Evaluation of ChatGPT's performance on answering pediatric urology questions based on association guidelines

Wyatt MacNevin[1], Nicholas Dawe[2], Laura Harkness[2], Budoor Salman[1], Daniel T. Keefe[1,3]

[1]Department of Urology, Dalhousie University, Halifax, NS, Canada; [2]Faculty of Medicine, Dalhousie University, Halifax, NS, Canada; [3]Department of Pediatric Urology, IWK Health Centre, Dalhousie University, Halifax, NS, Canada

Cite as: MacNevin W, Dawe N, Harkness L, et al. Evaluation of ChatGPT's performance on answering pediatric urology questions based on association guidelines. *Can Urol Assoc J* 2025;19(11):E362-7. http://dx.doi.org/10.5489/cuaj.9238

Published online July 28, 2025

Appendix available at *cuaj.ca*

ABSTRACT

INTRODUCTION: ChatGPT has been shown to provide accurate and complete responses to clinically focused questions, although its ability to successfully answer common pediatric urology-based questions remains unexplored. Furthermore, the concordance of ChatGPT's answers with association recommendations has yet to be analyzed.

METHODS: A list of common pediatric urology questions of varying difficulty was developed in association with publicly available guidelines and resources from the Canadian Urological Association (CUA), American Urological Association (AUA), and the European Association of Urology (EAU). Questions were administered individually using three separate functions, and responses were evaluated for comprehensiveness and accuracy using a Likert scale. Descriptive statistics and analysis of variance were used for statistical analysis.

RESULTS: ChatGPT performed best in the domain of phimosis (mean ± standard deviation: 2.32/3.00±0.57) and VUR (2.11/3.00±0.63), and worst in acute scrotal pathology (1.90/3.00±0.58) and cryptorchidism (1.92/3.00±0.56) (p=0.031). "Easy" questions (2.31/3.00±0.09) had greater comprehensiveness scores compared to "medium" (1.92/3.00±0.07, p=0.003) and "difficult" questions (1.86/3.00±0.101, p=0.003). Definition-based questions had greater comprehensiveness scores across all guidelines. ChatGPT was more accurate and in concordance with EAU-based information (2.10±0.41) compared to AUA (1.95±0.41, p=0.04).

CONCLUSIONS: ChatGPT answered questions with high levels of appropriateness and comprehensiveness. ChatGPT performed best in the areas of phimosis and VUR and worst in acute scrotal pathology. While ChatGPT performed well across all question domains, it performed best when referenced to EAU and CUA compared to AUA.

INTRODUCTION

In the field of health education, artificial intelligence (AI) has made a tremendous impact on improving medical knowledge accessibility for the general public.[1-3] Most notably, ChatGPT, a publicly available natural language processing model (NLP), has increased in popularity and now serves as a potential source of medical knowledge for individuals seeking guidance.[1,4,5] ChatGPT operates through the use of neural network-based deep learning to predict and generate text responses based on sequences and patterns identified in large collections of text data.[6] This basis has led to ChatGPT demonstrating accuracy in providing information in various medical and surgical fields, with recent applications in the field of urology.[5,7,8]

Early applications of ChatGPT have investigated the accuracy in answering questions related to urolithiasis, pediatric urology, and general urological concerns.[5,9,10] Furthermore, ChatGPT has been used to answer residency-based urology exams with signs of early success.[7,8] The field of pediatric urology has shown considerable interest in the use of AI and ChatGPT, with models being developed for diagnosis, predictive modelling, and patient information-gathering.[9,11] Despite this, there exists limited data on the use of ChatGPT in answering pediatric urology-based questions, and there is a paucity of data on how the performance of ChatGPT relates or differs based on which organization's guidelines and/or medical information resources are being referenced.

KEY MESSAGES

- ChatGPT answered pediatric urology questions with high levels of comprehensiveness.
- Phimosis and vesicoureteral reflux questions were answered most accurately by ChatGPT.
- ChatGPT performed most in concordance with Canadian and European urology guidelines and recommendations.

