Evaluating the Performance of a ChatGPT Model in Rheumatology Exams

Fadi Hassan MD[1,2]*, Basem Hijazi MD[2]*, and Mohammad E. Naffaa MD[1,2]

[1]Department of Rheumatology, Galilee Medical Center, Nahariya, Israel
[2]Azrieli Faculty of Medicine, Bar-Ilan University, Safed, Israel

*These authors contributed equally to this study.

ABSTRACT **Background:** Large language models (LLMs) are rapidly advancing, with the potential to improve healthcare. While LLM performance on medical licensing exams were studied extensively, their performance in rheumatology exams requires specific evaluation.

Objectives: To assess Chat Generative Pre-trained Transformer (ChatGPT) performance on 200 validated Israeli rheumatology board exam questions.

Methods: ChatGPT performance was evaluated using 200 multiple-choice questions from the 2023 and 2024 Israeli official rheumatology board examinations. Three gpt-4-turbo based variants were assessed: base model (Model 1), few-shot chain of thought (CoT) model (Model 2), and knowledge-augmented prompting model incorporating rheumatology guidelines (Model 3). Model 1 was assessed using both the original Hebrew and a validated English translated version, while Models 2 and 3 were assessed using the English version only.

Results: Overall, Model 3 achieved the highest numerical accuracy (81%), followed by Model 1 (English, 77%), Model 2 (75%), and Model 1 (Hebrew, 74.5%); however, these differences were not statistically significant. Performance varied markedly by question type. For text-only questions (n=177), accuracies ranged from 78.5% to 83.1%, with Model 3 showing the highest point estimate (83.1%). In contrast, all models demonstrated substantially lower performance on questions that included images (n=23), with accuracies ranging from 34.8% to 65.2%. Model 3 yielded the highest numerical accuracy (65.2%).

Conclusions: The study highlights the potential role of LLMs in rheumatology board examinations but also emphasizes their critical limitations. Future research should focus on addressing limitations, especially image interpretation and management of complex cases to enable efficient application of LLMs in rheumatology.

IMAJ 2026; 28: 162–167

KEY WORDS: artificial intelligence (AI), board examination, ChatGPT, multiple choice questions, rheumatology

The rapid advancement of large language models (LLMs) holds substantial promise for improving patient care and clinical decision making [1]. Recent studies have indicated LLMs can compete with, or even surpass, human physicians in tasks like diagnostic reasoning, treatment selection, and medical knowledge retrieval, highlighting their potential as an integral component of clinical practice [2,3].

LLMs have demonstrated a capability to pass rigorous medical licensing examinations across specialties, sometimes outperforming physician test-takers [4-9]. However, significant variability in performance exists across specialties, studies, and countries [4].

Rheumatology presents unique challenges for LLMs due to its diverse range of conditions affecting multiple systems and the common occurrence of overlapping, subtle signs and symptoms. Diagnosis relies on integrating patient history, physical findings, laboratory/serological tests, and imaging often require a clinician's knowledge, experience, and intuition in the absence of definitive tests [10-12].

Despite growing interest, LLMs' clinical reasoning capabilities within the complexities of rheumatology have not yet been thoroughly assessed. Previous studies such as Madrid-Garcia and colleagues [13] reported a 93.71% accuracy for GPT-4 on 143 rheumatology questions from the Spanish specialized medical training examination. Daungsupawong et al. [14] found an 86.9% accuracy for ChatGPT on 420 board-level questions from online question banks. More recently, Omar and co-authors [15] conducted a multimodal performance analysis, demonstrating that even state-of-the-art models like GPT-4o and Claude Sonnet 3.5 exhibit marked variability and substantial limitations in visual interpretation tasks central to rheumatologic diagnosis. Because examinations vary considerably in format, style, and content emphasis across regions, further examination of LLM performance utilizing other validated question sets in this complex field is still necessary.

In the present study, we evaluated ChatGPT's performance in answering 200 validated rheumatologic questions from the Israeli board examination in rheumatology.

