Vol. 8, Issue 2, August 2026, pp 333-345
https://doi.org/10.36590/jika.v8i2.2042
http:salnesia.id/index.php/jika
jika@salnesia.id, p-ISSN: 2337-9847, e-ISSN: 2686-2883
Penerbit: Sarana Ilmu Indonesia (salnesia)

RESEARCH ARTICLE

Accuracy and Agreement of ChatGPT Based START Triage in the Emergency Department of Gunung Jati Hospital

Analisis Keakuratan Hasil Penelitian START Triase oleh AI ChatGPT di IGD RSUD Gunung Jati

Efa Dwina[1*], Agil Putra Tri Kartika[1], Rizaluddin Akbar[1], Widuri Azmi Atmaji[1]
[1]Nursing Study Program, Universitas Muhammadiyah Cirebon, Cirebon, Indonesia

Abstract

Triage is an essential process in emergency care services aimed at determining patient treatment priorities based on the severity of clinical conditions. This study aimed to analyze the accuracy and level of agreement of Artificial Intelligence (AI), particularly ChatGPT, in triage classification at the Emergency Department of RSUD Gunung Jati compared with medical personnel as the reference standard. This study employed a quantitative observational approach with a retrospective cross-sectional design. Data were obtained from 420 adult patient medical records collected between March and May 2025. Triage classification was conducted using the Simple Triage and Rapid Treatment (START) method by both medical personnel and AI ChatGPT. Data analysis was performed using a confusion matrix and diagnostic tests, including sensitivity, specificity, accuracy, positive predictive value (PPV), negative predictive value (NPV), as well as agreement analysis using Cohen's kappa and weighted Cohen's kappa. The results showed that AI ChatGPT achieved an overall accuracy of 72.1% (95% CI: 67.8–76.4). The Cohen's kappa value was 0.313, while the weighted Cohen's kappa values were 0.338 using linear weighting and 0.384 using quadratic weighting, indicating a fair level of agreement. These findings suggested that ChatGPT had the potential as a supportive system in triage classification; however, it could not yet fully replace the role of medical personnel and still required further development, particularly in cases involving higher levels of emergency severity.

Keywords: artificial intelligence, triage, patient treatment

Article history:

Submitted 19 April 2026
Accepted 16 May 2026
Published 04 August 2026

PUBLISHED BY :
Sarana Ilmu Indonesia (salnesia)
Address :
Jl. Dr. Ratulangi No. 75A, Baju Bodoa, Maros Baru, Kabupaten Maros, Sulawesi Selatan.
Email :
info@salnesia.id, jika@salnesia.id
Phone :
+62 85255155883

Abstrak

Triase merupakan proses penting dalam pelayanan kegawatdaruratan yang bertujuan menentukan prioritas penanganan pasien berdasarkan tingkat keparahan kondisi klinis. Penelitian ini bertujuan menganalisis tingkat keakuratan dan kesesuaian Artificial Intelligence (AI), khususnya ChatGPT, dalam klasifikasi triase di Instalasi Gawat Darurat RSUD Gunung Jati dibandingkan dengan tenaga medis sebagai baku acuan. Penelitian menggunakan pendekatan kuantitatif observasional dengan desain retrospektif *cross-sectional*. Data penelitian diperoleh dari 420 rekam medis pasien dewasa periode Maret–Mei 2025. Proses klasifikasi triase dilakukan menggunakan metode *Simple Triage and Rapid Treatment* (START) oleh tenaga medis dan AI ChatGPT. Analisis data dilakukan menggunakan confusion matrix serta uji diagnostik yang meliputi sensitivitas, spesifisitas, akurasi, *positive predictive value* (PPV), *negative predictive value* (NPV), dan uji kesesuaian menggunakan *Cohen's kappa* serta *weighted Cohen's kappa*. Hasil penelitian menunjukkan bahwa AI ChatGPT memiliki tingkat akurasi sebesar 72,1% (95% CI: 67,8–76,4). Nilai *Cohen's kappa* sebesar 0,313, sedangkan *weighted Cohen's kappa* sebesar 0,338 pada pembobotan linear dan 0,384 pada pembobotan kuadratik, yang menunjukkan tingkat kesesuaian kategori *fair*. Temuan tersebut menunjukkan bahwa ChatGPT memiliki potensi sebagai sistem pendukung dalam klasifikasi triase, namun belum dapat menggantikan peran tenaga medis secara penuh dan masih memerlukan pengembangan lebih lanjut, terutama pada kasus dengan tingkat kegawatan yang lebih tinggi.