This study aimed to investigate the accuracy and reliability of ChatGPT in answering common pediatric urology questions. Secondly, this study examined the concordance of ChatGPT-generated answers with each urological association's statements (Canadian Urological Association [CUA], American Urological Association [AUA], and European Association of Urology [EAU]).

METHODS

Study design

A list of pediatric urology questions was developed based on review of the CUA, AUA, and EAU websites, educational resources, and patient information materials.[12-14] After cross-referencing between resources, a list of question areas was developed; these included phimosis, cryptorchidism, acute scrotal pathology, hypospadias, vesicoureteral reflux (VUR), and urolithiasis (Table 1). From these question areas, the authors (WM, DK) developed questions based on the topic's definition and basic clinical information that a patient might ask. Questions with subjective or ambiguous answers were excluded.

Question difficulty was determined as either easy, medium, or difficult based on author consensus (WM, BS), with disagreements settled by a third author (DK). "Easy" questions were defined as questions requiring minimal urologic expertise and those in which correct information could be readily found through online searches. "Medium" questions were defined as those requiring some urologic expertise and that could not be easily answered through online searches. "Difficult" questions were defined as those requiring urologic expertise with answers not readily available on the internet.

Questions were formatted into layperson terms to emulate how a patient would interact with a healthcare provider and/or an internet search. All questions were then administered individually into ChatGPT version 4 using unique chat functions, and responses were recorded.[15] To account for variability in responses, each question was input into ChatGPT three separate times in a new chat function.[16] Questions were input into ChatGPT sequentially and then recorded (July 5, 2024). Responses were then assessed by three urology residents/fellowship-trained pediatric urology attendings for comprehensiveness and accuracy.

A four-point Likert scale was used for evaluation (0=completely incorrect, 1=some correct and some incorrect, 2=correct but inadequate, and 3=comprehensive). "Appropriate" answers were defined as answers with comprehensiveness scores ≥2. Responses were assessed against reference answers derived from CUA, AUA, and EAU resources (Supplementary Table 1; available at *cuaj.ca*).

Research ethics approval exemption was granted by an institutional research ethics board, as patient data was not used in this study.

Statistical analysis

All data were compiled and imported into Statistical Package for Social Sciences (SPSS) version 29, and descriptive statistics were performed and expressed using frequencies and percentages.[17] Analysis of variance (ANOVA) was performed to compare the differences in means of each response based on difficulty, association reference used, and question topic. Furthermore, two-way ANOVA was used to differentiate the means between definition-based questions and general clinical knowledge-based questions. Levene's test of equality of error variances was used to assess the homogeneity of variance. Statistical significance was set at p=0.05 with a 95% confidence interval. Normality of data was assessed through interpretation of skewness and kurtosis. Inter-rater agreement was analyzed using Kappa statistics.

RESULTS

A total of 27 questions were developed, resulting in 81 unique responses from ChatGPT, which were assessed against CUA, AUA, and EAU resources by three authors (WM, ND, LH). Inter-rater agreement ranged from slight to fair agreement (rater 1 and rater 2=0.075, p=0.019; rater 1 and rater 3=0.116, p=0.005; rater 2 and rater 3=0.273, p=0.001). Overall, ChatGPT performed best across all guidelines in the domains of phimosis (mean ± standard deviation: 2.32/3.00±0.57) and VUR (2.11/3.00±0.63) (Table 2). ChatGPT per-