PATIENTS AND METHODS

STUDY DESIGN AND OVERVIEW

In this study, we assessed the performance of three distinct LLM variants based on the multimodal OpenAI gpt-4-turbo-2024-04-09 architecture, accepting both text and images, in answering questions from two full Israeli board examinations in rheumatology (2023 and 2024). The core 200-question dataset comprised single-best-answer multiple-choice questions. The exams were authored by board-certified rheumatologists, required a passing score ≥ 65, and predominantly featured clinically oriented case vignettes assessing diagnosis, management, and decision-making. Approximately 10–20% of questions include images (physical exam, imaging). All questions from the official examinations were used without any exclusions. Passing the exam is obligatory for board certification in Israel. The study objectives were to assess the models' accuracy in answering multiple-choice questions and to explore performance enhancements using few-shot chain of thought (CoT) and knowledge augmentation.

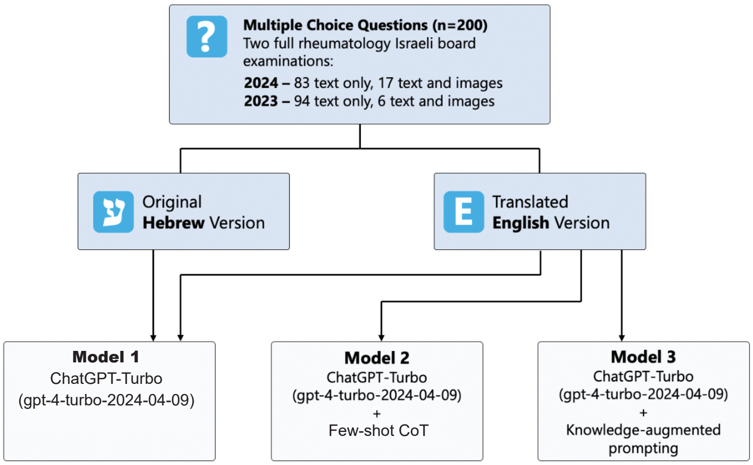
MODEL CONFIGURATIONS AND PROMPTING STRATEGIES

We examined three distinct LLM variants based on the same multimodal gpt-4-turbo-2024-04-09 architecture. All modifications were applied only at the prompt level [Figure 1]:

- *Model 1 (base model):* The variant is the unaltered base version (gpt-4-turbo-2024-04-09 model) that was used without any modifications to serve as the benchmark based on its pre-trained knowledge.
- *Model 2 (few-shot CoT):* This model incorporated few-shot CoT reasoning using five example questions and answers, including the process of reasoning that led to the correct answer. These examples were taken from previous examinations not including either the 2023 or 2024 examination sets. The exemplars were provided once at the beginning of the session, and the same prompt was used throughout. Few-shot CoT reasoning was used to enhance the model's ability to structure its reasoning in a more logical and systematic manner as physician test-takers. This method allows the model to generate intermediate reasoning steps as typically employed by clinicians.

Figure 1. Study methodology

ChatGPT = Chat generative pre-trained transformer, CoT = chain of thought


Multiple Choice Questions (n=200)
Two full rheumatology Israeli board examinations:
2024 – 83 text only, 17 text and images
2023 – 94 text only, 6 text and images
Original Hebrew Version
Translated English Version
Model 1
ChatGPT-Turbo
(gpt-4-turbo-2024-04-09)
Model 2
ChatGPT-Turbo
(gpt-4-turbo-2024-04-09)
+
Few-shot CoT
Model 3
ChatGPT-Turbo
(gpt-4-turbo-2024-04-09)
+
Knowledge-augmented prompting

- *Model 3 (rheumatologic knowledge):* We used a knowledge-augmented prompting approach, which is conceptually similar to retrieval-augmented generation (RAG) but implemented as a static, non-retrieving context window. A consolidated document consisting of American College of Rheumatology (ACR) and European Alliance of Associations for Rheumatology (EULAR) guideline summaries and selected textbook excerpts (Hochberg Rheumatology, 8th edition. Elsevier: 2022) was manually compiled into a single text file. This file was manually inserted into the system prompt at the start of the session.

In models 2 and 3, no assistants API, no retrieval tools, no embeddings search, and no automated RAG framework were used. No fine-tuning, supervised training, or parameter-level modification occurred in either mode. All answers were generated strictly from the combination of the static knowledge context and the question provided. The same knowledge prompt was used for all runs to ensure reproducibility.

HANDLING OF IMAGE-BASED QUESTIONS

All 23 image-based questions (e.g., clinical photos, radiographs) were presented using multimodal capability. The raw image files and accompanying text were uploaded directly to the model input. No images were omitted or converted to text descriptions.

