TURKISH JOURNAL OF EMERGENCY MEDICINE

Original Article

Access this article online

Quick Response Code:

Website: www.turkjemergmed.com

DOI: 10.4103/tjem.tjem_79_23

Performance of emergency triage prediction of an open access natural language processing based chatbot application (ChatGPT): A preliminary, scenario-based cross-sectional study

İbrahim Sarbay[1*], Göksu Bozdereli Berikol[2], İbrahim Ulaş Özturan[3,4]

[1]Department of Emergency Medicine, Keşan State Hospital, Edirne, Turkey, [2]Department of Emergency Medicine, Bakırköy Dr. Sadi Konuk Training and Research Hospital, İstanbul, Turkey, [3]Department of Emergency Medicine, Kocaeli University, Faculty of Medicine, Kocaeli, Turkey, [4]Department of Medical Education, Acibadem University, Institute of Health Sciences, Istanbul, Turkey
**Corresponding author*

Abstract:

OBJECTIVES: Artificial intelligence companies have been increasing their initiatives recently to improve the results of chatbots, which are software programs that can converse with a human in natural language. The role of chatbots in health care is deemed worthy of research. OpenAI's ChatGPT is a supervised and empowered machine learning-based chatbot. The aim of this study was to determine the performance of ChatGPT in emergency medicine (EM) triage prediction.

METHODS: This was a preliminary, cross-sectional study conducted with case scenarios generated by the researchers based on the emergency severity index (ESI) handbook v4 cases. Two independent EM specialists who were experts in the ESI triage scale determined the triage categories for each case. A third independent EM specialist was consulted as arbiter, if necessary. Consensus results for each case scenario were assumed as the reference triage category. Subsequently, each case scenario was queried with ChatGPT and the answer was recorded as the index triage category. Inconsistent classifications between the ChatGPT and reference category were defined as over-triage (false positive) or under-triage (false negative).

RESULTS: Fifty case scenarios were assessed in the study. Reliability analysis showed a fair agreement between EM specialists and ChatGPT (Cohen's Kappa: 0.341). Eleven cases (22%) were over triaged and 9 (18%) cases were under triaged by ChatGPT. In 9 cases (18%), ChatGPT reported two consecutive triage categories, one of which matched the expert consensus. It had an overall sensitivity of 57.1% (95% confidence interval [CI]: 34–78.2), specificity of 34.5% (95% CI: 17.9–54.3), positive predictive value (PPV) of 38.7% (95% CI: 21.8–57.8), negative predictive value (NPV) of 52.6 (95% CI: 28.9–75.6), and an F1 score of 0.461. In high acuity cases (ESI-1 and ESI-2), ChatGPT showed a sensitivity of 76.2% (95% CI: 52.8–91.8), specificity of 93.1% (95% CI: 77.2–99.2), PPV of 88.9% (95% CI: 65.3–98.6), NPV of 84.4 (95% CI: 67.2–94.7), and an F1 score of 0.821. The receiver operating characteristic curve showed an area under the curve of 0.846 (95% CI: 0.724–0.969, $P < 0.001$) for high acuity cases.

CONCLUSION: The performance of ChatGPT was best when predicting high acuity cases (ESI-1 and ESI-2). It may be useful when determining the cases requiring critical care. When trained with more medical knowledge, ChatGPT may be more accurate for other triage category predictions.

Keywords:

Chatbot, ChatGPT, emergency severity index, triage

Submitted: 18-03-2023
Revised: 13-04-2023
Accepted: 24-05-2023
Published: 26-06-2023

ORCID:
IS: 0000-0001-8804-2501
GBB: 0000-0002-4529-3578
IUO: 0000-0002-1364-5292

Address for correspondence:
Dr. İbrahim Sarbay, Department of Emergency Medicine, Keşan State Hospital, Aşağı Zaferiye Mahallesi Evreşe Caddesi Keşan, Edirne, Turkey.
E-mail: ibrahimsar@gmail.com

This is an open access journal, and articles are distributed under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 License, which allows others to remix, tweak, and build upon the work non-commercially, as long as appropriate credit is given and the new creations are licensed under the identical terms.

For reprints contact: WKHLRPMedknow_reprints@wolterskluwer.com

How to cite this article: Sarbay İ, Berikol GB, Özturan İU. Performance of emergency triage prediction of an open access natural language processing based chatbot application (ChatGPT): A preliminary, scenario-based cross-sectional study. Turk J Emerg Med 2023;23:156-61.

© 2023 Turkish Journal of Emergency Medicine | Published by Wolters Kluwer - Medknow

Box-ED section

What is already known on the study topic?

- Chatbots may help to improve the quality of patient care and accelerate health-care delivery.

What is the conflict on the issue? Has it importance for readers?

- Artificial intelligence and chatbots have made great progress in recent years. Evaluating the potential of these developments in emergency medicine is of interest.

How is this study structured?

