Evaluation of an Artificial Intelligence Chatbot for Delivery of Interventional Radiology Patient Education Material

Colin J. McCarthy, MD, Interventional Radiologist, Beth Israel Deaconess Medical Center, Harvard Medical School; Seth Berkowitz, MD; Vijay Ramalingam, MD; Muneeb Ahmed, MD

Introduction

To assess the potential role for ChatGPT for the delivery of medical information to patients.

Hypothesis

To assess the accuracy, completeness, and readability of patient educational material produced by a machine-learning model and compare the output to that provided by a Societal patient education website.

Methods

Content from the Society of Interventional Radiology (SIR) Patient Center website was retrieved, categorized and organized into discrete questions. These questions were entered into the ChatGPT platform, and the output was analyzed for word and sentence count, readability using multiple validated scales, factual correctness and suitability for patient education using the PEMAT-P instrument.

Results

21,154 words were analyzed, including 7,917 words from the website and 13,377 words representing the total output of the ChatGPT platform across twenty-two text passages. Compared to the Societal website, output from the ChatGPT platform was longer and more difficult to read on 4 of 5 readability scales. The ChatGPT output was incorrect for 12 of 104 (11.5%) questions. When reviewed using the PEMAT-P tool, the ChatGPT content scored lower than the website material. Content from both the website and ChatGPT were significantly above the recommended 5th or 6th grade-level for patient education, with mean Flesch Kincaid Grade Level of 11.1 (+/- 1.3) for the website and 11.9 (+/- 1.6) for the ChatGPT content.

