Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Quality, reliability, transparency, and disease-specific content of erectile dysfunction videos on four Chinese platforms: a cross-sectional content analysis

Jiangnan Xu
Affiliated Hospital of Nanjing University Medical School
Yuhui Zhang
Children's Hospital of Soochow University
Huixing Pan
Affiliated Hospital of Nanjing University Medical School
Zhenyu Xu
ksxzy2012@163.com

Kunshan Chinese Medicine Hospital, Nanjing University of Chinese Medicine

Article

Keywords: erectile dysfunction, health information, social media, content analysis, patient education, information quality

Posted Date: November 19th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11507708/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background: Videos about erectile dysfunction (ED) are a growing source of patient information on Chinese short-video platforms, yet their reliability, source transparency, and disease-specific clinical content have not been jointly assessed.

Methods: On October 8, 2026, we retrieved the first 100 default-ranked videos for the Chinese term "□□□□□□" ("erectile dysfunction") on each of four platforms; after deduplication and eligibility screening, 348 videos were included. Two physicians independently rated each video using the Global Quality Score (GQS), the original DISCERN reliability items 1-8, adapted JAMA benchmarks, and a pre-specified ten-item ED checklist informed by current guidelines. Analyses used consensus scores, rank-based comparisons, within-platform correlations with multiplicity correction, and HC3-adjusted regression.

Results: Platform comparisons depended on the assessment domain: Bilibili had the highest median DISCERN-8 score (31), whereas TikTok had the highest median adapted JAMA score (4); all platform tests were significant ($P < 0.001$). Correct mention of first-line PDE5 inhibitor treatment, medication safety, and other treatment options occurred in only 17.8%, 19.5%, and 8.9% of videos. Under the study rubric, traditional-medicine videos had higher unadjusted odds of exaggerated efficacy claims (OR 12.49, 95% CI 5.66-27.58) and primary-treatment substitution (OR 108.00, 95% CI 13.17-885.84). Professional uploaders scored lower on unadjusted DISCERN-8 but higher after adjustment (β 1.69, 95% CI 0.60-2.78), consistent with compositional confounding. Only four of 96 video-characteristic–score correlations survived within-platform correction, all in TikTok.

Conclusions: Generic and disease-specific instruments identified domain-dependent platform differences in this sample. Treatment-related content was limited, and there was limited evidence of a consistent association between engagement and information quality.

Background

Erectile dysfunction (ED), the persistent inability to attain or maintain an erection sufficient for satisfactory sexual performance, is a common men's health concern [1, 2, 3, 4]. Its management spans lifestyle intervention, pharmacotherapy, psychosexual counseling, and device- or surgery-based options, and treatment choice carries specific safety considerations—most notably the cardiovascular precautions and nitrate interactions of phosphodiesterase type 5 (PDE5) inhibitors [5, 6, 7]. The information patients encounter before, or instead of, a clinical consultation is therefore directly relevant to care.

Health information seeking has shifted decisively toward social media, which now functions as a routine source of medical content for the public [8, 9, 10]. The shift is particularly pronounced in China, where short-video applications had approximately 1.04 billion users as of December 2024 [11]. This scale creates opportunity and risk in parallel: short videos can disseminate accessible education, yet a substantial share of health-related social media content contains misinformation [12, 13, 14, 15]. For a stigmatized condition such as ED—where embarrassment may delay or replace clinical care-seeking—platform content can become a first, and sometimes only, information source [16, 17, 18].

Evaluating online health information has a two-decade methodological tradition, extended to video platforms through systematic reviews of YouTube content and, more recently, to short-video applications such as TikTok [19, 20, 21, 22, 23]. On Chinese short-video platforms, assessments of systemic lupus erythematosus, myelodysplastic syndromes, and lumbar spinal stenosis illustrate complementary designs: platform and uploader comparisons, reliability assessed alongside understandability and actionability, and separate evaluation of treatment recommendations [24, 25, 26]. Liver cancer video assessments have also compared quality and reliability across Chinese platforms [27]. For ED specifically, YouTube videos were predominantly of poor quality, and an analysis of ED treatment videos on YouTube and international TikTok found limited reliability and guideline-consistent content [28, 29]. Three gaps remain. First, ED content has not been jointly assessed across the mainland Chinese platforms on which most Chinese patients actually encounter it. Second, most video assessments apply generic instruments alone, which cannot establish whether disease-specific clinical content—first-line therapy, medication safety, or misleading claims—is present. Third, engagement metrics are routinely reported but rarely tested against quality under multiplicity correction and covariate adjustment.

We therefore evaluated ED-related videos from TikTok (mainland Douyin), Bilibili, Xiaohongshu, and Kwai (mainland Kuaishou). The design has three distinctive features. First, the four platforms were sampled simultaneously under a uniform top-100 default-ranking frame, permitting direct platform comparison within a single sampling day. Second, generic instruments—the Global Quality Score (GQS) [30], the original DISCERN reliability items 1–8 [31], and adapted JAMA benchmarks [32]—were complemented by a prespecified ten-item ED checklist informed by the EAU 2026 guideline and the Chinese 2022 andrology guideline, so that presentation quality, reliability, source transparency, and disease-specific clinical content are assessed as separate domains [33, 34]. Third, crude and adjusted associations are reported together, and engagement–quality correlations are tested within platforms under explicit multiplicity correction. We compared platforms and uploader groups, quantified selected clinical content and prespecified misleading-claim categories, and explored whether popularity tracks information quality.

Methods

2.1 Design and sampling

Platform labels follow the convention used by Zhao et al. [26]: TikTok, Bilibili, Xiaohongshu, and Kwai. Here, TikTok denotes the mainland Chinese Douyin platform and Kwai denotes the mainland Chinese Kuaishou platform; their international applications were not sampled. Xiaohongshu is also called Rednote in Gao et al. [24]. These are reporting labels and do not imply that mainland and international services are interchangeable.

This cross-sectional content analysis sampled videos on October 8, 2026, between 14:00 and 20:00 (China Standard Time, UTC + 8) in Yancheng, Jiangsu, China. Searches were conducted with Microsoft Edge 155.0.4283.45 (64-bit) on Windows 11 in private browsing mode, using a newly registered, logged-in account on each platform. Each platform was searched with the single Chinese diagnostic term "□□□□□□" ("erectile dysfunction"). The first 100 videos displayed in the default comprehensive ranking were selected as candidates on each platform, without additional search filters, yielding 400 candidate videos. The target of 100 candidates per platform was informed by Zhao et al. [26]. Duplicate content was removed before eligibility screening, including duplicate videos with different URLs. Excluded videos were not replaced by lower-ranked results. Two attending physicians specializing in andrology independently screened the videos. Disagreements were discussed with a chief physician specializing in andrology until consensus was reached. No minimum or maximum video duration was imposed.

Of the 400 candidate videos, 8 duplicates were removed, leaving 392 for eligibility screening. A further 44 videos were excluded: 39 were unrelated to ED and 5 had no Chinese speech. The final sample comprised 348 videos: 88 from TikTok, 92 from Bilibili, 86 from Xiaohongshu, and 82 from Kwai (Fig. 1).

Videos were eligible if they primarily discussed ED, contained Chinese speech, and could be accessed and played using the study accounts. Videos were excluded for duplicate content, lack of relevance to ED, or absence of Chinese speech. Videos without Chinese speech were excluded even when Chinese subtitles were available, as the study focused on spoken content for Chinese-speaking users. Videos containing product promotion were retained when they also provided substantive ED information.

