Exploring ChatGPT's potential in the context of colon cancer patient education: proof-of-concept study

Abdulwhhab Abu Alamrain, Mary Adewunmi, Mahmoud Abu Al Amrain, Ming Chao Wong, Kwang Chien Yee

Submitted to: JMIR Cancer
on: August 01, 2023

Table of Contents

Exploring ChatGPT's potential in the context of colon cancer patient education: proof-of-concept study

Abdulwhhab Abu Alamrain[1, 2] MD; Mary Adewunmi[3] PhD; Mahmoud Abu Al Amrain[4] MD; Ming Chao Wong[3] PhD; Kwang Chien Yee[3] MD, PhD

[1]Surgery Department Al-Aqsa Martyr's Hospital Palestinian Ministry Of Health Gaza Strip PS
[2]Faculty of Medicine Al-Quds University Gaza Strip PS
[3]CHM, School of Medicine University of Tasmania Hobart AU
[4]Faculty of Medicine Islamic University Gaza Strip PS

Corresponding Author:
Mary Adewunmi PhD
CHM, School of Medicine
University of Tasmania
TAS7000
Hobart
AU

Abstract

Background: ChatGPT is a large language model capable of generating human-like conversation. It has demonstrated promise as a tool for medical education for both professionals and patients. Previous research in medical oncology and colon cancer showed a glimpse of its application on topics like colonoscopy, colorectal surgery, and guideline-based treatment.

Objective: To evaluate ChatGPT's performance as a source of patient medical education for colon cancer

Methods: A set of twenty non-expert questions were prepared and fed to ChatGPT three times. Later, generated responses were evaluated by two doctors for accuracy, simplicity (0-10), and consistency (0,1). Mean, median and standard deviation were calculated for both accuracy and simplicity scores along with the Intraclass Correlation Coefficient and confidence interval for inter-rater agreement assessment. For consistency, rate, cohen's kappa, standard error, and confidence interval were calculated.

Results: Accuracy: Mean = 8.4, Median = 8.5, SD = 1.7. ICC: Avg. measures for absolute agreement = 0.7 (95% CI 0.25 to 0.88), for consistency = 0.74 (95% CI 0.34 to 0.9).
Simplicity: Mean = 8.55, Median = 9, SD = 1.69. ICC: Avg. measures for absolute agreement = 0.65 (95% CI 0.12 to 0.86), for consistency = 0.72 (95% CI 0.28 to 0.89).
Consistency: rate = 67.5%, Cohen's Kappa: 0.66 (SE = 0.18, 95% CI 0.31 to 1.0).

Conclusions: In this study, we assessed ChatGPT's capabilities of answering patients' questions about colon cancer. Findings showed significant and promising results of answers' accuracy, simplicity, and consistency in multiple trials. However, there is room for improvement. As ChatGPT continues to gain popularity among users, research studies on the impact of this technology on patient outcomes are needed urgently to guide clinical application.

(JMIR Preprints 01/08/2023:51444)
DOI: https://doi.org/10.2196/preprints.51444

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Exploring ChatGPT's potential in the context of colon cancer patient education: proof-of-concept study

Abstract

Background:
ChatGPT is a large language model (LLM) capable of generating human-like conversations. It has demonstrated promise as a medical education tool for professionals and patients. Previous research in medical oncology showed a glimpse of its application to topics like colonoscopy, colorectal surgery, and guideline-based treatment but has not been fully explored in patient education.
Objective:
To evaluate ChatGPT's performance as a source of patient medical education for colon cancer.
Methods:
A set of twenty non-expert questions was prepared and fed to ChatGPT three times. Later, generated responses were evaluated by two doctors for accuracy, simplicity (1-10), and consistency (0,1). Mean, median, and standard deviation were calculated for both accuracy and simplicity scores along with the Intraclass Correlation Coefficient (ICC) and confidence interval (CI) for inter-rater agreement assessment. For consistency, rate, cohen's kappa, standard error, and confidence interval were calculated.
Results:
Accuracy: mean = 8.4 (SD 1.7), median = 8.5. ICC: Avg. measures for absolute agreement = 0.7 (95% CI 0.25 to 0.88), for consistency = 0.74 (95% CI 0.34 to 0.9).
Simplicity: Mean = 8.55 (SD 1.69), Median = 9. ICC: Avg. measures for absolute agreement = 0.65 (95% CI 0.12 to 0.86), for consistency = 0.72 (95% CI 0.28 to 0.89).
Consistency: rate = 67.5%, Cohen's Kappa: 0.66 (SE = 0.18, 95% CI 0.31 to 1.0).
Conclusions:
This study aimed to evaluate the proficiency of ChatGPT in responding to inquiries from patients regarding colon cancer. Findings demonstrated promising results in terms of accuracy, simplicity, and consistency over several trials. However, challenges such as lack of referencing and outdated information exist. As ChatGPT gains popularity, further research is needed to understand its usefulness in patient education and clinical application

