MEDINFO 2025 — Healthcare Smart × Medicine Deep
M.S. Househ et al. (Eds.)

doi:10.3233/SHTI250844

Impact of an Evidence-Based Large Language Model (LLM) Diagnostic Decision Support System: A Randomised Controlled Trial

Sangah Ahn[a,#], Joongheum Park M.D. [b,c,#], Sujeong Hur Ph.D.[a,c], Kwang Yul Jung M.D., Ph.D.[d], Jae-ho Lee M.D., Ph.D.[e], Sae Won Choi M.D.[f], Meong Hi Son M.D.[g], Min-Jeoung Kang, Ph.D.[h], Youn-Jung Kim M.D., Ph.D. [e], Hanna Park M.D. [e], Won Chul Cha M.D., Ph.D.[a,g], and Junsang Yoo Ph.D. [a]

[a] *SAIHST, Sungkyunkwan University, Seoul, Republic of Korea;* [b] *Beth Israel Deaconess Medical Center, Boston, MA, USA;* [c] *AVOMD, Brooklyn, NY, USA;* [d] *Chung-ang University Gwangmyeong Hospital, Gwangmyeong, Gyung-gi, Republic of Korea;* [e] *Asan Medical Center, Seoul, Republic of Korea;* [f] *Seoul National University Hospital, Seoul, Republic of Korea;* [g] *Samsung Medical Center, Seoul, Republic of Korea;* [h] *Bringham and Women's Hospital, Boston, MA, USA;* [#] *Equally contributed.*

ORCiD ID: Sangah Ahn https://orcid.org/0009-0003-5668-8326

Abstract. Generative artificial intelligence (AI) influences clinical decision-making in healthcare by analyzing medical data and proposing personalized treatment options based on patient records. However, generative AI limited due to its inability to provide accurate evidence. Therefore, this study aims that the influence of AI-generated diagnostic suggestions on emergency healthcare providers' diagnostic patterns and decision-making, and evaluates the correlation between clinicians' adoption and diagnosis accuracy.

Keywords. Large Language Model (LLM), Clinical Decision Support System (CDSS), Emergency Medicine

1. Introduction

Generative artificial intelligence (AI) refers to AI that generates text through pre-training on large datasets using machine learning, with large language models (LLMs) being a notable subset of this technology [1]. LLMs learn textual data across diverse domains to create coherent, contextually relevant sentences, mimicking human-like expressions. Although AI adoption in healthcare is still in its early stages, it is rapidly advancing, with applications in medical imaging, recommendation about diagnoses, personalized treatment plans, and big data analysis to support medical research. Recent reviews suggest that LLMs have the potential to improve clinical decision-making by supporting accurate diagnoses, effective treatment planning, and anticipating patient prognoses [2].

While generative AI is being explored in various fields such as drug development, diagnostics, treatment planning, and patient-reported outcomes, its potential in clinical

decision-making—especially in high-stakes environments like emergency medicine—is particularly promising. This study focuses on its application in emergency medicine, where rapid and accurate decision-making is critical.

However, a critical barrier to realizing the full potential of LLM-based clinical decision support system (CDSS) is gaining the trust of medical professionals. Hallucination—where LLMs produce incorrect information not supported by facts or literature—is a significant concern in healthcare, where errors can have irreversible consequences [3]. This highlights the need for caution when applying AI-generated information, as it lacks explicit evidence and is based on patterns in training data rather than objective truth.

Emergency medicine requires accurate, rapid decision-making to save lives, with clinicians in emergency departments (EDs) often making decisions under time pressure and limited information [4]. Diagnostic errors in EDs account for approximately 10–15% of all errors, higher than in other specialties [5].

This study studies the impact of AI-generated diagnostic suggestions on the diagnostic patterns of emergency healthcare providers and the appropriateness of their decision-making. Furthermore, we assess the correlation between clinicians' adoption of these recommendations and the appropriateness of diagnoses.

2. Methods

2.1. AI-based diagnostic CDSS

Retrieval-augmented generation (RAG) is effective in addressing hallucination problems, but it has limitations in reasoning power. We developed a synergetic clinical LLM, a collaborative AI system, to complement RAG by enhancing its reasoning power. This system consists of 11 specialized AI units, each with distinct expertise, which navigates various facets of healthcare information management. Further technical details on the system cannot be revealed according to the developer company's policy. The system employs unique interplay mechanics through three steps: collaboration, competition, and critique.

When generating responses, the LLM primarily bases its answer on a single clinical guideline. However, it can also incorporate additional guidelines from the system’s extensive list of available sources, ensuring a more comprehensive and nuanced response.

2.2. Study Design

We conducted prospective, multicentre, diagnostic, randomized controlled trial in tertiary academic hospital in Republic of Korea. This study was approved by the Institutional Review Board (IRB) of Samsung Medical Center (IRB No. SMC 2024-03-052).

