Croat Nurs J. 2024; 8(1): 29-36

Professional paper

The Future of Triage: The Analysis of Traditional Methods Compared to ChatGPT

[1] **Helena Mayerhoffer**

[1] **University of Applied Health Sciences, Zagreb, Croatia**

Article received: 13.09.2023.

Article accepted: 12.12.2023.

Author for correspondence:
Helena Mayerhoffer
University of Applied Health Sciences, Mlinarska 38, Zagreb, Croatia
E-mail: mayerhoffer.helena@gmail.com

https://doi.org/10.24141/2/8/1/3

Keywords: ChatGPT, nursing, triage

Abstract

Introduction. Triage is the assessment of the patient's condition in order to determine the urgency of treatment. It is usually performed by a nurse, often using a five-level protocol.

Aim. To conduct a comparative analysis of the accuracy of categorization and diagnosis between ChatGPT (a chatbot which uses machine learning algorithms) and traditional medical triage, as well as to provide recommendations on how artificial intelligence can improve the work of medical professionals in patient triage.

Methods. The literature selected for comparison is "Emergency Nursing: 5-Tier Triage Protocols". The most common diagnoses for which patients present to the emergency department were selected for research. Then, triage categories were selected and case presentations were created. These cases were presented to ChatGPT, and its responses were compared with the literature.

Results. ChatGPT correctly categorizes triage cases in 43.33% of cases, with an average category difference of 0.7. Although it made mistakes in 1 or 2 categories in some cases, it assigned diagnoses to a higher category for patient safety.

Discussion. Comparison with other studies shows that errors occur in up to 40% of nurse decisions due to various factors such as inexperience, speed of work, and a large number of patients, which could be reduced by additional artificial intelligence assistance. It is necessary to take into account factors that artifi-

cial intelligence cannot take over and that it can only be a help, not a substitute for medical personnel.

Conclusion. ChatGPT has potential for usage in medical triage, but with improvements in specialized training of models on medical data and terminology to improve the accuracy and reliability of the model.

Introduction

Triage is a preliminary assessment of a patient's condition with the aim of determining the urgency of their need for treatment. It is a process which helps ensure that patients receive timely and appropriate care based on the severity of their condition. It is typically conducted by nurses using a set of established protocols designed to help them rapidly and accurately assess the patient's condition and determine the appropriate level of care (1). One of the triage protocols used is the five-level protocol. The level depends on the urgency at which the patient needs treatment and is divided into: the first level, where the patient requires immediate resuscitation, intubation, or emergency surgery; the second level, where there is a potentially life-threatening condition which requires urgent assessment and treatment (e.g., severe bleeding, chest pain, breathing difficulties); the third level, where the patient needs prompt attention, but is not in a life-threatening condition (e.g., bone fracture, moderate pain, fever); the fourth level, which includes conditions requiring medical attention within a few hours (e.g., minor cut, mild allergic reactions); and the fifth level, where the patient can wait for several hours or days (e.g., ankle sprain, minor rash). Protocols are categorized by the diagnose and are based on symptoms, with questions for each diagnosis which aid the nurse in assessing the patient's condition (1). With the development of technology and artificial intelligence capabilities, the question arises whether this process should be digitalized for faster processing. Chat robots like ChatGPT have been developed to assist with everyday human tasks. The question is whether artificial intelligence can aid in triage by using machine learning algorithms to analyze patients' symptoms and provide guidance to healthcare professionals. While chat robots have the potential to revolutionize the triage process, it's important to evaluate their effectiveness compared to traditional nurse triage protocols (2).

The aim of this research paper is to conduct a comparative analysis of traditional nurse triage protocols and ChatGPT's responses to case scenarios. The efficiency and accuracy of ChatGPT's responses to a prompt with case scenarios will be studied and compared with the existing data from the literature. The importance of this research is to evaluate the efficiency and accuracy of ChatGPT's categorization of medical cases and its ability to propose accurate diagnoses based on provided symptoms. This analysis is conducted to provide insights into the potential benefits and limitations of using ChatGPT for triage and to recommend how artificial intelligence can enhance the work of healthcare professionals. The research questions are: how does ChatGPT categorize medical cases into triage categories compared to categorization based on literature, what diagnoses does ChatGPT propose for given case presentations, and can ChatGPT be considered a useful tool for practical medical triage?

The hypotheses are that ChatGPT can effectively categorize medical cases into triage categories in accordance with recommendations from the literature, it can accurately provide diagnoses for given case presentations, and that it is a useful tool for practical medical triage.