Keywords: colon cancer; colorectal cancer; ChatGPT; LLM; patient education; medical education; screening, treatment

Introduction

ChatGPT, developed by OpenAI, is a large language model (LLM) capable of human-like conversation. It is built on Generative Pretrained Transformer (GPT) 3.5, a technology that was released in November 2022 [1]. Its potential in medical education has sparked interest among researchers, offering rapid access to information for professionals and patients alike.

ChatGPT in medical education

Medical education is a dynamic field with various pedagogical approaches and frameworks. It aims to produce competent healthcare professionals dedicated to patient care and continuous learning [2]. ChatGPT's ability to quickly provide medical information showed promise in various contexts, including, assisting medical students with USMLE questions, OSCE preparation, and even medical imaging tasks [3]–[6].

ChatGPT in cancer, colon cancer, and colon

Since its inception, ChatGPT has garnered significant popularity and demonstrated utility in providing accurate medical information including in the field of oncology [7], [8]. These range from debunking myths to aiding in screening, prevention, and treatment recommendations along with patient-provider communication for cancers like breast and pancreatic cancer [9]–[11]. Regarding colon-related matters, ChatGPT has proven effective in improving clinical communication about colonoscopy procedures and aids in adhering to post-colonoscopy surveillance protocols, providing surgical and medical treatment recommendations in line with cancer guidelines [12]–[15].

With these capabilities in mind, this study seeks to investigate the potential of using ChatGPT in the domains of colon cancer screening, diagnosis, and treatment as an information source for patient education.

Methods

In this proof-of-concept study, we assessed ChatGPT's potential as a tool for patient education about colon cancer. A set of twenty non-expert questions were prepared using both ChatGPT's recommended and author- generated questions. ChatGPT questions were acquired using the following prompt: "Top 10 most common questions people ask about colon cancer", while the other 10 questions were crafted based on author experience.

Questions were presented to the ChatGPT 3.5 version on July 23, 2023. Later, the generated responses were evaluated by two doctors, prof. Kwang, a senior gastroenterologist and associate professor of Medicine, and Dr. Abdulwhhab, a G.P. doctor, whom both have strong cancer research record. The doctors used a highly detailed grading criteria for evaluation. For accuracy and simplicity evaluation was based on answer content, answer organization, word choice, and overall structure.

To further assess the consistency of ChatGPT, all queries were repeated three times. Grading scores were assigned on a scale of 0 to 10 for accuracy and simplicity and on a scale of 0 to 1 for consistency, as shown in Appendix 1. A score of 10 represents the best performance, while a score of 0 indicates the lowest performance.

The statistical analysis was conducted using MedCalc Statistical Software version 22.009 [16]. The study focused on analyzing the mean, median, and standard deviation (SD) for both accuracy and simplicity scores. Additionally, the two-way model Intraclass Correlation Coefficient (ICC) was calculated for both absolute agreement and consistency to assess inter-rater agreement, along with the confidence interval (CI). To assess the consistency across the three trials of the same question, the consistency rate was calculated. The agreement on consistency was assessed using unweighted Cohen's Kappa, along with its standard error (SE) and CI.

By conducting this evaluation, we aim to assess the ability of the publicly available ChatGPT version to provide reliable and accurate information related to colon cancer.

Results

Emphasis on ChatGPT's answers

In this study, ChatGPT was presented with a total of twenty non-expert questions about colon cancer, covering etiology, diagnosis, treatment, and prognosis, and generated answers for them. For example,

it delivered a simple and informative definition of colon cancer (Q1). It also had very good general steps for preparation for the colonoscopy procedure (Q9) and an impressive answer to a question about the survival rate for colon cancer patients (Q20)

However, there were instances where the model's answers were highly accurate but complex, making them challenging for patients to comprehend, as observed in questions about colon cancer staging and chemotherapy's role in treatment (Q10, Q13). Regarding questions 18 and 19, ChatGPT provided almost the same answer for two different questions.

