EuroIntervention
2024;20:e496-e503
published online e-edition April 2024
DOI: 10.4244/EIJ-D-23-00643

A study of ChatGPT in facilitating Heart Team decisions on severe aortic stenosis

Adil Salihu[1], MD; David Meier[1], MD; Nathalie Noirclerc[1], MD; Ioannis Skalidis[1], MD; Sarah Mauler-Wittwer[1], MD; Frédérique Recordon[1], BSc; Matthias Kirsch[2], MD, PhD; Christan Roguelov[1], MD; Alexandre Berger[1], MD; Xiaowu Sun[3], MSc, PhD; Emmanuel Abbe[3], MS, PhD; Carlo Marcucci[4], MD; Valentina Rancati[4], MD; Lorenzo Rosner[4], MD; Emmanuelle Scala[4], MD; David C. Rotzinger[5], MD, PhD; Marc Humbert[6], MD; Olivier Muller[1], MD, PhD; Henri Lu[1,7], MD; Stephane Fournier[1]*, MD, PhD

A. Salihu and D. Meier contributed equally to this paper.
**Corresponding author: Department of Cardiology, Lausanne University Hospital, Rue du Bugnon 46, 1011, Lausanne, Switzerland. E-mail: stephane.fournier@chuv.ch*

The authors' affiliations can be found at the end of this article.

ABSTRACT

BACKGROUND: Multidisciplinary Heart Teams (HTs) play a central role in the management of valvular heart diseases. However, the comprehensive evaluation of patients' data can be hindered by logistical challenges, which in turn may affect the care they receive.

AIMS: This study aimed to explore the ability of artificial intelligence (AI), particularly large language models (LLMs), to improve clinical decision-making and enhance the efficiency of HTs.

METHODS: Data from patients with severe aortic stenosis presented at HT meetings were retrospectively analysed. A standardised multiple-choice questionnaire, with 14 key variables, was processed by the OpenAI Chat Generative Pre-trained Transformer (GPT)-4. AI-generated decisions were then compared to those made by the HT.

RESULTS: This study included 150 patients, with ChatGPT agreeing with the HT's decisions 77% of the time. The agreement rate varied depending on treatment modality: 90% for transcatheter valve implantation, 65% for surgical valve replacement, and 65% for medical treatment.

CONCLUSIONS: The use of LLMs offers promising opportunities to improve the HT decision-making process. This study showed that ChatGPT's decisions were consistent with those of the HT in a large proportion of cases. This technology could serve as a failsafe, highlighting potential areas of discrepancy when its decisions diverge from those of the HT. Further research is necessary to solidify our understanding of how AI can be integrated to enhance the decision-making processes of HTs.

KEYWORDS: aortic stenosis; innovation; TAVI

SUBMITTED ON 07/08/2023 - REVISION RECEIVED ON 1st 03/11/2023 / 2nd 23/12/2023 / 3rd 31/01/2024 - ACCEPTED ON 01/02/2024

The Heart Team (HT) has become a cornerstone in the management of valvular heart diseases, ensuring a multidisciplinary approach to decision-making, optimising patient care, and ultimately leading to improved outcomes. Current guidelines mandate the inclusion of HTs in order to make tailored decisions regarding the treatment of patients with aortic stenosis (AS), including options such as conservative management, surgical intervention, or percutaneous treatment[1,2].

However, in practical implementation, coordinating and bringing together the diverse expertise of the HT can be challenging and can occasionally result in incomplete attendance at meetings. Moreover, the high volume of patients and time constraints may hinder the detailed evaluation of each patient's clinical data. These factors can potentially limit the effectiveness of HT meetings, resulting in suboptimal patient management.

To overcome these constraints and enhance the efficiency of HTs, this study explores the potential application of artificial intelligence (AI), specifically large language models (LLMs). LLMs offer a promising solution to enhance the clinical decision-making process by providing an initial comprehensive evaluation of patients' clinical data[3,4]. This technology can serve as a failsafe, drawing attention to potential discrepancies when its decision differs from that of the clinicians, particularly in situations where specific specialists are unavailable to provide their input.

Our aim was to assess the capacity of LLMs to effectively address complex clinical scenarios discussed during HT meetings, specifically in the context of AS management.

Editorial, see page e465

Methods

STUDY POPULATION

Data were retrospectively collected from the last consecutive 150 patients with AS presented at the HT meetings of a single Swiss university hospital. Patients provided informed consent by either signing the general consent form for research (Consentement général pour la recherche), the consent form of the SwissTAVI registry or both. Our study was conducted in accordance with the principles of the Declaration of Helsinki. Ethical approval was given by the Vaud Canton Ethics Committee (decision CER-VD 211/13, dated 10 May, 2013).

