

Abstract

Can AI Improve Triage Quality? A Preliminary Assessment of ChatGPT Performance in Evaluating Triage Decisions

Christian Ramacciani Iseman¹, Simona Burresi², Sabina Innocenti³, Lorenzo Righi⁴

¹ Department of Nursing and Midwifery, “Riuniti della Valdichiana” Hospital, Azienda USL Toscana Sud Est, Arezzo, Italy

² Department of Nursing and Midwifery, “Misericordia” Hospital, Azienda USL Toscana Sud Est, Arezzo, Italy

³ Department of Nursing and Midwifery, “La Gruccia” Hospital, Azienda USL Toscana Sud Est, Arezzo, Italy

⁴ Department of Nursing and Midwifery, Azienda USL Toscana Sud Est, Arezzo, Italy

(Corresponding Author: ramacciani@gmail.com)

Citation: Ramacciani Iseman C, Burresi S, Innocenti S, Righi L. “Can AI Improve Triage Quality? A Preliminary Assessment of ChatGPT Performance in Evaluating Triage Decisions” (2026) *International Journal of Health Supplements* 1(Suppl. 1):43. Supp. DOI: <https://doi.org/10.82051/ijhs-2026-1>

Introduction. Accuracy and consistency in emergency department (ED) triage are essential for patient safety, yet variability in nurse-assigned triage levels persists despite standardized protocols. This study explored the potential of ChatGPT as a retrospective evaluator of triage appropriateness according to the Tuscan Triage System.

Methods. Fifty real-world clinical scenarios were randomly extracted from an urban ED in Tuscany, Italy. Two certified triage experts independently reviewed each case, with a third resolving disagreements. ChatGPT (GPT-4o, OpenAI) was then asked to assess whether the original triage level was appropriate. Model outputs were compared with the expert-defined reference standard. Agreement, Cohen’s kappa, macro-averaged precision, recall, F1-score, and class-specific sensitivity and specificity were calculated.

Results. Exact agreement between ChatGPT and experts was observed in 46% of cases. Under-triage discrepancies were more frequent than over-triage (38% vs 16%). Cohen’s kappa was 0.243. Performance was better in high-complexity cases ($\kappa=0.313$; $F1=0.704$) and lower in moderate- and low-complexity categories. Most errors involved adjacent triage levels.

Discussion. ChatGPT showed limited-to-moderate agreement with expert triage assessment, with better performance in more critical scenarios. Although unsuitable for autonomous triage, it may have potential as a retrospective quality assurance tool. Further refinement and clinical validation are needed before use in audit or decision-support processes.

Keywords: Emergency Department, Artificial Intelligence, Clinical Decision Support, Triage, Emergency Nursing

© The Author(s) 2026. This abstract is published by iEditore and distributed under the terms of the Creative Commons Attribution 4.0 International License (CC BY 4.0).

Funding: This research received no external funding. **Conflict of Interest:** The authors declare no conflict of interest.