Kata Kunci: kecerdasan buatan, triase, penanganan pasien

*Penulis Korespondensi:
Efa Dwina, email: efadwinaa@gmail.com

Highlight:

- AI ChatGPT showed adequate accuracy and could serve as a tool to support Simple Triage and Rapid Treatment (START) triage classification in the Emergency Department (ED) in a consistent and systematic way.
- The performance and agreement of AI ChatGPT with medical personnel's assessment remained in the fair-agreement category and showed limitations in recognizing high-severity cases.
- AI ChatGPT cannot yet replace medical personnel as the primary decision-makers in the ED, so its use is recommended only as a decision support tool that still requires medical verification and oversight.

INTRODUCTION

The Emergency Department (ED) is a hospital service unit that focuses on managing patients with emergency conditions who require prompt, appropriate, and timely medical intervention. Services in the ED require accurate clinical decision-making to determine treatment priorities according to the severity of each patient's condition (Widodo and Riani, 2026). One of the systems used to determine these priorities is triage, which is the process of categorizing patients according to the severity of their conditions in order to establish the appropriate priority of care (Pranoto and Wibowo, 2020). The Simple Triage and Rapid Treatment (START) method is one of

the widely used triage methods because it has a simple, rapid, and easy-to-implement procedure that can be applied effectively in emergency situations.

However, the implementation of triage remains susceptible to inaccuracies due to various factors, such as high patient volumes, limited resources, and differences in healthcare providers' experience and clinical judgment. These conditions may lead to undertriage or overtriage, both of which can affect the quality and effectiveness of emergency care services (Harrigan et al., 2023). Undertriage may result in delayed treatment for patients with severe conditions, potentially increasing the risk of adverse outcomes. In contrast, overtriage may increase the inefficient use of available resources and reduce the availability of services for patients who are more in need of immediate care (Huabbangyang et al., 2023).

Previous research has shown that an AI system for triage classification had a specificity of 88.2% in identifying non-emergency patients. However, a false-positive rate of 11.8% was still observed, indicating the occurrence of overtriage, in which non-emergency patients were classified as emergency patients. This condition may potentially lead to less efficient use of resources in emergency departments, although its clinical risk is considered lower than that of undertriage because patient safety remains relatively better protected. These findings indicate that the use of AI as a clinical decision-support system still requires further evaluation and careful implementation to ensure that its use does not compromise the quality and safety of emergency care (Kartika et al., 2025).

The development of Artificial Intelligence (AI), particularly language-based models such as ChatGPT, has created new opportunities for the use of technology to support clinical decision-making (Ayoub et al., 2023). ChatGPT has the ability to understand and process text-based information rapidly and systematically, making it potentially useful as an auxiliary tool for triage classification (Zuhairini et al., 2025). This capability may provide additional support in processing clinical information and generating triage classifications based on the information provided. Nevertheless, the use of AI in emergency care contexts requires further evaluation, particularly regarding its accuracy and agreement compared with healthcare providers' assessments as the reference standard.

Several studies have demonstrated that the use of AI in healthcare has the potential to improve the efficiency and accuracy of clinical decision-making. However, research on the use of ChatGPT for START triage classification remains limited, particularly studies that directly compare AI classification results with healthcare providers' assessments as the reference standard in the ED (Lee et al., 2025). To date, studies specifically evaluating the accuracy and agreement of ChatGPT AI in START triage classification compared with healthcare providers as the reference standard in the ED remain limited, particularly in healthcare service settings in Indonesia. In addition, similar studies conducted in healthcare facilities in Indonesia are still scarce. Based on these conditions, this study offers novelty by evaluating the accuracy and agreement of ChatGPT AI in START triage classification compared with healthcare providers in the ED of Gunung Jati Regional General Hospital, Cirebon City (Sandmann et al., 2024).

This study was conducted to evaluate the accuracy and agreement of START triage classification results generated by ChatGPT AI compared with healthcare providers' assessments in the ED of Gunung Jati Regional General Hospital, Cirebon City. The analysis was performed using a confusion matrix, diagnostic tests including sensitivity, specificity, accuracy, positive predictive value (PPV), and negative predictive value (NPV), as well as an agreement test using Cohen's kappa. These

analyses were used to provide a comprehensive assessment of the performance and agreement of AI-generated triage classifications with the reference-standard assessment. This study is expected to provide an objective overview of the potential of AI as a support system for triage classification while also assessing its limitations in supporting clinical decision-making in the ED.