CLINICAL EVALUATION AND HEART TEAM MEETING

In the context of severe AS, each patient underwent a comprehensive evaluation, including a transthoracic echocardiogram, a carotid ultrasound, a computed tomography (CT)-scan and a coronary angiogram. In addition, a geriatric evaluation was conducted to assess potential frailty. The results from these evaluations were then presented to the HT, comprising interventional cardiologists, imaging specialists, cardiac surgeons, a vascular surgeon, an anaesthesiologist, and a geriatrician. The combined expertise of the team was used to determine the most appropriate management strategy.

Impact on daily practice

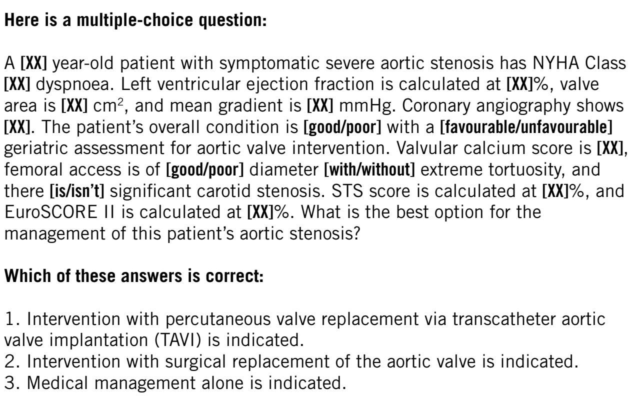
The use of large language models and particularly ChatGPT showed an interestingly high agreement rate of 77% with the decision of the Heart Team regarding the management of severe aortic stenosis. The integration of artificial intelligence (AI) in routine workflows can enhance efficiency by providing an initial comprehensive evaluation of patient data, allowing the Heart Team to focus on critical aspects of patient care and deliberations. We acknowledge that AI should not replace the healthcare provider's judgment, but rather serve as a valuable tool to support and expedite decision-making processes.

CLINICAL VIGNETTE PRESENTED TO CHATGPT

For each patient, a standardised clinical vignette was created with a total of 14 key variables derived from clinical evaluation, forming a standardised report. Treatment options were incorporated into a multiple-choice questionnaire **(Figure 1)**. This report was then submitted to a model devised by OpenAI, known as the Chat Generative Pre-trained Transformer (ChatGPT) version 4.0 (GPT-4), for analysis and processing. The set of 14 variables consisted of clinical data (patient's age, New York Heart Association [NYHA] Functional Class, and a geriatric assessment providing details about the patient's frailty condition and an overall evaluation of the assessment), echocardiographic data (left ventricular ejection fraction [LVEF], aortic valve area, and mean gradient across the aortic valve), cardiovascular assessment data (coronary angiography description, CT scan information related to aortic valve calcium score, femoral artery diameter and tortuosity, as well as carotid artery evaluation), and surgical risk scores (Society of Thoracic Surgeons [STS] and European System for Cardiac Operative Risk Evaluation [EuroSCORE] II). In order to standardise the clinical vignette, coronary lesions were categorised as being either significant (i.e., potentially needing revascularisation) or non-significant (i.e., with medical treatment being recommended). Likewise, in cases where transcatheter aortic valve implantation (TAVI) was assessed as a treatment option, a transfemoral approach was considered as either feasible or not feasible based on the diameter of the femoral arteries. Regarding the evaluation of carotid arteries, stenosis of <50% was reported as being non-significant. On the other hand, for any stenosis ≥50%, the exact degree of stenosis was documented. Finally, three possible treatment options were proposed for each patient: percutaneous intervention (TAVI), surgical aortic valve replacement (SAVR), or medical treatment. We then mirrored the question posed to the Heart Team in our query to ChatGPT, specifically asking, "What is the best management option for this patient?".

Abbreviations

AI	artificial intelligence	**HT**	Heart Team	**TAVI**	transcatheter aortic valve implantation
ChatGPT	Chat Generative Pre-trained Transformer	**LLM**	large language models		
		SAVR	surgical aortic valve replacement		



Here is a multiple-choice question:

A **[XX]** year-old patient with symptomatic severe aortic stenosis has NYHA Class **[XX]** dyspnoea. Left ventricular ejection fraction is calculated at **[XX]**%, valve area is **[XX]** cm^2, and mean gradient is **[XX]** mmHg. Coronary angiography shows **[XX]**. The patient's overall condition is **[good/poor]** with a **[favourable/unfavourable]** geriatric assessment for aortic valve intervention. Valvular calcium score is **[XX]**, femoral access is of **[good/poor]** diameter **[with/without]** extreme tortuosity, and there **[is/isn't]** significant carotid stenosis. STS score is calculated at **[XX]**%, and EuroSCORE II is calculated at **[XX]**%. What is the best option for the management of this patient's aortic stenosis?

