Chatbots in limb lengthening and reconstruction surgery. How accurate are the responses?

Christopher August Iobst[1], Anirejuoritse Bafor, Søren Kold[2], Kirsten Tulchin-Francis[1], Daryn Strub[3]
[1]Nationwide Children's Hospital, [2]Aalborg University Hospital, [3]Nationwide Childrens Hospital

INTRODUCTION:

In the last decade, internet search engines and online platforms have been a resource for patients, providing answers to questions relating to healthcare. In pediatric orthopedics, studies have shown that a significant percentage of parents use online search engines to find out more about the health condition of their children. The recent introduction of Chatbots has provided an interactive medium to answer patient questions. The accuracy of responses with these programs in limb lengthening and reconstruction surgery has not previously been determined. Therefore, the purpose of this study was to assess the accuracy of answers from 3 free AI chatbot platforms to 23 common questions regarding treatment for limb lengthening and reconstruction.

METHODS:

We generated a list of 23 common questions asked by parents before their child's limb lengthening and reconstruction surgery. Each question was posed to three different AI chatbots (ChatGPT 3.5 [OpenAI], Google Bard, and Microsoft Copilot [Bing!]) by three different answer retrievers on separate computers between November 17 and November 18, 2023. Responses were only asked one time to each chatbot by each answer retriever. Nine answers (3 answer retrievers x 3 chatbots) were randomized and platform-blinded prior to rating by three orthopedic surgeons. The 4-point rating system reported by Mika et al. was used to grade all responses.

RESULTS:

ChatGPT had the best response accuracy score (RAS) with a mean score of 1.73 ± 0.88 across all three raters (range of means for all three raters – 1.62 – 1.81) and a median score of 2. The mean response accuracy scores for Google Bard and Microsoft Copilot were 2.32 ± 0.97 and 3.14 ± 0.82, respectively. This ranged from 2.10 – 2.48 and 2.86 – 3.54 for Google Bard and Microsoft Copilot, respectively. The differences between the mean RAS scores were statistically significant ($p < 0.0001$). The median scores for Google Bard and Microsoft Copilot were 2 and 3, respectively.

DISCUSSION AND CONCLUSION:

Using the Response Accuracy Score, the responses from ChatGPT were determined to be satisfactory, requiring minimal clarification, while the responses from Microsoft Copilot were either satisfactory, requiring moderate clarification, or unsatisfactory, requiring substantial clarification. Although ChatGPT had satisfactory answers, the other AI chatbot platforms did not perform well. Therefore, if patients are using these alternative platforms for information, the surgeon may need to spend extra time with the patient to undo any misinformation they have been provided as well as ensure that a proper and accurate understanding of the limb lengthening journey is conveyed.