METHODS

This study employed a quantitative approach with an analytical observational design and a retrospective study design. The research data were obtained from the medical records of patients who visited the Emergency Department (ED) of Gunung Jati Regional General Hospital, Cirebon City, during the period from March to May 2025. This study aimed to evaluate the accuracy and agreement of *Simple Triage and Rapid Treatment* (START) triage classifications generated by ChatGPT AI compared with healthcare providers' assessments as the reference standard.

The study sample was selected using a *simple random sampling* technique with the assistance of *Microsoft Excel*. A total of 420 medical records of adult patients aged ≥18 years who met the inclusion criteria were analyzed in this study. The inclusion criteria consisted of complete clinical data, including age, sex, chief complaint, medical history, level of consciousness, and vital signs, such as blood pressure, pulse rate, respiratory rate, body temperature, and oxygen saturation. The exclusion criteria were medical records with incomplete clinical data, unclear or illegible information, or information that did not meet the requirements of the study.

Eligible medical record data were entered into ChatGPT AI, GPT-4 version, to generate triage classifications based on the *Simple Triage and Rapid Treatment* (START) method and were subsequently compared with the triage classifications made by healthcare providers in the ED. The reference standard (*ground truth*) in this study was determined based on the triage classification performed by ED triage nurses and documented in the patients' medical records. Clinical data were entered into the AI using a structured and standardized format to minimize interpretation bias during the classification process. The prompt contained clinical information including age, sex, chief complaint, level of consciousness, vital signs, and history of comorbidities. The AI was then instructed to determine the triage classification according to the START method. Each patient's data were entered using the same prompt format to maintain consistency throughout the classification process.

Data analysis was performed using IBM SPSS Statistics, with additional calculations using a *confusion matrix*. Univariate analysis was used to describe the characteristics of the subjects in terms of frequency distributions and percentages. The comparison of triage classifications between healthcare providers and ChatGPT AI was presented using a cross-tabulation table (*cross-tabulation/confusion matrix*). The performance of the AI was evaluated using diagnostic tests, including sensitivity, specificity, accuracy, positive predictive value (PPV), and negative predictive value (NPV). The level of agreement between the two classification methods was analyzed using Cohen's kappa.

Cohen's kappa was used to assess the level of agreement between the START triage classifications generated by ChatGPT AI and the healthcare providers' assessments as the reference standard. This test was considered appropriate for categorical data because it measures the degree of agreement while accounting for the possibility of agreement occurring by chance (Hanegraaf et al., 2024).

This study obtained ethical approval from the Ethics Committee of Gunung Jati Regional General Hospital, Cirebon City, under approval number No.044/LAIKETIK/KEPPKRSGJ/VIII/2025. All patient data used in the study were kept confidential by anonymizing the subjects' identities and were used solely for research purposes. Access to the research data was restricted to the researchers to maintain the security and confidentiality of patient information. In addition, the retrospective design of this study may have limitations related to the completeness and quality of the available medical record data.

RESULTS AND DISCUSSION

Demographic data of patients admitted to the ED

Based on Table 1, most subjects were in the adult age group of 30–59 years (258 subjects, 61.4%), with a mean age of 45 ± 15.1 years and an age range of 18–94 years. Males predominated, at 222 (52.9%).

Table 1. Demographic data of patients admitted to the ED of RSUD Gunung Jati, Cirebon City, March–May 2025 (n = 420)

Parameter	Category	F (%)	Mean (±SD)	Min–Max
Age	Adolescent (13–19 years)	7 (1.7)	45 ± 15.1	18–94
	Young adult (20–29 years)	90 (21.4)		
	Adult (30–59 years)	258 (61.4)		
	Elderly (60 years)	65 (15.5)		
Vital signs	Systolic (mmHg)		123 ± 28.5	73–255
	Diastolic (mmHg)		79 ± 17.6	41–162
	Pulse (beats/min)		101 ± 23.9	39–193
	Respiration (breaths/min)		24 ± 10.2	12–46
	Temperature		36.8 ± 1.2	35.2–41
	SpO2 (%)		97 ± 6	82–100
	GCS score		15 ± 1.2	3–15
Sex	Male	222 (52.9)		
	Female	198 (47.1)		
	Comorbidity			
	Hypertension	43 (10.2)		
	Diabetes mellitus	46 (11)		
	Coronary heart disease	88 (21)		
	Chronic kidney failure	15 (3.6)		
	Respiratory infection	57 (13.6)		
	Stroke	3 (0.7)		
	Dyspepsia	168 (40)		
Chief complaint	Abdominal pain	50 (11.9)		
	Chest pain	125 (29.8)		
	Shortness of breath	69 (16.4)		
	Fever	77 (18.3)		
	Nausea or vomiting	38 (9.0)		
	Bleeding	45 (10.7)		
	Weakness	16 (3.8)		

