronounced beneficial effect on diagnostic accuracy and the quantity of differential diagnoses considered, potentially attributable to its heightened computational capabilities.[12] Additionally, we seek to assess whether brief instructional training emphasising the importance of expanding the hypothesis space augments these effects. To achieve this, our primary focus is on modelling the dependent variables diagnostic accuracy and number of generated differential diagnoses using linear mixed-effects models[54] in R.[55]

We have been collecting data during an online experiment with medical students at the Charité Medical School in Berlin. Students have been invited to participate via mailing lists in exchange for financial remuneration (€35 per participant). Data collection began on 22 April 2024 and is planned to last until the end of June 2024. The study has a randomised, single-blind study design with a 2×2 factorial design, with the source of assistance (human coach vs ChatGPT) and training (training vs no training) as between-subjects factors (see figure 1). Participants are randomly assigned to the type of assistance they receive and the training/no training condition.

Sample size

A sample size of N=158 was determined using G*Power V.3.1.9.7[56] for a 2×2 analysis of variance (ANOVA), to detect a practically relevant medium effect size with α=0.05 and β=0.80. Each of the four subgroups is randomly assigned an approximately equal number of participants.

Inclusion and exclusion

All (N=640) fourth-year medical students (in a 6-year programme) from Charité Medical School in Berlin are eligible to take part in the study. Students are recruited via faculty mailing lists, posters and online platforms of the Charité Skills Lab. Students 18 years or older who sign the informed consent can be included. Coaches in the 'human condition' are two medical interns who have recently completed their sixth year of studies at the Charité Medical School, have passed their state examination and are now working in the hospital. Human coaches are thus 2 years more advanced than the participants. They are paid €20 per hour.

Main study procedures

Data collection is taking place remotely in two online sessions (see figure 1). In the first session, students provide their written informed consent (see online supplemental information) and watch a short general introduction video on the idea and methods of LLMs to level off potential differences in experience with LLMs among participants. For this, a freely available, up-to-date introductory video was chosen (https://youtu.be/2IK3DFHRFfw?si=uSnEBQv2mhPmIOis). Then, participants fill in a short baseline survey (via https://www.soscisurvey.de) on their medical expertise, attitudes towards and experience with ChatGPT and other forms of AI, and their demographics (see online supplemental e table 1 for an overview of all questionnaires and our OSF repository https://osf.io/cbpr3/?view_only=e5e94231ddd546b491c2e07f43f02c88 for all original items and their English translation). To ensure that participants completed the first session, they are asked to send a codeword ('Psychologie'), which is provided on the last slide of the survey, by email to the experimenter.

The second session is administered via MS Teams. Up to six students are invited to the same session. On arrival, participants are welcomed by the experimenter

