

Evaluating Cross-Platform Large Language Models for Patient Education in Hypertension and Chronic Kidney Disease: Quality, Readability, and Actionability

Zhiqiang Wang

Department of Cardiology, Huanggang Central Hospital, Huanggang

Mengqiu Hu

Department of Cardiology, Huanggang Central Hospital, Huanggang

Xiaoya Li

School of Medicine, Yangtze University

Zhiwen Zhang

fwzzw01@163.com

Central China Fuwai Hospital of Zhengzhou University

Muwei Li

Central China Fuwai Hospital of Zhengzhou University

Article

Keywords: generative artificial intelligence, large language models, hypertension, chronic kidney disease, patient education, readability

Posted Date: June 2nd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9358054/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

Objective: To compare the understandability, actionability, information quality, and readability of patient education texts on hypertension with chronic kidney disease generated by 12 cross-platform large language models, and to further examine their structural patterns using exploratory network analysis and latent profile analysis.

Methods: Based on the literature, clinical guidelines, and expert discussion, 20 frequently asked questions were developed and categorized into five thematic modules. On March 10, 2026, independent English prompts were submitted to 12 platforms, generating 240 responses. Texts were evaluated using PEMAT-P, EQIP-36, GQS, JAMA, DISCERN, AIRM, and seven English readability formulas. Data were summarized as medians and interquartile ranges. Exploratory Kruskal-Wallis tests, Spearman correlation analyses, regularized partial correlation network analysis, and latent profile analysis were then performed.

Results: Overall, median PEMAT-P understandability, actionability, and total scores were 69.20, 20.00, and 55.60, respectively; median EQIP-36, DISCERN, and JAMA scores were 35.71, 38.00, and 1.00; FRES and SMOG were 20.53 and 16.36, indicating limited actionability, poor source transparency, and substantial reading burden. Across platforms, ERNIE Bot 5.0 Preview showed higher PEMAT-P understandability and total scores; DeepSeek-V3.2 showed lower reading burden across multiple readability indices and a higher FRES; Perplexity showed the greatest reading burden. Performance also varied by module: “Lifestyle Intervention and Long-Term Management” scored higher on EQIP-36 and DISCERN, whereas “Risk Alerts and Special Considerations” scored higher on PEMAT-P total score. Network analysis identified JAMA and DISCERN Q1-Q8 as central nodes. Latent profile analysis supported five statistical phenotypes.

Conclusions: Under standardized English prompts, large language model-generated patient education texts for hypertension with chronic kidney disease showed relatively good understandability but limited actionability and source transparency, while overall reading burden remained high. At present, these models are better suited for draft generation and language support in patient education than for use as independent sources of advice without clinical review.

Introduction

Hypertension is a common chronic noncommunicable disease and a major risk factor for cardiovascular disease and chronic kidney disease (CKD). In 2024, approximately 1.4 billion adults worldwide had hypertension, yet the blood pressure control rate remained below 20%, imposing a substantial medical and economic burden^[1]. Hypertension is closely associated with CKD. Among patients with CKD, the prevalence of hypertension is high and increases with disease progression, rising from approximately 22% in CKD stage 1 to about 80% in CKD stage 4^[2]. In 2023, an estimated 788 million adults aged 20 years or older were living with CKD globally, with an age-standardized prevalence of 14.2%^[3]. The final report of the Systolic Blood Pressure Intervention Trial showed that, among individuals without diabetes or prior stroke but at elevated cardiovascular risk, a systolic blood pressure target of < 120 mmHg, compared with < 140 mmHg, further reduced the risks of major cardiovascular events and all-cause mortality^[4], supporting the benefit of intensive blood pressure lowering for improving cardiovascular and renal outcomes. Over the past three decades, incident CKD attributable to hypertension has increased by more than 60%, and in some regions the prevalence of hypertension-CKD comorbidity has approached 30%^[5–6]. Accordingly, patients with hypertension and CKD face a greater disease burden and a more urgent need for health education and long-term management.

Sustained blood pressure control depends not only on pharmacotherapy and clinical follow-up, but also on patients’ disease knowledge and self-management capacity. Evidence indicates that limited health literacy is associated with misunderstanding of medical information and a higher risk of emergency revisits^[7]. Among patients with hypertension, the proportion with both adequate health literacy and good medication adherence remains low, with older adults, low-income populations, and people with disabilities at a particular disadvantage^[7]. Existing patient education materials remain suboptimal in readability, understandability, and actionability. One study found that only 2.1% of patient education texts published in high-impact medical journals met a sixth-grade reading level^[8]. Other materials had a Flesch Reading Ease Score below 50 and a Simple Measure of Gobbledygook grade above 13, while understandability and actionability scores also fell below the commonly used 70% threshold^[9]. Together, these findings suggest that current patient education content remains insufficiently tailored to the cognitive characteristics of target populations and inadequately supports translation into health-related behaviors.

The development of generative large language models (LLMs) has opened a new avenue for optimizing medical education texts. Previous studies have shown that ChatGPT, Gemini, and Claude can reduce the reading grade level of American Heart Association patient education materials from approximately grade 10 to 5.6–7.6^[10]. GPT-4 has also been shown to reduce reading difficulty when rewriting patient materials in interventional radiology, although factual errors persisted and the recommended reading level was still not achieved^[11]. These findings indicate that LLMs have the potential to improve readability, but output quality varies across platforms, and factual inaccuracies may still occur; therefore, human review remains essential in medical applications^[11]. Further evidence suggests that Med-PaLM 2 can achieve an accuracy of up to 86.5% on the MedQA benchmark^[12]. However, in hypertension-related clinical scenarios, the response accuracy of GPT-4, Gemini, and MedLM remains below expert performance^[13]. In tasks such as citation generation for systematic reviews and clinical summary generation, LLMs may also produce hallucinations, omit key information, or distort content^[14–15]. Even when retrieval-augmented generation is introduced, evaluators may still disagree in their judgments of output quality^[16–17]. Therefore, whether LLMs can generate patient education content that is simultaneously accurate, understandable, and actionable requires systematic evaluation in specific disease contexts.

At present, cross-platform comparative evidence on LLM-generated patient education texts for patients with hypertension and CKD remains limited. In particular, comprehensive evaluations conducted under standardized prompts across understandability, actionability, information quality, source transparency, and readability are lacking. To address this gap, the present study included 12 conversational artificial intelligence platforms and conducted a multidimensional evaluation of generated texts in the context of health education and management for hypertension with CKD using the Patient Education

Materials Assessment Tool for Printable Materials (PEMAT-P), Ensuring Quality Information for Patients with 36 items (EQIP-36), Global Quality Score (GQS), Journal of the American Medical Association benchmark criteria (JAMA), DISCERN, Artificial Intelligence Response Measurement (AIRM), and seven English readability formulas. Exploratory network analysis and latent profile analysis were further performed to inform the development of intelligent patient education systems that integrate evidence-based content, understandability, and practical usability.

Methods

Study Design and Question Collection

This study aimed to systematically identify common public information needs regarding hypertension with CKD and to develop a clinically relevant and representative question dataset for evaluating the performance of LLMs in generating patient education texts. Based on the literature, authoritative clinical guidelines, and expert discussion, an initial pool of 20 frequently asked questions was developed (Table 1). According to the clinical management pathway for patient education in hypertension with CKD, the questions were organized into five thematic modules: Foundational Disease Knowledge; Screening, Diagnosis, and Stratification; Treatment Targets and Pharmacologic Management; Lifestyle Intervention and Long-Term Management; and Risk Alerts and Special Considerations. This development process was informed by previous similar studies to ensure comprehensive coverage and clinical applicability[18–20]. To ensure content validity and standardized wording, three cardiovascular experts from the author team independently reviewed the question list, focusing on alignment between individual questions and thematic modules, clinical accuracy, suitability for patient education, and the appropriateness of risk-related wording. These experts included the two corresponding authors and the first author; one corresponding author is a medical professor, one corresponding author is a medical postdoctoral researcher, and the first author is a clinician with more than 10 years of clinical experience. Discrepant items were revised through consensus discussion, yielding the final dataset.