Which of these answers is correct:

1. Intervention with percutaneous valve replacement via transcatheter aortic valve implantation (TAVI) is indicated.
2. Intervention with surgical replacement of the aortic valve is indicated.
3. Medical management alone is indicated.

Figure 1. *Standardised text that was used for each patient and that was submitted to ChatGPT. EuroSCORE: European System for Cardiac Operative Risk Evaluation; GPT: Generative Pre-trained Transformer; NYHA: New York Heart Association; STS: Society of Thoracic Surgeons*

OUTCOMES

We considered ChatGPT to be deterministic and only took into account the first response provided, without assessing or quantifying variability in the same scenario. Each clinical vignette was presented to ChatGPT in separate windows. This approach ensured that the answers generated by ChatGPT were independent of previous responses. The collected answers were then compared to those provided by the HT. The primary outcome was the overall agreement between ChatGPT and the HT regarding the choice of treatment option (TAVI, SAVR or medical treatment), while the secondary outcome was the overall agreement rate between the HT and a simple classifier using the European Society of Cardiology (ESC) and American Heart Association (AHA) guidelines' decision trees[1,2].

For this preliminary feasibility study, we focused only on the agreement rate between ChatGPT, the HT and these basic decision trees in the first instance, and the follow-up of the patients was not analysed.

STATISTICAL ANALYSES

Data were summarised using descriptive statistics, with mean±standard deviation (SD) for normally distributed continuous variables and median (interquartile range [IQR]) for non-normally distributed continuous variables. Frequencies with percentages were used for categorical variables. The agreement rate regarding the choice of treatment option (TAVI, SAVR or medical treatment) was assessed between the responses provided by the HT, the guideline-derived decision trees and ChatGPT overall and for each treatment option individually. This study was not powered to allow for advanced between-group comparisons. All analyses were carried out using SPSS Statistics (IBM).

As a high proportion of undetermined choice based on the current ESC guidelines was expected, a 5-fold cross-validation with a 120/30 train/test split, leveraging the scikit-learn toolkit in the Python package, was also applied in order to train the ESC guideline-derived decision tree. Gini impurity, as a criterion to measure the probability of incorrect classification at each node, was used. The decision trees evaluated variables as features, requiring a minimum of 2 samples for splitting a node, and followed scikit-learn's default settings for other parameters.

Results

PATIENT CHARACTERISTICS

A total of 150 consecutive patients with severe AS discussed in HT meetings between November 2021 and August 2023 were included. The mean age was 77±10 years, and the median NYHA dyspnoea Class was II [IQR II-III]. On echocardiographic evaluation, the mean valvular area was 0.7±0.2 cm^2, the median transvalvular gradient was 42 [IQR 32-50] mmHg and the mean LVEF was 57±13%. Coronary angiograms revealed significant coronary artery disease in 54% of the patients. Regarding the carotid artery evaluation, 10% of patients presented with a stenosis of 50% or greater. On the CT scan, the mean valvular calcium score was 3,072±1,960 arbitrary units. Geriatric assessment was in favour of an invasive intervention on the aortic valve (either TAVI or SAVR) in 83% of the cases. Finally, the surgical risk assessment, measured by the STS score and EuroSCORE II, averaged 3.7±2.6% and 3.8±2.9%, respectively.

HEART TEAM DECISIONS

According to the assessment conducted by the HT, 70 patients were considered eligible for TAVI, 60 patients for SAVR, and 20 patients for medical treatment. **Table 1** provides a comprehensive summary of the baseline characteristics of the overall study population, as well as a detailed analysis based on the decision given by the HT.

CHATGPT DECISIONS

GPT-4 provided an answer for all 150 patients, achieving an overall accuracy of 77%. Moreover, it displayed an agreement rate of 90% when dealing with patients eligible for TAVI, 65% for SAVR, and 65% for medical treatment. Details regarding the answers of GPT-4 and basic statistical analyses are reported in the **Central illustration**, **Figure 2** and **Table 2**.

GUIDELINE-DERIVED DECISIONS

The ESC guideline-derived decision tree provided an answer for 144 patients (96%), achieving an overall agreement of 73%. It displayed an agreement rate of 76% when dealing with patients eligible for TAVI, 73% for SAVR, and 60% for medical treatment. The AHA guideline-derived decision tree provided an answer for only 88 patients, achieving an overall agreement rate of 43%. It displayed an agreement rate of 49% when dealing with patients eligible for TAVI, 28% for SAVR, and 70% for medical treatment. For both decision trees, indeterminate answers were due to a grey zone where the patient was eligible for either SAVR or TAVI. Details are reported in **Figure 3**.