2.2 Data collection and classification

Extracted characteristics included video duration, upload date, uploader category, content theme, and platform-displayed counts of likes, favorites, shares, views, and comments. Engagement counts were recorded during the same 14:00–20:00 search window. Views were the platform-displayed playback counts at collection. Because counting conventions were not independently standardized across platforms, views were treated as platform-native measures. Days online were calculated from the fixed search date, with a minimum of one day. Daily engagement was defined as the sum of likes, favorites, shares, and comments divided by days online.

Uploader categories were medical professionals, health communicators, patients or other individuals, official medical institutions, and commercial organizations. The professional group combined medical professionals and official institutions. Uploader identity was assessed using profile pages and platform certification information. The two attending physicians independently classified content themes, with disagreements resolved through discussion with the chief physician. Each video received one

dominant content theme: awareness and diagnosis, conventional treatment, traditional Chinese medicine, lifestyle and psychology, or experience sharing and care seeking.

2.3 Assessment instruments and rating procedure

The Global Quality Score (GQS, 1–5) assessed overall structure and usefulness [30]. We used the eight reliability items from the original DISCERN instrument, each scored 1–5, yielding a total of 8–40. These items concerned clarity of aims, achievement of aims, relevance, sources, currency, balance, additional information sources, and uncertainty [31]. Four adapted JAMA benchmarks assessed authorship, attribution, disclosure, and currency (0–4) [32]. Ratings used video information and uploader profile pages.

The study-specific Erectile Dysfunction Guideline-Consistency Score (ED-GLS) contained ten binary items informed by the EAU 2026 Guidelines on Sexual and Reproductive Health and the Chinese 2022 guideline [33, 34]. The items and scoring rules were fixed before formal rating and remained unchanged. Items 1–8 assessed correct mention of selected ED information; absent, incorrect, or misleading content received zero. Items 9 and 10 recorded the absence of exaggerated efficacy claims and primary-treatment substitution by supplements, dietary therapy, folk remedies, or wellness devices under the study rubric. Videos not addressing treatment received one point on these items if neither specified claim occurred. The total therefore combines content coverage with absence of specified claims and is not a validated measure of comprehensive clinical accuracy. The operational criteria are provided in Additional file 1.

The same two attending physicians independently rated all 348 videos. Before formal rating, they completed calibration using 20 separately selected videos, five per platform. These training videos were outside the 400 candidate records and were excluded from the analysis sample, avoiding repeated assessment of training videos in the main study. Content assessment used exported videos without engagement metrics. Profile information was also consulted for JAMA ratings; engagement and follower counts were not visible to the raters during this assessment. Each rater remained unaware of the other rater's scores until independent rating was complete. Disagreements were discussed with the chief physician until a final consensus score was reached. Inter-rater reliability used the two pre-consensus total scores; all other analyses used final consensus scores.

2.4 Statistical analysis

Continuous variables were summarized by medians and interquartile ranges, with distribution-free median confidence intervals from binomial order statistics. Categorical variables were reported as counts and percentages. Reliability was assessed using a two-way random-effects, absolute-agreement, single-rating intraclass correlation coefficient, ICC [1, 2], with F-based confidence intervals, and quadratic-weighted Cohen kappa with 2,000 paired-video bootstrap replicates.

GQS, DISCERN-8, and JAMA were compared across platforms using tie-corrected Kruskal–Wallis tests, eta-squared H effect sizes, and six Dunn comparisons per outcome with Bonferroni correction. Professional-group comparisons used Mann–Whitney tests and rank-based effect sizes; Holm-adjusted results were added for this three-outcome family. ED total scores were described but not tested across the mixed-theme platform samples. Categorical comparisons used Pearson chi-square tests when expected counts were adequate, Fisher exact tests for sparse 2×2 tables, and conditional Monte Carlo Fisher–Freeman–Halton tests for larger sparse tables (99,999 draws). Cramer V was derived from the Pearson statistic.

Within each platform, Spearman correlations related six characteristics to four score totals; Benjamini–Hochberg correction covered 24 tests per platform. ED-total correlations were exploratory because content composition can influence the total. Separate correction families covered the ten checklist-item platform tests and three exploratory interaction tests. Checklist proportions received Wilson confidence intervals. For items 9 and 10, unadjusted odds ratios compared failure in traditional-medicine versus other videos, with Woolf confidence intervals and two-sided Fisher P values.

Multivariable ordinary least squares models for GQS, DISCERN-8, and JAMA included platform, professional status, log10(duration), and log10(daily engagement + 1). HC3 standard errors and t-based intervals were used. Sensitivity analyses used uploader-proxy clustered uncertainty, added theme and online age, and removed the JAMA authorship item. The proxy cluster was platform plus uploader display name; it is not a verified permanent account identifier. Exploratory theme-adjusted models included platform-by-engagement interactions, evaluated by joint three-degree-of-freedom Wald tests with clustered covariance.

Proportional-odds models were also fitted for GQS and JAMA, but cumulative-logit diagnostics showed sparse extreme thresholds or separation; these models were therefore not treated as a validation of the linear-model findings. No engagement outliers were removed. Software versions, exact results, analysis amendments, and reproducible code are described in Additional file 2, with supplementary results in Additional file 3. All analyses were conducted in R version 4.5.2 (R Foundation for Statistical Computing, Vienna, Austria) [35].

OpenAI Codex was used to assist with analysis-code development and manuscript editing. The authors reviewed the generated material and take responsibility for the analyses and the final manuscript.

Ethics approval

and compliance. The study was approved by the Ethics Committee of Yancheng First People's Hospital on September 26, 2026 (approval no. 2026-J-029). All methods were performed in accordance with the relevant guidelines and regulations. The committee waived the requirement for informed consent. The study analyzed publicly available platform content accessed through logged-in accounts, without interaction with platform users. Video URLs and uploader display names were collected for source verification; these direct identifiers are excluded from the shared analysis dataset.

Results

3.1 Sample composition and rating reliability

The final dataset included 245 medical-professional videos, 68 health-communicator videos, 31 patient/individual videos, one official-institution video, and three commercial-organization videos. Thus, 246 videos (70.7%) belonged to the professional group. Awareness and diagnosis accounted for 165 videos, conventional treatment for 67, traditional Chinese medicine for 35, lifestyle and psychology for 56, and experience/care seeking for 25. Traditional-medicine content was particularly frequent within the Kwai sample (24/82, 29.3%), compared with 2/88, 2/92, and 7/86 in the other platforms (Table 1). Duration ranged from 5 to 2,604 seconds; 40 videos were shorter than 30 seconds and were retained under the unrestricted-duration criterion.

