Disaster Medicine and Public Health Preparedness

www.cambridge.org/dmp

Repeatability, Reproducibility, and Diagnostic Accuracy of a Commercial Large Language Model (ChatGPT) to Perform Disaster Triage Using the Simple Triage and Rapid Treatment (START) Protocol

Jeffrey Michael Franc MD, MS(Stats), MSc(DM), FCFP(EM), D Sport Med[1,2], Atilla Hertelendy PhD[3], Lenard Cheng MD[3], Ryan Hata MD[3] and Manuela Verde MD, MSc[2]

[1]University of Alberta, Edmonton, AB, Canada; [2]Universita' del Piemonte Orientale, Novara, NO, Italy and [3]Beth Israel Deaconess Medical Center, Boston, MA, USA

Abstract

Cite this article: Franc JM, Hertelendy A, Cheng L, Hata R and Verde M (2024). Repeatability, Reproducibility, and Diagnostic Accuracy of a Commercial Large Language Model (ChatGPT) to Perform Disaster Triage Using the Simple Triage and Rapid Treatment (START) Protocol. *Disaster Medicine and Public Health Preparedness*, **18**, e183, 1
https://doi.org/10.1017/dmp.2024.194

Abstract

Objective: The release of ChatGPT in November 2022 drastically lowered the barrier to artificial intelligence with an intuitive web-based interface to a large language model. This study addressed the research problem: "Can ChatGPT adequately triage simulated disaster patients using the Simple Triage and Rapid Treatment (START) tool?"
Methods: Five trained disaster medicine physicians developed nine prompts. A Python script queried ChatGPT Version 4 with each prompt combined with 391 validated patient vignettes. Ten repetitions of each combination were performed: 35190 simulated triages.
Results: A valid START score was returned In 35102 queries (99.7%). There was considerable variability in the results. Repeatability (use of the same prompt repeatedly) was responsible for 14.0% of overall variation. Reproducibility (use of different prompts) was responsible for 4.1% of overall variation. Accuracy of ChatGPT for START was 61.4% with a 5.0% under-triage rate and a 33.6% over-triage rate. Accuracy varied by prompt between 45.8% and 68.6%.
Conclusions: This study suggests that the current ChatGPT large language model is not sufficient for triage of simulated patients using START due to poor repeatability and accuracy. Medical practitioners should be aware that while ChatGPT can be a valuable tool, it may lack consistency and may provide false information.

Supplementary material. The supplementary material for this article can be found at http://doi.org/10.1017/dmp.2024.194.

CAMBRIDGE
UNIVERSITY PRESS