- This is a preliminary cross-sectional study conducted with case scenarios generated by the researchers based on the Emergency Severity Index handbook v4 cases.

What does this study tell us?

- Although it needs to be validated with real large sample size data, ChatGPT appears to be useful in determining the cases requiring critical care in its present form.

Introduction

Chatbots are artificial intelligence software programs that imitate human speech using natural language processing methods.[1] Artificial intelligence companies that investigate natural language processing and neural network-based chatbots have recently been increasing their development of chatbots to improve the results of artificial intelligence in response to human speech, especially for social media.[2,3] These systems, which have been used for user support and consultancy on many online platforms for years, have recently been introduced in the academic community in the field of health care.[4-6]

Emergency triage is the classification of patients presenting to health-care services in terms of the urgency of the need to see a physician.[7] Triage may be based on presenting complaints, vital signs or need for resources.[8-10] The presenting complaints used for determining triage level are mostly defined from structured triage complaint lists.

Natural language processing is a set of computational techniques that are theoretically able to analyze texts at one or more levels of linguistic analysis for the purpose of human-like language processing for a set of tasks.[11] Currently, there are several studies into determining patient triage levels using natural language processing and machine learning, a task usually performed by specialist triage personnel.[12-15] In particular, estimation of outcome from data obtained from unstructured complaint data and prediction of outcomes, such as need for hospitalization, discharge, triage and mortality, from the minimum data obtained from the moment of admission to the emergency department have recently gained importance.[16-19] These studies aim to reduce the variability and complexity of triage and to increase the quality of patient care. Multiple triage scales are used, and one of the most popular among these is the Emergency Severity Index (ESI).[20] ESI is a five-level triage system categorizing patients from level 1 (most urgent) to level 5 (least urgent) based on the urgency of their medical conditions and the resources required.[21]

ChatGPT is designed as a supervised and empowered machine learning-based chatbot, trained through Reinforcement Learning from Human Feedback, developed by OpenAI.[22] With this dynamic model trained from big data on the Internet, it has become an application that can perform actions such as providing information, answering questions, and creating content with high accuracy.[23] In this study, we aimed to measure the performance of ChatGPT on triage prediction in an emergency medicine (EM) scenario.

Methods

This was a preliminary cross-sectional study conducted with the case scenarios generated by EM specialists and based on the cases provided in the ESI handbook v4.[21] The scenarios were limited to presenting symptoms and vital signs, and while they were similar to the ones in the handbook in terms of style, they were entirely different [Supplement 1]. Following the Ethical Approval, a total of 50 case scenarios were prepared. The case scenarios were electronically prepared and distributed to two independent EM specialists who were experts in the ESI triage scale to determine blindly their triage category as multiple-choice questions. A blinded third independent EM specialist was consulted in the case of conflicting categorization. The EM specialists each have an experience of more than 5 years in academic emergency departments and currently teaching emergency physicians about triage. They were not aware of how many cases were intended for each level, and cases were randomized to prevent them from being ranked in increasing order of urgency. The agreed result of each case scenario was assumed as the reference triage category. Subsequently, each case scenario was queried with the January 9 version of ChatGPT to determine the chatbot-determined triage category [Figure 1]. The answer given by the software was recorded as the index triage category [Supplement 2 – reference and index triage categories]. Inconsistent classifications between the ChatGPT and reference category were defined as over-triage (false positive) or under-triage (false negative). For the prediction performance of high acuity cases, a 2-tier classification was carried out. ESI-1 and ESI-2 were defined as "high acuity", while ESI-3, ESI-4,

and ESI-5 levels were defined to be "moderate and low acuity" levels.

Interrater reliability was determined using Cohen's Kappa. A confusion matrix was constructed with the predictions of ChatGPT and the reference triage categories. The sensitivity (Recall), specificity, positive values (PPV or Precision), and negative predictive values (NPV) and F1 scores for each triage category were evaluated. F1 scores were calculated as follows:

F1 score = 2 × ([precision × recall]/[precision + recall])

A receiver operating characteristic (ROC) curve was constructed to evaluate the predictive performance of ChatGPT [Figure 2].

Institutional review board approval was obtained for this study on February 9, 2023 (GOKAEK-2023/03.12).

Results

Fifty case scenarios were prepared and analyzed by EM specialists and subsequently analyzed by ChatGPT. Of the scenarios, nine cases were labeled as ESI category 1, 12 cases as category 2, 10 cases as category 3, six cases as category 4, and 13 cases as category 5, by the EM specialists. Reliability analysis showed a fair agreement between EM specialists and ChatGPT (Cohen's Kappa: 0.341). Eleven cases (22%) were over triaged, and nine (18%) cases were under triaged by Chat GPT. For nine cases (18%), ChatGPT reported two consecutive triage categories, one of which matched the expert consensus. ChatGPT had an overall sensitivity of 57.1% (95% confidence interval [CI]: 34–78.2), specificity of 34.5% (95% CI: 17.9–54.3), PPV of 38.7% (95% CI: 21.8–57.8), NPV of 52.6 (95% CI: 28.9–75.6), and an F1 score of 0.461. ChatGPT showed better predictive performance for ESI category 1 with a sensitivity of 88.9% (95% CI: 51.8–99.7), specificity of 95.1% (95% CI: 83.5–99.4), PPV of 80% (95% CI: 44.4–97.5), NPV of 97.5 (95% CI: 86.8–99.9), and an F1 score of 0.842. The performance for other triage categories is presented in Table 1. F1 scores for ESI 2-3-4-5 were 0.500, 0.435, 0.273, and 0.245, respectively. The confusion matrix is presented in Table 2.

