Cureus
Part of SPRINGER NATURE

Open Access Original Article

Evaluating Online Patient Information and AI-Generated Advice for Testicular Microlithiasis: Readability and Information Quality

Review began 10/06/2026
Review ended 11/01/2026
Published 11/04/2026

© **Copyright** 2026
Turacli et al. This is an open access article distributed under the terms of the Creative Commons Attribution License CC-BY 4.0., which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

DOI: 10.7759/cureus.119119

Huseyin Turacli [1], Emane Ouadam [1], Samuel Pathmanathan [1], Faiz A. Shaikh [1], Jas Kalsi [1]

1. Urology, Wexham Park Hospital, Frimley Health NHS Foundation Trust, Slough, GBR

Corresponding author: Emane Ouadam, emane1@hotmail.co.uk

Abstract

Background

Testicular microlithiasis (TM) is commonly found incidentally on scrotal ultrasound, often resulting in considerable anxiety due to conflicting malignancy associations. Patients increasingly use the internet and large language models (LLMs) for information on health. Health Education England advises patient materials be written at a reading age of 11. We assessed the readability and quality of patient-facing resources and AI-generated adviceregardingTM.

Methods

Google (Google LLC, Mountain View, CA, USA), Bing (Microsoft Corporation, Redmond, WA, USA), and Yahoo (Yahoo Inc., New York, NY, USA) were searched for "Testicular Microlithiasis," with the first two pages of results screened. After deduplication and exclusion of professional-facing sources, 13 English-language, patient- and public-facing websites were included. One response from each of GPT-5.2 Instant (OpenAI, San Francisco, CA, USA), Gemini 3 Flash (Google LLC, Mountain View, CA, USA), and Claude Haiku 4.5 (Anthropic, San Francisco, CA, USA) was generated using a standardized patient-information prompt. Readability was assessed using the Flesch Reading Ease Score (FRES), Flesch-Kincaid Grade Level (FKGL), and Gunning Fog Index (GFI), while information quality was assessed using DISCERN. Exploratory comparisons between website and AI-generated content and between Google search subgroups used exact permutation tests based on the Mann-Whitney U statistic, with Holm adjustment across four outcomes separately for each comparison.

Results

The 13 websites had poor readability (mean FRES 19.0; FKGL 15.3; GFI 17.8). AI responses had a higher mean FRES (41.2) and lower mean FKGL (11.1) and GFI (12.8), but none of the 16 resources met the adopted readability benchmarks. Mean summed DISCERN was 52.1. No website-AI comparison was statistically significant after Holm adjustment (all adjusted $p \geq 0.157$); no Google-subgroup comparison was significant either (all adjusted $p = 1.000$). Two commercial resources contained potentially misleading symptom or dietary advice. None of the three AI responses provided verifiable source citations.

Conclusions

Within this evaluated sample, patient-facing TM resources exceeded both stringent and standard readability benchmarks and included occasional misleading commercial advice. Descriptively, the three AI responses had lower mean textual complexity and aligned with European Society of Urogenital Radiology (ESUR) management guidance, but exploratory statistical comparisons with websites were inconclusive, and all AI outputs lacked verifiable citations. Given the small sample, these findings should not be generalized to all online or AI resources; however, urologists should consider directing patients to clinician-reviewed, plain-English materials at diagnosis.

Categories: Urology, Epidemiology/Public Health, Healthcare Technology
Keywords: generative ai, health literacy, infodemiology, large language models, patient education, readability, testicular microlithiasis

Introduction

Testicular microlithiasis (TM) is an ultrasonographic finding characterized by multiple small, non-shadowing echogenic foci inside the testicular parenchyma, corresponding histologically to microscopic calcifications within the seminiferous tubules [1]. It is oftendiagnosed as an incidentalfinding duringscrotal ultrasonography performed for unrelated symptoms or clinical concerns and may occur in otherwise asymptomatic individuals [2]. Although TM itself does not represent a testicular tumor, its clinical significance has been debated for several decades because of its reported link with testicular germ cell tumors and germ cell neoplasia in situ [3].

How to cite this article
Turacli H, Ouadam E, Pathmanathan S, et al. (November 04, 2026) Evaluating Online Patient Information and AI-Generated Advice for Testicular Microlithiasis: Readability and Information Quality. Cureus 18(11): e119119. DOI 10.7759/cureus.119119

Historically, this association generated considerable concern. Early observational reportsidentifiedarelatively highprevalence of concurrent testicular malignancy among patients with TM, and some authorsconsequentlydescribed TM as a potentially premalignant condition and advocated regular surveillance [4].Laterresearch hasdemonstratedthat an association between TM and malignancy does notestablishTM itself as a precursor to cancer. A systematic review concluded that, in the absence of additional risk factors, the risk of developing testicular germ cell tumor was similar to that of the general population,whereasthe presence ofadditionalrisk factors was associated with increased risk[5].European Society of Urogenital Radiology (ESUR) guidance therefore does not recommend routine ultrasonographic surveillance or biopsy for isolated TM, while additional follow-up may be appropriate in individuals with risk factors such as previous germ cell tumor, testicularmaldescentor orchidopexy, testicular atrophy, or a relevant family history [6].

This evolution in understanding makesaccuratepatient communication particularly important. A patient may be unexpectedly informed that an ultrasound has demonstrated testicular "microlithiasis" and may encounter online material describing historical associations with cancer without sufficient explanation of absolute risk, accompanying risk factors, or contemporary surveillance recommendations. Unnecessary investigation has itself previously been recognized as a potential source of patient anxiety [7].More broadly, online health-information seeking has been associated with health anxiety [8]. This highlights the importance of presenting potentially alarming findings in a balanced and comprehensible manner.