Table 1. List of generated questions and topics

Question number	Topic	Question	Question type	Difficulty
1	Phimosis	What is the definition of phimosis?	Definition	Easy
2		What is the first-line treatment in symptomatic phimosis?	General	Medium
3		How is phimosis treated surgically?	General	Easy
4		What is the best treatment for asymptomatic phimosis in infants with a risk of recurrent urinary tract infections?	General	Difficult
5		How is pediatric paraphimosis treated?	General	Medium
6	Cryptorchidism	What is the definition of undescended testicle?	Definition	Easy
7		What is the role of imaging studies in the investigation of undescended testicle?	General	Medium
8		When should treatment be initiated for undescended testicles?	General	Medium
9		What is the role of medical therapy for undescended testicle	General	Medium
10		What is the surgical approach to non-palpable testes?	General	Difficult
11	Acute scrotum	What is the definition of a pediatric acute scrotum?	Definition	Easy
12		What is the approach to managing pediatric testicular torsion?	General	Medium
13		Should contralateral orchidopexy be performed for treatment of testicular torsion?	General	Difficult
14		Does torsion of the appendix testis require surgery?	General	Medium
15	Hypospadias	What is the definition of hypospadias? How is it classified?	Definition	Medium
16		Which patients with hypospadias require complete workup to exclude differences in sexual development?	General	Difficult
17		What age is recommended for primary hypospadias repair?	General	Easy
18		What conditions should be monitored for in hypospadias repair follow-up?	General	Difficult
19	Vesicoureteral reflux (VUR)	How is VUR diagnosed?	Definition	Easy
20		Do pediatric patients with VUR need continuous antibiotic prophylaxis?	General	Medium
21		What surgical options exist and are recommended for VUR?	General	Difficult
22		How should low-grade VUR in pediatric patients be managed?	General	Medium
23	Urinary stone disease	What type of stones do pediatric patients develop? What type is most common?	General	Easy
24		What is the initial approach to medical management in a pediatric patient with calcium stones?	General	Medium
25		What is the role of imaging in pediatric stone disease?	General	Easy
26		What metabolic evaluations are done for pediatric patients with stones?	General	Medium
27		Do all pediatric patients with stones require surgery?	General	Easy

formed worse in the domains of acute scrotal pathology (1.90/3.00±0.58) and cryptorchidism (1.92/3.00±0.56) ($F(17,225)=1.503$, $p=0.031$). ChatGPT's appropriateness was 70.4% (n=19/27), 55.6% (n=15/27), and 74.1% (n=20/27) based on CUA, AUA, and EAU, respectively. The highest-scored question was, "What is the definition of an undescended testicle?" (2.78/3.00±0.33), and the lowest-scored question was, "What is the role of imaging studies in the investigation of an undescended testicle?" (0.37/3.00±0.51). No question had a mean score of 0 (completely incorrect).

There was a significant difference in ChatGPT's ability to answer questions based on assigned difficulty ($F(8,252)=7.421$, $p=0.001$). ChatGPT performed better across guidelines when answering "easy" questions (2.31/3.00±0.09) compared to "medium" (1.92/3.00±0.07, $p=0.003$) and "difficult" questions (1.86/3.00±0.101, $p=0.003$) (Figure 1). There was

no difference in ChatGPT's performance between "medium" and "difficult" questions (p=1.00). Similarly, when comparing the repeatability between ChatGPT iterations, there was no difference in average variance between questions (CUA: 0.445, AUA: 0.429, and EAU: 0.448, p=0.21).

Subanalysis on definition-based questions demonstrated greater mean levels of comprehensiveness (definition: 2.54±0.13 vs. non-definition: 1.93±0.05) across all guidelines, with the highest scores being for cryptorchidism (2.78/3.00±0.33) and hypospadias (2.70/3.00±0.35). This was further supported through ANOVA, with definition-based questions having significantly greater comprehensiveness scores across all guidelines (F(5,255)=17.68, p=0.0001).

When analyzing the ability of ChatGPT to answer questions aligned with the CUA, AUA, and EAU guidelines, there was a significant difference in adherence, with the greatest difference in means seen between the EAU and AUA (Figure 2). ChatGPT was more accurate when being referenced to EAU (2.10±0.41) compared to AUA (1.95±0.41) (F(86,696)=1.388, p=0.04).