In the 2-tier analysis, reliability analysis showed a good agreement between EM specialists and ChatGPT (Cohen's Kappa was 0.707). ChatGPT showed a sensitivity of 76.2% (95% CI: 52.8–91.8), specificity of 93.1% (95% CI: 77.2–99.2), PPV of 88.9% (95% CI: 65.3–98.6), NPV of 84.4 (95% CI: 67.2–94.7), and an F1 score of 0.821 for high acuity cases. The ROC curve showed an area under the curve of 0.846 (95% CI: 0.724–0.969, $P < 0.001$).

Discussion

There are many studies with chatbots and conversational agents used in communication, social media, computer engineering/sciences, and social arts.[1] Chatbots have been used in many fields including education, mental healthcare, maternity, chronic diseases by populations for screening in the COVID-19 pandemic.[6,23-25] The present study is the first use of ChatGPT for predicting triage categories in the literature, to the best of our knowledge. The overall performance of ChatGPT in

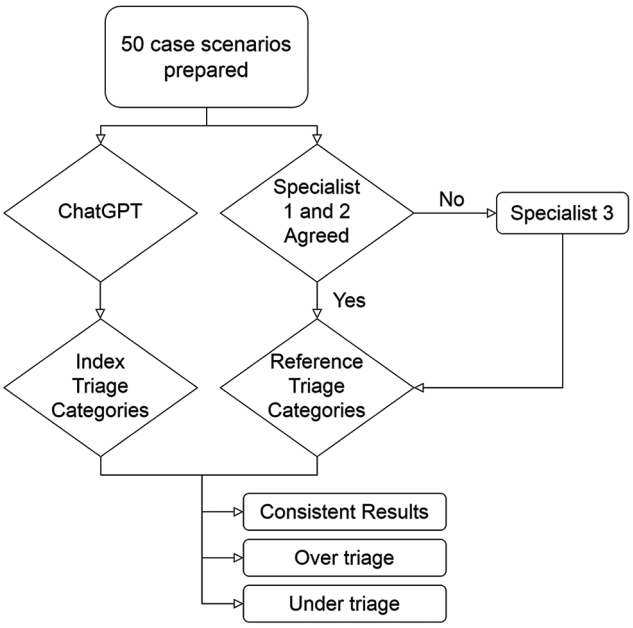

Figure 1: Study design

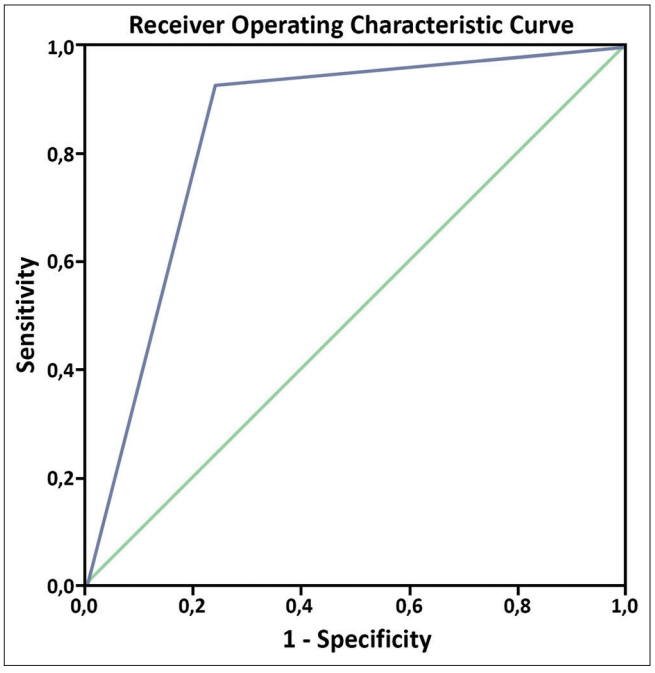

Figure 2: Receiver operating characteristic curve for high acuity cases

Table 1: Predictive performance of chat generative pretraining transformer for each emergency severity index triage category