Note: Secondary data obtained from the medical records of ED patients at RSUD Gunung Jati, Cirebon City, March–May 2025

For vital signs, Table 1 shows a mean systolic blood pressure of 123 ± 28.5 mmHg, diastolic blood pressure of 79 ± 17.6 mmHg, pulse rate of 101 ± 23.9 beats/min, respiratory rate of 24 ± 10.2 breaths/min, body temperature of 36.8 ± 1.2 °C, oxygen saturation of 97 ± 6.0%, and a GCS score of 15 ± 1.2. These findings describe the general clinical condition of the subjects at the time of their arrival at the ED. The most common comorbidity was dyspepsia (168 subjects, 40.0%), while the most common chief complaint was chest pain (125 subjects, 29.8%). The distribution of these clinical characteristics provides an overview of the varied conditions encountered among patients presenting to the ED.

The subjects' characteristics show that patients arriving at the ED had diverse clinical conditions in terms of age, vital signs, comorbidity, and chief complaint. This variety indicates that triage requires a fast, accurate, and consistent assessment to establish service priorities according to the severity of each patient's condition. The differences in clinical characteristics also require healthcare providers to consider multiple clinical findings when determining the appropriate triage category. Triage assessment is influenced not by a single clinical indicator but by a comprehensive interpretation of the combination of vital signs, level of consciousness, and comorbid conditions (NCHS, 2022). Therefore, a comprehensive assessment of the available clinical information is important to support appropriate prioritization of patients in the ED.

The most common chief complaints were chest pain, fever, and shortness of breath, which were generally related to cardiovascular conditions or respiratory disorders requiring prompt attention and management in the ED. These complaints may represent a range of clinical conditions with different levels of severity, making an accurate initial assessment essential for determining the appropriate priority of care. In addition, comorbidities such as coronary heart disease, diabetes mellitus, and respiratory infection can increase the complexity of clinical assessment because they may contribute to or worsen a patient's condition rapidly. The presence of these underlying conditions may therefore require healthcare providers to consider the patient's overall clinical status when determining triage priorities.

These findings demonstrate that accurate triage classification is important in emergency care to reduce the risk of delayed treatment or inappropriate prioritization. The use of Artificial Intelligence (AI) as a clinical decision-support system could therefore help improve the consistency of triage assessment by supporting the processing and interpretation of available clinical information, although its use still requires appropriate evaluation and oversight by medical personnel (Sax et al., 2025). The diversity of patient characteristics observed in this study further demonstrates that the triage system requires a consistent and systematic assessment process to support fast and accurate clinical decision-making in the ED. Such consistency is particularly important in emergency settings, where clinical decisions must be made promptly while considering multiple patient characteristics simultaneously.

Comparison of START triage categories between the Hospital and AI ChatGPT

Table 2 shows the distribution of START triage categories by the Hospital and AI ChatGPT. In both assessments, the most dominant category was yellow. However, AI ChatGPT tended to classify more patients into the red and green categories than the Hospital did. This indicates a difference in the distribution pattern of triage categories between the two methods.

Table 2. Distribution of START triage categories by the Hospital and AI ChatGPT

Category	Hospital		AI ChatGPT	
	n	%	n	%
Green	7	1.7	33	7.9
Yellow	353	84.0	278	66.2
Red	60	14.3	109	26.0
Black	0	0	0	0
Total	**420**	**100.0**	**420**	**100.0**

Note: n = number of subjects; % = percentage of the total 420 subjects; category distribution presented by Hospital and AI ChatGPT assessment

Based on Table 2, the most dominant START triage category in the Hospital assessment was the yellow category, comprising 353 subjects (84.0%), followed by the red category with 60 subjects (14.3%) and the green category with 7 subjects (1.7%). A similar pattern was also observed in the AI ChatGPT assessment, in which the yellow category remained the most frequent category, accounting for 278 subjects (66.2%), followed by the red category with 109 subjects (26.0%) and the green category with 33 subjects (7.9%). No patients were classified into the black category by either assessment method.