Table 2. Overall comprehensiveness/performance scores of ChatGPT based on general vs. definition-based questions and between reference guidelines

	CUA	AUA	EAU	Total
General questions				
Phimosis	2.56±0.43	2.22±0.71	2.18±0.57	2.32±0.57
Cryptorchidism	1.91±0.65	1.84±0.71	2.00±0.38	1.92±0.58
Acute scrotum	2.11±0.72	1.69±0.38	1.89±0.57	1.90±0.56
Hypospadias	2.03±0.44	1.92±0.55	2.28±0.50	2.07±0.49
Vesicoureteral reflux	1.72±0.76	2.31±0.54	2.31±0.61	2.11±0.63
Urinary stone disease	2.00±0.67	1.91±0.65	2.20±0.78	2.04±0.70
Definition-based questions				
Phimosis	2.22±0.44	2.44±0.53	2.67±0.50	2.44±0.49
Cryptorchidism	2.67±0.50	2.67±0.50	3.00±0.00	2.78±0.33
Acute scrotum	2.33±0.50	2.00±0.00	2.33±0.50	2.22±0.33
Hypospadias	2.56±0.53	2.56±0.53	3.00±0.00	2.70±0.35

Mean ± standard deviation.

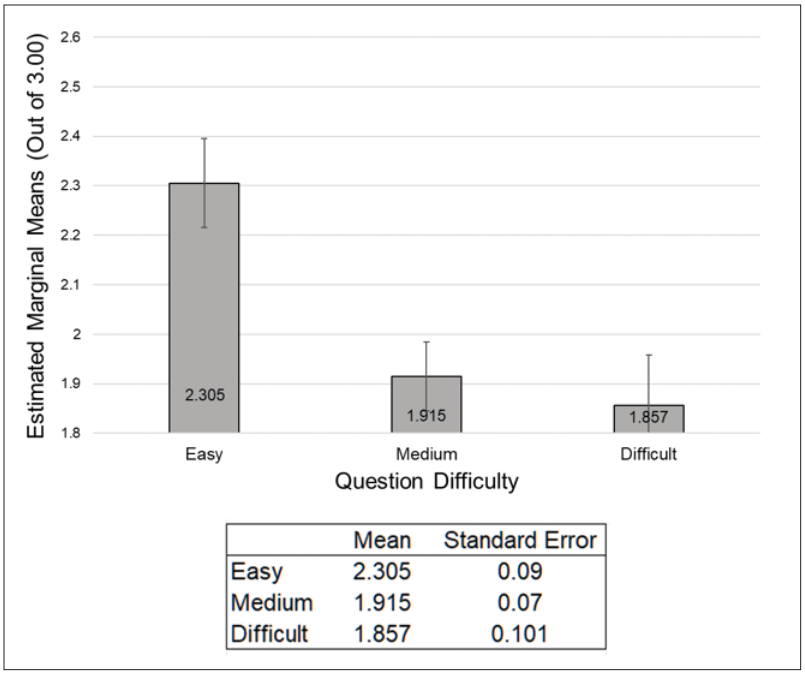


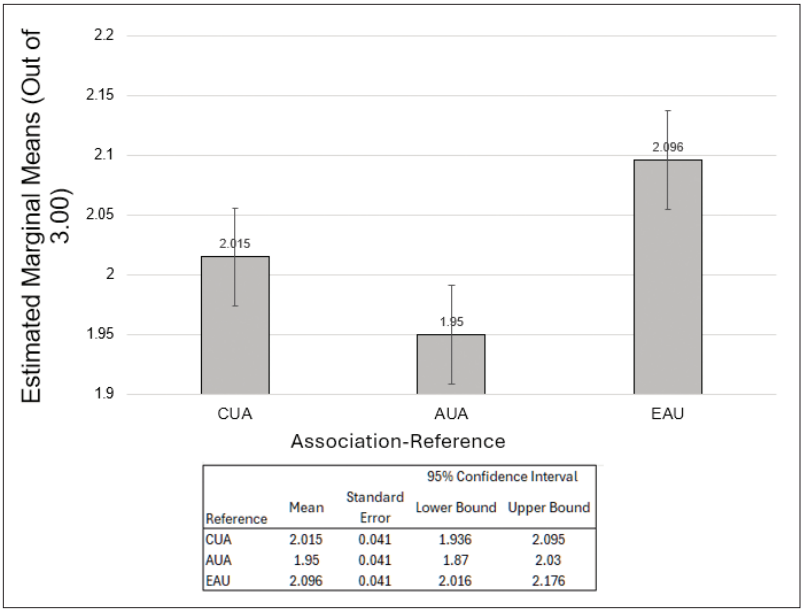
Figure 1. ChatGPT performance based on question topic.