Triage category	Percentage (95% CI)			
	Sensitivity	Specificity	PPV	NPV
ESI 1	88.9 (51.8–99.7)	95.1 (83.5–99.4)	80 (44.4–97.5)	97.5 (86.8–99.9)
ESI 2	41.7 (15.2–72.3)	92.1 (78.6–98.3)	62.5 (24.5–91.5)	83.3 (68.6–93)
ESI 3	50 (18.7–81.3)	80 (64.4–90.9)	38.5 (13.9–68.4)	86.5 (71.2–95.5)
ESI 4	50 (11.8–88.2)	70.5 (54.8–83.2)	18.8 (4–45.6)	91.2 (76.3–98.1)
ESI 5	15.4 (1.9–45.5)	97.3 (85.8–99.9)	66.7 (9.4–99.2)	76.6 (62–87.7)

ESI: Emergency severity index, PPV: Positive predictive value, NPV: Negative predictive value, CI: Confidence interval

Table 2: Confusion matrix

Index triage category	Reference triage category				
	1	2	3	4	5
1	7	1	0	0	0
2	2	8	1	0	0
3	0	2	6	1	1
4	0	1	3	4	8
5	0	0	0	1	4

five-level ESI triage categories was poor, while it showed better results in distinguishing high acuity cases.

Some studies have examined the use of natural language processing models in Emergency Triage in the past. These often focus on triage scores, hospitalization, and critical illness estimation.[13-15] In the study of Ivanov *et al*.,[15] the developed model was able to detect the ESI score with an accuracy of 75.9% which was higher than that of nurses (59.8%). Sterling *et al*.[14]'s model were successful in predicting the number of required resources based on machine learning of nursing triage notes and clinical data. In ESI triage classification, it is important to determine each case's exact level correctly, but correctly deciding whether the case has a triage level of 1–2 (high acuity) or 3–5 (moderate and low acuity) is vital as it determines whether the patient should be treated immediately.[15] ChatGPT showed a sensitivity of 76.2% (95% CI: 52.8–91.8), specificity of 93.1% (95% CI: 77.2–99.2) for high acuity cases in our study.

Kim *et al*.[4] studied the prediction of medical specialties via chatbot. These authors found that a single prediction yielded an accuracy of 70.6%, while three predictions showed a higher accuracy of 88.5%. Hirosawa *et al*.[26] studied the prediction of diagnosis, including five possible differential diagnoses, in ten patient scenarios. They found physicians' predictions to be 98% and 93.3% accurate while these values were 83.3% and 53.3% with ChatGPT-3, respectively.[26] These studies show the performance of chatbots increase when predicting a list for outcome, both for triage and diagnosis decisions.

Benoit studied 45 pediatric case scenarios written by the January version of ChatGPT with different points of view as parent, physician and grade 8 reading level person.[27] After generating the cases, the system's performance of triage (emergency, nonemergency care, and self-care) was 57.8% accurate. Low performance occurs not only when the result is requested as a single answer, but also when the triage result has been defined with the most urgent category in the original paper which the scenarios had been used.[27,28] Furthermore, using the same vignettes, another study yielded an accuracy of 51% with twelve different symptom checkers.[29] It was also reported that some symptom checkers showed higher performances for safe discharge. In our study, we found ChatGPT had higher specificity in predicting high acuity cases that may be useful for the correct selection of non-ESI 1–2 patients.

Ghosh *et al*.[30] studied natural language processing with symptom checker chatbots. These authors used 30 clinical patient vignettes, classified into emergency care, GP or self-care.[30] They found 83.3% and 66.6% accuracy when predicting one or two expected conditions. Notably, for emergency care they reported 100% recall. In the present study, among high acuity questions, we found 76.2% (95% CI: 52.8–91.8) recall. The overall precision average was 0.82 in the study of Ghosh *et al*., while in the present study, this was 0.89% (95% CI: 0.65–0.99).

Limitations

The first limitation is the number of questions and participants. There was 30% inter-rater disagreement in selecting the right triage category and this 30% needed a third opinion. The physician decision was achieved by majority voting, which may have introduced some bias. The reason for this may be the subjective nature of triage, relying on physician experience and varying by country or hospital conditions where triage is performed, as in the ESI guideline. We preferred to create new case scenarios instead of data that a trained system could have potentially encountered previously. The reliability of the results may be low when we evaluated the statistics as subgroups. Earlier studies compared nonemergency, emergency and self-care cases although the present study investigated chatbot performance in five-level triage. We suggest external validation with other applications to compare systems' performance. Further studies should be conducted with big data of real

patient records for the big language models. The second limitation is that it may not be generalizable to all natural language processing-based AI systems and chatbots, as the study used a single chatbot, chatGPT. However, the results may further be improved with training with more verified or real medical data. Since it is related to the data it receives from the performance database, we have limited knowledge of how much the content of the health database is and how much it has increased. We used the January 9 Version of ChatGPT and these results may change following updates of the system.

Conclusion

The performance of the chatbot, ChatGPT, was higher when predicting the high acuity triage categories (ESI-1 and ESI-2). Although it needs to be validated with real large sample size data, ChatGPT may be useful when determining cases requiring critical care. When trained with more medical knowledge, ChatGPT may be useful for general triage prediction.

