Original Research

DIGITAL HEALTH

DIGITAL HEALTH
Volume 10: 1–9

DOI: 10.1177/20552076241227132
journals.sagepub.com/home/dhj

Sage

Reliability of ChatGPT for performing triage task in the emergency department using the Korean Triage and Acuity Scale

Jae Hyuk Kim[1], Sun Kyung Kim[2,3], Jongmyung Choi[4] and Youngho Lee[4]

Abstract

Background: Artificial intelligence (AI) technology can enable more efficient decision-making in healthcare settings. There is a growing interest in improving the speed and accuracy of AI systems in providing responses for given tasks in healthcare settings.

Objective: This study aimed to assess the reliability of ChatGPT in determining emergency department (ED) triage accuracy using the Korean Triage and Acuity Scale (KTAS).

Methods: Two hundred and two virtual patient cases were built. The gold standard triage classification for each case was established by an experienced ED physician. Three other human raters (ED paramedics) were involved and rated the virtual cases individually. The virtual cases were also rated by two different versions of the chat generative pre-trained transformer (ChatGPT, 3.5 and 4.0). Inter-rater reliability was examined using Fleiss' kappa and intra-class correlation coefficient (ICC).

Results: The kappa values for the agreement between the four human raters and ChatGPTs were .523 (version 4.0) and .320 (version 3.5). Of the five levels, the performance was poor when rating patients at levels 1 and 5, as well as case scenarios with additional text descriptions. There were differences in the accuracy of the different versions of GPTs. The ICC between version 3.5 and the gold standard was .520, and that between version 4.0 and the gold standard was .802.

Conclusions: A substantial level of inter-rater reliability was revealed when GPTs were used as KTAS raters. The current study showed the potential of using GPT in emergency healthcare settings. Considering the shortage of experienced manpower, this AI method may help improve triaging accuracy.

Keywords

Artificial intelligence, KTAS, classification, emergency department, ChatGPT, inter-rater reliability

Submission date: 14 August 2023; Acceptance date: 28 December 2023

Introduction

Triage systems are commonly used in emergency departments (EDs) across the world. The Manchester Triage System (MTS), the Emergency Severity Index (ESI), and the Canadian Triage Acuity Scale (CTAS) are some of the examples of triage systems.[1–4] The triage algorithm can be used to prioritize and sort patients to determine acceptable medical waiting time and optimize the efficiency of emergency care.[5] Triaging entails the classification of patients according to background information and the presenting symptoms, enabling timely identification of patients who require urgent care.[5]

[1]Department of Emergency Medicine, Mokpo Hankook Hospital, Jeonnam, South Korea
[2]Department of Nursing, Mokpo National University, Jeonnam, South Korea
[3]Department of Biomedicine, Health & Life Convergence Sciences, Biomedical and Healthcare Research Institute, Jeonnam, South Korea
[4]Department of Computer Engineering, Mokpo National University, Jeonnam, South Korea

Corresponding authors:
Sun Kyung Kim, Department of Nursing and Department of Biomedicine, Health & Life Convergence Sciences, BK21 Four, Mokpo National University, 1666 Yeongsan-ro, Cheonggye-myeon, Muan-gun, Jeonnam 58554, South Korea.
Email: skkim@mnu.ac.kr

Jongmyung Choi, Department of Computer Engineering, Mokpo National University, 1666 Yeongdan-ro, Cheonggye-myeon, Muan-gun, Jeonnam 58554, South Korea.
Email:jmchoi@mnu.ac.kr

In Korea, the Korean Triage and Acuity Scale (KTAS) is universally used and applied to all patients admitted to EDs.[6] Patient overload and shortage of healthcare personnel and patient beds is a known challenge in the EDs across Korea. The KTAS was developed by the Ministry of Health and Welfare in 2012 to address this issue.[7] The KTAS has been implemented in emergency centers in Korea since 2016.[6] These facilities catered to ~7.2 million patients in 2022, corresponding to a patient–bed ratio of 1009.[8] The regional emergency medical centers at tertiary-level hospitals, catering to more critical patients, had a higher proportion of level 1 (requiring resuscitation) and level 2 (emergent) cases, and the shortage of workforce in these settings hampers the quality of care.[9]