DECISION COMPARISON

GPT-4 "misclassified" a total of 35 patients. Among them, GPT-4 recommended TAVI instead of SAVR for 21 patients, SAVR or medical treatment instead of TAVI for 7 patients, and TAVI instead of medical treatment for 7 patients.

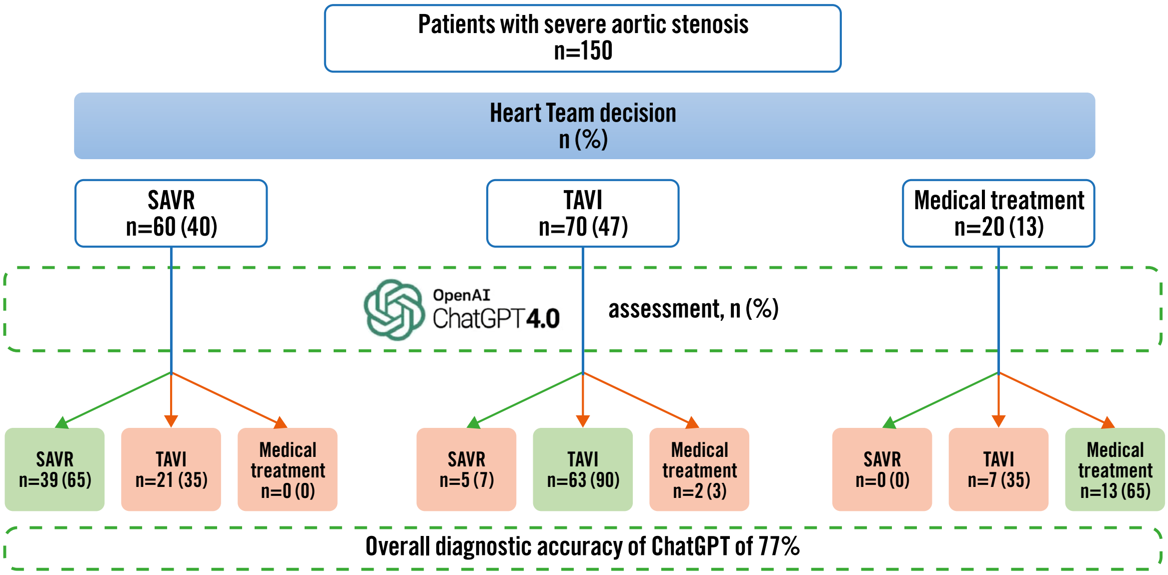
Table 1. Baseline characteristics of patients according to Heart Team decision.

	Overall N=150	TAVI N=70	SAVR N=60	Medical treatment N=20
Clinical data				
Age, years	77±10	82±6	68±9	86±8
NYHA dyspnoea Class	2 [2-3]	3 [2-3]	2 [2-3]	2 [2-3]
I	21 (14)	9 (13)	9 (15)	3 (15)
II	62 (41)	25 (36)	27 (45)	10 (50)
III	61 (41)	30 (43)	24 (40)	7 (35)
IV	6 (4)	6 (8)	0 (0)	0 (0)
Favourable geriatric assessment	124 (83)	58 (83)	60 (100)	6 (30)
Echocardiographic data				
LVEF, %	57±13	58±14	57±12	56±10
Surface valve area, cm^2	0.7±0.2	0.7±0.2	0.8±0.2	0.7±0.3
Median gradient, mmHg	42 [32-50]	41 [32-49]	45 [35-54]	39 [26-46]
Cardiovascular assessment				
CAD	81 (54)	39 (56)	30 (50)	12 (60)
Carotid stenosis	15 (10)	13 (19)	2 (3)	0 (0)
Valvular calcium score, AU	3,072±1,960	2,722±1,748	3,629±2,473	3,007±1,986
Risk score				
EuroSCORE II	3.7±2.6	4.1±2.4	2.5±2.1	5.2±2.8
STS score	3.8±2.9	4.4±2.8	2.3±1.8	6.2±3.8

Data are presented as mean±SD, median [IQR] or n (%). AU: arbitrary units; CAD: coronary artery disease; IQR: interquartile range; LVEF: left ventricular ejection fraction; NYHA: New York Heart Association; SAVR: surgical aortic valve replacement; SD: standard deviation; STS: Society of Thoracic Surgeons; TAVI: transcatheter aortic valve implantation

EuroIntervention **Central Illustration**

Heart Team decisions versus ChatGPT's decisions.