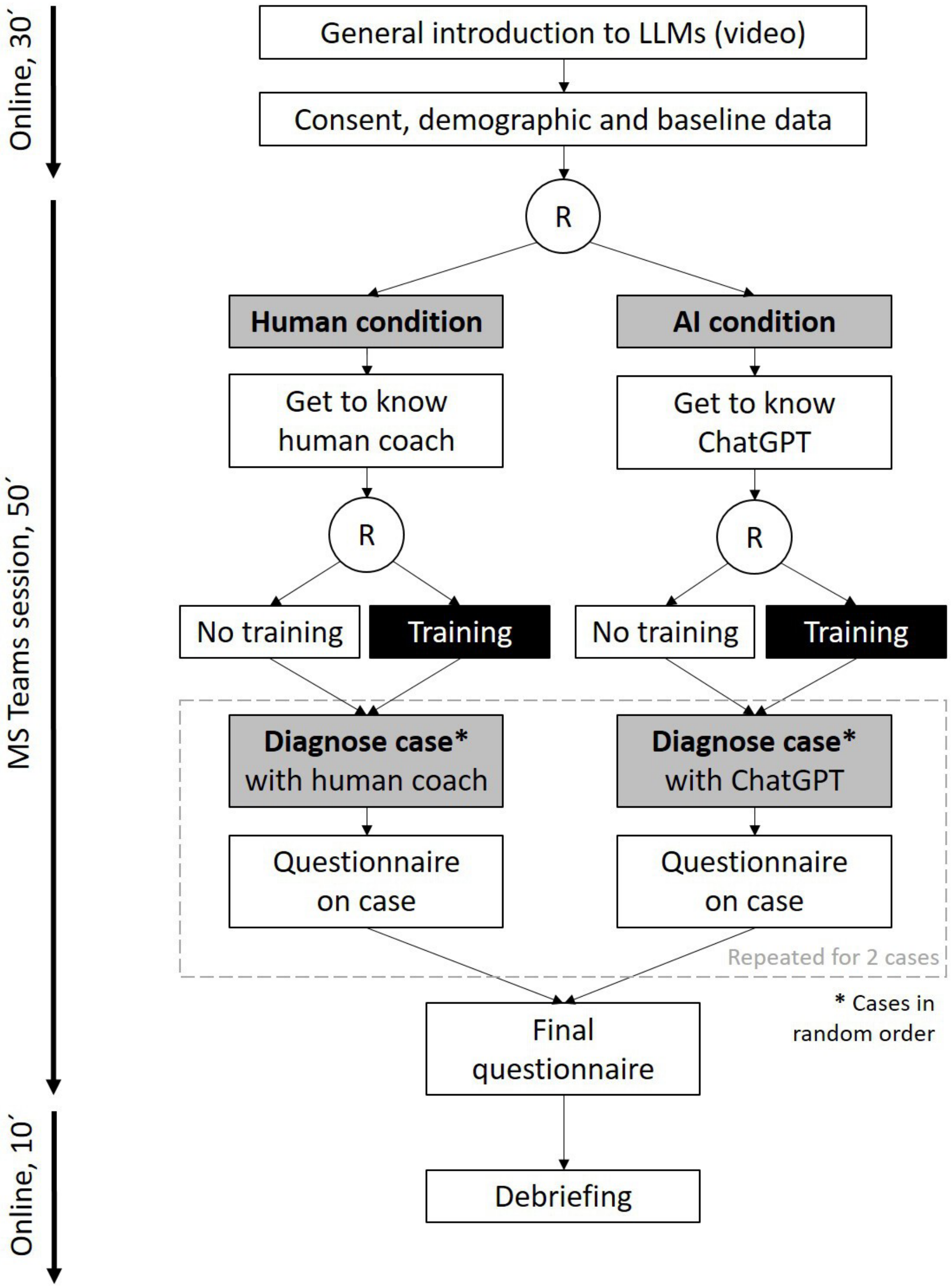


Figure 1 Study design. AI, artificial intelligence; ChatGPT, OpenAI's generative pre-trained transformer; LLMs, large language models; R, randomisation.

and receive a short introduction to the study. Then, participants are randomly assigned to the human or AI condition and training or no training subgroup by the experimenters using a computer-generated randomisation process. Participants are blinded to the training versus no training condition but are aware of the random allocation procedure to the human versus AI condition (from the general study information; see online supplemental information). Participants are sent to individual breakout rooms and receive a link to access their experimental session. They then work individually on the experiment in their breakout room with the opportunity to chat with the experimenter in case of problems or questions. After finishing, they return to the meeting room and are informed about the debriefing (which comes at a later date; see Debriefing below), thanked and dismissed. Experimenters note all deviations from the protocols, technical issues and participants' comments so that the quality of data collection can be evaluated.

Get to know

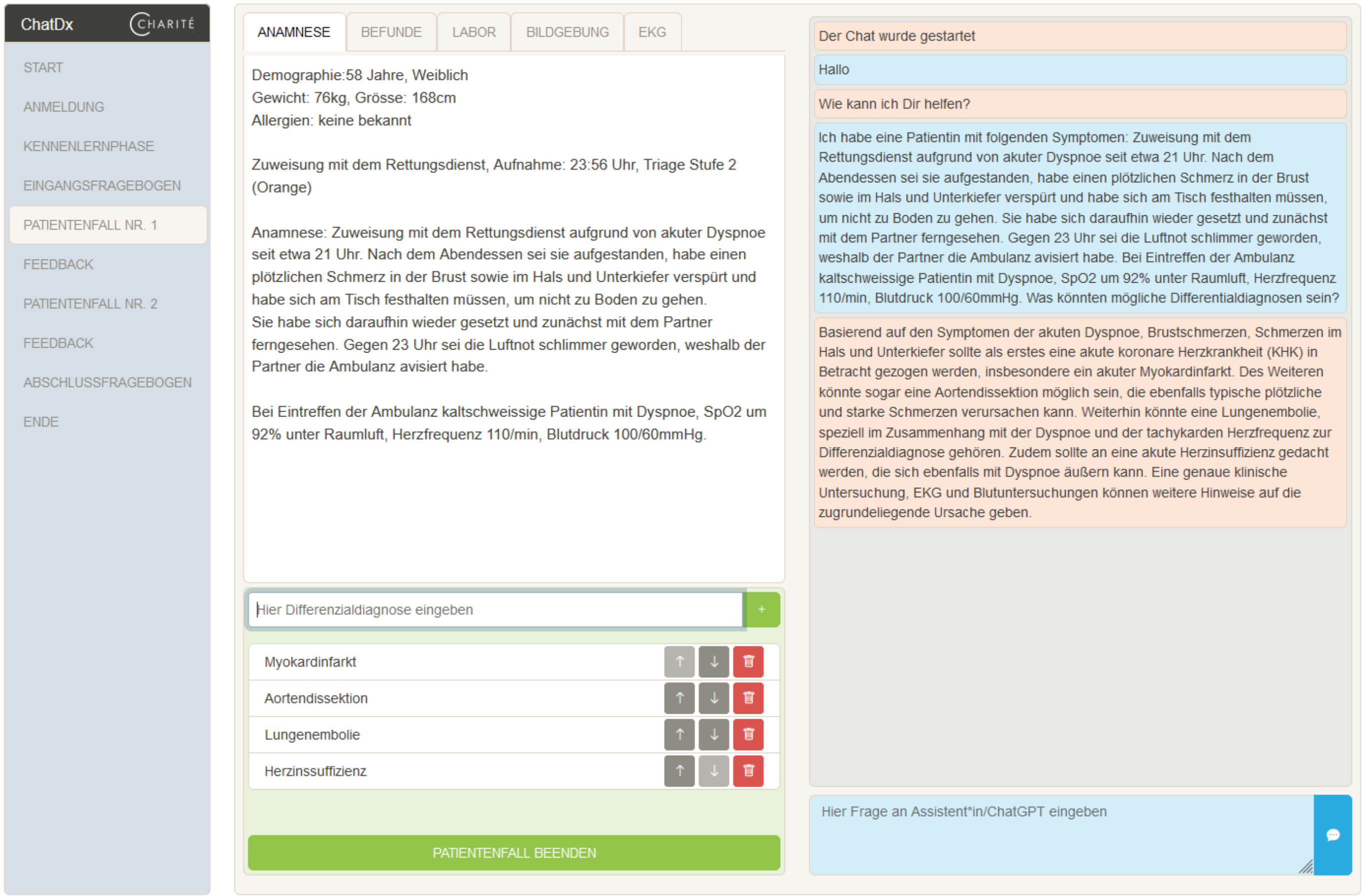
The experimental session starts with a get-to-know phase designed to acquaint participants with their respective mode of assistance, whether the human coach 'Toni' or ChatGPT. This short introduction highlights the strengths of each coach, such as Toni's background in medicine, including successful completion of medical studies and practical medical experience, and ChatGPT's expansive knowledge base (see online supplemental information). Participants are also made aware of the limitations inherent to each coach, such as Toni's potential knowledge gaps compared with a senior physician and the possibility of 'hallucinations' with ChatGPT. This initial step is crucial in addressing participants' onboarding needs, facilitating their evaluation of the capabilities and intentions of their human or ChatGPT coach.[57] By establishing familiarity and understanding of the strengths and limitations, participants can begin to develop trust in their respective coach, which is vital for effective collaboration and decision-making.[58] The get-to-know phase does not contain any examples of when and how to interact with the coach, which is only part of the training.