Evaluation Outcomes

Analysis of doctors' evaluations showed a promising result. For accuracy, the mean score was 8.4, with a median of 8.5 and SD of 1.7. ICC measurements indicated substantial agreement between the evaluators, with average measures of 0.7 (95% CI 0.25 to 0.88) for absolute agreement and 0.74 (95% CI of 0.34 to 0.9) for consistency.

On the other hand, the simplicity results were slightly better, with a mean score of 8.55 and a median of 9, along with an SD of 1.69. ICC average measures for simplicity were 0.65 (95% CI 0.12 to 0.86) for absolute agreement and 0.72 (95% CI 0.28 to 0.89) for consistency. (Figure 1) and (Figure 2) demonstrate the results and inter-rater agreement for both accuracy and simplicity.

ChatGPT also demonstrated a consistency rate of 67.5% in providing the same answers to the same questions in all three trials. Cohen's Kappa yielded a value of 0.66, with a SE of 0.18 and a 95% CI ranging from 0.31 to 1.0.

These findings suggest ChatGPT's ability to provide relatively accurate and simple responses to colon cancer questions, supported by agreement between evaluators and consistency across multiple trials. Overall, the model was able to comprehend the questions and provide acceptable answers, demonstrating its potential to assist patients in understanding complex medical topics. However, there is room for improvement.

Discussion

In this study, we evaluated ChatGPT's performance in providing accurate and simple answers for common non-expert questions about colon cancer. Overall, the model's performance was satisfactory. There were instances where it excelled in delivering both accurate and simple answers, while in some cases, it excelled in one aspect but fell short in the other. Notably, we did not encounter any responses that pose potential harm or danger to the patients.

Similar findings have been reported in other studies. For example, ChatGPT provided appropriate responses for most (88%) questions about breast cancer prevention and screening [10] and comparable responses in both cardiovascular diseases (84%)[17] and lung cancer (84%)[18].

However, we faced a significant challenge in assessing accuracy due to the lack of references to guidelines, raising concerns about potential bias. Nevertheless, this aspect was not within the scope of our study. Other limitation includes the small question set and evaluator pool, highlighting the need for larger-scale investigations.

Another limitation relates to the potential outdatedness of information due to the September of 2021 OpenAI cutoff for ChatGPT training data. This highlights an internal issue in the ChatGPT model itself and underscore the necessity for ongoing model development based on latest medical updates.

Our choice of ChatGPT 3.5 was motivated by its public availability, making it accessible to a wide audience, including both professionals and patients.

While ChatGPT shows promise in medical education, careful consideration of content oversight and fact-checking mechanisms is important. Ultimately, further research is essential to fully harness ChatGPT's potential in healthcare education. Similar application to other cancer types or patient education topics can also be done.

Conclusions

In this study, we evaluated ChatGPT's ability to address patient inquiries regarding colon cancer and its potential for patient education. Findings showed promising results in terms of accuracy, simplicity, and consistency across multiple trials of 20 questions, with mean scores of 8.4 (SD 1.7), 8.55 (SD 1.69) and rate of 67.5% respectively. However, challenges such as lack of referencing and outdated information exist. As ChatGPT gains popularity, further research is needed to understand its usefulness in patient education and clinical application.

Acknowledgements

Special thanks go to CARESAI collaborators for their support during the research.

Funding: This study was supported by the University of Tasmania, School of Medicine, (SoM), College of Health and Medicine (CHM), Hobart, Australia.

Conflicts of Interest: The authors have declared no competing interests.

Generative AI was used to generate both questions and answers.

All questions and answers generated, evaluated, and analyzed during this study are included in the supplementary file.

Authors contribution:
Abdulwhhab Abu Alamrain: research idea, methodology development, human evaluator, statistical analysis, data visualization, reference management, project management, overall editing, final submission
Mary Adewunmi: literature review, draft write-up (abstract, introduction, discussion, conclusion), supervision, funding acquisition
Mahmoud Abu Al Amrain: literature review, draft write-up (background)
Ming Chao Wong: supervision, and insights on the manuscript
Kwang Chien Yee: human evaluator, overall editing, and supervision.