Table 1
List of Frequently Asked Questions for Patients With Hypertension and CKD

Issue List
I. Foundational Disease Knowledge
1. Why can hypertension damage the kidneys?
2. Why can chronic kidney disease (CKD) worsen hypertension?
3. Which populations are at greater risk of developing coexisting hypertension and CKD?
4. Why is the coexistence of hypertension and CKD considered a high-risk condition?
II. Screening, Diagnosis, and Stratification
1. What is the most important basic test for screening for hypertension with CKD?
2. What is the time-based diagnostic criterion for CKD?
3. How is CKD staged according to estimated glomerular filtration rate (eGFR)?
4. How is CKD stratified according to the level of albuminuria?
III. Treatment Targets and Pharmacologic Management
1. What is the target blood pressure for patients with hypertension and CKD?
2. Which patients should be prioritized for treatment with an angiotensin-converting enzyme inhibitor (ACEI) or an angiotensin receptor blocker (ARB)?
3. After initiating an ACEI or ARB, when should serum creatinine be reassessed?
4. Which patients should be considered for treatment with a sodium-glucose cotransporter 2 (SGLT2) inhibitor?
IV. Lifestyle Intervention and Long-Term Management
1. Which dietary component should be most strictly restricted in patients with hypertension and CKD?
2. Are low-sodium salt substitutes appropriate for patients with hypertension and CKD?
3. How should home blood pressure monitoring be conducted in a standardized manner?
4. Which indicators should be prioritized during long-term follow-up?
V. Risk Alerts and Special Considerations
1. Which situations indicate the need for prompt referral to a nephrologist?
2. What is resistant hypertension?
3. Does the development of hyperkalemia necessarily require immediate discontinuation of an ACEI or ARB?
4. Which manifestations suggest substantial disease deterioration?

AI Platform Questioning and Data Collection

This cross-sectional study included 12 publicly available domestic and international conversational AI platforms: Ant Afu (Ant Group, December 15, 2025)^[21], ChatGPT 5.1-Thinking (OpenAI, November 12, 2025)^[22], ChatGPT 5.2-Thinking (OpenAI, December 11, 2025)^[23], DeepSeek-V3.2 (DeepSeek, December 1, 2025)^[24], Doubao (ByteDance/Volcano Engine, June 11, 2025)^[25], ERNIE Bot 5.0 Preview (Baidu, November 13, 2025)^[26], Qwen 3.5-plus (Alibaba Cloud, February 16, 2026)^[27], Baichuan Doctor Edition (Beijing Baichuan Intelligent Technology Co., Ltd., January 22, 2026)^[28], Baichuan Patient Edition (Beijing Baichuan Intelligent Technology Co., Ltd., January 22, 2026)^[28], Gemini 3-pro (Google LLC, November 18, 2025)^[29], Perplexity (Perplexity AI, Inc., December 7, 2022)^[30], and Tencent Yuanbao (Shenzhen Tencent Computer Systems Company Limited, May 30, 2024)^[31].

All queries were submitted in English to ensure prompt consistency. For each prespecified question, a new independent conversation was initiated on each platform, and prior chat history was cleared before every query to minimize residual contextual effects on the response. Within a unified time window on March 10, 2026, all 20 questions were entered sequentially across platforms, yielding 240 AI-generated texts. Basic textual features, including word count and sentence count, were recorded simultaneously. All responses were preserved verbatim without manual editing. Because the comparison focused on the actual output of publicly available default product versions, the findings are more appropriately interpreted as a platform-level comparison of patient education outputs rather than a strict head-to-head evaluation of fully standardized underlying model capabilities.

Evaluation Tools and Measures

To comprehensively evaluate outputs from the 12 platforms, this study used three categories of measures: (1) understandability/actionability: PEMAT-P; (2) information quality: EQIP-36, GQS, JAMA, DISCERN (Q1-Q15), and AIRM; and (3) objective readability: seven English readability formulas. Except for automatically calculated readability indices, PEMAT-P, EQIP-36, GQS, JAMA, DISCERN, and AIRM were independently rated by two trained reviewers. Before formal scoring, the reviewers developed scoring rules and examples based on the original manuals, then conducted pilot testing and calibration using 30 randomly selected texts. All texts were de-identified and randomly ordered before assessment, and both reviewers were blinded to platform source and thematic module. Interrater agreement in the pilot phase was high (overall kappa = 0.891). Because the included tools comprised binary items, ordinal ratings, and total scores, this index was used to reflect overall consistency of the scoring process. Any substantial disagreement on a given scale was adjudicated by a senior researcher with a cardiovascular background.

Understandability Assessment

PEMAT-P was used to assess understandability and actionability^[32]. Higher PEMAT-P scores indicate that the text is easier to understand and more actionable for readers^[32].

Quality Assessment

EQIP-36. EQIP-36 is a 36-item instrument for assessing the quality of patient information. It evaluates content, source identification, and structure, and provides an overall quality score from 0 to 100, with higher scores indicating better information quality^[33]. The EQIP has been used to assess the quality of online health information and has shown good reliability and validity in prior studies^[34].

GQS. GQS is a 5-point global rating of overall usefulness and quality, ranging from 1 (very poor) to 5 (excellent)^[35].

JAMA. The JAMA benchmark criteria are widely used to assess the quality and credibility of online health information. This tool evaluates four domains—authorship, attribution, disclosure, and currency—to reflect source transparency, traceability, and reporting standards. Total scores range from 0 to 4, with 1 point assigned for each criterion met; higher scores indicate more complete reporting and greater credibility^[36].

DISCERN. This study used the first 15 DISCERN items (Q1-Q15) to assess health information quality. Developed by Charnock et al., DISCERN is primarily used to evaluate written health information on treatment choices and has demonstrated good reliability and validity^[37]. Q1-Q8 assess reliability, whereas Q9-Q15 assess the quality of information on treatment options. Each item is scored from 1 to 5, with higher scores indicating better quality. The official Q1-Q15 version was used; Q16, the overall quality rating, was not included, to capture reliability and treatment-related content quality separately.

AIRM. AIRM is a 5-point global impression scale for evaluating the quality of AI-generated health education or patient education responses. It assesses suitability for general readers, consistency with current evidence and mainstream consensus, clarity of grammar and syntax, and content completeness. Higher scores indicate clearer, more complete responses that are more consistent with current evidence and clinical consensus. AIRM was first introduced by Anastasio et al. in a study of frequently asked patient questions in foot and ankle surgery and was subsequently applied in other patient education settings to evaluate the quality and usability of LLM responses, with some later studies reporting good interrater agreement^[38–39].

Readability Measures

Seven established English readability formulas were used to assess each response: ARI, CLI, FKGL, FRES, GFOG, Linsear Write, and SMOG^[40–45]. Except for FRES, these indices can generally be interpreted as approximate grade-level requirements, with higher values indicating greater textual complexity; for FRES, higher scores indicate easier readability. All indices were calculated using the same software tool to ensure consistency. Because different implementations may vary in the calculation and scaling of Linsear Write, this index was used primarily for relative ranking and distributional comparison rather than for stand-alone judgment of clinical readability.

Statistical Analysis

Continuous variables are presented as medians and interquartile ranges, denoted as $M (Q_1, Q_3)$. Because each question was repeatedly submitted across 12 platforms, creating clustering at the question level, rank-sum tests were used as exploratory tools for distributional comparison rather than for confirmatory inference under strict independence assumptions. Overall differences across platforms and thematic modules were examined using the Kruskal-Wallis H test, and correlations were assessed using Spearman rank correlation. All tests were two-sided, with $\alpha = 0.05$. The analysis focused on overall distributional differences and effect directions, without making strong inferences about platform superiority.

LPA was conducted in Mplus 8.11 by fitting models with one to five classes. As the number of classes increased, the log-likelihood improved, whereas the AIC, BIC, and aBIC decreased continuously and reached their lowest values for the five-class model. The BLRT was significant for the two- to five-class models, and entropy was highest for the five-class model (0.901). Based on the information criteria, BLRT, and classification accuracy, the five-class model was selected.

To characterize conditional dependence among quality dimensions in LLM-generated patient education texts for hypertension with CKD, an exploratory partial correlation network was estimated using a regularized Gaussian graphical model. Based on the observed results, the network included nine nodes: PEMAT-P understandability, PEMAT-P actionability, EQIP-36, GQS, AIRM, JAMA, DISCERN (Q1-Q8), DISCERN (Q9-Q15), and SMOG. The network was estimated in R 4.5.2 using EBICglasso, and a nonparanormal transformation was applied to improve robustness to skewed distributions and tied values. Edges represent the direction and strength of conditional correlations after adjustment for all other nodes. Stability of edge weights and centrality was assessed using 2,000 bootstrap resamples. Because network structure is sensitive to node selection and sample composition, these findings were interpreted primarily as exploratory structural evidence.