Training

Afterward, participants either see the training instructions on the screen (training condition) or not (no-training condition), depending on the subgroup they are randomly assigned to. The training instructions are designed to heighten awareness regarding the potential for diagnostic errors and delineate three prevalent factors contributing to diagnostic errors: limited knowledge, premature closure and overconfidence.[1 59] These are briefly explained. Additionally, the instructions provide exemplar inquiries that participants may pose to their respective coach (whether human or ChatGPT) to effectively navigate these three challenges (see online supplemental information for complete instructions). The training instructions are no longer available once the participant proceeds to the next page.

Task: diagnose cases

The main task is then to diagnose two patient cases (in random order). The cases are based on published cases of real patients[43 60] and represent ambiguous emergency cases with a known correct diagnosis but a main competing diagnosis that has to be considered (case 1: pulmonary embolism vs myocardial infarction; case 2: aortic dissection vs stroke). On the patient case page, patient information including ECGs, laboratory results of blood samples and patient history is presented in a patient chart. On the same page, participants have access to a field in which to chat with their coaches, who reply in real time. Participants are instructed not to use any other sources of information than those on the screen. Participants are asked to record all differential diagnoses considered in a separate field on the same page. All clicks, chats and entries are logged with time stamps. Figure 2 shows a screenshot of a patient case page (in



Demographie:58 Jahre, Weiblich
Gewicht: 76kg, Grösse: 168cm
Allergien: keine bekannt

Zuweisung mit dem Rettungsdienst, Aufnahme: 23:56 Uhr, Triage Stufe 2 (Orange)

Anamnese: Zuweisung mit dem Rettungsdienst aufgrund von akuter Dyspnoe seit etwa 21 Uhr. Nach dem Abendessen sei sie aufgestanden, habe einen plötzlichen Schmerz in der Brust sowie im Hals und Unterkiefer verspürt und habe sich am Tisch festhalten müssen, um nicht zu Boden zu gehen.
Sie habe sich daraufhin wieder gesetzt und zunächst mit dem Partner ferngesehen. Gegen 23 Uhr sei die Luftnot schlimmer geworden, weshalb der Partner die Ambulanz avisiert habe.

Bei Eintreffen der Ambulanz kaltschweissige Patientin mit Dyspnoe, SpO2 um 92% unter Raumluft, Herzfrequenz 110/min, Blutdruck 100/60mmHg.

- Myokardinfarkt
- Aortendissektion
- Lungenembolie
- Herzinssuffizienz

Der Chat wurde gestartet

Hallo

Wie kann ich Dir helfen?

Ich habe eine Patientin mit folgenden Symptomen: Zuweisung mit dem Rettungsdienst aufgrund von akuter Dyspnoe seit etwa 21 Uhr. Nach dem Abendessen sei sie aufgestanden, habe einen plötzlichen Schmerz in der Brust sowie im Hals und Unterkiefer verspürt und habe sich am Tisch festhalten müssen, um nicht zu Boden zu gehen. Sie habe sich daraufhin wieder gesetzt und zunächst mit dem Partner ferngesehen. Gegen 23 Uhr sei die Luftnot schlimmer geworden, weshalb der Partner die Ambulanz avisiert habe. Bei Eintreffen der Ambulanz kaltschweissige Patientin mit Dyspnoe, SpO2 um 92% unter Raumluft, Herzfrequenz 110/min, Blutdruck 100/60mmHg. Was könnten mögliche Differentialdiagnosen sein?

Basierend auf den Symptomen der akuten Dyspnoe, Brustschmerzen, Schmerzen im Hals und Unterkiefer sollte als erstes eine akute koronare Herzkrankheit (KHK) in Betracht gezogen werden, insbesondere ein akuter Myokardinfarkt. Des Weiteren könnte sogar eine Aortendissektion möglich sein, die ebenfalls typische plötzliche und starke Schmerzen verursachen kann. Weiterhin könnte eine Lungenembolie, speziell im Zusammenhang mit der Dyspnoe und der tachykarden Herzfrequenz zur Differenzialdiagnose gehören. Zudem sollte an eine akute Herzinsuffizienz gedacht werden, die sich ebenfalls mit Dyspnoe äußern kann. Eine genaue klinische Untersuchung, EKG und Blutuntersuchungen können weitere Hinweise auf die zugrundeliegende Ursache geben.

