ChatGPT and Orthopaedic Surgery: Assessing the Readability and Accuracy

Oscar Yuan-Jie Shen, Jayanth Sairam Pratap, Xiang Li, Neal C Chen[1], Abhiram Bhashyam
[1]Massachusetts General Hospital

INTRODUCTION:

The advent of artificial intelligence (AI) has the potential to enhance health literacy significantly, and ChatGPT, developed by OpenAI, is a prime example of this. The NIH recommends health information be written at a 6-7th grade reading level, and most health information still fails to meet this recommendation. Adoption of large language models like ChatGPT by patients represents a potential shift in the accessibility of health information; unlike traditional search engines, models like ChatGPT have the capability to aggregate and summarize complex information from multiple sources, presenting them in a single response. This has an unclear effect on the readability of the response. This study investigates the current readability and accuracy of three topics in orthopaedic surgery, comparing responses from ChatGPT with Google Search, the most frequently used search engine by patients.

METHODS:

We used repeated identical queries of ChatGPT, along with the same queries with Google Search, to explore the relationship of ChatGPT's answers to the top 20 Google Search results.

The following three questions were investigated. These three queries were selected due to their varying level of consensus in the medical community:

1. What is the cause of carpal tunnel syndrome? (High consensus)
2. What is the cause of tennis elbow? (Medium consensus)
3. Platelet rich plasma (PRP) for thumb arthritis? (Low consensus)

Each question was entered verbatim into ChatGPT-3·5 20 times. This was also repeated for GPT-4 which is the latest iteration in the series of generative language models, and it boasts significant improvements over its predecessor, GPT-3·5.

The questions were also entered verbatim into Google without being logged in, and the top 20 search results were identified. Websites were assigned the classification of manuscript for academic papers, academic for academic institutions or organizations, government for websites sponsored/run by the government, or private for any other type of website. Inclusion and exclusion of text from each website was done with the end goal of imitating the structure of ChatGPT responses.

Seven readability algorithms were used to score each text: Flesch Reading Ease, Flesch-Kincaid Reading Level, SMOG (Simple Measure of Gobbledygook) Grade, Coleman-Liau Index, Gunning Fog Index, Automated Readability Index, and Linsear Write Formula. The overall grade level for Google Search websites and ChatGPT responses was calculated by averaging the seven scores.

Two attending orthopaedic surgeons also assessed the coverage and accuracy of information provided by Google and ChatGPT. Each query had three to four pertinent clinical domains that were assessed. Coverage could be considered more than minimally addressed, minimally addressed, or not addressed. Accuracy could be graded as completely correct, mostly correct, or mostly incorrect.

RESULTS:

The scores are given as a mean and standard deviation across each category for each question. The overall reading level for CTS was 9.05 ± 1.93 for Google, 14.40 ± 1.27 for GPT-3.5, and 15.00 ± 0.92 for GPT-4. The overall reading level for TE was 9.30 ± 2.41 for Google, 14.30 ± 1.56 for GPT-3.5, and 14.20 ± 1.24 for GPT-4. The overall reading level for PRP was 13.00 ± 2.18 for Google, 14.40 ± 1.27 for GPT-3.5, 15.25 ± 1.07 for GPT-4. There was a significant difference in overall reading level between Google and GPT-3.5 and GPT-4. The difference in overall reading level between GPT-3.5 and GPT-4 only reached significance for PRP.

Accuracy is currently being assessed.

DISCUSSION AND CONCLUSION:

Our study illustrates that the responses that ChatGPT generates are significantly more difficult to read than Google Search results. For CTS and TE, the average reading level of a Google Search result was around that of a freshman in high school. Meanwhile for ChatGPT, it consistently provided responses at the reading level of a college sophomore or junior. If we find that responses generated by ChatGPT are also less accurate than those found on Google, our findings represent a severe limitation in the adoption of ChatGPT by the public for medical information. ChatGPT could still be a useful tool if it had the capability of summarizing and simplifying currently available health information to a middle school level, but its current complexity and potential inaccuracy could be potentially harmful as patients could be misled and/or confused by the information presented. Patients should be advised about these findings when using ChatGPT, and future versions of pre-trained language models should have the ability to simplify and summarize information being presented, without sacrificing accuracy of information.

	CTS			TE			PRP		
	Google	GPT-3.5	GPT-4	Google	GPT-3.5	GPT-4	Google	GPT-3.5	GPT-4
Flesch Reading Ease Score	59.7 ± 10.1	39.5 ± 5.4	32.3 ± 4.4	54.2 ± 14.2	35.9 ± 6.9	33.7 ± 4.1	39.4 ± 11.5	38.4 ± 5.3	31.2 ± 4.7
Gunning-Fog Index	11.0 ± 2.3	17.0 ± 1.4	17.7 ± 0.9	11.3 ± 3.4	17.2 ± 1.7	17.0 ± 1.3	15.7 ± 2.4	18.3 ± 1.5	19.5 ± 1.1
Flesch Kincaid Grade	8.6 ± 2.0	14.1 ± 1.3	14.5 ± 0.8	9.4 ± 2.6	19.7 ± 23.0	14.4 ± 1.1	12.8 ± 2.3	13.6 ± 1.3	14.6 ± 1.0
Coleman-Liau Index	11.05 ± 1.6	12.7 ± 0.9	14.5 ± 1.0	10.8 ± 2.0	11.4 ± 1.2	12.9 ± 1.0	12.6 ± 2.0	13.3 ± 1.0	14.5 ± 0.9
SMOG score	8.0 ± 1.6	12.5 ± 1.0	13.1 ± 0.6	8.3 ± 2.4	12.2 ± 1.2	12.1 ± 1.1	11.8 ± 1.8	13.4 ± 1.0	14.1 ± 0.9
Automated Readability Index	9.1 ± 2.3	15.6 ± 1.5	16.0 ± 1.0	8.9 ± 2.6	14.9 ± 1.9	15.1 ± 1.4	13.0 ± 2.7	14.8 ± 1.5	15.7 ± 1.2
Linsear Write Formula	8.6 ± 2.6	17.3 ± 1.9	16.7 ± 1.1	8.8 ± 3.3	17.4 ± 2.0	15.8 ± 3.2	14.4 ± 2.8	17.0 ± 2.2	17.4 ± 1.9
Overall	9.1 ± 1.9	14.4 ± 1.3	15.0 ± 0.9	9.3 ± 2.4	14.3 ± 1.6	14.2 ± 1.2	13.0 ± 2.2	14.4 ± 1.3	15.3 ± 1.1

Table 1. The mean (± standard deviation) score or grade for each readability measure for texts based on source.