Performance of an AI Large Language Model in Guiding Therapy in Coronary and Aortic Valve Disease

Brady Gunn DO, Maxwell F Kilcoyne DO, Zachary W. Sollie MD, Ahmed Alameldin, MS, John Del Gaizo, PhD, Roshan Mathi, BS, Brett Welch, MBA, Khaled Shorbaji, MD, Arman Kilic, MD
Medical University of South Carolina, Division of Cardiothoracic Surgery

Introduction

ChatGPT is a powerful AI large language model (LLM) whose role in healthcare decision-making remains unexplored. This study evaluated the performance of ChatGPT in recommending therapy for coronary artery disease (CAD) and aortic valve disease (AVD).

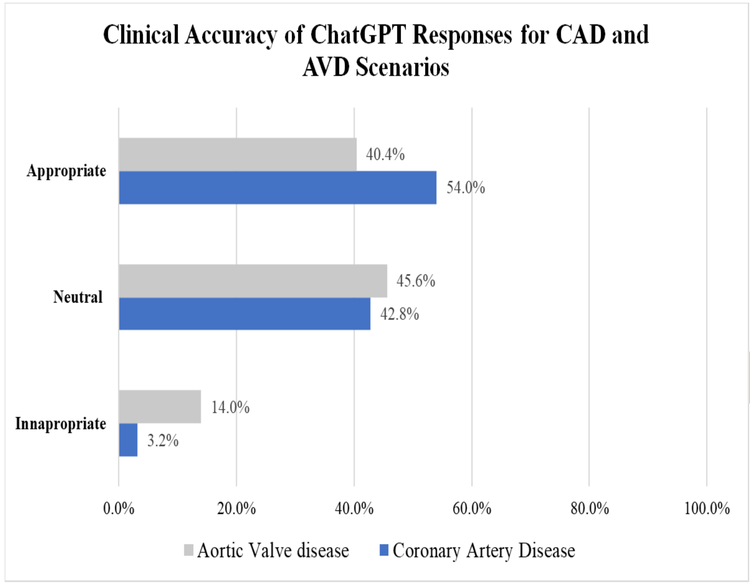
METHOD

- Clinicians created 100 common patient scenarios (CAD and AVD each n=50) with clear American College of Cardiology / American Heart Association (ACC/AHA) guideline-based therapy recommendations.
- Each scenario was presented individually to ChatGPT five times. The accuracy of responses was reviewed by clinical experts and graded individually and labeled as appropriate, neutral, or inappropriate based on the current ACC/AHA guidelines.
- Management pathways included coronary artery bypass grafting (CABG), percutaneous coronary intervention (PCI), or medical management. AVD management pathways included surgical aortic valve replacement with either mechanical or bioprosthetic prostheses, transcatheter aortic valve replacement, or continued medical management/surveillance.

RESULTS

- For CAD scenarios, the average age of the patient was 60.5 years old and the majority were male (71.7%, n=33). Approximately half described multi-vessel CAD (48.8%, n=20) and 14.6% (n=6) had left main disease. Regarding ChatGPT performance for CAD scenarios, 54.0% (n=135) were graded as appropriate, 42.8% (n=107) were graded as neutral, and 3.2% (n=8) were graded as inappropriate (Figure 1.). The average between-iteration concordance for each CAD scenario was 85.6% (SD=15.67) for ChatGPT, suggesting reasonable concordance between iterations. In the AVD scenarios, the average age was significantly older than the CAD scenarios (65.95 vs 60.5 years old, p=0.046). AV stenosis was the most common pathology described (62.0%, n=31), followed by AV insufficiency (34.0%, n=17) and mixed disease (4.0%, n=2). For AVD scenarios, 40.4% (n=101) were graded as appropriate, 45.6% (n=114) were graded as neutral, and 14.0% (n=35) were graded as inappropriate (Figure 2.) The average between-iteration concordance for each AVD scenario was 76.8% (SD=19.94) for ChatGPT, suggesting reasonable concordance as well between iterations. The between-iteration concordance was significantly better in CAD compared to AVD scenarios (t= 2.45, p = 0.016). The proportion of appropriate, neutral, and inappropriate responses by ChatGPT were improved in CAD vs AVD scenarios (p< 0.001).

	CAD Scenarios	AVD Scenarios	P-value
Age – mean (SD)	60.52 (14.82)	65.95 (11.87)	0.046
Male - % (n)	71.7 (33)	64.6% (31)	0.460
STS PROM included - % (n)	34.0 (17)	22% (11)	0.181
Diabetic - % (n)	20.0 (10)	-	-
Acute Coronary Syndrome - % (n)	26.0 (13)	-	-
Left main Disease - % (n)	14.6 (6)	-	-
Multi-vessel CAD	48.8 (20)	-	-
LVEF < 50 - % (n)	20.0 (10)	-	-
Bicuspid AV	-	20.0 (10)	-
AS Predominant	-	62.0 (31)	-
AI Predominant	-	34.0 (17)	-
Mixed AI/AS	-	4.0 (2)	-
AV Endocarditis	-	8.0 (4)	-



Conclusions

ChatGPT performed reasonably well in selecting guideline recommended therapies for both CAD and AVD with low rates of inappropriate suggestions. As this technology continues to evolve and improve, continual assessment of its accuracy will be essential to define its role in patient and provider education.

REFERENCES

1.Sarraju A, et al. Appropriateness of Cardiovascular Disease Prevention Recommendations Obtained From a Popular Online Chat-Based Artificial Intelligence Model. JAMA. 2023:329(10):842-844.

2. Gautam N, et al. Current and Future Applications of Artificial Intelligence in Coronary Artery Disease. Healthcare. 2022:10(2)