The contributions of this research include a comparative analysis of traditional nurse triage protocols and ChatGPT, investigating the efficiency, accuracy, benefits and limitations of ChatGPT, as well as recommendations for improving the work of healthcare professionals using artificial intelligence.

Methods

In this research, a thorough literature review was conducted to select relevant literature for comparing data with the ChatGPT model. The book which was chosen was the "Emergency Nursing: 5-Tier Triage Protocols", published in 2020 (1). The research included the most common diagnoses for which

patients come to emergency departments, including cardiac arrest, cerebrovascular insult, chest pain, abdominal pain, allergic reactions, traumatic injuries, viral and bacterial respiratory diseases, breathing difficulties, burns, hypothermia, and diabetes related issues. Under each diagnosis, the triage categories and symptoms per category are listed. For the purposes of the research, three categories (out of a total of five) were randomly selected for each diagnosis. The random selection of triage categories serves multiple purposes. Firstly, it prevents intentional bias in the selection process, avoiding patterns which could inadvertently favor the model's training data. Randomization reduces the risk of favoring specific cases and ensures a representative sample for assessing ChatGPT's generalization across diverse medical conditions. This approach introduces variability, exposing the model to a spectrum of complex and less complex cases, thereby challenging it and revealing strengths and weaknesses. By simulating a realistic scenario where healthcare professionals encounter diverse cases daily, randomization contributes to the generalizability of the evaluation. Moreover, it mitigates the risk of model overfitting to specific categories during training, ensuring a more comprehensive and unbiased assessment. For the diagnosis of stroke and breathing difficulties, categories 1, 2 and 4 were chosen; for allergic reactions, chest and abdominal pain, categories 1, 2 and 3 were chosen; for traumatic injuries, burns and diabetes related issues, categories 2, 3 and 4 were chosen; for hypothermia, categories 2, 3 and 5 were chosen and for respiratory infections, categories 1, 3 and 4 were chosen. Only three categories were chosen to challenge the ChatGPT model in the triaging process. If all five categories were included, the model would distinguish between the categories more easily, potentially deviating from real in-hospital situations. A total of 30 case scenarios were analyzed to determine how ChatGPT triages the given patients and what diagnoses it suggests based on the written symptoms. The case scenario was presented to ChatGPT as patient exhibiting symptoms listed under the selected category from the previously mentioned book "Emergency Nursing: 5-Tier Triage Protocols". Such case scenarios were presented to ChatGPT with the following prompt: "Here are 5 emergency cases, read each of them, then categorize them into triage categories using the 5-Tier Triage Protocol and suggest a potential diagnosis." To avoid errors due to the character limit which ChatGPT can process and the length of responses, 5 cases were written in one question. After recording ChatGPT's responses, they were compared to the literature to assess accuracy. The model used was the basic ChatGPT-3, available on the Internet for free.

Results

In Table 1, diagnoses taken from the literature, diagnoses which ChatGPT assumed based only on those symptoms, triage categories in which those symptoms are classified according to the literature, the category in which ChatGPT triaged, the accuracy of ChatGPT's responses compared to the literature (correct-incorrect), and the difference in categories between the literature and ChatGPT (how much ChatGPT has erred) are shown.

The results of this study show that ChatGPT can classify medical cases into triage categories in accordance with literature recommendations, but with an 43.33% accuracy. The average difference in categorization between ChatGPT and literature is 0.7. When individual cases are examined, ChatGPT made errors for 2 categories in the 4th category of stroke, the 4th category of breathing difficulties, the 4th category of burns, and the 3rd category of diabetes related issues. It made errors for 1 category in the 2nd and 3rd categories of traumatic injuries, the 2nd category of breathing difficulties, the 2nd and 3rd categories of burns, the 2nd and 3rd categories of hypothermia, the 2nd and 3rd categories of chest pain, the 4th and 3rd categories of respiratory infections, and the 2nd and 4th categories of diabetes related issues. Cases where ChatGPT categorized correctly include the 1st and 2nd categories of stroke, categories 1-3 of allergic reactions, the 4th category of traumatic injuries, the 1st category of breathing difficulties, the 5th category of hypothermia, the 1st category of chest pain, categories 1-3 of abdominal pain, and the 1st category of respiratory infections.

