Original Article

Chat generative pre-trained transformer's performance on dermatology-specific questions and its implications in medical education

James Behrmann[1]^, Ellen M. Hong[1], Shannon Meledathu[1], Aliza Leiter[1], Michael Povelaitis[1], Mariela Mitre[1,2]

[1]Hackensack Meridian School of Medicine, Hackensack Meridian Health, Nutley, NJ, USA; [2]Division of Dermatology, Department of Medicine, Hackensack University Medical Center, Hackensack, NJ, USA

Contributions: (I) Conception and design: J Behrmann, EM Hong; (II) Administrative support: J Behrmann; (III) Provision of study material or patients: J Behrmann, S Meledathu; (IV) Collection and assembly of data: J Behrmann, S Meledathu; (V) Data analysis and interpretation: All authors; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

Correspondence to: James Behrmann, BS. Hackensack Meridian School of Medicine, Hackensack Meridian Health, 123 Metro Blvd., Nutley, NJ 07110, USA. Email: James.behrmann@hmhn.org.

Background: Large language models (LLMs) like chat generative pre-trained transformer (ChatGPT) have gained popularity in healthcare by performing at or near the passing threshold for the United States Medical Licensing Exam (USMLE), but some limitations should be considered. Dermatology is a specialized medical field that relies heavily on visual recognition and images for diagnosis. This paper aimed to measure ChatGPT's abilities to answer dermatology questions and compare this sub-specialty accuracy to its overall scores on USMLE Step exams.

Methods: A total of 492 dermatology-related questions from Amboss were separated into their corresponding medical licensing exam (Step 1 =160, Step 2CK =171, and Step 3 =161). The question stem and answer choices were input into ChatGPT, and the answers, question difficulty, and the presence of an image omitted from the prompt were recorded for each question. Results were calculated and compared against the estimated 60% passing standard.

Results: ChatGPT answered 41% of all the questions correctly (Step 1 =41%, Step 2CK =38%, and Step 3 =46%). There was no significant difference in ChatGPT's ability to answer questions originally containing images or no image [P=0.205; 95% confidence interval (95% CI): 0.00 to 0.15], but it did score significantly lower compared to the estimated 60% threshold passing standard for USMLE exams (P=0.008; 95% CI: −0.29 to −0.08). Analyzing questions by difficulty level demonstrated a skewed distribution with easier-rated questions correctly more often (P<0.001).

Conclusions: Our findings demonstrate that ChatGPT answered fewer correct dermatology-specific questions when compared to its overall performance on the USMLE (41% and 60%, respectively). Interestingly, ChatGPT scored similarly whether or not the question had an associated image, which may provide insight into how it utilizes its knowledge base to select answer choices. Using ChatGPT in conjunction with deep learning systems that include image analysis may improve accuracy and provide a more robust educational tool in dermatology.

Keywords: Artificial intelligence (AI); dermatology; chat generative pre-trained transformer (ChatGPT)

Received: 12 May 2023; Accepted: 12 September 2023; Published online: 25 September 2023.

doi: 10.21037/jmai-23-47

View this article at: https://dx.doi.org/10.21037/jmai-23-47

^ ORCID: 0000-0001-5281-0660.

Introduction

Artificial intelligence (AI) has become an increasingly important topic in medicine. AI has many components, including machine learning, neural networks, and large language models (LLMs), all playing different roles in their applications (1). Despite rapid development in the field, there are limitations to validating the results produced by AI (2,3). There is also controversy over whether patients would trust interfacing with AI (4). Continued advances in AI may address some of these limitations.

LLMs, in particular, have gained popularity recently due to the emergence of chat generative pre-trained transformer (ChatGPT) (OpenAI, San Francisco, California, USA). ChatGPT is an iteration of GPTs from OpenAI that underwent reinforcement training with human feedback (RTHF) (5,6). This training is conducted by a human "labeler" initially demonstrating an example of an answer to a prompt to ChatGPT. Next, ChatGPT receives a prompt and is instructed to produce multiple responses, of which the human "labeler" ranks the responses from best to worst using a reward system. ChatGPT bases future responses using the framework of the reward system (7). Due to the increasing popularity of OpenAI's LLM, companies like Google and its Chinese counterpart, Baidu, have developed similar models (8,9).

