Exploration of Cardiology

Open Exploration

Open Access Original Article

Diagnostic performance of artificial intelligence in the syncope unit

Steven van Zanten[1,2*], Thomas T. Boel[1,3], Jelle SY de Jong[1,3], Egbert M Koomen[4], Babette Bais[5], Ako Dara[6], Freek Giele[7], Christiaan Geertsma[8], Richard Sutton[9], Mike G Scheffer[2], Joris R de Groot[1,3], Frederik J de Lange[1,3]

[1]Experimental Cardiology, Heart Center, Amsterdam University Medical Center, University of Amsterdam, Amsterdam, Noord-Holland 1105 AZ, The Netherlands

[2]Department of Cardiology, Reinier de Graaf Gasthuis, Delft, Zuid-Holland 2625 AD, The Netherlands

[3]Heart Failure and Arrhythmias, Amsterdam Cardiovascular Sciences, Amsterdam, Noord-Holland 1105 AZ, The Netherlands

[4]Department of Cardiology, Gelre Hospitals, Apeldoorn, Gelderland 7334 DZ, The Netherlands

[5]Reinier Academy, Reinier de Graaf Gasthuis, Delft, Zuid-Holland 2625 AD, The Netherlands

[6]Department of Neurology, Reinier de Graaf Gasthuis, Delft Zuid-Holland, 2625 AD, The Netherlands

[7]Rubicon B.V., Leusden, Utrecht 3832 RC, The Netherlands

[8]ICT Department, Amsterdam University Medical Center, Vrije Universiteit Amsterdam, Amsterdam, Noord-Holland 1081 HV, The Netherlands

[9]Department of Cardiology, Hammersmith Hospital Campus, National Heart & Lung Institute, Imperial College, W12 0NN London, United Kingdom

***Correspondence:** Steven van Zanten, Department of Cardiology, Reinier de Graaf Gasthuis, Reinier de Graafweg 5, Delft, Zuid-Holland 2625 AD, The Netherlands. stevenvz@rdgg.nl

Academic Editor: Teresa Tsang, University of British Columbia, Canada; Eugenio Picano, Italian National Research Council, Italy

Received: April 17, 2026 **Accepted:** June 3, 2026 **Published:** July 10, 2026

Cite this article: van Zanten S, Boel TT, de Jong JSY, Koomen EM, Bais B, Dara A, et al. Diagnostic performance of artificial intelligence in the syncope unit. Explor Cardiol. 2026;4:1012114. https://doi.org/10.37349/ec.2026.1012114

Abstract

Aim: Artificial intelligence may support syncope evaluation, but reliability of language models in structured syncope care remains uncertain. We evaluated diagnostic performance, safety, and within-case consistency of Generative Pre-trained Transformer-5 (GPT-5) in patients with transient loss of consciousness (T-LOC).

Methods: This prospective cohort study included 55 patients evaluated in syncope units. GPT-5 and a syncope-expert assessed identical case information after core evaluation (CE: history-taking, physical examination, active standing test, and 12-lead electrocardiogram) and extended evaluation (EE: CE plus additional testing when indicated). An expert panel adjudicated the final diagnosis after 18 months. Outcomes were diagnostic yield, final-diagnosis inclusion rate, diagnostic precision score (DPS), cardiac diagnostic safety, and within-case consistency across five repeated GPT-5 runs.

Results: Of 55 patients, 54 had complete follow-up for performance analyses. Diagnostic yield was 94% for the syncope-expert at CE and EE, and 100% and 96% for GPT-5 at CE and EE, respectively. GPT-5 included the final diagnosis in 52% (CE) and 57% (EE) of cases, versus 67% for the syncope-expert. DPS remained negative for GPT-5 at CE (mean −0.03, SD 0.54) and EE (mean −0.01, SD 0.49). Among four final cardiac syncope cases, GPT-5 selected the final diagnosis in one case and the syncope-expert in three. First-diagnosis consistency across five GPT-5 runs was 69% after CE and 74% after EE.

Conclusions: GPT-5 generated diagnostic outputs frequently but showed limited precision, cardiac safety concerns, and within-case variability. Its role in syncope evaluation should remain supportive within clinician-led pathways rather than autonomous.

Keywords

artificial intelligence, syncope unit, diagnostic accuracy, clinical decision, education, research

Introduction

Transient loss of consciousness (T-LOC) is a symptom with a broad differential diagnosis spanning neurology, cardiology, and internal medicine. This heterogeneity increases diagnostic uncertainty in acute and secondary care, hampering clinical decision making [1, 2].

Syncope units (SUs) address this diagnostic uncertainty through a guideline-based "initial evaluation" [1]. Protocols among SUs vary [3], but usually combine multidisciplinary input, in which the patient's presentation is assessed from different diagnostic perspectives with autonomic testing. Referral letters often lack essential symptom descriptions, physical findings, electrocardiogram (ECG) and medication histories. Furthermore, pre-referral non-pharmacological intervention and medication reviews are often not performed by referents [4]. After all, structured thorough history-taking remains the cornerstone of T-LOC assessment [1]. These essential findings, when documented accurately and comprehensively, support effective initial triage and provide the basis for subsequent risk stratification, in accordance with the European Society of Cardiology (ESC) syncope guidelines, thereby helping to ensure patient safety [1].

Artificial intelligence (AI) has increasingly been explored as an adjunctive diagnostic and decision-support approach in syncope evaluation [4–10]. Recent studies demonstrate potential applications across the syncope pathway, including recognition and classification of T-LOC, distinction of syncope from syncope mimics, emergency-department risk stratification, prediction of short-term adverse events, hospitalization and length of stay, and automated ECG interpretation [5–10]. However, robust assessment of performance and safety depends on systematic clinical follow-up resulting in a final or reference diagnosis [4, 7, 11–14]. More broadly, ChatGPT and other large language models may support clinical communication, documentation, education and workflow-related tasks, although concerns regarding hallucinations, inaccuracy, bias, overreliance, privacy and the need for further clinical validation remain [15–20].

We previously compared the performance of Generative Pre-trained Transformer-4 omni (GPT-4o) [21] to that of medical professionals using only referral letters to the SU [4]. GPT-4o's recall was high, but precision was poor. The most alarming cause of T-LOC, cardiac syncope, was often missed. While many therapeutic interventions were aligned with current recommendations of the ESC guidelines, the diagnostic output of the Large Language Model (LLM) approach lacked the clinical judgement necessary for safe autonomous implementation [4].

In this study, we evaluate the diagnostic performance, safety, and consistency of the Generative Pre-trained Transformer-5 (GPT-5) model using structured clinical reports derived from evaluations performed by a syncope-expert. Unlike many diagnostic evaluations of LLMs that focus primarily on a single model output, we also examined whether GPT-5 provided stable diagnostic conclusions when identical clinical information was presented repeatedly. The unique contribution of this study is therefore the combined assessment of diagnostic yield, diagnostic inclusion, diagnostic precision, cardiac diagnostic safety, and within-case consistency across five independent runs per case, using systematic follow-up and blinded expert-panel adjudication as the reference standard.

Materials and methods

Patients

We investigated a random sample of 55 general practitioner-referred patients evaluated in SUs across three general hospitals, selected from the prospectively collected consecutive Haaglanden SU database (April

2015–December 2022), the same cohort as in our previous study [4]. This database comprises patients with documented T-LOC who were referred for evaluation and were assessed by the same syncope-expert in a dedicated SU following a standardized protocol (as described in the following section).

SU evaluation

The study protocol was approved by the Medical Ethics Review Committee Leiden/The Hague and Delft (METC nr: 18-061). All participants provided written informed consent. GPT-5 was evaluated under controlled conditions at the Amsterdam University Medical Center, ensuring patient safety and data privacy. Ethical approval for the use of GPT-5 with anonymized data was obtained, including General Data Protection Regulation compliance and provisions for human oversight (METC, Amsterdam, nr: 2024.0124).

All patients were first evaluated in the SU by the same syncope-expert, using a standardized diagnostic work-up protocol (see Figure 1). The initial diagnostic assessment was performed in accordance with guideline-based syncope evaluation and is hereafter referred to as core evaluation (CE). CE consisted of thorough history-taking, physical examination, an active standing test, and a 12-lead ECG. Extended evaluation (EE) comprised CE plus transthoracic echocardiography (TTE) and exercise ECG (XECG) whenever feasible, with additional investigations (e.g., Holter monitoring and head-up tilt testing [HUTT]) performed only when clinically indicated, at the discretion of the syncope-expert (Figure 1). Findings from the evaluation were recorded in a fixed, structured format across all cases to minimise interpretative variation. The clinical report was limited to a structured summary of observed clinical and test data; no 'impression/assessment' or diagnostic interpretation was provided to GPT-5. In cases of diagnostic uncertainty, multidisciplinary consultation was readily available.

Follow-up

The syncope-expert and GPT-5 made a diagnosis after the CE and after the EE, using identical information. Thereafter, we followed the patients systematically for eighteen months and gathered information regarding T-LOC episodes and diagnostic changes. Structured questionnaires captured all recurrent syncope, further diagnostic testing, hospital admissions, and any change in diagnosis following the SU visit (Figure 1). Follow-up consult was performed at six and eighteen months. When appropriate, data were verified by additional phone consultation.

After follow-up, an expert panel of three physicians (FJdL, JSYdJ, TTB) with expertise in syncope reviewed all cases using all available data, without reference to the prior diagnostic conclusion. GPT-5 was evaluated in a non-interactive, case-based format and could not ask clarifying questions, request additional information, or interact with patients or clinicians. Both the final adjudication panel and GPT-5 were blinded to the diagnostic labels under comparison: the adjudication panel did not receive the initial diagnosis of the syncope-expert or GPT-5 diagnoses, and GPT-5 did not receive the syncope-expert diagnosis or the final adjudicated diagnosis. They reached a 'final diagnosis' following the same protocol as in prior studies [4, 22]. Cases in which no etiological final diagnosis could be supported after complete follow-up were classified as 'unexplained T-LOC', consistent with guideline-based syncope practice. The final diagnosis was employed as the reference standard to evaluate diagnostic precision and safety, as outlined in the following section (Figure 1). Separate clinical reports were drafted following the CE and EE, each encompassing the data corresponding to that particular assessment. An anonymized example of the structured CE/EE report template is provided in Supplementary material 1.

