Accuracy of a Large Language Model (ChatGPT) in Responding to Patient-Generated Queries Following Benign Prostatic Hyperplasia Surgeries

Authors: Jad Najdi,[1] Bilal Alameddine,[1] Alexandre Armache,[1] Marwan Zein,[1] William S Azar,[2] Towfik Sebai,[1] Yara Ghandour,[1] *Albert El-Hajj[1]

1. American University of Beirut Medical Center, Lebanon
2. Georgetown University School of Medicine, Washington, District of Columbia, USA

*Correspondence to ae67@aub.edu.lb

Disclosure: The authors have declared no conflicts of interest.

Keywords: AI, minimally invasive surgical procedures, natural language processing, postoperative care, prostatic hyperplasia.

Citation: EMJ Urol. 2025;13[1]:48-49. https://doi.org/10.33590/emjurol/NXXB4817

INTRODUCTION

The rapid advancements in AI, especially in large language models like ChatGPT (OpenAI, San Francisco, California, USA), hold potential for various applications in healthcare.[1-6] This study aims to assess the accuracy of ChatGPT in responding to post-operative patient enquiries after surgery for benign prostatic hyperplasia.

METHODS

Common post-operative questions were collected from discharge instruction booklets, online forums, and social media platforms. Surgeries of interest included transurethral resection of the prostate (TURP), simple prostatectomy, laser enucleation of the prostate, Aquablation, Rezum, greenlight photovaporisation of the prostate, Urolift, and iTIND. ChatGPT 3.5 outputs were graded by two independent senior urology residents using pre-defined evaluation criteria. A third senior reviewer resolved grading discrepancies. Response errors were categorised into different types. Categorical variables were analysed using the Chi-square test. Inter-rater agreement was measured using Cohen's Kappa coefficient.

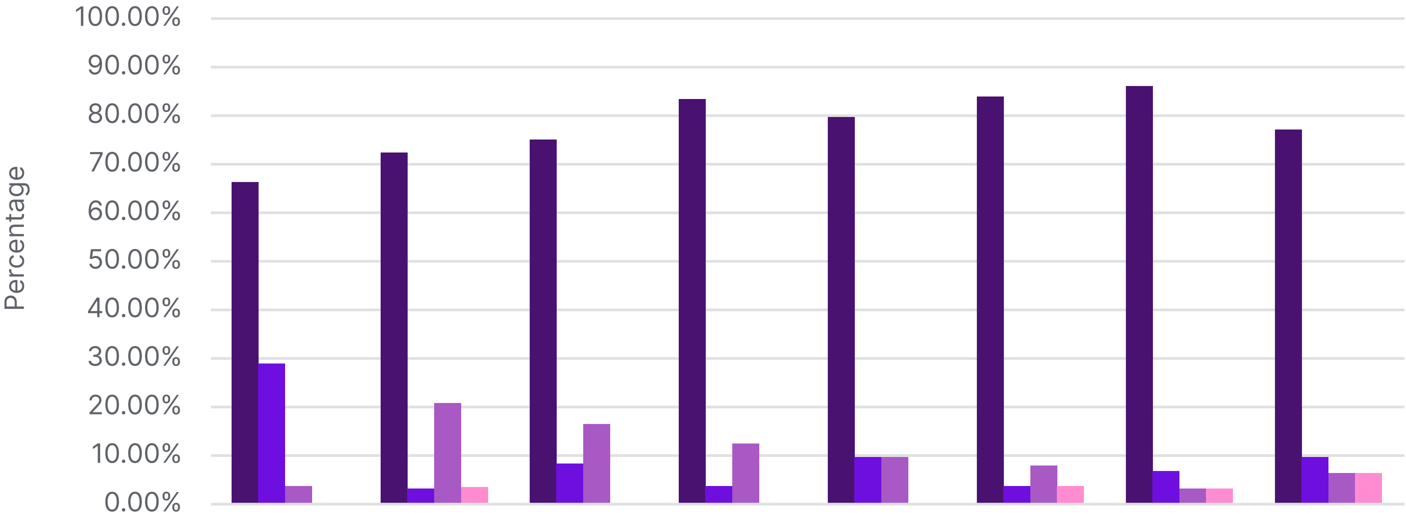
RESULTS

A total of 496 questions were evaluated by two reviewers, of which 280 were excluded. Of the 216 graded responses, 78.2% were comprehensive and correct, 9.3% were incomplete or partially correct, 10.2% were misleading or contained a mix of accurate and inaccurate information, and 2.3% were completely inaccurate (Figure 1). The highest percentage of correct answers was observed with newer procedures (Aquablation, Rezum, iTIND) as compared to older procedures (TURP, simple prostatectomy). Lack of context or incorrect information (36.6%) were the most common errors encountered.

CONCLUSION

ChatGPT demonstrates promise in providing accurate post-operative guidance for patients undergoing benign prostatic hyperplasia surgeries. However, incomplete or misleading responses raise concerns about its current clinical applicability, emphasising the need for further research to enhance its accuracy and ensure patient safety.

Figure 1: Percentage of answers in the four different grading categories divided by procedure type.



	TURP	Simple P	LEP	AQUA	Rezum	G-PVP	Urolift	iTIND
■ Comprehensive/ correct	66.70%	72.40%	75.00%	83.30%	80.00%	84.00%	86.10%	77.30%
■ Incomplete/ partially correct	29.20%	3.40%	8.30%	4.20%	10.00%	4.00%	6.90%	9.70%
■ Mixed with accurate and inaccurate data/ misleading	4.10%	20.70%	16.70%	12.50%	10.00%	8.00%	3.50%	6.50%
■ Completely inaccurate/ irrelevant	0%	3.5%	0%	0%	0%	4%	3.5%	6.5%

AQUA: aquablation; G-PVP: greenlight photovaporisation of the prostate; LEP: laser enucleation of the prostate; Simple P: simple prostatectomy; TURP: transurethral resection of the prostate.

References

1. Najdi et al. Accuracy of a large language model (ChatGPT) in responding to patient-generated queries following BPH surgery. Abstract A0901. EAU25, 21-24 March, 2025.
2. Wu J et al. The application of ChatGPT in medicine: a scoping review and bibliometric analysis. J Multidiscip Healthc. 2024;17:1681-92.
3. Wright BM et al. Is ChatGPT a trusted source of information for total hip and knee arthroplasty patients?. Bone Jt Open. 2024;5(2):139-46.
4. Kuşcu O et al. Is ChatGPT accurate and reliable in answering questions regarding head and neck cancer?. Front Oncol. 2023;13.
5. Harada Y et al. Performance evaluation of ChatGPT in detecting diagnostic errors and their contributing factors: an analysis of 545 case reports of diagnostic errors. BMJ Open Qual. 2024;13(2):e002654.
6. Briganti G. How ChatGPT works: a mini review. Eur Arch Otorhinolaryngol. 2024;281(3):1565-9.