Abbreviations
SD: standard deviation
ICC: Intraclass Correlation Coefficient
CI: confidence interval
SE: standard error
LLM: large language model
GPT: Generative Pretrained Transformer

NCCN: National Comprehensive Cancer Network
Q: question

References

[1] "Introducing ChatGPT." Accessed: Aug. 01, 2023. [Online]. Available: https://openai.com/blog/chatgpt

[2] T. Swanwick, "Understanding Medical Education," in *Understanding Medical Education*, John Wiley & Sons, Ltd, 2018, pp. 1–6. doi: 10.1002/9781119373780.ch1.

[3] T. H. Kung *et al.*, "Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models," *PLOS Digit Health*, vol. 2, no. 2, p. e0000198, Feb. 2023, doi: 10.1371/journal.pdig.0000198.

[4] R. Tsang, "Practical Applications of ChatGPT in Undergraduate Medical Education," *Journal of Medical Education and Curricular Development*, vol. 10, p. 23821205231178449, Jan. 2023, doi: 10.1177/23821205231178449.

[5] M. Hu, S. Pan, Y. Li, and X. Yang, "Advancing Medical Imaging with Language Models: A Journey from N-grams to ChatGPT," arXiv.org. Accessed: Aug. 01, 2023. [Online]. Available: https://arxiv.org/abs/2304.04920v1

[6] B. Anders, "Why ChatGPT is such a big deal for education," *C2C Digital Magazine*, vol. 1, no. 18, Jan. 2023, [Online]. Available: https://scholarspace.jccc.edu/c2c_online/vol1/iss18/4

[7] D. Uprety, D. Zhu, and H. J. West, "ChatGPT-A promising generative AI tool and its implications for cancer care," *Cancer*, vol. 129, no. 15, pp. 2284–2289, Aug. 2023, doi: 10.1002/cncr.34827.

[8] V. Sorin, Y. Barash, E. Konen, and E. Klang, "Large language models for oncological applications," *J Cancer Res Clin Oncol*, vol. 149, no. 11, pp. 9505–9508, Sep. 2023, doi: 10.1007/s00432-023-04824-w.

[9] S. Lukac *et al.*, "Evaluating ChatGPT as an adjunct for the multidisciplinary tumor board decision-making in primary breast cancer cases," *Arch Gynecol Obstet*, Jul. 2023, doi: 10.1007/s00404-023-07130-5.

[10] H. Haver, E. Ambinder, M. Bahl, E. Oluyemi, J. Jeudy, and P. Yi, "Appropriateness of Breast Cancer Prevention and Screening Recommendations Provided by ChatGPT," *Radiology*, vol. 307, p. 230424, Apr. 2023, doi: 10.1148/radiol.230424.

[11] Z. Moazzam, J. Cloyd, H. A. Lima, and T. M. Pawlik, "Quality of ChatGPT Responses to Questions Related to Pancreatic Cancer and its Surgical Care," *Ann Surg Oncol*, Jun. 2023, doi: 10.1245/s10434-023-13777-w.

[12] T.-C. Lee, K. Staller, V. Botoman, M. P. Pathipati, S. Varma, and B. Kuo, "ChatGPT Answers Common Patient Questions About Colonoscopy," *Gastroenterology*, vol. 165, no. 2, pp. 509-511.e7, Aug. 2023, doi: 10.1053/j.gastro.2023.04.033.

[13] Y. Gorelik, I. Ghersin, I. Maza, and A. Klein, "Harnessing Language Models for Streamlined Post-Colonoscopy Patient Management: A Novel Approach," *Gastrointest Endosc*, pp. S0016-5107(23)02675–5, Jun. 2023, doi: 10.1016/j.gie.2023.06.025.

[14] W. Li, Y. Zhang, and F. Chen, "ChatGPT in Colorectal Surgery: A Promising Tool or a Passing Fad?," *Ann Biomed Eng*, May 2023, doi: 10.1007/s10439-023-03232-y.

[15] B. Schulte, "Capacity of ChatGPT to Identify Guideline-Based Treatments for Advanced Solid Tumors," *Cureus*, vol. 15, no. 4, Apr. 2023, doi: 10.7759/cureus.37938.