The dominance of the yellow category in both the Hospital and AI ChatGPT assessments indicates that most patients in this study were at a level of severity requiring medical management but had not yet reached the most critical condition. In ED services, triage serves to identify the severity of patients' conditions, determine the priority of care, and support more effective utilization of available resources (Uly et al., 2025). Therefore, the distribution of triage categories presented in Table 2 can provide an initial overview of the pattern of patient severity before further analyses are conducted to assess the accuracy and agreement of the classifications.

Although the yellow category was the most frequent category in both assessment methods, AI ChatGPT showed higher proportions of red and green categories compared with the Hospital assessment. This difference indicates the possibility of classification shifts between the AI-generated results and the assessments made by healthcare providers. In the context of triage, such classification shifts require attention because they may lead to undertriage or overtriage. Undertriage may result in patients with serious conditions not receiving immediate treatment priority, whereas overtriage may lead to less efficient utilization of healthcare resources. Therefore, these distributional findings need to be followed by cross-tabulation analysis, diagnostic tests, and agreement testing to assess the accuracy and consistency of AI ChatGPT classifications against the Hospital reference standard.

This explanation is supported by Margaret (2023), who associated mistriage with definitions of undertriage and overtriage based on interventions and resource requirements. A study by Sari and Fajarini (2022) also showed that mistriage may occur in the form of undertriage or overtriage. Furthermore, Sax et al. (2025) discussed factors associated with undertriage and overtriage among trauma patients.

Table 3. Cross-tabulation of START triage categories between the Hospital and AI ChatGPT

Hospital triage	AI ChatGPT			
	Green	**Yellow**	**Red**	**Total**
Green	5	2	0	7
Yellow	28	257	68	353
Red	0	19	41	60
Total	**33**	**278**	**109**	**420**

Note: table presented as a confusion matrix

Based on Table 3, the highest level of agreement in START triage classification between the Hospital and AI ChatGPT was found in the yellow category, with 257 cases, followed by the red category with 41 cases and the green category with 5 cases. Classification discrepancies were mainly observed among patients who were categorized as yellow by the Hospital but were classified as green by AI ChatGPT in 28 cases and as red in 68 cases. In addition, there were 19 cases that were categorized as red by the Hospital but were assessed as yellow by AI ChatGPT. No cases were classified into the black category by either assessment method.

Table 3 shows that the highest level of agreement occurred in the yellow category. This means that most patients assessed as yellow by the Hospital were also assessed as yellow by AI ChatGPT. However, several classification differences were still observed, particularly among patients who were categorized as yellow by the Hospital but were assessed as either green or red by the AI. In addition, some patients who were categorized as red by the Hospital were assessed as yellow by the AI. These differences indicate that AI ChatGPT was not yet fully consistent in determining the appropriate triage category in several cases.

The cross-tabulation or confusion matrix in this study was used to describe the patterns of agreement and disagreement in START triage classification between the Hospital assessment and AI ChatGPT (Rainio, 2024). Through this table, it is possible to identify the categories that were most frequently classified similarly by the two methods, as well as the categories in which classification shifts occurred.

In the evaluation of classification models, the *confusion matrix* plays an important role because it can display the distribution of prediction results for each class and serve as a basis for assessing classification performance in greater detail (Hicks et al., 2022). Therefore, Table 3 not only presents the number of cases with matching and non-matching classifications but also provides an initial basis before diagnostic tests and agreement analyses are performed. This is consistent with Rainio (2024), who explained that the *confusion matrix* is used in the evaluation of both binary and multiclass classification to assess model performance in a structured manner.

In emergency care services, differences in classification such as these are important to consider because they may affect the priority of patient management. If a patient who should be classified into a more severe category is assessed as having a lower level of severity, *undertriage* may occur, potentially resulting in delayed treatment (Sax et al., 2025). Conversely, if a patient who is actually less critically ill is assessed as having a more severe condition, *overtriage* may occur, which can lead to less efficient utilization of ED resources (Sax et al., 2025). Therefore, the results of this cross-tabulation need to be followed by diagnostic tests and a kappa test to provide a

clearer assessment of the accuracy and agreement of *AI ChatGPT* compared with the Hospital assessment.

