

Large language models differ in patient education quality across the robot-assisted radical prostatectomy care pathway

Qiang Wang

Affiliated Hospital of North Sichuan Medical College

Xiao Wang

Affiliated Hospital of North Sichuan Medical College

Xiaolu Jiang

Affiliated Hospital of North Sichuan Medical College

Sheng Xie

Affiliated Hospital of North Sichuan Medical College

Haotian Hou

Affiliated Hospital of North Sichuan Medical College

Hangxuan Ren

Affiliated Hospital of North Sichuan Medical College

Yaxin Luo

Affiliated Hospital of North Sichuan Medical College

Huiyu Gao

Affiliated Hospital of North Sichuan Medical College

Qianru Chang

North Sichuan Medical University

Tao Wu

alhawking@163.com

Affiliated Hospital of North Sichuan Medical College

Article

Keywords: large language models, robot-assisted radical prostatectomy, prostate cancer, patient education, readability, artificial intelligence

Posted Date: October 11th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10994677/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

Patients considering robot-assisted radical prostatectomy (RARP) require information that spans diagnosis, treatment selection, functional recovery, and recurrence surveillance. We compared the quality of English-language RARP patient education generated by ChatGPT, DeepSeek, and Doubao. Twenty-five questions covering screening, diagnosis, treatment decision-making, perioperative care, rehabilitation, and follow-up were submitted individually to each model, yielding 75 responses. Three blinded clinicians evaluated the responses using the Global Quality Score, modified DISCERN, Quality Analysis of Medical Artificial Intelligence, Ensuring Quality Information for Patients, and the Patient Education Materials Assessment Tool; six readability indices were also calculated. Overall Global Quality Scores did not differ significantly across models. ChatGPT achieved a higher EQIP score (median, 75.00%) than DeepSeek and Doubao (65.83% each) and greater PEMAT actionability (59.00% vs. 54.33% and 47.67%, respectively). DeepSeek generated comparatively easier-to-read text, although all three models exceeded commonly recommended sixth- to eighth-grade reading levels for public-facing health materials. Under the majority-rater criterion, no factual errors or outdated recommendations were identified, whereas key omissions occurred mainly in DeepSeek and Doubao responses. Inter-rater agreement was low, indicating substantial uncertainty in multidimensional human assessment. LLMs may support the preparation of clinician-reviewed RARP educational materials but should not replace individualized counseling by urologists.

Introduction

Robot-assisted radical prostatectomy (RARP) is an established curative treatment for localized prostate cancer and selected cases of locally advanced disease. Across the care pathway, patients must interpret information about prostate-specific antigen (PSA), magnetic resonance imaging (MRI), pathological grading, nerve-sparing surgery, pelvic lymph node dissection, urinary continence, sexual function, biochemical recurrence, and salvage therapy. Randomized and comparative evidence indicates that surgical approaches can differ in perioperative and functional outcomes, while continence recovery evolves over time [1,2]. Effective patient education therefore needs to describe potential treatment benefits together with the expected recovery trajectory, interindividual variability, and remaining uncertainty.

Management of localized prostate cancer is preference-sensitive, making informed shared decision-making particularly important. Decision aids in uro-oncology can improve patient knowledge and participation, although their effects on decisional conflict, regret, and the treatment ultimately selected are inconsistent [3,4]. Inadequate comprehension or overly definitive risk communication may distort treatment expectations and adversely affect shared decisions, perioperative engagement, and adherence to long-term surveillance.

Large language models (LLMs) provide rapid conversational responses and are increasingly being evaluated for medical question answering and patient education [5–7]. Studies in urology have reported

generally moderate-to-high information quality, but performance remains variable for source attribution, evidence transparency, actionability, and readability [8,9]. Previous prostate cancer investigations have commonly examined a single model, limited prompt sets, or radiotherapy-focused questions [10–13]. Direct, paired comparisons across the broader RARP care pathway remain limited.

Factual accuracy alone does not determine the quality of patient education. Reliability, completeness, understandability, actionability, and reading burden are related but distinct attributes. The Quality Analysis of Medical Artificial Intelligence (QAMAI), Ensuring Quality Information for Patients (EQIP), and Patient Education Materials Assessment Tool (PEMAT) evaluate complementary aspects of AI-generated health information [14–16]. Readability formulas, by contrast, estimate linguistic complexity rather than medical correctness or actual patient comprehension.

We therefore conducted a cross-sectional comparison of ChatGPT, DeepSeek, and Doubao using 25 patient-centered questions spanning prostate cancer screening, RARP treatment decisions, postoperative recovery, and recurrence management. English responses were reviewed blindly with multiple assessment instruments, with each question serving as the paired unit for comparisons among models. Exploratory analyses by stage of care were also performed to characterize where relative model strengths and limitations might differ.

Results

Data completeness and inter-rater agreement

All 75 planned responses were available for analysis, with 25 generated by each model, and no primary outcome data were missing. Agreement among raters was low across the evaluation instruments. Single-measure intraclass correlation coefficients (ICCs) ranged from 0.030 to 0.215, whereas ICCs based on the mean of the three raters ranged from 0.084 to 0.452 (Table 1). These estimates indicate substantial between-rater variability in judgments of multidimensional patient-education quality.

Table 1
Inter-rater agreement across assessment instruments

Measure	ICC(2,1)	95% CI	ICC(2,3)	95% CI
GQS	0.215	0.059 to 0.365	0.452	0.157 to 0.633
mDISCERN	0.030	-0.084 to 0.153	0.084	-0.302 to 0.351
QAMAI	0.047	-0.002 to 0.097	0.129	-0.007 to 0.243
EQIP	0.147	0.034 to 0.248	0.342	0.095 to 0.497
PEMAT-U	0.119	-0.000 to 0.237	0.288	-0.001 to 0.482
PEMAT-A	0.043	-0.027 to 0.122	0.119	-0.087 to 0.294

Overall quality, reliability and patient usefulness

No significant overall difference was observed for the Global Quality Score (GQS; chi-square = 4.965, $P = 0.084$, Kendall's $W = 0.099$). In contrast, overall model effects were significant for modified DISCERN (mDISCERN), QAMAI, and EQIP. ChatGPT had a median mDISCERN score of 3.33 and exceeded DeepSeek (3.00) after multiplicity correction ($P < 0.05$); Doubao also scored higher than DeepSeek ($P < 0.05$). Median QAMAI scores were 23.67 for ChatGPT, 23.33 for DeepSeek, and 22.67 for Doubao. The largest separation was observed for EQIP ($W = 0.769$), for which ChatGPT achieved a median of 75.00%, compared with 65.83% for both DeepSeek and Doubao (both $P < 0.05$). No significant difference in EQIP was observed between DeepSeek and Doubao.