Conventional statistical analyses were performed in IBM SPSS Statistics 25, LPA in Mplus 8.11, network analysis in R 4.5.2, and visualization in GraphPad Prism 9 and related network packages.

Results

Text Characteristics

Characteristics of Patient Education Texts Generated by Different LLM Platforms for Hypertension With CKD

Exploratory Kruskal-Wallis comparisons showed distributional differences across the 12 LLM platforms in text quality, readability, actionability, and length (Table 2). Overall, the median PEMAT-P understandability, actionability, and total scores were 69.20, 20.00, and 55.60, respectively. The median EQIP-36 score was 35.71; the median DISCERN (Q1-Q8), DISCERN (Q9-Q15), and total DISCERN scores were 26.00, 11.00, and 38.00; and the median JAMA score was 1.00. Thus, although understandability was relatively better, the absolute levels of actionability and source transparency remained low. Descriptively, ERNIE Bot 5.0 Preview showed relatively high median PEMAT-P understandability (80.95) and total scores (64.70). DeepSeek-V3.2 also showed high understandability (76.90), together with lower values on several readability indices and a higher FRES (33.25). In contrast, Perplexity had higher ARI, FKGL, GFOG, and SMOG scores, indicating a greater English reading burden. Text length also varied substantially across platforms, with Ant Afu generating the longest texts and Qwen 3.5-plus the shortest.

Variables	Total (n = 240)	Baichuan Patient Edition (n = 20)	Baichuan Doctor Edition (n = 20)	Tencent Yuanbao (n = 20)	Ant Afu (n = 20)	ChatGPT 5.1- Thinking (n = 20)	ChatGPT 5.2- Thinking (n = 20)
PEMAT-P Understandability, M (Q ₁ , Q ₃)	69.20 (53.80, 76.90)	50.00 (38.50,61.50)	46.20 (38.50,55.72)	59.90 (50.00,69.20)	76.90 (69.20,85.95)	76.90 (69.20,77.68)	69.20 (61.50,76.90)
PEMAT-P Actionability, M (Q ₁ , Q ₃)	20.00 (0.00, 40.00)	20.00 (0.00,20.00)	40.00 (40.00,40.00)	20.00 (0.00,28.75)	20.00 (0.00,20.00)	40.00 (0.00,45.00)	20.00 (0.00,4
PEMAT-P score, M (Q ₁ , Q ₃)	55.60 (44.40, 61.10)	41.65 (33.30,50.00)	44.40 (38.90,55.60)	45.75 (35.60,56.40)	61.10 (54.93,67.23)	61.10 (55.60,68.08)	55.60 (50.00,62.00)
EQIP 36 score, M (Q ₁ , Q ₃)	35.71 (28.79, 42.65)	33.09 (27.21,38.86)	41.97 (39.09,43.17)	32.53 (27.50,39.11)	32.68 (29.70,36.95)	41.18 (33.96,50.00)	37.50 (29.70,42.12)
DISCERN-(Q1-Q8) Total, M (Q ₁ , Q ₃)	26.00 (24.00, 29.00)	26.00 (25.00,28.00)	26.00 (24.50,27.00)	23.50 (21.75,29.00)	31.00 (30.00,31.00)	26.00 (22.75,28.00)	27.50 (22.00,30.25)
DISCERN-(Q9-Q15) Total, M (Q ₁ , Q ₃)	11.00 (7.00, 19.00)	11.00 (7.00,20.25)	13.00 (9.50,21.25)	11.50 (7.00,21.00)	7.00 (7.00,17.00)	15.00 (8.00,18.50)	14.50 (8.00,1
DISCERN-(Q1-Q15) Total, M (Q ₁ , Q ₃)	38.00 (33.00, 47.00)	35.50 (33.75,47.50)	38.50 (33.50,47.75)	36.00 (29.75,49.25)	38.50 (37.00,48.00)	41.00 (32.75,46.00)	41.50 (32.00,48.25)
Words, M (Q ₁ , Q ₃)	460.50 (236.75, 714.25)	452.50 (386.00,514.25)	599.50 (542.00,663.00)	335.50 (113.50,777.25)	1094.50 (913.75,1419.75)	641.50 (382.00,834.75)	403.50 (358.25,614.7
Sentences, M (Q ₁ , Q ₃)	22.50 (12.00, 38.00)	20.50 (16.50,25.25)	29.50 (25.00,31.25)	10.50 (5.00,42.00)	68.50 (54.25,78.50)	31.00 (14.25,43.50)	18.00 (12.75,28.25)
Characters, M (Q ₁ , Q ₃)	2888.50 (1651.00, 4460.25)	2873.00 (2447.50,3367.75)	3758.50 (3499.75,4364.25)	2014.00 (690.25,5084.00)	7562.50 (6466.50,9600.25)	4189.50 (2445.00,5515.25)	2557.00 (2365.25,402
Letters, M (Q ₁ , Q ₃)	2642.00 (1407.50, 4078.25)	2595.00 (2211.75,3049.25)	3408.50 (3146.00,3847.75)	1887.50 (671.75,4678.75)	6848.50 (5782.25,8620.50)	3591.00 (2182.50,4809.75)	2303.00 (2020.00,349
Syllables, M (Q ₁ , Q ₃)	887.50 (466.50, 1358.00)	853.50 (757.75,1026.25)	1149.00 (1048.50,1289.75)	624.00 (229.50,1565.75)	2337.00 (1962.00,2871.00)	1192.50 (738.25,1606.50)	774.00 (678.25,1188
Complex Words, M (Q ₁ , Q ₃)	126.50 (65.75, 202.25)	119.50 (107.50,149.75)	162.00 (153.75,180.50)	87.50 (36.00,234.00)	374.50 (307.50,433.75)	175.50 (106.50,240.25)	122.50 (87.25,154.50)
ARI, M (Q ₁ , Q ₃)	18.70 (16.50, 20.46)	19.62 (19.09,20.69)	19.06 (18.29,19.99)	18.41 (17.16,19.92)	20.18 (19.33,20.70)	19.26 (18.30,22.59)	19.40 (17.76,23.31)
CLI, M (Q ₁ , Q ₃)	16.82 (15.29, 18.19)	16.89 (16.35,17.76)	16.35 (15.43,17.30)	17.48 (16.25,18.50)	18.36 (17.99,19.29)	16.07 (13.79,17.59)	16.48 (14.71,17.23)
FKGL, M (Q ₁ , Q ₃)	15.14 (13.88, 16.44)	16.01 (14.78,16.48)	15.28 (14.88,16.18)	16.26 (14.65,17.85)	15.85 (15.10,16.61)	14.46 (14.06,16.51)	15.88 (14.22,17.85)
FRES, M (Q ₁ , Q ₃)	20.53 (12.59, 29.32)	20.17 (14.18,24.86)	21.73 (19.92,27.00)	16.37 (9.66,22.30)	11.95 (8.64,18.18)	24.19 (15.96,34.04)	20.87 (13.86,31.17)
GFOG, M (Q ₁ , Q ₃)	19.16 (17.66, 20.83)	20.66 (19.00,20.95)	19.31 (18.81,20.51)	20.69 (18.43,22.22)	19.58 (18.98,20.61)	18.45 (17.86,20.64)	19.23 (17.80,21.60)
Linsear, M (Q ₁ , Q ₃)	80.00 (75.00, 84.00)	81.50 (75.25,84.00)	80.00 (76.75,82.25)	82.00 (78.00,86.12)	86.50 (81.75,90.25)	75.50 (69.50,77.25)	76.50 (74.75,84.75)
#. exploratory Kruskal-Wallis test; M: median; Q ₁ : 1st quartile; Q ₃ : 3rd quartile							