Figure 2 Screenshot of a patient case page. Starting on the left, there is a window showing the current step within the experiment and the patient chart with several subcategories, above the field for entering the differential diagnoses; on the right is the chat window (here, in the artificial intelligence condition).

German). When leaving the patient case page, participants are asked to assess the likelihood of each diagnosis generated (on a Visual Analogue Scale of 0–100), to provide a reason for their most likely diagnosis (open answer) and to report their intended next steps if this were a real patient (open answer).

Human versus AI coach

The LLM used in this study is OpenAI's ChatGPT (version gpt-4–0613, DeploymentName='GPT-4', MaxTokens=1000, Temperature=1.0f), accessed via the application programming interface provided by Microsoft Azure's cloud platform (hosted in the 'Switzerland North' data centre).

The human coach is randomly drawn from the two medical interns who serve as coaches and who received a 5-hour training on the study purpose, the chat system and the philosophy of peer teaching[30] and deliberate reflection,[61] as well as scripts with standard answers to frequent requests (as identified in a pilot phase) to ensure that they could reply quickly and in a standardised way. Both human coaches are introduced by the unisex name 'Toni' to avoid potential gender bias and to keep their identities confidential. Human coaches sit at their computer at home and chat via the experimental interface with the participant. The interface was created using Microsoft's 'Blazor Server App' web framework. Both ChatGPT and the human coach received the instruction to act as a medical coach and accompany fourth-year medical students through the diagnostic process, including asking guiding questions such as 'Which findings support/oppose your hypothesis?' following the logic of deliberate reflection[61 62] (for the complete instructions, see system prompt in online supplemental information).

Questionnaire per case

Following each patient case, participants respond to questions pertaining to their case perception, encompassing factors such as perceived difficulty and familiarity with the diagnosis, as well as their assessment of the competence and support provided by the coaches (online supplemental e table1).

Final questionnaire

A final questionnaire is administered after completion of both patient cases to assess the perceived usefulness of,[63] satisfaction with[64] and credibility of the coaches.[65]

Debriefing

On re-entering the virtual meeting room, participants are told about future debriefing, thanked and dismissed by the experimenter. Following the data collection phase, a comprehensive written debriefing will be provided. This debriefing will include solutions to the patient cases, an information package containing the training instructions (also in the no-training condition), as well as links to additional resources on clinical reasoning and LLMs.

Pilot study

In a pilot study involving N=11 fourth-year medical students and medical interns (M_{age}=26 years, SD=4.9, 55% female), the case material was tested for intelligibility and feasibility without assistance from a human coach or ChatGPT. Diagnoses were elicited as free text responses. For case 1, the correct diagnosis (pulmonary embolism) was listed by 27% of participants as the most likely diagnosis, and in case 2, the correct diagnosis (aortic dissection) by 0%, confirming that we had adequately selected difficult cases to prevent any ceiling effects.

Data to be analysed

Data will be in the form of questionnaires, process measures (eg, timestamps of clicks), chat protocols and ratings. Data will be entered into a web-based database that fulfils the requirements of the Swiss Human Research Act. Participants will be asked to generate a 'study ID,' which guarantees their anonymity but allows for matching baseline surveys with the data collected during the experimental session. All data will be digital. Only authorised study personnel will have access to personal information (eg, email address) during data collection. Any data shared with external parties (eg, collaborators) will be deidentified to remove all personally identifiable information. Only anonymised, coded data will be published together with DOIs in the OSF repository to make them findable. Primary and secondary endpoints as well as control variables are listed in table 1.

Statistical analyses

Data analysis will be conducted with R.[55] For statistical analyses, we will use generalised linear mixed models (GLMMs), complemented by suitable post hoc techniques, particularly for subgroup analyses. Standard descriptive statistics and graphical representations will be employed, along with normality testing to assess assumptions for the proper application of parametric testing methodologies. Prior to data analysis, data quality will be checked by, for example, range checks for data values. To evaluate the randomisation procedure to the conditions, we will compare the four groups regarding their demographics (eg, age, gender, prior experience with LLMs) with ANOVAs. To determine whether participants in the training condition read the training instructions, we will compare the time they spent on the page with a minimum reading time threshold. This threshold will be set slightly below the average time spent on the page by participants in the no-training condition.

To determine the accuracy of the differential diagnoses, first, they will be automatically coded to International Classification of Diseases (10th revision; ICD-10) codes using a proprietary German-language natural language processing engine (Averbis Health Discovery, https://averbis.com), which maps ICD-10-German modification codes to unstructured text. 50% of the diagnoses will be randomly selected for cross-checking by two expert raters, blinded to the condition, to ensure the accuracy of the