The KTAS is a symptom-based classification tool used for assessing patients. It involves evaluating the initial impression given by the patient, conducting basic interviews and examinations, and considering the presenting symptoms. The primary considerations that apply to common symptoms and secondary considerations specific to particular symptoms are used to determine the severity and urgency of the patient's condition.6 However, general triage systems are not sufficiently specific for every illness,[3,4] and the diverse back ground of the personnel conducting triaging is another concern. The complexity of symptoms as well as the capacity, knowledge, and experience levels of classifiers may lead to discrepancies in the outcomes of triage.[10]

The recommended time for the initial medical examination by an emergency physician is determined on five levels, ranging from immediate to within 120 min in the following order: (a) age (15 years as the age criterion for distinguishing between adults and pediatrics), (b) main symptom, (c) consciousness and vital signs, (d) pain pattern, and (e) other factors, e.g., pregnancy and psychological state. The system utilizes a total of 155 symptoms for adults and 165 symptoms for pediatric cases, classified into 17 common groups. The selection of symptoms and consideration of first and second criteria are designed such that the first consideration criteria include consciousness, hemodynamic status based on vital signs, degree of respiratory distress, fever, pain, presence of bleeding disorders, and the mechanism of injury—applicable uniformly across most symptoms. The second consideration criteria are applicable to specific symptoms. For instance, if "sudden changes in vision" is chosen as the primary symptom, it is further classified as level 2 in the second consideration criteria if there is "sudden change in vision." It is important to note that "sudden changes in vision" is a consideration limited to certain eye-related symptoms.[6]

Those with the highest priority will receive treatment first, while patients with lower classifications might wait for treatment.[11] Given the importance of accurate triage classifications, there is a need to strengthen the ability and acuity of classifiers. Triage systems have started to incorporate artificial intelligence (AI) technology to support the clinical decision-making process. A previous systematic review of various clinical decision-making support systems for triage classifications found improvements in decision-making, leading to better patient outcomes.[12]

Various techniques, from logistic regression to neural networks, have utilized machine learning approaches to improve the accuracy of patient prioritization.[13] Accurate remote triage classifications using wearable devices are expected to substitute for manpower.[13] Recent studies have demonstrated the potential of chat generative pre-trained transformer (ChatGPT)[14] in assisting informed decision-making by health providers.[15–17] ChatGPT analyzes various written materials fed into neural networks.[18,19] This AI technology understands and responds to conversational input in a human-like manner.

Developed by OpenAI, ChatGPT is a large language model (LLM).[14] GPT3.5 utilizes ~175 billion parameters and is trained on about 753.4 GB of data from various sources, including the World Wide Web, books, and Wikipedia. ChatGPT uses the GPT3.5 model, which is a version of GPT3 enhanced by reinforced learning with human feedback (RLHF) to improve response accuracy and expression. ChatGPT4 employs one trillion parameters and can comprehend and construct longer and more intricate sentences. The ability of GPT4 to retain more contextual data in its memory makes it more intelligent.[14]

ChatGPT not only gathers data from every source available, but it also establishes connections between various pieces of data.[11] Given its predictive power, previous studies have examined its performance in medical licensing exams[20] and the use of a data set for medical questions.[21] The present study was conducted to examine the reliability of ChatGPT ratings used to establish KTAS levels. The accuracy of triaging by ChatGPT was determined by assessing inter-rater reliability with human raters. A substantial agreement would indicate the potential of generalizing the use of ChatGPT in ED and healthcare services.

Methods

Design

This was an inter-rater reliability study performed using written triage case scenarios. The individual scenarios were constructed virtually based on KTAS version 1.6 guidelines, incorporating the clinical experience of an emergency medicine specialist (the first author). The assessment consisted of 17 major items and detailed sub-categories. Each case focused on clinical significance with an intention to ensure complexity.

Instrument

Triage scale. KTAS is a modified version of CTAS which is adapted to the emergency medical settings in Korea.[7] It is a five-level triage scale that classifies patients as KTAS 1–5 (1, resuscitation; 2, emergent; 3, urgent; 4, less urgent; and 5, non-urgent). The priority of care is determined based on this classification.