ChatGPT has also gained notoriety in academia due to its ability to pass standardized tests, including the United States Medical Licensing Exam (USMLE), "at or near" the approximate passing threshold of 60% (10,11). Researchers have been pushing ChatGPT's skills in other areas to evaluate its proficiency since that time (12-15). Educators debate whether ChatGPT should be banned from schools or adapted as a learning tool (16). One of the controversies and ethical concerns arises from how ChatGPT is trained, which emphasizes answering the user's question using plausible-sounding responses without focusing on logic or correctness (7). This has led to the creation of factually incorrect 'hallucinations' that sound believable enough to stump experts (17).

As ChatGPT becomes more widely accepted, it is crucial to understand its limitations in medicine and how its utility applies to specific medical subspecialties such as dermatology, which relies heavily on the visual appearance of conditions. Online symptom calculators have shown difficulty generating accurate dermatological diagnoses with text-only inputs (18). Deep learning systems incorporating image analysis have shown promise in correctly diagnosing common skin pathologies (19). Limitations are being addressed to increase the variety of skin tones represented in the training data used by these systems (20,21).

In general, the content in Step 1 consists of basic science content, while Steps 2 and 3 aim to test participants on the clinical applications of knowledge. Dermatology represents a small portion of the content tested on USMLE, ranging from 4–10%, depending on the exam (22). There is limited data available that suggests medical students perform worse on dermatology questions due to the limited exposure to dermatology in pre-clinical and clinical courses (23). ChatGPT theoretically should perform similarly across all content sections of USMLE because it doesn't rely on specialty training.

The aim of this paper was to investigate ChatGPT's performance on dermatology-specific board questions to evaluate its role in the future of medical education.

Highlight box

Key findings

- Chat generative pre-trained transformer (ChatGPT) scored significantly lower on dermatology-specific questions. Despite not having access to the images, ChatGPT performed similarly on questions with and without associated images, suggesting bias in the model's method for selecting answers.

What is known and what is new?

- Large language models like ChatGPT have demonstrated their ability to perform at or above passing scores on standardized tests such as medical licensing exams.
- ChatGPT provided reasonable explanations but struggled with next-step question stems, pointing towards its more appropriate use as a supplementary learning resource.

What is the implication, and what should change now?

- As medical education incorporates artificial intelligence, we must consider that it may not similarly apply to all medical disciplines.

Methods

ChatGPT

The version of ChatGPT (v3.5) used in this study was from the January 30th, 2023, release note. No specific training or priming was provided prior to the study. ChatGPT displayed a high level of concordance (>90%) from previous studies and was not directly tested in this study (10,24,25).

Amboss

Amboss is an online education company that provides resources such as content and questions for board exams and continuing medical education curriculums accredited through the Accreditation Council for Continuing Medical Education (26,27). It provides a question bank that prepares medical students for USMLE and National Board of Medical Examiners (NBME) examinations. It has an internal metric for measuring the difficulty of a question based on the number of students who answer it correctly.

Question filtering

A total of 5,626 questions from the Amboss question bank were filtered by specialty to include 492 questions related to dermatology. Those questions were separated into groups representing their corresponding medical licensing exam, 160 for USMLE Step 1, 171 for USMLE Step 2 Clinical Knowledge, and 161 for USMLE Step 3. Of the 492 questions, 154 (31.3%) included images. We included these in the study to maximize the number of tested questions but differentiated them during analysis.

Question input and recording

The question stems and answer choices were input into ChatGPT, and responses were imputed back into Amboss for evaluation. Question outcomes and difficulty levels (1 to 5 hammers denoted as level 1–5 here, with 1 being the easiest and 5 the hardest) were recorded (28). ChatGPT received no instruction on whether or not its response was correct. Questions were input as new instances so as not to bias answers from previous entries. Due to the inability to input images into ChatGPT, only the text portion of questions associated with images was input; the images were omitted. The number of questions with images was tallied, and answers were recorded for analysis.

Analysis

The primary outcome was to assess the correct percentages from the dermatology question bank and to compare them to the approximated passing score of 60% on the USMLE exams. The secondary outcome was to assess how ChatGPT performed depending on question difficulty. The difficulty level assigned by Amboss was recorded for each question (28). The percentage of correct answers was calculated by dividing the number of correct responses by the total number of questions for that level. We also quantified the questions for each difficulty level by exam.