For such information to support informed decision-making, it must be both reliable and accessible.Public health literacy guidelines, such as those fromHealth Education England, advocate that patient-facing literature targets an 11-year-old reading level[9].However, patients now obtain information not only from conventional websites but also from generative artificial intelligence systems and large language models (LLMs). The rapid adoption of these tools issubstantial: in June 2025, approximately 15.8 million UK online adults (32%) visited at least one major AI chatbot [10]. These tools may translate complex medical concepts into more accessible language, but concerns remain regarding accuracy, consistency, and transparency of information sources [11].

Given the clinical uncertainty historically surrounding TM and the potential consequences of misunderstanding its relationship with malignancy, the quality of patient-facing information is particularly relevant.The primary objective of this study was to evaluate the readability (using Flesch Reading Ease Score (FRES), Flesch-Kincaid Grade Level (FKGL), and Gunning Fog Index (GFI)) and information quality (using the DISCERN instrument) of publicly accessible, patient-facing websites (identified via Google (Google LLC, Mountain View, CA, USA), Bing (Microsoft Corporation, Redmond, WA, USA), and Yahoo (Yahoo Inc., New York, NY, USA)) and AI-generated patient information leaflets (produced by GPT-5.2 Instant (OpenAI, San Francisco, CA, USA), Gemini 3 Flash (Google LLC, Mountain View, CA, USA), and Claude Haiku 4.5 (Anthropic, San Francisco, CA, USA)) on TM. Secondary exploratory objectives were to compare readability and DISCERN scores between traditional websites and AI-generated responses and between the five highest-ranked Google websites and lower-ranked websites.

Materials And Methods

In this cross-sectional study, internet searches were conducted on 21 December 2025 through the Google, Bing, and Yahoo public search interfaces using the search term “Testicular Microlithiasis." The first two pages of results returned by each interface were screened. Original search positions were recorded before applying eligibility criteria. Considering Yahoo relies on Bing-powered indexing, it was included solely to reflect real-world consumer search pathways rather than as an independent search index. Across 49 total search appearances (16 Google, 16 Bing, 17 Yahoo), deduplication removed 28 repeats, leaving 21 unique websites. Resources were eligible when they provided English-language information about TM intended primarily for patients or the general public. Eligibility was not restricted to formal patient information leaflets or healthcare-provider websites. General reference and educational webpages were included when they explained the condition, diagnosis, risks, or management for nonspecialist readers. Wikipedia (Wikimedia Foundation, Inc., San Francisco, CA, USA) provided a general reference overview, Health Jade provided a health-information article covering causes, diagnosis, and management, and Biology Insights (Biology Insights LLC, Chicago, IL, USA) addressed the nature and seriousness of the condition for general readers. These resources were included under the general-public criterion rather than as formal patient information leaflets. Eligibility was assessed from the purpose and content of the resource; public accessibility alone was insufficient. Eight websites directed mainly at healthcare professionals were excluded, leaving 13 eligible patient/public-facing websites. Each included resource was analyzed once, while its original position within each interface was retained.

Three consumer-facing generative artificial intelligence models were evaluated separately: GPT-5.2 Instant through ChatGPT, Gemini 3 Flash through Google Gemini, and Claude Haiku 4.5 through Claude.These models were selected torepresentthree independently developed, publicly accessible consumer-facing generative AI platforms available at the time of data collection, allowing comparison across different model ecosystems.All models were accessed on 21December 2025 using their standard free consumer interfaces. A new conversation was opened for each model to minimize the effect ofpreviousconversational context. Each model received the identical prompt: "Write a patient information leaflet explaining the management of

testicular microlithiasis." No follow-up prompts weresubmitted, and no dedicated web-search or research mode was deliberately activated. The resulting responses were evaluated using the same readability and quality measures as the included websites.

Textual complexity was quantified via the FRES, FKGL, and GFI[12-14].Scores were calculated using the Readability Scoring System provided by ReadabilityFormulas.com[15].For readability analysis, the main body text was manually copied into plain-text format to strip HTML tags. Navigation menus, advertisements, cookie notices, author biographies, contact details, and reference lists were excluded, whereas substantive subheadings and bullet-point content were retained (with full stops appended to unpunctuated bullet items to prevent artificial sentence-length inflation). The same protocol was applied to the complete AI-generated responses without substituting the polysyllabic diagnostic term "testicular microlithiasis," thereby preserving the authentic text encountered by patients.

Higher FRES values reflect easier text, whereas lower FKGL and GFI values indicate simpler educational requirements [12-14]. A summary of the readability formulae, interpretation bands, and DISCERN scoring framework used in this study is provided in Appendix 1. Health Education England recommends that written patient information should be understandable at a typical 11-year-old reading age [9]. For contextual comparison, FKGL ⩽6 (corresponding to an 11- to 12-year-old reading age), GFI <8 (broadly accessible text), and FRES ⩾90 (a stringent upper benchmark adopted from a previous UK infodemiology study [16], alongside FRES ⩾60 as standard readability [12]) were used as reference benchmarks [16,17]. Content reliability was evaluated by applying the DISCERN criteria [18].Following a previously published classification, we adopted the categorization of summed DISCERN scores asvery poor, poor, fair, good, or excellent [19]. Questions 1-8 assess publication reliability, while Questions 9-15 assess the quality of information about treatment choices.Each item is scored from one to five. Scores for Questions 1-15 were summed to produce a total between 15 and 75 andwere classified asvery poor(15-26), poor (27-38), fair (39-50), good (51-62), or excellent (63-75)[19]. Question 16, which provides an overall quality rating from one to five, was analyzed separately and was not included in the summed score[18,19]. All DISCERN ratings were initially assigned by HT using the official DISCERN handbook descriptors and subsequently audited item-by-item by EO and SP. Because scoring used a sequential multi-author audit rather than blinded independent parallel rating, formal inter-rater reliability coefficients (such as Cohen's kappa) were not calculated; instead, any item where reviewers proposed a discrepancy of ⩾1 point was jointly re-evaluated against the source text until 100% consensus agreement was reached across all 16 resources (256 scored items). Both websites and AI-generated responses were also cross-checked against ESUR guidelines [6] to identify any clinically inaccurate statements regarding symptoms, diet, or surveillance.