Table 4. Diagnostic test parameters of AI ChatGPT against the Hospital triage assessment

Category	Sensitivity	Specificity	Accuracy	PPV	NPV	False Negative	False Positive
Green	71.4	93.2	92.9	15.2	99.5	28.6	6.8
Yellow	72.8	68.7	72.1	92.4	32.4	27.2	31.3
Red	68.3	81.1	79.3	37.6	93.9	31.7	18.9

Note: sensitivity, specificity, accuracy, PPV, and NPV were obtained using the Hospital triage result as the reference standard

Based on Table 4, the diagnostic test results show that AI ChatGPT's performance in START triage classification differed by category. For green, AI ChatGPT showed sensitivity 71.4%, specificity 93.2%, accuracy 92.9%, PPV 15.2%, NPV 99.5%, false negative 28.6%, and false positive 6.8%. For yellow, sensitivity was 72.8%, specificity 68.7%, accuracy 72.1%, PPV 92.4%, NPV 32.4%, false negative 27.2%, and false positive 31.3%. For red, sensitivity was 68.3%, specificity 81.1%, accuracy 79.3%, PPV 37.6%, NPV 93.9%, false negative 31.7%, and false positive 18.9%. Overall, AI ChatGPT performed fairly well on several parameters, especially for yellow and green, but had limitations in detecting the red category. The black category could not be evaluated because no cases were found under either method.

For sensitivity, the AI's ability to recognize the red category was lower than for other categories. This warrants attention because red represents higher-severity patients, so failing to detect this category can imply delayed treatment priority (Sax et al., 2025).

This assessment focuses on AI ChatGPT's ability to recognize patients who truly belong to a given triage category according to the Hospital as the reference standard (Hoyer and Zapf, 2021). Here, sensitivity is important because it relates to the AI's ability to detect patients who should receive priority according to their severity. Sensitivity is a key measure in diagnostic-accuracy studies because it reflects a method's ability to recognize positive cases against the reference standard.

For specificity, the discussion focuses on AI ChatGPT's ability to identify patients who do not belong to a given triage category. Specificity matters because it shows how well the AI avoids assigning an inappropriate category label (Hoyer and Zapf, 2021).

In this study, the higher specificity for green shows that the AI was relatively good at distinguishing patients who were not green. However, the lower specificity for yellow shows that the AI still tended to classify some patients as yellow even when, by the reference standard, they did not belong there. Specificity should therefore be read together with sensitivity, so that AI performance is judged not only by its ability to recognize a category but also by its ability to rule out an incorrect one (Lim, 2021).

For accuracy, the discussion assesses the overall correctness of AI ChatGPT's classification compared with the Hospital as the reference standard. Accuracy is important because it gives a general picture of the proportion of the AI's classifications that were correct across all study data (Hicks et al., 2022). Unlike sensitivity and specificity, which only assess the ability to recognize or rule out a given category, accuracy shows overall performance (Rainio, 2024). In this study, however, accuracy must be interpreted cautiously because the triage-category distribution was imbalanced

and dominated by yellow. Accuracy therefore cannot be the sole basis for judgment and should be read together with other parameters such as sensitivity, specificity, PPV, NPV, false negatives, and false positives, so that AI ChatGPT's performance can be understood more comprehensively.

For positive predictive value (PPV) and negative predictive value (NPV), the assessment focuses on the confidence in AI ChatGPT's classifications. PPV indicates the probability that a patient the AI classifies into a category truly matches the Hospital's assessment, while NPV indicates the probability that a patient deemed not in a category truly is not (Hoyer and Zapf, 2021).

In this study, the higher PPV for yellow shows that the AI's classification in that category was relatively more trustworthy. Conversely, the lower PPV for green and red shows that positive results in those categories still need cautious interpretation (Hicks et al., 2022). The better NPV for green and red shows that the AI's decision to rule out those categories was relatively more consistent, while the low NPV for yellow may reflect that category's dominance in the data. PPV and NPV should therefore be read together with other diagnostic parameters for a more thorough understanding of AI ChatGPT's performance (Hoyer and Zapf, 2021).

For false negatives and false positives, the discussion focuses on the classification errors still found in AI ChatGPT's triage results (Chen et al., 2024). A false negative describes a patient who should be in a given category according to the Hospital but is not recognized by the AI, while a false positive describes the AI placing a patient in a category when, by the reference standard, the patient does not belong there (Hicks et al., 2022).

In this study, false negatives in the red category are a concern because they can lead to undertriage when a higher-severity patient is rated milder than warranted. Conversely, false positives in the yellow category can indicate overtriage patients classified at a priority level not matching their actual condition (Chen et al., 2024). In the ED, undertriage risks delayed care, while overtriage can lead to less efficient resource use. Both parameters are therefore important for assessing the safety and accuracy of AI ChatGPT as a triage aid, especially in high-severity categories (Huabbangyang et al., 2023).

