Use of AI (GPT-4)-generated multiple-choice questions for the examination of surgical subspecialty residents: Report of feasibility and psychometric analysis

Jin Kyu Kim[1,2], Michael Chua[1,2], Armando Lorenzo[1,2], Mandy Rickard[2], Laura Andreacchi[1], Michael Kim[1], Douglas Cheung[1], Yonah Krakowsky[1,3], Jason Y. Lee[1,4]
[1]Division of Urology, Department of Surgery, University of Toronto, Toronto, ON, Canada; [2]Division of Urology, Department of Surgery, The Hospital for Sick Children, Toronto, ON, Canada; [3]Division of Urology, Department of Surgery, Women's College Hospital and Mount Sinai Hospital, Toronto, ON, Canada; [4]Division of Urology, Department of Surgery, University Health Network, Toronto, ON, Canada

Cite as: Kim JK, Chua M, Lorenzo A, et al. Use of AI (GPT-4)-generated multiple-choice questions for the examination of surgical subspecialty residents: Report of feasibility and psychometric analysis. *Can Urol Assoc J* 2025 February 24; Epub ahead of print. http://dx.doi.org/10.5489/cuaj.9020

Published online February 24, 2025

Corresponding author: Dr. Jin Kyu Kim, Division of Urology, Hospital for Sick Children, Toronto, ON, Canada; jjk.kim@mail.utoronto.ca

***

ABSTRACT

Introduction: Multiple-choice questions (MCQs) are essential in medical education and widely used by licensing bodies. They are traditionally created with intensive human effort to ensure validity. Recent advances in AI, particularly large language models (LLMs), offer the potential to streamline this process. This study aimed to develop and test a GPT-4 model with customized instructions for generating MCQs to assess urology residents.
Methods: A GPT-4 model was embedded using guidelines from medical licensing bodies and reference materials specific to urology. This model was tasked with generating MCQs designed to mimic the format and content of the 2023 urology examination outlined by the Royal College of Physicians and Surgeons of Canada (RCPSC). Following generation, a selection of MCQs underwent expert review for validity and suitability.

KEY MESSAGES

- A GPT-4 model trained using urology-related guidelines and reference materials produced 123 MCQs; 60 were used for a formal exam administered to residents in various postgraduate years.
- Results showed a correlation between resident performance and level of training, with a noticeable progression of scores by cohort. The study also analyzed the difficulty, discriminatory power, and effectiveness of distractors in the questions.
- AI-driven models like GPT-4 can enhance the efficiency of MCQ creation for medical exams while maintaining question quality, although ethical considerations and further validation are essential for implementation.

Results: From an initial set of 123 generated MCQs, 60 were chosen for inclusion in an exam administered to 15 urology residents at the University of Toronto. The exam results demonstrated a general increasing performance with level of training cohorts, suggesting the MCQs' ability to effectively discriminate knowledge levels among residents. The majority (33/60) of the questions had discriminatory value that appeared acceptable (discriminatory index 0.2–0.4) or excellent (discriminatory index >0.4).
Conclusions: This study highlights AI-driven models like GPT-4 as efficient tools to aid with MCQ generation in medical education assessments. By automating MCQ creation while maintaining quality standards, AI can expedite processes. Future research should focus on refining AI applications in education to optimize assessments and enhance medical training and certification outcomes.

INTRODUCTION

Since its invention in 1914, multiple-choice questions (MCQ) have been widely used for testing across many disciplines.[1] This is without exception in medicine, where urology, along with any other medical specialities, rely on this testing tool for examination leading to board-certification and licensure in North America.[2,3] Creating a well-formulated MCQ is a difficult task, especially when designed for testing high-level and high-stake examinations, such as critical test to determine advancement in training or granting medical license.[4] However, there are principles to creating suitable MCQs, sentiment is transparent and shared across different licensing organizations.[5,6] In the context of medical licensing exams, these principles include having a clear lead-in question that is of appropriate difficulty with clinically relevant scenarios, adequate testing application of medical knowledge, and generating fair distractors.