Table 1 Overview of variables of interest

Variable	Values	Explanation, example	Type of data
Primary endpoints			
Diagnostic accuracy of the most likely diagnosis/3 most likely diagnoses	Range 0–n	Number of steps required within the ICD taxonomy to get from the listed diagnosis to the correct diagnosis[70]	Automated coding (see text), checked by 2 blinded physicians
Number of diagnoses	Range 1–n	Number of differential diagnoses generated	Process measure
Secondary endpoints			
Information search	Range 1–5	Number of pieces of diagnostic information acquired	Process measure
Diversity of diagnoses		Average distance between ICD-10 codes of differential diagnoses generated	Calculated post hoc
Confidence in diagnoses	Range 0–100	Rated likelihood per diagnosis per case	Survey
Satisfaction with coach/perceived usefulness of coach	Range 1–5	Average of 2 items	Survey
Time on case		Duration in min:sec	Process measure
Control variables			
Gender	male, female, other		Survey
Medical knowledge	Range 0–200; median split (competent vs less competent)	Average score of last 3 progress tests medicine[71]	Survey
Experience working with LLMs		Descriptive statistics of frequency, confidence	Survey
General trust in LLMs	Range 1–5	Average value of 6-item scale	Survey
Credibility of ChatGPT during case	Range 1–5	Average value of 5-item credibility subscale of the TMS scale[65]	Survey
Exploratory analyses of			
Type of reasons for most likely diagnosis	Categories to be determined	For example, occurrence of typical symptoms, advice of coach, guessing	Categorised by 2 trained blinded raters
Type of next steps	Categories to be determined	For example, initiation of specific therapeutic steps, further consultation with senior physician	Categorised by 2 trained blinded raters
Perception of case and coach	Range 1–7	Descriptive statistics of case difficulty, perceived own and coach expertise	Survey
Chat with coach	Categories to be determined	For example, time point, frequency of different types of requests (eg, confirmation, falsification), length of chats	Categorised by 2 trained blinded raters
Accuracy of answers of coach		Type of mistakes by human coaches vs ChatGPT	Categorised by 2 trained blinded raters
Reactions to mistakes by participant	Categories to be determined	For example, clarifying questions	Categorised by 2 trained blinded raters
Impact of coach on diagnostic list	Categories to be determined	For example, adding or removing diagnoses	Categorised by 2 trained blinded raters

ChatGPT, OpenAI's generative pre-trained transformer; ICD, International Classification of Diseases; LLM, large language model; TMS, transactive memory system.

automated ICD matching. If accuracy of this automated matching turns out to be below 95%, the proportion will be increased to 60%, 70% and so forth for human cross-checking. Then, these codes will be compared with the correct codes of the two cases. Accuracy will be calculated as the number of steps required within the ICD taxonomy to get from one diagnosis to the other, as described elsewhere.[62]

To assess the impact of the type of assistance and training on the primary and secondary outcome variables, we will conduct successively more complex GLMMs,[54] starting with participant ID and item ID as random intercepts, and gender and conditions as fixed effects. The dependent variables will include diagnostic accuracy, the number of differential diagnoses and the secondary endpoints (see table 1). Sensitivity analyses are planned to check the robustness of our findings. These will include alternative model specifications, assessing interaction effects, applying different methods for handling missing data (eg, imputation methods, complete case analysis) and subgroup analyses. For example, we will successively include more control variables, such as participants' medical competence[41 42] and general trust in LLMs,[58 66] to account for potential confounders and gain a deeper understanding of the conditions under which LLMs are most effective.

In preparation for the qualitative analysis of prompts and usage patterns of coaches, all chat interactions and open answers will be coded using MAXQDA software. Coding categories (eg, confirmatory or knowledge questions) will be derived inductively and deductively by trained raters with domain knowledge. Two trained raters will independently code the material, blinded to the conditions (human coach vs ChatGPT, training vs no training). Rater agreement will be reported as coefficient kappa. Exploratory analyses and subgroup analyses will be conducted to characterise successful and unsuccessful prompts and the differences between consulting a human coach versus ChatGPT. Further, the timing of using the coach (early or late in the process), the frequency and type of errors made by the coaches and the impact of the (correct or incorrect) diagnoses proposed by the coaches on the diagnoses listed by the participants will be explored.

Patient and public involvement

We intend to disseminate the main results to the participants and public in a format that is suitable for a non-specialist audience. There was no patient nor public involvement in the design and conduct of the study.

ETHICS AND DISSEMINATION

This is a prospective, randomised controlled experimental study. Participant anonymity for participants will be respected at all times by anonymisation of their data. The Bern Cantonal Ethics Committee considered the study exempt from full ethical review (BASEC No: Req-2023-01396). All methods will be carried out in accordance with relevant guidelines and regulations. All students will participate voluntarily and will sign an informed consent after receiving written and oral information about the study.

Results will be presented at scientific meetings. Results will be published in peer-reviewed scientific journals and authorship will be determined according to International Committee of Medical Journal Editors guidelines.

DISCUSSION

Our study has several strengths. First, it is a prospective randomised controlled experiment involving advanced medical students diagnosing complex patient cases, allowing us to investigate both diagnostic outcomes and processes. Second, the study compares consultations with either an LLM or a human coach, both of which are practically relevant advisors for medical students solving complex cases. Third, the detailed analysis of both the diagnostic process and its outcomes will provide a deeper insight into the research findings.

