Embargoed for release until 12:01 a.m. PST Feb. 12, 2024

For more information, contact:

Deanna Killackey	847-384-4035	630-815-5195	killackey@aaos.org
Lauren P. Riley	847-384-4031	708-227-1773	pearson@aaos.org

Studies Show AI Chatbots Provide Inconsistent Accuracy for Musculoskeletal Health Information

- **Researchers agree: Orthopaedic surgeons remain the most reliable source of information**
- **All chatbots displayed significant limitations, omitting critical steps in workup**
- **Researchers summarize: ChatGPT is still not an adequate resource to answer patient questions; further work is needed to develop an accurate orthopaedic-focused chatbot is needed**

SAN FRANCISCO (Feb. 12, 2024)—With the growing popularity of large language model (LLM) chatbots, a type of artificial intelligence (AI) used by ChatGPT, Google Bard and BingAI, it is important to outline the accuracy of musculoskeletal health information they provide. Three new studies presented at the 2024 Annual Meeting of the American Academy of Orthopaedic Surgeons (AAOS) analyzed the validity of the information chatbots gave to patients for certain orthopaedic procedures, assessing the accuracy of how chatbots present research advancements and clinical decision making.

While the studies found that certain chatbots provide concise summaries across a wide spectrum of orthopaedic conditions, each demonstrated limited accuracy depending on the category. Researchers agree that orthopaedic surgeons remain the most reliable source of information. The findings will help those in the field understand the efficacy of these AI tools, if the use by patients or non-specialist colleagues could introduce bias or misconceptions and how future enhancements can make chatbots a potentially valuable tool for patients and physicians in the future.

STUDY OVERVIEWS AND OUTCOMES

Potential misinformation and dangers associated with clinical use of LLM chatbots

This study, led by Branden Sosa, a fourth-year medical student at Weill Cornell Medicine, assessed the accuracy of Open AI ChatGPT 4.0, Google Bard and BingAI chatbots to explain basic orthopaedic concepts, integrate clinical information and address patient queries. Each chatbot was prompted to answer 45 orthopaedic-related questions spanning categories of “Bone Physiology,” “Referring Physician,” and “Patient Query” and then assessed for accuracy. Two independent, blinded reviewers scored responses on a scale of 0-4, assessing accuracy, completeness and useability. Responses were analyzed for strengths and limitations within categories and across chatbots. The research team found the following trends:

- When prompted with orthopedic questions, OpenAI ChatGPT, Google Bard and BingAI provided correct answers that covered the most critical salient points in 76.7%, 33% and 16.7% of queries, respectively.
- When providing clinical management suggestions, all chatbots displayed significant limitations by deviating from the standard of care and omitting critical steps in workup such as ordering antibiotics before cultures or neglecting to include key studies in diagnostic workup.
- When asked less complex patient queries, ChatGPT and Google Bard were able to provide mostly accurate responses but often failed to elicit critical medical history pertinent to fully address the query.
- A careful analysis of citations provided by chatbots revealed an oversampling of a small number of references and 10 faulty links that were nonfunctional or led to incorrect articles.

Is ChatGPT ready for prime time? Assessing the accuracy of AI in answering common arthroplasty patient questions

Researchers, led by Jenna A. Bernstein, MD, orthopaedic surgeon at Connecticut Orthopaedics, sought to investigate how accurately ChatGPT 4.0 answered patient questions by developing a list of 80 commonly asked patient questions about knee and hip replacements. Each question was queried two times in ChatGPT; first asking the questions as written, and then prompting the ChatGPT to answer the patient questions "as an orthopaedic surgeon." Each surgeon on the team evaluated the accuracy of each set of answers and rated these on a scale of one to four. Agreement was assessed between the two surgeons' evaluation of each set of ChatGPT answers. The association between the question prompt and response accuracy were both assessed using two means of statistical analysis (Cohen's kappa and Wilcoxon signed-rank test, respectively). The findings included:

- When assessing the quality of the ChatGPT responses, 26% (21 of 80 responses) had an average scale of three (partially accurate, but incomplete) or less when asked without a prompt, and 8% (six of 80 responses) had an average grade of less than three when preceded by a prompt. As such, researchers summarized that ChatGPT is still not an adequate resource to answer patient questions and further work to develop an accurate orthopaedic-focused chatbot is needed.
- ChatGPT performed substantially better when appropriately prompted to answer patient questions "as an orthopaedic surgeon" with 92% accuracy.

Can ChatGPT 4.0 be used to answer patient questions concerning the Latarjet procedure for anterior shoulder instability?

Researchers at the Hospital for Special Surgery in New York, led by Kyle Kunze, MD, assessed the propensity for ChatGPT 4.0 to provide medical information about the Latarjet procedure for patients with anterior shoulder instability. The overall goal of this study was to understand whether this chatbot could demonstrate potential to serve as a clinical adjunct and help both patients and providers through providing accurate medical information.

To answer this question, the team first conducted a Google search using the query "Latarjet" to extract the top ten frequently asked questions (FAQs) and associated sources concerning the procedure. They then asked ChatGPT to perform the same search for FAQs to identify the questions and sources provided by the chatbot. Highlights of the findings included:

- ChatGPT demonstrated the ability to provide a broad range of clinically relevant questions and answers and derived information from academic sources 100% of the time. This is in opposition to Google, which included a small percentage of academic resources, combined with information found on surgeons' personal websites and larger medical practices.
- The most common question category for both ChatGPT and Google was technical details (40%); however, ChatGPT also presented information concerning risks/complications (30%), recovery timeline (20%) and evaluation of surgery (10%).

# # #

2024 AAOS Annual Meeting Disclosure Statement

About the AAOS

With more than 39,000 members, the American Academy of Orthopaedic Surgeons is the world's largest medical association of musculoskeletal specialists. The AAOS is the trusted leader in advancing musculoskeletal health. It provides the highest quality, most comprehensive education to help orthopaedic surgeons and allied health professionals at every career level to best treat patients in their daily practices. The AAOS is the source for information on bone and joint conditions, treatments and related musculoskeletal healthcare issues; and it leads the healthcare discussion on advancing quality.

Follow the AAOS on Facebook, X, LinkedIn and Instagram.