Adil Salihu *et al.* • *EuroIntervention* 2024;20:e496-e503 • DOI: 10.4244/EIJ-D-23-00643

GPT: Generative Pre-trained Transformer; SAVR: surgical aortic valve replacement; TAVI: transcatheter aortic valve implantation

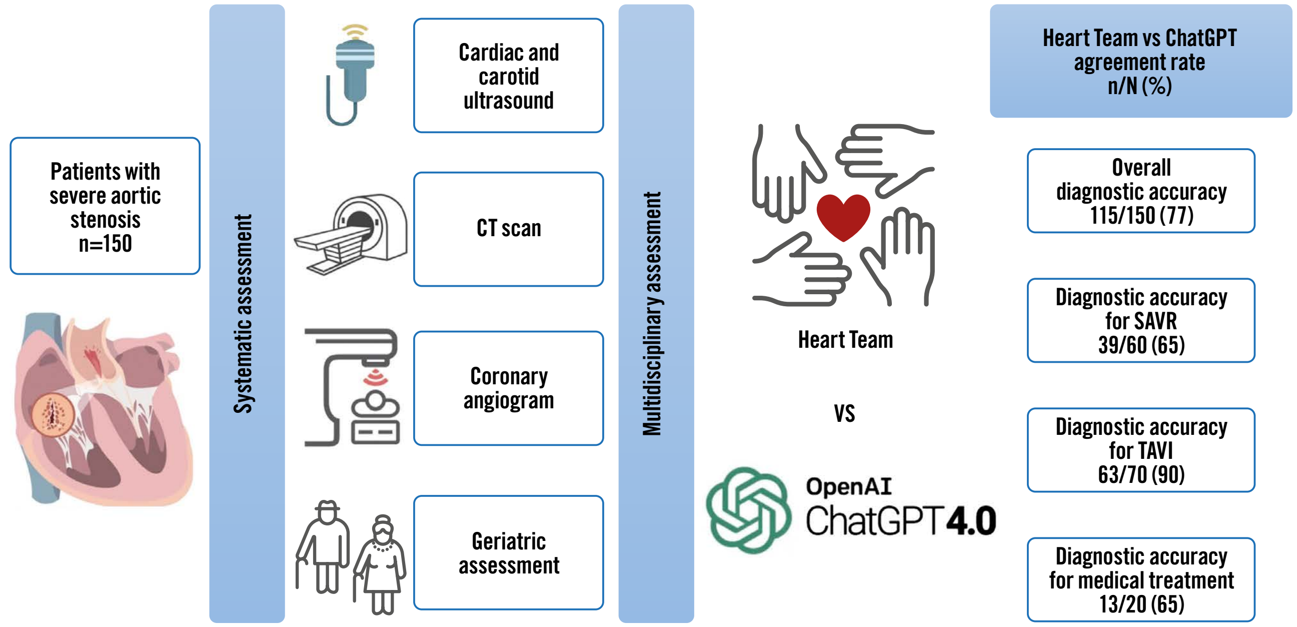

Patients with severe aortic stenosis n=150
Systematic assessment
Cardiac and carotid ultrasound
CT scan
Coronary angiogram
Geriatric assessment
Multidisciplinary assessment
Heart Team
VS
OpenAI ChatGPT4.0
Heart Team vs ChatGPT agreement rate n/N (%)
Overall diagnostic accuracy 115/150 (77)
Diagnostic accuracy for SAVR 39/60 (65)
Diagnostic accuracy for TAVI 63/70 (90)
Diagnostic accuracy for medical treatment 13/20 (65)

Figure 2. *Workflow diagram with the agreement rate of Heart Team and GPT-4 decisions. CT: computed tomography; GPT: Generative Pre-trained Transformer; SAVR: surgical aortic valve replacement; TAVI: transcatheter aortic valve implantation*

Table 2. Performance of ChatGPT.

	ChatGPT for TAVI	ChatGPT for SAVR	ChatGPT for medical treatment
Sensitivity	63/70 (90)	39/60 (65)	8/20 (40)
Specificity	47/80 (59)	85/90 (94)	128/120 (98)
Predictive positive value	63/96 (66)	39/44 (87)	8/10 (80)
Predictive negative value	47/54 (87)	84/106 (80)	128/140 (91)

Data are presented as n/N (%). GPT: Generative Pre-trained Transformer; SAVR: surgical aortic valve replacement; TAVI: transcatheter aortic valve implantation

The ESC guideline-derived decision tree performance was inferior to that of GPT-4. It "misclassified" a total of 41 patients. Among them, the guideline-derived decision tree recommended TAVI instead of SAVR for 12 patients, SAVR or medical treatment instead of TAVI for 15 patients, and TAVI instead of medical treatment for 8 patients. Indeterminate recommendations were identified in 4 patients scheduled for surgery and 2 patients set to undergo TAVI. After training the decision trees based on ESC guidelines, the overall accuracy was 67% and remained lower than that of GPT-4.