Virtual patient scenarios. This study used 202 clinical case scenarios requiring triage assessment. Using the confidence interval (CI) approach, suggested by Rotondi and Donner,[22] following previous study,[23] the anticipated value of K was set at .66 with an upper bound of .56 and lower bound of .76. The minimum sample size for four raters was 81. Considering the comparative analysis of the characteristics of virtual patients, a total of 202 virtual patient scenarios (100 cases with text and 102 cases without text) were used for statistical analysis. Using the method described by Walter et al.,[24] and factoring a target intra-class correlation coefficient (ICC) of .80, this sample size was sufficient.

The scenario describes the demographic characteristics of patients (age, sex), chief complaints, vital signs, medical history, and other information (e.g., numerical rating scale (NRS) score for pain). Based on previous studies conducted in emergency centers of tertiary hospitals,[25–27] estimates of the prevalence of individual five triage levels were determined. The mean age of virtual patients was 56.04 years with adults accounting for 92.1% of patients and pediatric patients accounting for 7.9%. The most common reasons for admission were pain (23.3%), followed by trauma (10.9%) and dyspnea (7.9%) (Table 1).

Table 1. Demographic and clinical characteristics of virtual patients ($N = 202$).

Characteristics	Categories	N (%) or $M \pm SD$
Age (years)		56.04 ± 23.67
Division	Adult	186 (92.1%)
	Pediatric	16 (4.9%)
Gender	Male	111 (55.0)
	Female	91 (45.0)
Mental status	Alert	182 (90.1)
	Drowsy	10 (5.0)
	Stupor	7 (3.5)
	Semi coma	0 (0.0)
	Coma	3 (1.5)
Chief complaint	Pain	47 (23.3)
	Chest pain	9 (4.5)
	GI trouble	14 (6.9)
	Trauma	22 (10.9)
	Altered mentality	8 (4.0)
	Fever	13 (6.4)
	Dyspnea	16 (7.9)
	URI symptoms	8 (4.0)
	Others	65 (32.2)

Participants

All healthcare professionals currently working at ED and assuming triage role were eligible. Assessment of inter-rater reliability requires a minimum of two raters. The sample size for the number of raters was not guided by any previous study; yet, a previous study showed that increasing the number of raters improves the precision.[28] This study used a purposive sample of four staff members selected by the head of the ED from among staff who were willing to participate in this study. All participants had undergone training and were experienced in the use of KTAS before the study and had a working experience of at least 3 years in the ED.

KTAS rating

Scoring was conducted by four human raters (one emergency medicine specialist and three emergency medical technicians) and ChatGPT3.5 and 4.0. The four human raters independently conducted KTAS classifications based on clinical data on vital signs, Glasgow coma scale (GCS) scores, NRS scores, chief complaints, and other relevant parameters. All human raters were currently working at a regional emergency medical center in Korea.

KTAS rating by ChatGPT

To evaluate how accurately ChatGPT determined the condition of emergency patients, we asked GPT3.5 and GPT4 to rate the patient's urgency based on the KTAS classification. ChatGPT3.5 was asked to evaluate the KTAS classification using its application programming interfaces (APIs).[29] In this case, the model was GPT3.5-turbo, and the temperature

variable was set to 0. To evaluate ChatGPT4, we asked questions on the web using GPT PLUS, where we could not change the temperature variable. The questions asked of GPT3.5 and GPT4, including patient information and basic medical information, were all in Korean and were organized as:

> Please classify the severity of the Korean Triage and Acuity Scale (KTAS) scores of patients with the following symptoms and answer in the KTAS form, for example, sex: F, age: 26, main symptom: headache, consciousness: alert, GCS score: 15, vital signs: blood pressure 180/100-110-15-36.5, NRS score for pain: 3, and pregnancy at 34 weeks.

