

Assessing the accuracy and consistency of generative pretrained transformers in assigning Eastern Cooperative Oncology Group performance status

Chun En Yau^{*1}, Qihuang Xie^{*2}, Ren Yi Jonas Ho³ MRCP, Chun Yi Yau¹ MBBS, Elaine Guan¹, Dawn Yi Xin Lee⁴, Xinyan Zhou¹, Gerald Gui Ren Sng⁵ MRCP, Joshua Yi Min Tung⁶ MRCS, Andrew Fu Wah Ho^{7,8} PhD, Ryan Shea Ying Cong Tan^{**9,10} MRCP, Daniel Yan Zheng Lim^{**3,11} MRCP

Dear Editor,

The Eastern Cooperative Oncology Group (ECOG) is a commonly used performance status (PS) scale in oncology. It influences cancer treatment decisions and clinical trial recruitment. However, there can be significant inter-rater variability in ECOG-PS scoring, due to subjectivity in human scoring and innate cognitive biases.^{1,2} We propose that generative pretrained transformers (GPT), a foundational large language model (LLM), can accurately and reliably score ECOG-PS.

We used 16 fictional scripts (Supplementary Table S1, Set 1) from studies by Datta et al. and Azam et al., both of which assessed the inter-rater reliability of ECOG-PS scoring by oncology health professionals.^{2,3} We used the OpenAI Action Programming Interface to query the GPT-3.5-turbo model with the scripts. A standardised prompt (I would like you to assume the role of an oncologist. You will review the patient case history that I give you and score the ECOG score) was used. No further instructions were provided. We queried GPT-3.5-turbo with 16 scenarios sequentially. Each scenario was reassessed for 50 iterations, and all outputs were captured and documented precisely. For the 12 scenarios from Datta et al., the most appropriate ECOG score for each vignette was decided in the original study through consensus rating among the 3 oncologists who designed the vignettes. Any discrepancy in the ratings was resolved through feedback and discussion. For the scenarios from Azam et al., the original study did

not provide a gold standard ECOG score for each of the 4 vignettes. Thus, the most appropriate ECOG score was determined by an author of the current study, RSYCT, a consultant oncologist. We did an initial qualitative appraisal of the interpretability and validity of the responses. We then quantitatively assessed the accuracy (percentage correct ECOG-PS) and consistency (Fleiss' kappa⁴) of the responses. GPT's scores were compared with human scores using the Mann-Whitney U test. The cut-off for statistical significance was set at $P < 0.001$ (after Bonferroni correction).

Since GPT pretraining data used text corpora before 2021, there is a possibility that the original vignettes were included as pretraining data. Thus, we also assessed GPT responses to modified versions of the vignettes. We created 2 new sets of vignettes (Supplementary Table S1, Sets 2–3), with 1 set paraphrasing the script and another modifying numerical values. We ensured that the textual and numerical changes did not change the overall ECOG-PS for both sets.

Qualitatively, GPT yielded clear, comprehensible answers in response to all scripts. A single valid ECOG-PS was assigned for all the scenarios. Although unprompted, GPT further justified the ECOG-PS assigned in all cases. All script responses were reviewed qualitatively and summarised (Supplementary Table S2).

Notably, it demonstrated the ability to distinguish between a patient's functional ability for activities of daily living and their ability to perform work

The Annals is an open access journal, allowing non-commercial use under CC BY-NC-SA 4.0.

¹ Yong Loo Lin School of Medicine, National University of Singapore, Singapore

² Department of Pharmacy, National University of Singapore, Singapore

³ Department of Gastroenterology and Hepatology, Singapore General Hospital, Singapore

⁴ School of Medicine, Dentistry and Nursing, University of Glasgow, Glasgow, United Kingdom

⁵ Department of Endocrinology, Singapore General Hospital, Singapore

⁶ Department of Urology, Singapore General Hospital, Singapore

⁷ Department of Emergency Medicine, Singapore General Hospital, Singapore

⁸ Pre-Hospital and Emergency Research Centre, Duke-NUS Medical School, Singapore

⁹ Division of Medical Oncology, National Cancer Centre, Singapore

¹⁰ Duke-NUS Medical School, Singapore

¹¹ Medicine Academic Clinical Programme, Duke-NUS Medical School, Singapore

* Joint first authors

** Joint last authors

Correspondence: Dr Daniel Yan Zheng Lim, Department of Gastroenterology and Hepatology, Singapore General Hospital, Outram Road, Singapore 169608.