The task of generating MCQs demands qualified individuals to spend several hours formulating such questions and discussing their validity.[7] However, with the 'rules' available to make appropriate MCQs, artificial intelligence (AI), specifically large language models (LLM) such as ChatGPT, may be instructed to create these questions in accordance with these regulations. Previous studies have shown utility of LLM in medical education, including generating and answering MCQs across various disciplines. For instance, broader MCQ generation using LLMs like ChatGPT has been explored in recent literature, emphasizing its potential and limitations.[8] Moreover, with the advent of newer LLM that can analyze the contents of files such as portable document format (PDF), LLMs can now access knowledge to assist creation of MCQ with reliable content. Herein, we report the novel case of high level MCQ creation using GPT-4 model with customized instructions for the purpose of testing surgical subspecialty (urology) residents on their knowledge.

METHODS

Following institutional research ethics board approval (#1000081309), a GPT-4 model with customized instructions using ChatGPT-4 from OpenAI (openai.com) was created with instructions specific to create MCQ questions, based on established guidelines on how to write well-designed MCQ items, available from the Royal College of Physicians and Surgeons of Canada (RCPSC) and the National Board of Medical Examiners (NBME®). Custom instructions included rules for MCQ building, recommendations for drafting good stems and lead-ins, description of suitable distractors, and number of options, which was limited to four (Appendix A).

Sample MCQ were generated in December 2023 using this GPT-4 model with customized instructions by providing a PDF of publicly-available guidelines and reference study materials according to 2023 urology examination format provided by the RCPSC. Through retrieval-augmented generation (RAG), LLMs such as GPT-4 extract relevant text data from uploaded documents to generate responses. In RAG, the model retrieves information from external documents and combines it with its existing knowledge base to create more contextually accurate outputs. This approach enables the model to use specific, credible sources, thereby ensuring that the generated MCQs are well-aligned with current guidelines.[9] We included Canadian Urological Association (CUA) clinical guidelines, CUA best practice reports, and CUA Journal review articles (Example sample question creation: https://chatgpt.com/share/45e7bab5-d942-4c21-9ebd-3742de166a7d; for this specific question, CUA guideline for neurogenic lower urinary tract dysfunction was uploaded to LLM chat).[10] The American Urological Association (AUA) guidelines were not included due to new policy enforcing restrictions on AUA-related content exposure to LLM models. While not explicitly stated as a reference study material by RCPSC, European Urological Association (EUA) guidelines were also used.

Generated MCQ were screened by one or more Royal College certified author(s) for content validity. Questions of appropriate difficulty were identified and a randomly selected 60-question MCQ exam was created with its contents mimicking the anticipated proportion of question topics per 2023 urology examination format provided by RCPSC. This exam was distributed among urology residents training at University of Toronto on June 7, 2024 in a monitored setting. The proportion of questions answered correctly was evaluated by year of training and topic for psychometric analysis, which included difficulty assessment, discriminatory index calculation (formula defined as % correct by top 27th percentile performers minus % correct by bottom 27th percentile performers; top/bottom 27th percentile cut-off were used per previously established conventional cut-off for high- and low-performers in literature), and functional distractor evaluation (defined as options chosen by >5% of participants).[11-13]

Privacy concerns surrounding LLM use was addressed by opting out of model training based on contents provided by the user account, as well as utilizing the ChatGPT Team account, on which OpenAI guarantees privacy without training based on uploaded content.[14]

RESULTS

Exam construction

A total of 123 sample questions were generated using our GPT-4 model. Following screening, 6 questions (4.9%, 6/123) had undergone modifications per at least one screener's suggestion due to possible ambiguity of clinical scenario (3/6, 50%) or answer choices (3/6, 50%). The remaining 117 questions appeared suitable for testing with good construct validity (Questions requiring modifications to improve ambiguity are shown in Supplementary Table 1).

Of the 123 questions, 60 were chosen for testing by the author group. Questions were selected based on topic distribution suggested by 2023 RCPSC urology examination format and randomly chosen within our question bank of 123 questions. The proportion of topics in this exam are summarized in Figure 1 and the examined questions are available in Supplementary File 1.