QUESTIONS: VERSIONS AND VALIDATION

Questions were first provided in their original Hebrew version. A translated English version was created by two expert rheumatologists not involved in the study to ensure that the English version maintained the integrity and accuracy of the original Hebrew version. The two versions, Hebrew and English, were presented to base model (Model 1), while only the English version was presented to the other two models [Figure 1].

DATA COLLECTION AND EVALUATION

Each LLM variant was presented with the same set of 200 questions in the same order. Their responses were recorded for analysis. The performance of the models was evaluated according to the following parameters:

- *Accuracy of responses*: The number of correct answers of each model was assessed by comparing its responses to the official correct answers.
- *Comparative analysis:* The performance of the three models were compared to each other.

STATISTICAL ANALYSIS

Descriptive statistics were used to summarize the accuracy of the models' responses and the quality of their reasoning. Performance was reported as overall accuracy percentages and analyzed for differences between the three models. The chi-square test was used to compare overall accuracy of the models. In addition, 95% confidence intervals were calculated for accuracy estimates.

RESULTS

Model 3 demonstrated numerically higher performance, achieving an accuracy of 81%, compared to 77% (Model 1 [English]), 75% (Model 2) and 74.5% (Model 1 [Hebrew]) [Table 1], although these differences did not reach statistical significance. Subsequently, the performance of the different models was evaluated separately for the 2023 and 2024 questions, specifically to examine the consistency of results across different examinations. Again, Model 3 showed the highest point estimates of accuracy (82% and 80%, respectively), although without statistical significance [Table 1]. A comparison of overall accuracy between the four model configurations demonstrated no statistically significant difference in total correct answers across the four models (chi-square = 3.54, *P* = 0.316).

Furthermore, the performance of models on answering questions containing only text versus questions containing text combined with images was assessed [Table 2]. For text-only questions (n=177), Model 1 achieved an accuracy of 82.5% and 79.7% in English and Hebrew, respectively. Model 2 achieved 78.5% accuracy, while Model 3 achieved the highest numerical accuracy (83.1%), although differences were non-significant (chi-square = 1.64, *P* = 0.651). For questions containing text combined with images (n=23),

Table 1. Comparative accuracy of models in answering examination questions, with year-specific (2023 vs. 2024) chi-square analyses

Model	Language	Years	Questions	Correct answers	95% confidence interval	Incorrect answers	Chi-square	*P*-value
Model 1	English	**2023–2024**	200	154 (77.0%)	70.7-82.3%	46 (23.0%)	0.00	1.00
		2023	100	77 (77.0%)		23 (23.0%)		
		2024	100	77 (77.0%)		23 (23.0%)		
Model 1	Hebrew	**2023–2024**	200	149 (74.5%)	68.1-80.1%	51 (25.5%)	0.237	0.626
		2023	100	73 (73.0%)		27 (27.0%)		
		2024	100	76 (76.0%)		24 (24.0%)		
Model 2	English	**2023–2024**	200	150 (75.0%)	68.7-80.6%	50 (25.0%)	0.00	1.00
		2023	100	75 (75.0%)		25 (25.0%)		
		2024	100	75 (75.0%)		25 (25.0%)		
Model 3	English	**2023–2024**	200	162 (81.0%)	75.2-86.0%	38 (19.0%)	0.13	0.72
		2023	100	82 (82.0%)		18 (18.0%)		
		2024	100	80 (80.0%)		20 (20.0%)		

A separate chi-square test comparing overall accuracy across the four model configurations (Model 1 English, Model 1 Hebrew, Model 2, Model 3) showed no statistically significant difference between the models (chi-square = 3.54, *P* = 0.316)

Table 2. Model performance for text-only and text combined with images questions

Type	Model	Language	Questions	Correct answers	95% confidence interval	Incorrect answers	Chi-square	*P*-value
Text only	Model 1	English	177	146 (82.5%)	76.4–87.3%	31 (17.5%)	1.64	0.651
	Model 1	Hebrew	177	141 (79.7%)	73.3–85.0%	36 (20.3%)		
	Model 2	English	177	139 (78.5%)	72.0–84.0%	38 (21.5%)		
	Model 3	English	177	147 (83.1%)	77.0–87.8%	30 (16.9%)		
Text and images	Model 1	English	23	8 (34.8%)	17.2–57.2%	15 (65.2%)	5.78	0.123
	Model 1	Hebrew	23	8 (34.8%)	17.2–57.2%	15 (65.2%)		
	Model 2	English	23	11 (47.8%)	26.8–69.4%	12 (52.2%)		
	Model 3	English	23	15 (65.2%)	42.7–83.6%	8 (34.8%)		

Model 1 demonstrated 34.8% accuracy in both English and Hebrew, Model 2 achieved 47.8% accuracy while Model 3 achieved highest numerical accuracy 65.2% accuracy. Again, no statistically significant difference was observed among the models (chi-square = 5.78, $P = 0.123$) [Table 2].