Variables	Total (n = 240)	Baichuan Patient Edition (n = 20)	Baichuan Doctor Edition (n = 20)	Tencent Yuanbao (n = 20)	Ant Afu (n = 20)	ChatGPT 5.1- Thinking (n = 20)	ChatGPT 5.2- Thinking (n = 20)
SMOG, M (Q ₁ , Q ₃)	16.36 (15.29, 17.69)	17.48 (16.60,18.00)	16.86 (16.35,17.57)	17.59 (15.70,18.80)	16.44 (15.98,17.16)	16.12 (15.36,17.84)	16.61 (15.38,18.55)
GQS score, M (Q ₁ , Q ₃)	4.00 (4.00, 4.00)	4.00 (3.75,4.00)	4.00 (4.00,4.00)	4.00 (3.00,4.00)	4.00 (4.00,4.00)	4.00 (4.00,4.25)	4.00 (4.00,4.2
AIRM, M (Q ₁ , Q ₃)	4.00 (3.00, 4.00)	4.00 (3.00,4.00)	3.00 (3.00,3.00)	4.00 (3.00,4.00)	4.00 (3.00,4.00)	4.00 (3.75,4.00)	4.00 (3.00,4.0
JAMA score, M (Q ₁ , Q ₃)	1.00 (0.00, 1.00)	1.00 (1.00,1.00)	1.00 (1.00,1.00)	0.00 (0.00,0.00)	1.00 (1.00,1.00)	0.00 (0.00,1.00)	1.00 (0.00,1.0
#: exploratory Kruskal-Wallis test; M: median; Q ₁ : 1st quartile; Q ₃ : 3rd quartile							

Characteristics of Patient Education Texts Generated Across Thematic Modules for Hypertension With CKD

When grouped by content domain, LLM-generated texts showed overall distributional differences in quality, readability, actionability, and length (Table 3). Descriptively, the Lifestyle Intervention and Long-Term Management module had the highest overall scores for EQUIP-36 [39.15 (34.29, 43.36)], DISCERN (Q1-Q8) [29.00 (26.75, 30.00)], DISCERN (Q9-Q15) [18.00 (13.00, 21.00)], and total DISCERN [47.00 (40.75, 50.00)], suggesting relatively better overall information quality and reliability. The Risk Alerts and Special Considerations module showed higher PEMAT-P understandability [69.20 (61.50, 78.00)] and total PEMAT-P [59.10 (52.17, 66.70)] scores. In contrast, Foundational Disease Knowledge and Screening, Diagnosis, and Stratification tended to score lower on PEMAT-P actionability and total DISCERN. For readability, differences in ARI, CLI, FRES, and Linsear Write were statistically significant; notably, Screening, Diagnosis, and Stratification had a higher FRES, indicating easier reading. By contrast, FKGL, GFOG, SMOG, GQS, and JAMA did not differ significantly across modules; therefore, thematic comparisons should be interpreted cautiously.

Table 3
Characteristics of Patient Education Texts for Hypertension With CKD Generated by LLM Platforms Across Thematic Modules

Variables	Total (n = 240)	Treatment Targets and Pharmacologic Management (n = 48)	Foundational Disease Knowledge (n = 48)	Risk Alerts and Special Considerations (n = 48)	Lifestyle Intervention and Long-Term Management (n = 48)	Screening, Diagnosis, and Stratification (n = 48)	Statistic
PEMAT-P Understandability, M (Q ₁ , Q ₃)	69.20 (53.80, 76.90)	61.50 (52.47,69.20)	61.50 (46.70,76.90)	69.20 (61.50,78.00)	69.20 (57.17,76.90)	64.10 (53.67,76.90)	$\chi^2=11.55\#$
PEMAT-P Actionability, M (Q ₁ , Q ₃)	20.00 (0.00, 40.00)	20.00 (0.00,40.00)	0.00 (0.00,40.00)	20.00 (0.00,40.00)	20.00 (15.00,40.00)	0.00 (0.00,25.00)	$\chi^2=13.37\#$
PEMAT-P score, M (Q ₁ , Q ₃)	55.60 (44.40, 61.10)	50.00 (38.90,61.10)	50.00 (38.00,62.00)	59.10 (52.17,66.70)	55.60 (44.40,65.20)	50.00 (38.90,56.48)	$\chi^2=17.35\#$
EQIP 36 score, M (Q ₁ , Q ₃)	35.71 (28.79, 42.65)	38.52 (31.29,44.37)	29.29 (24.70,39.29)	35.61 (30.30,42.54)	39.15 (34.29,43.36)	33.17 (27.50,40.30)	$\chi^2=20.66\#$
DISCERN-(Q1-Q8) Total, M (Q ₁ , Q ₃)	26.00 (24.00, 29.00)	26.00 (23.00,28.00)	25.00 (23.00,26.25)	27.00 (24.75,29.00)	29.00 (26.75,30.00)	26.00 (23.00,30.00)	$\chi^2=18.64\#$
DISCERN-(Q9-Q15) Total, M (Q ₁ , Q ₃)	11.00 (7.00, 19.00)	16.50 (10.00,21.00)	7.00 (7.00,10.00)	15.50 (9.00,21.00)	18.00 (13.00,21.00)	7.00 (7.00,8.00)	$\chi^2=88.85\#$
DISCERN-(Q1-Q15) Total, M (Q ₁ , Q ₃)	38.00 (33.00, 47.00)	43.00 (34.00,48.25)	33.00 (30.75,37.00)	40.50 (36.00,49.00)	47.00 (40.75,50.00)	34.00 (30.00,38.00)	$\chi^2=65.28\#$
Words, M (Q ₁ , Q ₃)	460.50 (236.75, 714.25)	521.50 (333.50,722.50)	350.50 (157.00,709.75)	599.00 (356.75,885.00)	437.50 (255.00,641.25)	473.00 (224.25,766.00)	$\chi^2=10.15\#$
Sentences, M (Q ₁ , Q ₃)	22.50 (12.00, 38.00)	27.00 (15.75,38.00)	18.00 (6.00,32.50)	28.00 (15.00,52.50)	21.00 (12.00,35.50)	22.00 (11.75,38.25)	$\chi^2=10.29\#$
Characters, M (Q ₁ , Q ₃)	2888.50 (1651.00, 4460.25)	3383.00 (2296.75,4704.25)	2122.50 (899.00,4346.50)	3751.00 (2344.00,5497.50)	2884.00 (1692.25,4060.50)	2834.50 (1414.25,4824.25)	$\chi^2=12.65\#$
Letters, M (Q ₁ , Q ₃)	2642.00 (1407.50, 4078.25)	3018.50 (2019.00,4161.00)	1985.00 (854.00,3901.50)	3502.00 (2054.75,5012.50)	2585.00 (1544.00,3665.25)	2642.00 (1199.75,4321.75)	$\chi^2=12.74\#$
Syllables, M (Q ₁ , Q ₃)	887.50 (466.50, 1358.00)	1009.00 (705.75,1436.75)	673.50 (283.25,1338.75)	1153.00 (671.00,1673.75)	878.50 (511.00,1249.75)	886.00 (391.00,1461.75)	$\chi^2=11.92\#$
Complex Words, M (Q ₁ , Q ₃)	126.50 (65.75, 202.25)	144.50 (108.25,206.00)	93.00 (40.00,192.50)	151.00 (100.75,235.75)	131.00 (69.50,175.75)	110.50 (50.50,206.25)	$\chi^2=12.93\#$
ARI, M (Q ₁ , Q ₃)	18.70 (16.50, 20.46)	19.27 (16.85,20.13)	17.86 (16.32,19.15)	19.13 (17.41,20.66)	19.02 (16.89,21.01)	17.93 (15.57,19.79)	$\chi^2=11.40\#$
CLI, M (Q ₁ , Q ₃)	16.82 (15.29, 18.19)	17.20 (16.07,18.21)	15.46 (14.00,17.54)	17.56 (16.96,19.29)	17.14 (16.04,18.46)	15.42 (13.68,17.34)	$\chi^2=41.56\#$
FKGL, M (Q ₁ , Q ₃)	15.14 (13.88, 16.44)	15.75 (14.72,16.65)	14.75 (13.58,16.16)	15.28 (14.43,16.50)	15.53 (13.72,16.55)	14.34 (13.22,15.98)	$\chi^2=8.83\#$
FRES, M (Q ₁ , Q ₃)	20.53 (12.59, 29.32)	17.44 (10.08,23.13)	23.76 (17.23,36.37)	17.61 (11.53,25.96)	18.62 (12.76,25.96)	26.66 (15.25,35.26)	$\chi^2=21.14\#$
GFOG, M (Q ₁ , Q ₃)	19.16 (17.66, 20.83)	19.59 (18.26,20.60)	18.89 (17.46,20.90)	19.18 (17.96,20.68)	19.57 (17.87,20.87)	18.55 (16.84,20.94)	$\chi^2=3.61\#$
#: exploratory Kruskal-Wallis test							
M: median; Q ₁ : 1st quartile; Q ₃ : 3rd quartile							