Table 1
Characteristics of the included videos by platform

Characteristic	Total (*n* = 348)	TikTok (*n* = 88)	Bilibili (*n* = 92)	Xiaohongshu (*n* = 86)	Kwai (*n* = 82)	Test	Statistic	*P*	Effect
Likes	289 (45, 1715.75)	321.5 (104.5, 1560.75)	175.5 (23, 811.5)	34 (6.25, 120)	2930 (1260.25, 8942.5)	KW	154.99	< 0.001	η^2H = 0.442
Favorites	236.5 (28, 1503)	112.5 (33.75, 490.25)	303.5 (41.25, 1092.25)	21.5 (4, 58.75)	3270.5 (1435, 9097.25)	KW	175.73	< 0.001	η^2H = 0.502
Shares	112 (18, 1631.75)	84 (20.75, 471.75)	52 (6, 212)	30 (5, 111)	4944.5 (1963.5, 12501.2)	KW	158.68	< 0.001	η^2H = 0.453
Views	22438.5 (3959.25, 157541)	23722.5 (7170, 90759.2)	12491 (2115.5, 76805.2)	2978.5 (912, 6301)	320500 (110500, 1,080,750)	KW	152.91	< 0.001	η^2H = 0.436
Comments	27 (3, 224.5)	23.5 (4, 86)	18 (4, 169)	1 (0, 8)	355.5 (150.25, 1110.25)	KW	153.58	< 0.001	η^2H = 0.438
Duration (s)	66 (40, 149.25)	52 (36.75, 78)	429.5 (163.25, 917.25)	53 (36.25, 80)	47 (36, 65.75)	KW	147.56	< 0.001	η^2H = 0.420
Days online	397.5 (144.25, 666)	631 (399.25, 699)	330.5 (86.75, 896.25)	216.5 (40, 389)	444 (203, 1183.75)	KW	61.21	< 0.001	η^2H = 0.169
Daily engagement	3.09 (0.60, 25.36)	1.42 (0.59, 6.14)	3.94 (0.85, 20.90)	0.58 (0.26, 2.23)	33.12 (12.09, 126.12)	KW	110.95	< 0.001	η^2H = 0.314
Medical professional	245/348 (70.4%)	84/88 (95.5%)	28/92 (30.4%)	78/86 (90.7%)	55/82 (67.1%)	MC Fisher	126.81	< 0.001	*V* = 0.349
Health communicator	68/348 (19.5%)	4/88 (4.5%)	43/92 (46.7%)	6/86 (7.0%)	15/82 (18.3%)	—	—	—	—
Patient / individual	31/348 (8.9%)	0/88 (0.0%)	20/92 (21.7%)	0/86 (0.0%)	11/82 (13.4%)	—	—	—	—
Official medical institution	1/348 (0.3%)	0/88 (0.0%)	1/92 (1.1%)	0/86 (0.0%)	0/82 (0.0%)	—	—	—	—
Commercial organization	3/348 (0.9%)	0/88 (0.0%)	0/92 (0.0%)	2/86 (2.3%)	1/82 (1.2%)	—	—	—	—
Awareness and diagnosis	165/348 (47.4%)	60/88 (68.2%)	29/92 (31.5%)	40/86 (46.5%)	36/82 (43.9%)	Pearson	111.02	< 0.001	*V* = 0.326

Continuous variables: median (Q1, Q3). Counts: n/N (%). KW=Kruskal-Wallis. Categorical tests use Fisher-Freeman-Halton Monte Carlo when any expected count < 5 (99,999 samples, seed 42); statistic and V remain Pearson chi-square summaries. Views are reported platform-native counts confirmed by the investigator and are not standardized exposures. No duration restriction. TikTok = mainland Douyin; Kwai = mainland Kuaishou. International apps were not sampled. H is the KW statistic; categorical statistics are Pearson χ^2. η^2H, KW effect size; V, Cramér's V. Category tests refer to the entire uploader/theme distribution, not the first category alone. Daily engagement = (likes + favorites + shares + comments)/days online.

Characteristic	Total (*n* = 348)	TikTok (*n* = 88)	Bilibili (*n* = 92)	Xiaohongshu (*n* = 86)	Kwai (*n* = 82)	Test	Statistic	*P*	Effect
Conventional treatment	67/348 (19.3%)	21/88 (23.9%)	15/92 (16.3%)	21/86 (24.4%)	10/82 (12.2%)	—	—	—	—
Traditional Chinese medicine	35/348 (10.1%)	2/88 (2.3%)	2/92 (2.2%)	7/86 (8.1%)	24/82 (29.3%)	—	—	—	—
Lifestyle and psychology	56/348 (16.1%)	3/88 (3.4%)	27/92 (29.3%)	14/86 (16.3%)	12/82 (14.6%)	—	—	—	—
Experience and care seeking	25/348 (7.2%)	2/88 (2.3%)	19/92 (20.7%)	4/86 (4.7%)	0/82 (0.0%)	—	—	—	—

Continuous variables: median (Q1, Q3). Counts: n/N (%). KW=Kruskal-Wallis. Categorical tests use Fisher-Freeman-Halton Monte Carlo when any expected count < 5 (99,999 samples, seed 42); statistic and V remain Pearson chi-square summaries. Views are reported platform-native counts confirmed by the investigator and are not standardized exposures. No duration restriction. TikTok = mainland Douyin; Kwai = mainland Kuaishou. International apps were not sampled. H is the KW statistic; categorical statistics are Pearson χ^2. η^2H, KW effect size; V, Cramér's V. Category tests refer to the entire uploader/theme distribution, not the first category alone. Daily engagement = (likes + favorites + shares + comments)/days online.

All main outcomes, engagement variables, themes, and rater total scores were complete. The data contained 223 platform-plus-display-name clusters. ICC estimates were 0.884 (95% CI 0.859–0.905) for GQS, 0.980 (0.975–0.983) for DISCERN-8, 0.928 (0.912–0.942) for JAMA, and 0.959 (0.950–0.967) for ED-GLS (Additional file 3: Table S1).

3.2 Platform and uploader differences

Median GQS scores were 3, 4, 4, and 3 for TikTok, Bilibili, Xiaohongshu, and Kwai, respectively. Corresponding DISCERN-8 medians were 23, 31, 27, and 22, whereas JAMA medians were 4, 3, 3, and 3. The three platform comparisons were significant (all $P < 0.001$), with η^2 of 0.285, 0.371, and 0.202 (Table 2; Fig. 2). Thus, the platform with the highest reliability score was not the one with the highest median transparency score. ED total medians were 6, 5, 5, and 4, presented descriptively because platform theme composition differed.

Table 2
Quality and transparency scores by platform

Outcome	TikTok	Bilibili	Xiaohongshu	Kwai	*H*	*P*	η²H
GQS	3 (3, 4) [a] 95% CI 3–4	4 (4, 4) [b] 95% CI 4–4	4 (3, 4) [a] 95% CI 3–4	3 (3, 3) [c] 95% CI 3–3	101.20	< 0.001	0.285
DISCERN-8	23 (20, 29) [a] 95% CI 22–26	31 (29, 34) [b] 95% CI 30–32	27 (25, 29) [c] 95% CI 26–28	22 (21, 24) [d] 95% CI 21–22	130.54	< 0.001	0.371
JAMA	4 (3, 4) [b] 95% CI 3–4	3 (2, 4) [a] 95% CI 2–3	3 (3, 4) [c] 95% CI 3–3	3 (2, 3) [a] 95% CI 2–3	72.34	< 0.001	0.202
ED-GLS	6 (5, 7) 95% CI 5–6	5 (4, 6) 95% CI 5–5	5 (4, 5) 95% CI 4–5	4 (3, 5) 95% CI 4–4	Descriptive	Not tested	Not tested

Values: median (Q1, Q3); distribution-free 95% median CI from binomial order statistics. Shared letters indicate pairwise Dunn-Bonferroni $P \geq 0.05$, not equivalence. Six pairwise tests per outcome. ED-GLS total is descriptive only; its domain coverage depends on topic. No fixed six-point ceiling is assumed. TikTok = mainland Douyin; Kwai = mainland Kuaishou. International apps were not sampled. GQS, Global Quality Score; DISCERN-8, original DISCERN questions 1–8; JAMA, Journal of the American Medical Association benchmark criteria; ED-GLS, study-specific ED guideline-based score; CI, confidence interval.

Professional uploaders had higher JAMA scores than non-professional uploaders (median 3 versus 2; r = 0.463), but lower unadjusted DISCERN-8 scores (25 versus 28; r = 0.131; unadjusted P = 0.015, Holm-adjusted P = 0.030). GQS did not show clear evidence of a professional-group difference (P = 0.070). Joint platform-by-uploader summaries revealed substantial imbalance and several empty or very small cells (Table 3; Table 4; Fig. 3), limiting interpretation of isolated institution or commercial categories.

