Emergency departments (EDs) are experiencing historic crowding. A lack of patient guidance for where to seek care for acute illness further exacerbates ED strain. Consequently, patients often arrive at the ED at a rate faster than the nurse can triage. In addition, the rapid and uncertain nature of ED arrivals makes human triage subject to biases that could result in both undertriage and overtriage. General purpose language models trained on clinical notes have demonstrated early potential as all-purpose prediction engines in health care, which raises the possibility that they could make triage in the ED more efficient, accurate, and equitable.1 However, before deploying such systems in practice, evidence in the clinical setting about how well such tools might perform is needed.

Williams et al2 present their analysis of the performance of a large language model (LLM) for the task of estimating which clinical history for a patient presenting to the ED would be assigned a higher triage acuity level defined by the Emergency Severity Index (ESI). In their analysis, the authors present the LLM with the chief concern, history of presenting illness, and review of systems from 10,000 deidentified pairs of clinical records drawn from the electronic health record (EHR) of a single health system. The LLM was then asked to identify which of the 2 histories was associated with a higher acuity, defined by the initial nurse-triaged ESI. The performance of the general purpose model, without any fine-tuning to this clinical task, is impressive. Despite this high performance, much more work is needed to determine the potential to change clinical practice because of how the study was designed, variables used, and outcomes chosen. These limitations highlight challenges and opportunities common in the evaluation of clinical artificial intelligence (AI) systems.

First, the medical record elements presented to the model were created by an experienced emergency triage nurse and clinician. Consider the chief concerns a triage nurse might assign 3 patients, each of whom presented with a headache: migraine, headache, and stroke alert. The nurse’s choice conveys their implicit assessment of severity for the model to use, yet that information would not be available to a model attempting to assign a triage score before the nurse’s assessment. Similarly, the clinical history used in this study was created by a clinician and therefore relies on an ED clinician’s full evaluation, the results of laboratory tests and imaging studies, and judgment of the patient’s status during the intervening time from the moment of triage to when the note was written, hours to days after the ESI score was assigned. As the authors cautiously note, unlike the LLM in this assessment, the triage nurse does not have access to future information available to the ED clinician writing the note. Thus, the resident and attending assessments to which model performance is compared may be more accurate than the triage score criterion standard in this study.

Second, while ESI is a clinically relevant variable, the predicted outcome for the studied task is not clinically useful. The chosen outcome only compares 2 patients (a related but distinct task from sorting dozens) and excludes ties. The LLM, although it performed well at the task designed by the authors, is not emulating a task that any ED triage nurse or clinician is being asked, or that would be useful, to perform. Notably, the model’s performance drops off markedly in differentiating between adjacent acuity levels, suggesting it is not yet ready for the more important and common task of deciding whom among the dozen patients assessed as ESI level 2 (emergent) in the waiting room will be the next person to get a bed. A more informative task would be predicting eventual severity of illness, ED disposition, or resource use from the ED clinician’s early documentation.
Third, the performance metric presented is appropriate to assess the accuracy of the model but is not appropriate for clinical use. Accuracy scores combine elements of sensitivity and specificity into a single number, which weights those outcomes arbitrarily rather than by the relative harms of overtriage and undertriage. Each clinical problem to which an LLM is applied should determine those harms quantitatively and assess model performance for that particular task accordingly. An assessment of the model’s calibration, or how close a predicted ESI score was to the observed ESI score, would have provided more insight into the LLM’s readiness for triage tasks.

The authors appropriately define a narrow clinical setting in which LLM-based triage might be useful: flagging high-risk patients during the ever-more-common delay between when a patient arrives to the ED and when the triage nurse has the capacity to assess them. Such a model could also be used to provide a suggested triage level to the triage nurse, with caution to avoid biasing the triage nurse when the model’s prediction is inaccurate.

The lessons from this study also plot a path for using research to improve triage more broadly. Models can be compared using transcribed patient-provided histories on downstream outcomes such as adverse events and admission rates. These insights would allow the development of models that could identify patients who could be safely routed to other destinations before presenting to the ED, and models that could identify patients who might not otherwise present to the ED but would benefit from timely hospital care.

In the US, we provide little guidance as to the site of care, but in France those calling the equivalent of 911 are triaged by telephone, a process that can include speaking immediately with a physician on call, being transported that same day to their primary care physician’s clinic, or having an ambulance sent. In the US, nurse-led telephone triage does not cause adverse patient outcomes but is poorly adopted due to cost. The LLM-based at-home triage could reduce this cost by capturing symptoms and history while people are still at home and guide patients in choosing the ED or alternatives such as urgent care, primary care, or telemedicine.

During COVID-19, several institutions developed scripted chatbots that helped patients determine whether to present for care if their COVID-19 symptoms progressed. This saved lives, likely by encouraging patients to present to the ED before their illness was too far along. This approach was successful not by replacing humans but by complementing humans in situations with a high volume of patients that caused capacity issues. The chatbot was carefully supervised by clinicians, updated frequently, and focused on a single triage task. Such studies highlight lessons for building future LLM-based triage systems that are effective, safe, and integrated into clinical workflows.

New models of human-machine collaboration will have to be developed in support of such a tool. For instance, the model will need to be monitored and direct human interaction initiated when the uncertainty around the model’s assessment of the severity of the situation is high. Since LLMs do not natively form any assessment of severity, this will require significant advances.

Fortunately, demand for these models, if developed, would be substantial. Insurers are acutely attuned to the costs of ED visits. Patients do not want to wait hours in the ED to see a clinician if they could receive a diagnosis or referral for further testing from home. The work by Williams et al. points the way toward that future, even as it demonstrates that the future is not yet here.
The Leonard Davis Institute, University of Pennsylvania, Philadelphia (Friedman, Delgado, Weissman); Department of Biostatistics, Epidemiology, and Informatics, University of Pennsylvania, Philadelphia (Delgado, Weissman); Center for Health Care Transformation and Innovation, University of Pennsylvania, Philadelphia (Delgado); Palliative and Advanced Illness Research Center, University of Pennsylvania Perelman School of Medicine, Philadelphia (Weissman); Pulmonary, Allergy, and Critical Care Division, University of Pennsylvania Perelman School of Medicine, Philadelphia (Weissman).

Conflict of Interest Disclosures: Dr Friedman reported receiving grant funding from the National Institute on Aging during the conduct of the study. Dr Delgado reported receiving grant funding from the Agency for Health Care Quality and Research, Patient Centered Outcomes Research Institute, National Institutes of Health, the Abramson Family Foundation, and the AAA Foundation for Traffic Safety outside the submitted work. No other disclosures were reported.

REFERENCES