PEMAT understandability showed no statistically significant overall model effect ($P = 0.053$). By contrast, PEMAT actionability differed across models (chi-square = 15.093, $P < 0.001$, $W = 0.302$). The median actionability score was 59.00% for ChatGPT, 54.33% for DeepSeek, and 47.67% for Doubao. Adjusted pairwise comparisons showed higher scores for ChatGPT than for DeepSeek and Doubao, and for DeepSeek than for Doubao (all $P < 0.05$) (Table 2; Figs. 2 and 3).

Table 2
Descriptive statistics and overall comparisons of primary outcomes across the three models

Measure	ChatGPT	DeepSeek	Doubao	chi-square	P value	Kendall W
GQS	3.67 (3.67, 4.00)	3.67 (3.33, 4.00)	3.67 (3.33, 3.67)	4.965	0.084	0.099
mDISCERN	3.33 (3.33, 3.67)	3.00 (2.67, 3.33)	3.33 (3.00, 3.33)	16.545	< 0.001	0.331
QAMAI	23.67 (23.00, 24.00)	23.33 (22.33, 23.67)	22.67 (21.33, 23.00)	18.258	< 0.001	0.365
EQIP	75.00 (74.17, 76.67)	65.83 (62.50, 69.17)	65.83 (63.33, 66.67)	38.449	< 0.001	0.769
PEMAT-U	73.00 (69.67, 77.33)	70.00 (64.00, 74.33)	68.00 (64.33, 70.33)	5.879	0.053	0.118
PEMAT-A	59.00 (55.33, 63.67)	54.33 (47.67, 59.00)	47.67 (38.00, 56.33)	15.093	< 0.001	0.302
ARI	16.48 (14.49, 17.40)	15.50 (13.94, 16.97)	17.19 (16.04, 18.90)	8.880	0.012	0.178
FRES	26.00 (24.00, 31.00)	32.00 (21.00, 37.00)	21.00 (15.00, 30.00)	18.330	< 0.001	0.367
GFI	15.80 (15.20, 16.30)	15.30 (13.80, 16.40)	16.50 (15.40, 17.20)	13.879	< 0.001	0.278
FKGL	14.90 (13.58, 15.86)	13.95 (12.31, 14.59)	15.55 (14.27, 16.55)	11.120	0.004	0.222
CLI	15.43 (14.70, 16.45)	14.99 (13.80, 16.81)	16.70 (15.87, 18.19)	16.080	< 0.001	0.322
SMOG	13.26 (12.63, 14.44)	12.54 (11.77, 13.40)	13.62 (12.64, 14.97)	14.480	< 0.001	0.290

Readability of English answers

All six readability metrics differed significantly across the three models. DeepSeek had the highest median Flesch Reading Ease score (32.00), compared with ChatGPT (26.00, $P < 0.05$) and Doubao (21.00, $P < 0.05$), indicating comparatively easier text. Its median Flesch-Kincaid Grade Level was also lower (13.95) than those of ChatGPT (14.90, $P < 0.05$) and Doubao (15.55, $P < 0.05$). Doubao showed greater estimated reading difficulty on the automated readability index, Gunning Fog Index, Coleman-Liau Index, and SMOG Index. Nevertheless, the outputs of all three models remained above the sixth- to eighth-grade levels commonly recommended for public-facing health information, with technical terminology and complex sentence structures remaining important barriers to readability.

Content defects

Using the prespecified majority criterion, in which a defect required agreement by at least two raters, no factual errors or outdated medical recommendations were identified for any model. One Doubao response was classified as containing a hallucination. Key-information omissions were identified in two DeepSeek responses and four Doubao responses, but in none of the ChatGPT responses. Between-model differences were not statistically significant for either hallucinations or key-information omissions (both $P > 0.05$) (Table 3).

Table 3
Content defects identified for each model

Defect type	ChatGPT	DeepSeek	Doubao
Factual error	0/25 (0.0%)	0/25 (0.0%)	0/25 (0.0%)
Hallucination	0/25 (0.0%)	0/25 (0.0%)	1/25 (4.0%)
Key information omission	0/25 (0.0%)	2/25 (8.0%)	4/25 (16.0%)
Outdated medical advice	0/25 (0.0%)	0/25 (0.0%)	0/25 (0.0%)

Stage-stratified findings and correlations

Exploratory comparisons across six stages of care suggested that relative model performance varied by stage. Because each stage contained only three to six questions, statistical power was limited and the stage-specific results were interpreted cautiously. Stage-standardized scores are presented in Fig. 5.

Spearman analysis showed positive associations among GQS, QAMAI, EQIP, and PEMAT measures. The automated readability index, Gunning Fog Index, Flesch-Kincaid Grade Level, Coleman-Liau Index, and SMOG Index were positively correlated with one another, whereas Flesch Reading Ease varied in the opposite direction. Associations between information-quality measures and readability were not consistently aligned, indicating that higher quality did not systematically correspond to easier reading (Fig. 4). These correlations were calculated from 75 responses and were not interpreted as causal.

Discussion

This study identified distinct performance profiles among the three LLMs across multiple dimensions of RARP patient education. Four principal findings emerged. First, ChatGPT achieved higher EQIP and PEMAT actionability scores and also outperformed Doubao on QAMAI, whereas overall GQS did not differ significantly among models. Second, DeepSeek generated comparatively more readable English, while Doubao showed greater difficulty on most grade-level indices; however, all models exceeded recommended reading levels for public health materials. Third, content defects were infrequent under the majority-rater definition, with key-information omissions occurring mainly in DeepSeek and Doubao

responses. Fourth, the low level of agreement among the three raters introduced substantial measurement uncertainty into the multidimensional human evaluation.

ChatGPT's higher EQIP score may reflect stronger coverage of information structure, contextual explanation, practical usefulness, and support for patient decision-making. Its advantage in PEMAT actionability likewise indicates that its responses more often translated information into concrete patient-facing actions. This pattern is compatible with previous work on common urological malignancy queries, in which chatbot responses were often of moderate-to-high quality while actionability remained weaker than understandability [8]. Even so, ChatGPT, the highest-scoring model for actionability in the present study, reached a median PEMAT actionability score of only 59.00%, demonstrating considerable scope for improvement.

Differences in mDISCERN should be interpreted cautiously. Although ChatGPT and Doubao both scored higher than DeepSeek, the instrument rewards explicit discussion of sources, uncertainty, and routes to additional information; a higher score therefore does not, by itself, establish greater medical accuracy. QAMAI and EQIP measure broader but partly different constructs, and the resulting model rankings were not identical across instruments. This divergence reinforces the need to evaluate LLM-generated patient education with multiple complementary measures rather than a single scale.