Participants

Fifteen urology residents training at University of Toronto were tested using GPT-4 generated MCQ exam on June 7, 2024. Tested cohorts were from post-graduate year (PGY) 1 to 4, near the end of their academic year (July 1, 2023-June 30, 2024). The median score was 37/60 (range 31-44).

There were five PGY-1s who wrote the exam. Among the PGY-1 cohort, two were international medical graduates who have already completed urology residency training in another country prior to joining our PGY-1 cohort and were evaluated separately from their non-IMG peers. In addition, there were three PGY-2s, four PGY-3s, and three PGY-4s who wrote the exam. PGY-5s were not tested as they have already passed their RCSPC examination and have contributed to screening of the questions.

Psychometric analysis

Among the questions, 21 were deemed difficult with less than 50% correct response rates. In contrast, 14 questions were deemed easy, with >80% correct response rates.

The discriminatory index was calculated for each question. Twenty-seven questions (45%) had poor discriminatory value (<0.2), and thirty-three questions 65% had acceptable or excellent discriminatory values (20 acceptable [0.2-0.4], 13 excellent [>0.4]).

There were 98 functional distractors (98/180, 54%) with >5% selection rates from participants. There was poor correlation with number of functional distractors and discriminatory index (Pearson r=0.150; Supplementary Figure 1).

When evaluating percentage correct per topic, residents performed well on pediatrics, trauma/fistula, and basic sciences, with >70% average score. In contrast, they performed poorly on transplantation/nephrology/renovascular disease, diagnostic imaging, infection/inflammation, and neurology/voiding dysfunction with <60% correct rate (Figure 3).

There was no significant correlation to discriminatory index and proportions correct questions in each topic. However, there was a significant correlation between the median

functional distractors per topic and the proportion of questions correct per topic (Pearson r=-0.759, p=0.004; Supplementary Table 2).

Exam results

The scores, per cohort, are shown in Figure 2A/Table 1. This showed progression of improved scores based on the level of training in urology. The PGY-1 IMGs who have more experience than their cohort performed better in their examination and their scores were comparable to the senior residents. While PGY-3s performed very well, the PGY-4s have performed with less variation in their scores. When excluding MCQs that had poor discriminatory index, there was a noticeable stepwise increase in scores with progressing PGY level (Figure 2B).

DISCUSSION

Until present, LLM have been primarily used to answer MCQ to assess performance on standardized exams. Recent GPT-4 models have performed remarkably well on standardized examination questions, suggesting it can be well-trained on various topics including medicine.[15] The accuracies are often reported to be beyond the passing score of 70% across several disciplines.[16-19] This suggests that LLM such as ChatGPT have significant contextual data and can create questions that are clinically relevant and reasonable for the examinees. There has also been improving performance in examinations specific to urology, but its performance as a subspecialty expert is still inferior to top performing human counterparts.[20,21] There are preliminary reports of using GPT-4 to create MCQs designed to test surgical subspecialty residents. However, these authors did not customize the GPT-4 model with specific instructions around MCQ building and also did not provide appropriate reference materials. This led to increased susceptibility to "hallucination" and poorer question quality, with only 25% of their questions reaching discriminatory index value of >0.2.[22] To our knowledge, this is the first study to create high-level MCQs designed for testing medical subspecialty residents and to evaluate the feasibility to implement this in a formal examination setting.