In all cases where ChatGPT misclassified, it tended to categorize at a higher level, indicating a positive pattern where it prioritizes patient safety when there is insufficient information or when it perceives that the symptoms mentioned may indicate a more seri-

Table 1. **Comparison of ChatGPT's responses with sources from literature**

DIAGNOSIS LITERATURE	DIAGNOSIS CHATGPT	CATEGORY LITERATURE	CATEGORY CHATGPT	ACCURACY OF CHATGPT RESPONSE	DIFFERENCE BETWEEN CHATGPT AND LITERATURE CATEGORY
Stroke	Anaphylaxis or allergic reaction	1	1	+	0
Stroke	Stroke	2	2	+	0
Stroke	Neurological disorder	4	2	-	2
Allergic reaction	Acute airway obstruction	1	1	+	0
Allergic reaction	Anaphylaxis or angioedema	2	2	+	0
Allergic reaction	Allergic reaction or viral disease	3	3	+	0
Traumatic injury	Fracture or dislocation	2	1	-	1
Traumatic injury	Fracture	3	2	-	1
Traumatic injury	Fracture	4	4	+	0
Breathing difficulties	Heart attack or asthma attack	1	1	+	0
Breathing difficulties	Airway obstruction, severe allergic reaction	2	1	-	1
Breathing difficulties	Respiratory infection or pneumonia	4	2	-	2
Burn	3rd degree burn	2	1	-	1
Burn	Severe burn, possible airway involvement	3	2	-	1
Burn	2nd degree burn with signs of deep infection	4	2	-	2
Hypothermia	Hypothermia	2	1	-	1
Hypothermia	2nd degree frostbite	3	2	-	1
Hypothermia	Mild hypothermia	5	5	+	0
Chest pain	Cardiac arrest	1	1	+	0
Chest pain	Cardiac arrest	2	1	-	1
Chest pain	Deep vein thrombosis or cardiac arrest	3	2	-	1
Abdominal pain	Respiratory arrest, sepsis, cerebrovascular insult	1	1	+	0

Table 1. **Comparison of ChatGPT's responses with sources from literature**

DIAGNOSIS LITERATURE	DIAGNOSIS CHATGPT	CATEGORY LITERATURE	CATEGORY CHATGPT	ACCURACY OF CHATGPT RESPONSE	DIFFERENCE BETWEEN CHATGPT AND LITERATURE CATEGORY
Abdominal pain	Gastrointestinal bleeding	2	2	+	0
Abdominal pain	Appendicitis	3	3	+	0
Respiratory infection	Acute asthmatic attack, pneumonia, COVID-19	1	1	+	0
Respiratory infection	Respiratory infection	3	2	-	1
Respiratory infection	Bacterial infection	4	3	-	1
Diabetes related issued	Severe hypoglycemia, diabetic ketoacidosis	2	1	-	1
Diabetes related issued	Diabetic ketoacidosis	3	1	-	2
Diabetes related issued	Did not provide a diagnosis	4	3	-	1
TOTAL				43.33%	0.7

ous problem. When it comes to ChatGPT's ability to provide diagnoses based on provided symptoms, it performed well in most cases, even offering multiple diagnoses in some instances. The only case where it didn't provide a diagnosis was in the 4th category of diabetes related issues. The results presented in Table 1 indicate that ChatGPT is not a reliable tool for diagnosis. For instance, in the reference book, symptoms listed for stroke in category 1 include severe respiratory distress, paleness, diaphoresis, lightheadedness or weakness, and unresponsiveness. These symptoms also overlap with those of a category 1 allergic reaction. Consequently, ChatGPT failed to distinguish between the two due to the similarity in symptoms. This same behavior can be seen with some other diagnoses because ChatGPT lacks specific medical training and has limited understanding of the context in which the situation is happening.

Discussion

Research conducted in Turkey showed that the accuracy rate of decisions made by nurses in triage was 59.3%, meaning that 40.7% of decisions were incorrect (3). Additionally, research by Chen J. C. and colleagues (4) reported a 40% inaccuracy rate in nurses triage decisions, while in the study by Jordi K. and colleagues (5), it was 40.4%. The research from Turkey suggests that the number of patients in the emergency department significantly affects triage accuracy, with larger patient volumes leading to lower triage accuracy (3). It has also been demonstrated that nurses sometimes struggle with patient categorization and require more time, with approximately half of the patients who presented to the emergency department being placed in the 3rd category (3). Another factor affecting triage is the experience of the nurse, with those having less than a year of experience making about 10% more errors than those with around 4 years of experience (3). When taking these factors into con-

sideration, even though ChatGPT's categorization accuracy is 43.33%, the average difference in categories between ChatGPT and literature is only 0.7, which is a relatively small difference considering the impact of the mentioned factors on triage speed and accuracy. It's important to note that in cases where ChatGPT misclassified, it tended to categorize at a higher level, thereby prioritizing patient safety (3).