DISCUSSION

The use of AI and NLP in healthcare knowledge transfer shows great promise, with opportunities to provide patients with high-quality, comprehensive, and reliable medical information.[1] When examining parental healthcare-seeking behaviors, the internet and social media are well-established media in which patients gather information.[18,19] Use of NLPs for medical information, although still in its infancy, will only become more prevalent.[20] ChatGPT serves as the forerunner in investigating the capabilities of NLPs as a way to better inform patients in the community.[20]

With regard to pediatric urology, ChatGPT has shown early success in providing satisfactory responses to general inquiries with reference to EAU guidelines.[5] In a study by Caglar et al, ChatGPT was able to answer general pediatric urology questions with a 92.0% "completely correct" rate and strong repeatability.[5] Of note, in that study, ChatGPT avoided answering any question "completely incorrect," thus only the level of comprehensiveness in the answers varied.[5]

When compared to the accuracy and comprehensiveness of information provided by other social media platforms, the accuracy of medical information provided has been shown to vary from 20–50%.[21,22] This demonstrates the considerable improvement in comprehensiveness and accuracy that ChatGPT provides in contrast to other mediums.[20] This is further supported by our findings of ChatGPT providing overall comprehensive results in 56–74% of pediatric urology-based inquiries, with no answer being completely incorrect.



Reference	Mean	Standard Error	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
CUA	2.015	0.041	1.936	2.095
AUA	1.95	0.041	1.87	2.03
EAU	2.096	0.041	2.016	2.176

Figure 2. ChatGPT performance based on association reference. AUA: American Urological Association; CUA: Canadian Urological Association; EAU: European Association of Urology.

Our findings suggest that ChatGPT provides answers to pediatric urology questions with a greater concordance to information provided by the EAU as compared to the AUA and CUA. This novel finding may provide context when examining the accuracy of ChatGPT responses provided in future studies. Furthermore, although agreement exists in most cases between CUA, AUA, and EAU recommendations, this may also further elucidate discrepancies and nuances in the manner these organizations present information for patients.[12-14] Due to the nature of how ChatGPT functions, this may also represent an artifact related to the volume of text surveyed by ChatGPT and the potential European/EAU focus on the topic.[5]

ChatGPT demonstrated the greatest performance in answering questions related to phimosis and VUR. This may be due to the relative frequency which phimosis occurs and its common existence within both the adult and pediatric literature.[23] Furthermore, more comprehensive results generated for VUR may be due to the algorithmic nature of diagnosing and treating this urologic abnormality.[24]

ChatGPT performed worst in the area of acute scrotal pathology, which is of interest and concern. Acute scrotal pathology, inclusive of testicular torsion, is one area of pediatric urology with the greatest time sensitivity and urologic consequence if patients delay emergent presentation.[25] Both pediatric patients and parents/caregivers have low levels of understanding of testicular torsion and the time-sensitive nature of the pathology.[26,27] Furthermore, pre-hospital delay remains the greatest risk to testicular salvage.[25,28] Therefore, reliance on ChatGPT, which may provide suboptimal information on acute scrotal pathology, could potentially delay care and increase the risk of testicular loss. Conversely, in our study, ChatGPT approached the "correct but inadequate" level of comprehensiveness, which may signify more specific and accurate medical information and advice compared to generic internet search responses.

ChatGPT demonstrated greater performance when answering definition-based questions and "easy" questions when compared to general knowledge-based questions and "medium" or "difficult" questions. This is likely due to the interpretation of the prompt by ChatGPT and the more objective criteria for scoring definition-based responses. There were no differences between "medium" and "difficult" questions by ChatGPT, which may be related to the subjective internal scoring of the questions in this study. Future studies evaluating the use of ChatGPT should focus on defining questions based on difficulty and complexity to better classify ChatGPT performance.

Limitations

This study highlights the promise and complexity of ChatGPT for providing medical information to the general public but is not without limitations.

Although question difficulty was defined, there exists subjectivity intrinsic to this assignment, which may introduce bias and variability in our results when compared with others. Additionally, although expert review and discretion was used for output scoring, there exists subjectivity in the interpretation of the ChatGPT outputs and the corresponding score assignments.