Author contributions CReDiT statement

İbrahim Sarbay: Conceptualization, methodology, investigation, software, resources, data curation, visualization, writing – review and editing, supervision, Project administration.

Göksu Bozdereli Berikol: Conceptualization, methodology, investigation, software, resources, data curation, visualization, writing – review and editing.

İbrahim Ulaş Özturan**:** Conceptualization, methodology, formal analysis, investigation, writing – review and editing.

Conflict of interest

None Declared.

Ethical approval

Institutional review board approval was obtained for this study on 09.02.2023. (Kocaeli University Non-Interventional Clinical Research Ethics Committee - GOKAEK-2023/03.12).

Consent to participate

No informed consent was required for this study.

Funding

None.

References

1. Caldarini G, Jaf S, McGarry K. A literature survey of recent advances in chatbots. Information 2022;13:41.
2. Kolter JZ. AlphaCode and "data-driven" programming. Science 2022;378:1056.
3. Li Y, Choi D, Chung J, Kushman N, Schrittwieser J, Leblond R, *et al*. Competition-level code generation with AlphaCode. Science 2022;378:1092-7.
4. Kim Y, Kim JH, Kim YM, Song S, Joo HJ. Predicting medical specialty from text based on a domain-specific pre-trained BERT. Int J Med Inform 2023;170:104956.
5. King MR. The future of AI in medicine: A perspective from a chatbot. Ann Biomed Eng 2023;51:291-5.
6. Tzelios C, Contreras C, Istenes B, Astupillo A, Lecca L, Ramos K, *et al*. Using digital chatbots to close gaps in healthcare access during the COVID-19 pandemic. Public Health Action 2022;12:180-5.
7. Wuerz R, Fernandes CM, Alarcon J. Inconsistency of emergency department triage. Emergency department operations research working group. Ann Emerg Med 1998;32:431-5.
8. Baumann MR, Strout TD. Evaluation of the emergency severity index (version 3) triage algorithm in pediatric patients. Acad Emerg Med 2005;12:219-24.
9. Bullard MJ, Musgrave E, Warren D, Unger B, Skeldon T, Grierson R, *et al*. Revisions to the Canadian emergency department triage and acuity scale (CTAS) guidelines 2016. CJEM 2017;19:S18-27.
10. Unger B, Afilalo M, Boivin JF, Bullard M, Grafstein E, Schull M, *et al*. Development of the Canadian emergency department diagnosis shortlist. CJEM 2010;12:311-9.
11. Travers DA, Haas SW. Evaluation of emergency medical text processor, a system for cleaning chief complaint text data. Acad Emerg Med 2004;11:1170-6.
12. Tootooni MS, Pasupathy KS, Heaton HA, Clements CM, Sir MY. CCMapper: An adaptive NLP-based free-text chief complaint mapping algorithm. Comput Biol Med 2019;113:103398.
13. Sterling NW, Patzer RE, Di M, Schrager JD. Prediction of emergency department patient disposition based on natural language processing of triage notes. Int J Med Inform 2019;129:184-8.
14. Sterling NW, Brann F, Patzer RE, Di M, Koebbe M, Burke M, *et al*. Prediction of emergency department resource requirements during triage: An application of current natural language processing techniques. J Am Coll Emerg Physicians Open 2020;1:1676-83.
15. Ivanov O, Wolf L, Brecher D, Lewis E, Masek K, Montgomery K, *et al*. Improving ED emergency severity index acuity assignment using machine learning and clinical natural language processing. J Emerg Nurs 2021;47:265-78.e7.
16. Lee S, Mohr NM, Street WN, Nadkarni P. Machine learning in relation to emergency medicine clinical and operational scenarios: An overview. West J Emerg Med 2019;20:219-27.
17. Lee SH, Levin D, Finley PD, Heilig CM. Chief complaint classification with recurrent neural networks. J Biomed Inform 2019;93:103158.
18. Thompson DA, Eitel D, Fernandes CM, Pines JM, Amsterdam J, Davidson SJ. Coded chief complaints – Automated analysis of free-text complaints. Acad Emerg Med 2006;13:774-82.
19. Beitel AJ, Olson KL, Reis BY, Mandl KD. Use of emergency department chief complaint and diagnostic codes for identifying respiratory illness in a pediatric population. Pediatr Emerg Care 2004;20:355-60.
20. Kuriyama A, Urushidani S, Nakayama T. Five-level emergency triage systems: Variation in assessment of validity. Emerg Med J 2017;34:703-10.
21. Gilboy N, Tanabe P, Travers DA. The emergency severity index version 4: Changes to ESI level 1 and pediatric fever criteria. J Emerg Nurs 2005;31:357-62.
22. ChatGPT: Optimizing Language Models for Dialogue. Available from: https://openai.com/blog/chatgpt/. [Last accessed on 2023 Feb 25].
23. Kittipimpanon K, Noyudom A, Panjatharakul P, Visudtibhan PJ. Use of and satisfaction with mobile health education during the COVID-19 pandemic in Thailand: Cross-sectional study. JMIR Form Res 2023;7:e43639.
24. He Y, Yang L, Zhu X, Wu B, Zhang S, Qian C, *et al*. Mental health chatbot for young adults with depressive symptoms during the COVID-19 pandemic: Single-Blind, three-arm randomized controlled trial. J Med Internet Res 2022;24:e40719.
25. Goonesekera Y, Donkin L. A cognitive behavioral therapy chatbot (Otis) for health anxiety management: Mixed methods pilot study. JMIR Form Res 2022;6:e37877.