2.3. Participants

We recruited residents and specialists from emergency medicine, internal medicine, and family medicine who worked in ED within the last three years. Physicians without recent ED experience were excluded.

Participants were randomly assigned in a 1:1:1 ratio to one of three study groups: diagnostic decision-making with AI support and rationale provided; diagnostic decision-making with AI support but no rationale; and solely diagnostic decision-making without AI support.

2.4. Procedure

This study created three clinical vignettes illustrating acute exacerbation situations for chronic kidney disease, chronic obstructive lung disease, and Crohn's disease. The emergency medicine specialist selected three diagnoses from the array of clinical criteria and diagnoses held by the agent used in the study. The vignettes were first composed by a nurse specialized in digital health, adhering to clinical criteria, and subsequently evaluated and modified by a faculty member in emergency care from a tertiary university hospital. The completed vignettes were input into an AI-driven clinical decision support system, which produced diagnostic and testing suggestions. Four faculty members independently analyzed the system's replies for conformity with clinical criteria and assessed the clinical validity of any inconsistencies.

Following randomization and written consent, participants submitted open-ended replies for each vignette, identifying three primary diagnoses and three differential diagnoses. Emergency medical faculty evaluated the suitability of each response using a five-point Likert scale. Diagnostic appropriateness was reclassified as positive (4–5), neutral (3), or negative (1–2), and inter-rater reliability was assessed using average percent agreement, with a threshold of 0.67. Delphi rounds were executed in instances when agreement was not attained. Secondary outcomes, such as trust, acceptance, explanation satisfaction, and usability, were assessed using four translated and back-translated questionnaires: the Trust Scale for XAI [6], the Acceptance Scale for AI deployment [7], the Explanation Satisfaction Scale [6], and the System Usability Scale [8]. The overall schematic of the study is shown in figure 1.

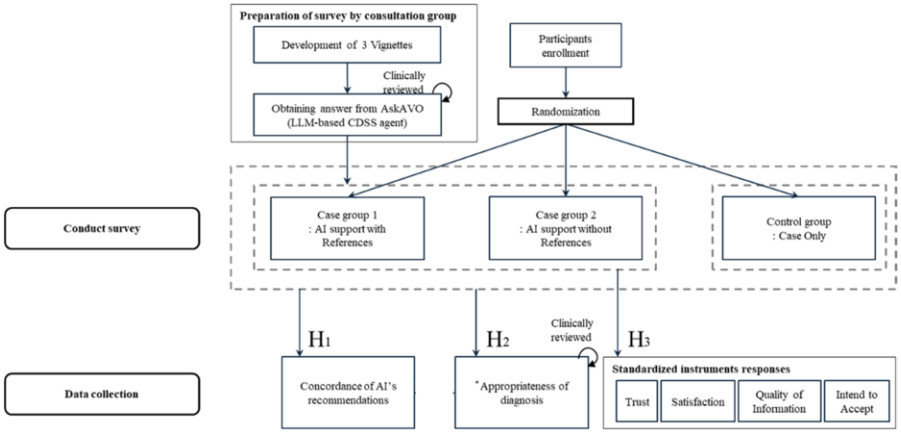

Figure 1. Overview of Three-Arm Randomized Controlled Trial Simulation Setting

3. Results

The study with 118 participants revealed no significant disparities in age, gender distribution, professional position, or departmental specialty. Among the 118 participants, 61.0% were residents, 52.5% specialized in internal medicine, and 38.1% in emergency medicine. Merely 1.7% possessed prior experience in medical AI creation, whereas 54.2% had utilized AI in clinical practice.

An assessment was performed to evaluate the conformity of AI-generated diagnostic recommendations with recognized clinical criteria. Four emergency department faculty members from three tertiary academic hospitals in Korea independently evaluated the AI-generated diagnoses and their corresponding evidence. 44.4% had direct support from primary recommendations, 9.7% conformed to secondary standards, and 45.8% were deemed clinically valid.

We analyzed the concordance of physicians' diagnoses with the recommendations of the AI-based CDSS across case and control groups. The AI-assisted physician group demonstrated much greater agreement with AI-generated diagnoses compared to those who relied exclusively on their own diagnostic capabilities. The same trend was similarly noted for differential diagnosis. Physicians in the experimental group exhibited a higher concordance with AI system recommendations compared to those in the control group (Figure 2). Three independent reviewers reviewed the appropriateness of each diagnosis. Following two rounds of consensus, 631 diagnoses (94.89%) achieved the agreement criterion, whilst 34 diagnoses (5.11%) were eliminated from further study. Physicians utilizing the diagnostic AI agent attained a much superior diagnostic appropriateness score (4.07 ± 0.0227) compared to those diagnosing autonomously (3.99 ± 0.0338, $p = 0.046$) (Figure 3). The impact size was 0.098, suggesting a minor influence of its aid on diagnostic appropriateness. The survey indicated no significant disparities in physicians' trust, satisfaction, perceived information quality, or willingness to accept AI recommendations upon reference visibility.