FRES, FKGL, GFI, and summed DISCERN scores were summarized using the mean, sample standard deviation, range, and 95% confidence interval of the mean. Confidence intervals were calculated using Student’s t-distribution. DISCERN domain scores and Question 16 ratings were summarized as mean ± sample standard deviation. Statistical analyses added during revision were exploratory. Two-sided exact permutation tests based on the Mann-Whitney U statistic compared the four score distributions between the 13 websites and three AI responses and between the five highest-positioned eligible Google resources and the remaining eight websites. Tied values received average ranks. All 560 possible website-AI label allocations and all 1,287 possible Google-subgroup label allocations were enumerated. Two-sided p-values were twice the smaller inclusive permutation-tail probability, capped at 1. Unadjusted and Holm-adjusted p-values across the four outcomes were reported separately for each comparison; an adjusted p-value below 0.05 was considered statistically significant. Analyses were performed in Python 3.12 (Python Software Foundation, Fredericksburg, VA) with SciPy 1.17.0 using the resource-level scores in Appendix 2. Given the purposive sample and one response per AI platform, inferential findings were interpreted cautiously. The study analyzed publicly accessible online material and AI-generated text and did not include human participants, patient records, or biological samples.

Results

The search log recorded 49 website-result appearances: 16 from Google, 16 from Bing, and 17 from Yahoo. Removal of 28 repeat appearances left 21 unique websites. Eight professional-facing websites were excluded, leaving 13 eligible patient/public-facing websites. These were analyzed alongside three separately generated AI responses, giving 16 evaluated resources. Individual readability and DISCERN scores are provided in Appendix 2. The 13 websites appeared nine times on Google, 10 on Bing, and 11 on Yahoo (which added no unique websites beyond Bing). The highest-positioned eligible resource was the NHS Tayside leaflet, ranked second on Google and first on both Bing and Yahoo. The five highest-positioned eligible Google resources appeared at original positions 2, 3, 4, 9, and 11. Source categories and original search positions are summarized in Table *1*.

Resource	Source Category	Google Rank	Bing Rank	Yahoo Rank
NHS Tayside Patient Leaflet	Government / NHS	2	1	1
Mayo Clinic	Clinical provider (Non-Profit)	3	7	8
Wikipedia	Online encyclopedia	4	3	3
Prof. Dr. Emin Özbek Clinic	Private clinical practice	9	9	9
Urologenzentrum	Private clinical practice	11	NR	NR
Staffordshire Urology Clinic	Private clinical practice / consultant website	13	NR	NR
Medicover Hospitals	Private hospital network	14	NR	17
Mentone General Practice	General practice	15	13	15
Happiest Health	Health news / wellness portal	16	11	14
Health Jade	Online health encyclopedia	NR	16	10
NatuVitro	Private fertility clinic	NR	12	16
The Kingsley Clinic	Private telemedicine clinic	NR	4	4
Biology Insights	Educational / science website	NR	15	13

TABLE 1: Search-result positions and source categories of included patient/public-facing resources for testicular microlithiasis

Positions are the original unfiltered search-result ranks recorded before application of eligibility criteria. Yahoo uses Bing-generated results, so Bing and Yahoo positions are not independent rankings. Resources appearing across interfaces were analyzed once. AI responses were generated separately by direct prompting and are not included in these rankings.

NR: not returned within the evaluated search results; NHS: National Health Service

Google (Google LLC, Mountain View, CA, USA); Yahoo (Yahoo Inc., New York, NY, USA); Bing (Microsoft Corporation, Redmond, WA, USA); Wikipedia(Wikimedia Foundation, Inc., San Francisco, CA, USA); Happiest Health (Happiest Health Systems Pvt. Ltd., Bengaluru, Karnataka, India); Biology Insights(Biology Insights LLC, Chicago, IL, USA)

Across all 16 resources, the mean FRES was 23.2 (95% CI 14.7-31.6), while the mean FKGL and GFI were 14.5 (95% CI 12.9-16.1) and 16.9 (95% CI 15.1-18.6), respectively. The mean summed DISCERN score was 52.1 (95% CI 47.7-56.6). Using the prespecified FRES interpretation bands, 11 of 16 resources (68.8%) were classified asvery difficultto read, four (25.0%) as difficult, and one (6.3%) asfairly difficult. No resource achieved a standard, fairly easy, easy, or very easy reading level. Among the 13 websites, 11 (84.6%) were classified as very difficult to read. Across all 16 resources, none met any of the adopted benchmarks of FRES ⩾90, FKGL ⩽6, or GFI <8 (0/16), nor did any achieve a standard reading ease score of FRES ⩾60 (0/16). Aggregate readability and information-quality outcomes are summarized in Table *2*.

