Can Patients Differentiate Responses by Experts and ChatGPT Regarding Total Hip Arthroplasty?

Perry Lee Lim[1], Anoop Prasad, Amy Zhang Blackburn, David Ensor, Marc Succi, Kyle Alpaugh, Christopher Michael Melnic, Karen Sepucha[1], Hany S Bedair[1]

[1]Massachusetts General Hospital

INTRODUCTION:

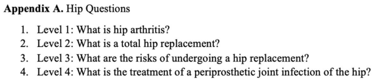
The digital era's abundance of health information poses great opportunities for health literacy gains by patients, but also represents challenges such as misinformation. Artificial intelligence (AI)-driven chatbots such as Chat Generative Pre-trained Transformer 3 (ChatGPT3) from OpenAI, blur the human-AI content creation boundary, influencing healthcare literacy. This study aimed to assess perceived message credibility by patients when comparing healthcare descriptions by ChatGPT3 with orthopaedic attendings. This was assessed across four levels of increasing complexity. We hypothesized that patients would be able to differentiate ChatGPT3 response from an expert clinician as the complexity of topic described increased.

METHODS: We conducted a prospective survey of 160 respondents from October 2023 to March 2024. The study assessed hip arthroplasty topics across four difficulty levels: medical student, intern, resident, and fellow/attending **(see Appendix A)**. Demographics and news preferences were collected. Each patient was randomly assigned one of the four questions and asked to compare four answers: one from ChatGPT and three from expert orthopedic clinicians. Patients rated each answer (1-7) on credibility (accuracy, authenticity, believability) and indicated their most trusted answer. Multilevel modeling, incorporating varying intercepts for patient-level observations, surgeon interactions, and credibility domains, was used to provide estimated scores for each surgeon and ChatGPT.

RESULTS:

The credibility scores by question are displayed in **Table 1.** For Question #1, ChatGPT and surgeons performed similarly in accuracy and authenticity, while ChatGPT showed better believability than Surgeon B (6.0 vs. 5.4, P = 0.035) and Surgeon C (6.0 vs. 5.1, P = 0.003), with scores comparable to Surgeon A. Question #2 demonstrated significant differences across all three credibility domains, with ChatGPT outperforming Surgeons A and C in accuracy (5.9 vs. 5.3, P = 0.024; 5.9 vs. 4.2, P < 0.001) and authenticity (5.8 vs. 4.9, P = 0.006; 5.8 vs. 3.5, P < 0.001) but showing similar believability scores. In Question #3, ChatGPT scored higher than Surgeon C in accuracy (6.0 vs. 4.5, P < 0.001) and in authenticity compared to Surgeon B (5.9 vs. 5.0, P = 0.007) and Surgeon C (5.9 vs. 4.3, P < 0.001). In believability, ChatGPT outperformed all three surgeons (Surgeon A: 6.2 vs. 5.5, P = 0.016; Surgeon B: 6.2 vs. 5.0, P < 0.001; Surgeon C: 6.2 vs. 4.5, P < 0.001). Regarding Question #4, ChatGPT's performance was similar across all credibility domains compared to the three surgeons. Multilevel modeling revealed ChatGPT scored the highest across all three credibility domains **(Figure 1)**. Notably, ChatGPT's answer was the was the most trusted response by patients in two out of the four questions (Question #1 and #3). Sub analysis stratifying credibility scores by sex, education level, and news input revealed no difference.

DISCUSSION AND CONCLUSION: The study suggests that patients perceive ChatGPT as highly credible, particularly in terms of accuracy, authenticity, and believability compared to expert orthopedic clinicians. These findings underscore the potential of AI-driven chatbots of improving healthcare literacy and patient decision-making. However, further research is warranted to explore the nuances of patient trust and preferences in AI-generated healthcare content.



Appendix A. Hip Questions

1. Level 1: What is hip arthritis?
2. Level 2: What is a total hip replacement?
3. Level 3: What are the risks of undergoing a hip replacement?
4. Level 4: What is the treatment of a periprosthetic joint infection of the hip?