Our study showed that the exam did have some discriminatory value in evaluating the knowledge of urology residents, based on their training year. Due to the small number of participants, evaluation of statistical differences among cohorts were not performed. There are improvements that can be made. There were nine questions (15%) of the exam that were of limited utility in evaluating resident knowledge as there was 100% or 0% correct rates, with 45% of questions having poor discrimination. Moreover, some topics were more difficult than others, likely in part due to higher proportion of functional distractors that were present in questions covering those topics. As there were fewer numbers of questions in some of these topics, caution is needed to interpret the resident knowledge-base based on these questions alone. However, this may also reflect the design of the residency program exposure (early exposure to pediatric rotations, surgical foundations, and trauma rotation as a PGY-1/2 leading to higher marks). Additional evaluation of "difficult" and "easy" questions, more difficult questions were testing concepts that are not frequently encountered or tested during clinical settings such as pheochromocytoma follow up and different catheter types that may reduce urinary tract

infections. Despite this, most of the questions were acceptable or excellent in discriminatory index classification, suggesting that there may be great value in utilizing LLMs in supporting the creation of MCQs. This is further corroborated when evaluating performance based on questions with acceptable or excellent discriminatory value, there is a clear trend to higher scores in the more senior cohort. While PGY-3s performed better than the PGY-4s as an average, there was less variation across PGY-4 marks, suggesting that as a complete cohort, that they may be better prepared compared to the PGY-3s. Drafting MCQs for examination is a complex process and there are measures by organizations such as American Board of Urology to experiment with a proportion of their MCQ exam in order to safeguard question quality. It is undisclosed how many of such questions become testable questions in the future, but the questions created by our custom GPT performed well in creating a good mix of easy, moderate, and difficult questions.[23]

The correlation analysis reveals key factors that contribute to both the difficulty of a question and its discriminatory power. A strong negative correlation (-0.78) between the percentage of correct responses and the number of functional distractors suggests that questions with more effective distractors are generally harder, as fewer respondents can choose the correct answer. Regarding discriminatory power, the data indicate that the number of functional distractors has a positive correlation (0.15) with the discriminatory index, suggesting that questions with more plausible distractors are better at distinguishing between high- and low-performing respondents. The relatively weak correlations between the discriminatory index and other factors, such as the percentage of correct responses (-0.15), indicate that the ability of a question to discriminate between performance levels is not solely dependent on how many people answer correctly, but rather on the quality of the distractors. Future LLMs used for this purpose should be modified to ensure that at least one or more plausible distractors are included in its options (with or without help of humans) to enhance its discriminatory index. Moreover, this may be an opportunity to study the functional distractors that are frequently chosen by residents to identify potential common misconceptions and create tailored educational content. By augmenting LLMs to assess individual trainee weaknesses, we can also potentially create a system that generates personalized questions to address these gaps.

Barriers to using presented GPT-4 based models in the future include limited number of custom GPT-4 use by non-subscribers to OpenAI's ChatGPT service. There have also been privacy concerns with LLM, with models being trained on subscription or propriety content without permission. In the era of artificial intelligence and LLM, there has been increasing efforts to protect propriety data. In case of American Urological Association (AUA), whose guideline contents are also testable resources according to RCPSC, they have released a recent policy that their contents should not be uploaded to LLM without permission. The results of our study suggest that useable MCQ creation can be made significantly less cumbersome using LLM. Thus, agencies that may benefit from such increased efficiency should aim to develop partnership with those with expertise in LLM, as well as organizations that have propriety content of interest to ensure creation of appropriate MCQ without breach in privacy policies.

There are several limitations to this investigation. While it shows feasibility of LLM use in creation of useable and high quality MCQs for subspecialty resident examination, this evaluation was performed in a small cohort of 15 individuals from a single residency program in Canada.

Further internal and external validity for ongoing reliable use of such tools is necessary prior to routine use. Moreover, this investigation did not include the graduating cohort of PGY5s who have already passed their licensing examination and would be the ideal gold standard to assess whether the exam is representative of the RCPSC urology examination. There may also be a selection bias with the questions that were chosen – however, we aimed to minimize this by selecting the questions randomly from the database of questions that were deemed appropriate in quality by screeners.