The results of Benoit J.R.A.'s study show that ChatGPT is successful in diagnosing simple cases in 71.1% of cases and correctly triaging 57.8% of cases (2). The higher percentage in this study is due to several factors. The author categorized cases into three types: emergencies, non-emergencies, and cases which can be managed at home. This is not a standard categorization, and it is unclear how emergency cases were ranked. The experiment's data do not specify how case scenarios were presented, only that they were simple scenarios. This study has raised questions about ChatGPT's capabilities in triage and presents opportunities for further research (2).

A cross-sectional study by İbrahim et al. evaluated the performance of ChatGPT in predicting triage categories in an Emergency Room (ER) setting. The researchers generated case scenarios based on the Emergency Severity Index. Two independent ER specialists categorized the cases, and a third specialist resolved any conflicting categorizations. ChatGPT was then used to predict triage categories, and its performance was compared to expert classifications. The study found fair agreement between ChatGPT and ER specialists, with a Cohen's Kappa of 0.341. The sensitivity for high acuity cases was 76.2%, while specificity was 93.1%. The study suggests that ChatGPT, while showing promise in distinguishing high acuity cases, has limitations in accurately predicting triage categories overall. The researchers recommend further validation with larger datasets and highlight the importance of considering the subjective nature of triage and potential biases in the decision-making process (6).

A study conducted by Gebrael G. et al. demonstrated a diagnostic performance of 87.5% using ChatGPT in emergency cases for patients with metastatic prostate cancer. The study highlighted limitations in determining the need for hospital admission. The researchers underscored the significance of developing an AI model for this purpose and emphasized the potential benefits of utilizing AI in emergency room settings (7).

Researchers who compared the diagnostic and triage accuracy of ChatGPT 3.5, ChatGPT 4.0, Ada and WebMD showed that in the diagnostic analysis, ChatGPT 3.5 exhibited the highest diagnostic accuracy, with a top-3 diagnostic match rate of 63%. However, it also had a concerning high unsafe triage rate of 41%, signifying instances where the triage recommendations could be potentially harmful or inappropriate. On the other hand, ChatGPT 4.0 demonstrated lower diagnostic accuracy compared to ChatGPT 3.5, with a top-3 diagnostic match rate of 50%. However, it presented a notably lower unsafe triage rate of 22% and achieved the highest triage agreement rate (76%) with the physicians among all models. This suggests that ChatGPT 4.0, despite its reduced diagnostic accuracy, performed better in terms of providing triage recommendations which align with physician's assessments while minimizing unsafe suggestions (8).

Despite its low accuracy, ChatGPT has demonstrated some positive results in diagnosing and categorizing cases. This suggests the potential for improvement through additional specialized training and model optimization for medical purposes. ChatGPT has the potential to expedite triage and assist less experienced healthcare professionals, but its accuracy is not high enough to rely solely on its responses.

Of course, there are aspects which ChatGPT cannot replace, such as empathy, human touch and comforting words which are essential when patients are in panic or pain and seeking help. Individuals also display a range of nonverbal signals, such as body language and facial expressions, which can offer valuable insights into a patient's condition, which require human observation. There is also a responsibility issue, i.e. if a healthcare provider relies on ChatGPT, they must bear the responsibility if a negative outcome occurs. In general, while ChatGPT has the potential to assist in medical triage, it should be considered a tool which adds to the knowledge and work of nurses rather than replacing them.

Conclusion

By analyzing the presented research results, we can assess how ChatGPT categorizes medical cases in triage compared to literature-based categorization and the diagnoses it provides for given case presentations. The overall accuracy of ChatGPT's responses is 43.33%, with an average category difference of 0.7 compared to the literature. Based on these results, we can examine the hypotheses put forth. The first hypothesis was that ChatGPT can effectively categorize medical cases into triage categories according to literature recommendations. The results show variable accuracy of ChatGPT in categorizing medical cases into triage categories. While some cases align with the literature, there are deviations in others. Therefore, we cannot fully confirm this hypothesis at this time. The second hypothesis was that ChatGPT can accurately diagnose given case presentations. ChatGPT provides different diagnoses for given case presentations, but the accuracy of these diagnoses varies. Therefore, this hypothesis is also partially confirmed, with room for improvement. The third hypothesis was that ChatGPT is a useful tool for practical medical triage. Considering the variable accuracy and consistency of ChatGPT in diagnosing and categorizing cases, we cannot currently consider ChatGPT as an accurate tool for practical medical triage. Therefore, this hypothesis cannot be confirmed at this time. However, the research results highlight the potential of artificial intelligence in the clinical environment, and with further improvement and model adaptation, better accuracy and reliability can be achieved. In conclusion, ChatGPT shows potential for usage in medical triage, but its current level of accuracy and consistency does not justify its independent use in medical practice. Based on the findings, future research is recommended to focus on specialized training of the model using medical data and terminology to enhance its accuracy and reliability in the context of medical triage. It would be valuable to assess ChatGPT's accuracy on real case presentations, explore its applicability in the emergency department, and make comparisons between the triage results of nurses and ChatGPT.

