AI Clinical Reasoning Training Agent on Medical Students' Clinical Reasoning Skills and Case-based Learning Experience: A Cluster Randomized Controlled Trial

Protocol

Principal Institution: Peking Union Medical College Hospital, Chinese Academy of Medical Sciences & Peking Union Medical College

Principal Investigator: Professor Li Yue, Department of Education / Department of Gastroenterology, Peking Union Medical College Hospital, Chinese Academy of Medical Sciences & Peking Union Medical College

2026-03-30

1 Background

Traditional medical education has long emphasized one-way transmission of theoretical knowledge, which presents limitations in the systematic cultivation of clinical reasoning skills among medical students. Miller's pyramid of clinical competence divides the learning process into four levels: knows, knows how, shows how, and does. It emphasizes the gradual transformation from theoretical knowledge to clinical practice ability [1]. Within this framework, effective training of clinical reasoning not only relies on theoretical instruction, but also requires repeated case-based training and feedback.

Case-based learning (CBL), as a teaching method centered on real or simulated clinical cases, is a key strategy to address the above limitations. By guiding students to analyze, discuss, and solve clinical problems in cases, CBL promotes the deep integration of theoretical knowledge and clinical decision-making, thereby systematically exercising students' clinical reasoning skills [2]. CBL is not merely a supplement to theoretical teaching; it is an indispensable practical step in the transition of clinical reasoning from "knowing" to "doing".

In recent years, artificial intelligence (AI) has gradually emerged in medical education, particularly showing potential in simulating clinical scenarios and providing personalized feedback. AI-assisted clinical reasoning training tools can overcome time and space constraints, and offer students repeatable, adaptive, and real-time feedback case training, thereby reinforcing the sustained role of CBL in clinical reasoning development. Currently, it still lacks high-quality evidence from randomized controlled trials on the impact of AI agents on medical students' clinical reasoning skills.

This study plans to evaluate the impact of an AI clinical reasoning training agent on students' clinical reasoning training outcomes and CBL learning experience. We will conduct a cluster randomized controlled trial (cRCT) in a real teaching environment. We hope to provide empirical evidence and replicable implementation experience for the integration of AI and medical education.

2 Objectives

Primary Objective: To evaluate the impact of the AI agent on student learning outcomes (clinical reasoning test scores, clinical scenario simulation scores).

Secondary Objective: To investigate students' learning experience with the AI agent, including average learning time per case, course exam review time, and AI technology acceptance (perceived usefulness, perceived ease of use, satisfaction, and intention to use).

3 Methods

3.1 Study Design

This study adopts a two-arm parallel cluster randomized controlled trial design. The trial is designed and reported in accordance with the CONSORT statement [3].

3.2 Study Population and Sample Size

The study population will recruit Class of 2021 medical students (8-year program) from Peking Union Medical College and Class of 2020 medical students (8-year program) from Tsinghua University School of Medicine. Both cohorts are officially enrolled in the "Comprehensive Clinical Course" for the 2025-2026 academic year, have consistent foundational knowledge in basic medicine and diagnostics, and are in the phase of clinical medicine theory learning, not yet having entered clinical practice.

Inclusion criteria: (1) full-time registered and enrolled in the "Comprehensive Clinical Course"; (2) signed informed consent, voluntary participation in this study and completion of relevant tests and questionnaires. Exclusion criteria: planned suspension of studies, withdrawal, or major transfer during the study period.

The sample size calculation formula per arm is $N=DE\times[2\times(Z_{1-\alpha/2}+ Z_{1-\beta})^2\times\sigma^2/\delta^2]$ [4]. Considering clinical reasoning test scores and clinical scenario simulation scores as primary outcomes, the difference in mean scores between intervention and control groups is set at $\delta=2$ points (out of 100) based on teaching experience; the standard deviation $\sigma=3$ points is set based on students' baseline diagnostic scores. Using dormitory as the smallest cluster unit, with average cluster size $M=3$. Based on previous studies [5,6] and considering that small cluster sizes have little impact on total sample size, the intracluster correlation coefficient (ICC, ρ) is set at 0.02, giving

a design effect DE=1+(M-1)×ICC=1.04. The significance level α=0.05 (two-sided), statistical power 1-β=0.8, $Z_{1-\alpha/2}$=1.96, $Z_{1-\beta}$=0.84. Using PASS 2025 software, the sample size per arm for the cRCT is 39, with number of clusters per arm K=N/M =13, Considering a 10% attrition or exclusion rate, the target recruitment is 88 participants.

Recruitment will be conducted through class information sessions, where the study purpose and design are explained in detail. Participating students who enroll and complete the study will receive an incentive. Upon enrollment, all students in both intervention and control groups will complete a questionnaire on AI literacy and AI use interest as baseline data. The AI literacy section is designed based on the "Expert Consensus on the Artificial Intelligence Proficiency Competency List and Assessment Framework for Medical Students (2025 Edition) [7].