26. Hirosawa T, Harada Y, Yokose M, Sakamoto T, Kawamura R, Shimizu T. Diagnostic accuracy of differential-diagnosis lists generated by generative pretrained transformer 3 chatbot for clinical vignettes with common chief complaints: A pilot study. Int J Environ Res Public Health 2023;20:3378.
27. Benoit JR. ChatGPT for clinical vignette generation, revision, and evaluation. medRxiv 2023.02.04.23285478. Available from: https://doi.org/10.1101/2023.02.04.23285478. [Last accessed on 2023 June 5].
28. Semigran HL, Linder JA, Gidengil C, Mehrotra A. Evaluation of symptom checkers for self diagnosis and triage: Audit study. BMJ 2015;351:h3480.
29. Ceney A, Tolond S, Glowinski A, Marks B, Swift S, Palser T. Accuracy of online symptom checkers and the potential impact on service utilisation. PLoS One 2021;16:e0254088.
30. Ghosh S, Bhatia S, Bhatia A. Quro: Facilitating user symptom check using a personalised chatbot-oriented dialogue system. Stud Health Technol Inform 2018;252:51-6.

Supplement 1: Triage Scenarios

1. A 30-year-old female patient is brought to the ED due to sudden onset of unconsciousness. On her arrival she appears comatose. Pupils were miotic, GCS: 6, BP: 100/60 mmHg, HR: 110, RR: 14, SpO2: 90%, T: 98° F. First responders told you that they found an empty syringe at the scene
2. A 60-year-old male patient presents to the emergency department with complaints of fainting and impaired consciousness. Vital signs on arrival: BP: 70/30 mmHg, HR: 25/min, RR: 15, SpO2: 99%, T: 98° F
3. A 90-year-old male patient is brought to the ED by ambulance. It is stated that the patient fainted in the toilet and his stool seemed bloody. He seems pale. Intense red blood is seen on the patient's diaper. Vital signs on arrival: BP: 65/30 mmHg, HR: 140/min, RR: 25. Medication history is notable for anticoagulants.
4. A 25-year-old male patient is brought to the ED by ambulance due to a gunshot wound to the abdomen. On arrival, he is conscious and oriented. The bullet entrance wound is located on the left upper quadrant of the abdomen. No other wounds were found in the primary evaluation. BP: 80/50 mmHg, HR: 120/min, RR: 20/min. It is learned that up to 500 ml of fluid was given during the transfer
5. A 60-year-old female patient with a diagnosis of COPD is brought to the ED with respiratory distress. On arrival, her oxygen saturation is measured as 70%. BP: 90/55 mmHg, HR: 115/min, RR: 25/min
6. A 65-year-old female patient presents with pressure-like pain located in the middle of the chest that has been going on for 1 hour. She appears to have a cold sweat. BP: 60/palp, heart rate: 140/min, RR: 25
7. A 50-year-old patient with a history of heart failure and hypertension is brought to the ED with shortness of breath. BP: 190/90 mmHg, HR: 100/min, Spo2: 78%, RR: 30/min. A gasping sound when breathing is notable
8. A 20-year-old male patient was found unconscious at home. Empty boxes of prescribed drugs he used for the treatment of depression were found next to him. On arrival he appears comatose. Pupils are miotic, GCS: 5, BP: 90/60 mmHg, HR: 110, RR: 10, SpO2: 89%
9. EMS is activated for a 30-year-old female patient who was found unconscious at home. First responders found the patient was in cardiac arrest and spontaneous circulation was established with successful CPR. On arrival to the ED BP: 90/50 mmHg, HR: 130, spo2: 99%
10. A 70-year-old patient who presented to the ED with abdominal pain. Seems pale. Capillary refill time is prolonged. Vital signs on arrival: BP: 60/palp, HR: 130, spo2: 97%, Temp: 36 C
11. A 36-year-old female presents to the ED. She has a history of severe allergic reaction due to a bee sting. She told you that a bee stung her arm. In a couple of minutes, she felt dizzy and nauseous. Vital Signs: BP 145/74, HR 117, RR 19, SpO2: 98%, T: 97° F
12. 12-A 50-year-old male patient is admitted to the ED with complaints of palpitation and fainting that has been going on for a few hours. Vital Signs: BP: 100/40 mmHg, HR: 190, RR: 15, SpO2: 99%, T: 98° F
13. A 40-year-old female patient presents with pressure-like pain in the middle of the chest that has been going on for 2 hours. BP: 130/80 mmHg, heart rate: 70/min, RR: 14, T: 36.5 C
14. A 30-year-old female patient comes with a sudden onset of severe headache and nausea after lifting a heavy object. She is conscious, oriented and cooperative and states that the pain is the most severe pain of her life. BP: 120/75 mmHg, HR: 80, RR: 15, T: 36.5 C
15. A 25-year-old physician comes to the ED after a needle used while administering medication to her patient accidentally stabbed her hand
16. A 26-year-old patient, who is 20 weeks pregnant, presents to the ED with the complaint of pain and "water breaking ". BP: 120/85 mmHg, HR: 90, RR: 15, T: 36.5 C
17. A 50-year-old male patient presents to the ED with complaints of high fever, and chills. It is learned that he is on chemotherapy for lung cancer. BP: 120/69 mmHg, HR: 105, SpO2: 96, RR: 16, T: 38.8 C
18. An 18-year-old female patient was rescued by her relatives after an attempt to hang herself at home and brought to the ED by ambulance. There was no loss of consciousness. No markings are present on her neck. BP: 130/75 mmHg, HR: 100, SpO2: 98, RR: 14, T: 36.8 C
19. A 90-year-old male patient is brought to the ED because of the confusion that started on the same day. On arrival he is conscious, but gives confused answers. BP: 120/69 mmHg, HR: 80, SpO2: 98, RR: 14, T: 36.6 C
20. A 40-year-old female patient presents with severe left flank pain. Her skin looks pale. In her medical history, it is learned that she presented to the ED due to kidney stones before. Pain score is 9/10. BP: 140/90 mmHg, HR: 110, SpO2: 98, RR: 17, T: 36.6 C
21. A 20-year-old male patient presents to the ED with complaints of nausea and vomiting that started in the morning and pain localized to the right lower quadrant within hours. Pain score is 6/10. BP: 130/90 mmHg, heart rate: 100, SpO2: 98, RR: 16, T: 36.4 C