To assess agreement in response patterns between models, pairwise 2 × 2 contingency tables were generated [Table 3]. The match rate, the proportion of questions for which two models produced the same classification outcome, ranged from 82.5% for Model 1 (English vs. Hebrew) to 91.0% for Model 2 vs. Model 3. Chi-square tests indicated statistically significant differences across all model pairs ($P < 0.001$); however, the absolute differences in match rates were small.

Table 3. Agreement between models: 2 × 2 contingency structure for each model pair

Classification outcome	N (%)	Chi-square	*P*-value
Model 1 (English) vs. Model 1 (Hebrew)			
Both correct	134 (67%)		
Both incorrect	31 (15.5%)		
English correct / Hebrew incorrect	20 (10%)		
English incorrect / Hebrew correct	15 (7.5%)		
Match rate	82.5%	82.5%	0.001
Model 1 (English) vs. Model 2 (CoT)			
Both correct	141 (70.5%)		
Both incorrect	37 (18.5%)		
English correct / Hebrew incorrect	13 (6.5%)		
English incorrect / Hebrew correct	9 (4.5%)		
Match rate	89%	97.9	0.001
Model 2 (CoT) vs. Model 3 (knowledge-augmented prompting)			
Both correct	147 (73.5%)		
Both incorrect	35 (17.5%)		
English correct / Hebrew incorrect	3 (1.5%)		
English incorrect / Hebrew correct	15 (7.5%)		
Match rate	91.0%	112.7	0.001

Match rate represents the percentage of items for which both models produced the same classification outcome (either both correct or both incorrect). This measure reflects concordance in response patterns rather than accuracy. Although chi-square testing demonstrated statistically significant differences (all $P < 0.001$), the absolute differences between model pairs were small (82.5–91%), indicating broadly similar behavior across models.

CoT = few-shot chain of thought

DISCUSSION

In this study, we evaluated the performance of three LLM variants that surpassed the mandatory 65% passing threshold of the official Israeli Board Examinations in Rheumatology (2023 and 2024) required for board certification. Performance was consistent across both examination years while Model 3 (knowledge augmented) achieved the highest numerical accuracy, although the difference across the three prompting strategies (base, few-shot CoT, knowledge-augmented) was not statistically significant. This finding, coupled with a high degree of concordance in the response patterns across all models, suggests a potential intrinsic limitation in the ability of prompt-based techniques to substantially enhance the performance of the most advanced models in complex medical reasoning.

The models' performance, exceeding the 65% passing threshold on the rheumatology certification exam, is consistent with LLM effectiveness reported across various medical specialties [4-9]. However, performance benchmarks vary significantly between different studies and domains. For example, our maximum text-only score was 83.1%, lower than the 93.7% reported in a similar text-based study [13]. This variability likely reflects inherent differences between medical domains and methodological factors, such as different LLM architecture and version, differing examination structures, question emphasis, and prompt engineering techniques.

Our study demonstrates that despite the continuous improvements (e.g., GPT-4: 81% accuracy vs. GPT-3.5: 58% [13]), current LLMs still have limitations. The lack of significant performance gains after implementing few-shot CoT or knowledge-augmented prompting highlights potential intrinsic limitations in complex reasoning and knowledge application. Furthermore, the high agreement among models in correct/incorrect responses suggests prompt-based approaches have a limited ability to effectively utilize external knowledge or the CoT process. Although LLMs can process images, the suboptimal performance of all models on questions combining text and images (e.g., physical findings, diagnostic imaging) reflects that this multimodal capability is still in its early stages [15]. This inherent limitation in interpreting visual medical data restricts the clinical utility of current models in specialties where visual pattern recognition is essential for diagnosis and management.