Our study also has several limitations. First, it focuses solely on fourth-year medical students, which may restrict the generalisability of the results to a broader medical student population or to residents and practising physicians. Also, the study is set within a medical education context, involving complex cases that are challenging for this level of training. Second, only approximately half of our questionnaires have been validated by previous research. This is due to the lack of suitable instruments, given the novelty of our study's focus. For instance, we were unable to find scientifically validated questions that assess trust in an AI chat partner. Third, although we plan to conduct in-depth qualitative analyses of the interactions between participants and either human coaches or ChatGPT, insights into the underlying mechanisms of how AI influences decision-making processes will still be limited to our setting. More research in various medical (education) contexts is needed to better understand the way users perceive and interact with AI tools.[24 25 31 67 68] Last, we acknowledge that integrating AI into medical diagnostics is not just a technological upgrade but also introduces complex ethical dilemmas and practical implementation challenges that require thorough exploration.[19 69] In our study, we point participants to the limitations and potential biases of ChatGPT (and human coaches), but any considerations to integrate ChatGPT into medical education need to be accompanied by additional ethical considerations and dedicated training programmes as part of the medical curriculum.

Author affiliations
[1]Department of Emergency Medicine, Inselspital University Hospital Bern, University of Bern, Bern, Switzerland
[2]Institute for Medical Informatics (I4MI), Bern University of Applied Sciences, Bern, Switzerland
[3]Department of Anesthesiology and Operative Intensive Care Medicine CCM & CVK, Charité Universitätsmedizin Berlin, Berlin, Germany
[4]Lernzentrum (Skills Lab), Charité Universitätsmedizin Berlin, Berlin, Germany
[5]Department of Management, Technology, and Economics, ETH Zurich, Zurich, Switzerland

X Juliane E Kämmer @julianekaemmer

Acknowledgements The authors would like to thank Anita Todd for language editing as well as Tobias Bolte, Phillip Cyrenius, Robin Runge, Leonie Dahms and Anna Wittenstein for their support with data collection.

Contributors JEK, WEH, GK, TCS, DP, TB and NB designed the study. GK programmed the experiment. JEK and NB will analyse the data. JEK wrote the manuscript. All authors revised the manuscript critically. All authors gave their approval of the final version of the manuscript and agreed to be accountable for all aspects of the published work. JEK is the guarantor. ChatGPT3.5 was used for language editing for selected parts of the manuscript to improve language, and input was cross-checked by a native speaker.

Funding This work is supported by funding from the Swiss National Science Foundation NRP77 Digital Transformation Programme (Grant no. 187331).

Disclaimer The funding organisation had no role in the design of the study, collection, analysis, and interpretation of data, or in writing the manuscript.

Competing interests None declared.

Patient and public involvement statement Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

Patient consent for publication Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

ORCID iDs

Juliane E Kämmer http://orcid.org/0000-0001-6042-8453
Wolf E Hautz http://orcid.org/0000-0002-2445-984X
Gert Krummrey http://orcid.org/0000-0002-8397-2336
Thomas C Sauter http://orcid.org/0000-0002-6646-5789
Dorothea Penders http://orcid.org/0000-0003-1795-3791
Tanja Birrenbach http://orcid.org/0000-0002-3046-0900
Nadine Bienefeld http://orcid.org/0000-0003-4200-8695

REFERENCES

1 Berner ES, Graber ML. Overconfidence as a cause of diagnostic error in medicine. *Am J Med* 2008;121:S2–23.

2 Newman-Toker DE, Peterson SM, Badihian S, *et al*. Diagnostic errors in the emergency department: a systematic review. In: *Agency for healthcare research and quality (AHRQ)*. 2022. Available: https://effectivehealthcare.ahrq.gov/products/diagnostic-errors-emergency/research

3 Singh H, Meyer AND, Thomas EJ. The frequency of diagnostic errors in outpatient care: estimations from three large observational studies involving US adult populations. *BMJ Qual Saf* 2014;23:727–31.

4 Miller BT, BaloghEP, eds. Committee on diagnostic error in health care, board on health care services, Institute of medicine, the National academies of sciences, engineering, and medicine. In: *Improving diagnosis in health care*. Washington, D.C: National Academies Press, 2015. Available: http://www.nap.edu/catalog/21794 [accessed 15 Nov 2019].

5 Hautz WE, Kämmer JE, Hautz SC, *et al*. Diagnostic error increases mortality and length of hospital stay in patients presenting through the emergency room. *Scand J Trauma Resusc Emerg Med* 2019;27:54.

6 Zwaan L, de Bruijne M, Wagner C, *et al*. Patient record review of the incidence, consequences, and causes of diagnostic adverse events. *Arch Intern Med* 2010;170:1015–21.

7 Wallace E, Lowry J, Smith SM, *et al*. The epidemiology of malpractice claims in primary care: a systematic review. *BMJ Open* 2013;3:e002929.

8 Riches N, Panagioti M, Alam R, *et al*. The effectiveness of electronic differential diagnoses (DDX) generators: a systematic review and meta-analysis. *PLOS ONE* 2016;11:e0148991.

9 Gilson A, Safranek CW, Huang T, *et al*. How does ChatGpt perform on the United States medical licensing examination? The implications of large language models for medical education and knowledge assessment. *JMIR Med Educ* 2023;9:e45312.

10 Kung TH, Cheatham M, Medenilla A, *et al*. Performance of ChatGpt on USMLE: potential for AI-assisted medical education using large language models. *PLOS Dig Health* 2023;2:e0000198.

11 Berşe S, Akça K, Dirgar E, *et al*. The role and potential contributions of the artificial intelligence language model ChatGpt. *Ann Biomed Eng* 2024;52:130–3.

12 Goh E, Gallo R, Hom J, *et al*. Influence of a large language model on diagnostic reasoning: a randomized clinical vignette study health Informatics. *medRxiv* [Preprint] 2024.