The findings are consistent with the broader literature on LLM-generated patient education. Previous reviews have commonly assessed accuracy, completeness, readability, and understandability, but cross-study comparisons are constrained by differences in model versions, prompts, question sources, and rater expertise [7,13]. Within the RARP pathway, LLMs may be most useful for preparing initial drafts of preoperative education and follow-up question-and-answer materials. Clinical review remains necessary to verify evidence, incorporate individualized risk stratification, and specify when patients should seek medical assessment; treatment decisions should not be delegated to an LLM.

All three models generated English text at reading levels substantially above those generally recommended for public-facing health materials, consistent with earlier studies of prostate cancer information. Hershenhouse et al. and Trapp et al. found that LLM-generated prostate cancer content could be accurate or complete while remaining difficult to read [11,12]. A systematic review similarly reported that online health information frequently falls between approximately the tenth- and fifteenth-grade levels, despite recommendations that patient-facing materials target about the sixth-grade level [27].

Although DeepSeek performed better on most readability indices, it did not obtain the highest EQIP or QAMAI scores. This pattern may indicate tension between linguistic simplicity and informational completeness. Readability formulas are strongly influenced by sentence length, word length, and syllable count, and necessary terms such as RARP, PSMA PET/CT, ISUP Grade Group, and biochemical recurrence inevitably increase calculated reading difficulty. Improving accessibility should therefore not rely on removing essential medical concepts. More appropriate strategies include shorter sentences,

plain-language definitions, layered explanations, concise key-point summaries, and adaptation to the patient's health literacy.

Patient counseling throughout the RARP pathway must address treatment selection, the balance between oncological control and functional preservation, urinary and sexual recovery, and long-term surveillance for recurrence. Randomized evidence indicates that robotic and open prostatectomy can differ in perioperative, oncological, and functional outcomes [1], while different RARP approaches may influence early continence recovery even when longer-term differences are less consistent [2]. Accordingly, educational responses should avoid absolute assurances or isolated recovery percentages and should instead explain how outcome definitions, follow-up intervals, operative technique, and baseline function affect reported results.

Contemporary guidelines provide additional context for counseling. The European Association of Urology indicates that PSA should decline to an undetectable level after radical prostatectomy and that follow-up testing is typically performed every six months for the first three years and annually thereafter [19,21]. In suspected recurrence, PSMA PET/CT is most informative when the result could alter management, and salvage radiotherapy should not be postponed solely to await a higher PSA threshold in patients who are already appropriate candidates [22]. Because these recommendations are time-sensitive and clinically consequential, even high-quality LLM responses require interpretation by a urologist who can integrate pathology, PSA kinetics, age, comorbidity, and patient preferences.

The absence of majority-rated factual errors or outdated advice should not be interpreted as evidence that the models are inherently safe. The defect analysis was based on only 25 fixed questions, and requiring agreement from at least two raters may conceal clinically important concerns raised by a single expert. In addition, LLM outputs are probabilistic and may change as models are updated. Their appropriate role is therefore as clinician-reviewed educational support rather than an independent source of clinical guidance.

Low inter-rater agreement was a notable methodological finding. Even when established instruments are used, raters may differ in how they interpret completeness, clarity, usefulness, source adequacy, visual support, and action-oriented content. The low ICCs reduce confidence in small between-model differences, particularly for GQS and understandability. Future evaluations would benefit from structured rater training, calibration cases, explicit item-level scoring definitions, and formal procedures for resolving disagreement. Incorporating patient evaluations and expert reference responses could further strengthen external validity.

Several limitations constrain interpretation. The study evaluated 25 English-language questions with a single response from each model, limiting both coverage of the full range of RARP scenarios and assessment of response reproducibility. Results may also change with model versions, system updates, and platform access settings. Because only English outputs were analyzed, direct generalization to Chinese-language patient use is uncertain. Inter-rater agreement was low, and several instruments involve subjective judgments. We did not benchmark responses against a urologist-authored reference

answer or evaluate actual patient comprehension, knowledge gain, decisional conflict, or adherence. Finally, the small number of questions within each care stage makes the stage-stratified findings exploratory.

In conclusion, ChatGPT, DeepSeek, and Doubao showed different strengths as sources of RARP patient education. ChatGPT provided higher-quality patient information on several measures and more actionable guidance, whereas DeepSeek produced comparatively easier-to-read English. All three models generated text above recommended reading levels, and omission of key information remained a concern. The low agreement among raters further underscores uncertainty in current multidimensional human assessment. LLMs may assist in drafting materials for professional review, but they should not replace urologist-led risk communication, shared decision-making, or individualized follow-up.

Methods

Study design

This cross-sectional observational evaluation was performed in Sichuan, China, on 18 August 2026. Reporting was guided by the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement [17] and the Chatbot Assessment Reporting Tool (CHART) guidance [18], including documentation of model identity and settings, prompts, query procedures, outcome measures, and statistical methods.

The evaluation was limited to publicly framed questions and text generated by the models. No patients, animals, human tissue, or identifiable personal data were involved; institutional ethics approval was therefore not required. Clinical experts participated in the study as reviewers and evaluators. Informed consent was obtained from all the experts involved in the study.

Question development and stage grouping

Candidate RARP questions were assembled from three sources. First, Google Trends data from 2020 to 2026 were searched using terms including prostate cancer, robot-assisted radical prostatectomy, and RARP to identify topics of high patient interest. Second, commonly asked questions were collected from major Chinese health-information platforms, including Baidu Health, Dingxiang Doctor, and Good Doctor Online, together with patient-education pages from health-care institutions. Third, clinically important topics were derived from the 2026 European Association of Urology prostate cancer guidelines and related peer-reviewed updates, the 2026 American Urological Association/American Society for Radiation Oncology guideline amendment for clinically localized prostate cancer, and guidance on PSA persistence, biochemical recurrence, and salvage treatment after radical prostatectomy [19–22].

Duplicate items were removed, closely related questions were consolidated, and the wording was reviewed by expert urology clinicians. The resulting set comprised 25 questions (Table 4) covering screening and diagnosis, grading and preoperative assessment, treatment and surgical decision-making,

perioperative management, postoperative complications and functional recovery, pathological interpretation, recurrence management, follow-up, and long-term health.