Furthermore, with use of LLMs, one must be cognizant of how it was trained and developed.[24] While it is uncertain what training data GPT-4 models had access to, we ensured that proper resources were provided by augmenting our custom GPT with reliable resources. Moreover, RAG is a complex process that relies on the model to locate and extract the most pertinent information in a vast array of data, which is particularly challenging due to 'needle in a haystack' nature of identifying the most relevant information from extensive data sources and LLM's ability to perform RAG is often compared between different models for their performances. It has been shown that LLMs may miss information, especially as length of provided content becomes longer.[25] Nonetheless, we suspect that LLMs' capacity to perform RAG will also continue to improve in the future with advanced techniques such as improved indexing, relevance ranking, and context-aware retrieval strategies.[9,26] While studies reporting use of older LLMs may soon become obsolete as newer LLMs are continually introduced, we focused on applicability of LLMs that will continue to be relevant for future LLM users. As use of AI in medical education becomes increasingly common, there may be benefit in collaboration of computer scientists and AI-experts to evolve our methodology and create higher level use of LLMs through computer programming. We hope to continue validation of our tool in our residency program cohorts, as well as other institutions in the future for external validation.

CONCLUSIONS

This study highlights the feasibility and utility of utilizing AI-driven models, such as GPT-4, in the creation of MCQs for medical education and assessment. By leveraging AI, the process of MCQ generation can be expedited while maintaining standards of validity and relevance. One should be cognizant of ethics around using LLMs and ensuring questions are developed from reliable sources. Future research could further explore and refine AI applications in educational assessment, potentially revolutionizing the efficiency and effectiveness of medical training and certification examinations.

REFERENCES

1. National Education Association. History of standardized testing in the United States. *NEA* 2020. Accessed June 6, 2024. https://www.nea.org/resource-library/history-standardized-testing-united-states
2. Royal College of Physicians and Surgeons of Canada. Multiple-choice question (MCQ) format information. *Royal College of Physicians and Surgeons of Canada* 2019. Accessed June 6, 2024. https://www.royalcollege.ca/en/credentials-exams/assessment-documents/multiple-choice-question-format-information.html
3. American Board of Urology. Qualifying (Part 1) exam content. *American Board of Urology*. Accessed June 6, 2024. https://abu.org/certification/qualifying-examination/exam-content
4. Javaeed A. Assessment of higher ordered thinking in medical education: multiple-choice questions and modified essay questions. *MedEdPublish* 2018;7:128. https://doi.org/10.15694/mep.2018.0000128.1
5. Royal College of Physicians and Surgeons of Canada. Guidelines for the development of multiple-choice questions. *Royal College of Physicians and Surgeons of Canada* 2008. Accessed June 6, 2024. https://www.canadiancriticalcare.org/resources/Documents/GuidelinesforDevelopmentMCQRoyalCollege.pdf
6. National Board of Medical Examiners (NBME). Item writing guide. *NBME*. Accessed June 6, 2024. https://www.ucns.org/common/Uploaded%20files/Help/NBME%20Item%20Writing%20Guide.pdf
7. Jørgensen M, Savran MM, Christakopoulos C, et al. Development and validation of a multiple-choice questionnaire-based theoretical test in direct ophthalmoscopy. *Acta Ophthalmol* 2019;97:700-6. https://doi.org/10.1111/aos.14065
8. Cheung BHH, Lau GKK, Wong GTC, et al. ChatGPT versus human in generating medical graduate exam multiple-choice questions—a multinational prospective study (Hong Kong S.A.R., Singapore, Ireland, and the United Kingdom). *PLoS One* 2023;18:e0290691. https://doi.org/10.1371/journal.pone.0290691
9. Lewis P, Perez E, Piktus A, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Adv Neural Inf Process Syst* 2020;33:9459-74.
10. Kavanagh A, Baverstock R, Campeau L, et al. Canadian Urological Association guideline: diagnosis, management, and surveillance of neurogenic lower urinary tract dysfunction—full text. *Can Urol Assoc J* 2019;13:E157-76. https://doi.org/10.5489/cuaj.5912
11. Rezigalla AA, Eleragi AMESA, Elhussein AB, et al. Item analysis: the impact of distractor efficiency on the difficulty index and discrimination power of multiple-choice items. *BMC Med Educ* 2024;24:445. https://doi.org/10.1186/s12909-024-05433-y
12. Kumar D, Jaipurkar R, Shekhar A, et al. Item analysis of multiple-choice questions: a quality assurance test for an assessment tool. *Med J Armed Forces India* 2021;77:S85-9. https://doi.org/10.1016/j.mjafi.2020.11.007
13. Tarrant M, Ware J, Mohammed AM. An assessment of functioning and non-functioning distractors in multiple-choice questions: a descriptive analysis. *BMC Med Educ* 2009;9:40. https://doi.org/10.1186/1472-6920-9-40