Assessment Metric	Overall Mean Score	Standard Deviation	95% CI	Range (Min - Max)	Reference/Interpretation Criterion
Flesch Reading Ease (FRES)	23.2	15.8	14.7 - 31.6	4.4 - 52.0	≥90=very easy readability; adopted patient-facing benchmark [16]
Flesch-Kincaid Grade Level (FKGL)	14.5	3.0	12.9 - 16.1	9.4 - 18.2	US Grade 6 or below: adopted patient-facing benchmark [16]
Gunning-Fog Index (GFI)	16.9	3.3	15.1 - 18.6	11.6 - 20.6	<8: broadly accessible text [17]
Total DISCERN Score	52.1	8.3	47.7-56.6	31.0 - 65.0	63-75: excellent information quality under adopted classification [19]

TABLE 2: Aggregate readability and information-quality metrics across all evaluated resources (N=16)

Readability reference criteria are described in Appendix1and should be interpreted according to thecited readability guidance. The DISCERN criterionrepresentsthe prespecified summed quality classification used in this study and is not a patient-literacy threshold.

CI: confidence interval; SD: standard deviation; FRES: Flesch Reading Ease Score; FKGL: Flesch-Kincaid Grade Level; GFI: Gunning-Fog Index

The three AI-generated responses had lower textual complexity on average than the 13 websites across all three readability measures. Mean FRES was 41.2 (range 35.8-47.9) for AI responses compared with 19.0 (range 4.4-52.0) for websites, while mean FKGL was 11.1 (range 9.6-11.9) compared with 15.3 (range 9.4-18.2), and mean GFI was 12.8 (range 11.6-14.1) compared with 17.8 (range 11.9-20.6). All three AI responses were classified as difficult to read. Mean summed DISCERN scores were 54.0 (range 52.0-57.0) for AI responses and 51.7 (range 31.0-65.0) for websites. Exploratory exact permutation tests gave unadjusted p-values of 0.057 for FRES, 0.057 for FKGL, 0.039 for GFI, and 0.779 for summed DISCERN. Corresponding Holm-adjusted p-values were 0.171, 0.171, 0.157, and 0.779, respectively; none were statistically significant after adjustment. Group summaries and exploratory comparisons are presented in Table *3*.

Assessment Metric	Websites (n=13), mean ± SD	Websites 95% CI	AI Responses (n=3), mean ± SD	AI Responses 95% CI	Unadjusted p-value	Holm-adjusted p-value
Flesch Reading Ease (FRES)	19.0 ± 14.4	10.3 - 27.7	41.2 ± 6.1	26.0 - 56.5	0.057	0.171
Flesch-Kincaid Grade Level (FKGL)	15.3 ± 2.7	13.7 - 17.0	11.1 ± 1.3	7.9 - 14.2	0.057	0.171
Gunning-Fog Index (GFI)	17.8 ± 2.9	16.1 - 19.6	12.8 ± 1.3	9.7 - 15.9	0.039	0.157
Total DISCERN Score	51.7 ± 9.2	46.1 - 57.2	54.0 ± 2.6	47.4 - 60.6	0.779	0.779

TABLE 3: Comparative analysis of readability and quality outcomes between traditional websites and artificial intelligence language models

Data are presented as mean ± SD with 95% CIs. Higher FRES scoresindicateeasier readability,whereaslower FKGL and GFI scoresindicatelower reading complexity. The wide confidence intervals for the generative AI group reflect the small number of evaluated models (n=3). P-values are from exploratory two-sided exact permutation tests based on the Mann-Whitney U statistic, using average ranks for ties. Holm adjustment accounts for the four website–AI outcomes. No comparison was statistically significant after adjustment.

AI: artificial intelligence; CI: confidence interval; SD: standard deviation; FRES: Flesch Reading Ease Score; FKGL: Flesch-Kincaid Grade Level; GFI: Gunning-Fog Index

Eight of the 13 websites (61.5%) were classified as good or excellent according to the prespecified summed DISCERN categories, while four were classified as fair and one as poor. All three AI responses were classified as good by the summed DISCERN score. At the domain level, websites had a higher mean reliability score than AI-generated responses (29.9/40 vs 27.3/40), whereas AI responses had a higher mean treatment-information score (26.7/35 vs 21.8/35). Mean global DISCERN ratings were 3.4/5 for websites and 4.0/5 for AI

responses. None of the three AI-generated responses provided verifiable source citations.The corresponding DISCERN domain scores are summarized in Table 4.Domain-level and Question 16 comparisons were descriptive.

Resource Group	Section 1: Reliability, mean ± SD /40	Section 2: Treatment information, mean ± SD /35	Summed DISCERN, mean ± SD /75	Question 16: Overall quality, mean ± SD /5
Traditional Websites (n=13)	29.9 ± 6.3	21.8 ± 3.8	51.7 ± 9.2	3.4 ± 1.4
AI Responses (n=3)	27.3 ± 1.5	26.7 ± 1.2	54.0 ± 2.6	4.0 ± 0.0

TABLE 4: Thematic quality evaluation utilising the DISCERN framework domains

Values are mean ± sample SD across the evaluated resources. DISCERN Section 1 comprises Questions 1–8 and assesses reliability (maximum 40 points); Section 2 comprises Questions 9–15 and assesses information about treatment choices (maximum 35 points). The summed score comprises Questions 1–15 (maximum 75 points). Question 16 is a separate overall quality rating from 1 to 5 and is excluded from the summed score [18,19]. Domain-level and Question 16 comparisons were descriptive.