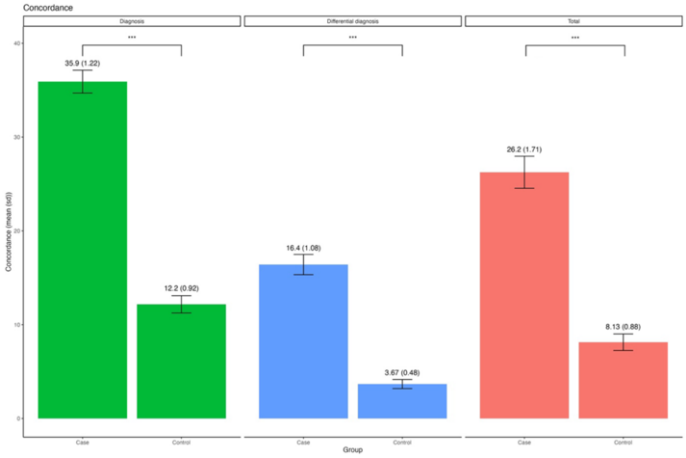

Figure 2. Concordance of physicians' diagnoses with recommendations from AI-based CDSS between the case and control groups

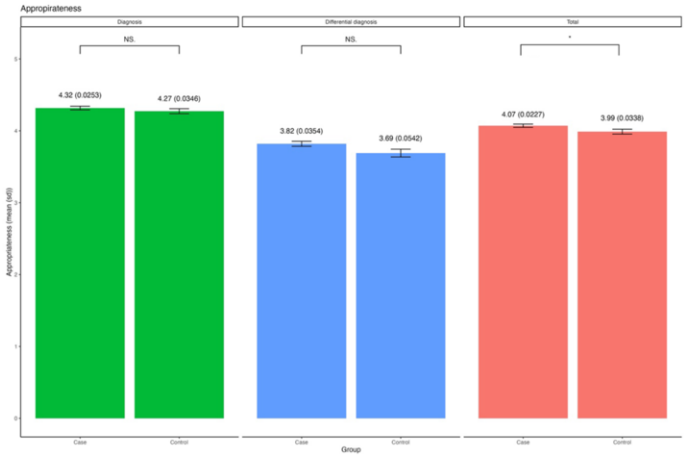

Figure 3. Appropriateness of physicians' diagnoses between the case and control groups

4. Discussion and Conclusions

The study revealed that clinicians utilizing decision help from the AI agent, Ask Avo, had improved diagnosis patterns compared to those relying exclusively on their own judgments. The diagnostic concordance with AI suggestions and appropriateness were much superior in the AI-assisted group compared to the control group. Nonetheless, the results may not be immediately applicable to real-world emergency room practices, as the study concentrated on acute exacerbation situations of particular chronic illnesses. Future study should investigate the presentation of information by AI-based diagnostic Clinical Decision Support Systems (CDSS), doctors' cognitive responses to this information, and the correlation between their acceptance behavior and clinical appropriateness. Formulating a thorough mental model for AI-assisted decision-making is crucial for enhancing its incorporation into healthcare procedures.

References

[1] Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training.
[2] Korea Health Industry Development Institute. (2023). Global health industry trends Vol. 482 (2023.08.07).
[3] Kumar, M., Mani, U. A., Tripathi, P., Saalim, M., Roy, S., & Roy Sr, S. (2023). Artificial hallucinations by Google bard: think before you leap. Cureus, 15(8).
[4] Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F., & Ting, D. S. W. (2023). Large language models in medicine. Nature medicine, 29(8), 1930-1940.
[5] Stone, E. L. (2019). Clinical decision support systems in the emergency department: opportunities to improve triage accuracy. Journal of emergency nursing, 45(2), 220-222.
[6] Hoffman, R. R., Mueller, S. T., Klein, G., & Litman, J. (2018). Metrics for explainable AI: Challenges and prospects. arXiv preprint arXiv:1812.04608.
[7] Dong, Z., Wang, J., Li, Y., Deng, Y., Zhou, W., Zeng, X. & Yu, H. (2023). Explainable artificial intelligence incorporated with domain knowledge diagnosing early gastric neoplasms under white light endoscopy. NPJ Digital Medicine, 6(1), 64.
[8] Holzinger, A., Carrington, A., & Müller, H. (2020). Measuring the quality of explanations: the system causability scale (SCS) comparing human and machine explanations. KI-Künstliche Intelligenz, 34(2), 193-198.