Study Title

Impact of Responsibility Allocation Structures on Diagnostic Quality in AI-Assisted Diagnosis: An Individually Randomized Four-Arm Parallel-Group Controlled Experiment

Background

Artificial intelligence (AI) is increasingly being used to support clinical diagnosis, but its effectiveness depends not only on the performance of the algorithm itself, but also on how clinicians interpret, verify, and use AI-generated recommendations. In human–AI collaborative diagnostic settings, one important concern is the diffusion of responsibility: when responsibility boundaries are unclear, clinicians may rely excessively on AI and reduce independent verification and careful judgment, thereby affecting diagnostic accuracy as well as the alignment between confidence and actual correctness.

Different responsibility allocation frameworks may influence how clinicians adopt AI recommendations, engage in verification behavior, and ultimately make diagnostic decisions. However, experimental evidence on whether different responsibility structures improve or impair diagnostic quality in AI-assisted diagnosis remains limited. In particular, it is still unclear whether emphasizing full clinician responsibility, equal shared responsibility between clinicians and AI, or dynamically prompting responsibility at key decision points leads to differential effects on diagnostic accuracy and confidence calibration.

Therefore, this study aims to systematically examine the effects of different responsibility allocation structures on decision quality in AI-assisted diagnosis through a randomized controlled experiment conducted in a simulated clinical environment.

Objective

The aim of this study is to evaluate the effects of different responsibility allocation frameworks on diagnostic performance in AI-assisted diagnosis. The specific objectives are as follows:

(1) To compare differences in final diagnostic accuracy across four conditions: dynamic responsibility, clinician-full responsibility, equal shared responsibility between clinicians and AI, and a control condition.
(2) To compare differences in confidence calibration across conditions.
(3) To examine differences in clinician–AI agreement and diagnostic adjustment across conditions.
(4) To assess differences in participants' subjective evaluations of the diagnostic process across conditions.

Study Design

This study is an individually randomized, four-arm, parallel-group controlled experiment

conducted in an online simulated clinical environment. Participants were randomly assigned to one of the following four study groups:

(1) Dynamic responsibility group

(2) Full responsibility group

(3) Equal shared responsibility group

(4) Control group

All participants completed the same standardized clinical vignette tasks. The only between-group difference lay in how responsibility allocation frameworks were presented during the AI-assisted diagnostic process.

Participants

Participants were hospital doctors with professional qualifications or verified credentials recruited through the Credamo platform.

Inclusion Criteria:

(1) Licensed or verified practicing clinicians;

(2) Ability to read and complete the study tasks online;

(3) Completion of electronic informed consent prior to participation.

Exclusion Criteria:

(4) Failure to complete the study tasks in full;

(5) Failure to complete the required diagnostic tasks.

A total of 96 doctors were included in the experiment.

Randomization

This study adopted individual randomization, whereby participants were randomly assigned to one of four parallel groups.

Random assignment was implemented through the online experimental platform, ensuring that each participant entered only one responsibility allocation condition. Group assignment was completed before the tasks began and remained unchanged throughout the experiment.

Intervention Arms

(1) Dynamic Responsibility Group

During the task process, participants were presented with the AI system's confidence level along with corresponding responsibility prompts. In particular, when a participant's initial diagnosis was inconsistent with the AI recommendation, the system prompted the doctor to actively verify the information and take responsibility for the final decision in key

disagreement situations.

(2) Full Responsibility Group

Throughout the task process, participants were explicitly informed that the clinician bore sole responsibility for the final diagnostic decision, regardless of whether AI was involved.

(3) Equal Shared Responsibility Group

Throughout the task process, participants were informed that responsibility for the diagnostic decision was shared equally between the clinician and the AI system.

(4) Control Group

Participants completed the same AI-assisted diagnostic tasks but did not receive any explicit responsibility allocation framing.

Outcome Measures

(1) Primary Outcomes

Diagnostic Accuracy： Diagnostic accuracy refers to whether the diagnosis made by the participant for each case was consistent with the case-specific reference standard. This measure was coded dichotomously, with a correct diagnosis coded as 1 and an incorrect diagnosis coded as 0.

Final Diagnostic Confidence Calibration： Final diagnostic confidence calibration refers to the degree of alignment between participants' confidence in their final diagnosis and the actual correctness of that diagnosis. Confidence in the final diagnosis was rated on a 1–7 scale and linearly rescaled to the 0–1 range. Final diagnostic accuracy was coded as 1 for a correct diagnosis and 0 for an incorrect diagnosis. Calibration was defined as the absolute difference between confidence and accuracy, with values closer to 0 indicating better calibration.

(2) Secondary Outcomes

Clinician–AI Agreement： Clinician–AI agreement refers to whether participants endorsed the key elements of the AI reasoning. This measure was coded dichotomously, with endorsement coded as 1 and non-endorsement coded as 0.

Diagnostic Adjustment： Diagnostic adjustment refers to whether participants modified their original diagnostic conclusion after receiving AI advice. This measure was also coded dichotomously, with adjustment coded as 1 and no adjustment coded as 0.

Post-Task Subjective Evaluations： Post-task subjective evaluations refer to participants' assessments of subjective experiences during the task, including collaborative efficacy, cognitive load, and perceptions of attribution. Relevant items were measured using 1–7 Likert-type scales, where 1 indicated "strongly disagree" and 7 indicated "strongly agree."

Study Procedure

This study was conducted between November 2025 and January 2026 through the Credamo online platform in a simulated clinical decision-making environment.

After entering the study, participants first read the study instructions and completed electronic informed consent. They were then randomly assigned to one of four responsibility allocation conditions. Each participant completed 10 validated clinical vignette tasks and used an AI diagnostic support tool during the task process.

For each task, participants first reviewed the case information and made an initial diagnostic judgment. They then viewed the AI recommendation, after which they could make or revise their final diagnosis and report their corresponding confidence in that diagnosis. In the dynamic responsibility group, when the participant's initial diagnosis was inconsistent with the AI recommendation, the system additionally displayed a responsibility prompt. In the other responsibility-framing groups, the system presented the corresponding responsibility statement according to the assigned group. In the control group, no responsibility allocation prompt was provided.

After completing all tasks, participants filled out a post-task subjective evaluation questionnaire.

Statistical Analysis

This study compared the primary and secondary outcomes across the four experimental groups.

First, descriptive statistics were used to summarize participants' basic characteristics and each outcome measure. Subsequently, appropriate statistical models were used to compare differences among the four groups, with a particular focus on the effects of different responsibility allocation frameworks on final diagnostic accuracy and confidence calibration. Where appropriate, pairwise comparisons were conducted to test differences between specific experimental groups. Statistical significance was evaluated using conventional thresholds. All analyses were conducted after data collection had been completed.

Ethics and Informed Consent Statement

The participants in this study were practicing doctors. The study was conducted in the form of an online simulated diagnostic experiment using virtual clinical cases and did not involve real patient data.

Before participating in the study, participants read the study information and took part voluntarily. All participants completed electronic informed consent before entering the formal task procedure.

Because this study used simulated cases and did not involve real patient medical record data,

the overall level of risk was low.

Preregistration Statement

This study was preregistered on the AsPredicted platform before formal analysis. The preregistration record is as follows: **AsPredicted #6433; https://aspredicted.org/6433**