22. A 20-year-old male patient presents to the ED with the suspicion of a fracture of the forearm after falling on his arm while playing basketball. A prominent displaced fracture is observed on the proximal wrist. Radial pulse is palpable and sensation and circulation are evaluated as normal
23. A 34-year-old female patient comes with complaints of high fever and cough for two days. Describes chest pain that increases with coughing. She seems healthy. BP: 120/90 mmHg, heart rate: 100, SpO2: 98, RR: 16, T: 38.4 C
24. A 40-year-old female patient presents with the complaint of pain in the upper quadrants of the abdomen that started after lunch today. She has a history of gallstones and DM in her history. BP: 130/80 mmHg, heart rate: 87, SpO2: 98, RR: 14, T: 36.2 C
25. A 27-year-old female patient presents to the ED with complaints of nausea, vomiting and diarrhea that started about 1 day ago. She describes abdominal pain that is relieved by diarrhea and is not very severe. A few hours before her symptoms started, she had eaten a meal that "tasted strange". Diarrhea is watery and includes mucus and does not contain blood. Her skin appears dry but turgor is normal. BP: 140/90 mmHg, heart rate: 89, SpO2: 99, RR: 15, T: 36.4 C
26. A 35-year-old female patient is presented to the ED due to the pain in her left leg for 2 days. She previously had traveled on intercontinental flights. The left leg appears to be edematous. Distal pulses are palpable. On examination, the Homans test is positive. BP: 130/88 mmHg, heart rate: 85, SpO2: 97, RR: 15, T: 36.4 C
27. A 55-year-old male patient comes to the ED with the complaint of shortness of breath during exertion. No active complaints at the time of presentation. He states that he has arrhythmia and hypertension in his medical history. He doesn't use prescribed drugs. BP: 120/60 mmHg, heart rate: 97, SpO2: 95, RR: 16, T: 36.4 C
28. A 45-year-old female patient is brought to the ED by ambulance for falling from the first floor balcony. She has an open fracture in her left leg. She states that she fell on her left foot and describes no other injury. She has not lost consciousness. Pain Score: 9/10. BP: 110/80 mmHg, heart rate: 100, SpO2: 99, RR: 15, T: 36.4 C
29. A 3-year-old male patient is brought in with the complaint of slipping his foot and hitting his head on the cabinet door. There is a superficial laceration in the frontal head. He is alert. Family does not describe loss of consciousness. The family was worried because he vomited several times
30. A 24-year-old female patient presents to the ED with severe abdominal pain. It is learned that there is a delay in menstruation and that the pregnancy test she did 2 days ago was positive. BP: 110/64 mmHg, heart rate: 70, RR: 14, T: 36.5 C
31. A 22-year-old male patient comes with pain around the ankle following falling while running. No open wounds or deformity. Pain score is 8/10
32. A 30-year-old male patient presents with swelling in the nail bed of the left index finger. The appearance of the abscess formed within two days. He complaints of pain and throbbing. He has no fever
33. A 40-year-old female patient comes with complaints of frequent and painful urination. No known medical condition. She does not describe abdominal pain. BP: 120/90 mmHg, heart rate: 70, SpO2: 99, RR: 14, T: 36.4 C
34. A 15-year-old male patient comes with a complaint of edema in the nose after being hit with a ball while playing football. No deformity. No active nosebleeds. No history of syncope. No additional complaints. Oriented and cooperative. BP: 110/60 mmHg, heart rate: 70, SpO2: 99, RR: 14, T: 36.5 C
35. A 23-year-old female patient fell down while jogging. Presents to the ED due to a superficial laceration on the forearm. Range of motion of joints is normal. She does not describe pain in the extremity. There is no active bleeding from the wound. There is a simple superficial laceration of approximately 3-4 cm in length
36. A 30-year-old male patient comes with the complaint of eye itching after walking in windy weather. Thinks a particle of dust got in his eye. No known medical history. Pain score: 3/10
37. A 77-year-old female patient with a history of heart valve replacement and on anticoagulants presents to the ED with a couple of bruises in her legs. No additional complaints. Describes no bleeding. BP: 110/75 mmHg, heart rate: 80, SpO2: 99, RR: 14, T: 36.5 C
38. A 43-year-old male patient came to the ED because of an accidental knife cut on his index finger. The bleeding stopped with the dressing he applied at home. The cut seems superficial but may need some sutures. Distal sensory and circulatory examination is normal. Tendons and nerves seem intact. BP: 110/80 mmHg, heart rate: 80, SpO2: 99, RR: 14, T: 36.5 C
39. A 60-year-old female patient comes because a splinter penetrated her foot while walking. It is seen that there is a splinter located on the sole of the foot which is protruding from the skin
40. A 20-year-old male patient, who describes he has lost taste and smell for two days and suspects that he has contacted with COVID-19, is presented to the ED. He has no medical history. Respiratory sounds are normal. Seems healthy. BP: 130/85 mmHg, HR: 73, SpO2: 99, RR: 14, T: 36.7 C
41. An 8-year-old male patient presents with a burn with an area of 3 cm2 due to scald injury on his left forearm the day before. No bullae. Pain score: 3/10