Risk-stratification

Risk features were defined and categorized using the 2018 version of the ESC syncope guidelines [1]. Specifically, low-risk, high-minor, and high-major risk features were evaluated across history, physical examination, and ECG. Minor high-risk features were subsequently interpreted according to the conditional ESC criteria: minor high-risk features related to the syncopal event were classified as clinically high-risk only when accompanied by structural heart disease and/or an abnormal ECG, whereas minor high-risk electrocardiographic features were classified as clinically high-risk only when the history was consistent

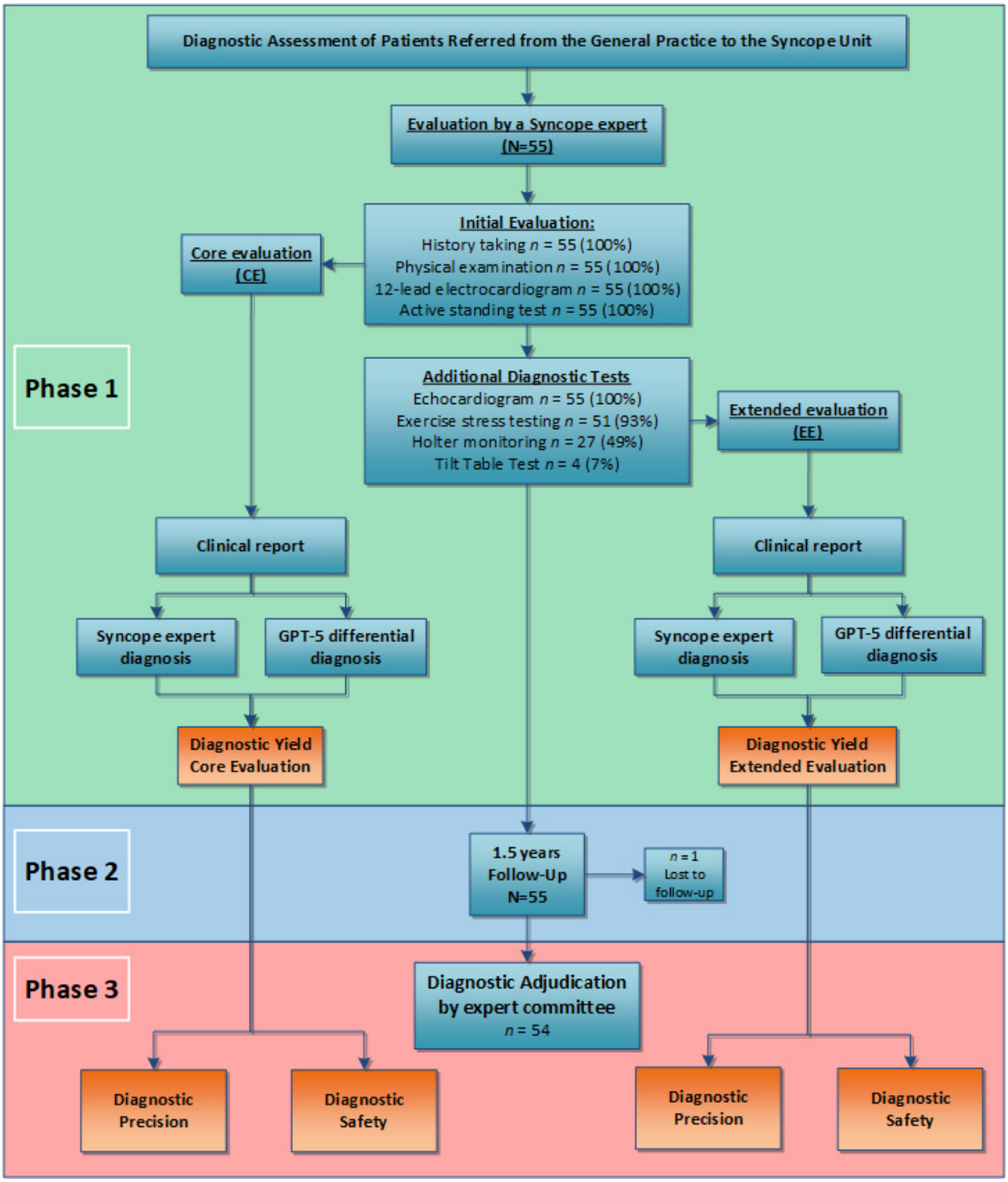

Figure 1. Diagnostic flowchart. Phase-1: Core evaluation (CE) in the syncope unit by the syncope-expert and GPT-5. Phase-2: Follow-up of 1.5 years. Phase-3: Diagnostic adjudication by an expert committee after 1.5 years of follow-up, serving as the reference standard diagnosis, resulting in the final diagnosis. In addition to adjudication, diagnostic precision and diagnostic safety were also assessed.

with arrhythmic syncope (see Table 1) [23]. For each SU letter, we assessed whether these features were explicitly documented. This assessment was performed to evaluate the completeness of reporting; letters were not excluded on the grounds of missing information, and no information was added.

Instructions to LLM

LLMs such as GPT-5 are a specific type of AI based on transformer architectures, trained on massive text corpora to predict the next word (token) in a sequence; by repeating this prediction step, they can generate, summarize, and reformulate clinical text and answer free-text questions in a probabilistic, pattern-based manner [17–19].

Table 1. Risk-stratification in the syncope unit.

Category	Risk Level	Feature	Total per item	Total per group
Syncopal event	Low-risk	Associated with prodrome typical of reflex syncope (e.g. light-headedness, feeling of warmth, sweating, nausea, vomiting)	27	76
	Low-risk	After sudden unexpected unpleasant sight, sound, smell, or pain	7	
	Low-risk	After prolonged standing or crowded, hot places	11	
	Low-risk	During a meal or postprandial	3	
	Low-risk	Triggered by cough, defaecation, or micturition	8	
	Low-risk	With head rotation or pressure on carotid sinus (e.g. tumour, shaving, tight collars)	1	
	Low-risk	Standing from supine/sitting position	19	
	High-risk (Minor)	No warning symptoms or short (< 10 seconds) prodrome	31	48
	High-risk (Minor)	Family history of Sudden Cardiac Death at young age	6	
	High-risk (Minor)	Syncope in the sitting position	11	
	High-risk (Major)	New onset of chest discomfort, breathlessness, abdominal pain, or headache	1	14
	High-risk (Major)	Syncope during exertion or when supine	11	
	High-risk (Major)	Sudden onset palpitation immediately followed by syncope	2	
Past medical history	Low-risk	Long history (years) of recurrent syncope with low-risk features with the same characteristics of the current episode	14	60
	Low-risk	Absence of structural heart disease	46	
	High-risk (Major)	Severe structural or coronary artery disease (heart failure, low Left Ventricular Ejection Fraction or previous myocardial infarction)	6	6
Physical examination	Low-risk	Normal examination	45	45
	High-risk (Major)	Unexplained systolic blood pressure in the emergency department < 90 mmHg	0	10
	High-risk (Major)	Suggestion of gastrointestinal bleed on rectal examination	0	
	High-risk (Major)	Persistent bradycardia (< 40 b.p.m.) in awake state and in absence of physical training	0	
	High-risk (Major)	Undiagnosed systolic murmur	10	
Electrocardiogram	Low-risk	Normal electrocardiogram	43	43
	High-risk (Minor)	Mobitz I second-degree Atrioventricular block block and 1°degree Atrioventricular block with markedly prolonged PR interval	0	3

Table 1. Risk-stratification in the syncope unit. (*continued*)

Category	Risk Level	Feature	Total per item	Total per group
	High-risk (Minor)	Asymptomatic inappropriate mild sinus bradycardia (40–50 b.p.m.), or slow atrial fibrillation (40–50 b.p.m.)	1	
	High-risk (Minor)	Paroxysmal Supraventricular Tachycardia or Atrial Fibrillation	1	
	High-risk (Minor)	Pre-excited QRS complex	0	
	High-risk (Minor)	Short QTc interval (≤ 340 ms)	0	
	High-risk (Minor)	Atypical Brugada patterns	1	
	High-risk (Minor)	Negative T waves in right precordial leads, epsilon waves suggestive of arrhythmogenic right ventricular cardiomyopathy	0	
	High-risk (Major)	Electrocardiogram changes consistent with acute ischaemia	0	10
	High-risk (Major)	Mobitz II second- and third-degree Atrioventricular block	0	
	High-risk (Major)	Slow Atrial Fibrillation (< 40 b.p.m.)	0	
	High-risk (Major)	Persistent sinus bradycardia (< 40 b.p.m.), or repetitive sinoatrial block or sinus pauses >3 seconds in awake state and in absence of physical training	0	
	High-risk (Major)	Bundle branch block, intraventricular conduction disturbance, ventricular hypertrophy, or Q waves consistent with ischaemic heart disease or cardiomyopathy	10	
	High-risk (Major)	Sustained and non-sustained ventricular tachycardia	0	
	High-risk (Major)	Dysfunction of an implantable cardiac device (pacemaker or Implantable Cardioverter-Defibrillator)	0	
	High-risk (Major)	Type I Brugada pattern	0	
	High-risk (Major)	ST-segment elevation with type I morphology in leads V1–V3 (Brugada pattern)	0	
	High-risk (Major)	QTc > 460 ms in repeated 12-lead electrocardiogram indicating LQTS	0	

Guideline-based risk-stratification features per category (syncopal event, medical history, physical examination, and electrocardiogram). Totals are shown per feature and summed per risk-group. High-risk (minor) features were interpreted according to the conditional ESC criteria. Minor high-risk features related to the syncopal event were considered high-risk only when associated with structural heart disease and/or an abnormal electrocardiogram. Minor high-risk electrocardiographic features were considered high-risk only when the clinical history was consistent with arrhythmic syncope. Adapted with permission from [1].

To facilitate this analysis, a single pre-specified structured prompt was applied to all cases. It was developed based on the 2018 ESC syncope guideline and its web-based Practical Instructions [1, 24]. Prompt design followed the RICCE framework (Role, Instructions, Context, Constraints, and Examples), as previously applied in syncope-related GPT-4o evaluation, with the aim of improving standardization and reproducibility of model input [4]. GPT-5 was allowed to indicate diagnostic uncertainty and to state when the available information was insufficient to support a clear diagnosis. The prompt had been developed and tested previously on separate cases not included in the present study and was fixed before analysis. The same prompt was then applied across all study cases and across all five runs per case [10, 25, 26]. No additional model training or fine-tuning was performed prior to this analysis. GPT-5 was provided with the clinical report after the CE and after EE. Building on our prior study with GPT-4o using SU referral letters, we applied the same evaluation framework to the newer GPT-5 model, expected to perform better with richer, guideline-concordant SU data (CE and EE).