3.3 Randomization and Blinding

Considering potential heterogeneity in baseline between students from the two schools, and possible contamination due to discussions among dormitory mates during the intervention, this study will adopt stratified cluster randomization, first stratifying by school, then using dormitory as the smallest randomization unit. A random number will be generated for each dormitory using WPS spreadsheet; dormitories will be sorted by the random number, with the first half allocated to the intervention group and the second half to the control group. The allocation scheme will be kept by research personnel not directly involved in teaching. Participants' group assignment will be revealed via unique student ID only after baseline data collection and informed consent are completed. Participants cannot be blinded to their use of the AI tool during the intervention but will be instructed not to share accounts or learning materials across groups. The primary outcome assessor will be blinded.

3.4 Intervention and Control

This study will select five topics from the "Comprehensive Clinical Course": "Infectious Diarrhea," "Viral Hepatitis," "Bloodstream Infection," "Infective Endocarditis," and "Central Nervous System Infection". Standardized cases will be provided by the teaching faculty, with two cases per topic, totaling 10 cases. The case design template consists of case information, supplementary diagnostic and treatment

process information, clinical reasoning questions, and answer keys, breaking down the entire diagnostic and treatment process into multiple steps. The cases will be reviewed and approved by two senior clinical faculty members to ensure appropriate content and difficulty level. These 10 cases will be developed into an AI agent on the Rain Classroom platform. The agent guides students step-by-step through history taking, physical examination, ancillary test selection, diagnostic reasoning, etc., through interactive dialogue, providing real-time personalized feedback. It is accessible on mobile and computer devices and supports voice interaction. In developing the agent, special emphasis is placed on the "reasoning guidance" function, limiting scenarios where "students ask for answers and get them directly." The answer key (script) will be released systematically after training, highlighting the roles of the learning materials as a "reference book" and the agent as a "practice field." Multiple rounds of testing with teachers and students will be conducted to fix potential issues in logic, interaction, stability, etc.

3.4.1 Intervention Measures

After class, the AI agent training tasks will be sent to students in the intervention group. Each intervention group student is required to complete training on all 10 cases. The system backend will automatically record each student's AI interaction logs (e.g., training duration, number of interactions, logical order of history taking, types and frequency of AI feedback, etc.). These data will serve as adherence verification evidence and process materials for reasoning training, used for quality control and subsequent outcome analysis. Before the intervention begins, students will receive training on how to use the AI agent. The technical team will provide full technical support throughout the study period.

3.4.2 Control Measures

The same 10 cases including case information, questions, and answer keys will be distributed as learning materials to the control group after class. Control group students are also required to complete self-study of these cases. At the end of the study, information on control group students' material learning duration, learning frequency, and use of external AI tools will be collected via questionnaire for comparison with

the intervention group. The AI agent will be made available to all students after the study concludes to ensure educational equity.

3.6 Outcomes and Data Collection

3.6.1 Primary Outcomes

(1) Clinical reasoning test score: One month after the course, a clinical reasoning test will be administered. The test includes 2 A3 case-cluster MCQs, 3 A4 case-series best-answer MCQs, and 2 case analysis questions, with a total score of 50. All test questions are also provided by the teaching faculty, covering core diseases from the five topics. The questions are reviewed by the same two senior clinical faculty members to ensure appropriate case selection and difficulty. A pilot study will be conducted to confirm the difficulty index (P value) and discrimination index (D value) of the test questions, preventing ceiling or floor effects. A3 and A4 questions are objective and will be scored by computer; case analysis questions are subjective and will be graded by the question setter.

(2) Clinical scenario simulation score: Students will be organized into groups for clinical scenario simulations. Instructors will score each student's performance, with a total score of 50. All instructors will receive standardized training to enhance scoring consistency. A pilot study will be conducted to calculate the intraclass correlation coefficient (ICC) for independent scoring by instructors, performing rater consistency testing.

The above scores are used only for data analysis in this study and are not included in students' overall course grades. This approach follows the principle of fairness in educational research and minimizes the Hawthorne effect and evaluation anxiety on study results, aiming to improve internal validity.

3.6.2 Secondary Outcomes

(1) AI technology acceptance: After the test, an electronic questionnaire will be administered to intervention group students to assess their acceptance of the AI agent technology. The questionnaire is based on the Technology Acceptance Model (TAM) and includes four dimensions: perceived usefulness (perceived enhancement of clinical reasoning training by the AI agent), perceived ease of use (interface

friendliness and operational convenience of the AI agent), satisfaction (satisfaction with the AI agent), and intention to use (willingness to use and recommend the AI agent in the future). A 5-point Likert scale is used (1 = strongly disagree, 5 = strongly agree). The questionnaire will be pilot-tested for reliability and validity. After the questionnaire survey, purposive sampling will be used to recruit participants for qualitative interviews. The final number of interviewees will be determined based on the principle of information saturation. Semi-structured interviews will be conducted to supplement understanding of intervention group students' experiences with the AI agent, including usage issues and reasoning guidance capabilities.