Figure 1. Multilevel Modeling Scores

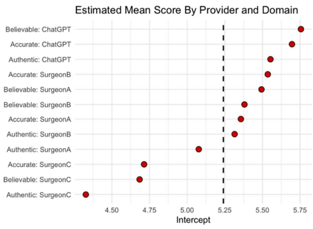
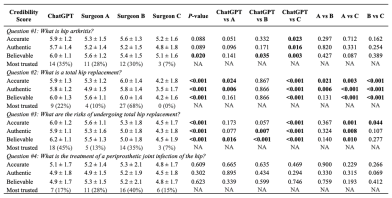


Table 1. Credibility Scores by Question



Credibility Score	ChatGPT	Surgeon A	Surgeon B	Surgeon C	*P*-value	ChatGPT vs A	ChatGPT vs B	ChatGPT vs C	A vs B	A vs C	B vs C
Question #1: What is hip arthritis?											
Accurate	5.9 ± 1.2	5.3 ± 1.5	5.6 ± 1.3	5.2 ± 1.6	0.088	0.051	0.332	**0.023**	0.297	0.712	0.162
Authentic	5.7 ± 1.4	5.2 ± 1.4	5.2 ± 1.5	4.8 ± 1.8	0.089	0.096	0.171	**0.016**	0.820	0.331	0.254
Believable	6.0 ± 1.1	5.6 ± 1.2	5.4 ± 1.5	5.1 ± 1.6	**0.020**	0.141	**0.035**	**0.003**	0.427	0.087	0.389
Most trusted	14 (35%)	11 (28%)	12 (30%)	3 (7%)	NA	NA	NA	NA	NA	NA	NA
Question #2: What is a total hip replacement?											
Accurate	5.9 ± 1.3	5.3 ± 1.2	6.0 ± 1.4	4.2 ± 1.8	**<0.001**	**0.024**	0.867	**<0.001**	**0.021**	**0.003**	**<0.001**
Authentic	5.8 ± 1.2	4.9 ± 1.5	5.8 ± 1.4	3.5 ± 1.7	**<0.001**	**0.006**	0.866	**<0.001**	**0.006**	**<0.001**	**<0.001**
Believable	6.0 ± 1.3	5.6 ± 1.1	6.0 ± 1.4	4.2 ± 1.6	**<0.001**	0.161	0.866	**<0.001**	0.131	**<0.001**	**<0.001**
Most trusted	9 (22%)	4 (10%)	27 (68%)	0 (0%)	NA	NA	NA	NA	NA	NA	NA
Question #3: What are the risks of undergoing total hip replacement?											
Accurate	6.0 ± 1.2	5.6 ± 1.1	5.3 ± 1.8	4.5 ± 1.7	**<0.001**	0.173	0.057	**<0.001**	0.367	**0.001**	**0.044**
Authentic	5.9 ± 1.1	5.3 ± 1.6	5.0 ± 1.8	4.3 ± 1.8	**<0.001**	0.077	**0.007**	**<0.001**	0.324	**0.008**	0.107
Believable	6.2 ± 1.1	5.5 ± 1.3	5.0 ± 1.8	4.5 ± 1.9	**<0.001**	**0.016**	**<0.001**	**<0.001**	0.140	**0.010**	0.277
Most trusted	18 (45%)	5 (13%)	14 (35%)	3 (7%)	NA	NA	NA	NA	NA	NA	NA
Question #4: What is the treatment of a periprosthetic joint infection of the hip?											
Accurate	5.1 ± 1.7	5.2 ± 1.4	5.3 ± 2.1	4.8 ± 1.7	0.609	0.665	0.635	0.469	0.900	0.229	0.266
Authentic	4.9 ± 1.8	4.9 ± 1.5	5.2 ± 1.9	4.5 ± 1.8	0.302	0.895	0.434	0.294	0.330	0.315	0.069
Believable	4.9 ± 1.7	5.3 ± 1.5	5.2 ± 2.1	4.8 ± 1.7	0.623	0.339	0.599	0.746	0.759	0.193	0.412
Most trusted	7 (17%)	11 (28%)	16 (40%)	6 (15%)	NA	NA	NA	NA	NA	NA	NA