References

1. Briggs JK, Aarne GV. Emergency Nursing: 5-Tier Triage Protocols. 2nd ed. New York: Springer; 2020.
2. Benoit JRA. ChatGPT for clinical vignette generation, revision and evaluation. MedRxiv. 2023.
3. Cetin SB, Eray O, Cebeci F, Coskun M, Gozkaya M. Factors affecting the accuracy of nurse triage in tertiary care emergency departments. Turk J Emerg Med. 2020;20(4):163-7. doi: 10.4103/2452-2473.297462.
4. Chen SS, Chen JC, Ng CJ, Chen PL, Lee PH, Chang WY. Factors that influence the accuracy of triage nurses' judgement in emergency departments. Emerg Med Journal. 2010;27:4515. https://doi.org/10.1136/emj.2008.059311
5. Jordi K, Grossmann F, Gaddis GM, Cignacco E, Denhaerynck K, Schwendimann R, et al. Nurses' accuracy and self-perceived ability using the Emergency Severity Index triage tool: a cross-sectional study in four Swiss hospitals. Scand J Trauma Resusc Emerg Med. 2015;23:62. https://doi.org/10.1186/s13049-015-0142-y
6. 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(3):156-61. https://doi.org/10.4103/tjem.tjem_79_23
7. Gebrael G, Sahu KK, Chigarira B, Tripathi N, Mathew Thomas V, Sayegh N, et al. Enhancing triage efficiency and accuracy in emergency rooms for patients with metastatic prostate cancer: a retrospective analysis of artificial intelligence-assisted triage using ChatGPT 4.0. Cancers (Basel). 2023;15(14):3717. https://doi.org/10.3390/cancers15143717
8. Fraser H, Crossland D, Bacher I, Ranney M, Madsen T, Hilliard R. Comparison of diagnostic and triage accuracy of ada health and WebMD symptom checkers, ChatGPT, and physicians for patients in an emergency department: clinical data analysis study. JMIR Mhealth Uhealth. 2023;11:e49995. https://doi.org/10.2196/49995

BUDUĆNOST TRIJAŽE: ANALIZA TRADICIONALNIH POSTUPAKA U USPOREDBI S CHATGPT-jem

Sažetak

Uvod. Trijaža je procjena stanja pacijenta u cilju utvrđivanja hitnosti liječenja. Trijažu obično provodi medicinska sestra, najčešće s pomoću protokola u pet razina.

Cilj. Provesti komparativnu analizu točnosti kategorizacije i dijagnostike ChatGPT-ja (robota za chat koji primjenjuje algoritme strojnog učenja) u medicinskoj trijaži, kao i dati preporuke na koji način umjetna inteligencija može poboljšati rad medicinskih djelatnika u trijaži pacijenata.

Metode. Literatura odabrana za usporedbu jest Emergency Nursing: 5-Tier Triage Protocols. Za istraživanje su odabrane najčešće dijagnoze zbog kojih se pacijenti javljaju u hitnu službu. Zatim su odabrane kategorije trijaže koje će se primijeniti te su formirani prikazi slučajeva. Ti su se slučajevi postavili ChatGPT-ju, nakon čega su se uspoređivali njegovi odgovori s literaturom.

Rezultati. ChatGPT točno razvrstava slučajeve trijaže u 43,33 %, s prosječnom razlikom kategorije od 0,7. Iako je negdje pogriješio za jednu ili dvije kategorije, postavio je dijagnoze na višu kategoriju u svrhu osiguravanja sigurnosti pacijenta.

Rasprava. Pregledom literature utvrđeno je da se pogreške događaju u čak 40 % odluka medicinskih sestara zbog raznih čimbenika poput neiskustva, brzine rada i velikog broja pacijenata, što bi mogla umanjiti dodatna pomoć umjetne inteligencije. Potrebno je uzeti u obzir čimbenike koje stroj ipak ne može preuzeti te da može biti samo pomoć, a ne i zamjena medicinskog djelatnika.

Zaključak. ChatGPT ima potencijal za primjenu u medicinskoj trijaži, ali uz poboljšanja u smislu specijalizirane obuke modela na medicinskim podacima i terminologiji kako bi se poboljšala točnost i pouzdanost modela.

Ključne riječi: ChatGPT, sestrinstvo, trijaža