Table 3
Quality scores by professional uploader status

Outcome	Professional (*n* = 246)	Non-professional (*n* = 102)	*U*	*Z*	*P*	*r*	Rank-biserial *r*
GQS	3 (3, 4)	4 (3, 4)	11131.5	-1.809	0.070	0.097	-0.113
DISCERN-8	25 (22, 29)	28 (23, 31)	10467	-2.437	0.015	0.131	-0.166
JAMA	3 (3, 4)	2 (1, 3)	19572	8.644	< 0.001	0.463	0.560

Professional = medical professionals plus official medical institutions (codes 1 or 4). Mann-Whitney U with tie and continuity correction; r=|Z|/sqrt(N). Rank-biserial correlation is positive for higher professional scores. Unadjusted professional differences are confounded by platform composition (see Discussion). ED total is not compared across mixed themes. P values are unadjusted exploratory comparisons; Holm-adjusted values are in Additional file 3: Table S4. GQS, Global Quality Score; DISCERN-8, original DISCERN questions 1–8; JAMA, Journal of the American Medical Association benchmark criteria; ED-GLS, study-specific ED guideline-based score; CI, confidence interval.

Table 4
Exploratory comparisons across five uploader categories

Uploader category	*n*	GQS	DISCERN-8	JAMA
Medical professional	245	3 (3, 4)	25 (22, 29)	3 (3, 4)
Health communicator	68	4 (3, 4)	29 (25.75, 31.25)	2 (1, 3)
Patient / individual	31	3 (3, 4)	25 (22.5, 29.5)	2 (1, 2.5)
Official medical institution	1	3 (3, 3)	28 (28, 28)	1 (1, 1)
Commercial organization	3	3 (2.5, 3)	22 (20, 22.5)	2 (1.5, 2.5)
Kruskal–Wallis *H*	—	19.057	16.410	82.703
P	—	< 0.001	0.003	< 0.001
η^2H	—	0.044	0.036	0.229

Median (Q1, Q3). Institution n = 1 and commercial n = 3 preclude stable isolated estimates. KW is exploratory; no sparse-category pairwise claim. Full unrounded results in Additional file 3: Table S4b. GQS, Global Quality Score; DISCERN-8, original DISCERN questions 1–8; JAMA, Journal of the American Medical Association benchmark criteria; ED-GLS, study-specific ED guideline-based score; CI, confidence interval.

3.3 Disease-specific coverage and misleading claims

Correct mention of ED definition, etiology or risk factors, and lifestyle intervention occurred in 221 (63.5%), 249 (71.6%), and 197 (56.6%) videos. First-line PDE5 inhibitor treatment was mentioned correctly in 62 (17.8%), medication safety in 68 (19.5%), psychological or partner factors in 70 (20.1%), professional care seeking in 188 (54.0%), and other treatment options in 31 (8.9%). These percentages describe the sampled videos, rather than the accuracy of all statements within each video (Table 5; Fig. 4; Additional file 4: Figure S1).

Table 5
ED-GLS domain coverage and absence of misleading claims

Item	Total	TikTok	Bilibili	Xiaohongshu	Kwai	Test	χ^2	*P*	*V*	Q (10 items)
Q1. Definition	221/348 (63.5%)	84/88 (95.5%)	68/92 (73.9%)	41/86 (47.7%)	28/82 (34.1%)	Pearson	82.86	< 0.001	0.488	< 0.001
Q2. Etiology / risk factors	249/348 (71.6%)	76/88 (86.4%)	66/92 (71.7%)	63/86 (73.3%)	44/82 (53.7%)	Pearson	22.51	< 0.001	0.254	< 0.001
Q3. Lifestyle intervention	197/348 (56.6%)	43/88 (48.9%)	63/92 (68.5%)	45/86 (52.3%)	46/82 (56.1%)	Pearson	8.08	0.044	0.152	0.063
Q4. PDE5 inhibitors as first-line treatment	62/348 (17.8%)	18/88 (20.5%)	19/92 (20.7%)	15/86 (17.4%)	10/82 (12.2%)	Pearson	2.70	0.440	0.088	0.499
Q5. Medication safety	68/348 (19.5%)	22/88 (25.0%)	16/92 (17.4%)	17/86 (19.8%)	13/82 (15.9%)	Pearson	2.65	0.449	0.087	0.499
Q6. Psychological / partner factors	70/348 (20.1%)	22/88 (25.0%)	24/92 (26.1%)	17/86 (19.8%)	7/82 (8.5%)	Pearson	10.20	0.017	0.171	0.028
Q7. Seek professional care	188/348 (54.0%)	72/88 (81.8%)	44/92 (47.8%)	37/86 (43.0%)	35/82 (42.7%)	Pearson	37.23	< 0.001	0.327	< 0.001
Q8. Other treatment options	31/348 (8.9%)	8/88 (9.1%)	5/92 (5.4%)	9/86 (10.5%)	9/82 (11.0%)	Pearson	2.06	0.560	0.077	0.560
Q9. No exaggerated efficacy claims	309/348 (88.8%)	88/88 (100.0%)	74/92 (80.4%)	82/86 (95.3%)	65/82 (79.3%)	Pearson	28.76	< 0.001	0.287	< 0.001
Q10. No supplement / folk-remedy substitution	338/348 (97.1%)	88/88 (100.0%)	91/92 (98.9%)	85/86 (98.8%)	74/82 (90.2%)	MC Fisher	18.47	< 0.001	0.230	0.001

Q1–Q8: correct explicit mention. Q9–Q10: absence of a specified claim under the study rubric; these are not content-coverage measures. Pearson statistics and Cramer V shown even where conditional Monte Carlo Fisher supplies *P*. BH across 10 item tests. Non-significance does not demonstrate equal coverage. TikTok = mainland Douyin; Kwai = mainland Kuaishou. International apps were not sampled. ED-GLS, study-specific ED guideline-based score; PDE5, phosphodiesterase type 5; BH, Benjamini–Hochberg; MC, Monte Carlo; V, Cramér's V. Values are n/N (%). Equivalent wording was accepted for Q1; a stated three-month threshold was not required. Q8 required correct presentation of at least one eligible modality.

Thirty-nine videos (11.2%) contained a prespecified exaggerated efficacy claim and ten (2.9%) recommended a supplement, dietary therapy, folk remedy, or wellness device as the primary ED treatment under the study rubric. Exaggerated claims occurred in 17/35 traditional-medicine videos versus 22/313 other videos (OR 12.49, 95% CI 5.66–27.58). Primary-treatment substitution occurred in 9/35 versus 1/313 (OR 108.00, 95% CI 13.17–885.84); both Fisher tests yielded P < 0.001 (Table 6; Table 7).

Table 6
Misleading-claim failures across content themes

Content theme	*n*	Q9 failures, *n*/N (%)	Q10 failures, *n*/N (%)
Awareness and diagnosis	165	1/165 (0.6%)	0/165 (0.0%)
Conventional treatment	67	3/67 (4.5%)	0/67 (0.0%)
Traditional Chinese medicine	35	17/35 (48.6%)	9/35 (25.7%)
Lifestyle and psychology	56	10/56 (17.9%)	0/56 (0.0%)
Experience and care seeking	25	8/25 (32.0%)	1/25 (4.0%)

Event = item score 0. Sparse theme-by-failure comparisons use conditional Monte Carlo Fisher; Pearson statistic and V describe association. Q9: chi-square = 84.14, df = 4, $P < 0.001$, V = 0.492. Q10: chi-square = 74.06, df = 4, $P < 0.001$, V = 0.461.