Future studies should adopt definition-based and difficulty-based scoring criteria and also compare the quality of ChatGPT answers with non-ChatGPT internet answers to allow for direct comparison in answer quality.

This study also only used the publicly accessible version of ChatGPT. It is important to note that response accuracy may be improved with higher versions of ChatGPT or with AI-search engines that better use retrieval-augmented generation.

Although ChatGPT shows high levels of comprehensiveness and promise, further work is required before referring patients to ChatGPT as a clinical information source for pediatric urology concerns. Furthermore, guidelines or recommendations to best support patients

in using ChatGPT for medical information-gathering will become increasingly important with further adoption of AI-based medical information-gathering by patients.

CONCLUSIONS

ChatGPT demonstrates promise for answering common pediatric urology questions and may serve as a potential alternative for gathering information. In our study, ChatGPT was able to answer common pediatric urology questions with high levels of appropriateness and comprehensiveness. ChatGPT performed worst in the area of acute scrotal pathology, which highlights an area of improvement. Furthermore, ChatGPT performed best when referenced to EAU-based resources when compared to AUA. With further refinement, ChatGPT may one day be seen as a reliable tool for the general public seeking more information on pediatric urology topics.

COMPETING INTERESTS: The authors do not report any competing personal or financial interests related to this work.

REFERENCES

1. Wei Q, Yao Z, Cui Y, et al. Evaluation of ChatGPT-generated medical responses: A systematic review and meta-analysis. *J Biomed Inform* 2024;151:104620. https://doi.org/10.1016/j.jbi.2024.104620
2. Meo SA, Al-Masri AA, Alotaibi M, et al. ChatGPT knowledge evaluation in basic and clinical medical sciences: Multiple-choice question examination-based performance. *Healthcare* 2023;11:2046. https://doi.org/10.3390/healthcare11142046
3. Reynolds K, Tejasvi T. Potential use of ChatGPT in responding to patient questions and creating patient resources. *JMIR Dermatol* 2024;7:e48451. https://doi.org/10.2196/48451
4. Braga AVNM, Nunes CN, Santos EN, et al. Use of ChatGPT in urology and its relevance in clinical practice: Is it useful? *Int Braz J Urol* 2024;50:192-8. https://doi.org/10.1590/s1677-5538.ibju.2023.0570
5. Caglar U, Yildiz O, Meric A, et al. Evaluating the performance of ChatGPT in answering questions related to pediatric urology. *J Pediatr Urol* 2024;20:26.e1-26.e5. https://doi.org/10.1016/j.jpurol.2023.08.003
6. Ray PP. ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations, and future scope. *Internet Things Cyber-Phys Syst* 2023;3:121-54. https://doi.org/10.1016/j.iotcps.2023.04.003
7. Deebel NA, Terlecki R. ChatGPT performance on the American Urological Association self-assessment study program and the potential influence of artificial intelligence in urologic training. *Urology* 2023;177:29-33. https://doi.org/10.1016/j.urology.2023.05.010
8. Touma NJ, Caterini J, Liblk K. Performance of artificial intelligence on a simulated Canadian urology board exam: Is ChatGPT ready for primetime? *Can Urol Assoc J* 2024;18:329-32. https://doi.org/10.5489/cuaj.8800
9. Cakir H, Caglar U, Yildiz O, et al. Evaluating the performance of ChatGPT in answering questions related to urolithiasis. *Int Urol Nephrol* 2023;56:17-21. https://doi.org/10.1007/s11255-023-03773-0
10. Whiles BB, Bird VG, Canales BK, et al. Caution! AI bot has entered the patient chat: ChatGPT has limitations in providing accurate urologic healthcare advice. *Urology* 2023;180:278-84. https://doi.org/10.1016/j.urology.2023.07.010
