Treatment of Acute Compartment Syndrome: Comparing Appropriate Use Criteria with Large Language Model Recommendations Utilizing Chat Generative Pre-Trained Transformer

Katrina Nietsch, Sarah Lu, Akiro H Duey[1], Bashar Zaidat, Nancy Shrestha, Laura Chelsea Mazudie Ndjonko, Pierce Joseph Ferriter, Jun Sup Kim[2], Samuel Kang-Wook Cho
[1]Icahn School of Medicine At Mount Sinai, [2]Mt Sinai School of Medicine Affl Hosps

INTRODUCTION:

Acute Compartment Syndrome (ACS) is a unique condition where prompt action can prevent ischemia and tissue necrosis, thus decreasing the risk of amputation. Given scenario-specific diagnostic markers for ACS, physicians can utilize the power of Chat Generative Pre-trained Transformer-4 (ChatGPT-4.0) to assist with making decisions on the appropriate treatment. These treatment methods include: conduct a fasciotomy, consider an alternate diagnosis, conduct frequent/serial observation, obtain/repeat serum biomarkers, and obtain/repeat pressure measurements. The purpose of this study was to evaluate the accuracy of ChatGPT by comparing its appropriateness scores for ACS treatments given various clinical scenarios.

METHODS:

The Major Extremity Trauma Research Consortium (METRC) and the American Academy of Orthopaedic Surgeons (AAOS) developed the Appropriate Use Criteria (AUC) for ACS to serve as an aid for physicians to best inform their clinical decision-making process when determining which treatments are most appropriate. The AUC, which is used to implement the AAOS Clinical Practice Guidelines, was determined to be the gold standard. The evidence-based indications for treatment included clinical symptoms compatible with ACS (symptoms, no applicable symptoms, or symptoms unknown, unreliable, or obtunded), perfusion pressure (<30 mmHg, >30 mmHg, or not obtained), and biomarkers (abnormal, normal, or unknown). A numerical scale was utilized to quantify the appropriateness of various treatments for ACS. An appropriate treatment is one for which the expected health benefits exceed the expected negative consequences by a sufficiently wide margin. For each set of indications, a score of 1-9 was assigned to each treatment method based on its appropriateness, as deemed by the METRC and AAOS panel using the modified Delphi method. A score from 7-9 signifies “Appropriate,” 4-6 signifies “May Be Appropriate,” and 1-3 signifies “Rarely Appropriate.” ChatGPT was prompted to assign a score for each treatment option based on the indications evaluated by the panel. To determine the error, the ChatGPT scores were subtracted from the AUC scores and the mean error, mean absolute error, and mean squared error were calculated. Pearson correlation and paired t-tests were used to determine statistical significance with alpha <.05.

RESULTS:

Twenty-seven indication variations, or patient scenarios, were evaluated among 5 different treatment options for a total of 135 paired scores. The mean error was 0 ± 1.4 for fasciotomy, -1.6 ± 1.1 for considering an alternate diagnosis, -0.8 ± 1.4 for frequent/serial observation, -4.2 ± 1.5 for obtaining/repeating serum biomarkers, and -0.7 ± 1.8 for obtaining/repeating pressure measurements. The mean absolute error was 1.0 ± 0.9 for fasciotomy, 1.7 ± 1.6 for considering an alternate diagnosis, 1.1 ± 1.1 for frequent/serial observation, 4.2 ± 4.1 for obtaining/repeating serum biomarkers, and 1.4 ± 1.3 for obtaining/repeating pressure measurements. The mean squared error was 2.0 ± 2.8 for fasciotomy, 3.8 ± 3.9 for considering an alternate diagnosis, 2.6 ± 4.2 for frequent/serial observation, 20.0 ± 13.5 for obtaining/repeating serum biomarkers, and 3.7 ± 5.1 for obtaining/repeating pressure measurements (Table 1). Pearson correlation testing found that there was a significant positive correlation between AAOS and ChatGPT scores for fasciotomy (.82, P<.001), considering an alternate diagnosis (.63, P<.001), frequent/serial observation (.46, P=.016), and obtaining/repeating pressure measurements (.50, P=.009). There was a nonsignificant weakly positive correlation between scores for obtaining/repeating serum biomarkers (.26, P=.191) (Table 2). Using a paired t-test, there were statistically significant differences between scores for considering an alternate diagnosis (P<.001), frequent/serial observation (P=.010), and obtaining/repeating serum measurements (P<.001) (Table 3).

DISCUSSION AND CONCLUSION:

The Appropriate Use Criteria is designed to minimize missed ACS diagnoses and unnecessary fasciotomy procedures. A delayed or failed diagnosis may result in systemic illness, limb amputation, and loss of function. The appropriateness scores for fasciotomy, considering an alternate diagnosis, frequent/serial observation, and obtaining/repeating pressure measurements determined by ChatGPT were weakly positively correlated with the AUC scores. ChatGPT underestimated the appropriateness of considering an alternate diagnosis, frequent/serial observation, obtaining/repeating serum biomarkers, and obtaining/repeating pressure measurements. Although there was no difference between fasciotomy scores, the scores to consider an alternate diagnosis, frequent/serial observation, and obtain/repeat serum measurements were non-equivalent. Although it demonstrated minimal capacity to evaluate treatment options for ACS based on varying indications, ChatGPT requires improvement. The potential for ChatGPT to make individualized, time-sensitive decisions surrounding ACS must be further evaluated.

Table 1. Mean error, mean absolute error, and mean squared error between AAOS and ChatGPT scores for Acute Compartment Syndrome treatment options.

Treatment	Mean Error	Mean Absolute Error	Mean Squared Error
Fasciotomy	0 (1.4)	1.0 (0.9)	2.0 (2.8)
Consider alternate diagnosis	-1.6 (1.1)	1.7 (1.6)	3.8 (3.9)
Frequent/serial observation	-0.8 (1.4)	1.1 (1.1)	2.6 (4.2)
Obtain/repeat serum biomarkers	-4.2 (1.5)	4.2 (4.1)	20.0 (13.5)
Obtain/repeat pressure measurements	-0.7 (1.8)	1.4 (1.3)	3.7 (5.1)

*Standard deviations provided in parentheses

Table 2. Pearson correlation coefficients and p-values comparing AAOS and ChatGPT scores for Acute Compartment Syndrome treatment options.

Treatment	Pearson's r	P-Value
Fasciotomy	0.82	**<.001**
Consider alternate diagnosis	0.63	**<.001**
Frequent/serial observation	0.46	**0.016**
Obtain/repeat serum biomarkers	0.26	0.191
Obtain/repeat pressure measurements	0.50	**0.009**

Table 3. Paired t-test results comparing AAOS and ChatGPT scores for Acute Compartment Syndrome treatment options.

Treatment	P-Value
Fasciotomy	1.000
Consider alternate diagnosis	**<.001**
Frequent/serial observation	**0.010**
Obtain/repeat serum biomarkers	**<.001**
Obtain/repeat pressure measurements	0.071