Analysis

We used several methods to assess the diagnostic performance of GPT-5 compared to the syncope-expert. In accordance with the Haaglanden database, the syncope-expert documented a single diagnosis for each patient rather than a differential diagnosis comprising multiple aetiologies. Outcomes were diagnostic yield, the diagnostic inclusion rate, and for GPT-5 the diagnostic precision score (DPS).

Diagnostic yield was defined as the proportion of cases in which any etiological diagnosis (reflex syncope, orthostatic hypotension (OH), cardiac syncope, or other specified cause) or 'unexplained T-LOC' was assigned by the syncope-expert or GPT-5 after CE and after EE. Yield therefore reflects etiological classification rather than direct physiological observation during the event.

The diagnostic inclusion rate quantified whether the final diagnosis appeared anywhere within GPT-5's differential diagnosis list (coverage), whereas the DPS (see next section) quantified the conciseness of that list by rewarding correct inclusion and penalizing over-inclusive differentials (precision).

The DPS was calculated using the methodology previously described [4]. In short: DPS rewards inclusion of the correct final diagnosis (+1 point) while penalizing over-inclusive differentials through a fixed deduction per incorrect diagnosis. For this we calculated a penalty score to determine the DPS for GPT-5 after both the CE and EE. For example, if GPT-5 provided the largest possible list of six differential diagnoses, including the final diagnosis and five incorrect ones, the penalty score for each incorrect diagnosis would be $1 \div 5 = 0.20$. The DPS for that case would then be: $+1 - (5 \times 0.20) = 0$. If the final diagnosis was not included at all, no point was awarded, but the penalty score is still applied to all diagnoses in the list, resulting in a negative DPS of $(5 \times 0.20) = -1.00$. This method ensures that concise, accurate diagnostic reasoning was rewarded, while excessive or unfocused differential lists were penalized accordingly (see also Supplementary Table S1).

Diagnostic safety

To assess diagnostic safety, we analyzed a subgroup of cases with a final diagnosis of cardiac syncope. Diagnostic safety was defined as how often the first diagnosis of cardiac syncope (the most alarming cause of T-LOC) was correctly identified by the syncope-expert and GPT-5 (CE and EE), compared to the final diagnosis. Secondly, all cases with a final cardiac diagnosis were traced back to their initial diagnostic classification to quantify how often cardiac syncope was missed at the first diagnostic assessment of the syncope-expert and GPT-5.

Within-case consistency of GPT-5

Per case, a single GPT-5 analysis was initially planned and executed. In one case, an additional analysis was performed inadvertently. The discrepant result prompted a protocol amendment. Five independent GPT-5 analyses (with the same prompt) per case were generated for all cases, and for both the CE and EE phases using identical case input, the same prompt template, and the same model parameters as in the primary analysis [4].

Within-case diagnostic consistency across the five runs was assessed by counting, for each run, how often the first diagnosis of the differential list was replicated. For discordant cases (≥ 2 different first diagnoses), specific diagnosis combinations were enumerated (e.g., OH with reflex, cardiac with reflex, OH with cardiac, etc.) to determine in which diagnostic domains the model showed inconsistency or mismatch. Also, we recorded the proportion of cardiac diagnoses besides reflex and OH that appeared at least once in the differential diagnosis.

Outcomes for comparison with the final diagnosis were summarized at two levels.

1. Case level: For each case, the number of runs was counted in which the first diagnosis of the differential diagnosis matched the final diagnosis (0/5–5/5).
2. Group level: All individual runs were pooled per phase (CE and EE, 54 cases × 5 analyses = 270 runs per phase), and the proportion of runs where the first differential diagnosis matched the final diagnosis was calculated.

Determinants and diagnostic inconsistency

We evaluated the determinants and their association with GPT-5's diagnostic inconsistency for CE and EE, including demographics; characteristics of the syncopal event before, during, and after the episode; vital signs; results of additional diagnostic tests; risk scores; and features of the syncope referral letter, as detailed in Supplementary Table S2.

Statistical analysis

Continuous variables are presented as means with standard deviations (SD) or medians with IQR, depending on the distribution of the data. Categorical variables are presented as counts and percentages. Diagnostic performance is reported as a percentage of total cases, representing the proportion of correct classifications.

The diagnostic correctness of GPT-5 and the syncope-expert (each assessed against the final diagnosis) a paired exact McNemar test was used; results are reported as a matched-pairs odds ratio (OR) with 95% confidence interval (CI), separately for CE and EE.

A Sankey plot was generated to visually represent the flow of data and the distribution of diagnostic outcomes between the syncope-expert and GPT-5, highlighting the transitions and associations across the various stages of the study.

Univariable logistic regression was performed to evaluate the association between each candidate factor and the AI outcome; results are presented as OR with 95% CIs and *P*-values.

Results

The study population consisted of 55 patients referred for evaluation of T-LOC and evaluated by a single syncope-expert (Figure 1). All underwent the CE and EE phases. Subsequently, syncope-expert clinical reports were evaluated by GPT-5 (Figure 1). The clinical characteristics are shown in Table 2. Data about the syncopal event and vital signs are shown in Table 3.

Risk-stratification

All 55 clinical reports after CE by the syncope-expert contained documentation of the syncopal event, physical examination, past medical history, and ECG findings. Across these domains, risk features were recorded per patient and event (Table 1) [23].

- Syncopal event: 76 low-risk; 62 high-risk (48 minor, 14 major).
- Physical examination: 45 low-risk; 10 high-risk.
- Past medical history: 60 low-risk; 6 high-risk.
- ECG findings: 43 low-risk; 13 high-risk (3 minor, 10 major).

Table 2. Patient characteristics.

Variable	Mean (SD)	Median (Q1, Q3)	Min–Max	*n* (%)
Demographics and anthropometrics				
Age (years)	61 (17)	63 (50, 75)	21–87	—
Female sex	—	—	—	30 (54.5)
Height (cm)	173 (9)	172 (167, 180)	155–193	—
Weight (kg)	77 (13)	75 (66, 87)	54–105	—
Body mass index (kg/m^2)	26 (3.7)	25 (23, 28)	18–36	—
Syncope characteristics				
Syncopal episodes in previous 12 months	2.11 (2.71)	1	0–20	—
Self-reported duration of loss of consciousness 0–5 min	—	—	—	48 (87.3)
Self-reported duration of loss of consciousness > 5 min	—	—	—	6 (10.9)
Recovery time after syncope (min)	8.09 (25.20)	0	0–180	—
Immediately reoriented after syncope	—	—	—	46 (83.6)
Prodromal symptoms present	—	—	—	30/54 (55.6)
Prodrome duration (s)	41.09 (116.23)	2	0–720	—
Previous diagnostic evaluation by general				
Number of diagnostic tests performed before evaluation syncope-expert	1.93 (1.37)	2	0–7	—
Fluid intake				
Daily fluid intake (mL)	1,636 (476)	1,750	500–2,500	—
Fluid intake assessed	—	—	—	1 (1.8)
Lifestyle				
Non-smoker	—	—	—	31 (56.4)
Ex-smoker	—	—	—	14 (25.5)
Current smoker	—	—	—	10 (18.2)
Alcohol intake	—	—	—	24 (43.6)
Substance use	—	—	—	1 (1.8)
Medical history				
Cardiac genetic predisposition	—	—	—	29 (52.7)
Hypertension	—	—	—	16 (29.1)
Family members deceased before age 55 years	—	—	—	6 (10.9)
Diabetes mellitus	—	—	—	6 (10.9)
Coronary artery disease	—	—	—	5 (9.1)
Hypercholesterolaemia	—	—	—	4 (7.3)
Cerebrovascular accident	—	—	—	4 (7.3)
Pulmonary embolism	—	—	—	3 (5.5)
Heart valve disease	—	—	—	1 (1.8)
Congenital heart disease	—	—	—	0 (0.0)
Parkinson’s disease	—	—	—	0 (0.0)
Epilepsy	—	—	—	0 (0.0)
Deep vein thrombosis	—	—	—	0 (0.0)
Medication use				
Any medication use	—	—	—	37 (67.3)
Number of medications per patient	2.98 (3.26)	2	0–14	—
Antihypertensive, heart-rate-lowering or antianginal medication	—	—	—	18 (32.7)
Gastrointestinal medication and laxatives	—	—	—	15 (27.3)
Endocrine, bone and vitamin supplementation	—	—	—	13 (23.6)
Antithrombotic medication	—	—	—	13 (23.6)
CNS-active medication	—	—	—	12 (21.8)

Table 2. Patient characteristics. (*continued*)

Variable	Mean (SD)	Median (Q1, Q3)	Min–Max	*n* (%)
Lipid-lowering medication	—	—	—	12 (21.8)
Allergy, ENT or ophthalmic medication	—	—	—	5 (9.1)
Hormonal contraception	—	—	—	5 (9.1)
Diabetes medication	—	—	—	4 (7.3)
Respiratory medication	—	—	—	4 (7.3)
Analgesic or anti-inflammatory medication	—	—	—	4 (7.3)
Rheumatologic, immunosuppressive or gout medication	—	—	—	4 (7.3)
Urogenital medication	—	—	—	2 (3.6)
Smoking cessation medication	—	—	—	2 (3.6)
Anti-infective medication	—	—	—	1 (1.8)

Fifty-five patients analyzed in the syncope unit by the syncope-expert.

Table 3. Clinical features of the syncopal event and vital signs.

Variable	Mean	SD	Minimum	Maximum	Median
Syncopal event					
T-LOC episodes lifetime	6.65	9.4	1	45	3
T-LOC episodes last year	2.13	2.27	0	20	1
Previous diagnostic tests	1.93	1.37	0	7	2
Previous consulted specialist	0.4	0.19	0	1	0
Duration prodrome before syncope (s)	41.09	116.23	0	720	2
Duration of loss of consciousness (min)	1.98	1.28	1	6	2
Duration of recovery (min)	8.24	25.41	0	180	0
Vital signs					
Systolic blood pressure (mmHg)	119.87	19.045	80	160	120
Diastolic blood pressure (mmHg)	76.76	10.472	50	110	80
Heart rate (beats per minute)	76.81	15.63	45	130	73.5
Left ventricular ejection fraction (%)	59.98	4.223	45	65	60

Syncopal event features, and vital signs of 55 patients. Continuous data are shown as mean (SD), minimum, maximum, and median; categorical data as counts.