Variables	Total (n = 240)	Treatment Targets and Pharmacologic Management (n = 48)	Foundational Disease Knowledge (n = 48)	Risk Alerts and Special Considerations (n = 48)	Lifestyle Intervention and Long-Term Management (n = 48)	Screening, Diagnosis, and Stratification (n = 48)	Statistic
Linsear, M (Q ₁ , Q ₃)	80.00 (75.00, 84.00)	82.50 (79.75,87.25)	76.00 (72.75,82.00)	80.00 (75.00,85.00)	81.50 (75.75,85.25)	75.00 (70.00,80.00)	$\chi^2=36.18\#$
SMOG, M (Q ₁ , Q ₃)	16.36 (15.29, 17.69)	16.46 (15.45,17.58)	16.26 (15.31,17.83)	16.34 (15.43,17.40)	16.61 (15.19,17.76)	16.07 (14.65,17.37)	$\chi^2=1.87\#$
GQS score, M (Q ₁ , Q ₃)	4.00 (4.00, 4.00)	4.00 (4.00,4.00)	4.00 (4.00,4.00)	4.00 (4.00,5.00)	4.00 (4.00,4.00)	4.00 (3.00,4.00)	$\chi^2=9.30\#$
AIRM, M (Q ₁ , Q ₃)	4.00 (3.00, 4.00)	4.00 (3.00,4.00)	4.00 (4.00,4.00)	4.00 (3.00,4.00)	4.00 (3.00,4.00)	4.00 (3.00,4.00)	$\chi^2=17.79\#$
JAMA score, M (Q ₁ , Q ₃)	1.00 (0.00, 1.00)	0.00 (0.00,1.00)	0.00 (0.00,1.00)	1.00 (0.00,1.00)	1.00 (0.00,1.00)	0.50 (0.00,1.00)	$\chi^2=4.35\#$
#: exploratory Kruskal-Wallis test							
M: median; Q ₁ : 1st quartile; Q ₃ : 3rd quartile							

Text Quality

Ridge plots showed marked heterogeneity in the distribution of text quality across platforms (Fig. 1). Overall, ChatGPT 5.1-Thinking, ChatGPT 5.2-Thinking, DeepSeek-V3.2, ERNIE Bot 5.0 Preview, and Gemini 3-pro showed right-shifted distributions for PEMAT-P understandability, total PEMAT-P, EQIP-36, and total DISCERN, indicating relatively better overall quality. In contrast, both Baichuan versions, Doubao, and Qwen 3.5-plus showed relatively left-shifted distributions. Notably, PEMAT-P actionability remained low and widely dispersed across nearly all platforms, suggesting that being understandable does not necessarily mean being actionable. By comparison, GQS and AIRM were more tightly distributed, with most platforms clustering in the mid-to-high score range, whereas JAMA showed greater dispersion, indicating inconsistent source disclosure and transparency across platforms.

Half-violin plots showed heterogeneous distributions of text quality across thematic modules (Fig. 2). Overall, Lifestyle Intervention and Long-Term Management showed higher total DISCERN [47.00 (40.75, 50.00)], EQIP-36 [39.15 (34.29, 43.36)], and AIRM scores, indicating relatively better overall quality and information reliability. Risk Alerts and Special Considerations performed better on PEMAT-P understandability [69.20 (61.50, 78.00)] and total PEMAT-P [59.10 (52.17, 66.70)]. Treatment Targets and Pharmacologic Management also ranked relatively high on EQIP-36 and total DISCERN. In contrast, Foundational Disease Knowledge and Screening, Diagnosis, and Stratification tended to score lower on PEMAT-P actionability and total DISCERN, suggesting that these themes were more likely to provide adequate explanation but limited support for action.

Text Readability

Radar plots showed marked heterogeneity in readability across thematic modules and platforms (Fig. 3). Across content domains, Risk Alerts and Special Considerations and Lifestyle Intervention and Long-Term Management showed higher values on difficulty indices, including ARI, CLI, FKGL, GFOG, and SMOG, whereas Screening, Diagnosis, and Stratification showed lower values and a higher FRES, indicating easier readability. Across platforms, Perplexity showed consistently higher ARI, FKGL, GFOG, and SMOG scores, indicating the greatest English reading difficulty. Ant Afu also showed relatively high complexity, with higher CLI and Linsear Write scores and a lower FRES. By contrast, DeepSeek-V3.2 showed lower values across multiple difficulty indices and the highest FRES, with ERNIE Bot 5.0 Preview also performing relatively well. Importantly, despite these relative differences, the overall median FRES was only 20.53 and the median SMOG was 16.36, indicating that the texts remained difficult in absolute terms and were not well suited for direct reading by the general patient population. Because Linsear Write values were substantially higher than those of other grade-level formulas, this index is better interpreted as a supplementary relative measure.

Correlation Analysis

The correlation heatmap showed clear clustering among the evaluation metrics (Fig. 4). Within the quality domain, PEMAT-P understandability was strongly correlated with total PEMAT-P ($r = 0.89$). DISCERN (Q9-Q15) showed the strongest correlation with total DISCERN ($r = 0.91$), and DISCERN (Q1-Q8) was also strongly correlated with total DISCERN ($r = 0.71$). EQIP-36 showed moderately strong positive correlations with DISCERN (Q9-Q15) and total DISCERN ($r = 0.56$ and 0.59 , respectively). Text length measures were highly correlated with one another, with words, sentences, characters, letters, syllables, and complex words all showing high pairwise correlations ($r = 0.92$ - 1.00). Among readability formulas, ARI, FKGL, GFOG, and SMOG were strongly positively correlated, whereas FRES was significantly negatively correlated with these indices ($r = -0.63$ to -0.79), indicating good consistency across readability measures. Notably, PEMAT-P understandability, total PEMAT-P, and GQS were generally negatively correlated with ARI, FKGL, GFOG, and SMOG, but weakly positively correlated with FRES, suggesting that more readable texts tended to have higher understandability and subjective overall quality. JAMA was positively correlated with ARI, FKGL, GFOG, and SMOG, but negatively correlated with PEMAT-P understandability, GQS, and AIRM. Where these directions differ from the partial correlation edges observed in the subsequent network, they should be interpreted as reflecting differences between marginal correlations and conditional partial correlations rather than contradictory findings.

Network Analysis

Network Structure

The exploratory network plot showed a relatively dense conditional dependence structure among the nine nodes, with both positive and negative associations (Fig. 5). Overall, the strongest positive edge was observed between JAMA and DISCERN (Q1-Q8), suggesting strong concordance between credibility-related features—such as source transparency, authorship, and citation standards—and the basic reliability captured by the first part of DISCERN. Other relatively strong positive edges were observed between EQIP-36 and DISCERN (Q9-Q15), GQS and DISCERN (Q1-Q8), and EQIP-36 and PEMAT-P actionability, indicating coordinated variation among overall material quality, subjective global quality, and actionability. Negative edges were concentrated between JAMA and SMOG, JAMA and PEMAT-P understandability, and AIRM and PEMAT-P actionability, suggesting that credibility, understandability, actionability, and readability did not increase in parallel across platforms and may instead reflect structural trade-offs. Given the dependence of network structure on node selection and sample composition, these findings should be interpreted as exploratory structural evidence.

Centrality Analysis

Centrality analysis showed differences in the relative importance of nodes within the network (Fig. 6). Using strength as the primary index, JAMA and DISCERN (Q1-Q8) had the highest centrality, indicating the strongest overall connectivity with other nodes and identifying them as the core metrics in the network. GQS and EQIP-36 ranked next and also showed relatively high connectivity. By contrast, SMOG had the lowest strength, suggesting a more peripheral position in the overall network. These findings indicate that, within the current indicator framework, source transparency and basic reliability may play a more central role in linking other quality dimensions.