Statistical analysis

To analyze the data, a 2-tailed paired *t*-test was used to compare ChatGPT's performance on the dermatology question sets (all questions, questions without images, and questions with images). A one-sample paired *t*-test analyzed the performance on question sets compared to the estimated "at or near" passing score of 60% on the USMLE. A 95% confidence interval (CI) was calculated with each of the *t*-tests. A chi-square test was used to evaluate ChatGPT's performance by question difficulty level.

Results

ChatGPT yields lower accuracy on dermatology-based questions than on USMLE

ChatGPT correctly answered 204 out of the 492 total questions (41%), with 65 out of 160 (41%) on the USMLE Step 1 questions, 65 out of 171 (38%) on the USMLE Step 2CK questions, and 74 out of 161 (46%) on USMLE Step 3 questions. When adjusting by removing questions with images, ChatGPT answered 148 questions correctly out of 338 (44%), 49 out of 107 (46%) on USMLE Step 1, 47 out of 121 (39%) on USMLE Step 2CK, and 52 out of 110 (47%) on USMLE Step 3 questions. ChatGPTs results were significantly different when contrasting against the approximated 60% passing mark for USMLE [all questions: P=0.008, 95% CI: −0.29 to −0.08; with image: P=0.012, 95% CI: −0.31 to −0.16; without image: P=0.012, 95% CI: −0.27 to −0.05; *Table 1*]. There was no significant difference between scores on questions with or without images (P=0.205; 95% CI: 0.00 to 0.15; *Table 2*).

Question difficulty level analysis demonstrated a skewed distribution toward easier questions

We utilized a chi-squared analysis with an overall percent correct score of 41% to generate the expected number of questions ChatGPT should have answered correctly by difficulty level. There was a significant difference to what was observed during the study (Table S1), with ChatGPT answering lower-difficulty questions with higher accuracy and higher-difficulty questions with lower frequency

Table 1 Comparison of chat generative pre-trained transformer's performance on dermatology questions

USMLE	Correct	Total	Correct (%)	Approximated passing score	Mean difference	P value	95% CI
All questions				0.60	–0.18	0.008	–0.29 to –0.08
Step 1	65	160	41				
Step 2CK	65	171	38				
Step 3	74	161	46				
Total	204	492	41				
Questions without images				0.60	–0.16	0.012	–0.27 to –0.05
Step 1	49	107	46				
Step 2CK	47	121	39				
Step 3	52	110	47				
Total	148	338	44				
Questions with images				0.60	–0.24	0.012	–0.31 to –0.16
Step 1	16	53	30				
Step 2CK	18	50	36				
Step 3	22	51	43				
Total	56	154	36				

Analysis of the questions by breaking them down into three groups: (I) the results for all questions in the data set; (II) only the questions in the data set that were never associated with an image; and (III) the questions that initially contained images. The mean difference was calculated between the three groups and compared to the passing score. There was a significant difference between all three groups and the approximated USMLE passing score of 60%. USMLE, United States Medical Licensing Exam; 95% CI, 95% confidence interval.

Table 2 Comparison of chat generative pre-trained transformer's performance on questions with and without images

	Mean difference	P value	95% CI
All questions *vs.* without images	0.02	0.232	–0.07 to 0.02
All questions *vs.* with images	0.05	0.198	–0.01 to 0.11
Without images *vs.* with images	0.08	0.205	0.00 to 0.15

A comparison of the groups from *Table 1* to one another instead of a passing score. There was no significant difference between any of the groups in this study. 95% CI, 95% confidence interval.

(P<0.001).

For the questions assigned to the Step 1 designation, ChatGPT correctly answered 14 of the 21 (67%) level 1 questions (easiest), 22 of the 56 (39%) level 2 questions, 22 of the 59 (37%) level 3 questions, 6 of the 19 (32%) level 4 questions, and 1 of the 5 (20%) level 5 questions (most difficult).

