Study Title: Trust or Verify? Automation Bias in Physician-LLM Diagnostic Reasoning (A Randomized Clinical Trial)

Authors: Ihsan Ayyub Qazi, PhD; Ayesha Ali, PhD; Asad Ullah Khawaja, MBBS; Muhammad Junaid Akhtar, MBBS; Ali Zafar Sheikh, MBBS; Muhammad Hamad Alizai, PhD

NCT Number: NCT06963957

Date: May 28, 2025

Study Protocol and Statistical Analysis Plan

Brief Summary: This study aims to systematically measure the extent and patterns of automation bias among physicians when utilizing ChatGPT-4o in clinical decision-making.

Condition or disease: Diagnosis

Intervention/treatment: Other: ChatGPT-4o

Phase: Not Applicable

Detailed Description:

Diagnostic errors represent a significant cause of preventable patient harm in healthcare systems worldwide. Recent advances in Large Language Models (LLMs) have shown promise in enhancing medical decision-making processes. However, there remains a critical gap in our understanding of how automation bias -- the tendency to over-rely on technological suggestions -- influences medical doctors' diagnostic reasoning when incorporating these AI tools into clinical practice.

Automation bias presents substantial risks in clinical environments, particularly as AI tools become more integrated into healthcare workflows. Although LLMs such as ChatGPT-4o offer potential advantages in reducing errors and improving efficiency, their lack of rigorous medical validation raises concerns about potentially amplifying cognitive biases through the generation of incorrect or misleading information.

Multiple contextual factors can exacerbate automation bias in medical settings: time constraints in high-volume clinical settings, financial incentives that prioritize efficiency over thoroughness, cognitive fatigue during extended shifts, and diminished vigilance when confronting diagnostically challenging cases. These factors may interact with psychological mechanisms that include the diffusion of responsibility, overconfidence in technological solutions, and cognitive offloading---collectively increasing the risk of uncritical acceptance of AI-generated recommendations.

This randomized controlled trial (RCT) aims to systematically measure the extent and patterns of automation bias among physicians when utilizing ChatGPT-4o in clinical decision-making. The investigators will assess how access to LLM-generated information influences diagnostic reasoning through a novel methodology that precisely quantifies automation bias. In this study, participants will be randomly assigned to one of two groups. The treatment group will receive LLM-generated recommendations containing deliberately introduced errors in a subset of cases, while the control group will receive LLM-generated recommendations without such deliberately introduced errors. Participants will evaluate six clinical vignettes randomly sequenced to prevent detection patterns. The flawed vignettes provided to the treatment group will incorporate subtle yet clinically significant errors that should be identifiable by trained doctors. This will enable investigators to quantify the degree of automation bias by measuring the differential in diagnostic accuracy scores between the treatment and control groups.

Prior to participation, all physicians will complete a comprehensive training program covering LLM capabilities, prompt engineering techniques, and output evaluation strategies. Responses will be evaluated by blinded reviewers using a validated assessment rubric specifically designed to detect uncritical acceptance of erroneous information, with greater score disparities indicating stronger automation bias. This naturalistic approach will yield insights directly applicable to real clinical workflows, where mounting cognitive demands may progressively impact diagnostic decision quality.

Study Design

Study Type: Interventional (Clinical Trial)

Actual Enrollment: 50 participants

Allocation: Randomized. Randomization list was created by Ihsan Ayyub Qazi using the Sealed Envelope program.

Intervention Model: Parallel Assignment

Intervention Model Description: The trial will be designed as a randomized, two-arm, single-blind parallel group study.

Masking: Single (Outcomes Assessor)

Masking Description: The grading of responses will be performed by assessors blinded to participant identity and treatment assignment.

Primary Purpose: Diagnostic

Official Title: Trust or Verify? Automation Bias in Physician-LLM Diagnostic Reasoning

Actual Study Start Date: June 5, 2025

Actual Primary Completion Date: July 1, 2025

Actual Study Completion Date: July 15, 2025

Arms and Interventions

Arm	Intervention/Treatment
Active Comparator: ChatGPT-4o Recommendations with Hallucinations *Participants will evaluate six clinical vignettes. During the trial, they will have access to clinical recommendations from a specific, commercially available LLM (ChatGPT-4o) in addition to conventional diagnostic resources. LLM recommendations for three vignettes will contain deliberately flawed diagnostic information and for three vignettes it will contain accurate recommendations). The cases will be presented in random order.*	**Other: ChatGPT-4o Recommendations with Hallucinations** *ChatGPT-4o's differential diagnoses of six clinical vignettes, three of which will contain deliberately introduced inaccurate information.*
No Intervention: ChatGPT-4o Recommendations without Hallucinations *Participants will evaluate the same six clinical vignettes as in the intervention arm. During the trial, they will have access to clinical recommendations from a specific, commercially available LLM (ChatGPT-4o) in addition to conventional diagnostic resources. However, the LLM-generated recommendations will not contain any deliberately introduced errors. The cases will be presented in random order.*	

Outcome Measures

Primary Outcome Measures:

Diagnostic reasoning [Time Frame: Assessed at a single time point for each case, during the scheduled diagnostic reasoning evaluation session, which takes place between 0-4 days after participant enrollment.]