Of the 48 minor high-risk features related to the syncopal event, thirteen met clinically relevant high-risk criteria (structural heart disease (n = 5) and high-risk ECG abnormalities (n = 8)). No cases fulfilled the minor ECG criteria of recurrent falls presumed to be arrhythmogenic in origin [23].

Diagnostic phases

Diagnostic yield

The diagnostic yield (i.e., the production of any diagnosis including “unexplained T-LOC”) of the syncope-expert was 94% for both the CE and EE assessments (Figure 2). The diagnostic yield of GPT-5 was 100% for the CE and 96% for the EE assessments (Figure 2). Diagnostic yield was prespecified as the production of any diagnostic label and was analyzed separately from final-diagnosis inclusion and the DPS, which are reported below.

Follow-up

In one case, the follow-up was not complete, leaving 54 cases available for assessment of the diagnostic performance.

Final diagnosis

The baseline characteristics of this cohort have been described in detail previously [4], the distribution of final diagnosis (n = 54) revealed reflex syncope (46%, n = 25), unexplained T-LOC (24%, n = 13), OH (20%, n = 11), cardiac syncope (7%, n = 4), and functional neurological disorder with apparent T-LOC (2%, n = 1).

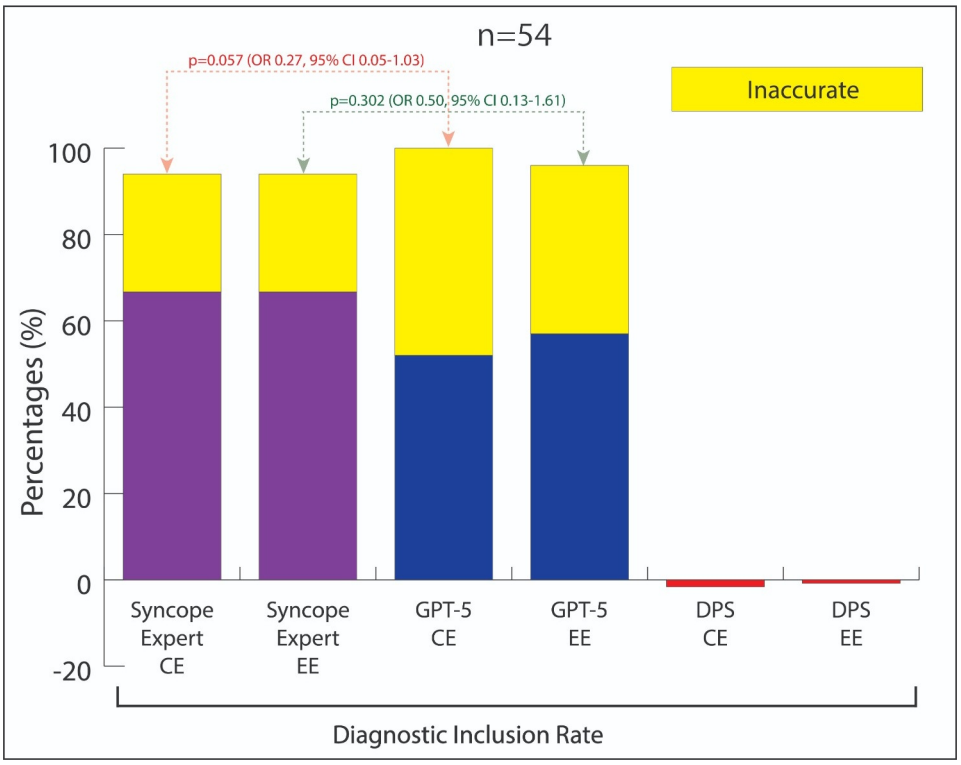

Figure 2. Diagnostic yield, inclusion rate and diagnostic precision score. The colored and yellow bars together represent the diagnostic yield of syncope-expert and GPT-5. The purple bars show the inclusion rate for syncope-expert, and the blue bars show the inclusion rate for GPT-5. The inclusion rate is defined as the proportion of final diagnosis that was included in the list of the differential diagnosis of GPT-5. The yellow bar: the proportion of diagnosis that did not included the final diagnosis. The red bar shows the diagnostic precision score: (%). CE: core evaluation; EE: extended evaluation.

Notably, three of the four final diagnoses adjudicated as cardiac syncope were arrhythmic in origin, while one was attributed to structural heart disease. Mean age (years) was 61 (SD 17; IQR 50–75) and 54.5% were female (n = 30).

Diagnostic inclusion rate (syncope-expert and GPT-5)

One case was lost to follow-up, leaving 54 cases for the analysis of the diagnostic inclusion rate and the DPS. The syncope-expert included the final diagnosis in 67% (n = 36) of cases in both the CE and EE protocols. GPT-5 included the final diagnosis in 52% (n = 28) of CE cases and 57% (n = 31) of EE cases (Figure 2). No significant difference between the syncope-expert and GPT-5 was observed in CE (exact McNemar p = 0.057) or EE (p = 0.302). In CE (n = 54), the syncope-expert alone included the final diagnosis in 11 cases, whereas GPT-5 was correct in three cases (matched-pairs OR 0.27, 95% CI 0.05–1.03); in EE (n = 54) these were 10 vs 5 (OR 0.50, 95% CI 0.13–1.61).

DPS (GPT-5)

The maximum number of diagnoses provided by GPT-5 was six. Consequently, the penalty-score for each incorrect diagnosis in the list of differential diagnoses was 0.20 (see for explanation also in Materials and methods section and Supplementary Table S1). Using this methodology, we calculated the DPS for every case. The cumulative DPS in the CE phase was −2% (mean −0.03, SD 0.54). After EE information, cumulative DPS was −1% (mean −0.01, SD 0.49) (see Figure 2).

Diagnostic safety (syncope-expert and GPT-5)

To quantify diagnostic safety, GPT-5 assigned a cardiac diagnosis in 17% (n = 9) of 54 cases after CE. This matched the final adjudicated diagnosis in 1 of 9 cases, corresponding to a positive predictive value of 11%. The remaining cases had a final diagnosis of unexplained T-LOC (44%, n = 4), OH (22%, n = 2), and reflex syncope (22%, n = 2) (Figure 3). After EE GPT-5 assigned a cardiac diagnosis in 9% (n = 5) of 54 cases. In 20% (1/5), this matched the final diagnosis, corresponding to a positive predictive value of 20%. The remaining cases had a final diagnosis of unexplained T-LOC (60%, n = 3) and OH (20%, n = 1) (Figure 3).

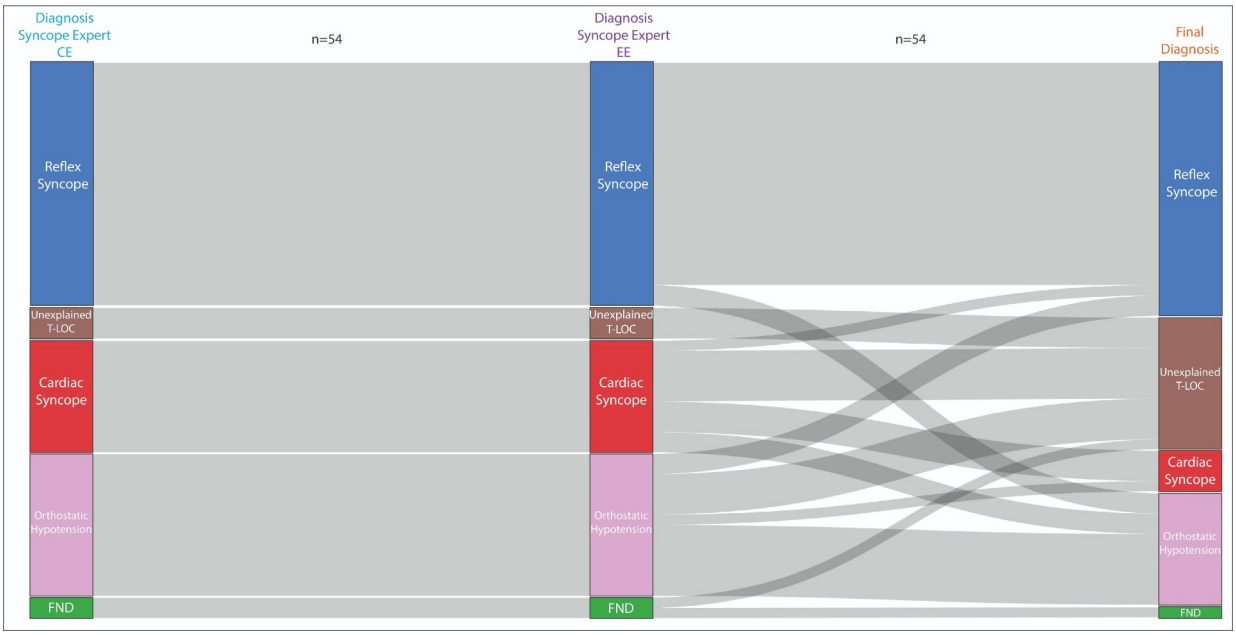

Figure 3. Diagnostic flowchart of the diagnosis from the syncope-expert compared to the final diagnosis. Diagnostic flowchart comparing the diagnostic classification of the syncope-expert with the final diagnosis. The left panel shows the diagnosis after the CE evaluation based on only history-taking, physical examination, ECG, and active standing test. The middle panel presents the EE diagnosis based on CE with additional testing (echocardiography, exercise stress testing, Holter monitoring, and tilt-table test). The right panel shows the final diagnosis. Lines indicate the diagnostic concordance or shift between diagnoses. FND: functional neurological disorder; CE: core evaluation: EE: extended evaluation.

The diagnoses established by the syncope-expert during the CE phase remained unchanged after the subsequent EE evaluation in all cases. A cardiac cause was assigned in 20% of cases (11/54). In 27% (3/11), this diagnosis matched the final diagnosis. In the remaining cases, the syncope-expert either withheld a diagnosis (46%, 5/11), misclassified the case as OH (18%, 2/11), or as reflex syncope (9%, 1/11) (Figure 4).