AI: artificial intelligence; SD: standard deviation

When evaluated against ESUR clinical guidance [6], two commercial websites contained potentially misleading statements: one associated TM with dysuria and hematuria, while another recommended dietary calcium restriction. By contrast, all three AI-generated responses aligned with ESUR management recommendations, advising reassurance and routine testicular self-examination for isolated TM while indicating specialist follow-up or ultrasound surveillance only when additional germ cell tumor risk factors are present.

The five highest-positioned eligible patient/public-facing resources identified on Google had a mean FRES of 18.1 (SD 19.9, range 4.4-52.0, 95% CI −6.6-42.8), a mean FKGL of 15.6 (SD 3.7, range 9.4-18.2, 95% CI 11.0-20.2), and a mean GFI of 18.1 (SD 3.6, range 11.9-20.6, 95% CI 13.7-22.5). None achieved the prespecified patient readability target, and four of the five had FRES values below 30, placing them within the very difficult readability category. Their mean summed DISCERN score was 54.2 (SD 7.7, range 46.0-65.0, 95% CI 44.6-63.8), with one resource classified as excellent, two as good, and two as fair. The readability and DISCERN scores of these resources are shown in Table 5.

Google Position	Source Designation	FRES	FKGL	GFI	DISCERN Score
2	NHS Tayside Patient Leaflet	52.0	9.4	11.9	65 - excellent
3	Mayo Clinic	19.6	14.8	18.2	59 - good
4	Wikipedia	8.6	17.6	19.9	49 - fair
9	Prof. Dr Emin Özbek Clinic	4.4	17.8	19.9	52 - good
11	Urologenzentrum	5.7	18.2	20.6	46 - fair
Mean ± SD	-	18.1 ± 19.9	15.6 ± 3.7	18.1 ± 3.6	54.2 ± 7.7

TABLE 5: Readability and information quality of the five highest-positioned eligible patient-facing Google resources for testicular microlithiasis

Google positionrepresentsthe original unfiltered search-result position recorded before application of eligibility criteria. Lower Google position numbersindicategreater search prominence. Higher FRES scoresindicateeasier readability,whereaslower FKGL and GFI scoresindicatelower reading complexity.DISCERN categories follow the classification described in Methods [19]. The final row gives the mean ± sample SD across the five resources.

FRES: Flesch Reading Ease Score; FKGL: Flesch-Kincaid Grade Level; GFI: Gunning-Fog Index

Google (Google LLC, Mountain View, CA, USA); Wikipedia(Wikimedia Foundation, Inc., San Francisco, CA, USA)

The assessment frameworks and individual resource scores are provided in Appendices 1, 2, respectively. For

comparison, the remaining eight eligible websites had a mean FRES of 19.6 (SD 11.2, range 10.8-44.5, 95% CI 10.2-29.0), mean FKGL of 15.2 (SD 2.2, range 10.4-17.7, 95% CI 13.3-17.0), mean GFI of 17.7 (SD 2.6, range 12.2-20.2, 95% CI 15.5-19.8), and mean summed DISCERN score of 50.1 (SD 10.2, range 31.0-62.0, 95% CI 41.6-58.6). Three of the five highest-positioned Google resources (60.0%) were classified as good or excellent by DISCERN despite none meeting the prespecified patient readability target. Exact permutation tests comparing the two groups gave unadjusted p-values of 0.435 for FRES, 0.524 for FKGL, 0.497 for GFI, and 0.803 for summed DISCERN; all four Holm-adjusted p-values were 1.000. These exploratory comparisons did not provide evidence of group differences, but the small sample does not rule out an association between search position and readability or information quality.

Discussion

This study identified uniformly poor readability across patient-facing online resources on TM, with neither traditional websites nor AI-generated responses achieving the recommended 11-year-old reading level. Information quality was stronger than readability, with the mean DISCERN score falling within the 'good' category. The problem, therefore, appears to be not simply a lack of relevant content but difficulty communicating it in a form that patients can readily understand and verify.

This distinction is important because TM is usually an incidental ultrasonographic finding rather than a condition requiring active treatment. Contemporary evidence indicates that isolated TM, in the absence of recognized additional risk factors, does not itself justify routine ultrasonographic surveillance or biopsy, while follow-up is directed by other risk factors for germ cell tumor [5,6]. Poorly contextualized information may lead patients to overestimate the significance of the finding, particularly when historical associations with malignancy are presented without explaining current risk stratification.A systematic reviewand meta-analysishasalso reported a positive association between online health-information seeking and health anxiety[8]. Patient information on TM should therefore contextualise risk and provide clear safety-netting rather than simply listingpossible associations.

The readability gap was substantial, consistent with prior evidence that online urological patient materials frequently exceed recommended reading levels [20]. Although the adopted FRES ⩾90 benchmark is notably stringent (approximating a US Grade 5 level), no evaluated resource reached even the standard readability band (FRES ⩾60) or the 11-year-old reading age target (FKGL ⩽6). Readability was inevitably penalized by the 10-syllable diagnostic term "testicular microlithiasis," where both words contain ⩾3 syllables and directly inflate FRES, FKGL, and GFI scores [12-14]. While substituting a monosyllabic placeholder during analysis would artificially improve scores, we evaluated the intact text to reflect the real-world linguistic burden faced by patients, alongside substantial heterogeneity across source types, ranging from concise NHS patient leaflets to dense online encyclopedias, wellness portals, and private clinic pages.

Exploratory comparison between the five highest-ranked Google websites and the remaining eight websites showed no statistically significant differences in readability or DISCERN scores after Holm adjustment. Thus, higher search prominence did not guarantee greater accessibility or clinical usefulness [21]. Clinicians should therefore consider signposting patients to specific reviewed resources.