On the questions assigned to Step 2CK, ChatGPT correctly answered 26 of the 45 (58%) level 1 questions, 17 of the 42 (40%) level 2 questions, 12 of the 48 (25%) level 3 questions, 9 of the 28 (32%) level 4 questions, and 1 of the 8 (13%) level 5 questions. Finally, of the questions assigned to Step 3, ChatGPT correctly answered 31 of the 48 (65%) level 1 questions, 24 of the 44 (55%) level 2 questions, 14 of the 42 (33%) level 3 questions, 3 of the 20 (15%) level 4 questions, and 2 of the 7 (29%) level 5 questions (*Figure 1*).

There were thirteen questions from the question bank that ChatGPT answered with an answer that was not listed. An additional attempt was made by assisting ChatGPT in these 13 scenarios to select from one of the answer choices listed, which still led to an incorrect answer by the bot. ChatGPT's rationale had the correct explanation but failed to identify the correct next-step answer choice. All the questions were marked as incorrect.

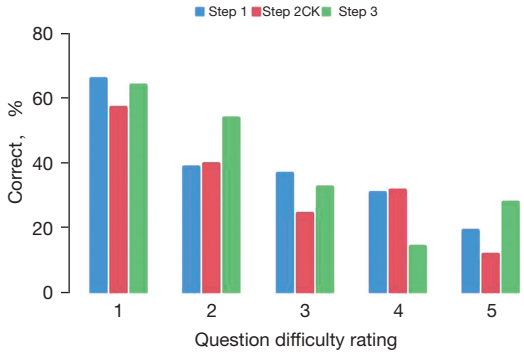

Figure 1 Overall percent correct by question difficulty level. Graphical demonstration of the overall % correct by difficulty level for associated USMLE. USMLE, United States Medical Licensing Exam.

Discussion

In this study, we have evaluated ChatGPT's medical knowledge competency in answering dermatology-specific exam questions. ChatGPT performed less accurately when answering dermatology-specific examination bank questions than its previously published performance on the USMLE licensing exams. Our results highlight an important finding when considering future applications of LLMs and AI in medical subspecialties that may not have been extensively represented in its current armamentarium of knowledge.

While the observed 8% difference in ChatGPT's ability to answer questions with or without images was not statistically significant, it may be due to a limited number of available questions leading to insufficient power in the study. ChatGPT's inability to process images and score similarly on questions with and without images may indicate how it selects answers. We recognized that ChatGPT referenced demographic information provided in the question stem in its rationale, and oftentimes, the response included high-probability selections for the patient demographic. It frequently failed to recognize or discuss key critical findings for the correct diagnosis. Its similar performance, albeit not passing, demonstrates its tendency to select high-probability answers. That may also explain why there was a trend of answering easier questions more accurately than more difficult ones, the latter requiring recognition of complicated question stem details and appropriately incorporating images into the differential diagnosis.

There were also cases in which ChatGPT provided an extensive rationale for its answer selection but ultimately selected the incorrect choice. These questions were frequently related to management algorithms, demonstrating ChatGPT's understanding of the general principle but limitations in determining the most critical next step in evaluation. In medical education, understanding the sequential steps to arrive at a diagnosis or treatment is emphasized to help students avoid ordering unnecessary tests and wasting healthcare resources. Following treatment algorithms may seem straightforward for ChatGPT to execute, but the nuance of determining where the patient was in the algorithm seemed underdeveloped.

While the exact mechanism by which ChatGPT selects a response is complicated, its training process does give us some insight. Incorrect answer selections could be due to a lack of dermatology-specific information in the training data corpora. Also, OpenAI employs human "labelers" that reward a particular type of response during the reinforcement learning process. It adds variability based on each labeler's professional or educational background, limiting the model from producing the response it may feel fits best. The model is also trained to err on the side of being overly cautious, further influencing its answer selections even though it might not have been its highest-rated response (7). As AI continues to be integrated into medicine, an important question may be how to integrate medical professionals into the AI developmental process.

Overall, ChatGPT provided coherent information on various topics in dermatology, supporting its use as an additional resource in medical education. Language models are trained to provide outputs regardless of their level of certainty, leading to their well-known tendency to provide false statements confidently. Further development may be warranted before utilizing it as a primary resource for students or patients during the early stages of learning. In the model's current state, its use as a secondary resource in conjunction with well-established education methods may be more appropriate.