Table 4

Common patient questions about robot-assisted radical prostatectomy used as exact prompts across all LLM platforms

Question No.	Question category	Question
1	Basic understanding, etiology and risk	Why did I develop prostate cancer, and how do age, family history, and genetic factors affect my risk of developing it?
2	Diagnosis, staging and preoperative assessment	If my PSA is elevated, does that necessarily mean I have prostate cancer, and what further evaluations do I need before deciding whether to undergo a biopsy?
3	Clinical presentation and indications to seek care	I do not have urinary frequency, urgency, or difficulty urinating. Could I still have localized prostate cancer?
4	Clinical presentation and indications to seek care	If I develop new, persistent bone pain, should I be evaluated for possible bone metastases from prostate cancer?
5	Diagnosis, staging and preoperative assessment	What tests are usually needed before prostate cancer is diagnosed, and what are the respective roles of PSA, digital rectal examination, MRI, and biopsy?
6	Diagnosis, staging and preoperative assessment	How will my prostate MRI results affect whether I need a biopsy, where the biopsy is targeted, and subsequent RARP surgical planning?
7	Diagnosis, staging and preoperative assessment	What do the Gleason score and ISUP Grade Group in my biopsy report mean, and how do they affect whether I choose RARP?
8	Diagnosis, staging and preoperative assessment	How does my doctor use my PSA, clinical stage, and pathological grade to determine my prostate cancer risk group?
9	Diagnosis, staging and preoperative assessment	Before deciding on RARP, do I need PSMA PET/CT, CT, or a bone scan to determine whether there is lymph node or distant metastasis?
10	Treatment options and RARP decision-making	For me, what are the main advantages and disadvantages of active surveillance, radiotherapy, and RARP, and how should I participate in treatment decision-making?
11	Treatment options and RARP decision-making	Compared with open or conventional laparoscopic radical prostatectomy, what differences in cancer control and functional recovery should I really focus on with RARP?
12	Treatment options and RARP decision-making	During RARP, can my nerves be spared, and how does the surgeon balance recovery of sexual function with the safety of complete cancer removal?
13	Treatment options and RARP decision-making	Do I need pelvic lymph node dissection during RARP, and what factors will my doctor use to decide?
14	Treatment options and RARP decision-making	How will my MRI findings and risk of extracapsular extension affect the extent of resection during RARP and my risk of positive surgical margins?

Question No.	Question category	Question
15	Postoperative complications and functional recovery	After RARP, what common complications and long-term functional effects do I most need to understand?
16	Postoperative complications and functional recovery	Why can urinary incontinence occur after RARP, and when should I start pelvic floor muscle training?
17	Postoperative complications and functional recovery	Why can erectile dysfunction occur after RARP, and how important is nerve sparing for my recovery of sexual function?
18	Postoperative complications and functional recovery	If my urinary incontinence does not improve over the long term, when should I undergo further evaluation and consider a male sling or artificial urinary sphincter?
19	Prognosis, recurrence and salvage treatment	In my pathology report after RARP, how do surgical margin status, pT stage, lymph node status, and ISUP grade each affect my risk of recurrence?
20	Prognosis, recurrence and salvage treatment	After RARP, to what level should my PSA usually fall, and why do I need long-term regular monitoring?
21	Prognosis, recurrence and salvage treatment	If my PSA rises again after RARP, what findings suggest biochemical recurrence, and do I need PSMA PET/CT?
22	Prognosis, recurrence and salvage treatment	If my PSA begins to suggest recurrence, why might my doctor recommend discussing salvage radiotherapy early rather than waiting until the PSA becomes very high?
23	Daily management, follow-up and quality of life	How often should I have my PSA and functional recovery assessed after RARP?
24	Daily management, follow-up and quality of life	How should I arrange pelvic floor muscle training and continence rehabilitation after RARP so that I can better return to daily life?
25	Daily management, follow-up and quality of life	After RARP, how should I manage exercise, weight, diet, smoking, and psychological support over the long term?

Model question-and-answer procedure

The same 25 English-language questions were submitted to ChatGPT 5.6 Sol, DeepSeek-V3.2, and Doubao 2.1 Turbo. Each question was entered in a separate chat to minimize carryover from preceding prompts. Web access, search functions, plugin functions, and reasoning mode were enabled on all three platforms during data collection. Responses were retained exactly as generated, producing a dataset of 75 outputs (25 questions x 3 models).

Assessment tools

GQS was scored from 1 to 5 and was used to assess overall information quality and usefulness to patients. The modified DISCERN (mDISCERN) measure was based on core DISCERN principles and comprised five binary items addressing a clearly stated aim, source reliability, balance, directions to further information, and acknowledgment of uncertainty; total scores ranged from 0 to 5 [23]. Analyses used the original binary structure contained in the source spreadsheet rather than the older 5-to-25-point scoring scheme.

QAMAI comprises six domains - accuracy, clarity, relevance, completeness, source, and usefulness - each rated from 0 to 5, yielding a total score of 0 to 30; its development study reported construct validity, internal consistency, and inter-rater reliability [14]. EQIP contains 20 items, and raw scores were converted to percentages [15]. PEMAT was used to derive separate percentages for understandability and actionability [16]. These instruments were prespecified as complementary measures rather than treated as interchangeable indicators of a single construct.

English readability was quantified using the automated readability index (ARI), Flesch Reading Ease score (FRES), Gunning Fog Index (GFI), Flesch-Kincaid Grade Level (FKGL), Coleman-Liau Index (CLI), and SMOG Index. FRES, FKGL, and SMOG were computed according to their original formulas [24–26]. Higher values on the grade-level measures generally indicate greater reading difficulty, whereas higher FRES values indicate easier text. These formulas primarily capture sentence length and lexical or syllabic complexity and do not directly measure factual accuracy or actual comprehension [27].

Content defects were categorized as factual errors, hallucinations, omissions of key information, or outdated medical advice. Each clinician independently recorded whether each defect was present. A defect was assigned to a response only when at least two of the three raters marked it as present.

Assessment process

Three clinicians serving as expert raters independently evaluated the responses after de-identification, random ordering, and removal of model labels. Inter-rater agreement was quantified with ICCs, interpreted as poor below 0.50, moderate from 0.50 to 0.75, good from 0.75 to 0.90, and excellent above 0.90 [28]. Score differences of at least one point were reviewed through discussion in an effort to reach consensus; unresolved cases were adjudicated by a senior physician.

Statistical analysis

All analyses were conducted in R version 4.5.0. The patient question, rather than an individual rater score, was the independent paired unit ($n = 25$). Continuous outcomes were summarized using the mean, standard deviation, median, and interquartile range. Given the ordinal properties of several measures, the sample size, and distributional characteristics, overall comparisons among the three models used the Friedman test, with Kendall's W as the effect size. Pairwise comparisons used paired Wilcoxon signed-rank tests with Holm-adjusted P values and paired rank-biserial correlations as effect sizes. Content-defect frequencies were compared using Cochran's Q test followed by Holm-adjusted paired McNemar tests. Analyses by stage of care were considered exploratory. Spearman correlations

were used to describe monotonic associations among measures without causal interpretation. Two-sided adjusted P values < 0.05 were considered statistically significant.