Statistical analysis

Inter-rater reliability was analyzed using Fleiss' kappa and the ICC. KTAS agreement was compared both as categorical variables and ordinal scores of urgent ratings. Fleiss' kappa was used to measure inter-rater reliability for categorical data KTAS classifications (1, resuscitation; 2, emergent; 3, urgent; 4, less urgent; and 5, non-urgent). The ICC was used to assess the degree of overall reliability between ChatGPT and gold standard (rater 1) as an ordinal variable (KTAS score 1–5). Fleiss' kappa values were interpreted according to Landis and Koch[30] as fair (.21–.40), moderate (.41–.60), substantial (.61–.80), and almost perfect (.81–1.00). An ICC below .50 was considered to indicate poor, >.50 and ≤.75 moderate, >.75 and ≤.90 good, and >.90 excellent correlations.[31] Additional calculations were conducted for each analysis according to whether the questions included text with descriptions of the patient's condition.

Results

The characteristics of the four human raters are presented in Table 2. Excluding the gold standard, the mean age and career experience of human raters were 26.7 and 3.7 years, respectively.

According to ratings by the gold standard, the largest proportion of patients was categorized in levels 3 (41.1%) and 2 (34.2%) (Table 3). Five percent of virtual patient scenarios were categorized as level 1 and one percent as level 5 (Table 4).

Overall, there was substantial agreement between human raters (kappa = .646, 95% CI = .610–.682) (Table 4). However, the value was lower when ChatGPT3.5 (kappa = .320, 95% CI = .294–.346) and 4.0 (kappa = .523, 95% CI = .496–.551) were added as raters. When the cases were divided according to the inclusion of a text description, the kappa values for cases without text were better between human raters when ChatGPT3.5 and 4.0 were added. The lowest agreement was when ChatGPT3.5 was added in triage level 1 cases with texts (kappa = .067, 95% CI = .006–.129).

Table 2. General characteristics of the human raters ($n = 4$).

	Gender	Occupation	Age (Years)	Career (Years)
Rater 1 (gold standard)	M	Emergency physician	47	17
Rater 2	M	Emergency medical technician	27	4
Rater 3	M	Emergency medical technician	26	4
Rater 4	M	Emergency medical technician	27	3

The ICC for the inter-rater reliability of triage levels between rater 1 and ChatGPT3.5 was classified as moderate. There was good inter-rater reliability between rater 1 and ChatGPT4.0. The values were higher when cases had text descriptions (Table 5).

Discussion

With technological advances, several attempts have been made to resolve the existing barriers preventing the optimal use of AI in healthcare. Timely treatment of patients is a key imperative in the ED. Therefore, developing a system to support the accurate and speedy classification of patients is of utmost importance. Triage is a tool to establish treatment priority by accurate grading of patients. Leveraging AI technology for patient triage can ensure objectivity. This study evaluated the inter-rater reliability for triage-level determinations between human raters and ChatGPT3.5 and 4.0. There was substantial inter-rater reliability among human raters for determining the level of urgency of cases. The inter-rater reliability of ChatGPT and human raters varied depending on the version of ChatGPT and the KTAS level.

Overall, the performance of ChatGPT4.0 in determining patient levels was much better than that of ChatGPT3.5 and was close to that of the gold standard and human raters. This result indicates the potential of incorporating this AI technology into clinical decision-making support systems. The results were consistent with those of a previous study in which the performance of ChatGPT4.0 far exceeded that of ChatGPT3.5, and ChatGPT4.0 exhibited a consistent

Table 3. Triage distribution by raters ($N = 202$).

Triage	Level 1	Level 2	Level 3	Level 4	Level 5
Gold standard (rater 1)	10 (5.0)	69 (34.2)	83 (41.1)	38 (18.8)	2 (1.0)
Rater 2	15 (7.4)	68 (33.74)	74 (36.6)	40 (19.8)	5 (2.5)
Rater 3	15 (7.4)	69 (34.2)	76 (37.6)	37 (18.3)	5 (2.5)
Rater 4	4 (2.0)	67 (33.2)	84 (41.6)	44 (21.8)	3 (1.5)
GPT3.5	100 (49.5)	97 (48.0)	5 (2.5)	0 (0.0)	0 (0.0)
GPT4	12 (5.9)	55 (27.2)	71 (35.1)	59 (29.2)	5 (2.5)

performance with a substantial understanding of complex clinical information.[32,33] However, another study revealed higher unsafe triage by ChatGPT4.0 when compared to other software (e.g., Ada or WebMD).[34] The authors suggested the need to conduct more rigorous and large sample assessments using real-world clinical data for validating ChatGPT prior to its application in clinical settings.

