Evaluating the Diagnostic and Treatment Capabilities of Generative Artificial Intelligence in Cervical Spine Pathology: A Comparative Study with Fellowship-Trained Surgeons

George Abdelmalek[1], Harjot Singh Uppal, Neil Patel[1], Daniel Coban, Stuart Changoor, Nikhil Sahai[2], Kumar Gautam Sinha[2], Ki S Hwang[2], Arash Emami
[1]St. Joseph's University Medical Center, [2]University Spine Center

INTRODUCTION: Generative artificial intelligence (GAI) has demonstrated its usefulness in various healthcare applications, including patient education, imaging analysis, billing, and coding. GAI's ability to process and generate relevant outputs from large data sets is valuable. The expertise required for diagnosing and treating surgical spine pathology involves many years of training. There is increasing interest in GAI's potential to employ clinical reasoning for patient diagnosis; however, few studies have evaluated GAI's diagnostic and treatment capabilities in clinical settings. Our study compares the effectiveness of GAI, specifically ChatGPT 3.5 and 4.0, in diagnosing and selecting treatment modalities for spine conditions against the determinations made by two fellowship-trained spine surgeons.

METHODS: This comparative analysis involved patients presenting with cervical neck pain at a single institution from January 1, 2023, to December 31, 2023. Patients were excluded if they were following up after surgical intervention or an emergency department visit. The study analyzed 50 patients. Three-sentence clinical scenarios were created using information from each patient's electronic medical record (EMR). The first sentence detailed the patient's demographics, chief complaint, chronicity, associated symptoms, and factors that relieved or exacerbated their condition. The second sentence described the positive physical exam findings, and the third included radiographic and advanced imaging impressions. These scenarios were input into ChatGPT 3.5 and 4.0, which were asked to produce the three most likely diagnoses in descending order of likelihood. The software was then prompted to select each patient's most beneficial treatment option. These results were compared to two spine surgeons' primary diagnoses and treatment plans.

RESULTS: ChatGPT 3.5, ChatGPT 4.0, and the two spine surgeons identified 9, 5, and 7 unique diagnoses for the 50 patient scenarios, respectively. There was a fair, statistically significant interrater agreement on likely diagnoses between ChatGPT 3.5 and 4.0 (□=0.233, p=0.04). Both versions of ChatGPT demonstrated slight, statistically significant interrater agreement with the primary diagnoses made by the spine surgeons (ChatGPT 3.5: □=0.110, p=0.037; ChatGPT 4.0: □=0.148, p=0.042). However, ChatGPT 4.0 showed fair interrater agreement with the spine surgeons regarding treatment selection (□=0.289, p< 0.001), while ChatGPT 3.5 showed slight interrater agreement (□=0.104, p=0.023).

DISCUSSION AND CONCLUSION: ChatGPT 3.5 and 4.0 have a slight interrater agreement with spine surgeons regarding likely diagnoses. ChatGPT 4.0, however, demonstrated superior interrater agreement with spine surgeons regarding treatment selection compared to ChatGPT 3.5. These findings suggest the potential for GAI to serve as a supportive diagnostic tool for spine pathology.