While LLMs can successfully pass rheumatology examinations, they do not replace clinician-level reasoning and lack the comprehensive skills of an expert rheumatologist. It is vital to distinguish exam-oriented performance from real-world clinical competence. Board examinations primarily test factual recall, which LLMs replicate well. However, clinical rheumatology requires context-dependent decisions, longitudinal assessment, and intuitive clinical judgment [16]. For example, while an LLM can correctly answer a classic exam vignette (e.g., rheumatoid arthritis via keyword matching), it lacks the deeper reasoning necessary for real patients who present with atypical or incomplete symptoms, requiring assessment of disease evolution and exclusion of mimickers. Similarly, LLMs may fail to appreciate acuity, urgency, co-morbidity considerations, or misclassification risks in critical scenarios like suspected giant cell arteritis. Furthermore, board exams use dichotomous right-or-wrong answers, but rheumatologic diseases are often complex, rare, and heterogeneous, lacking clear-cut distinctions [17]. LLMs also rely on internet data, which may overrepresent atypical or rare cases, leading to potentially biased or inaccurate outputs. The nuanced practice of rheumatology, demanding interpretation of subtle physical findings, integration of multimodal data, and navigation of diagnostic uncertainty, requires expertise, judgment, and experiential knowledge that current LLM iterations do not possess.

While LLMs may assist in routine settings, their weaknesses and susceptibility to inaccuracies necessitate caution. A substantial underperformance was noted across all models on image-based questions (e.g., clinical photographs, radiographs, dermatologic findings). Such items require nuanced pattern recognition and visual-spatial interpretation, tasks for which current LLMs and vision-language architectures are limited. Contributing factors likely include the scarcity of rheumatology-specific imaging data in training sets and the reliance on caption-based vision systems over true pixel-level medical image interpretation. This finding is consistent with literature showing that while LLMs may outperform physicians in text-only scenarios, human performance gain is significantly larger when images are added, suggesting LLMs rely predominantly on textual cues [18]. This marked drop in accuracy on multimodal questions reflects an intrinsic limitation of current GPT-4-turbo vision capabilities, rather than incomplete input. As these models are rapidly advancing [9], ongoing reassessment is essential to monitor improvements and determine the safe integration of LLMs into rheumatology practice.

The study's limitations include evaluating performance solely using multiple-choice questions, which limits the representation of real-world clinical complexity, and the absence of a detailed qualitative analysis of the models' reasoning. Statistical power was constrained by the fixed number of questions, especially the small subset of image-based items, increasing the risk of Type II error. Furthermore, the evaluation was restricted to ChatGPT-based models, limiting generalizability; future research should evaluate other multimodal and domain-adapted LLMs. Finally, it is crucial to recognize that LLMs are susceptible to socio-demographic and ethnic biases from training data imbalances [19,20]. Since disease manifestations in rheumatology vary across populations, incorporating bias assessment is essential for future integration.

CONCLUSIONS

Our study highlights the potential role of LLMs in rheumatology. Future research should focus on addressing limitations, especially image interpretation and handling complex cases to enable efficient application of LLMs in rheumatology.

Correspondence

Dr. F. Hassan
Dept. of Rheumatology, Galilee Medical Center, Nahariya 2210001, Israel
Phone: (972-4) 910-7038
Fax: (972-4) 910-7593
Email: fadihh@gmail.com