Stability Analysis

The stability of network centrality was assessed using a case-dropping bootstrap with 2,000 iterations (Fig. 7). The correlation stability coefficient for strength was 0.71, exceeding the commonly recommended threshold of 0.50, indicating good stability of the estimated centrality under the current sample and node specification. In addition, across different sampling proportions, the mean correlation between node centrality estimates and the original-sample results remained high, with limited fluctuation overall. This stability supports the robustness of the current network estimation, but does not imply causal interpretability of the observed network relationships.

Latent Profile Analysis

Model Fit and Class Enumeration

The LPA results showed that as the number of classes increased from 1 to 5, the AIC, BIC, and aBIC decreased continuously, indicating progressive improvement in model fit (Table 4). Except for the 1-class model, the BLRT was significant for all 2- to 5-class models (all $P = 0.01$). Entropy increased from 0.769 for the 2-class model to 0.901 for the 5-class model, indicating improved classification accuracy. Based on the information criteria, BLRT, and entropy, the 5-class model provided the best overall balance and was therefore selected as the final model.

Table 4 Fit indices for latent profile analysis models					
Class	AIC	BIC	aBIC	Entropy	BLRT p-value
1	4104.53	4146.298	4108.261	1	NA
2	3930.329	3996.461	3936.235	0.769	0.01
3	3882.371	3972.867	3890.454	0.771	0.01
4	3381.398	3496.259	3391.657	0.881	0.01
5	3360.259	3499.485	3372.694	0.901	0.01

Phenotypic Definition of the Latent Classes

The standardized mean profile plot of the 5-class model showed marked heterogeneity across latent classes in DISCERN, EQIP-36, JAMA, PEMAT-P actionability, PEMAT-P understandability, and SMOG. Class 1 scored below the overall mean on DISCERN, EQIP-36, JAMA, and PEMAT-P actionability, with PEMAT-P understandability near the mean and slightly lower SMOG, and was therefore defined as the overall low-quality profile. Class 2 showed higher EQIP-36, JAMA, and PEMAT-P actionability, but lower DISCERN, near-average PEMAT-P understandability, and lower SMOG, and was defined as the high-actionability/heterogeneous-quality profile. Class 3 had the highest or near-highest DISCERN and EQIP-36 scores, with both PEMAT-P dimensions also above average, but the lowest JAMA and lowest SMOG, and was defined as the high-quality, readable profile with limited transparency. Class 4 scored above average on both DISCERN and JAMA, with moderately elevated EQIP-36 and PEMAT-P actionability, near-average PEMAT-P understandability, and slightly higher SMOG, and was defined as the comprehensively high-quality profile. Class 5 showed relatively high JAMA, but below-average DISCERN, EQIP-36, and both PEMAT-P dimensions, together with the highest SMOG, and was defined as the low-quality, difficult-to-read profile. These classes should be interpreted as statistical phenotypes derived from the current indicators rather than fixed natural subtypes.

Discussion

A Dissociated Structure in LLM-Based Patient Education: Improved Understandability but Lagging Actionability and Transparency

This study showed that, in patient education texts for hypertension with CKD, LLM outputs did not demonstrate synchronized performance across understandability, actionability, transparency, and overall quality. Rather than reflecting a unidimensional gradient of overall superiority, their performance was characterized by a multidimensional pattern. Most platforms were capable of producing fluent and structurally coherent explanatory texts, suggesting advantages in linguistic simplification and information reorganization. However, these strengths were largely restricted to readability and did not consistently extend to concrete behavioral guidance, disclosure of supporting evidence, or indicators of credibility. Will et al. reported that LLMs can improve the readability of patient education materials, although performance varies substantially across platforms and outputs still require human review^[10]. Clackett et al. similarly found that LLM-based rewriting may improve readability but cannot replace clinical experts in ensuring accuracy and safety^[11]. In clinical tasks requiring contextual judgment and guideline translation, current models still underperform experts in accuracy, safety, and guideline concordance^[13, 34]. Because patient education materials have long carried an excessive reading burden, LLMs may still have practical value for language simplification^[8]. Taken together, these findings suggest that LLMs are better suited to the initial transcription, compression, and reorganization of complex medical information than to the independent generation of high-quality patient education content. Without clinical review and risk calibration, linguistic fluency alone cannot ensure safety or appropriateness.

In a Comorbidity Setting, Limited Actionability May Pose Greater Clinical Risk Than Unclear Wording

For patient education in hypertension with CKD, the key task is to translate evidence-based recommendations into concrete, executable behaviors rather than merely explain disease concepts. The 2021 KDIGO blood pressure guideline emphasizes standardized blood pressure measurement, home blood pressure monitoring, and treatment targets, while the 2024 KDIGO CKD guideline further specifies GFR and albuminuria assessment, risk stratification, and recognition of complications^[46–47]. These features make patient education in this setting numerically explicit, monitoring-intensive, and exception-sensitive. If a text remains limited to general advice such as low-salt diet and regular medication use, without specifying follow-up timing, abnormality recognition, or precautions related to kidney function stratification, patients may understand the content yet still fail to implement appropriate management^[46–47]. Understandability does not equal actionability, and clear wording does not necessarily translate into behavioral execution. Previous studies have linked limited health literacy to poorer self-management, worse outcomes, and higher health care utilization^[48]. Boonstra et al. further showed that patients with CKD and low health literacy face greater barriers to disease understanding, active participation, and behavior change^[49]. Patient education should therefore move beyond lowering reading difficulty alone and toward constructing safe, specific behavioral guidance. Without clear action instructions, its clinical utility remains limited.

Exploratory Network Analysis Suggests That Source Transparency and Basic Reliability Are Key Nodes

The network analysis showed that source disclosure, traceability of evidence, and basic reliability occupied central positions in the quality network, indicating that transparency is a structural determinant of patient education quality rather than a secondary attribute. Readable text is not necessarily trustworthy. Without authorship, update information, and evidence sources, even concise and accessible language is difficult to regard as truly useful educational material^[34]. This also helps explain why improving expression alone is insufficient to raise overall educational quality. Walker et al. identified inadequate evidence disclosure as a major weakness of LLM-based patient education^[34]. A systematic review by Huo et al. further showed substantial heterogeneity in reporting quality and safety presentation across studies of LLM-generated health advice, underscoring the need to prioritize transparency and standardized evaluation in clinical integration^[50]. For high-risk populations such as patients with hypertension and CKD, this limitation may have more direct consequences. If source traceability and guideline concordance are lacking, fluent language alone may produce texts that are easy to accept but difficult to verify. Patient-friendly language detached from evidence transparency is unlikely to improve educational effectiveness and may instead increase the risks of misinformation and inappropriate use.

Practical Implications for Intelligent Patient Education Systems for Hypertension With CKD

Based on these findings, the appropriate role of LLMs in patient education for hypertension with CKD is as tools for draft generation and structural reorganization within a supervised workflow, rather than as independent systems operating outside clinical review. A more feasible approach is to build scenario-specific prompt templates grounded in authoritative guidelines, embedding key elements such as blood pressure targets, home monitoring, and sodium restriction^[46–47]. LLMs can then perform patient-oriented language translation, after which human review should ensure consistency with evidence and clinical recommendations. This approach aligns well with the educational needs of patients with CKD. Allen et al. reported that patients with CKD expect personalized disease education and are receptive to digital formats^[51]. Boonstra et al. likewise showed that patients with low health literacy face greater barriers to disease understanding and behavior change^[49]. Future system optimization should therefore focus on improving the patient usability of evidence by increasing specificity, actionability, and traceability. Subsequent work may advance in four directions: standardized prompt templates, guideline-based retrieval augmentation, structured modules for action instructions, and pre-positioned risk information. Only by establishing a closed loop of language translation, evidence binding, and risk verification can LLM-based patient education systems meet the requirements of clinical use.