14. OpenAI. How your data is used to improve model performance. *OpenAI Help Center*. Accessed June 6, 2024. https://help.openai.com/en/articles/5722486-how-your-data-is-used-to-improve-model-performance
15. Newton P, Xiromeriti M. ChatGPT performance on multiple-choice question examinations in higher education: A pragmatic scoping review. *Assess Eval High Educ* 2024;1-18. https://doi.org/10.1080/02602938.2023.2299059
16. Alexandrou M, Mahtani AU, Rempakos A, et al. Performance of ChatGPT on ACC/SCAI interventional cardiology certification simulation exam. *JACC Cardiovasc Interv* 2024;17:1292-3. https://doi.org/10.1016/j.jcin.2024.03.012
17. Alessandri-Bonetti M, Liu HY, Donovan JM, et al. A comparative analysis of ChatGPT, ChatGPT-4, and Google Bard performances at the Advanced Burn Life Support Exam. *J Burn Care Res* 2024. https://doi.org/10.1093/jbcr/irae044
18. Rojas M, Rojas M, Burgess V, et al. Exploring the performance of ChatGPT versions 3.5, 4, and 4 with vision in the Chilean Medical Licensing Examination: Observational study. *JMIR Med Educ* 2024;10:e55048. https://doi.org/10.2196/55048
19. Abbas A, Rehman MS, Rehman SS. Comparing the performance of popular large language models on the National Board of Medical Examiners sample questions. *Cureus* 2024;16:e55991. https://doi.org/10.7759/cureus.55991
20. Sherazi A, Canes D. Comprehensive analysis of the performance of GPT-3.5 and GPT-4 on the American Urological Association self-assessment study program exams from 2012-2023. *Can Urol Assoc J* 2023.
21. Touma NJ, Caterini J, Liblk K. Performance of artificial intelligence on a simulated Canadian urology board exam: Is ChatGPT ready for primetime? *Can Urol Assoc J* 2024;18:329-32. https://doi.org/10.5489/cuaj.8800
22. Touma NJ, Skinner T, Leveridge M, et al. Artificial intelligence as a discriminator of competence in urologic training: Are we there? *J Urol* 2024. Epub ahead of print.
23. The American Board of Urology. Qualifying (Part 1) exam content. *The American Board of Urology* 2024. Accessed June 25, 2024. https://www.abu.org/certification/qualifying-examination/exam-content
24. Kim JK, Chua M, Rickard M, et al. ChatGPT and large language model (LLM) chatbots: The current state of acceptability and a proposal for guidelines on utilization in academic medicine. *J Pediatr Urol* 2023;19:598-604. https://doi.org/10.1016/j.jpurol.2023.05.018
25. Chaudhury S, Dan S, Das P, et al. Needle in the haystack for memory-based large language models. *arXiv preprint* arXiv:2407.01437. 2024.
26. Karpukhin V, Oğuz B, Min S, et al. Dense passage retrieval for open-domain question answering. 2020. https://doi.org/10.18653/v1/2020.emnlp-main.550

FIGURES AND TABLES

Figure 1. Breakdown of proportion of examination content (percentages within brackets are the suggested proportion of questions according to 2023 Royal College Urology Examination).

