Performance of ChatGPT Versus Spine Surgeons as an Emergency Department Spine Call Consultant

Taha Mahmet Taka[1], Sarah Meng, Seena Sebt, Andrew John Cabrera, David Shin, Vahe Yacoubian, Weyjuin E Chao[2], Daniel J Rossie[2], Zhengle Xu[2], Melissa Maria Erickson, Brett Rocos, Khoi D Than, Elizabeth M Yu, Nicholas Utchan Ahn, Christopher M Bono, Wayne K Cheng, Olumide A Danisa
[1]Orthopaedic Surgery, [2]Emergency Medicine

INTRODUCTION: Large language models (LLMs) like ChatGPT are increasingly being recognized as credible tools for use across diverse healthcare settings. While artificial intelligence (AI) use has previously been evaluated in emergency medicine, its use in subspecialty care - particularly spine surgery - remains underexplored. This study evaluates the clinical accuracy, management appropriateness, completeness, helpfulness, and overall quality of ChatGPT responses compared to those of board-certified, spine surgeons in response to common emergency department (ED) consultations.

METHODS: A seven-part questionnaire was developed based on common ED spine consultations (e.g. Cauda Equina Syndrome, compression fracture in elderly patients, purulent drainage from surgical wound, acute lumbar disc herniation, incomplete spinal cord injury, epidural abscess, and metastatic spine disease). Each case included 3–4 questions pertaining to examination, diagnosis, management, and counseling. Responses from ChatGPT and seven board-certified spine surgeons were restricted to 3-4 sentences per question. Three emergency medicine physicians rated each de-identified questionnaire response using a 5-point Likert scale. Statistical analysis was conducted using a two-sample T-test with unequal variance. Inter-rater reliability was assessed using pairwise weighted Cohen's kappa coefficient (κ).

RESULTS: When comparing AI responses versus spine surgeon responses to proposed ED consultations, AI responses were rated to be superior across all five metrics of clinical accuracy, management appropriateness, completeness, helpfulness, and overall quality ($p < 0.05$). Inter-rater reliability was assessed using average pairwise weighted Cohen's kappa coefficient which showed substantial agreement ($\kappa = 0.76$).

DISCUSSION AND CONCLUSION:

ChatGPT responses to emergency department spine consultations were rated as significantly higher compared to board-certified spine surgeons by emergency medicine providers. Though further validation is warranted, these findings suggest that ChatGPT can be a useful clinical adjunct for spine-related emergency department consultations.