Table 5. Agreement test between the Hospital triage and AI ChatGPT

Agreement test	Value	*p-value*	Interpretation
Cohen's Kappa	0.313	<0.001	Fair

Note: the Cohen's kappa value was obtained from SPSS analysis

Based on Table 5, the agreement test using Cohen's kappa showed a value of 0.313 with $p < 0.001$. This indicates statistically significant agreement between the triage classifications by medical personnel and AI ChatGPT. However, the kappa value in the fair-agreement category shows that the AI's agreement with medical personnel's assessment was not yet optimal.

Cohen's kappa was used because this study compared two categorical classifications: the triage results of medical personnel and AI ChatGPT. This method is more appropriate than a simple agreement percentage because it accounts for agreement that could occur by chance (Rainio, 2024). This result aligns with the confusion matrix and diagnostic tests, which still showed classification shifts in some triage categories, especially between yellow and red.

These findings show that AI ChatGPT could be used as a tool in triage

classification, but its use still requires verification by medical personnel to reduce the risk of misclassification (Dettori and Norvell, 2020). This also agrees with earlier literature stating that a Cohen's kappa in the 0,21–0,40 range falls in the fair-agreement category (Li et al., 2023).

CONCLUSION

This study shows that AI ChatGPT has potential as a support system in Simple Triage and Rapid Treatment (START) triage classification in the Emergency Department, as it showed a certain level of accuracy and agreement with medical personnel's assessment. However, the consistency of the AI's classifications was not yet fully optimal, so its use cannot yet replace medical personnel as the primary decision-makers in determining patient severity. AI ChatGPT is therefore more appropriately used as a tool to support the triage process more consistently and systematically in emergency care. The findings indicate that further development and evaluation are still needed, particularly to improve classification accuracy in high-severity cases. Beyond its practical implications for supporting clinical decision-making in the ED, the results can inform health-care facilities in developing policies on using Artificial Intelligence (AI) as a support system in emergency care. Future research is encouraged to involve larger samples, more varied clinical cases, and application across different health-care settings to obtain a more comprehensive picture of AI's reliability in supporting ED triage.

ACKNOWLEDGMENTS

The authors thank Universitas Muhammadiyah Cirebon and RSUD Gunung Jati, Cirebon City, for the support, cooperation, and contributions provided throughout the research and the preparation of this manuscript.

REFERENCES

Ayoub, M., Ballout, A.A., Zayek, R.A., Ayoub, N.F., 2023. Mind + Machine: ChatGPT as a Basic Clinical Decisions Support Tool. Cureus Journal of Medical Science 15(8), 8–11. Https://Doi.Org/10.7759/Cureus.43690

Chen, Q., Qin, Y., Jin, Z., Zhao, X., He, J., Wu, C., 2024. Enhancing Performance of The National Field Triage Guidelines using Machine Learning: Development of A Prehospital Triage Model to Predict Severe Trauma. Journal of Medical Internet Research 26, 1–15. Https://Doi.Org/10.2196/58740

Dettori, J.R., Norvell, D.C., 2020. Kappa and Beyond: Is There Agreement? Global Spine Journal 10(4), 499–501. Https://Doi.Org/10.1177/2192568220911648

Hanegraaf, P., Wondimu, A., Mosselman, J.J., Jong, R.D., Abogunrin, S., Queiros, L., Lane, M., Postma, M.J., 2024. Reviewer Reliability of Human Literature Reviewing and Implications Assisted for The Introduction of Machine Systematic Reviews: A Mixed Methods Review. BMJ Open 14(3), 1–10. Https://Doi.Org/10.1136/Bmjopen-2023-076912

Harrigan, M.E., Boremski, P.A., Collier, B.R., Tegge, A.N., Gillen, J.R., 2023. Impact of Nonphysician, Technology-Guided Alert Level Selection on Rates of Appropriate Trauma Triage in The United States: A Before and After Study. Journal of Trauma and Injury 36(3), 231–241.

