Navigating Neuroclinical Information: A Comparative Analysis of LLM and Google Search Quality for Patient Queries on BMP-2

Jonathan P. Japa, BS[1]; Shankar S Thiru, BS[2]; Nicholas E Aksu, MD[3]; Jamie Lee, BS[4]; Mark Ehioghae, MS[1]; Kevin Yoon, BS[1]; Aaron Phung BS[4]; Justin Hyde, BS[4]; Sean Bae, BS[4]; Ala Alshomali, BS[4]; Addisu Mesfin, MD[1]

Jonathan P. Japa: jonapaul1@gmail.com

1. MedStar Orthopaedic Institute, Washington Hospital Center
2. Department of Orthopaedics, Georgetown University School of Medicine
3. Department of Orthopaedics, MedStar Georgetown University Hospital
4. Georgetown University School of Medicine

Abstract

Introduction: As the use of advanced biologics like BMP-2 becomes more widespread, patients increasingly seek to understand such treatments independently. Over the past two decades, there has been a significant increase in the number of individuals seeking medical information online[1]. More recently, generative artificial intelligence (AI) tools – specifically Large Language Models (LLMs) such as ChatGPT, Google Gemini, and Meta's LLaMA – have emerged as alternative sources for patient education[2]. These tools are capable of producing natural-sounding, personalized responses to user queries, potentially improving accessibility. However, LLMs are known to 'hallucinate,' fabricating information with high confidence, which poses a unique risk in the medical domain, particularly in high-stakes fields like spine surgery. The purpose of this study was to systematically evaluate the quality, reliability, and factual accuracy of patient-facing information regarding Bone Morphogenetic Protein-2 (BMP-2) in spinal surgery, as provided by leading Large Language Models (LLMs) and Google Searches.

Methods: This cross-sectional content analysis compared responses to 20 common patient queries about BMP-2 from three LLMs (ChatGPT (GPT-4), Gemini (Gemini 1.5 Pro), LLaMA-3) and Google Search. Responses were assessed using the DISCERN instrument, the Journal of the American Medical Association (JAMA) benchmark criteria, a 5-point clinical accuracy scale, and a detailed methodology for detecting hallucinations. Two independent reviewers with clinical backgrounds conducted evaluations, and inter-rater reliability was assessed using ICCs.

Results: Significant differences were observed across platforms. Google Search yielded the highest mean DISCERN (51.74 ± 11.06) and JAMA (3.51 ± 1.07) scores. Among LLMs, Gemini demonstrated the highest mean accuracy (4.28) and the lowest overall hallucination rate (0.27% of factual claims). ChatGPT showed moderate performance in most domains, while LLaMA-3 consistently underperformed in quality, accuracy, and hallucination frequency. Notably, while individual factual claims were hallucinated, no LLM-generated response contained greater than one hallucination. The most common hallucination type was overgeneralization/exaggeration.

Discussion: LLMs and Google Search exhibit significant variability in the quality, accuracy, and transparency of patient-facing information on BMP-2. While Google provided the most transparent content, Gemini offered the most accurate LLM-generated responses with the lowest hallucination rate. This study has limitations. Primarily, while the modest (n=20), the number of queries may not encompass the full spectrum of BMP-2-related questions posed by patients.

Significance/Clinical Relevance: The persistence of inaccuracies and hallucinations across platforms raises concerns about their suitability for unsupervised patient education in complex neurosurgical contexts. Developers must prioritize transparency and accuracy to ensure these tools are reliable for clinical use.

Table 1. Accuracy Scores by Platform

Platform	Mean Accuracy ± SD	% Accuracy Statements
ChatGPT	3.93 ± 1.25	90.00%
Gemini	4.28 ± 0.96	95.00%
LLaMA	2.83 ± 1.11	62.50%
p-value (ANOVA)	**<0.001***	

* = statistical significance