Limitations

We encountered several limitations during this study. One challenge was entering questions into ChatGPT, which used tables with laboratory values. Incorporating tables of laboratory values was problematic, and although ChatGPT's response addressed the lab results, it was difficult to assess if it was correctly processing this type of data. Another limitation of this study is that the questions were from a different source than previous studies. However, the context performance of medical students is similar across

these databases. Model performance and generalizability are challenges that should be acknowledged and improved in further network iterations. ChatGPT's ability to create hallucinations that trick experts should be considered, as well, when analyzing the veracity of its explanations. Lastly, ChatGPT is limited to data from before 2022 which may affect answer accuracy due to updated medical guidelines.

Conclusions

The remarkable improvement of ChatGPT in performance and usability comes with a word of caution regarding its current limitations. Incorporating ChatGPT and other LLMs into medical education might not be similar for all medical specialties. We observed in this study that ChatGPT answered correctly fewer dermatology questions as compared to the general USMLE licensing exams observed by prior publications. Another recent study in gastroenterology produced similar findings, highlighting the need for further improvement of ChatGPT in medical fields (29). While the potential for ChatGPT to be incorporated into dermatology education and practice is of great interest, future iterations may need to incorporate image recognition and computer vision, as well as larger dermatology-based training data, to be of additional utility. LLMs are part of natural language processing which inherently trains ChatGPT and other LLMs via text databases to create text outputs. The reinforcement training emphasizes responses that ChatGPT presumes the user would like to know, using plausible-sounding responses without focusing on logic or correctness. Subsequent versions of ChatGPT that utilize datasets of labeled images combined with text under supervised training may be a more accurate tool for image-based diagnosis. There are applications for ChatGPT and LLMs generally to be used in medical education, but caution is warranted when utilizing them as a primary learning source.

Acknowledgments

Funding: None.

Footnote

Data Sharing Statement: Available at https://jmai.amegroups.com/article/view/10.21037/jmai-23-47/dss

Peer Review File: Available at https://jmai.amegroups.com/article/view/10.21037/jmai-23-47/prf

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at https://jmai.amegroups.com/article/view/10.21037/jmai-23-47/coif). The authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

References

1. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019;25:44-56.
2. Kantarjian H, Yu PP. Artificial Intelligence, Big Data, and Cancer. JAMA Oncol 2015;1:573-4.
3. Majumder A, Sen D. Artificial intelligence in cancer diagnostics and therapy: current perspectives. Indian J Cancer 2021;58:481-92.
4. Xiang Y, Zhao L, Liu Z, et al. Implementation of artificial intelligence in medicine: Status analysis and development suggestions. Artif Intell Med 2020;102:101780.
5. Bavarian M, Jun H, Tezak N, et al. Efficient Training of Language Models to Fill in the Middle. ArXiv 2022. Available online: https://doi.org/10.48550/arXiv.2207.14255
6. Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback. ArXiv 2022. Available online: https://doi.org/10.48550/arXiv.2203.02155
7. OpenAI. ChatGPT: Optimizing Language Models for Dialogue. OpenAI. Published November 30, 2022 Accessed February 27, 2023. Available online: https://openai.com/blog/chatgpt/
8. Thorbecke C. Google unveils its ChatGPT rival | CNN Business. CNN. Published February 6, 2023. Accessed