Declarations

Competing interests

The authors declare no competing interests.

Funding

This work was supported by the Health Commission of Sichuan Province Medical Science and Technology Program (24LCYJPT15), Medical Research Project of Sichuan Medical Association (S2024066), Sichuan Medical Science and Technology Innovation Research Association Project (2025YCZD007), and Sichuan Provincial Science and Technology-Education Joint Fund of the Sichuan Science and Technology Department (2026NSFSC2081).

Author Contribution

Q.W. conceived and designed the study. Q.W., X.W., and X.J. performed the analyses and interpreted the data. S.X., H.H., H.R., Y.L., H.G., and Q.C. curated the dataset. Q.W. and X.W. prepared the initial manuscript draft. T.W. provided critical intellectual revision, supervised the study, and obtained funding. All authors reviewed and approved the final manuscript and accept responsibility for the integrity of the work.

Data availability

The standardized prompts used in the study are listed in Table 4. The de-identified LLM responses, scoring sheets, and analysis scripts are available from the corresponding author upon reasonable request.

Code availability

The R scripts used for the analyses are available from the corresponding author upon reasonable request.

References

1. Nahas, W. C. et al. Perioperative, oncological, and functional outcomes between robot-assisted laparoscopic prostatectomy and open radical retropubic prostatectomy: a randomized clinical trial. *J. Urol.* 212, 32–40 (2024). <https://doi.org/10.1097/JU.0000000000003967>

2. Ficarra, V. et al. Impact of transperitoneal anterior, Retzius-sparing, extraperitoneal, transvesical and perineal approaches on urinary continence recovery after robot-assisted radical prostatectomy: a systematic review and meta-analysis of comparative studies. *Prostate Cancer Prostatic Dis.* 28, 328–341 (2025). <https://doi.org/10.1038/s41391-025-00943-2>
3. Grune, B. et al. Decision aids for shared decision-making in uro-oncology: a systematic review. *Eur. Urol. Focus* 8, 851–869 (2022). <https://doi.org/10.1016/j.euf.2021.04.013>
4. Violette, P. D. et al. Decision aids for localized prostate cancer treatment choice: systematic review and meta-analysis. *CA Cancer J. Clin.* 65, 239–251 (2015). <https://doi.org/10.3322/caac.21272>
5. Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F. & Ting, D. S. W. Large language models in medicine. *Nat. Med.* 29, 1930–1940 (2023). <https://doi.org/10.1038/s41591-023-02448-8>
6. Busch, F. et al. Current applications and challenges in large language models for patient care: a systematic review. *Commun. Med.* 5, 26 (2025). <https://doi.org/10.1038/s43856-024-00717-2>
7. Aydin, S., Karabacak, M., Vlachos, V. & Margetis, K. Large language models in patient education: a scoping review of applications in medicine. *Front. Med.* 11, 1477898 (2024). <https://doi.org/10.3389/fmed.2024.1477898>
8. Musheyev, D. et al. How well do artificial intelligence chatbots respond to the top search queries about urological malignancies? *Eur. Urol.* 85, 13–16 (2024). <https://doi.org/10.1016/j.eururo.2023.07.004>
9. Szczesniewski, J. J., Tellez Fouz, C., Ramos Alba, A. et al. ChatGPT and most frequent urological diseases: analysing the quality of information and potential risks for patients. *World J. Urol.* 41, 3149–3153 (2023). <https://doi.org/10.1007/s00345-023-04563-0>
10. Gibson, D. et al. Evaluating the efficacy of ChatGPT as a patient education tool in prostate cancer: multimetric assessment. *J. Med. Internet Res.* 26, e55939 (2024). <https://doi.org/10.2196/55939>
11. Hershenhouse, J. S. et al. Accuracy, readability, and understandability of large language models for prostate cancer information to the public. *Prostate Cancer Prostatic Dis.* 28, 394–399 (2025). <https://doi.org/10.1038/s41391-024-00826-y>
12. Trapp, C. et al. Patient- and clinician-based evaluation of large language models for patient education in prostate cancer radiotherapy. *Strahlenther. Onkol.* 201, 333–342 (2025). <https://doi.org/10.1007/s00066-024-02342-3>
13. Al-Khanaty, A. et al. Quality and usability of prostate cancer information generated by artificial intelligence chatbots: a comparative analysis. *Cancers* 18, 906 (2026). <https://doi.org/10.3390/cancers18060906>
14. Vaira, L. A. et al. Validation of the Quality Analysis of Medical Artificial Intelligence (QAMAI) tool: a new tool to assess the quality of health information provided by AI platforms. *Eur. Arch. Otorhinolaryngol.* 281, 6123–6131 (2024). <https://doi.org/10.1007/s00405-024-08710-0>
15. Moul, B., Franck, L. S. & Brady, H. Ensuring quality information for patients: development and preliminary validation of a new instrument. *Health Expect.* 7, 165–175 (2004).

- <https://doi.org/10.1111/j.1369-7625.2004.00273.x>
16. Shoemaker, S. J., Wolf, M. S. & Brach, C. Development of the Patient Education Materials Assessment Tool (PEMAT). *Patient Educ. Couns.* 96, 395–403 (2014).
<https://doi.org/10.1016/j.pec.2014.05.027>
 17. von Elm, E. et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *J. Clin. Epidemiol.* 61, 344–349 (2008). <https://doi.org/10.1016/j.jclinepi.2007.11.008>
 18. The CHART Collaborative. Reporting guideline for chatbot health advice studies: the CHART statement. *JAMA Netw. Open* 8, e2530220 (2025).
<https://doi.org/10.1001/jamanetworkopen.2025.30220>
 19. Cornford, P. et al. EAU-EANM-ESTRO-ESUR-ISUP-SIOG guidelines on prostate cancer-2026 update. Part I: screening, diagnosis, and local treatment with curative intent. *Eur. Urol.* (2026).
<https://doi.org/10.1016/j.eururo.2026.04.013>
 20. Eastham, J. A. et al. Clinically localized prostate cancer: AUA/ASTRO guideline amendment (2026). *J. Urol.* 216, 2–11 (2026). <https://doi.org/10.1097/JU.0000000000005060>
 21. Tilki, D. et al. EAU-EANM-ESTRO-ESUR-ISUP-SIOG guidelines on prostate cancer. Part II-2026 update: treatment of relapsing and metastatic prostate cancer. *Eur. Urol.* (2026).
<https://doi.org/10.1016/j.eururo.2026.04.014>
 22. Morgan, T. M. et al. Salvage therapy for prostate cancer: AUA/ASTRO/SUO guideline part I: introduction and treatment decision-making at the time of suspected biochemical recurrence after radical prostatectomy. *J. Urol.* 211, 509–517 (2024).
<https://doi.org/10.1097/JU.0000000000003892>
 23. 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* 53, 105–111 (1999).
 24. Flesch, R. A new readability yardstick. *J. Appl. Psychol.* 32, 221–233 (1948).
 25. Kincaid, J. P., Fishburne, R. P., Rogers, R. L. & Chissom, B. S. Derivation of new readability formulas for Navy enlisted personnel. Naval Technical Training Command, Millington, TN (1975).
 26. McLaughlin, G. H. SMOG grading: a new readability formula. *J. Read.* 12, 639–646 (1969).
 27. Daraz, L. et al. Readability of online health information: a meta-narrative systematic review. *Am. J. Med. Qual.* 33, 487–492 (2018). <https://doi.org/10.1177/1062860617751639>
 28. Koo, T. K. & Li, M. Y. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J. Chiropr. Med.* 15, 155–163 (2016). <https://doi.org/10.1016/j.jcm.2016.02.012>