The AHA guideline-derived decision tree performance was also inferior to that of GPT-4. It "misclassified" a total of 85 patients. Among them, the guideline-derived decision tree recommended TAVI treatment instead of SAVR for 5 patients, medical treatment instead of TAVI for 14 patients, and TAVI instead of medical treatment for 4 patients. Indeterminate recommendations were identified in 38 patients scheduled for surgery, 22 patients set to undergo TAVI and 2 patients recommended medical treatment.

Discussion

To our knowledge, this is the first study evaluating the performance of AI, specifically the ChatGPT model, in aiding decision-making for the management of severe AS. These results illustrate a promising role for LLMs in enhancing the decision-making process within an HT.

In this study, based on 150 consecutive patients, we demonstrated, with the use of only 14 standardised variables, the feasibility of incorporating AI into the preliminary assessment of patients' data. The overall accuracy of GPT-4 in this particular setting was found to be 77%. Nevertheless, the agreement was not uniform across the 3 final treatment decisions made by the HT, and this high agreement rate was mainly driven by the good accuracy of GPT-4 for patients selected for TAVI: 90% for TAVI, 65% for SAVR and 65% for medical treatment.

Regarding patients for whom the HT proposed medical treatment only, the sample size is too small (20 patients) to allow for any definitive conclusion, but several interesting observations can be made: ChatGPT did not suggest surgery for any of these patients, which confirms its ability to identify patients at high or prohibitive surgical risk. However, the algorithm still suggested TAVI for 7 patients assigned to medical treatment, illustrating the difficulty and uncertainty surrounding such a decision. Usually, the decision to choose medical management is mainly related to excessive frailty, the number of comorbidities, or expected limited life expectancy, suggesting the futility of an invasive procedure. However, this final decision is complex and

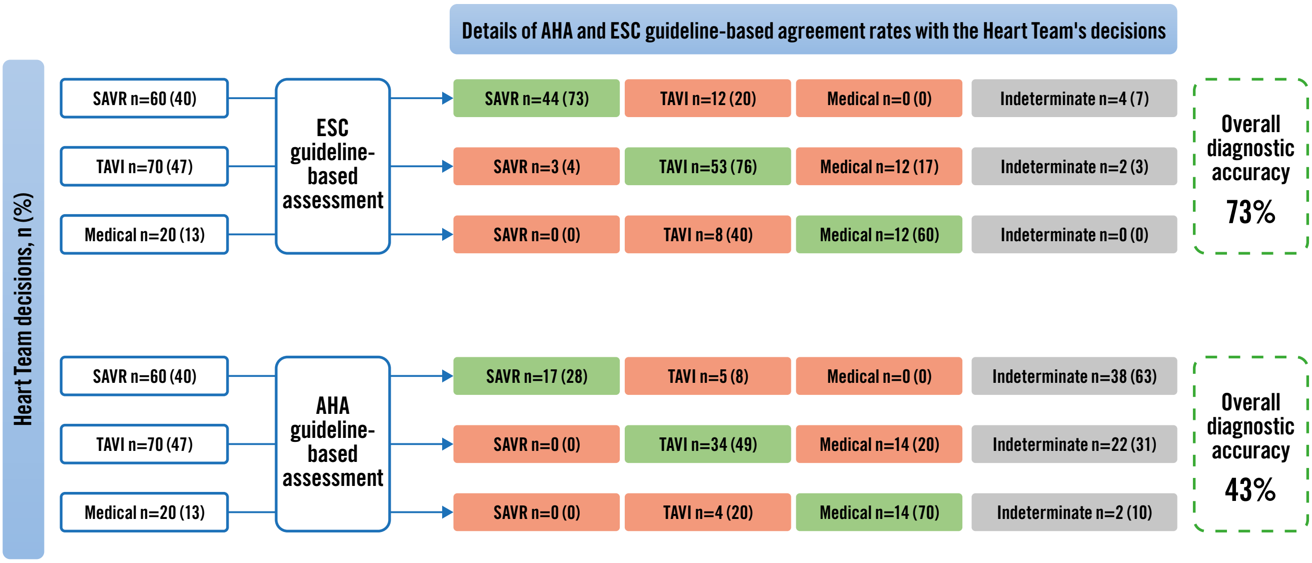


Figure 3. *Flowchart of ESC and AHA guideline-based assessments and their agreement rate with the Heart Team. AHA: American Heart Association; ESC: European Society of Cardiology; SAVR: surgical aortic valve replacement; TAVI: transcatheter aortic valve implantation*

potentially not fully represented by the 14 variables used in the present study. Indeed, among these 7 patients, the assessment revealed 3 new oncological situations with an uncertain prognosis, and 4 patients had a significant perioperative risk due to comorbidities not reflected in the geriatric assessment.