Table 7
Misleading-claim failures in traditional-medicine versus other videos

Item	TCM (*n* = 35)	Other themes (*n* = 313)	OR (95% CI)	Fisher *P*
Q9	17/35 (48.6%)	22/313 (7.0%)	12.49 (5.66, 27.58)	< 0.001
Q10	9/35 (25.7%)	1/313 (0.3%)	108.00 (13.17, 885.84)	< 0.001

Event = item score 0. Unadjusted OR; Woolf log-normal 95% CI; two-sided Fisher exact *P*. Q10 has one event in the reference group, giving substantial uncertainty. This is a content-category association, not a comparative clinical-efficacy conclusion. TCM, traditional Chinese medicine; OR, odds ratio; CI, confidence interval. The substitution item overlaps conceptually with the traditional-remedy theme; ORs do not estimate risk ratios or causal effects.

3.4 Engagement and adjusted associations

Four of the 96 correlations retained significance after within-platform correction, all in TikTok: favorites with GQS (ρ = 0.328), favorites with DISCERN-8 (ρ = 0.304), shares with GQS (ρ = 0.338), and views with ED total (ρ = −0.375). No correlation survived the corresponding 24-test correction on the other three platforms (Table 8; Fig. 5).

Table 8
Platform-specific associations between video characteristics and quality

Platform	Video characteristic	GQS	DISCERN-8	JAMA	ED-GLS
TikTok	Likes	0.270*	0.240*	0.011	-0.087
TikTok	Favorites	0.328**†	0.304**†	0.017	-0.039
TikTok	Shares	0.338**†	0.255*	0.048	-0.092
TikTok	Views	0.195	0.121	0.008	-0.375***†
TikTok	Comments	0.101	0.110	-0.048	-0.068
TikTok	Duration (s)	0.032	-0.054	-0.069	-0.078
Bilibili	Likes	0.096	0.129	-0.001	-0.082
Bilibili	Favorites	0.110	0.149	-0.005	-0.137
Bilibili	Shares	0.148	0.191	0.097	-0.008
Bilibili	Views	0.055	0.069	0.041	-0.061
Bilibili	Comments	0.125	0.152	0.204	0.175
Bilibili	Duration (s)	0.205	0.262*	0.027	0.006
Xiaohongshu	Likes	0.071	0.138	0.059	0.036
Xiaohongshu	Favorites	-0.005	0.081	0.067	0.008
Xiaohongshu	Shares	0.093	0.147	0.093	-0.030
Xiaohongshu	Views	0.176	0.140	0.155	0.130
Xiaohongshu	Comments	0.235*	0.108	-0.222*	-0.077
Xiaohongshu	Duration (s)	0.042	-0.034	0.034	0.001
Kwai	Likes	-0.037	0.002	-0.069	-0.265*
Kwai	Favorites	-0.102	-0.099	-0.081	-0.171
Kwai	Shares	-0.093	0.020	0.071	-0.224*
Kwai	Views	-0.075	-0.075	-0.029	-0.233*
Kwai	Comments	-0.034	-0.003	-0.056	-0.272*
Kwai	Duration (s)	0.059	0.140	0.334**	0.145

Spearman rho. * P < 0.05; ** P < 0.01; *** P < 0.001 (unadjusted); † BH q < 0.05 among 24 tests within platform. Confirmatory significance is not inferred from stars alone. ED-GLS correlations are exploratory. Additional file 3: Table S5 gives exact P, q, n and 95% CI. TikTok = mainland Douyin; Kwai = mainland Kuaishou. International apps were not sampled. GQS, Global Quality Score; DISCERN-8, original DISCERN questions 1–8; JAMA, Journal of the American Medical Association benchmark criteria; ED-GLS, study-specific ED guideline-based score; CI, confidence interval.

In HC3 models, Bilibili had higher GQS and DISCERN-8 scores than TikTok, with adjusted differences of 0.51 (95% CI 0.26–0.76) and 6.28 (4.20–8.36), respectively. Professional status was positively associated with DISCERN-8 after adjustment (β 1.69, 0.60–2.78), reversing the unadjusted comparison, and with JAMA (β 0.96, 0.70–1.23). The overall engagement coefficient showed no clear evidence of association with any of the three outcomes at P < 0.05. Model R^2 values were 0.296, 0.368, and 0.345, and all VIFs were at most 3.10 (Table 9; Fig. 6).

Table 9
Adjusted associations with quality and transparency

Predictor	VIF	GQS		DISCERN-8		JAMA	
		β (95% CI)	*P*	β (95% CI)	*P*	β (95% CI)	*P*
Bilibili vs TikTok	3.10	0.51 (0.26, 0.76)	< 0.001	6.28 (4.20, 8.36)	< 0.001	-0.48 (-0.84, -0.12)	0.009
Xiaohongshu vs TikTok	1.52	0.16 (-0.04, 0.35)	0.114	2.31 (0.82, 3.79)	0.002	-0.44 (-0.65, -0.23)	< 0.001
Kwai vs TikTok	1.97	-0.52 (-0.74, -0.31)	< 0.001	-2.53 (-4.23, -0.84)	0.003	-0.83 (-1.07, -0.59)	< 0.001
Professional uploader	1.48	0.13 (-0.03, 0.28)	0.112	1.69 (0.60, 2.78)	0.002	0.96 (0.70, 1.23)	< 0.001
log10(duration)	2.09	0.16 (-0.01, 0.33)	0.068	1.06 (-0.12, 2.24)	0.078	0.13 (-0.13, 0.39)	0.330
log10(daily engagement + 1)	1.48	0.07 (-0.03, 0.18)	0.173	0.68 (-0.07, 1.43)	0.076	-0.01 (-0.14, 0.13)	0.940

Unstandardized OLS coefficients, HC3 robust SE and t-based 95% CI. Reference: TikTok and non-professional uploaders. Duration coefficient is per tenfold increase in seconds; engagement coefficient per tenfold increase in daily engagement + 1. ED-GLS pooled models are not included because of topic dependence; within-theme and legacy reproductions are in Additional file 3: Tables S8 and S13. R^2: GQS 0.296; DISCERN-8 0.368; JAMA 0.345. TikTok = mainland Douyin; Kwai = mainland Kuaishou. International apps were not sampled. GQS, Global Quality Score; DISCERN-8, original DISCERN questions 1–8; JAMA, Journal of the American Medical Association benchmark criteria; ED-GLS, study-specific ED guideline-based score; CI, confidence interval. OLS, ordinary least squares; HC3, heteroscedasticity-consistent covariance estimator; SE, standard error; VIF, variance inflation factor.

Exploratory joint interaction tests supported platform heterogeneity in the engagement–DISCERN association (χ^2 = 12.23, 3 df; raw P = 0.0066; BH-adjusted P = 0.0199). The GQS and JAMA interactions did not meet the specified BH threshold (q = 0.0505 and 0.6052, respectively). The professional–DISCERN association remained positive under uploader-clustered uncertainty (β 1.69, 95% CI 0.63–2.75) and after adding theme and online age (β 1.50, 0.33–2.67). Removing the authorship item reduced the professional–JAMA association to 0.32 points (0.12–0.53). Full sensitivity coefficients, platform-specific slopes, and diagnostic limitations are reported in Additional file 3: Tables S8–S12; the analysis procedures are described in Additional file 2. Proportional-odds model results are reported descriptively in Table 10.

Table 10
Exploratory proportional-odds sensitivity models

Predictor	GQS		JAMA	
	OR (95% CI)	*P*	OR (95% CI)	*P*
Bilibili vs TikTok	5.45 (2.34, 12.70)	< 0.001	0.36 (0.16, 0.82)	0.015
Xiaohongshu vs TikTok	1.92 (1.06, 3.46)	0.031	0.33 (0.18, 0.59)	< 0.001
Kwai vs TikTok	0.15 (0.07, 0.35)	< 0.001	0.14 (0.07, 0.27)	< 0.001
Professional uploader	1.53 (0.86, 2.73)	0.150	8.72 (4.85, 15.69)	< 0.001
log10(duration)	1.63 (0.93, 2.86)	0.086	1.23 (0.70, 2.14)	0.468
log10(daily engagement + 1)	1.27 (0.91, 1.79)	0.163	1.01 (0.74, 1.37)	0.961

Odds of a higher rating; same predictors/references as Table 9. Model validity is conditional on proportional odds; see Additional file 3: Table S12. This table is not proof that OLS results are robust. TikTok = mainland Douyin; Kwai = mainland Kuaishou. International apps were not sampled. GQS, Global Quality Score; DISCERN-8, original DISCERN questions 1–8; JAMA, Journal of the American Medical Association benchmark criteria; ED-GLS, study-specific ED guideline-based score; CI, confidence interval. OR, odds ratio.