References

1. Jiang F, Jiang Y, Zhi H, et al. Artificial intelligence in healthcare: past, present and future. *Stroke Vasc Neurol* 2017; 2 (4): 230-43.
2. Liu M, Shen X, Pan W. Deep reinforcement learning for personalized treatment recommendation. *Stat Med* 2022; 41: 4034-56.
3. Katz U, Cohen E, Shachar E, et al. GPT versus resident physicians–a benchmark based on official board Scores. *N Engl J Med AI* 2024; 1: AIdbp2300192.
4. Gilson A, Safranek CW, Huang T, et al. How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? The implications of large language models for medical education and knowledge assessment. *JMIR Med Educ* 2023; 9: e45312.
5. Jang D, Yun TR, Lee CY, Kwon YK, Kim CE. GPT-4 can pass the Korean National Licensing Examination for Korean Medicine Doctors. *PLOS Digit Health* 2023; 2: e0000416. Erratum in PLOS Digit Health 2025; 4: e0001000.
6. Lai UH, Wu KS, Hsu TY, Kan JKC. Evaluating the performance of ChatGPT-4 on the United Kingdom Medical Licensing Assessment. *Front Med (Lausanne)* 2023; 10: 1240915.
7. Krusche M, Callhoff J, Knitza J, Ruffer N. Diagnostic accuracy of a large language model in rheumatology: comparison of physician and ChatGPT-4. *Rheumatol Int* 2024; 44: 303-6.
8. Cipolletta E, Fiorentino MC, Vreju FA, Moccia S, Filippucci E. Editorial: artificial intelligence in rheumatology and musculoskeletal diseases. *Front Med (Lausanne)* 2024; 11: 1402871.
9. Tarabanis C, Zahid S, Mamalis M, Zhang K, Kalampokis E, Jankelson L. Performance of publicly available large language models on internal medicine board-style questions. *PLOS Digit Health* 2024; 3: e0000604.
10. Pincus T, Yazici Y, Sokka T. Complexities in assessment of rheumatoid arthritis: absence of a single gold standard measure. *Rheum Dis Clin North Am* 2009; 35: 687-97.
11. Correia de Sa A, Batista M, Ferreira AL, Casanova D, Faria B, Cotter J. Diagnostic challenges of systemic lupus erythematosus. *Cureus* 2023; 15: e50132.
12. Connolly MK. Systemic sclerosis (scleroderma): remaining challenges. *Ann Transl Med* 2021; 9: 438.
13. Madrid-García A, Rosales-Rosado Z, Freites-Nuñez D, et al. Harnessing ChatGPT and GPT-4 for evaluating the rheumatology questions of the Spanish access exam to specialized medical training. *Sci Rep* 2023; 13: 22129.
14. Daungsupawong H, Wiwanitkit V. Comparative performance of artificial intelligence models in rheumatology board-level questions: evaluating Google Gemini and ChatGPT-4o: correspondence. *Clin Rheumatol* 2024; 43: 4015-16.
15. Omar M, Agbareia R, Klang E, Naffaa ME. Large language models in rheumatologic diagnosis: a multimodal performance analysis. *J Rheumatol* 2025; 52: 187-8.
16. Guillen-Grima F, Guillen-Aguinaga S, Guillen-Aguinaga L, et al. Evaluating the efficacy of ChatGPT in navigating the Spanish Medical Residency Entrance Examination (MIR): promising horizons for AI in clinical medicine. *Clin Pract* 2023; 13: 1460-87.
17. Mbakwe AB, Lourentzou I, Celi LA, Mechanic OJ, Dagan A. ChatGPT passing USMLE shines a spotlight on the flaws of medical education. *PLOS Digit Health* 2023; 2: e0000205.
18. Agbareia R, Omar M, Soffer S, Glicksberg BS, Nadkarni GN, Klang E. Visual-textual integration in LLMs for medical diagnosis: a preliminary quantitative analysis. *Comput Struct Biotechnol J* 2024; 27: 184-9.
19. Omiye JA, Lester JC, Spichak S, Rotemberg V, Daneshjou R. Large language models propagate race-based medicine. *NPJ Digit Med* 2023; 6: 195.
20. Omar M, Sorin V, Agbareia R, et al. Evaluating and addressing demographic disparities in medical large language models: a systematic review. *Int J Equity Health* 2025; 24: 57.

Capsule

A quantitative DOPA decarboxylase biomarker for diagnosis in Lewy body disorders

Accurate diagnosis of dementia with Lewy bodies (DLB) remains challenging, with misdiagnosis potentially leading to harmful treatment decisions. DOPA decarboxylase (DDC) shows promise as a cerebrospinal fluid (CSF) biomarker for DLB and Parkinson's disease (PD), but quantitative assays are needed for its clinical implementation. **Bolsewig** and colleagues reported on the development of two DDC immunoassays and the extensive clinical validation of DDC across three clinical cohorts (n = 740), one biologically defined cohort (n = 253), one cohort with detailed dopamine transporter imaging information (n = 102), and one autopsy-confirmed cohort (n = 78). CSF DDC levels were significantly higher in DLB and PD (up to 2.5-fold vs. controls; 1.9-fold vs. AD), showing area under the curve values > 0.9 for differential diagnosis. Elevated CSF DDC was linked to the presence, but not severity, of motor impairment. In autopsy-confirmed DLB, higher CSF DDC correlated with progressing α-synuclein pathology and immunohistochemistry in DLB and PD brain tissue revealed colocalization of DDC and α-synuclein in the substantia nigra. These findings underscore DDC's value to support DLB and PD diagnosis, paving the way for its clinical implementation using the here-presented developed immunoassays.

Nat Med 2026; https://doi.org/10.1038/s41591-026-04212-0
Eitan Israeli