As for patients for whom the HT proposed TAVI or SAVR, several cases were classified differently by ChatGPT. For TAVI cases, only 10% received an alternate treatment recommendation, while for patients for whom SAVR was proposed, the rate of different recommendations was 35%. Patients inaccurately classified by GPT-4, those designated initially to undergo SAVR following the HT meeting, predominantly ranged in age from 70 to 80 years old. Likewise, among patients intended for TAVI after the HT decision, the age of misclassified patients tended to be in the same age range, between 70 and 80. This observation regarding age range is interesting when considering the differences between European and US guidelines regarding the age cutoffs on which the choice of either TAVI or SAVR is based[1,2]. The ESC recommends surgical management for patients under the age of 75 with low surgical risk, whereas AHA guidelines set a cutoff point of less than 65 years old for surgical management and of more than 80 years old for TAVI, with case-by-case assessment for patients between these age ranges.

In comparing HT decisions with the management algorithms in the ESC and AHA guidelines, we noted variations in accuracy and in the rates of indeterminate responses concerning surgical or percutaneous recommendations. Specifically, the agreement rate between HT decisions and the ESC guidelines' algorithm was 73%, with 6% of responses remaining indeterminate (unable to decisively recommend TAVI or SAVR). In contrast, the AHA guidelines' algorithm aligned with HT decisions 43% of the time only, exhibiting a significantly higher indeterminate response rate of 41% (either TAVI or SAVR). This difference in indeterminate responses suggests a more flexible decision-making approach within the AHA guidelines, which might contribute to the variability in GPT-4's responses and could potentially lead to ambiguity, impacting its overall accuracy.

These performances, although close, did not surpass the accuracy achieved by ChatGPT, which was 77%. These different findings are possibly explained by the fact that ChatGPT's access to current guidelines is a contributing factor to its performance which closely aligns with the results derived from a guideline-based approach. In contrast to the guideline-based classifier, which often demonstrates a degree of uncertainty, ChatGPT consistently provides clear and decisive recommendations. This distinction underscores ChatGPT's enhanced capability for decision support in clinical scenarios, offering insights beyond the capacities of a simple guideline-based classifier, especially in uncertain cases. This extended analysis highlights ChatGPT's potential as a more effective decision-support tool in such clinical scenarios.

The introduction of ChatGPT in November 2022 has been widely covered by popular media and scientific publications[3,4], including in the field of cardiology. We previously reported that machine learning demonstrated superior accuracy in predicting future myocardial infarction events in comparison to human assessment or even angiographic parameters[5] and that ChatGPT could successfully pass the European Cardiology board examination[6]. The success of ChatGPT, which was not specifically trained for medical purposes, opens up possibilities for its potential application in clinical decision-making. However, employing such systems for diagnosis and treatment carries potential risks. There is a concern regarding the accuracy of generated responses, as our team recently reported: variables such as the version of ChatGPT or limited information about the user's identity can influence the output from LLM-based tools[7]. This can be attributed to the non-deterministic nature of ChatGPT which leads to variability in ChatGPT's responses. Some researchers have also noted that these models have a tendency to generate incorrect or fabricated information, a phenomenon referred to as "hallucinations"[3,8]. This raises the question of the decision-making process, which can be opaque and acts as a "black box", all of which could ultimately undermine the clinician's understanding of and trust towards AI. Therefore, the clinician's observational and critical thinking retain their utmost importance in assessing such scenarios and the clinical solutions generated by AI.

Limitations

This study has several limitations that need to be discussed. First, the simplification of certain parameters that may have had an impact on the results, such as coronary lesions or the quality of femoral or carotid arteries, must be acknowledged. Furthermore, the number of variables used in this study to form the clinical vignettes was limited and may not have reflected the complexity of some cases. However, the vignettes used reproduce real-life clinical situations in the way that they are often presented during a Heart Team meeting, where data are often presented in a binary fashion and the members of the HT are then asked: What is the best management option for this patient?

To overcome this limitation, the incorporation of more comprehensive and even unstructured clinical data may enhance the performance of ChatGPT and could provide a more accurate reflection of a patient's clinical status. This would, however, raise ethical questions about data anonymisation. In addition, one could also imagine implementing a system where each patient case is entered into the chat interface, assigned a numerical label, and then systematically followed up allowing ChatGPT to adjust its decision-making based on the continuous input and corrective feedback, thereby aligning more closely with local clinical guidelines and protocols over time.

Second, it should be noted that ChatGPT's available information is limited to data collected prior to September 2021. Therefore, the model does not have access to the most recent evidence or developments regarding AS management.

Lastly, a divergence between the responses of ChatGPT and human experts does not necessarily indicate the presence of an error in assessing the patient's case. Variability could be attributed to potential limitations in the data provided to ChatGPT or the existence of a "grey zone" in the management of these patients, as discussed above.