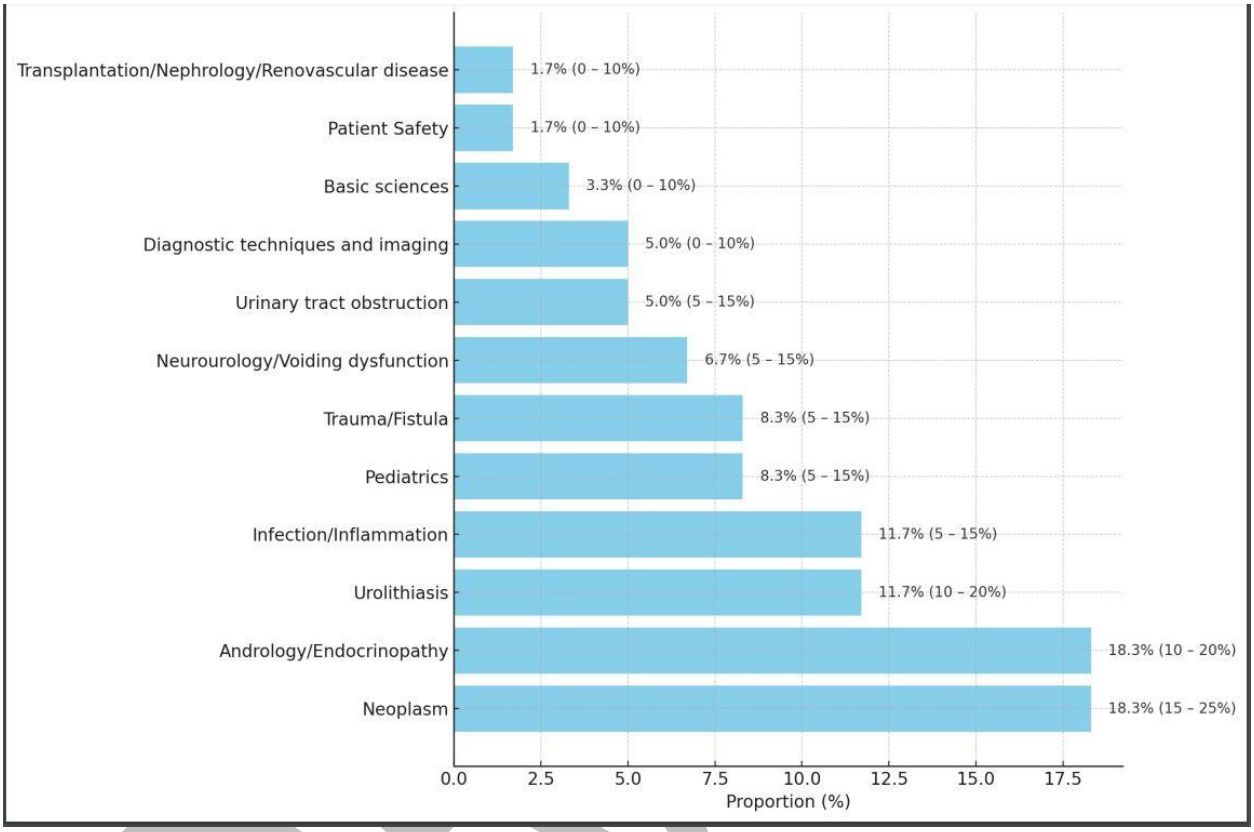

Figure 2. Summary of raw scores per postgraduate year training cohort: (A) summary of results for all questions (n=60); (B) summary of results for questions with acceptable and high discriminatory value (n=33).