The primary outcome will be the percent correct for each case, ranging from 0 to 100%, where higher scores indicate better diagnostic performance. For each case, participants will be asked for their three leading diagnoses, findings that support each diagnosis, and findings that oppose each diagnosis. For each plausible diagnosis, participants will receive 1 point. Findings supporting the diagnosis and findings opposing the diagnosis will also be graded based on correctness, with 1 point for each correct response. Participants will then be asked to name their top diagnosis they believe is most likely, earning 9 points for a reasonable response and 18 points for the most accurate response. Finally participants will be asked to name up to 3 next steps to further evaluate the patient with 0.5 point awarded for a partially correct response and 1 point for a completely correct response. The primary outcome will be compared at the case-level between the randomized groups.

Secondary Outcome Measures:

Top choice diagnosis accuracy score [Time Frame: Assessed at a single time point for each case, during the scheduled diagnostic reasoning evaluation session, which takes place between 0-4 days after participant enrollment.]

The secondary outcome will measure participants' performance in identifying the most likely diagnosis for each clinical vignette. After evaluating each case, participants will select their single most likely diagnosis, which will be scored on a pre-specified Three-Tier Diagnostic Accuracy Scale: 18 points for the most accurate diagnosis, 9 points for a clinically reasonable alternative, and 0 points for an incorrect diagnosis. For each participant, a Top Choice Diagnosis Accuracy Score is calculated as (total points earned ÷ maximum possible points) × 100, yielding a 0–100 % range in which higher scores indicate greater diagnostic accuracy. This percentage score will be compared at the case-level between randomized groups to quantify the impact of automation bias on diagnostic decision-making.

Eligibility Criteria

Ages Eligible for Study: All

Sexes Eligible for Study: All

Accepts Healthy Volunteers: Yes

Criteria

Inclusion Criteria:

- Completed Bachelor of Medicine, Bachelor of Surgery (MBBS) Exam. The equivalent degree of MBBS in US and Canada is called Doctor of Medicine (MD).
- Full or Provisionally Registered Medical Practitioners with the Pakistan Medical and Dental Council (PMDC).
- Participants must have completed a structured training program on the use of ChatGPT (or a comparable large language model), totaling at least 10 hours of instruction. The program must

include hands-on practice related to LLM's aspects, specifically prompt engineering and content evaluation.

Exclusion Criteria:

- Any other Registered Medical Practitioners (Full or Provisional) with PMDC (e.g., Professionals with Bachelor of Dental Surgery or BDS).

Potential Risks and Harms

This study poses minimal risk to participants. Some participants may experience mild discomfort or frustration, particularly those in the treatment group who may find clinical vignettes more challenging due to flawed LLM recommendations in a select set of vignettes.

Safeguards for Addressing Risks or Harms to Subjects

To minimize potential discomfort, participants will receive comprehensive instructions and ongoing support throughout the study. Participants will be explicitly informed that individual performance is not being evaluated, and that data will be analyzed for aggregate trends only. All participants will be advised of their right to withdraw from the study at any time without penalty or explanation. Research staff will monitor for signs of distress and provide appropriate support as needed.

Informed Consent

Informed consent will be obtained from all participants before enrollment in the study.

Power Analysis:

The minimum target sample size of 50 participants (25 participants per arm) was predetermined based on a prior study.[1] The power analysis, conducted using Python version 3.11.9 (Python Software Foundation), employed the statsmodels.stats.power module from statsmodels version 0.14.4 (Statsmodels Developers) and indicated that a total sample of 200 to 250 completed cases (approximately 4 to 5 cases per participant) would provide at least 80% power to detect an 8-percentage-point mean difference in diagnostic reasoning scores, assuming a two-sided α of .05. The analysis employed mixed-effects models suitable for cluster-randomized designs, considering an intraclass correlation coefficient (ICC) ranging from 0.05 to 0.15 and a standard deviation of 16.2%.

Statistical Analysis:

Descriptive analysis will be performed by comparing participant characteristics and automation bias behaviors between the treatment and control groups. Any categorical variables will be reported using counts and proportions, while continuous variables will be reported as means and standard deviations or, if their distribution is non-normal, as medians with interquartile ranges.