The findings of the Fleiss' kappa analysis indicated the lowest agreement at KTAS levels 1 and 5 among human raters, as well as when ChatGPT3.5 and 4.0 were used. The disagreement in the distribution of triage outcomes is due to over-triage tendency of ChatGPT. Given the importance of minimizing risk and over-utilization of resources, the accepted goals for under- and over-triage are suggested as <5% and between 25 and 30%, respectively.[35] In previous studies, misconceptions regarding symptoms and the selection of incorrect items were identified as the reasons for the disparity in KTAS grading between health workers.[36,37] ChatGPT can also make the same errors.[38] For these reasons, prior research suggested the need for supervised learning of ChatGPT, in which questions and the correct answers are created by humans.[39] This rigorous supervised learning can improve the accuracy of the answers. Inputting correct answers for a large number of real-world cases would improve the accuracy of AI ratings.

However, the question of how to construct the best prompt remains unresolved. In previous studies investigating the reliability of GPT4, repeating clear prompts improved its ability to generate a consistent rating, while inappropriate prompts reduced the consistency.[40] It is important to develop a comprehensive and clear prompt because the quality of its meaning and translation is tremendously affected by it.[41] A previous study identified varying performances depending on the prompt used to optimize LLMs.[42] In particular, the performance of ChatGPT4.0 improved when queries provided additional text-based information within the prompt.[42]

Even though ChatGPT4 has exhibited sufficient medical knowledge to pass medical examinations,[43,44] there are still concerns regarding its application in the field of medicine because it is a general AI model. Therefore, it is not yet good enough to rate KTAS scores correctly, and there is a need for it to learn more information related to emergency medicine. In the future, ChatGPT can be improved for emergency medical use via few-short learning[45] and fine-tuning.[46] It is also possible to develop a new LLM dedicated to emergency medicine similar to Impression ChatGPT.[47]

Future implementation

The current study demonstrated the possibility of using ChatGPT as a KTAS rater and its application in healthcare settings. However, the use of virtual scenarios was a limitation of this study and our findings may not be generalizable to real-world settings. Indeed, the application of ChatGPT is not free of challenges. Further, there was inconsistency in its rating depending on the complexity of the scenarios. For example, it was significantly worse at identifying patients who require resuscitation. In this context, prior studies have explored the approach to forming human–AI teams in medical settings.[48] Despite the increase in AI capabilities, the situational awareness provided by human raters is critical in medical settings. Extensive research on the informational needs of human raters and end-user research on AI technology is required prior to its application in clinical settings.

Processes using ChatGPT and other AI tools are anticipated to be slightly faster compared to human ratings, utilizing input data to achieve more accurate and objective assessments. However, concerns pertaining to the reliability of such a system remain. Thus, AI systems need to be adequately trained and tested to mitigate the potential for any negative public health impact. Identified problems, such as artificial hallucinations (a phenomenon of AI-generated information that does not correspond to any real-world input),[49] pose a significant risk for its applications in healthcare. Previous research has

Table 4. Fleiss' kappa values.