Figures

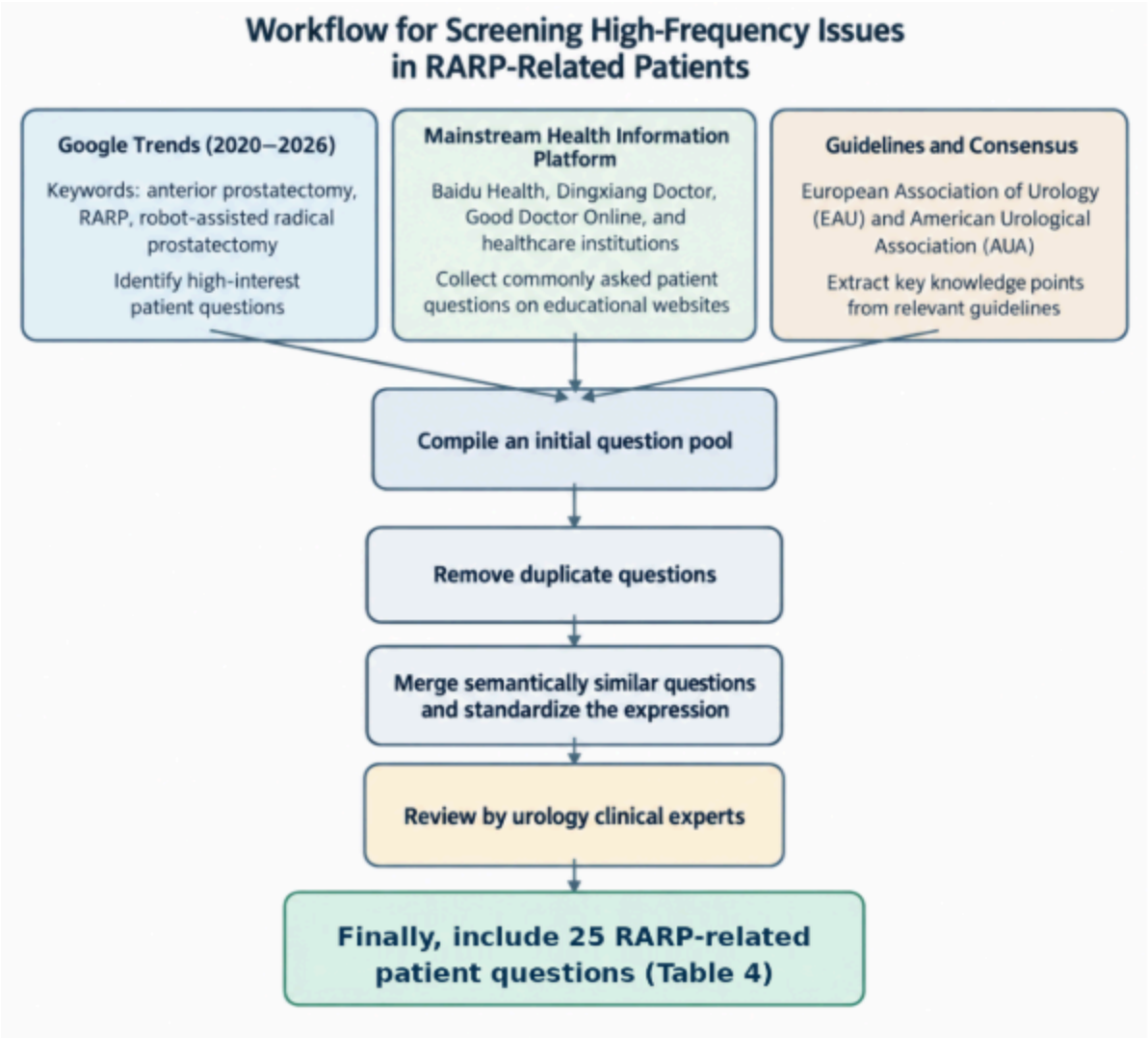


Figure 1

Workflow used to identify and screen high-frequency patient questions related to RARP.