February 13, 2023. Available online: https://www.cnn.com/2023/02/06/tech/google-bard-chatgpt-rival/index.html
9. Cheng JL. Baidu leaps to 11-month high as it reveals plan to launch ChatGPT-style "Ernie Bot." CNBC. Available online: https://www.cnbc.com/2023/02/07/baidu-shares-leaps-as-it-reveals-plan-for-chatgpt-style-ernie-bot.html
10. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digit Health 2023;2:e0000198.
11. Scoring & Score Reporting. USMLE. Available online: https://www.usmle.org/bulletin-information/scoring-and-score-reporting. Accessed February 8, 2023.
12. Terwiesch C. Would ChatGPT-3 Get a Wharton MBA? Mack Institute for Innovation Management. Published January 24, 2023. Accessed February 8, 2023. https://mackinstitute.wharton.upenn.edu/2023/would-chat-gpt3-get-a-wharton-mba-new-white-paper-by-christian-terwiesch/
13. Huh S. Are ChatGPT's knowledge and interpretation ability comparable to those of medical students in Korea for taking a parasitology examination?: a descriptive study. J Educ Eval Health Prof 2023;20:1.
14. Antaki F, Touma S, Milad D, et al. Evaluating the Performance of ChatGPT in Ophthalmology: An Analysis of Its Successes and Shortcomings. Ophthalmol Sci 2023;3:100324.
15. Rao A, Kim J, Kamineni M, et al. Evaluating ChatGPT as an Adjunct for Radiologic Decision-Making. medRxiv 2023:2023.02.02.23285399. doi: 10.1101/2023.02.02.23285399.
16. Roose K. Don't Ban ChatGPT in Schools. Teach With It. The New York Times. Published January 12, 2023. Accessed February 9, 2023. Available online: https://www.nytimes.com/2023/01/12/technology/chatgpt-schools-teachers.html
17. Else H. Abstracts written by ChatGPT fool scientists. Nature 2023;613:423.
18. Berry NA, Harvey JA, Pittelkow MR, et al. Online symptom checkers lack diagnostic accuracy for skin rashes. J Am Acad Dermatol 2023;88:487-8.
19. Liu Y, Jain A, Eng C, et al. A deep learning system for differential diagnosis of skin diseases. Nat Med 2020;26:900-8.
20. Chen ML, Rotemberg V, Lester JC, et al. Evaluation of diagnosis diversity in artificial intelligence datasets: a scoping review. Br J Dermatol 2023;188:292-4.
21. Chan HP, Samala RK, Hadjiiski LM, et al. Deep Learning in Medical Image Analysis. Adv Exp Med Biol 2020;1213:3-21.
22. United States Medical Licensing Examination. Step Exams. Accessed April 21, 2023. Available online: https://www.usmle.org/step-exams
23. Ulman CA, Binder SB, Borges NJ. Assessment of medical students' proficiency in dermatology: Are medical students adequately prepared to diagnose and treat common dermatologic conditions in the United States? J Educ Eval Health Prof 2015;12:18.
24. Huynh LM, Bonebrake BT, Schultis K, et al. New Artificial Intelligence ChatGPT Performs Poorly on the 2022 Self-assessment Study Program for Urology. Urol Pract 2023;10:409-15.
25. Khurana S, Vaddi A. ChatGPT From the Perspective of an Academic Oral and Maxillofacial Radiologist. Cureus 2023;15:e40053.
26. AMBOSS GmbH. AMBOSS Dermatology QBank. Available online: https://amboss.com/. Accessed February 4, 2023.
27. AMBOSS | ACCME. www.accme.org. Accessed February 13, 2023. Available online: https://www.accme.org/find-cme-provider/amboss
28. Website: Question Difficulty. AMBOSS. Accessed February 10, 2023. Available online: https://support.amboss.com/hc/en-us/articles/360035679652-Question-difficulty
29. Suchman K, Garg S, Trindade AJ. Chat Generative Pretrained Transformer Fails the Multiple-Choice American College of Gastroenterology Self-Assessment Test. Am J Gastroenterol 2023. [Epub ahead of print]. doi: 10.14309/ajg.0000000000002320.

doi: 10.21037/jmai-23-47
Cite this article as: Behrmann J, Hong EM, Meledathu S, Leiter A, Povelaitis M, Mitre M. Chat generative pre-trained transformer's performance on dermatology-specific questions and its implications in medical education. J Med Artif Intell 2023;6:16.

Supplementary

Table S1 Chi-squared test for question difficulty distribution

Difficulty level	Correct	Incorrect	Total	P value
Observed				<0.001
1	71	43	114	
2	63	79	142	
3	48	101	149	
4	18	49	67	
5	4	16	20	
Total	204	288	492	
Expected				<0.001
1	47	67	114	
2	59	83	142	
3	62	87	149	
4	28	39	67	
5	8	12	20	
Total	204	288	492	

Observed versus expected values for difficulty level distribution by test. Expected values were calculated from a chi-squared test using the 41% overall correct ratio (204/492).

https://dx.doi.org/10.21037/jmai-23-47