Discussion

In this four-platform sample of 348 ED videos, three findings stand out. First, platform comparisons were domain-dependent: Bilibili showed the highest median reliability (DISCERN-8 31), whereas TikTok showed the highest median source transparency (JAMA 4), and no platform led on all dimensions. Second, disease-specific treatment content was uncommon: correct mention of first-line PDE5 inhibitor therapy, medication safety, and other treatment options appeared in only 17.8%, 19.5%, and 8.9% of videos, and pre-specified misleading claims were concentrated in traditional-medicine videos. Third, there was limited evidence of a consistent association between engagement and quality: only four of 96 video-characteristic–score correlations survived within-platform correction, all in TikTok, and the overall adjusted engagement coefficient showed no clear association with any outcome. These findings support assessing individual information dimensions rather than assigning a single overall platform ranking.

The domain-dependent ranking complements prior Chinese short-video assessments that compared platforms using general quality and reliability instruments [24, 25, 26, 36]. Adjusted estimates showed that Bilibili scored higher than TikTok on GQS (0.51 points, 95% CI 0.26–0.76) and DISCERN-8 (6.28 points, 4.20–8.36), despite TikTok leading on the adapted JAMA benchmarks. One plausible explanation is structural: the JAMA items reward visible authorship and disclosure practices, which platform certification systems and account pages shape independently of content reliability. Uploader mix, audience preferences, and recommendation or moderation practices may contribute further, though these were not directly measured. Platform-native view counts also likely follow different counting conventions, so displayed views cannot establish proportional differences in audience reach; likewise, differences in disease, sampling, and instrument versions limit numerical comparison with prior work [24, 26].

Professional uploaders—70.7% of the sample—illustrate why crude and adjusted results must be read together. They had lower unadjusted DISCERN-8 scores than non-professional uploaders (median 25 vs. 28) but higher scores after adjustment for platform, duration, and engagement (β 1.69, 95% CI 0.60–2.78). This reversal is consistent with compositional differences and does not establish a causal benefit of professional status, and it remained positive under uploader-clustered uncertainty and after adding theme and online age. The adapted JAMA advantage (β 0.96, 0.70–1.23) partly shares a measurement basis with professional identity: removing the authorship item reduced the coefficient to 0.32 (0.12–0.53). Neither association establishes the accuracy of every clinical statement—a caution consistent with systematic reviews showing that visible credentials and content accuracy frequently diverge [19]. Prior ED video analyses likewise found professional-source content to be better on average, yet still of limited quality [28].

Our approach follows a small but growing line of work using disease-specific tools, as in Yao et al.'s assessment of dry eye disease treatment videos on Chinese short-video platforms [37]. Our ED checklist likewise separated general explanation from treatment-specific coverage. General topics were common—etiology or risk factors in 71.6% of videos, the ED definition in 63.5%, and lifestyle intervention in 56.6%—but treatment specifics were rare: first-line PDE5 inhibitor therapy in 17.8%, medication safety in 19.5%, and other treatment options in 8.9%. Prior studies in other disease domains have likewise argued for evaluating treatment recommendations independently [25, 38, 39]. Our finding of limited treatment coverage also echoes ED assessments on YouTube and international TikTok [28, 29]. Our rubric nevertheless measures selected coverage rather than the appropriateness of all advice: a zero on items 1–8 can represent omission or error, while a one on items 9–10 may simply indicate that a video did not discuss treatment. Narrowly focused videos need not cover an entire management pathway, and these proportions identify gaps in the sampled information without showing that viewers experienced clinical harm.

Prespecified misleading claims showed the sharpest content contrast. Exaggerated efficacy claims occurred in 17 of 35 traditional-medicine videos versus 22 of 313 other videos (OR 12.49, 95% CI 5.66–27.58), and primary-treatment substitution in 9 of 35 versus 1 of 313 (OR 108.00, 95% CI 13.17–885.84); the single reference-group event makes the latter estimate particularly imprecise. Traditional-medicine content was itself unevenly distributed, accounting for 29.3% of the Kwai sample but 8.1% or less elsewhere. These comparisons are sensitive to the rubric: traditional-remedy themes overlap conceptually with the substitution item, treatment-focused videos provide more opportunity for either claim, the same assessors performed theme classification and scoring, and the estimates are unadjusted. The associations describe the sampled content under the study definitions and do not establish that a theme independently causes misinformation or that all traditional or adjunctive treatments are inappropriate.

The engagement analyses provide limited evidence for any consistent relationship between popularity and information quality. Four correlations survived correction within TikTok—favorites with GQS (ρ = 0.328) and DISCERN-8 (ρ = 0.304), shares with GQS (ρ = 0.338), and views with the descriptive ED total (ρ = −0.375)—while none survived the corresponding correction on the other platforms; this does not establish absence of association, particularly for modest effects. The exploratory platform-by-engagement interaction was supported for DISCERN-8 (BH q = 0.0199) but not GQS or JAMA. A popularity–quality disconnect is consistent with two decades of web-based evidence that visibility is a poor proxy for accuracy [12, 40, 41]. Algorithmic exposures were not measured, and the findings identify no recommendation-system mechanism. These findings suggest that engagement metrics alone should not be used to judge information quality.

This study has several strengths. It provides a joint assessment of ED content across four mainland Chinese short-video platforms under a uniform sampling frame. The prespecified ten-item ED checklist complements generic instruments and identifies item-specific content gaps that generic scores do not describe. Two physicians rated all videos independently after calibration on a separate sample, with good-to-excellent agreement (ICC 0.884–0.980). Finally, reporting crude and adjusted associations side by side, together with multiplicity-controlled correlation analyses and clustered sensitivity models, makes the dependence of each conclusion on sample composition explicit.

Several limitations qualify these findings. First, the single-day sample of default-ranked results was not a probability sample of all videos or users; newly registered accounts and private browsing cannot eliminate location-, device-, or platform-level selection, and the single diagnostic search term with a Chinese-speech requirement may have omitted videos retrievable with colloquial terms or presented only through text. Second, the ED checklist and the JAMA adaptations have not undergone independent validation; coverage opportunities, the mixed meaning of the checklist total, and overlap between themes and claim definitions limit interpretation. Third, independent theme-classification records were not retained, so classification agreement could not be quantified. Fourth, primary HC3 regression addresses heteroskedasticity but not dependence among videos from the same uploader; uploader-proxy clustering was used in sensitivity and interaction models, whereas most other tests treated videos as independent, and display names are imperfect account identifiers. Fifth, sparse uploader categories and claim events limited detailed comparisons, and unstable cumulative-logit thresholds restricted the ordinal sensitivity analysis. Finally, engagement counts are platform-native cumulative measures; the analysis cannot establish temporal direction, individual exposure, treatment decisions, or clinical harm.

For practice, the content gaps suggest that clinicians should anticipate patients arriving with fragmentary or skewed ED information and ask what they have watched. For platforms, the results support clearer source attribution, visible credentialing, and content review targeted at exaggerated-efficacy and primary-treatment-substitution claims; the checklist items offer a starting template, subject to prospective validation. For research, the domain-dependence observed here argues against single-score platform comparisons, and prospective designs are needed to test whether moderation or educational interventions change patient behavior or outcomes.