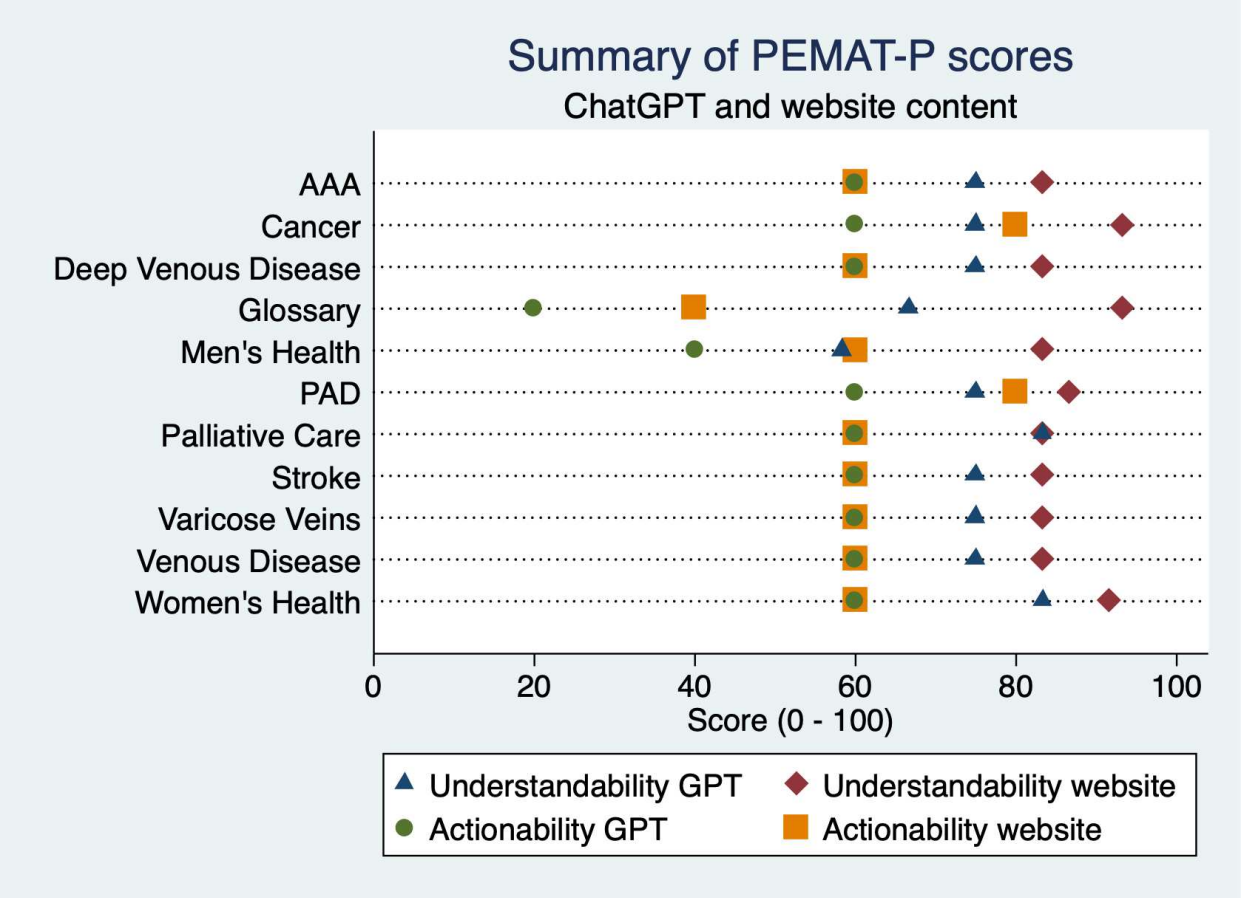
Conclusion

The ChatGPT platform may produce incomplete or inaccurate patient educational content, and providers should be familiar with the limitations of the system in its current form. Opportunities may exist to fine-tune existing large language models, which could be optimized for the delivery of patient educational content.

Statement of Impact

1. Evaluation of this large language Chatbot model as a patient education tool showed that incomplete or incorrect content was provided in 12 of 104 (11.5%) questions. 2. The absence of visual aids in the generated output results in material that is not fully optimized for patient educational purposes. 3. Compared to the Societal website, content from the ChatGPT platform was longer, contained more difficult words and longer sentences, and was more difficult

to read when assessed using several validated reading scales. Content from both the ChatGPT model and Societal website was written at a grade level higher than that recommended for patient education.



Performance of the ChatGPT model against the SIR website, using the PEMAT-P instrument.

Table 3 – overview of analyzed text passages.

	Word Count		Sentence Count		Long Sentence Count (i)		Reading Time (mm:ss)		Dale-Chall Difficult Words (ii)		Automated Readability Index (iii)		Passive Voice Count	
Topic	ChatGPT	SIR	ChatGPT	SIR	ChatGPT	SIR	ChatGPT	SIR	ChatGPT	SIR	ChatGPT	SIR	ChatGPT	SIR
Abdominal Aortic Aneurysm (AAA)	487	426	25	23	21	18	2:09	1:53	191	147	12.36	11.06	4	3
Cancer	873	465	39	28	32	19	3:52	2:04	339	185	13.94	11.01	12	4
Deep venous disease	979	856	52	47	42	32	4:21	3:48	356	309	10.61	10.03	9	9
Glossary of IR Terms	3801	1783	189	77	179	71	16:53	7:55	1340	680	11.31	13.77	97	28
Men's Health	988	738	46	38	39	26	4:23	3:16	421	308	13.93	12.88	15	5
Peripheral artery disease (PAD)	1275	871	68	43	52	29	5:40	3:52	438	288	10.65	12.35	17	11
Palliative care	1199	400	59	24	46	14	5:19	1:46	471	149	12.74	10.92	15	4
Stroke	1026	421	49	23	41	14	4:33	1:52	338	158	11.45	10.89	9	7
Varicose veins	845	469	53	31	42	16	3:45	2:05	356	193	10.61	10.04	21	5
Venous Disease	261	234	17	14	14	7	1:09	1:02	93	87	8.77	9.31	3	0
Women's Health	1603	1254	76	70	59	49	7:07	5:34	676	508	13.37	11.77	20	16
Grand Totals	13337	7917	673	418	567	295	59:11	35:07	5019	3012	129.74	124.03	222	92
Mean	1212.5	719.7	61.2	38	51.5	26.8	5:22	3:11	456.3	273.8	11.8	11.3	20.2	8.4

i) Long sentence count defined as sentences containing > 20 syllables. ii) Dale-Chall Difficult Words are those not found on a list of about 3,000 common words, which 80% or more of US 4th grade students are familiar with. iii) The calculation for the Automated Readability Index (ARI) is outlined in Table XXX, and assesses the approximate U.S. grade level required to read a piece of text.

Keywords

ChatGPT; Patient information; Chatbot; AI; Interventional Radiology