[16] F. Schoonjans, "MedCalc statistical software - free trial available," MedCalc. Accessed: Jul. 28, 2023. [Online]. Available: https://www.medcalc.org/

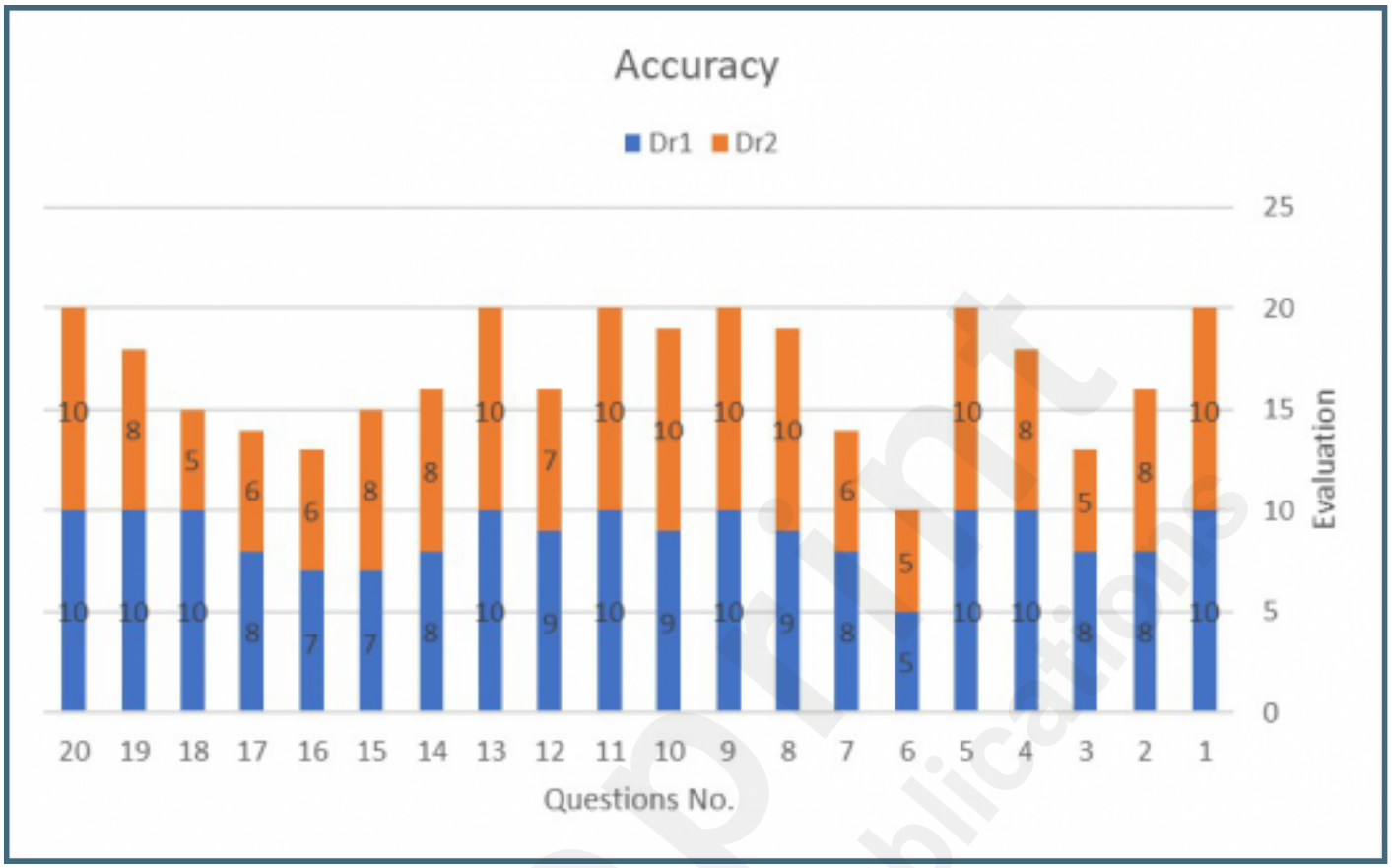
[17] A. Sarraju, D. Bruemmer, E. Van Iterson, L. Cho, F. Rodriguez, and L. Laffin, "Appropriateness of Cardiovascular Disease Prevention Recommendations Obtained From a Popular Online Chat-Based Artificial Intelligence Model," *JAMA*, vol. 329, no. 10, pp. 842–

844, Mar. 2023, doi: 10.1001/jama.2023.1044.
[18] H. L. Haver, C. T. Lin, A. Sirajuddin, P. H. Yi, and J. Jeudy, “Evaluating ChatGPT’s Accuracy in Lung Cancer Prevention and Screening Recommendations,” *Radiol Cardiothorac Imaging*, vol. 5, no. 4, p. e230115, Jul. 2023, doi: 10.1148/ryct.230115.