Whereas all three AI responses accurately reflected ESUR management guidelines [6], two commercial websites contained misleading claims associating TM with dysuria and hematuria or advising dietary calcium restriction. TM is generally asymptomatic, and ESUR patient guidance states that testicular calcium deposits are unrelated to diet [1,6]. Such advice may misdirect patients towards irrelevant symptoms or unnecessary lifestyle changes. This also illustrates why a global quality score cannot substitute for factual review. A resource may performreasonably acrossDISCERN items while stillcontainingan individual statement capable of misleading a patient.

The AI findings present a different balance of advantages and limitations. AI-generated responses had a higher mean FRES and lower FKGL and GFI than websites, consistent with evidence that LLMs can simplify patient education material [11,22]. None of the four exploratory website-AI comparisons was statistically significant after Holm adjustment, so these observed differences should be interpreted cautiously. Nevertheless, all three remained above the target reading level, so AI should not be assumed to produce plain-language material automatically. Although overall DISCERN scores were similar between groups, AI responses scored descriptively higher on treatment-choice items and lower on reliability. This treatment-domain advantage should be interpreted cautiously, as our prompt explicitly requested a management leaflet, whereas the included websites encompassed heterogeneous formats (such as general reference encyclopedias and clinic overviews) not uniformly structured around management.

None of the three AI responses provided verifiable source citations. This limits the ability of patients and clinicians to distinguish evidence-based guidance from unsupported model output. Similar findings have been reported in other studies evaluating LLM-generated patient information, highlighting inconsistent readability, a lack of citations, and the importance of clinical oversight [23].LLMs may generate plausible but factually inaccurate statements, highlighting the need for human review of AI-generated healthcare information [22]. These findings support AI as a tool for drafting or simplifying patient information rather than as a standalone clinical information source. A safer approach would combine AI-assisted plain-

language rewriting with guideline-based source material and final clinician review.

Future TM resources should clearly define the condition, distinguish it from renal or urinary calculi, explain the significance of cancer associations in the context of individual risk factors, and state when further assessment is or is not recommended [6].Authorship, a review date, and links or citations to supporting guidance should also be visible [18].A concise resource provided at the point of diagnosis may be especially valuable because it can frame the finding before patientsencounterheterogeneous online material.

Strengths include examination of three public search interfaces, preservation of original search positions, resource-level deduplication, three complementary readability measures, and a standardized AI prompt. However, several limitations warrant emphasis. First, the purposive sample was small (13 heterogeneous websites and a single output from each of three AI models), meaning all between-group statistical tests were strictly exploratory and underpowered to detect modest differences. Second, screening only the first two results pages for a single English search term at one time point introduces potential selection bias and limits generalizability. Third, syllable-based readability formulae only estimate syntactic and lexical complexity rather than actual patient comprehension, which is also shaped by layout, visual aids, prior health literacy, and anxiety [24]; future studies should perform sensitivity analyses removing polysyllabic medical terms and directly test patient comprehension. DISCERN was developed for written information about treatment choices [18],whereasisolated TM often requires explanation and risk stratification rather than active treatment, so its treatment-focused structure may not capture every feature of an ideal TM resource.

Although DISCERN ratings were reviewed by two additional authors, initial scoring was undertaken by a single assessor rather than by independent raters, so some subjectivity remains. Only one response from each AI model was assessed. Generative AI systems are updated frequently, and their outputs are probabilistic, meaning that identical prompts submitted later or in another session may not reproduce the responses evaluated here[25]. Future studies should include repeated AI sampling, direct patient assessment of understanding, and a prespecified guideline-concordance accuracy measure.Furthermore, the use of a single standardized prompt improvedcomparability between models but may also limit generalizability, as LLM outputs can vary according to prompt formulation[25]. More specific instructions, such as requesting a defined reading level, shorter sentences, or supporting references, may produce different readability and quality outcomes [22,25].Laterstudies could systematically compare different prompting strategies and repeated responses todeterminehow prompt design influences patient-facing information on TM. Additionally,the findingsmay not generalize to other generative AI platforms or later model versions.Browser and personalization settings were not recorded, which limits the reproducibility of the observed search rankings.

Overall, within the specific resources assessed, descriptive findings showed reasonable average information quality alongside universally poor readability and occasional commercial inaccuracies. While the three evaluated AI responses were descriptively less complex and guideline-concordant, exploratory statistical comparisons with websites were non-significant, and all AI outputs lacked verifiable sourcing. Since these findings reflect a small, cross-sectional sample rather than all online or LLM-generated health advice, conclusions regarding AI performance remain preliminary. In clinical practice, signposting a concise, clinician-reviewed TM leaflet at diagnosis may help reduce patient reliance on complex websites and uncited AI outputs.

Conclusions

Within the 16 TM resources evaluated in this study, readability was consistently poor despite generally good DISCERN quality scores, with no resource meeting either the stringent (FRES ⩾90) or standard (FRES ⩾60; FKGL ⩽6) readability benchmarks. Descriptively, the three AI-generated responses demonstrated lower mean textual complexity than the 13 websites and aligned with ESUR clinical guidance, whereas two commercial websites contained misleading symptom or dietary claims. However, exploratory statistical comparisons between websites and AI responses were inconclusive after Holm adjustment, and all AI outputs lacked verifiable source citations.

Given the restricted search strategy, single prompt per AI platform, and evaluation of textual complexity rather than actual patient comprehension, these conclusions are limited to the sampled resources. They should not be generalized to all online materials or generative AI systems. Nevertheless, our findings highlight the practical need for clinician-reviewed, plain-English patient resources at the point of diagnosis that distinguish TM from urinary calculi, contextualise malignancy risk factors in line with ESUR guidance, and provide clear safety-netting.