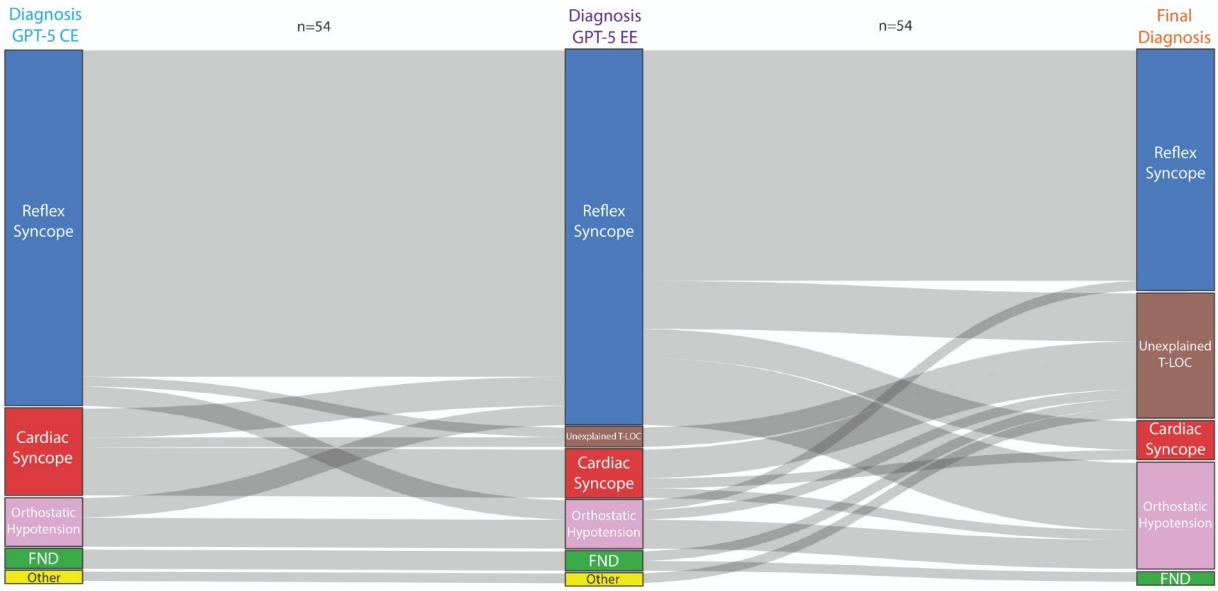

Figure 4. Diagnostic flowchart of GPT-5-based classification compared to the final diagnosis. Diagnostic flowchart comparing the GPT-5-based classification with the final diagnosis. The left panel displays GPT-5's diagnostic classification based on information of the clinical report of CE: containing only the guideline-based initial evaluation (i.e., history-taking, physical examination, ECG, and active standing test). The middle panel shows GPT-5's diagnostic classification based on information of the clinical report containing the guideline-based CE + additional testing: EE: Additional diagnostic tests included: echocardiography, exercise stress testing, Holter monitoring, and tilt-table testing. The right panel shows the final diagnosis. Lines indicate the diagnostic concordance or shift between diagnosis. FND: functional neurological disorder; CE: core evaluation: EE: extended evaluation.

Secondly, we examined how cases that were ultimately adjudicated as cardiac syncope were initially classified. Four of 54 patients had a final diagnosis of cardiac syncope. Of these four cases, GPT-5 classified 75% (3/4) as reflex syncope and 25% (1/4) as cardiac in both the CE and EE (Figure 3). The syncope-expert diagnosed 75% (3/4) of cases as cardiac syncope, and 25% (1/4) as OH (Figure 4) after both the CE and EE.

Diagnostic consistency

Across five GPT-5 runs per case on CE data, a single consistent diagnosis was provided in 69% (n = 37) of cases, while 30% (n = 16) yielded two different diagnoses, and 2% (n = 1) yielded three. In these 17 cases with discordant diagnoses among the different runs the different combinations of first diagnoses in the list were paired for each case: OH + reflex (47%, n = 8), cardiac + reflex (24%, n = 4), OH + cardiac (12%, n = 2), functional neurological disorder + reflex (6%, n = 1), traumatic + cardiac (6%, n = 1), and cardiac + OH + reflex (6%, n = 1). In total, a cardiac diagnosis was part of the differential diagnosis (after 5 runs) at least once in eight of the seventeen (47%) cases where GPT-5 was inconsistent, see Figure 5A.

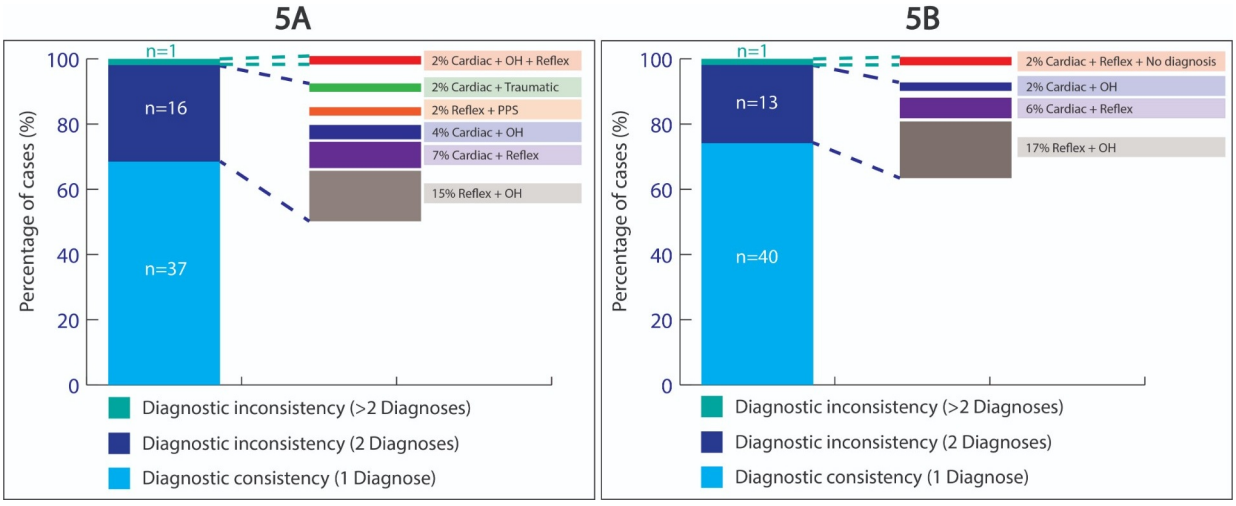

Figure 5. Diagnostic consistency. Within-case diagnostic consistency of GPT-5 across five independent runs per case in patients for CE (**A**) and for EE (**B**). In each panel, the stacked bar on the left shows the proportion of cases with 1, 2, or > 2 distinct diagnoses across the five runs, while dashed connectors from the inconsistent segments link to the corresponding diagnostic combinations on the right. All values are displayed as percentages with case counts. CE: core evaluation; EE: extended evaluation.

Across five GPT-5 runs on EE data (n = 54), a single consistent diagnosis was provided in 74% (n = 40), while 24% (n = 13) yielded two different diagnoses, and 2% (n = 1) yielded three. The combination of proposed diagnoses in these fourteen discordant cases was: OH + reflex (64%, n = 9), cardiac + reflex (21%, n = 3), OH + cardiac (7%, n = 1), and cardiac + reflex + unexplained T-LOC (7%, n = 1). In these 14 inconsistent cases, a cardiac diagnosis appeared at least once in the differential diagnosis during the five runs for five patients (36%), see Figure 5B.

Diagnostic consistency and agreement with the final diagnosis were summarized at two levels:

As shown in Figure 6, cases were categorized according to the number of GPT-5 runs (0–5/5) in which the final diagnosis was ranked first in the differential list.

Analysis of the CE phase revealed that the final diagnosis was never ranked first across the five runs in 35% of cases (n = 19), whereas it achieved consistent first-rank status in 41% of cases (n = 22). In the remaining 24% (n = 13) of cases, the final diagnosis was ranked first in all 1/5, 2/5, and 3/5 runs in 7% (n = 4), and 4/5 runs in 2% (n = 1) of cases, respectively. When all 270 CE runs were pooled, the final diagnosis occupied the first position in 51% (n = 138) of runs (Figure 6).

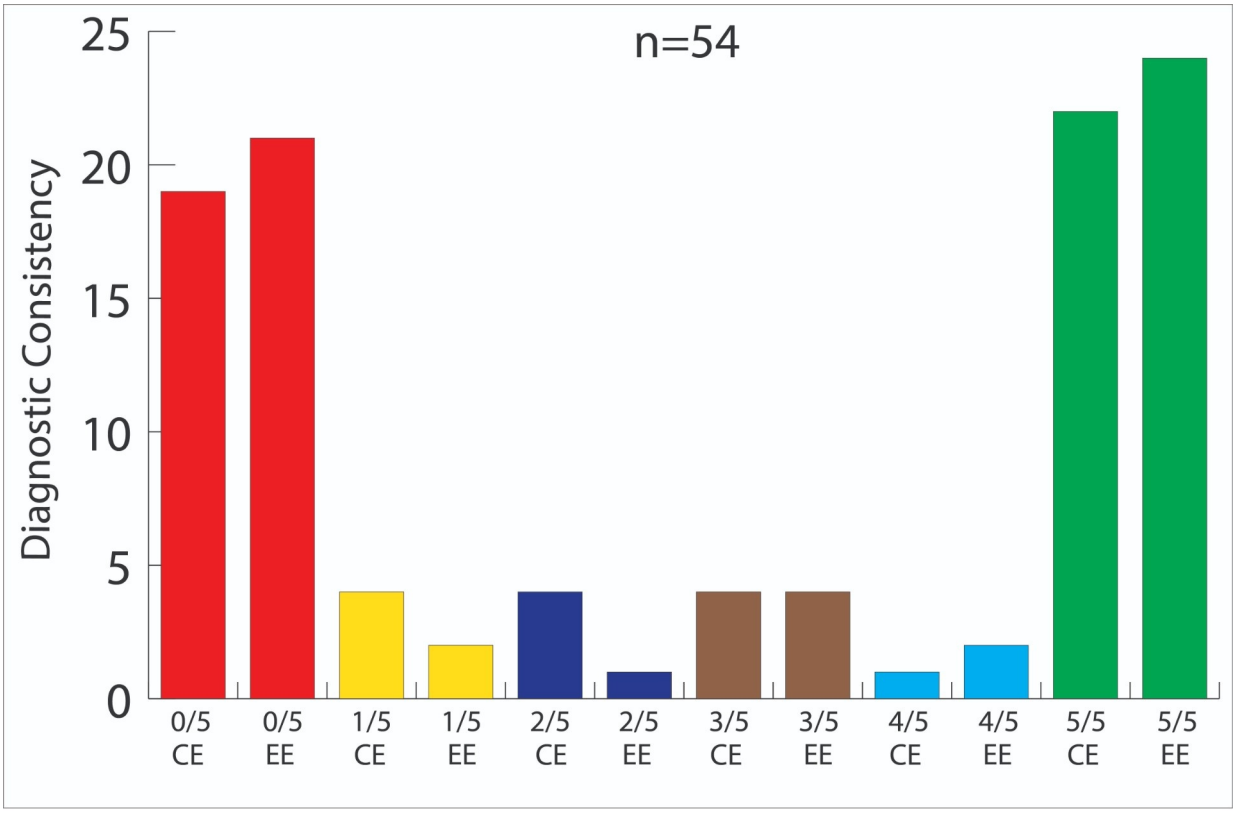

Figure 6. Reproducibility of GPT-5 diagnoses across multiple runs. Distribution of cases according to the number of five independent GPT-5 runs (all runs with identical clinical reports) in which the final diagnosis was ranked first in the differential diagnosis list. Bars illustrate whether the final diagnosis was never, in some, or in all runs ranked first, separately for the initial ESC guideline–based evaluation without additional testing (CE) and for the CE including additional diagnostic testing (EE). CE: core evaluation; EE: extended evaluation.