Perspectives

This feasibility study does not delve into outcomes such as the long-term prognosis of patients for whom there was a discrepancy between the Heart Team's decision and ChatGPT's recommendation nor into the reasons behind the discrepancy observed, for which one could consider asking ChatGPT the rationale behind each decision. Investigating the real-world outcomes and implications of such differing decision paths is a significant avenue for future research, providing deeper insights into the potential benefits or pitfalls of AI-assisted clinical decision-making.

Conclusions

This study suggests that AI, specifically ChatGPT-4, could potentially play a role in the decision-making process within an HT. This lays the ground for future larger studies with a multicentre and prospective design. These studies should aim to comprehensively examine the factors contributing to the occasional divergences between ChatGPT's evaluations and the HT's decisions. Additionally, they should assess the patient outcomes associated with instances where such discrepancies are present. Despite the good performance observed, it is crucial to remember that AI tools are not intended to replace clinicians but rather to support them in their decision-making process. The final clinical decision should remain in the hands of the healthcare provider, considering the patient's unique clinical status and preferences. AI technologies have the potential to revolutionise healthcare, making it more efficient, personalised, and patient centred. Nevertheless, in order to fully achieve this potential, it is essential to tackle the challenges related to the interpretability, legal implications and ethical considerations of AI use in healthcare. As AI continues to evolve, we can anticipate an increasingly prominent role of these tools, with the ultimate goal of pushing the boundaries of what can be achieved in patient care.

Authors' affiliations

1. Department of Cardiology, Lausanne University Hospital, University of Lausanne, Lausanne, Switzerland; 2. Department of Cardiovascular Surgery, Lausanne University Hospital, University of Lausanne, Lausanne, Switzerland; 3. Institute of Mathematics and School of Computer and Communication Sciences, EPFL, Lausanne, Switzerland; 4. Department of Anesthesiology, Lausanne University Hospital, University of Lausanne, Lausanne, Switzerland; 5. Department of Radiology, Lausanne University Hospital, University of Lausanne, Lausanne, Switzerland; 6. Department of Geriatrics, Lausanne University Hospital, University of Lausanne, Lausanne, Switzerland; 7. Division of Cardiovascular Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA

Conflict of interest statement

The authors have no conflicts of interest to declare in relation to this manuscript.

References

1. Vahanian A, Beyersdorf F, Praz F, Milojevic M, Baldus S, Bauersachs J, Capodanno D, Conradi L, De Bonis M, De Paulis R, Delgado V, Freemantle N, Haugaa KH, Jeppsson A, Jüni P, Pierard L, Prendergast BD, Sádaba JR, Tribouilloy C, Wojakowski W. 2021 ESC/EACTS Guidelines for the management of valvular heart disease. *EuroIntervention.* 2022; 17:e1126-96.

2. Otto CM, Nishimura RA, Bonow RO, Carabello BA, Erwin JP 3rd, Gentile F, Jneid H, Krieger EV, Mack M, McLeod C, O'Gara PT, Rigolin VH, Sundt TM 3rd, Thompson A, Toly C. 2020 ACC/AHA Guideline for the Management of Patients With Valvular Heart Disease: Executive Summary: A Report of the American College of Cardiology/

American Heart Association Joint Committee on Clinical Practice Guidelines. *Circulation*. 2021;143:e35-71.

3. Lee P, Bubeck S, Petro J. Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine. *N Engl J Med*. 2023;388:1233-9.

4. Haug CJ, Drazen JM. Artificial Intelligence and Machine Learning in Clinical Medicine, 2023. *N Engl J Med*. 2023;388:1201-8.

5. Mahendiran T, Thanou D, Senouf O, Meier D, Dayer N, Aminfar F, Auberson D, Raita O, Frossard P, Pagnoni M, Cook S, De Bruyne B, Muller O, Abbé E, Fournier S. Deep learning-based prediction of future myocardial infarction using invasive coronary angiography: a feasibility study. *Open Heart*. 2023;10:e002237.

6. Skalidis I, Cagnina A, Luangphiphat W, Mahendiran T, Muller O, Abbe E, Fournier S. ChatGPT takes on the European Exam in Core Cardiology: an artificial intelligence success story? *Eur Heart J Digit Health*. 2023;4:279-81.

7. Salihu A, Gadiri MA, Skalidis I, Meier D, Auberson D, Fournier A, Fournier R, Thanou D, Abbé E, Muller O, Fournier S. Towards AI-assisted cardiology: a reflection on the performance and limitations of using large language models in clinical decision-making. *EuroIntervention*. 2023;19:e798-801.

8. Alkaissi H, McFarlane SI. Artificial Hallucinations in ChatGPT: Implications in Scientific Writing. *Cureus*. 2023;15:e35179.