Fleiss' K (95% CI)	Between Four Human Raters			Human Raters and ChatGPT3.5			Human Raters and ChatGPT4.0		
	All Cases	Text	No Text	All Cases	Text	No Text	All Cases	Text	No Text
Overall	.646	.577	.702	.320	.272	.355	.523	.482	.546
	(.610–.682)	(.522–.631)	(.654–.750)	(.294–.346)	(.233–.310)	(.320–.391)	(.496–.551)	(.441–.524)	(.508–.583)
Level 1 (resuscitation)	.696	.488	.828	.182	.067	.294	.565	.322	.750
	(.639–.752)	(.409–.568)	(.748–.908)	(.138–.226)	(.006–.129)	(.232–.356)	(.522–.609)	(.261–.384)	(.688–.812)
Level 2 (emergent)	.710	.671	.743	.281	.282	.256	.600	.565	.610
	(.654–.767)	(.591–.750)	(.663–.823)	(.238–.325)	(.221–.343)	(.194–.318)	(.557–.644)	(.503–.626)	(.548–.672)
Level 3 (urgent)	.593	.539	.649	.359	.333	.386	.443	.440	.446
	(.537–.650)	(.459–.618)	(.569–.729)	(.316–403)	(.272–.394)	(.324–.448)	(.400–.487)	(.378–.501)	(.384–.508)
Level 4 (less urgent)	.616	.505	.685	.429	.358	.467	.592	.468	.559
	(.560–.673)	(.426–.584)	(.605–.765)	(.385–.473)	(.296–.419)	(.405–.529)	(.485–.573)	(.406–.529)	(.497–.621)
Level 5 (non-urgent)	.660	.330	.708	.492	.247	.526	.464	.247	.481
	(.604–.717)	(.251–.409)	(.628–.788)	(.449–.536)	(.186–.308)	(.464–.588)	(.421–.508)	(.186–.308)	(.419–.543)

Table 5. Intra-class correlation with the gold standard (rater 1).

ICC with Rater 1 (Gold Standard)	GPT3.5	95% CI	GPT4.0	95% CI
All cases ($n = 202$)	.520	.366-.636	.805	.743-.852
Cases with additional text ($n = 102$)	.516	.283-.673	.802	.707-.866
Cases without text ($n = 100$)	.511	.273-.671	.791	.690-.860

highlighted the need for monitoring and validation of ChatGPT-generated outcomes.[38] Thus, more research is required to demonstrate its reliability and ensure safe use in healthcare settings.

In a previous study, ChatGPT outperformed humans in making basic text classifications, demonstrating its potential for increasing efficiency. The use of algorithm-based deep learning was shown to improve the use of ChatGPT in emergency healthcare situations.[50] The limited exposure of under-trained health professionals and paramedics to KTAS rating and their relatively low patient encounter rates pose challenges to evaluating input data. Specifically, the implementation of KTAS grading (Pre-KTAS) in a pre-hospital setting in Korea is planned for December 2023. Leveraging AI assistance during data input for KTAS grading may yield more objective classification results than in emergency care settings, which often rely on under-trained human raters.

Conclusion

This study was the first attempt to explore the synchrony between human raters and GPTs using KTAS scores. Our results demonstrated the possibility of using ChatGPT as a KTAS rater and revealed some degree of reliability. Some factors affected its performance, including prompts and KTAS levels. However, the ChatGPT did not achieve sufficient agreement with other human raters. Therefore, further research on how ChatGPT learns can help improve the accuracy of the ratings. A precise design for the application of ChatGPT in emergency systems could be optimized in a pre-hospital setting where fast decisions by under-trained health providers are needed.

Consent statement: Not aplicable because the dataset is virtual patient scenarios.

Contributorship: J.H.K., J.C., and S.K.K. researched the literature and conceived the study. J.H.K. and S.K.K. were involved in protocol development, gaining ethical approval, patient recruitment, and data analysis. S.K.K. wrote the first draft of the manuscript. All authors reviewed and edited the manuscript and approved the final version of the manuscript.

Declaration of conflicting interests: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Ethical approval: Ethical approval was not applied in this study because no actual patient data was used.

Funding: The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by a grant (20025162) of the Regional Customized Disaster-Safety R&D Program funded by the Ministry of the Interior and Safety (MOIS, Korea) and by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (no. 2022R1A5A8033794).

Guarantor: SKK.

ORCID iD: Sun Kyung Kim https://orcid.org/0000-0001-8839-5577

Supplemental material: Supplemental material for this article is available online.