Appendices

Appendix 1

Measure	Formula or derivation	Direction of score	Interpretation used in this study
Flesch Reading Ease Score (FRES)	206.835 - 1.015 × (words/sentences) - 84.6 × (syllables/words) [12]	Higher scores indicate easier reading	<30: very difficult; 30-49.9: difficult; 50-59.9: fairly difficult; 60-69.9: standard; 70-79.9: fairly easy; 80-89.9: easy; ≥90: very easy. A FRES of ≥90 was used as the patient-facing readability benchmark [16].
Flesch-Kincaid Grade Level (FKGL)	0.39 × (words/sentences) + 11.8 × (syllables/words) - 15.59 [13]	Lower scores indicate a lower required educational level	The numerical score corresponds approximately to a US school grade level. FKGL ≤6 was used as the adopted patient-facing benchmark [13,16]
Gunning-Fog Index (GFI)	0.4 × [(words/sentences) + 100 × (complex words/words)] [14]	Lower scores indicate easier reading	Estimates the years of formal education required for comprehension. A score below 8 was used as the patient-facing reference level [14,17].
DISCERN	Sum of Questions 1-15, each scored from 1 to 5; Question 16 is considered separately	Higher summed scores indicate higher information quality according to the DISCERN criteria	15-26: very poor; 27-38: poor; 39-50: fair; 51-62: good; 63-75: excellent. Question 16 provides a separate overall quality rating from 1 to 5 [18,19].

TABLE 6: Readability and information-quality assessment frameworks used in the study

Flesch Reading Ease Score [12]; FKGL: Flesch-Kincaid Grade Level [13]; GFI: Gunning-Fog Index [14];The DISCERN summed score used in this study comprises Questions 1-15; Question 16 was reported separately and was not included in the 75-point total [18,19].

Appendix 2

Source Platform	FRES	FKGL	GFI	DISCERN Score
NHS Tayside Patient Leaflet (https://www.nhstaysidecdn.scot.nhs.uk/NHSTaysideWeb/idcplg?IdcService=GET_SECURE_FILE&dDocName=PROD_355268&Rendition=web&RevisionSelectionMethod=LatestReleased&noSaveAs=1)	52.0	9.4	11.9	65.0
Mayo Clinic (https://www.mayoclinic.org/diseases-conditions/testicular-cancer-care/expert-answers/microlithiasis/faq-20058074)	19.6	14.8	18.2	59.0
Wikipedia (https://en.wikipedia.org/wiki/Testicular_microlithiasis)	8.6	17.6	19.9	49.0
Prof. Dr. Emin Özbek Clinic (https://dreminozbek.com/en/testicular-microlithiasis/)	4.4	17.8	19.9	52.0
Urologenzentrum (https://www.urologenzentrum.at/en/portfolio/testicular-microlithiasis/)	5.7	18.2	20.6	46.0
Staffordshire Urology Clinic (https://www.staffordshireurologyclinic.co.uk/journal-watch/follow-up-of-testicular-microlithiasis/)	26.4	14.0	15.7	56.0
Medicover Hospitals (https://www.medicoverhospitals.in/diseases/microlithiasis/)	14.8	16.0	19.1	39.0
Mentone General Practice (https://mentonegeneralpractice.com.au/what-is-testicular-microlithiasis/)	44.5	10.4	12.2	50.0
Happiest Health (https://www.happiesthealth.com/articles/mens-wellness/testicular-microlithiasis-when-calcium-deposits-form-in-the-testicles)	10.8	17.7	20.2	56.0
Health Jade (https://healthjade.net/testicular-microlithiasis/)	11.1	16.7	19.0	62.0
NatuVitro (https://natuvitro.com/en/blog/microlithiase-testiculaire)	19.2	15.4	18.6	31.0
The Kingsley Clinic (https://thekingsleyclinic.com/resources/testicular-microlithiasis-causes-symptoms-and-treatment-options/)	15.1	15.2	17.6	53.0
Biology Insights (https://biologyinsights.com/what-is-testicular-microlithiasis-and-is-it-serious/)	14.9	16.1	18.9	54.0
ChatGPT - GPT-5.2 Instant	40.0	11.9	12.7	52.0
Google Gemini - Gemini 3 Flash	47.9	9.6	11.6	53.0
Claude - Haiku 4.5	35.8	11.7	14.1	57.0

TABLE 7: Raw evaluation matrix detailing individual metric scores for all 16 evaluated resources

AI responses were generated using GPT-5.2 Instant, Gemini 3 Flash, and Claude Haiku 4.5 through the standard free consumer interfaces of ChatGPT, Google Gemini, and Claude, respectively, on 21 December 2025. Direct URLs for all included websites are listed above; complete unfiltered search logs, including excluded professional-facing websites, are available from the corresponding author upon reasonable request.