13 Lee P, Bubeck S, Petro J. Limits, and risks of GPT-4 as an AI chatbot for medicine. *N Engl J Med* 2023;388:1233–9.

14 Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. *Healthc (Basel)* 2023.:887.

15 Wachter RM, Brynjolfsson E. Will generative artificial intelligence deliver on its promise in health care? *JAMA* 2024;331:65.

16 Alessandri Bonetti M, Giorgino R, Gallo Afflitto G, *et al*. How does ChatGpt perform on the Italian residency admission national exam compared to 15,869 medical graduates? *Ann Biomed Eng* 2023;52:745–9.

17 Scaioli G, Lo Moro G, Conrado F, *et al*. Exploring the potential of ChatGpt for clinical reasoning and decision-making: a cross-sectional study on the Italian medical residency exam. *Ann Ist Super Sanita* 2023;59:267–70.

18 Ferdush J, Begum M, Hossain ST. ChatGpt and clinical decision support: scope, application, and limitations. *Ann Biomed Eng* 2023;52:1119–24.

19 Bienefeld N, Boss JM, Lüthy R, *et al*. Solving the explainable AI conundrum by bridging clinicians' needs and developers' goals. *NPJ Digit Med* 2023;6:94.

20 Bienefeld N, Kolbe M, Camen G, *et al*. Human-AI teaming: leveraging transactive memory and speaking up for enhanced team effectiveness. *Front Psychol* 2023;14:1208019.

21 Kerstan S, Bienefeld N, Grote G. Choosing human over AI doctors? How comparative trust associations and knowledge relate to risk and benefit perceptions of AI in healthcare. *Risk Anal* 2024;44:939–57.

22 Marcin T, Hautz SC, Singh H, *et al*. Effects of a computerised diagnostic decision support tool on diagnostic quality in emergency departments: study protocol of the DDx-BRO multicentre cluster randomised cross-over trial. *BMJ Open* 2023;13:e072649.

23 Grunhut J, Marques O, Wyatt ATM. Needs, challenges, and applications of artificial intelligence in medical education curriculum. *JMIR Med Educ* 2022;8:e35587.

24 Zhang S, Yu J, Xu X. Rethinking human-AI collaboration in complex medical decision making: a case study in sepsis diagnosis. In: Proceedings of the CHI Conference on Human Factors in Computing Systems, CHI 24; Honolulu HI USA, May 11, 2024:1–18. 10.1145/3613904.3642343 Available: https://dl.acm.org/doi/proceedings/10.1145/3613904

25 Blease C, Worthen A, Torous J. Psychiatrists' experiences and opinions of generative artificial intelligence in mental healthcare: an online mixed methods survey. *Psychiatry Res* 2024;333:115724.

26 Hirosawa T, Harada Y, Yokose M, *et al*. Diagnostic accuracy of differential-diagnosis lists generated by generative pretrained transformer 3 chatbot for clinical vignettes with common chief complaints: a pilot study. *Int J Environ Res Public Health* 2023;20:3378.

27 Norman GR, Monteiro SD, Sherbino J, *et al*. The causes of errors in clinical reasoning: cognitive biases, knowledge deficits, and dual process thinking. *Acad Med* 2017;92:23–30.

28 Lu Z, Wang D, Yin M. Does more advice help? The effects of second opinions in AI-assisted decision making. *Proc ACM Hum-Comput Interact* 2024;8:1–31.

29 Kämmer JE, Choshen-Hillel S, Müller-Trede J, *et al*. A systematic review of empirical studies on advice-based decisions in behavioral and organizational research. *Decision*2023;10:107–37.

30 Ten Cate O, Durning S. Dimensions and psychology of peer teaching in medical education. *Med Teach* 2007;29:546–52.

31 Mollick ER, Mollick L. Assigning AI: seven approaches for students, with prompts. *SSRN J* 2023.

32 Hautz WE, Hautz SC, Kämmer JE. Whether two heads are better than one is the wrong question (though sometimes they are). *Adv Health Sci Educ Theory Pract* 2020;25:905–11.

33 Graber ML, Franklin N, Gordon R. Diagnostic error in internal medicine. *Arch Intern Med* 2005;165:1493–9.

34 Newman-Toker DE, Schaffer AC, Yu-Moe CW, *et al*. Serious misdiagnosis-related harms in malpractice claims: the 'big three'–vascular events, infections, and cancers. *Diagnosis (Berl)* 2019;6:227–40.

35 Norman GR. Research in clinical reasoning: past history and current trends. *Med Educ* 2005;39:418–27.

36 Pelaccia T, Tardif J, Triby E, *et al*. How and when do expert emergency physicians generate and evaluate diagnostic hypotheses? A qualitative study using head-mounted video cued-recall interviews. *Ann Emerg Med* 2014;64:575–85.

37 Evans KK, Haygood TM, Cooper J, *et al*. A half-second glimpse often lets radiologists identify breast cancer cases even when viewing the mammogram of the opposite breast. *Proc Natl Acad Sci U S A* 2016;113:10292–7.

38 Kostopoulou O, Mousoulis C, Delaney B. Information search and information distortion in the diagnosis of an ambiguous presentation. *Judgm decis mak* 2009;4:408–19.

39 Kostopoulou O, Russo JE, Keenan G, *et al*. Information distortion in physicians' diagnostic judgments. *Med Decis Making* 2012;32:831–9.