References

1. Huibers L, Giesen P, Wensing M, et al. Out-of-hours care in western countries: assessment of different organizational models. *BMC Health Serv Res* 2009; 9: 1–8.
2. Gratton RJ, Bazaracai N, Cameron I, et al. Acuity assessment in obstetrical triage. *J Obstet Gynaecol Can* 2016; 38: 125–133.
3. Zachariasse JM, Seiger N, Rood PP, et al. Validity of the Manchester Triage System in emergency care: a prospective observational study. *PLOS ONE* 2017; 12: e0170811.
4. Zachariasse JM, van der Hagen V, Seiger N, et al. Performance of triage systems in emergency care: a systematic review and meta-analysis. *BMJ Open* 2019; 9: e026471.
5. Van Ierland Y, Van Veen M, Huibers L, et al. Validity of telephone and physical triage in emergency care: The Netherlands Triage System. *Fam Pract* 2011; 28: 334–341.
6. Park J and Lim T. Korean Triage and Acuity Scale (KTAS). *J Korean Soc Emerg Med* 2017; 28: 547–551.
7. Lee KH, Cho SJ, Lee JE, et al. Study for standardization of Korean Triage and Acuity Scale. Ministry of Health and Welfare. 2012. 2012.
8. National Emergency Medical Center. 2022 Statistical Yearbook of Emergency Medical Services, Ministry of Health and Welfare. 2022.
9. Choi SK. The view of emergency medicine physician over the Korean emergency medical system; problems and improvements. *Publ Health Affairs* 2019; 3: 177–183.
10. Choi HJ, Kim HJ, Lee HJ, et al. Comparison with in-hospital Korean Triage and Acuity Scale (KTAS) and prehospital