In the EE group, the final diagnosis was never ranked first (0/5) in 39% (n = 21) of cases and was ranked first in all five runs (5/5) in 44% (n = 24) of cases. In the remaining 17% (n = 9) of cases, the final diagnosis was ranked first in 1/5, 2/5, 3/5, and 4/5 runs in 4% (n = 2), 2% (n = 1), 7% (n = 4), and 4% (n = 2) of cases, respectively. Overall, across 270 EE runs, the final diagnosis was ranked first in 53% (n = 144) of runs (Figure 6).

Determinants of GPT-5 diagnostic inconsistency

In univariable logistic regression analyses, in the CE data: total peri-procedural factors (OR 3.24, 95% CI 1.20–8.69; P = 0.020) and fatigue (OR 11.20, 95% CI 1.28–97.71; P = 0.029) were significantly associated with GPT-5 inconsistency, while medical history (including risk-factors and medication) and total word count in the SU history showed a trend towards an association with GPT-5 inconsistency (OR 1.009, 95% CI 1.000–1.018; P = 0.059 and OR 1.008, 95% CI 1.000–1.015; P = 0.054, respectively) (Supplementary Table S2).

EE: pain after syncope (OR 0.15, 95% CI 0.03–0.82; P = 0.029) and injury related to syncope (OR 0.13, 95% CI 0.03–0.72; P = 0.019) were inversely associated with GPT-5 inconsistency, while total post-procedural factors and total peri-procedural factors showed trends towards inverse and positive associations with GPT-5 inconsistency, respectively (OR 0.78, 95% CI 0.61–1.00; P = 0.050 and OR 2.51, 95% CI 0.97–6.49; P = 0.058) Supplementary Table S2).

Discussion

This study evaluated the diagnostic performance of GPT-5 against a structured, guideline-based SU workflow by syncope-expert using CE and EE datasets.

Diagnostic performance

Diagnostic yield (defined as the production of any diagnosis, including "unexplained T-LOC") was reported as a measure of diagnostic completeness rather than correctness. GPT-5 yield was 100% after CE, and 96% after EE, versus 94% for the syncope-expert in both phases. Conversely, GPT-5 included the final diagnosis in 52% (CE) and 57% (EE) versus 67% for the syncope-expert in both phases. Outcomes are similar to previous research, where GPT achieved 54% diagnostic accuracy (here: diagnostic inclusion rate), versus 75% for specialists [27]. This comparison is contextual only, because GPT-4o was not re-tested on the present dataset under identical conditions. Therefore, the present study cannot determine whether GPT-5 provides incremental diagnostic improvement over GPT-4o. There was no statistically significant difference between GPT-5 and the syncope-expert, despite the syncope-expert demonstrating higher diagnostic precision than GPT-5.

Additional cardiac function tests syncope-expert

Among the 55 patients, additional testing (mainly TTE and exercise ECG) was frequent, but never changed the syncope-expert's initial diagnosis. Prior work from de Jong et al. [11] showed that 93% of final diagnoses were already established after the guideline initial evaluation. Additional testing never changed these diagnoses, underscoring that a diagnosis is predominantly determined by first-line clinical information rather than routine downstream testing [13]. In the present study, the syncope-expert's core diagnosis remained unchanged after TTE and exercise ECG, whereas GPT-5's performance showed modest improvement only after access to additional clinical information. This is consistent with the relatively favorable cardiac health profile of the cohort (Table 2), implying a low pre-test probability for clinically relevant structural heart disease. Consequently, routine TTE and exercise ECG provided no incremental diagnostic yield beyond the CE, emphasizing that additional investigations should be reserved for clearly indicated scenarios rather than performed to "complete" the diagnostic work-up. This is in accordance with the ESC syncope guidelines [1], supports indication driven testing [28–32]. We conclude that thorough history-taking, physical examination, active standing test and ECG remain the cornerstone of T-LOC assessment, as also emphasized in the ESC syncope guidelines [1].

Risk-stratification and diagnostic safety

Across all 55 cases evaluated in the SU, documentation of the core domains was complete and standardized (event history, physical examination including vital signs and active standing test, relevant medical history/medication, and ECG), such that every patient met the minimum prerequisites for a valid and reproducible risk-assessment [1, 2, 13, 22, 24, 33–35]. This contrasts with our earlier work in which referrals from primary care frequently lacked essential baseline information, highlighting substantial scope for improvement in primary care through a systematic CE and structured reporting. For secondary care, these results emphasize the need to further standardize SU practice (uniform intake proforma and explicit ECG interpretation), to transparently link risk-features to management decisions (admission vs outpatient care and targeted downstream testing), and to provide consistent feedback to primary caretakers to improve the quality of future referrals and preserve SU capacity for patients in whom specialist assessment offers true incremental value [12, 36–38].

Despite all data being available, safety diverged on cardiac syncope diagnosis detection: GPT-5 diagnosed a cardiac cause in only (1/4), whereas the syncope-expert scored (3/4). GPT-5 misclassifications skewed to the more benign reflex or OH, risking false reassurance. Additional EE data did not improve discrimination, indicating that limitations in reasoning rather than lack of appropriate data caused this pattern [4]. The syncope-expert's willingness to defer a diagnosis in the case of clinical uncertainty likely functions as a safety buffer against early closure. Obviously, missing cardiac causes of T-LOC may induce risks of harm that potentially could have been prevented [39–41].

Effect of additional diagnostic testing on diagnostic performance of GPT-5

More comprehensive clinical data (thorough history-taking, active standing test, ECG and standard additional diagnostic tests after guideline-based CE) only modestly improved GPT-5's diagnostic performance (52% to 57%), suggesting that complete data alone are insufficient for diagnostic evaluation, without mechanisms that embed pathophysiological priors or calibrated uncertainty. Likewise, targeted instruction by prompting GPT-5 with contemporary syncope guidelines [1] and web practical instruction [24] yielded only limited gains in diagnostic performance, underscoring again that access to codified knowledge alone does not replace internalized clinical reasoning [4].

DPS, interpretability and trustworthiness of LLM outputs

In contrast to the clearly positive DPS achieved by human clinicians in the same dataset [4], GPT-5 remained negative (near zero) in both phases, mainly due to penalties because of a large number of differential diagnoses [42]. We speculate that GPT-5's differential diagnoses approximate an educated guess, with substantial probability retained on incorrect options. LLMs generate comprehensive differentials while their internal ranking/probability structure is not transparent and may be poorly calibrated [42–44]. This is consistent with previous research [4]. Although GPT-5 can generate plausible diagnostic explanations, these should be interpreted as probabilistic post-hoc narratives rather than as evidence of traceable causal or clinical pathophysiological reasoning. This distinction is clinically relevant, as prior vignette-based work with GPT-4 showed that access to an LLM did not necessarily improve physicians' diagnostic reasoning, despite superior standalone model performance [20]. The explanatory depth required for clinical use is task dependent. For screening, triage, or documentation-support tasks, transparent feature ranking or identification of missing information may be sufficient. In contrast, diagnostic classification of T-LOC and treatment-directed decisions, particularly when cardiac syncope is possible, require causal and pathophysiological reasoning, calibrated uncertainty, and the ability to defer a diagnosis when the available evidence is insufficient. We argue that any clinical use of GPT-5 would require auditable evidence trails, calibrated confidence estimates, and an explicit capacity to abstain or defer a definite diagnosis in high-risk cases when multiple options are still open [45].

Implications for LLM integration in clinical pathways

From the above, it follows that to date autonomous GPT-5 use in syncope pathways is not yet justified. The speed, breadth, and documentation inside clinician-led, protocol-bound support by LLM models with guardrails (e.g., mandatory escalation on cardiac flags or low confidence) could be employed within clinical decision making. We suggest that the use of LLM models should be restricted to the support of human clinical decision making. An example of pragmatic use of LLMs' could be a role to improve clinician diagnostic reasoning, triage or documentation augmentation with human oversight and verification for high-risk phenotypes [26, 46]. In addition, practical implementation depends not only on the availability of explanations, but also on their timing and cognitive burden. In time-sensitive clinical settings, explanations that are generated too late, or that require extensive clinician interpretation, may have limited practical value. Similarly, overly long or non-prioritized explanation reports may increase cognitive load and obscure the key diagnostic uncertainty or safety signal. Therefore, LLM-generated explanations should be concise, prioritized, and actionable, highlighting the information that changes risk assessment or management rather than providing exhaustive narrative justification.

Fivefold reanalysis

The fivefold reanalysis extends the primary single-output evaluation by assessing whether GPT-5 produced stable diagnostic conclusions when identical clinical information was presented repeatedly. Across five re-analyses per case, EE resulted in limited improvement: a fully correct first diagnosis was established in 41% after one run and increased to 44% after multiple runs. The presence of the final diagnosis anywhere in the five runs increased from 51% to 53%, insufficient for dependable, patient-level decisions. In the clinically important subgroup with final adjudicated cardiac syncope, GPT-5 selected the final diagnosis in 1

of 4 patients, whereas the syncope-expert did so in 3 of 4 patients. Given the very small number of final cardiac syncope cases, this observation should not be interpreted as a precise estimate of diagnostic sensitivity or safety. Conversely, the safety analysis also showed false-positive cardiac classification: after CE, only 1 of 9 GPT-5 cardiac classifications matched the final adjudicated diagnosis, and after EE, only 1 of 5 matched the final adjudicated diagnosis. Thus, the safety limitation was bidirectional, comprising both under-recognition of final cardiac syncope and over-classification of cardiac etiology in patients with non-cardiac final diagnoses. These findings reinforce the need for clinician oversight, particularly for potentially high-risk cardiac etiology, and do not support the use of GPT-5 as a stand-alone diagnostic tool in syncope evaluation.