Conclusions

In this sample of ED-related videos from four mainland Chinese short-video platforms, platform differences were domain-dependent: Bilibili had the highest median reliability score and TikTok the highest median source-transparency score. Treatment-specific clinical content (first-line therapy, medication safety, and other treatment options) was infrequently covered in the pooled sample, and prespecified misleading claims were more frequent in traditional-medicine videos under the study rubric. There was limited evidence of a consistent association between engagement and information quality. Generic instruments and a prespecified disease-specific checklist described complementary dimensions of the sampled content; these findings do not establish effects on patient behavior or clinical outcomes.

Abbreviations

BH
Benjamini–Hochberg
CI

confidence interval
DISCERN-8
original DISCERN reliability items 1–8
ED
erectile dysfunction
ED-GLS
Erectile Dysfunction Guideline-Consistency Score
GQS
Global Quality Score
HC3
heteroscedasticity-consistent covariance estimator
ICC
intraclass correlation coefficient
JAMA
Journal of the American Medical Association
OLS
ordinary least squares
OR
odds ratio
PDE5
phosphodiesterase type 5
SE
standard error
VIF
variance inflation factor.

Declarations

Ethics approval and consent to participate

The study was approved by the Ethics Committee of Yancheng First People's Hospital on September 26, 2026 (approval no. 2026-J-029). All methods were performed in accordance with the relevant guidelines and regulations. The committee waived the requirement for informed consent. The study analyzed publicly available platform content accessed through logged-in accounts, without interaction with platform users. Video URLs and uploader display names were collected for source verification; these direct identifiers are excluded from the shared analysis dataset.

Consent for publication

Not applicable.

Funding

This research received no funding.

Author Contribution

J.X. curated the data, conducted the investigation and formal analysis, and wrote the original draft. Y.Z. and H.P. curated the data and conducted the investigation. Z.X. supervised the study and reviewed and edited the manuscript. All authors read and approved the final manuscript.

Acknowledgements

Not applicable.

Data Availability

The analysis dataset and code supporting the findings of this study are available in the Open Science Framework (OSF) repository at https://doi.org/10.17605/OSF.IO/NAD3B. Direct identifiers, including video URLs, titles, uploader display names, and explicit upload dates, have been excluded from the shared dataset. Days online are retained for reproducibility. Removal of direct identifiers does not guarantee anonymity, as combinations of numerical variables may remain linkable to publicly available videos.

References

1. Bright, U. et al. Multi-ancestry investigation of the genomics of erectile dysfunction. *Nat. Commun.* **16** (1), 11602 (2025).
2. Zhang, X., Yang, B., Li, N. & Li, H. Prevalence and Risk Factors for Erectile Dysfunction in Chinese Adult Males. *J. Sex. Med.* **14** (10), 1201–1208 (2017).
3. Carlomagno, F. et al. Iron overload disorders in adults: a comprehensive review of gonadal function, reproductive, and sexual health. *Hum. Reprod. Update*. **32** (5), 663–690 (2026).
4. Amand, C. et al. A population-based survey of self-reported and IIEF-defined erectile dysfunction among adult men in the United States in 2021. *BMC public. health*. **25** (1), 4172 (2025).
5. Qaseem, A. et al. Hormonal testing and pharmacologic treatment of erectile dysfunction: a clinical practice guideline from the American College of Physicians. *Ann. Intern. Med.* **151** (9), 639–649 (2009).
6. Sansone, A. et al. Future perspectives for PDE5 inhibitors bridging the gap between cardiovascular health and psychological status. *Basic. Clin. Androl.* **35** (1), 3 (2025).
7. Li, J. et al. Safety profile and signal detection of tadalafil: a real-world analysis based on the Food and Drug adverse event reporting system. *Sex. Med.* **14** (2), qfag014 (2026).
8. Moorhead, S. A. et al. A new dimension of health care: systematic review of the uses, benefits, and limitations of social media for health communication. *J. Med. Internet. Res.* **15** (4), e85 (2013).
9. Smailhodzic, E., Hooijsma, W., Boonstra, A. & Langley, D. J. Social media use in healthcare: A systematic review of effects on patients and on their relationship with healthcare professionals. *BMC Health Serv. Res.* **16** (1), 442 (2016).
10. Zhao, Y. & Zhang, J. Consumer health information seeking in social media: a literature review. *Health Inf. Libr. J.* **34** (4), 268–283 (2017).
11. China Internet Network Information Center (CNNIC). The 55th Statistical Report on China's Internet Development. Beijing: CNNIC. (2025). Available at: https://www.cnnic.net.cn/NMediaFile/2025/0117/MAIN1737106895767 21DFTGKEAD.pdf (Accessed October 16, 2026).
12. Suarez-Lledo, V. & Alvarez-Galvez, J. Prevalence of Health Misinformation on Social Media: Systematic Review. *J. Med. Internet. Res.* **23** (1), e17187 (2021).
13. Shi, Y. & Mao, W. Quality and reliability of osteoporosis-related health information on short-video platforms: a multi-source cross-sectional study. *BMC public. health* ;**26**(1). (2026).
14. Zhu, Y., Cui, L., Zhou, T. & Xu, H. The quality and reliability of short videos about pulmonary hypertension on Bilibili and TikTok: a cross-sectional study. *BMC Cardiovasc. Disord.* ;**26**(1). (2026).
15. Zhang, X., Wen, D., Liang, J. & Lei, J. How the public uses social media wechat to obtain health information in china: a survey study. *BMC Med. Inf. Decis. Mak.* **17** (Suppl 2), 66 (2017).
16. Rodler, S. et al. Epidemiology and Treatment Barriers of Patients With Erectile Dysfunction Using an Online Prescription Platform: A Cross-Sectional Study. *Sex. Med.* **8** (3), 370–377 (2020).