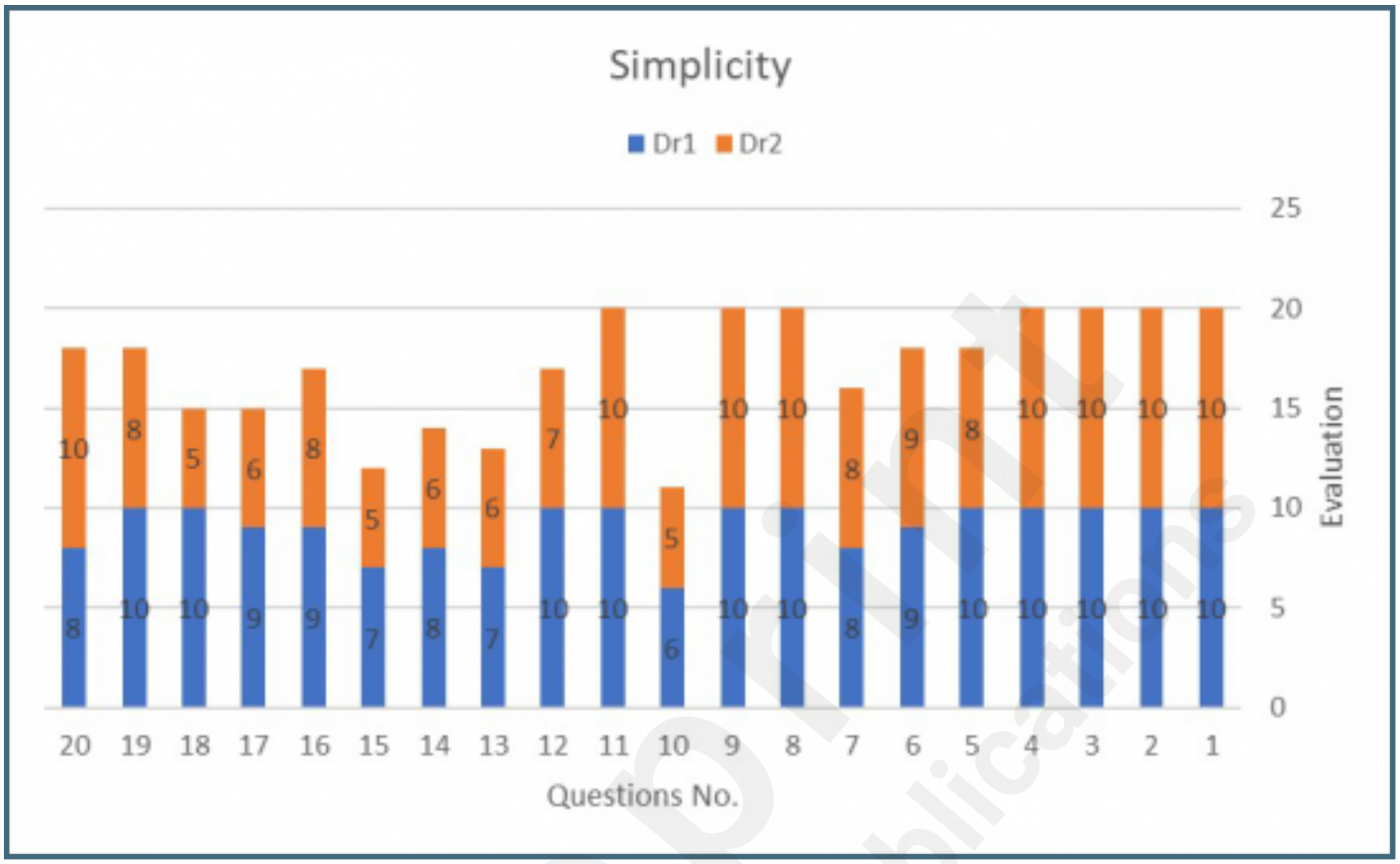
Supplementary Files

Figures

Accuracy results.



Simplicity results.



Multimedia Appendixes

Grading criteria.
URL: http://asset.jmir.pub/assets/19c8e4a443a4d37b92398b2e378fef31.pdf

Questions and Answers.
URL: http://asset.jmir.pub/assets/394d6d0d3c887fe0fe6d384d2fb67b10.pdf