11. Khondker A, Kwong JCC, Ahmad I, et al. A living scoping review and online repository of artificial intelligence models in pediatric urology: Results from the AI-PEDURO collaborative. *J Pediatr Urol* 2025. https://doi.org/10.1016/j.jpurol.2025.01.035
12. American Urological Association. American Urological Association - Educational Resources. Available at: https://www.urologyhealth.org/educational-resources?product_format=466%7C&language=1122%7C. Accessed Mar 5, 2025
13. European Association of Urology. European Association of Urology - Patient Information. Available at: https://patients.uroweb.org/. Accessed Mar 5, 2025
14. Canadian Urological Association. Canadian Urological Association - Patient Information and Brochures. Available from: https://www.cua.org/patient-information-and-brochures. Accessed Mar 5, 2025
15. OpenAI. ChatGPT. Accessed Mar 5, 2025
16. Sarraju A, Bruemmer D, Van Iterson E, et al. Appropriateness of cardiovascular disease prevention recommendations obtained from a popular online chat-based artificial intelligence model. *JAMA* 2023;329:842. https://doi.org/10.1001/jama.2023.1044
17. IBM Corp. Statistical Package for the Social Sciences (SPSS). Armonk, NY, USA; Accessed Mar 1, 2025
18. Frey E, Bonfiglioli C, Brunner M, et al. Parents' use of social media as a health information source for their children: A scoping review. *Acad Pediatr* 2022;22:526-39. https://doi.org/10.1016/j.acap.2021.12.006
19. Bryan MA, Evans Y, Morishita C, et al. Parental perceptions of the internet and social media as a source of pediatric health information. *Acad Pediatr* 2020;20:31-8. https://doi.org/10.1016/j.acap.2019.09.009
20. Walker HL, Ghani S, Kuemmerli C, et al. Reliability of medical information provided by ChatGPT: Assessment against clinical guidelines and patient information quality instrument. *J Med Internet Res* 2023;25:e47479. https://doi.org/10.2196/47479
21. O'Sullivan NJ, Nason G, Manecksha RP, et al. The unintentional spread of misinformation on 'TikTok': A paediatric urological perspective. *J Pediatr Urol* 2022;18:371-5. https://doi.org/10.1016/j.jpurol.2022.03.001
22. Toksoz A, Duran MB. Analysis of videos about vesicoureteral reflux on YouTube. *J Pediatr Urol* 2021;17:858.e1-858.e6. https://doi.org/10.1016/j.jpurol.2021.10.006
23. Aziz Filho AM, Soares de Azevedo LM, Carrijo Rochael M, et al. Frequency of lichen sclerosis in children presenting with phimosis: A systematic histological study. *J Pediatr Urol* 2022;18:529.e1-529.e6. https://doi.org/10.1016/j.jpurol.2022.06.030
24. Chirico V, Tripodi F, Lacquaniti A, et al. Therapeutic management of children with vesicoureteral reflux. *J Clin Med* 2023;13:244. https://doi.org/10.3390/jcm13010244
25. Zee RS, Bayne CE, Gomella PT, et al. Implementation of the accelerated care of torsion pathway: A quality improvement initiative for testicular torsion. *J Pediatr Urol* 2019;15:473-9. https://doi.org/10.1016/j.jpurol.2019.07.011
26. MacDonald C, Burton M, Carachi R, et al. Data and data illustrations supporting the analysis of transcripts from interviews exploring the views and experiences of young men and their parents/guardians regarding testicular health. *Data Brief* 2020;32:106106. https://doi.org/10.1016/j.dib.2020.106106
27. MacDonald C, Burton M, Carachi R, et al. Why adolescents delay with presentation to hospital with acute testicular pain: A qualitative study. *J Pediatr Surg* 2021;56:614-9. https://doi.org/10.1016/j.jpedsurg.2020.06.041
28. Madsen SMD, Rawashdeh YF. Assessing timeline delays associated with utilization of ultrasound diagnostics in paediatric acute scrotum, pre and per COVID-19 pandemic. *J Pediatr Urol* 2023;19:653.e1-653.e7. https://doi.org/10.1016/j.jpurol.2023.07.003

CORRESPONDENCE: Dr. Daniel T. Keefe, Department of Pediatric Urology, IWK Health Centre, Dalhousie University, Halifax, NS, Canada; Daniel.keefe@iwk.nshealth.ca