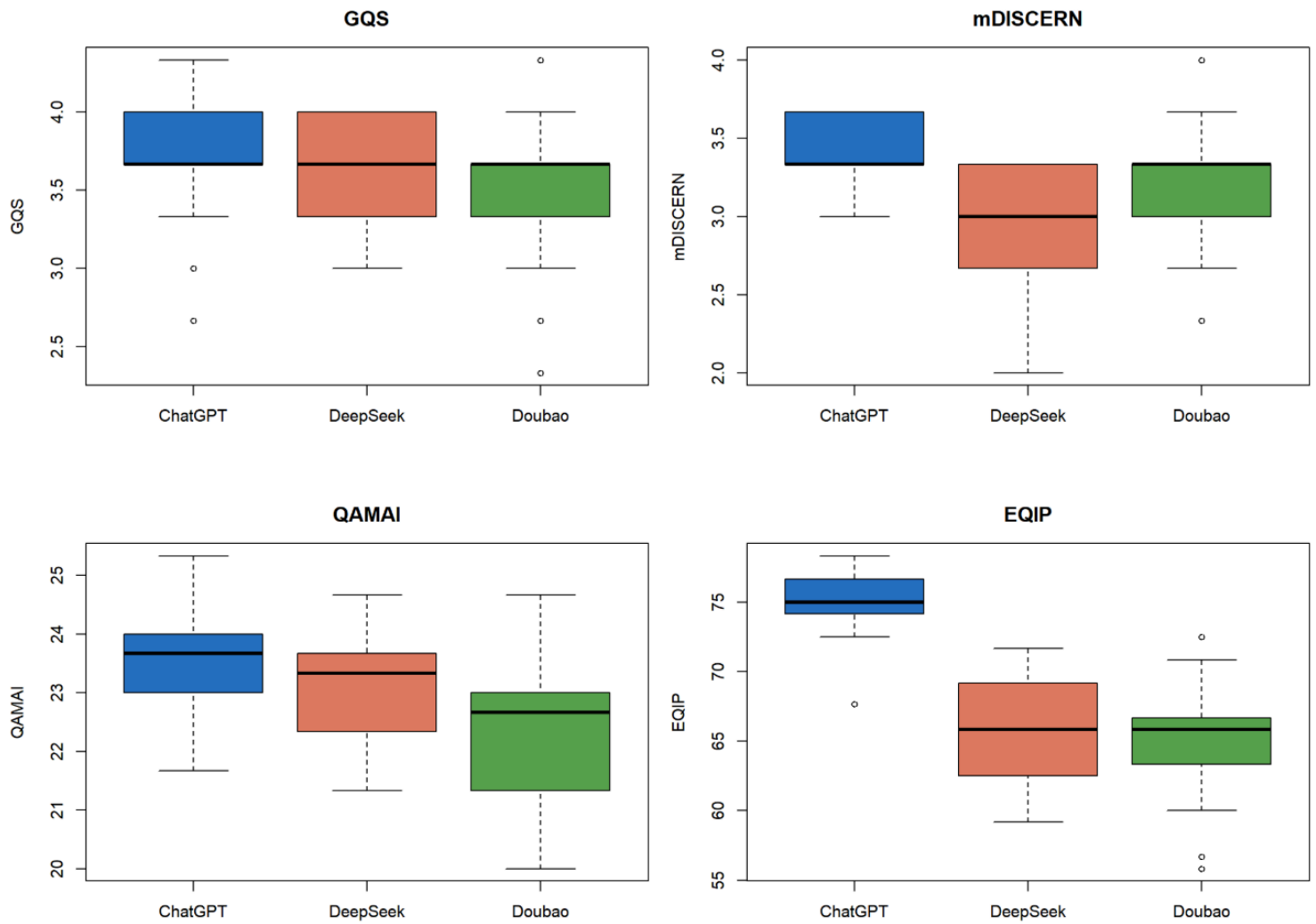


Figure 2

Paired distributions of quality and reliability measures for ChatGPT, DeepSeek, and Doubao.

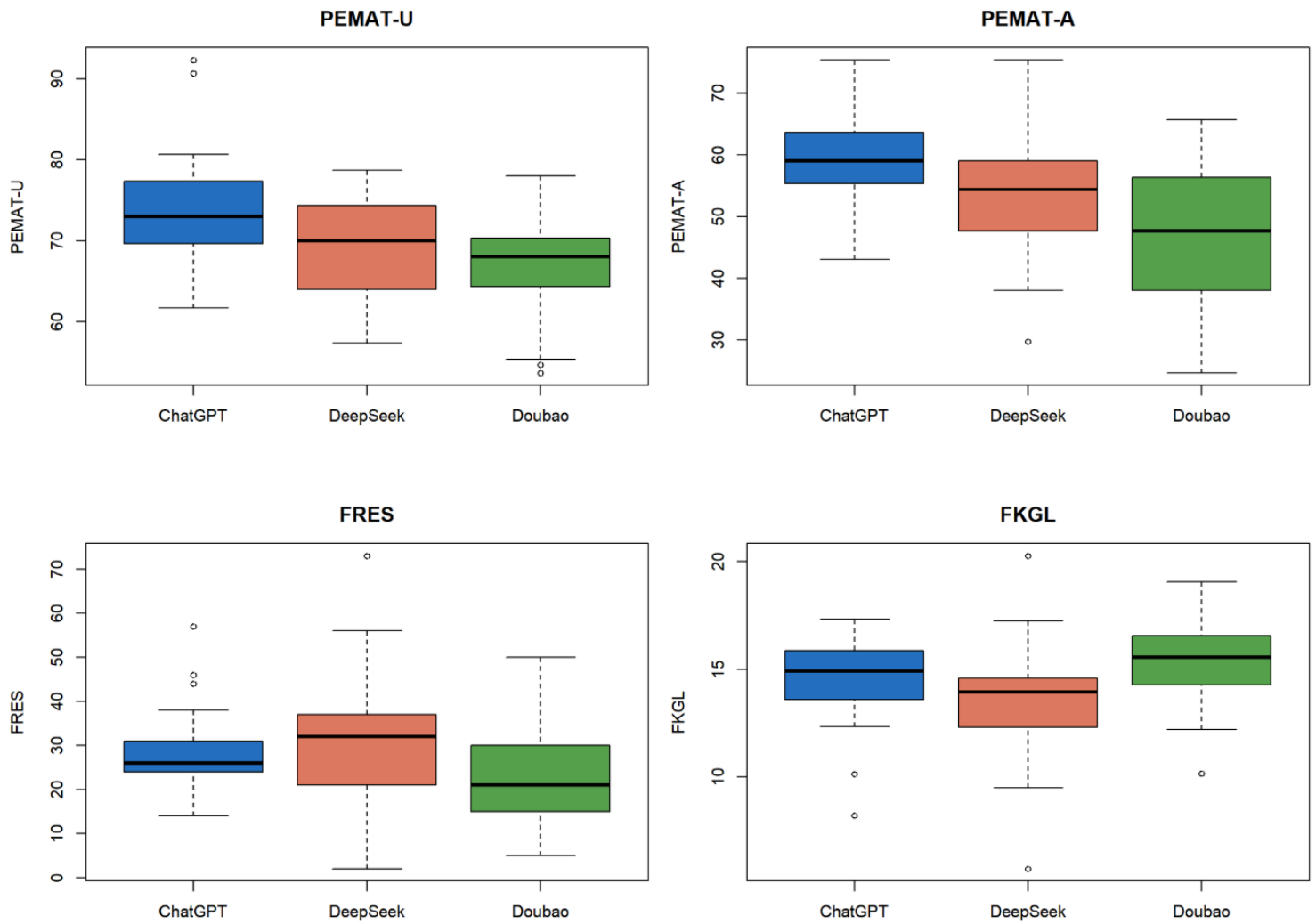


Figure 3

Understandability, actionability, and principal readability measures across the three models.