(2) Average learning time per case: AI usage behavior data (training duration, training frequency, etc.) for the intervention group will be retrieved from the system backend. Information on control group's learning duration, learning frequency, and use of external AI tools will be obtained through questionnaires.

(3) Course exam review time: Self-reported course exam review time for both intervention and control groups will be obtained through questionnaires.

3.7 Data Analysis

Quantitative data will be analyzed using SPSS 25.0 software, with significance level $\alpha=0.05$. Qualitative data will be analyzed using NVivo 14.0 software.

(1) Baseline data comparison: Descriptive statistics will be used to describe baseline characteristics, comparing demographic information, diagnostic scores, AI literacy, and AI use interest between intervention and control groups to do baseline comparison. Normally distributed continuous variables will be expressed as mean ± standard deviation, with independent samples t-test for between-group comparisons. Categorical variables will be expressed as frequency (percentage), with chi-square test for between-group comparisons. If $P > 0.05$, randomization is considered successful, with no significant baseline differences. If $P < 0.05$, these variables will be included as covariates in subsequent analyses.

(2) Primary outcome analysis: Independent samples t-test will be used to compare mean scores between intervention and control groups. Multiple linear regression models will be constructed with the score as the dependent variable,

intervention group as the main independent variable, and baseline AI literacy, diagnostic score, and self-reported "use of external AI tools" as covariates. Intention-to-treat (ITT) and per-protocol (PP) analyses will be conducted as sensitivity analyses. Criteria for PP analysis: (1) completion of all 10 case trainings; (2) at least 6 interaction rounds per case; (3) completion of all preset questions.

(3) Secondary outcome analysis:** Average learning time per case and course exam review time will be expressed as frequency ± standard deviation, with independent samples t-test for between-group comparisons. Likert scale scores for AI technology acceptance are expected to be non-normally distributed continuous variables; Mann-Whitney U test will be used for between-group comparisons. Thematic analysis will be applied to interview transcripts to extract core themes such as usefulness, ease of use, satisfaction, barriers, and suggestions. Two researchers will independently code and check theme consistency.

(4) Exploratory analysis: Correlations among AI literacy, AI interaction performance, AI technology acceptance, and course exam review time will be explored.

4 Quality Control

(1) Personnel training: All research personnel involved will receive standardized training to ensure they understand towards the study protocol, including recruitment, informed consent, data collection, and AI agent usage guidance.

(2) Data quality: Before the formal study, questionnaires will be pilot-tested for reliability and validity. All data entry will be double-checked to ensure accuracy. AI usage behavior data recorded in the backend will be verified for completeness and accuracy.

(3) Intervention adherence: Adherence of intervention group students will be assessed using data from the backend.

(4) Study process monitoring: Regular research team meetings will be held to discuss progress, resolve issues, and monitor study procedures to ensure compliance with the protocol.

5 Ethics and Privacy Protection

All participants will sign an informed consent form, clearly stating the study purpose, procedures, and right to withdraw. Student scores, questionnaire results, and other data collected during the study will be stored and analyzed anonymously. Study results will be used only for academic publication, without involving personal privacy or commercial interests, and will not affect students' final course assessments.

This study has been approved by the Research Ethics Committee of Peking Union Medical College Hospital (Approval No.: I-26PJ0851).

6 Reference

[1] Miller GE. The assessment of clinical skills/competence/performance. Acad Med. 1990; 65:S63-7.

[2] Case-Based Learning in Medical Education. Nature Research Intelligence [EB/OL]. [2025-12-21]. https://www.nature.com/research-intelligence/nri-topic-summaries/case-based-learning-in-medical-education-micro-5160.

[3] Hopewell S, Chan A, Collins G S, et al. CONSORT 2025 statement: updated guideline for reporting randomised trials. BMJ. 2025;389:e081123.

[4] Leyrat C, Eldridge S, Taljaard M, et al. K. Practical considerations for sample size calculation for cluster randomized trials. J Epidemiol Popul Health. 2024 Feb;72(1):202198.

[5] Kul, S., Vanhaecht, K. & Panella, M. Intraclass correlation coefficients for cluster randomized trials in care pathways and usual care: hospital treatment for heart failure. BMC Health Serv Res. 2014;14(84).

[6] Metcalfe JJ, Prescott MP, Schumacher M, et al. Community-based culinary and nutrition education intervention promotes fruit and vegetable consumption. Public Health Nutrition. 2022;25(2):437-449.

[7] Gong MC, Pan H, Liu H, et al. Expert Consensus on the artificial intelligence proficiency competency list and assessment framework for medical students. Acta Academiae Medicinae Sinicae. 2026;48(1):13-23.