Email: limydz@gmail.com

This article was first published online on 13 September 2024 at annals.edu.sg.

activities, which is important in differentiating ECOG-PS scores of 1 and 2. Furthermore, some responses displayed a level of inference of how a patient's symptoms might impact function, as shown in Supplementary Fig. S1, in which ascites and pedal oedema were inferred to contribute to limited mobility. Hallucination was observed in a few cases. This included a few instances of misinterpretation of the prompt. For instance, 1 case in which the patient's work activity of running a store was not considered, and she was given a score of ECOG-PS 3 (Supplementary Fig. S2).

GPT scored ECOG-PS correctly more often than human raters (Table 1). The difference in performance was statistically significant in 7 scripts (44%). Ten scripts (63%) had $\geq 90\%$ correct scores. GPT had excellent consistency, with Fleiss' kappa of 0.785 on Datta et al.'s scripts and 0.739 on Azam et al.'s. Conversely, human raters had less consistency, with kappa of 0.167 on Datta's scripts (Azam et al. did not report a kappa value). There was no substantial difference between the GPT responses on the original scripts (Set 1) and the modified scripts (Sets 2–3) for most cases (Supplementary Tables S3–S4). For the cases with a difference in performance, there were more correct responses on the variant scripts. Only patient 5 in Azam et al. had a decline in performance on the variant scripts. The consistency of GPT on variant scripts remained excellent (Fleiss' kappa: set 2 Datta et al., 0.776; set 3 Datta et al., 0.771; set 2 Azam et al., 0.843; set 3 Azam et al., 0.665).

To the authors' knowledge, this is the first study to examine whether GPT can be used to determine the ECOG-PS and compare its performance to human raters. Our results show that GPT performs ECOG-PS scoring more accurately and consistently than human raters in fictional patient scripts. GPT performed comparably to human raters on Set 1 in cases 2, 4, 7, 8, 9, 10 and 12. It also had significantly more correct responses in cases 1, 3, 5, 6 and 11 than human raters who had considerable experience in oncology (mean = 4.4 ± 4.2 years, $>80\%$ having at least 1 year of experience in Datta et al.).

Apart from assigning ECOG-PS with greater accuracy, our results demonstrate that GPT can consistently determine ECOG-PS with high inter-rater reliability across 3 sets of fictional scripts (kappa >0.7). Our results also suggest a good basic

performance level of GPT that could be improved with task-specific fine-tuning.

Important limitations of this study include the fact that GPT is a general language model not specifically trained for ECOG-PS scoring. At the time of writing, newer models such as GPT-4-turbo had not been released, and they were not assessed in this study. The quality information on ECOG in the underlying pretraining data of GPT is not public and cannot be independently verified. However, the relatively good performance in this study suggests that GPT processes this area of medicine well. Another limitation is that ECOG scores are subjective assessments dependent on the experience and opinion of the rater. While we have attempted to assign the most appropriate ECOG score by using consensus scoring and expert opinion from an experienced oncologist, they remain subjective measures of PS. Currently, only the zero-shot prompting technique was experimented. However, if necessary, LLM effectiveness may be enhanced through various prompting techniques and fine-tuning. Some prompting techniques that can be explored in future work range from in-context learning examples,⁵ to more intricate methods like the chain-of-thought,⁶ and even a prompting technique (MedPrompt) that yielded comparable performance in answering medical school examination questions to LLM pretrained on medical text.⁷ We also recognise that our study was performed on a small set of fictional patient scenarios. Future work should examine scoring on actual patient clinical notes with concurrent human evaluation (including clinicians across the spectrum of experience and practices) to better assess the capabilities of LLMs. With more refinements to LLM algorithms, training data and interpretability, LLM-determined clinical scoring may become even more consistent and can be used as a useful adjunct to counterpoise human bias for such scoring in the future.

Declaration

The authors declare there are no affiliations with or involvement in any organisation or entity with any financial interest in the subject matter or materials discussed in this manuscript. Patients or the public were not involved in the design, conduct, reporting or dissemination plans of this research.