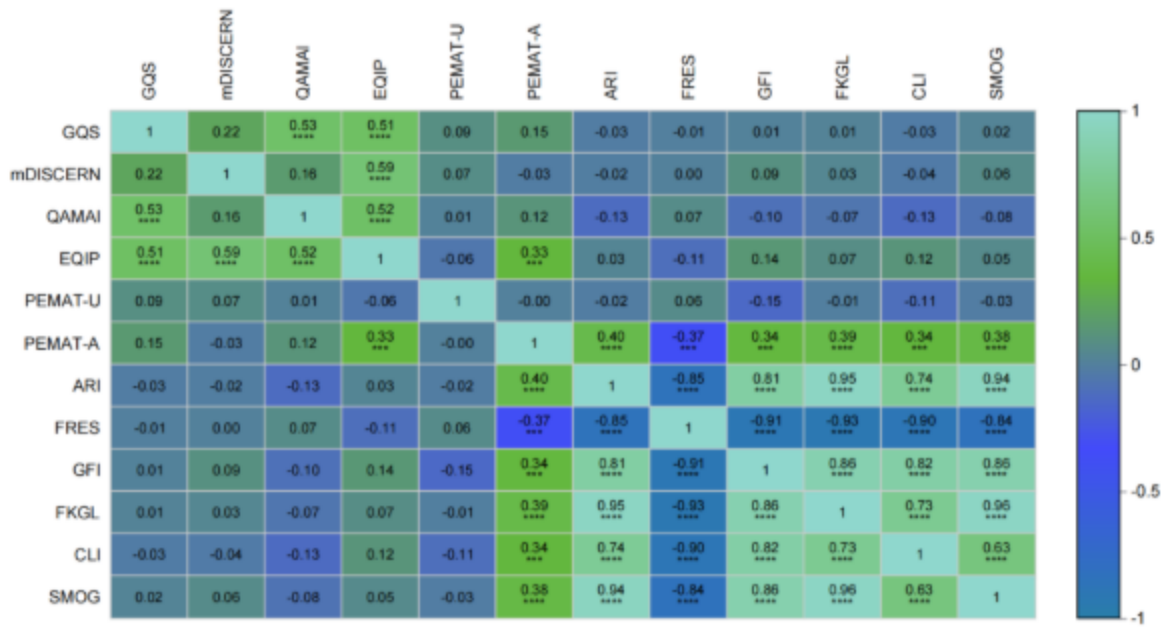


Figure 4

Spearman correlation heat map showing associations among quality, reliability, understandability, actionability, and readability measures.

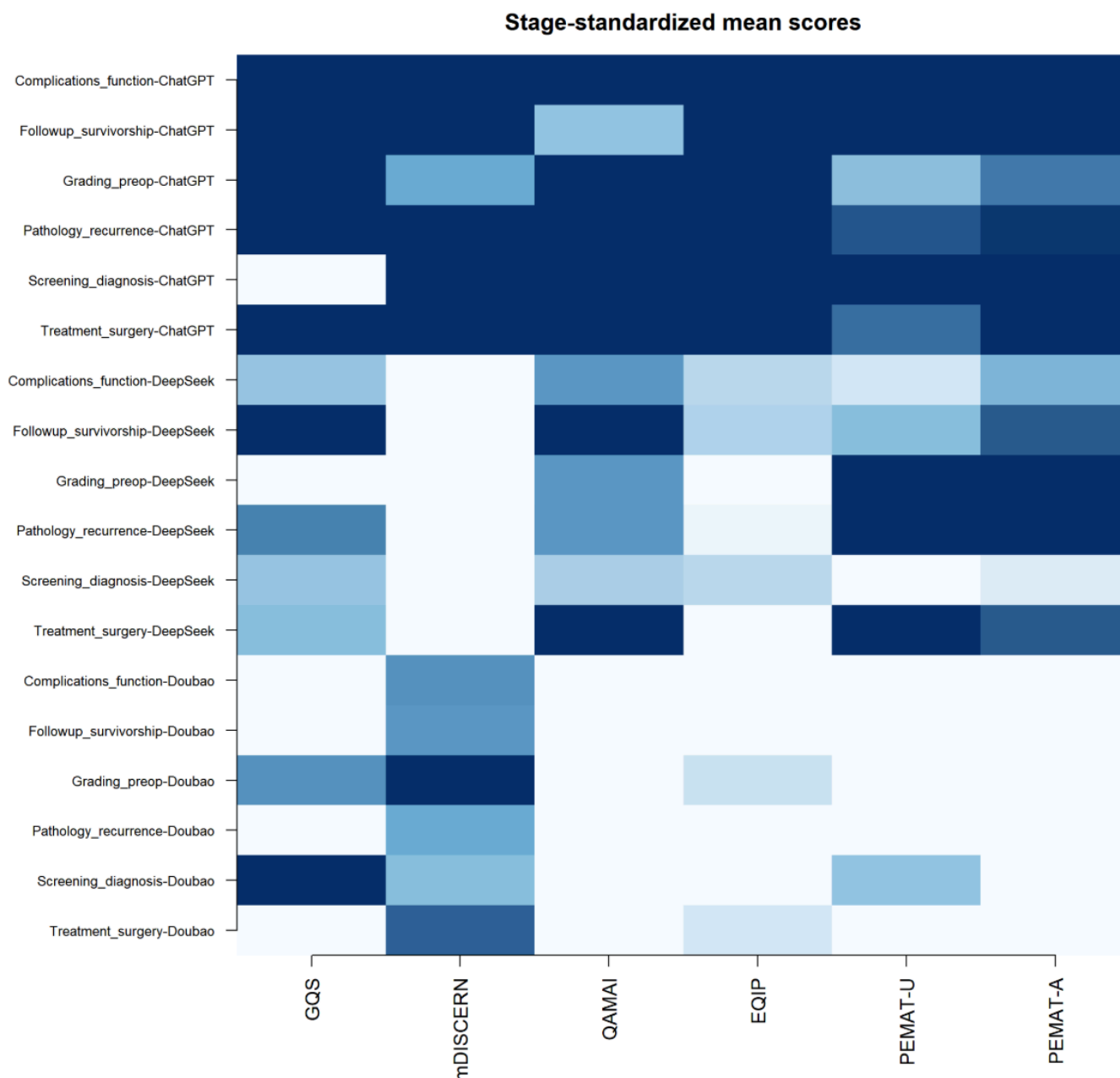


Figure 5

Stage-standardized mean scores across six stages of care. Heat map of stage-standardized mean scores for GQS, mDISCERN, QAMAI, EQIP, PEMAT understandability (PEMAT-U), and PEMAT actionability (PEMAT-A) across the six stages of care. Stage-level comparisons were exploratory because each stage included only 3-6 questions.