FRES: Flesch Reading Ease Score; FKGL: Flesch-Kincaid Grade Level; GFI: Gunning-Fog Index

Additional Information

Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

Concept and design: Huseyin Turacli, Faiz A. Shaikh, Jas Kalsi

Acquisition, analysis, or interpretation of data: Huseyin Turacli, Emane Ouadam, Samuel Pathmanathan

Drafting of the manuscript: Huseyin Turacli, Emane Ouadam

Critical review of the manuscript for important intellectual content: Huseyin Turacli, Samuel Pathmanathan, Faiz A. Shaikh, Jas Kalsi

Supervision: Jas Kalsi

Disclosures

Human subjects: All authors have confirmed that this study did not involve human participants or tissue.
Animal subjects: All authors have confirmed that this study did not involve animal subjects or tissue.
Conflicts of interest: In compliance with the ICMJE uniform disclosure form, all authors declare the following: **Payment/services info:** All authors have declared that no financial support was received from

any organization for the submitted work. **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work. **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

References

1. Balawender K, Orkisz S, Wisz P: Testicular microlithiasis: what urologists should know. A review of the current literature. Cent European J Urol. 2018, 71:310-4. 10.5173/ceju.2018.1728
2. Shanmugasundaram R, Singh JC, Kekre NS: Testicular microlithiasis: is there an agreed protocol? . Indian J Urol. 2007, 23:234-9. 10.4103/0970-1591.33442
3. van Casteren NJ, Looijenga LH, Dohle GR: Testicular microlithiasis and carcinoma in situ overview and proposed clinical guideline. Int J Androl. 2009, 32:279-87. 10.1111/j.1365-2605.2008.00937.x
4. Berger A, Brabrand K: Testicular microlithiasis--a possibly premalignant condition. Report of five cases and a review of the literature. Acta Radiol. 1998, 39:583-6. 10.1080/02841859809172231
5. Aoun F, Slaoui A, Naoum E, et al.: Testicular microlithiasis: systematic review and clinical guidelines . Prog Urol. 2019, 29:465-73. 10.1016/j.purol.2019.07.001
6. Richenberg J, Belfield J, Ramchandani P, et al.: Testicular microlithiasis imaging and follow-up: guidelines of the ESUR scrotal imaging subcommittee. Eur Radiol. 2015, 25:323-30. 10.1007/s00330-014-3437-x
7. Pourbagher MA, Kilinc F, Guvel S, Pourbagher A, Egilmez T, Ozkardes H: Follow-up of testicular microlithiasis for subsequent testicular cancer development. Urol Int. 2005, 74:108-12; discussion 113. 10.1159/000083279
8. Wang Z, Hu Y, Huang B, Zheng G, Li B, Liu Z: Is there a relationship between online health information seeking and health anxiety? A systematic review and meta-analysis. Health Commun. 2024, 39:2524-38. 10.1080/10410236.2023.2275921
9. Health Education England: Health literacy 'how to' guide . (2020). Accessed: October 1, 2026: https://library.nhs.uk/wp-content/uploads/sites/4/2020/08/Health-literacy-how-to-guide.pdf.
10. Ofcom: The era of answer engines . (2025). Accessed: October 1, 2026: https://www.ofcom.org.uk/siteassets/resources/documents/research-and-data/online-research/other/the-era-of-answer-eng....
11. Aydin S, Karabacak M, Vlachos V, Margetis K: Large language models in patient education: a scoping review of applications in medicine. Front Med (Lausanne). 2024, 11:1477898. 10.3389/fmed.2024.1477898
12. FL R: A new readability yardstick. J Appl Psychol. 1948, 32:221-33. 10.1037/h0057532
13. Kincaid JP, Fishburne RP Jr, Rogers RL, Chissom BS: Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel. Chief of Naval Technical Training, Millington, TN; 1975. 10.21236/ADA006655
14. Gunning R: The Technique of Clear Writing . McGraw-Hill, New York; 1952.
15. ReadabilityFormulas.com: Readability Scoring System. (2026). Accessed: October 1, 2026: https://readabilityformulas.com/readability-scoring-system.php.
16. Shannon C, Millar BC, Moore JE: Improving biomedical science literacy and patient-directed knowledge of tuberculosis (TB): a cross-sectional infodemiology study examining readability of patient-facing TB information. Br J Biomed Sci. 2024, 81:13566. 10.3389/bjbs.2024.13566
17. Watad A, Bragazzi NL, Brigo F, et al.: Readability of Wikipedia pages on autoimmune disorders: systematic quantitative assessment. J Med Internet Res. 2017, 19:e260. 10.2196/jmir.8225
18. Charnock D, Shepperd S, Needham G, Gann R: DISCERN: an instrument for judging the quality of written consumer health information on treatment choices. J Epidemiol Community Health. 1999, 53:105-11. 10.1136/jech.53.2.105
19. Som R, Gunawardana NP: Internet chemotherapy information is of good quality: assessment with the DISCERN tool. Br J Cancer. 2012, 107:403; author reply 404. 10.1038/bjc.2012.223
20. Colaco M, Svider PF, Agarwal N, Eloy JA, Jackson IM: Readability assessment of online urology patient education materials. J Urol. 2013, 189:1048-52. 10.1016/j.juro.2012.08.255
21. Eysenbach G, Köhler C: How do consumers search for and appraise health information on the world wide web? Qualitative study using focus groups, usability tests, and in-depth interviews. BMJ. 2002, 324:573-7. 10.1136/bmj.324.7337.573
22. Will J, Gupta M, Zaretsky J, Dowlath A, Testa P, Feldman J: Enhancing the readability of online patient education materials using large language models: cross-sectional study. J Med Internet Res. 2025, 27:e69955. 10.2196/69955
23. Liu Y, Li H, Ouyang J, et al.: Evaluating large language models for preoperative patient education in superior capsular reconstruction: comparative study of Claude, GPT, and Gemini. JMIR Perioper Med. 2025, 8:e70047. 10.2196/70047
24. Badarudeen S, Sabharwal S: Assessing readability of patient education materials: current role in orthopaedics. Clin Orthop Relat Res. 2010, 468:2572-80. 10.1007/s11999-010-1380-y
25. Meskó B: Prompt engineering as an important emerging skill for medical professionals: tutorial . J Med Internet Res. 2023, 25:e50638. 10.2196/50638