Within-case reliability (fivefold reanalysis)

Only 69% of the assessments by GPT-5 in the CE phase and 74% in the EE phase were stable across five runs [47]. Between 16% and 32% of diagnoses varied despite identical inputs, putting serious limitations on single-run outputs. Discordant outputs clustered along hemodynamic phenotypes (orthostatic vs reflex), indicating that the boundaries used by the model are not strict and have limited pathophysiological basis.

The generation of multiple, internally inconsistent diagnosis by GPT-5 is inherently unsafe. Diagnosing the same event simultaneously as a benign (e.g., orthostatic or reflex) and as high-risk cardiac syncope invites anchoring to the reassuring explanation, diminishes the perceived urgency of the "cardiac" signal, and may lead to inappropriate de-escalation of monitoring and work-up.

The particular implication of the above is that the syncope-expert currently outperforms AI for safety-critical detection and governance. Therefore, it is crucial to maintain human oversight, use multi-run consensus, escalate on any cardiac "hit" ("one-vote-cardiac"), and abstain when unanimity is absent [45, 48]. Until sensitivity and stability improve, the present findings support the use of GPT-5 only as supervised decision support within syncope-expert-led pathways.

From univariable logistic regression analyses of predefined clinical determinants, GPT-5's diagnostic behavior appeared to vary systematically with the clinical context of the case and with the amount of structured clinical information available at the time of assessment. The overall burden of documented determinants at presentation was significantly associated with the model's diagnostic output, whereas determinants recorded after the event showed only a borderline association. Taken together, these findings suggest that GPT-5's diagnostic output was more strongly influenced by syncopal episodes described with numerous or more severe symptoms, particularly when multiple clinical factors were present at the time of syncope. In other words, GPT-5 seems to add more value in complex, highly symptomatic presentations than in cases with only few or mild symptoms.

Temperature setting LLM

In this study, we verified that the model was operating with a low temperature setting, which minimizes stochastic variability and promotes reproducible outputs across repeated runs for the same case, and therefore did not adjust this parameter. In LLMs, the temperature controls randomness in token sampling: lower values constrain the model to high-probability continuations and yield more deterministic responses, whereas higher values increase variability of the output given the same input [49, 50].

Strengths and limitations

The principal strength of this analysis is the head-to-head comparison with a syncope-expert, GPT-5, and adjudication, across the predefined CE and EE phases. Because syncope mechanisms are rarely captured with simultaneous ECG and blood pressure at the time of T-LOC, we used a final diagnosis after structured follow-up as the best available reference standard, acknowledging that some cases will remain unexplained despite extensive evaluation. Although the final reference diagnosis was established by an independent expert panel using all available follow-up data, the study was performed in a single-center syncope-unit cohort and used one experienced syncope-expert as the clinical comparator. The adjudication process strengthens the internal validity of the reference standard and reduces the risk of confirmation bias, but it

does not fully address external generalizability. Diagnostic performance may differ across centers, referral pathways, patient case mix, and clinician expertise. Therefore, external validation in larger multicenter cohorts, preferably including multiple independent syncope-experts, is required before these findings can be generalized to broader syncope-care settings. The small number of cardiac outcomes limits precision for safety-critical estimates; however, this mirrors real-world prevalence and probably preserves external validity. Second, we did not perform formal test–retest assessment of the syncope-expert index rater, precluding direct comparison of within-rater variability between the syncope-expert and GPT-5. Third, although a single GPT-5 run per case was prespecified, an inadvertent repeat run in one case yielded discrepant output, underscoring potential stochastic variability in LLM-based classification and the need for prespecified strategies (e.g., deterministic settings or multi-run consensus) in future evaluations. Accordingly, we performed fivefold re-analysis, demonstrating within-case variability that would not be detected in single-run designs. GPT-5 narrows the diagnostic performance gap with expert SU workflows on aggregate metrics yet underperforms where safety matters most, that is, the case of cardiac syncope. Diagnostic safety in this study was operationalized as identification of cardiac syncope; we did not assess potential harm from overdiagnosis or unnecessary management triggered by incorrect non-cardiac classifications. Moreover, GPT-5 could not ask for more information and depended solely on the report provided by the syncope-expert. Thereby, the full capabilities of AI may not have been evaluated. Future work should prioritize hybrid architectures that fuse LLMs' outputs with pathophysiological inference, calibrated uncertainty and abstention, and prospectively validated escalation rules for cardiac red-flags. Larger, diverse cohorts and multi-run evaluation protocols with the use of diagnostic adjudication in order to test real life diagnostic performance or accuracy will be essential to quantify variance and to certify dependable behavior at patient level.

Conclusion

GPT-5 demonstrated a high diagnostic yield and diagnostic performance approaching that of the syncope-expert in parts of the structured diagnostic pathway. However, run-to-run inconsistency and the exploratory observation in the small cardiac syncope subgroup underscore the need for clinician oversight, particularly for safety-critical diagnoses. Structured, clinician-led reasoning remains essential in syncope evaluation, especially when cardiac or otherwise high-risk causes must be considered. In our cohort, the addition of TTE and exercise ECG data did not materially improve GPT-5 diagnostic refinement, supporting the continued primacy of guideline-based CE: careful history-taking, physical examination, active standing testing, and 12-lead ECG. Despite the rapid evolution of LLMs, their role in syncope care should remain supportive within clinician-led pathways rather than determinative, unless future systems demonstrate calibrated uncertainty, reliable abstention, stable multi-run consensus, and clinically grounded pathophysiological reasoning.

Abbreviations

AI: artificial intelligence

CE: core evaluation (guideline-based thorough history, physical examination, 12-lead electrocardiogram and active standing test)

CI: confidence interval

DPS: diagnostic precision score

ECG: 12-lead electrocardiogram

EE: extended evaluation (CE + additional diagnostic tests: echocardiogram, exercise stress testing, Holter, tilt test)

GPT-4o: Generative Pre-trained Transformer-4 omni

GPT-5: Generative Pre-trained Transformer-5

LLM: Large Language Model

OH: orthostatic hypotension

OR: odds ratio

SUs: syncope units

T-LOC: transient loss of consciousness

TTE: transthoracic echocardiography

Supplementary materials

The supplementary materials for this article are available at: https://www.explorationpub.com/uploads/Article/file/1012114_sup_1.pdf.

Declarations

Author contributions

SvZ: Conceptualization, Investigation, Writing—original draft, Writing—review & editing. TTB: Writing—review & editing. JSYdJ: Supervision, Writing—review & editing. EMK: Writing—review & editing. BB: Writing—review & editing. AD: Writing—review & editing. FG: Writing—review & editing. CG: Writing—review & editing. RS: Writing—review & editing. MGS: Writing—review & editing. JRdG: Writing—review & editing. FJdL: Supervision, Writing—review & editing. All authors read and approved the submitted version.

Conflicts of interest

The authors declare that they have no conflicts of interest.

Ethical approval

The study protocol was approved by the Medical Ethics Review Committee Leiden/The Hague and Delft (METC nr: 18-061) and Amsterdam Ethics Committee (nr: 2024.0124), and complies with the Declaration of Helsinki, including General Data Protection Regulation compliance and provisions for human oversight.

Consent to participate

Informed consent to participate in the study was obtained from all participants.

Consent to publication

Not applicable.

Availability of data and materials

The data of this manuscript could be available from the corresponding authors upon reasonable request.

Funding

Not applicable.

Publisher's note

References

1. Brignole M, Moya A, de Lange FJ, Deharo JC, Elliott PM, Fanciulli A, et al. 2018 ESC Guidelines for the diagnosis and management of syncope. Eur Heart J. 2018;39:1883–948. [DOI]
2. Kenny RA, Brignole M, Dan GA, Deharo JC, van Dijk JG, Doherty C, et al. Syncope Unit: rationale and requirement – the European Heart Rhythm Association position statement endorsed by the Heart Rhythm Society. Europace. 2015;17:1325–40. [DOI] [PubMed]
3. van Zanten S, de Jong JSY, Scheffer MG, Kaal ECA, de Groot JR, de Lange FJ. A cross-sectional nationwide survey of guideline based syncope units in the Netherlands: the SU-19 score—a novel validation for best practices. Europace. 2024;26:euae002. [DOI] [PubMed] [PMC]
4. van Zanten S, Boel TT, de Jong JS, Bais B, Fedorowski A, Sutton R, et al. Clinical Decision-Making of Artificial Intelligence vs Medical Professionals in Patients With Syncope. JACC Adv. 2025;5:102426. [DOI]
5. Lee S, Reddy Mudireddy A, Kumar Pasupula D, Adhaduk M, Barsotti EJ, Sonka M, et al. Novel Machine Learning Approach to Predict and Personalize Length of Stay for Patients Admitted with Syncope from the Emergency Department. J Pers Med. 2022;13:7. [DOI] [PubMed] [PMC]
6. Olshansky B, Gebska MA, Johnston SL. Syncope—Do We Need AI? J Pers Med. 2023;13:740. [DOI] [PubMed] [PMC]
7. Statz GM, Evans AZ, Johnston SL, Adhaduk M, Mudireddy AR, Sonka M, et al. Can Artificial Intelligence Enhance Syncope Management? JACC: Adv. 2023;2:100323. [DOI] [PubMed] [PMC]
8. Dipaola F, Gebska MA, Gatti M, Levra AG, Parker WH, Menè R, et al. Will Artificial Intelligence Be "Better" Than Humans in the Management of Syncope? JACC Adv. 2024;3:101072. [DOI] [PubMed] [PMC]
9. Aamir A, Jamil Y, Bilal M, Diwan M, Nashwan AJ, Ullah I. Artificial Intelligence in Enhancing Syncope Management - An Update. Curr Probl Cardiol. 2024;49:102079. [DOI] [PubMed]
10. Levra AG, Gatti M, Mene R, Shiffer D, Costantino G, Solbiati M, et al. A large language model-based clinical decision support system for syncope recognition in the emergency department: A framework for clinical workflow integration. Eur J Intern Med. 2025;131:113–20. [DOI] [PubMed]
11. de Jong JSY, van Zanten S, Thijs RD, van Rossum IA, Harms MPM, de Groot JR, et al. Syncope Diagnosis at Referral to a Tertiary Syncope Unit: An in-Depth Analysis of the FAST II. J Clin Med. 2023;12:2562. [DOI] [PubMed] [PMC]
12. van Wijnen VK, Gans ROB, Wieling W, Ter Maaten JC, Harms MPM. Diagnostic accuracy of evaluation of suspected syncope in the emergency department: usual practice vs. ESC guidelines. BMC Emerg Med. 2020;20:59. [DOI] [PubMed] [PMC]
13. van Dijk N, Boer KR, Colman N, Bakker A, Stam J, van Grieken JJ, et al. High Diagnostic Yield and Accuracy of History, Physical Examination, and ECG in Patients with Transient Loss of Consciousness in FAST: The Fainting Assessment Study. J Cardiovasc Electrophysiol. 2007;19:48–55. [DOI] [PubMed]
14. Bertens LC, Broekhuizen BD, Naaktgeboren CA, Rutten FH, Hoes AW, van Mourik Y, et al. Use of Expert Panels to Define the Reference Standard in Diagnostic Research: A Systematic Review of Published Methods and Reporting. PLoS Med. 2013;10:e1001531. [DOI] [PubMed] [PMC]
15. Luo Y, Miao Y, Zhao Y, Li J, Wu Y. Exploring the Current Applications and Effectiveness of ChatGPT in Nursing: An Integrative Review. J Adv Nurs. 2025;81:3473–84. [DOI]
16. Li J, Dada A, Puladi B, Kleesiek J, Egger J. ChatGPT in healthcare: A taxonomy and systematic review. Comput Methods Programs Biomed. 2024;245:108013. [DOI] [PubMed]
17. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med. 2023;29:1930–40. [DOI] [PubMed]
18. Meng X, Yan X, Zhang K, Liu D, Cui X, Yang Y, et al. The application of large language models in medicine: A scoping review. iScience. 2024;27:109713. [DOI] [PubMed] [PMC]