40 Kourtidis P, Nurek M, Delaney B, *et al*. Influences of early diagnostic suggestions on clinical reasoning. *Cogn Res Princ Implic* 2022;7:103.

41 Barrows HS, Norman GR, Neufeld VR, *et al*. The clinical reasoning process of randomly selected physicians in general medical practice. *Clin Invest Med* 1982;5:49–55.

42 Hobu PPM, Schmidt HG, Boshuizen HPA, *et al*. Contextual factors in the activation of first diagnostic hypotheses: expert-novice differences. *Med Educ* 1987;21:471–6.

43 Kumar B, Kanna B, Kumar S. The pitfalls of premature closure: clinical decision-making in a case of aortic dissection. *BMJ Case Rep* 2011;2011:bcr0820114594.

44 Kruglanski AW, Webster DM. Motivated closing of the mind: 'seizing' and 'freezing'. *Psychol Rev* 1996;103:263–83.

45 Eva KW, Cunnington JPW. The difficulty with experience: does practice increase susceptibility to premature closure. *J Contin Educ Health Prof* 2006;26:192–8.

46 Norman G. The bias in researching cognitive bias. *Adv in Health Sci Educ* 2014;19:291–5.

47 Saposnik G, Redelmeier D, Ruff CC, *et al*. Cognitive biases associated with medical decisions: a systematic review. *BMC Med Inform Decis Mak* 2016;16:138.

48 Ely JW, Kaldjian LC, D'Alessandro DM. Diagnostic errors in primary care: lessons learned. *J Am Board Fam Med* 2012;25:87–97.

49 Singh H, Giardina TD, Meyer AND, *et al*. Types and origins of diagnostic errors in primary care settings. *JAMA Intern Med* 2013;173:418–25.

50 Kostopoulou O, Lionis C, Angelaki A, *et al*. Early diagnostic suggestions improve accuracy of family physicians: a randomized controlled trial in Greece. *Fam Pract* 2015;32:323–8.

51 Kostopoulou O, Devereaux-Walsh C, Delaney BC. Missing celiac disease in family medicine: the importance of hypothesis generation. *Med Decis Making* 2009;29:282–90.

52 Meskó B. Prompt engineering as an important emerging skill for medical professionals: tutorial. *J Med Internet Res* 2023;25:e50638.

53 Nori H, Lee YT, Zhang S, *et al*. Can generalist foundation models outcompete special-purpose tuning? Case study in medicine. 2023. Available: http://arxiv.org/abs/2311.16452

54 Bates D, Mächler M, Bolker B, *et al*. Fitting linear mixed-effects models using Lme4. *J Stat Softw* 2014.

55 R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing; 2018. Available: https://www.R-project.org/

56 Faul F, Erdfelder E, Lang A-G, *et al*. G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behav Res Methods* 2007;39:175–91.

57 Cai CJ, Winter S, Steiner D, *et al*. Hello AI: uncovering the onboarding needs of medical practitioners for human-AI collaborative decision-making. *Proc ACM Hum-Comput Interact* 2019;3:1–24.

58 Schrah GE, Dalal RS, Sniezek JA. No decision-maker is an island: integrating expert advice with information acquisition. *J Behav Decis Making* 2006;19:43–60.

59 Gäbler M. Denkfehler BEI diagnostischen entscheidungen. *Wien Med Wochenschr* 2017;167:333–42.

60 Kunina-Habenicht O, Hautz WE, Knigge M, *et al*. Assessing clinical reasoning (ASCLIRE): instrument development and validation. *Adv Health Sci Educ* 2015;20:1205–24.

61 Mamede S, Schmidt HG, Penaforte JC. Effects of reflective practice on the accuracy of medical diagnoses. *Med Educ* 2008;42:468–75.

62 Mamede S, Schmidt HG. Deliberate reflection and clinical reasoning: founding ideas and empirical findings. *Med Educ* 2023;57:76–85.

63 Nagendran M, Festor P, Komorowski M, *et al*. Quantifying the impact of AI recommendations with explanations on prescription decision making. *NPJ Digit Med* 2023;6:206.

64 Jo H, Bang Y. Analyzing ChatGpt adoption drivers with the TOEK framework. *Sci Rep* 2023;13:22606.

65 Lewis K. Measuring transactive memory systems in the field: scale development and validation. *J Appl Psychol* 2003;88:587–604.

66 Lee J, Moray N. Trust, control strategies and allocation of function in human-machine systems. *Ergonomics* 1992;35:1243–70.

67 Wang D, Churchill E, Maes P, *et al*. From human-human collaboration to human-AI collaboration: designing AI systems that can work together with people. Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems; August 22, 2020:1–6. 10.1145/3334480.3381069 Available: https://dl.acm.org/doi/10.1145/3334480.3381069

68 Tangadulrat P, Sono S, Tangtrakulwanich B. Using ChatGpt for clinical practice and medical education: cross-sectional survey of medical students' and physicians' perceptions. *JMIR Med Educ* 2023;9:e50658.

69 Bienefeld N, Keller E, Grote G. Human-AI teaming in the ICU: a comparative analysis of data scientists' and clinicians' assessments on AI augmentation and automation at work. *J Med Internet Res* [Preprint].

70 Hautz WE, Kündig MM, Tschanz R, *et al*. Automated identification of diagnostic labelling errors in medicine. *Diagnosis (Berl)* 2021;9:241–9.

71 Osterberg K, Kölbel S, Brauns K. The progress test medizin. *GMS J Med Educ* 2006;23:Doc46.