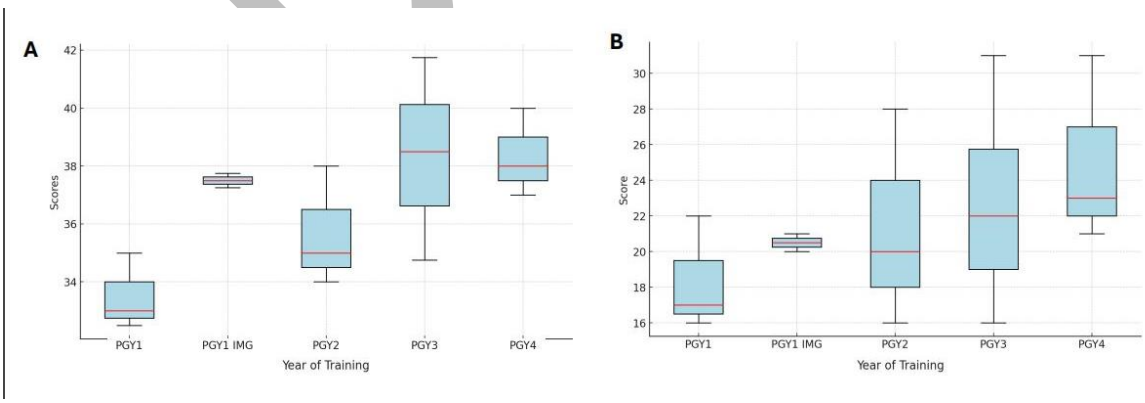

Figure 3. Summary of the proportion of questions correct per topic.

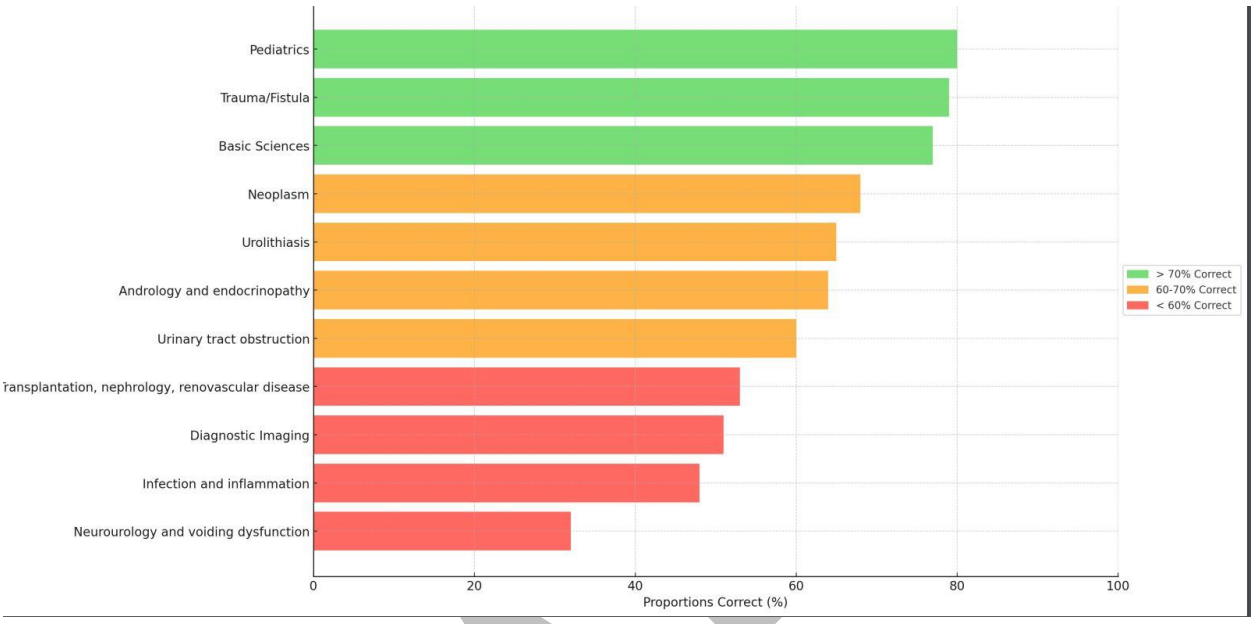

Table 1. Summary of test results			
Results (all questions)			
Year of training	**Median score (out of 60)**	**25th percentile**	**75th percentile**
PGY1	33	32.5	35
PGY1 IMG	37.5	37.25	37.75
PGY2	35	34	38
PGY3	38.5	34.75	41.75
PGY4	38	37	40
Results (with poor discriminatory index questions removed)			
Year of training	**Median score (out of 33)**	**25th percentile**	**75th percentile**
PGY1	17	16.5	19.5
PGY1 IMG	20.5	20.25	20.75
PGY2	20	18	24
PGY3	22	19	25.75
PGY4	23	22	27

PGY: postgraduate year.