Limitations and Future Directions

Limitations

This study has several limitations. First, to ensure cross-platform comparability, all queries were submitted in English and outputs were evaluated primarily with English readability formulas. Although this design improved internal consistency, it does not fully reflect real-world interactions of Chinese-speaking patients. Therefore, the findings are more applicable to cross-platform comparisons under standardized English prompts than to the actual reading experience of Chinese patient education materials. Second, because the same questions were repeatedly submitted across platforms, the data had both question-level clustering and repeated-measures structures. Accordingly, the rank-based tests used here are better interpreted as exploratory distributional comparisons, and future studies should validate these findings using mixed-effects models or generalized estimating equations (GEE) with questions treated as a random effect. Third, this study compared the actual outputs of publicly available platforms in their default product forms. Because native retrieval functions, system prompts, length controls, and other product settings may differ across platforms, the results are more appropriately interpreted as platform-level output comparisons than as purely standardized comparisons of underlying model capability. Fourth, although this study systematically assessed understandability, actionability, information quality, source transparency, and readability, and further used network analysis and LPA to characterize structural relationships and heterogeneity across indicators, it did not establish a sentence-level gold standard for factual verification. Nor did it grade the clinical severity of errors, potential harm, or consequences of deviation from guidelines. The results therefore represent statistical characterizations of output quality rather than definitive validation of clinical safety. Fifth, both network analysis and LPA are structure-discovery methods, and their results depend on the selected indicators, sample composition, and model specification. Although the stability analysis suggested reasonable robustness of network centrality, these methods do not support causal inference, and the latent classes should not be interpreted as fixed natural subtypes. Sixth, the evaluation relied mainly on professional raters and did not include real-world usability testing by patients, family caregivers, or individuals with limited health literacy. It also did not examine whether generated texts improved disease knowledge, home blood pressure monitoring, diet and medication management, or follow-up adherence.

Future Directions

In light of these limitations, future research should move beyond cross-sectional quality comparison toward an integrated strategy combining standardized evaluation, system optimization, and real-world validation. First, a standardized evaluation framework for patient education in hypertension with CKD should be established. With model version locking, prompt disclosure, repeated sampling, and multi-round retesting, future studies should systematically assess the stability and drift of outputs across prompt templates, languages, and interaction rounds, while preferentially using mixed-effects models to account for repeated measurements at the question level and improve interpretability and comparability across studies. Second, research should shift from comparing general-purpose models to evaluating customized patient education systems. Guideline statements, medication-monitoring rules, thresholds for abnormal indicators, referral criteria, and recognition of red-flag symptoms should be incorporated into retrieval-augmented or rule-constrained systems, with LLMs used primarily for patient-friendly language generation. This approach may improve accuracy, verifiability, and risk control simultaneously. Third, future work should incorporate real Chinese-language interaction settings, adopt readability or linguistic complexity metrics more suitable for Chinese corpora, and test their effects on knowledge acquisition, risk recognition, and disease-management behaviors in real patients. Finally, multicenter real-world studies are needed to compare three approaches—human-written materials, direct output from general-purpose LLMs, and workflows combining retrieval augmentation, rule constraints, and human review—with respect to safety, acceptability, comprehension, and behavior change. Such evidence would provide stronger external validity for the regulated use of generative artificial intelligence in chronic disease patient education.

Conclusion

This study showed that, under standardized English prompts, patient education texts for hypertension with CKD generated by cross-platform LLMs exhibited marked multidimensional heterogeneity. Relative to actionability and source transparency, understandability was generally better; however, in absolute terms, the overall reading burden remained high. Most platforms produced fluent, well-structured explanatory texts, but shared limitations in evidence disclosure, content traceability, and translation of recommendations into concrete action instructions. Exploratory network analysis suggested that source transparency and basic reliability had high connectivity within the overall quality structure. LPA further indicated that platform outputs were better characterized as several statistical quality phenotypes rather than a simple high-versus-low hierarchy.

These findings are broadly consistent with previous evidence showing that LLMs can improve the wording and readability of patient education materials, but such gains do not necessarily extend to reliability, actionability, or clinical usability^[11–12]. For disease-management tasks such as hypertension with CKD, which depend heavily on guideline translation, risk stratification, and individualized judgment, improved understandability alone is insufficient to support safe and accurate self-management. At present, LLMs are better positioned as supportive tools within evidence-based patient education workflows for draft generation, language simplification, and structural reorganization, rather than as independent sources of advice without clinical review. Future research and practice should focus on source transparency, completeness of action instructions, applicability in Chinese-language settings, and safety verification mechanisms to establish more targeted multidimensional evaluation and intervention frameworks.

Declarations

Ethics approval and consent to participate

Not applicable. This study did not involve human participants, human data, or human tissue. The three cardiovascular experts involved in developing and reviewing the question list were members of the author team rather than external participants; therefore, no separate informed consent was required.

Consent for publication

Not applicable.

Availability of data and materials

The datasets generated and/or analyzed during the current study are available in the Supplementary Materials. Further inquiries can be directed to the corresponding author.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by Key Research and Development Program of Henan Province(261111313900), National Natural Science Foundation of China (82470344, 82270474), and Henan Cardiovascular Disease Center (Central China Subcenter of National Center for Cardiovascular Diseases) (2023-FZX18).

Author contributions

HM and ZW contributed equally to this work. HM and ZW conceived and designed the study, collected and curated the data, performed the statistical analyses, prepared the figures and tables, and drafted the initial manuscript. XL contributed to data curation, literature review, and manuscript revision. ZZ and ML supervised the project, provided critical intellectual input, and revised the manuscript for important content. All authors contributed to manuscript revision, read, and approved the submitted version.

Acknowledgements

Not applicable.

References

1. World Health Organization. Global report on hypertension 2025: high stakes: turning evidence into action. Geneva: World Health Organization. (2025). <https://www.who.int/publications/i/item/9789240115569> [Accessed March 30, 2026].
2. Erviti, J. et al. Blood pressure targets for hypertension in people with chronic renal disease. *Cochrane Database Syst. Rev.* **10** (10), CD008564. 10.1002/14651858.CD008564.pub3 (2024).
3. GBD 2023 Chronic Kidney Disease Collaborators. Global, regional, and national burden of chronic kidney disease in adults, 1990–2023, and its attributable risk factors: a systematic analysis for the Global Burden of Disease Study 2023. *Lancet* **406** (10518), 2461–2482. 10.1016/S0140-6736(25)01853-7 (2025).
4. SPRINT Research Group et al. Final report of a trial of intensive versus standard blood-pressure control. *N Engl. J. Med.* **384**, 1921–1930. 10.1056/NEJMoa1901281 (2021).
5. Lai, Y. et al. Global burden and risk factors of chronic kidney disease due to hypertension in adults aged 20 plus years, 1990–2021. *Front. Public. Health.* **13**, 1503837. 10.3389/fpubh.2025.1503837 (2025).
6. Yirga, G. K. et al. Chronic kidney disease among patients with hypertension in sub-Saharan Africa: a systematic review and meta-analysis. *BMC Public. Health.* **25**, 1603. 10.1186/s12889-025-22828-8 (2025).
7. Shahid, R. et al. Impact of low health literacy on patients' health outcomes: a multicenter cohort study. *BMC Health Serv. Res.* **22**, 1148. 10.1186/s12913-022-08527-9 (2022).
8. Rooney, M. K. et al. Readability of patient education materials from high-impact medical journals: a 20-year analysis. *J. Patient Exp.* **8**, 2374373521998847. 10.1177/2374373521998847 (2021).
9. Veeramani, A., Johnson, A. R., Lee, B. T. & Dowlatshahi, A. S. Readability, understandability, usability, and cultural sensitivity of online patient educational materials (PEMs) for lower extremity reconstruction: a cross-sectional study. *Plast. Surg. (Oakv).* **32** (3), 452–459. 10.1177/22925503221120548 (2024).
10. Will, J. et al. Enhancing the readability of online patient education materials using large language models: cross-sectional study. *J. Med. Internet Res.* **27**, e69955. 10.2196/69955 (2025).
11. Clackett, W., Zealley, I. A., Yang, Z., Salahia, G. & White, R. D. Better understanding: can a large language model safely improve readability of patient information leaflets in interventional radiology? *BMJ Health Care Inf.* **32** (1), e101512. 10.1136/bmjhci-2025-101512 (2025).
12. Singhal, K. et al. Toward expert-level medical question answering with large language models. *Nat. Med.* **31** (3), 943–950. 10.1038/s41591-024-03423-7 (2025).
13. Zand, J. et al. Performance of large language models in analyzing common hypertension scenarios. *Hypertension* **83** (1), 225–234. 10.1161/HYPERTENSIONAHA.125.25492 (2026).