To understand automation bias behavior the following descriptive statistics will be reported for both the study groups:

- *Proportion of clinical vignettes in which participants accessed LLM suggestions*: For each of the six vignettes, a binary indicator will be recorded (1 = participant clicked to view the LLM's suggestions; 0 = no access). The final measure is the mean proportion of vignettes (range 0–1) per participant for which suggestions were accessed.

- *Proportion of clinical vignettes in which participants copy-pasted LLM suggestions*: For each of the six vignettes, a binary indicator is recorded:
    - 1 = participant directly copy-pasted the LLM's suggestion into their diagnostic answer (i.e., accepted verbatim without any edits)
    - 0 = no direct copy-paste occurred

  The final measure is calculated as the mean proportion of vignettes (range 0–1) per participant in which suggestions were copy-pasted.
- *Mean fraction of questions per vignette answered via direct copy-paste of LLM suggestions*: For each vignette, we will calculate the fraction of questions that were copy-pasted by dividing the number of copy-pasted answers by the total of five questions per vignette. The final measure is the average of these vignette-level fractions across all six vignettes, yielding a single measure per participant that ranges from zero (no copy-pasting) to one (all questions copy-pasted).

The primary outcome analysis will be conducted at the case level, with clustering by participant under an intention-to-treat framework. The primary analysis will be based on cases with completed responses only. We will first summarize the mean and standard deviation of scores (standardized on a 0–100 scale) and report the mean and standard deviation for the time spent on each case, both for the overall cohort and separately for each of the study arms. To evaluate the impact of the treatment (i.e., providing access to recommendations some which contain inaccurate information), we will apply generalized mixed-effect models, including a random effect for the participant to account for potential within-participant correlation between cases, and a random effect for cases to account for differences in difficulty across cases. To improve precision, we will include the following covariates as control variables: past experience in LLM use, gender, and years of practice post MBBS. We will also conduct the following sensitivity analyses to assess the robustness of the findings:

1) Effect sizes without controls
2) The primary outcome analysis will be repeated including incomplete cases (if any) to evaluate the potential impact of missing data (e.g., using multiple imputation).
3) Inter-rater reliability for the diagnostic reasoning scores across at least two raters will be assessed using Cohen's kappa or Fleiss' kappa depending on the final number of raters.
4) The internal consistency of the diagnostic reasoning score across the six vignettes will be assessed using Cronbach's alpha.
5) Assessment of normality assumptions

For the secondary outcome (top choice diagnosis accuracy score), a mixed-effects ordinal logistic regression model will be used, with the same random effects structure as the primary analysis, and the dependent variable being the three-level final diagnosis accuracy (Incorrect, Partially Correct, Correct).

A secondary analysis will be conducted to measure the sequential trend in automation bias. A linear mixed-effects model will be used, with random intercepts for participants and for vignettes, and fixed effects for vignette order, treatment arm, and their interaction. The dependent variable will be the vignette-level diagnostic reasoning performance (percentage); the sequence-by-treatment interaction will then quantify how the treatment arm's trajectory deviates from control, reflecting any progressive automation bias across successive vignettes. We will repeat this analysis for the secondary outcome as well.

Subgroup analyses will be performed based on experience with LLMs, gender and years of practice post MBBS. Scores for each case question will also be compared descriptively between the two randomized groups.

Exploratory Analyses

- Within-subject difference in diagnostic reasoning score between the 3 clinical vignettes without flawed LLM recommendations and 3 vignettes with flawed LLM recommendations and the corresponding difference in vignettes in the control group.
- Standalone LLM Performance: The standalone performance of the LLM will be compared to the control group using an independent samples t-test

Missing data patterns and extent will be described for all variables. The primary analysis will follow an intention-to-treat framework using complete case analysis. If substantial missing data occurs (>5% for any key variable), multiple imputation will be implemented and results compared with the complete case approach to assess robustness of findings. Given the small sample size, no interim analyses are planned.

All statistical analyses will be performed using Python software, version 3.11.12 (Python Software Foundation) with pandas for data manipulation and statsmodels version 0.14.4 for mixed-effects modeling, R (version 4.3.2 or later) or similar packages. Statistical significance will be based on a p value.

Contacts and Locations

Locations

Pakistan, Lahore

Lahore University of Management Sciences

Lahore, Punjab 54792

Sponsors and Collaborators

Lahore University of Management Sciences

Investigators

Principal Investigator: Ihsan Ayyub Qazi, PhD, Lahore University of Management Sciences

Principal Investigator: Ayesha Ali, PhD, Lahore University of Management Sciences

Principal Investigator: Muhammad Asadullah Khawaja, MBBS, King Edward Medical University

Principal Investigator: Ali Zafar Sheikh, MBBS, Lahore General Hospital

Principal Investigator: Muhammad Junaid Akhtar, MBBS, Children's Hospital, Lahore