triage system in a metropolitan city. *J Korean Soc Emerg Med* 2018; 29: 391–398.
11. Gilboy N, Tanabe P and Travers DA. The Emergency Severity Index version 4: changes to ESI level 1 and pediatric fever criteria. *J EmergNurs* 2005; 31: 357–362.
12. Fernandes M, Vieira SM, Leite F, et al. Clinical decision support systems for triage in the emergency department using intelligent systems: a review. *Artif Intell Med* 2020; 102: 101762.
13. Kim D, You S, So S, et al. A data-driven artificial intelligence model for remote triage in the prehospital environment. *PLOS ONE* 2018; 13: e0206006.
14. OpenAI. Introducing ChatGPT, https://openai.com/blog/chatgpt (accessed 19th May, 2023).
15. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. *Healthcare* 2023; 11: 887.
16. Biswas SS. Role of chat GPT in public health. *Ann Biomed Eng* 2023; 51: 868–869.
17. Javaid M, Haleem A and Singh RP. ChatGPT for healthcare services: an emerging stage for an innovative perspective. *BenchCouncil Trans Benchmarks Stand Evaluat* 2023; 3: 100105.
18. Di Giorgio AM and Ehrenfeld JM. Artificial intelligence in medicine &ChatGPT: de-tether the physician. *J Med Syst* 2023; 47: 32.
19. Johnson SB, King AJ, Warner EL, et al. Using ChatGPT to evaluate cancer myths and misconceptions: artificial intelligence and cancer information. *JNCI Cancer Spectr* 2023; 7: pkad015.
20. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLOS Digit Health* 2023; 2: e0000198.
21. Liévin V, Hother CE and Winther O. Can large language models reason about medical questions?. *arXiv preprint arXiv* 2022; 2207. 08143.
22. Rotondi MA and Donner A. A confidence interval approach to sample size estimation for interobserver agreement studies with multiple raters and outcomes. *J Clin Epidemiol* 2012; 65: 778–784.
23. Bullard MJ, Unger B, Spence J, et al. Revisions to the Canadian Emergency Department Triage and Acuity Scale (CTAS) adult guidelines. *CJEM* 2008; 10: 136–151.
24. van der Wulp I. Reliability and validity of emergency department triage systems. *Utrecht University 2010*.
25. Fayyaz J, Khursheed M, Mir MU, et al. Pattern of emergency department visits by elderly patients: study from a tertiary care hospital, Karachi. *BMC Geriatr* 2013; 13: 83.
26. Kim JH, Kim JW, Kim SY, et al. Validation of the Korean Triage and Acuity Scale compare to triage by Emergency Severity Index for emergency adult patient: preliminary study in a tertiary hospital emergency medical center. *J Korean Soc Emerg Med* 2016; 27: 436–441.
27. Lee IH, Kim OH, Kim CS, et al. Validity analysis of Korean Triage and Acuity Scale. *J Korean Soc Emerg Med* 2018; 29: 13–20.
28. Saito Y, Sozu T, Hamada C, et al. Effective number of subjects and number of raters for inter-rater reliability studies. *Stat Med* 2006; 25: 1547–1560.
29. OpenAI. API Referece, https://platform.openai.com/docs/api-reference (accessed 19th May. 2023).
30. Landis JR and Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977; 33: 159–174.
31. TK K and Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med* 2016; 15: 155–163.
32. Oh N, Choi GS and Lee WY. ChatGPT goes to the operating room: evaluating GPT-4 performance and its potential in surgical education and training in the era of large language models. *Ann Surg Treat Res* 2023; 104: 269–273.
33. Ueda D, Walston S, Matsumoto T, et al. Evaluating GPT-4-based ChatGPT's clinical potential on the NEJM quiz. *medRxiv* 2023; 2023: 05.
34. Fraser H, Crossland D, Bacher I, et al. 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; 3: e49995.
35. Hoyt DB and Schneidman DS. The American College of Surgeons: an enduring commitment to quality and patient care. *Am J Surg* 2015; 209: 436–441.
36. Kim HI, Oh SB and Choi HJ. Inter-rater agreement of Korean Triage and Acuity Scale between emergency physicians and nurses. *J Korean Soc Emerg Med* 2019; 30: 309–317.
37. Lupton JR, Davis-O'Reilly C, Jungbauer RM, et al. Under-triage and over-triage using the field triage guidelines for injured patients: a systematic review. *Prehosp Emerg Care* 2023; 27: 38–45.
38. Egli A. ChatGPT, GPT-4, and other large language models–the next revolution for clinical microbiology? *Clin Infect Dis*. Epub ahead of print 3 July 2023. DOI:10.1093/cid/ciad407
39. Chowdhury A, Rosenthal J, Waring J, et al. Applying self-supervised learning to medicine: review of the state of the art and medical implementations. *Informatics* 2021; 8: 59.
40. Hackl V, Müller AE, Granitzer M, et al. Is GPT-4 a reliable rater? Evaluating consistency in GPT-4 text ratings. *arXiv preprint arXiv* 2023; 2308. 02575.
41. Lyu Q, Tan J, Zapadka ME, et al. Translating radiology reports into plain language using ChatGPT and GPT-4 with prompt learning: promising results, limitations, and potential. *arXiv Preprint ArXiv* 2023; 2303: 09038.
42. Doshi R, Amin K, Khosla P, et al. Utilizing large language models to simplify radiology reports: a comparative analysis of ChatGPT3. 5, ChatGPT4. 0, Google Bard, and Microsoft Bing. *medRxiv* 2023; 2023: 06.
43. Nori H, King N, McKinney SM, et al. Capabilities of gpt-4 on medical challenge problems. *arXiv Preprint ArXiv* 2023; 2303: 13375.
44. Yanagita Y, Yokokawa D, Uchida S, et al. Accuracy of ChatGPT on medical questions in the national medical licensing examination in Japan: evaluation study. *JMIR Format Res* 2023; 7: e48023.
45. Dai H, Liu Z, Liao W, et al. AugGPT: leveraging ChatGPT for text data augmentation. *arXiv Preprint ArXiv* 2023; 2302: 13007.
46. OpenAI platform. Fine-tuning guides: learn how to customize a model for your application, https://platform.openai.com/docs/guides/fine-tuning (accessed 10August 2023).

47. Ma C, Wu Z, Wang J, et al. ImpressionGPT: an iterative optimizing framework for radiology report summarization with ChatGPT. *arXiv Preprint ArXiv* 2023; 2304: 08448.
48. Bansal G, Nushi B, Kamar E, et al. Beyond accuracy: the role of mental models in human-AI team performance. *Proc AAAI Conf Hum–Comput Crowdsour* 2019; 7: 2–11.
49. Alkaissi H and McFarlane SI. Artificial hallucinations in ChatGPT: implications in scientific writing. *Cureus* 2023; 15: e35179.
50. Gilardi F, Alizadeh M and Kubli M. ChatGPT outperforms crowd-workers for text-annotation tasks. *arXiv Preprint ArXiv* 2023; 2303: 15056.