17. Shabsigh, R., Perelman, M. A., Laumann, E. O. & Lockhart, D. C. Drivers and barriers to seeking treatment for erectile dysfunction: a comparison of six countries. *BJU Int.* **94** (7), 1055–1065 (2004).
18. Zhang, K., Yu, W., He, Z. J. & Jin, J. Help-seeking behavior for erectile dysfunction: a clinic-based survey in China. *Asian J. Androl.* **16** (1), 131–135 (2014).
19. Osman, W., Mohamed, F., Elhassan, M. & Shoufan, A. Is YouTube a reliable source of health-related information? A systematic review. *BMC Med. Educ.* **22** (1), 382 (2022).
20. Song, S. et al. Short-Video Apps as a Health Information Source for Chronic Obstructive Pulmonary Disease: Information Quality Assessment of TikTok Videos. *J. Med. Internet. Res.* **23** (12), e28318 (2021).
21. Luo, G., Liu, Y., Wu, X. & Guo, S. Credibility of health-related content on Chinese short-video platforms: A scoping review on accuracy, risks, and quality metrics. *Digit. health*. **12**, 20552076261428037 (2026).
22. Keelan, J., Pavri-Garcia, V., Tomlinson, G. & Wilson, K. YouTube as a source of information on immunization: a content analysis. *Jama* **298** (21), 2482–2484 (2007).
23. Kong, W., Song, S., Zhao, Y. C., Zhu, Q. & Sha, L. TikTok as a Health Information Source: Assessment of the Quality of Information in Diabetes-Related Videos. *J. Med. Internet. Res.* **23** (9), e30409 (2021).
24. Gao, L. et al. Quality and reliability assessment of systemic lupus erythematosus related health information on short video platforms. *Front. Digit. health*. **8**, 1875817 (2026).
25. Yang, R. et al. Quality and reliability of myelodysplastic syndromes-related health information on Chinese short-video platforms: a multi-platform cross-sectional analysis. *Front. Digit. health*. **8**, 1956383 (2026).
26. Zhao, W. et al. Stratified and combined analysis of the quality of lumbar spinal stenosis-related videos on major Chinese short video platforms. *Front. Digit. health*. **8**, 1769121 (2026).
27. Zheng, S. et al. Quality and Reliability of Liver Cancer-Related Short Chinese Videos on TikTok and Bilibili: Cross-Sectional Content Analysis Study. *J. Med. Internet. Res.* **25**, e47210 (2023).
28. Fode, M. et al. Quality of Information in YouTube Videos on Erectile Dysfunction. *Sex. Med.* **8** (3), 408–413 (2020).
29. Babar, M. et al. Cross-sectional and comparative analysis of videos on erectile dysfunction treatment on YouTube and TikTok. *Andrologia* **54** (5), e14392 (2022).
30. Bernard, A. et al. A systematic review of patient inflammatory bowel disease information resources on the World Wide Web. *Am. J. Gastroenterol.* **102** (9), 2070–2077 (2007).
31. 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. Commun. Health*. **53** (2), 105–111 (1999).
32. Silberg, W. M., Lundberg, G. D. & Musacchio, R. A. Assessing, controlling, and assuring the quality of medical information on the Internet: Caveant lector et viewor–Let the reader and viewer beware. *Jama* **277** (15), 1244–1245 (1997).
33. European Association of Urology. EAU Guidelines on Sexual and Reproductive Health. Chapter: Management of Erectile Dysfunction. (2026). edition https://uroweb.org/guidelines/sexual-and-reproductive-health/chapter/management-of-erectile-dysfunction
34. Erectile Dysfunction Diagnosis and Treatment Guideline Writing Group of the Andrology Branch of the Chinese Medical Association. Guidelines for the diagnosis and treatment of erectile dysfunction. *Natl. J. Androl.* **28** (8), 722–755 (2022). [Article in Chinese].
35. R Core Team. R: A Language and Environment for Statistical Computing. Vienna: R Foundation for Statistical Computing. (2025). Available at: https://www.R-project.org/
36. Zheng, X., Li, Q., Jin, L., Shi, K. & Deng, M. Quality, reliability, and dissemination of lung cancer information on short-video platforms in China: a cross-platform content analysis of TikTok, Kwai, and Rednote. *Front. public. health*. **13**, 1683561 (2025).
37. Yao, R., Ye, J., Zhu, Z. & Li, C. Insights into the quality and guideline consistency of dry eye disease treatment videos on Chinese short-video platforms. *Digit. health*. **12**, 20552076261458906 (2026).
38. Zhang, J. et al. Short video platforms as sources of health information about cervical cancer: A content and quality analysis. *PloS one*. **19** (3), e0300180 (2024).

39. Lei, Y., Liao, F., Li, X. & Zhu, Y. Quality and reliability evaluation of pancreatic cancer-related video content on social short video platforms: a cross-sectional study. *BMC public. health*. **25** (1), 1919 (2025).
40. Wang, Y., McKee, M., Torbica, A. & Stuckler, D. Systematic Literature Review on the Spread of Health-related Misinformation on Social Media. Social science & medicine 2019;240:112552. (1982).
41. Gabarron, E., Fernandez-Luque, L., Armayones, M. & Lau, A. Y. Identifying Measures Used for Assessing Quality of YouTube Videos with Patient Health Information: A Review of Current Literature. *Interact. J. Med. Res.* **2** (1), e6 (2013).

Figures

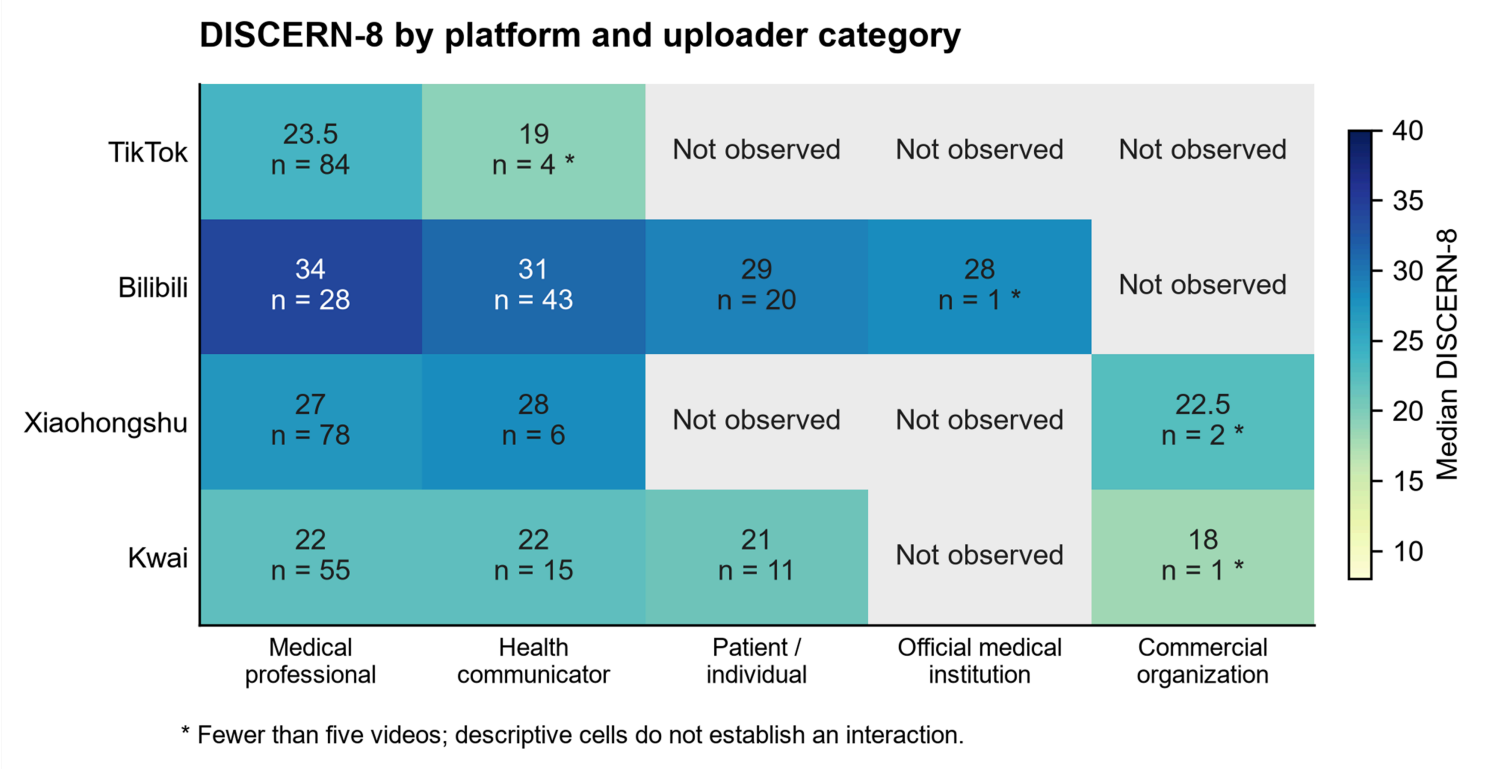


Figure 3

DISCERN reliability scores by platform and uploader category
Cells show the median DISCERN-8 score and sample size. Gray cells indicate unobserved combinations; asterisks identify fewer than five videos. Color uses a common 8–40 scale. The chart describes joint strata, not a fitted platform-by-uploader interaction. Official-institution and commercial categories are too small for stable isolated estimates.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile1ScoringProtocol.docx
- Additionalfile2CodeandReproducibility.docx