https://doi.org/10.20408/jti.2023.0020
Hicks, S.A., Strümke, I., Thambawita, V., Hammou, M., Riegler, M.A., Halvorsen, P., Parasa, S., 2022. On Evaluation Metrics for Medical Applications of Artificial Intelligence. Scientific Reports 12(5979), 1–9. Https://Doi.Org/10.1038/S41598-022-09954-8
Hoyer, A., Zapf, A., 2021. Studies for The Evaluation of Diagnostic Tests. Deutsches Arzteblatt 118, 555-560. Https://Doi.Org/10.3238/Arztebl.M2021.0224
Huabbangyang, T., Rojsaengroeng, R., Tiyawat, G., Silakoon, A., Vanichkulbodee, A., On, J.S., Buathong, S., 2023. Associated Factors of Under and Over-Triage Based on The Emergency Severity Index; A Retrospective Cross- Sectional Study. Archives of Academic Emergency Medicine 11(1), 1–11. https://doi.org/10.22037/aaem.v11i1.2076
Kartika, A.P.T., Akbar, R., Atmaji, W.A., Dwina, E., Gennie, A., Krisna, A.A., 2025. Triase Instalasi Gawat Daruat Berbasis Artificial Intelligence Berdasarkan Emergency Severity Index. Jurnal Ners 10(1), 2522–2529. https://journal.universitaspahlawan.ac.id/index.php/ners/article/view/54523
Lee, S., Jung, S., Park, J.H., Cho, H., Moon, S., Ahn, S., 2025. Performance of ChatGPT, Gemini and Deepseek for Non-Critical Triage Support Using Real-World Conversations in Emergency Department. BMC Emergency Medicine 25(1), 1-20. https://doi.org/10.1186/s12873-025-01337-2
Li, M., Gao, Q., Yu, T., 2023. Kappa Statistic Considerations in Evaluating Inter-Rater Reliability Between Two Raters: Which, When, and Context Matters. BMC Cancer 23(1), 1–5. https://doi.org/10.1186/s12885-023-11325-z
Lim, C., 2021. Methods for Evaluating The Accuracy of Diagnostic Tests. Cardiovascular Prevention and Pharmacotherapy 3(1), 15–20. https://e-jcpp.org/journal/view.php?doi=10.36011/cpp.2021.3.e2
Margaret, E., 2023. Evaluation of The Version 4 of the Emergency Severity Index in US Emergency Departments for The Rate of Mistriage. JAMA Netw Open 6(3), 1-19. Https://Doi.Org/10.1001/Jamanetworkopen.2023.3404
[NCHS] National Center for Health Statistics., 2022. National Hospital Ambulatory Medical Care Survey: 2022 Emergency Department Summary Tables. National Center for Health Statistics, Washington.
Pranoto, Y.A., Wibowo, S.A., 2020. Aplikasi Desktop Sistem Triase untuk Pendukung Prioritas Tingkat Kegawatan. Jurnal Ilmu Komputer dan Teknok Informatika 3(1), 1–6. https://ejournal.itn.ac.id/index.php/mnemonic/article/view/2319
Rainio, O., Teuho, J., Klen, R., 2024. Evaluation Metrics and Statistical Tests for Machine Learning. Scientific Reports 14(6086), 1–14. Https://Doi.Org/10.1038/S41598-024-56706-X
Sandmann, S., Riepenhausen, S., Plagwitz, L., Varghese, J., 2024. Systematic Analysis of ChatGPT, Google Search and Llama 2 for Clinical Decision Support Tasks. Nature Communications 15, 1–8. Https://Doi.Org/10.1038/S41467-024-46411-8
Sari, S.R., Fajarini, M., 2022. The Emergency Severity Indeks (ESI) Usage: Triage Accuracy and Causes Of Mistriage. Jurnal Ilmu Kesehatan 7(S1), 243–248. Https://Doi.Org/10.30604/Jika.V7is1.1190
Sax, D.R., Warton, E.M., Mark, D.G., Reed, M.E., 2025. Emergency Department Triage Accuracy and Delays in Care for High-Risk Conditions. JAMA Network Open 8(5), 1–12. Https://Doi.Org/10.1001/Jamanetworkopen.2025.8498
Uly, R.G.Z., Sriyono, S., Qona'ah, A., 2025. Triage in Emergency Management in the Emergency Departement: A Systematic Review. Indonesian Journal of Global

Health Research 7(2), 527–534. https://jurnal.globalhealthsciencegroup.com/index.php/IJGHR/article/view/5607

Widodo, S,P., Riani, D.A., 2026. Literatur Desain Instalasi Gawat Darurat (IGD) pada Rumah Sakit di Bekasi. Jurnal Riset Ilmiah 3(2), 513–522. https://manggalajournal.org/index.php/SINERGI/article/view/2328

Zuhairini, R., Natasyah, W., Nurfadillah, W., Putri, I.F., Sapikah, N., 2025. Pemanfaatan Artificial Intelligence dalam Pengambilan Keputusan Klinis dan Manajemen Pasien Gawat Darurat di Bidang Anestesiologi: Sebuah Scoping Review. Jurnal Siti Rufaidah 3(4), 310-326. https://journal.ppniunimman.org/index.php/JASIRA/article/view/271