19. Yu E, Chu X, Zhang W, Meng X, Yang Y, Ji X, et al. Large Language Models in Medicine: Applications, Challenges, and Future Directions. Int J Med Sci. 2025;22:2792–801.
20. Goh E, Gallo R, Hom J, Strong E, Weng Y, Kerman H, et al. Influence of a Large Language Model on Diagnostic Reasoning: A Randomized Clinical Vignette Study. JAMA Netw Open. 2024;7:e2440969. [DOI] [PubMed] [PMC]
21. OpenAI. GPT-4o Technical Report. San Francisco, CA: OpenAI; 2024.
22. de Jong JSY, Blok MRS, Thijs RD, Harms MPM, Hemels MEW, de Groot JR, et al. Diagnostic yield and accuracy in a tertiary referral syncope unit validating the ESC guideline on syncope: a prospective cohort study. EP Eur. 2020;23:797–805. [DOI] [PubMed] [PMC]
23. Boel TT, Peeters SYG, van Alem AP, Boogers JMJ, Bootsma M, van den Dorpel MA, et al. Design and rationale of the RISC-Trial: multicenter RCT to assess immediate discharge of syncope patients admitted to the (cardiac) emergency room. Eur Heart J. 2026;27:euaf085.683.
24. Brignole M, Moya A, de Lange FJ, Deharo JC, Elliott PM, Fanciulli A, et al. Practical Instructions for the 2018 ESC Guidelines for the diagnosis and management of syncope. Eur Heart J. 2018;39:e43–80. [DOI] [PubMed]
25. Harada Y, Suzuki T, Harada T, Sakamoto T, Ishizuka K, Miyagami T, et al. Performance evaluation of ChatGPT in detecting diagnostic errors and their contributing factors: an analysis of 545 case reports of diagnostic errors. BMJ Open Qual. 2024;13:e002654. [DOI] [PubMed] [PMC]
26. Tu T, Schaekermann M, Palepu A, Saab K, Freyberg J, Tanno R, et al. Towards conversational diagnostic artificial intelligence. Nature. 2025;642:442–50. [DOI] [PubMed] [PMC]
27. Maiorana NV, Marceglia S, Treddenti M, Tosi M, Guidetti M, Creta MF, et al. Large Language Models in Neurological Practice: Real-World Study. J Med Internet Res. 2025;27:e73212. [DOI] [PubMed] [PMC]
28. Chiu DT, Shapiro NI, Sun BC, Mottley JL, Grossman SA. Are Echocardiography, Telemetry, Ambulatory Electrocardiography Monitoring, and Cardiac Enzymes in Emergency Department Patients Presenting with Syncope Useful Tests? A Preliminary Investigation. J Emerg Med. 2014;47:113–8. [DOI] [PubMed]
29. Bazan V, Cediel G, Llibre C, Sarrias A, Romeo I, Ibars S, et al. Contemporary Yield of 24-hour Holter Monitoring: Role of Inter-Atrial Block Recognition. J Atr Fibrillation. 2019;12:2225.
30. Han SK, Yeom SR, Lee SH, Park SC, Kim HB, Cho YM, et al. Transthoracic echocardiogram in syncope patients with normal initial evaluation. Am J Emerg Med. 2017;35:281–4. [DOI] [PubMed]
31. Madeira CL, Craig MJ, Donohoe A, Stephens JR. Things We Do For No Reason: Echocardiogram in Unselected Patients with Syncope. J Hosp Med. 2017;12:984–8. [DOI] [PubMed]
32. Sun BC, Emond JA, Camargo CA Jr. Direct medical costs of syncope-related hospitalizations in the United States. Am J Cardiol. 2005;95:668–71. [DOI] [PubMed]
33. Croci F, Brignole M, Alboni P, Menozzi C, Raviele A, Del Rosso A, et al. The application of a standardized strategy of evaluation in patients with syncope referred to three syncope units. Europace. 2002;4:351–5. [DOI] [PubMed]
34. Sarasin FP, Pruvot E, Louis-Simonet M, Hügli OW, Sztajzel JM, Schläpfer J, et al. Stepwise evaluation of syncope: A prospective population-based controlled study. Int J Cardiol. 2008;127:103–11. [DOI] [PubMed]
35. Boel T, Peeters SYG, Hemels MEW, Samim M, Koomen EM, Houtgraaf J, et al. Rationale and design of the RISC trial: a multicenter RCT to assess in hospital 24 hour observation with telemetry of syncope patients admitted to the cardiac emergency room and emergency department. Europace. 2025;27: e27. [DOI]
36. Firouzbakht T, Shen ML, Groppelli A, Brignole M, Shen WK. Step-by-step guide to creating the best syncope units: From combined United States and European experiences. Auton Neurosci. 2022;239: 102950. [DOI] [PubMed]

37. Anderson TS, Thombley R, Dudley RA, Lin GA. Trends in Hospitalization, Readmission, and Diagnostic Testing of Patients Presenting to the Emergency Department With Syncope. Ann Emerg Med. 2018;72: 523–32. [DOI] [PubMed]
38. Mazzella AJ, Wood BS, Doad J, Hendrickson MJ, Rosman L, Gehi AK. Interhospital variability in hospital admissions for patients with low-risk syncope presenting to the emergency department. Heart Rhythm O2. 2024;5:435–42. [DOI] [PubMed] [PMC]
39. Soteriades ES, Evans JC, Larson MG, Chen MH, Chen L, Benjamin EJ, et al. Incidence and Prognosis of Syncope. N Engl J Med. 2002;347:878–85. [DOI] [PubMed]
40. Koene RJ, Adkisson WO, Benditt DG. Syncope and the risk of sudden cardiac death: Evaluation, management, and prevention. J Arrhythmia. 2017;33:533–44. [DOI] [PubMed] [PMC]
41. Toarta C, Mukarram M, Arcot K, Kim SM, Gaudet S, Sivilotti MLA, et al. Syncope Prognosis Based on Emergency Department Diagnosis: A Prospective Cohort Study. Acad Emerg Med. 2018;25:388–96. [DOI] [PubMed]
42. Bridges JM. Computerized diagnostic decision support systems – a comparative performance study of Isabel Pro vs. ChatGPT4. Diagnosis. 2024;11:250–8. [DOI] [PubMed]
43. Savage T, Wang J, Gallo R, Boukil A, Patel V, Safavi-Naini SAA, et al. Large language model uncertainty proxies: discrimination and calibration for medical diagnosis and treatment. J Am Med Inform Assoc. 2024;32:139–49. [DOI] [PubMed] [PMC]
44. de Oliveira R, Garber M, Gwinnutt JM, Rashidi E, Hwang JS, Gilmour W, et al. A study of calibration as a measurement of trustworthiness of large language models in biomedical natural language processing. JAMIA Open. 2025;8:ooaf058. [DOI]
45. Socrates V, Wright DS, Huang T, Fereydooni S, Dien C, Chi L, et al. Identifying Deprescribing Opportunities With Large Language Models in Older Adults: Retrospective Cohort Study. JMIR Aging. 2025;8:e69504. [DOI] [PubMed] [PMC]
46. McDuff D, Schaekermann M, Tu T, Palepu A, Wang A, Garrison J, et al. Towards accurate differential diagnosis with large language models. Nature. 2025;642:451–7. [DOI] [PubMed] [PMC]
47. Yang H, Hu M, Most A, Hawkins WA, Murray B, Smith SE, et al. Evaluating accuracy and reproducibility of large language model performance on critical care assessments in pharmacy education. Front Artif Intell. 2025;7:1514896. [DOI] [PubMed] [PMC]
48. Lucas MM, Yang J, Pomeroy JK, Yang CC. Reasoning with large language models for medical question answering. J Am Med Inform Assoc. 2024;31:1964–75. [DOI] [PubMed] [PMC]
49. Windisch P, Dennstädt F, Koechli C, Schröder C, Aebersold DM, Förster R, et al. The Impact of Temperature on Extracting Information From Clinical Trial Publications Using Large Language Models. Cureus. 2024;16:e75748. [DOI] [PubMed] [PMC]
50. Jarrett PC, Hill J, Howell M, Grabow Moore K, Thoppil JJ, Vargas Ortiz L, et al. Piloting Temperature-Driven Variability in Emergency Diagnostic Accuracy Using a Leading Large Language Model. Cureus. 2025;17:e94476. [DOI] [PubMed] [PMC]