42. A 10-year-old male patient comes to the Emergency Department with the complaint of itching and redness in both eyes. He has no fever. Vital parameters are normal. Does not describe pain
43. A 40-year-old female patient presents to the ED with a sore throat for 3 days. No swallowing difficulties or hoarseness. Examination of the throat is normal. BP: 120/75 mmHg, heart rate: 80, SpO2: 98, RR: 14, T: 36.8 C
44. A 20-year-old female patient has a history of nickel allergy. After wearing a metal bracelet, she complains of itching and redness on the area which contacts with the bracelet. No systemic reaction
45. A 50-year-old female patient with caries on her left molar teeth comes with the complaint of pain in the same area that starts while eating. No facial swelling. Pain score: 4/10. No fever
46. A 2-year-old female patient is brought to the ED with a runny nose and vomiting. She looks active and alert. Her mother states that her appetite is normal. Seems hydrated
47. A 27-year-old female patient presents with complaints of pain in both legs with movement and palpation, which started one day after heavy exercise. BP: 120/90 mmHg, HR: 70, SpO2: 99, RR: 14, T: 36.4 C. No difference in diameter between the two legs. Urine color is normal
48. A 10-year-old male patient presents with swelling in the nail bed of the right hand thumb. There is no pus-filled blister. He has no fever
49. An 18-year-old female patient presented to the ED due to stepping on a rusty nail. On inspection, you can not see a wound. She has no additional complaints and no fever
50. A 70-year-old male patient comes with pain on the left big toe while walking. He states that his complaints have been going on for about a few weeks. On palpation, a hard dark lesion is present located under the left big toe.

Supplement 2: Reference and ındex triage categories

Scenario ID	Reference triage category	Index triage category
1	1	1
2	1	1
3	1	1
4	1	1
5	2	2
6	1	1
7	2	1
8	1	1
9	1	1
10	1	1 or 2
11	1	2 or 3
12	2	1
13	2	2
14	2	3
15	4	3
16	2	2
17	3	2
18	2	2
19	2	3
20	2	2
21	3	3
22	3	4
23	3	3
24	3	3
25	5	4
26	3	3
27	3	2
28	2	3
29	3	3
30	2	3
31	2	4
32	4	4
33	5	4
34	5	3
35	4	3
36	4	4
37	3	4 or 5
38	4	4
39	3	5
40	4	3
41	5	4
42	5	4
43	5	5
44	5	5
45	5	4 or 5
46	5	4 or 5
47	5	4 or 5
48	5	4 or 5
49	5	4 or 5
50	5	4 or 5