Keywords: cancer, ChatGPT, ECOG, large language model, performance status

Table 1. Comparison of GPT and human rater responses on original scripts provided in Datta et al. and Azam et al. ^{2,3}

Source	Patient	ECOG PS					Most common ECOG PS	Suggested answer ^a	Accuracy (% correct)	P value ^b
		0	1	2	3	4				
Datta et al.	1	12	45	14	1	0	1	2	19.4	Ref
GPT		0	2	48	0	0	2	2	96.0	<0.001*
Datta et al.	2	3	14	35	20	0	2	3	27.8	Ref
GPT		0	0	50	0	0	2	3	0.0	0.632
Datta et al.	3	6	10	21	30	5	3	2	29.2	Ref
GPT		0	13	37	0	0	2	2	74.0	<0.001*
Datta et al.	4	11	28	30	3	0	2	2	41.7	Ref
GPT		0	48	2	0	0	1	2	4.0	0.003
Datta et al.	5	3	22	23	22	2	2	2	31.9	Ref
GPT		0	0	20	30	0	3	2	40.0	<0.001*
Datta et al.	6	0	1	0	17	54	4	3	23.6	Ref
GPT		0	0	3	46	1	3	3	92.0	<0.001*
Datta et al.	7	1	12	24	31	4	3	2	33.3	Ref
GPT		0	0	50	0	0	2	2	100.0	0.001
Datta et al.	8	1	15	20	31	5	3	2	27.8	Ref
GPT		0	0	45	5	0	2	2	90.0	0.033
Datta et al.	9	8	41	21	2	0	1	1	56.9	Ref
GPT		0	49	1	0	0	1	1	98.0	0.02
Datta et al.	10	21	39	7	5	0	1	1	54.2	Ref
GPT		0	49	0	0	1	1	1	98.0	0.09
Datta et al.	11	1	2	24	39	6	3	3	16.7	Ref
GPT		0	0	0	49	1	3	3	98.0	<0.001*
Datta et al.	12	0	26	31	12	3	2	1	36.1	Ref
GPT		0	3	47	0	0	2	1	6.0	0.248
Azam et al.	1	4	18	25	1	0	2	2	54.3	Ref
GPT		0	0	50	0	0	2	2	100.0	<0.001*
Azam et al.	2	0	2	28	8	0	2	2	56.0	Ref
GPT		0	20	30	0	0	2	2	60.0	<0.001*
Azam et al.	3	0	1	12	33	1	3	3	66.0	Ref
GPT		0	0	2	48	0	3	3	96.0	0.004
Azam et al.	4	0	1	7	19	14	3	3	38.0	Ref
GPT		0	0	1	49	0	3	3	98.0	0.088

ECOG: Eastern Cooperative Oncology Group; GPT: generative pretrained transformers; PS: performance status

^a based on the original description by Datta et al.^b Mann-Whitney U-test for differences in score distribution

Supplementary materials

Fig. S1. Example prompt and response from GPT.

Fig. S2. A response with erroneous score by GPT.

Table S1. Set 1 (original vignettes by Datta et al. and Azam et al.), Set 2 (vignettes with variation in language) and Set 3 (vignettes with variation in numeric values).

Table S2. Qualitative summary of GPT's responses (compiled from Sets 1, 2 and 3).

Table S3. Comparison of GPT responses for Sets 1, 2 and 3 using the patient vignettes provided in Datta et al.

Table S4. Comparison of GPT responses for Sets 1, 2 and 3 using the patient vignettes provided in Azam et al.

REFERENCES

1. Chow R, Bruera E, Temel JS, et al. Inter-rater reliability in performance status assessment among healthcare professionals: an updated systematic review and meta-analysis. *Support Care Cancer* 2020;28:2071-8.
2. Datta SS, Ghosal N, Daruvala R, et al. How do clinicians rate patient's performance status using the ECOG performance scale? A mixed-methods exploration of variability in decision-making in oncology. *Ecancermedalscience* 2019;13:913.
3. Azam F, Latif MF, Farooq A, et al. Performance Status Assessment by Using ECOG (Eastern Cooperative Oncology Group) Score for Cancer Patients by Oncology Healthcare Professionals. *Case Rep Oncol* 2019;12:728-36.
4. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;33:159-74.
5. Brown TB, Mann B, Ryder N, et al. Language models are few-shot learners. *Adv Neural Inf Process Syst* 2020; 33:1877-901.
6. Wei J, Wang X, Schuurmans D, et al. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *Adv Neural Inf Process Syst* 2022;35:24824-37.
7. Nori H, Lee YT, Zhang S, et al. Can Generalist Foundation Models Outcompete Special-Purpose Tuning? Case Study in Medicine. <https://arxiv.org/pdf/2311.16452>. Accessed 10 March 2023.