14. Chelli, M. et al. Hallucination rates and reference accuracy of ChatGPT and Bard for systematic reviews: comparative analysis. *J. Med. Internet Res.* **26**, e53164. 10.2196/53164 (2024).
15. Asgari, E. et al. A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation. *NPJ Digit. Med.* **8**, 274. 10.1038/s41746-025-01670-7 (2025).
16. Wang, D. et al. Enhancing large language models for improved accuracy and safety in medical question answering: comparative study. *JMIR Med. Educ.* **11**, e70190. 10.2196/70190 (2025).
17. Vinufrancis, A. et al. Assessing the quality and reliability of AI-generated responses to common hypertension queries. *Cureus* **16** (8), e66041. 10.7759/cureus.66041 (2024).
18. Wang, Z., Li, X., Ma, C. & Zhang, Z. Challenges of using generative AI for patient education in chronic heart failure: an evaluation of content quality, readability, and actionability in cross-platform LLM-generated texts. *Front. Public. Health.* **14**, 1801829. 10.3389/fpubh.2026.1801829 (2026).
19. Luo, Z., Lin, C., Kim, T. H., Shin, Y. S. & Ahn, S. T. Quality and readability analysis of artificial intelligence-generated medical information related to prostate cancer: a cross-sectional study of ChatGPT and DeepSeek. *World J. Mens Health.* **43**, e70. 10.5534/wjmh.250144 (2025).
20. 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. 10.2196/55939 (2024).
21. Xinhua News Agency. Volcano Engine releases Doubao Large Model 1.6, accelerating large-scale application of AI agents [in Chinese]. (2025). <https://www.news.cn/tech/20250611/db9b3c0ef97d49058413563b1c434951/c.html> [Accessed March 30, 2026].
22. OpenAI GPT-5.1: A smarter, more conversational ChatGPT (2025). <https://openai.com/index/gpt-5-1/> [Accessed March 30, 2026].
23. OpenAI. Introducing GPT-5.2. (2025). <https://openai.com/index/introducing-gpt-5-2/> [Accessed March 30, 2026].
24. DeepSeek. DeepSeek-V3.2 Release. (2025). <https://api-docs.deepseek.com/news/news251201> [Accessed March 30, 2026].
25. Xinhua News Agency. Ant Group's AQ is upgraded to. Ant A-Fu: from medical care to health, and from AI tools to AI companions [in Chinese] (2025). <https://www.news.cn/tech/20251215/14b427d3eca34bbe9f10709f4da3542c/c.html> [Accessed March 30, 2026].
26. Wang, H. et al. ERNIE 5.0 Technical Report. arXiv [Preprint] (2026). 10.48550/arXiv.2602.04705
27. Alibaba Group. Alibaba Open-Sources Qwen3.5, A Natively Multimodal Model Built For High-Efficiency Inference. (2026). <https://www.alibabagroup.com/en-US/document-1960233590314762240> [Accessed March 30, 2026].
28. Baichuan Intelligence. Baichuan large model: aggregate world knowledge and create brilliant writing [in Chinese]. (2026). <https://www.baichuan-ai.com/> [Accessed March 30, 2026].
29. Google, A. I., for Developers. Release notes | & Gemini, A. P. I. (2026). <https://ai.google.dev/gemini-api/docs/changelog> [Accessed March 30, 2026].
30. Perplexity. What is Perplexity? <https://www.perplexity.ai/help-center/en/articles/10352155-what-is-perplexity> [Accessed March 30, 2026].
31. Tencent Yuanbao [in Chinese]. <https://yuanbao.tencent.com/> [Accessed March 30, 2026].
32. Shoemaker, S. J., Wolf, M. S. & Brach, C. Development of the Patient Education Materials Assessment Tool (PEMAT): a new measure of understandability and actionability for print and audiovisual patient information. *Patient Educ. Couns.* **96** (3), 395–403. 10.1016/j.pec.2014.05.027 (2014).
33. Charvet-Berard, A. I., Chopard, P. & Perneger, T. V. Measuring quality of patient information documents with an expanded EQIP scale. *Patient Educ. Couns.* **70** (3), 407–411. 10.1016/j.pec.2007.11.018 (2008).
34. Walker, H. L. et al. Reliability of medical information provided by ChatGPT: assessment against clinical guidelines and patient information quality instrument. *J. Med. Internet Res.* **25**, e47479. 10.2196/47479 (2023).
35. Bernard, A. et al. A systematic review of patient inflammatory bowel disease information resources on the World Wide Web. *Am. J. Gastroenterol.* **102**, 2070–2077. 10.1111/j.1572-0241.2007.01325.x (2007).
36. Silberg, W. M., Lundberg, G. D. & Musacchio, R. A. Assessing, controlling, and assuring the quality of medical information on the Internet: Caveant lector et viewer—let the reader and viewer beware. *JAMA* **277** (15), 1244–1245. 10.1001/jama.1997.03540390074039 (1997).
37. 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** (2), 105–111. 10.1136/jech.53.2.105 (1999).
38. Anastasio, A. T., Mills, F. B., Karavan, M. P. & Adams, S. B. Evaluating the quality and usability of artificial intelligence-generated responses to common patient questions in foot and ankle surgery. *Foot Ankle Orthop.* **8** (4), 24730114231209919. 10.1177/24730114231209919 (2023).
39. King, B. W. et al. Evaluating artificial intelligence-generated responses to patient questions regarding orthobiologic injections. *Orthop. J. Sports Med.* **14** (2), 23259671251414852. 10.1177/23259671251414852 (2026).
40. Zamanian, M. & Heydari, P. Readability of texts: state of the art. *Theory Pract. Lang. Stud.* **2**, 43–53. 10.4304/tpls.2.1.43 (2012).
41. DuBay, W. H. The principles of readability. Costa Mesa, CA: Impact Information (2004). <https://files.eric.ed.gov/fulltext/ED490073.pdf> [Accessed March 30, 2026].
42. Kher, A., Johnson, S. & Griffith, R. Readability assessment of online patient education material on congestive heart failure. *Adv. Prev. Med.* **2017**, 9780317. 10.1155/2017/9780317 (2017).
43. Flesch, R. A new readability yardstick. *J. Appl. Psychol.* **32** (3), 221–233. 10.1037/h0057532 (1948).
44. Kincaid, J. P., Fishburne, R. P. Jr, Rogers, R. L. & Chissom, B. S. Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy enlisted personnel. Research Branch Report 8–75. Millington, TN: Naval Technical Training Command, Research Branch (1975). <https://eric.ed.gov/?id=ED108134> [Accessed March 30, 2026].
45. McLaughlin, G. H. SMOG grading—A new readability formula. *J. Read.* **12** (8), 639–646 (1969).

46. Kidney Disease: Improving Global Outcomes (KDIGO) Blood Pressure Work Group. KDIGO 2021 clinical practice guideline for the management of blood pressure in chronic kidney disease. *Kidney Int.* **99** (3S), S1–S87. 10.1016/j.kint.2020.11.003 (2021).
47. Kidney Disease: Improving Global Outcomes (KDIGO) CKD Work Group. KDIGO 2024 clinical practice guideline for the evaluation and management of chronic kidney disease. *Kidney Int.* **105** (4S), S117–S314. 10.1016/j.kint.2023.10.018 (2024).
48. Billany, R. E. et al. Associations of health literacy with self-management behaviours and health outcomes in chronic kidney disease: a systematic review. *J. Nephrol.* **36** (5), 1267–1281. 10.1007/s40620-022-01537-0 (2023).
49. Boonstra, M. D. et al. A longitudinal qualitative study to explore and optimize self-management in mild to end stage chronic kidney disease patients with limited health literacy: perspectives of patients and health care professionals. *Patient Educ. Couns.* **105** (1), 88–104. 10.1016/j.pec.2021.05.016 (2022).
50. Huo, B. et al. Large language models for chatbot health advice studies: a systematic review. *JAMA Netw. Open.* **8** (2), e2457879. 10.1001/jamanetworkopen.2024.57879 (2025).
51. Allen, R. J. et al. Exploring patient needs and preferences in CKD education: a cross-sectional survey study. *Kidney360.* **5**(3):344–351. (2024). 10.34067/KID.0000000000000369

Figures

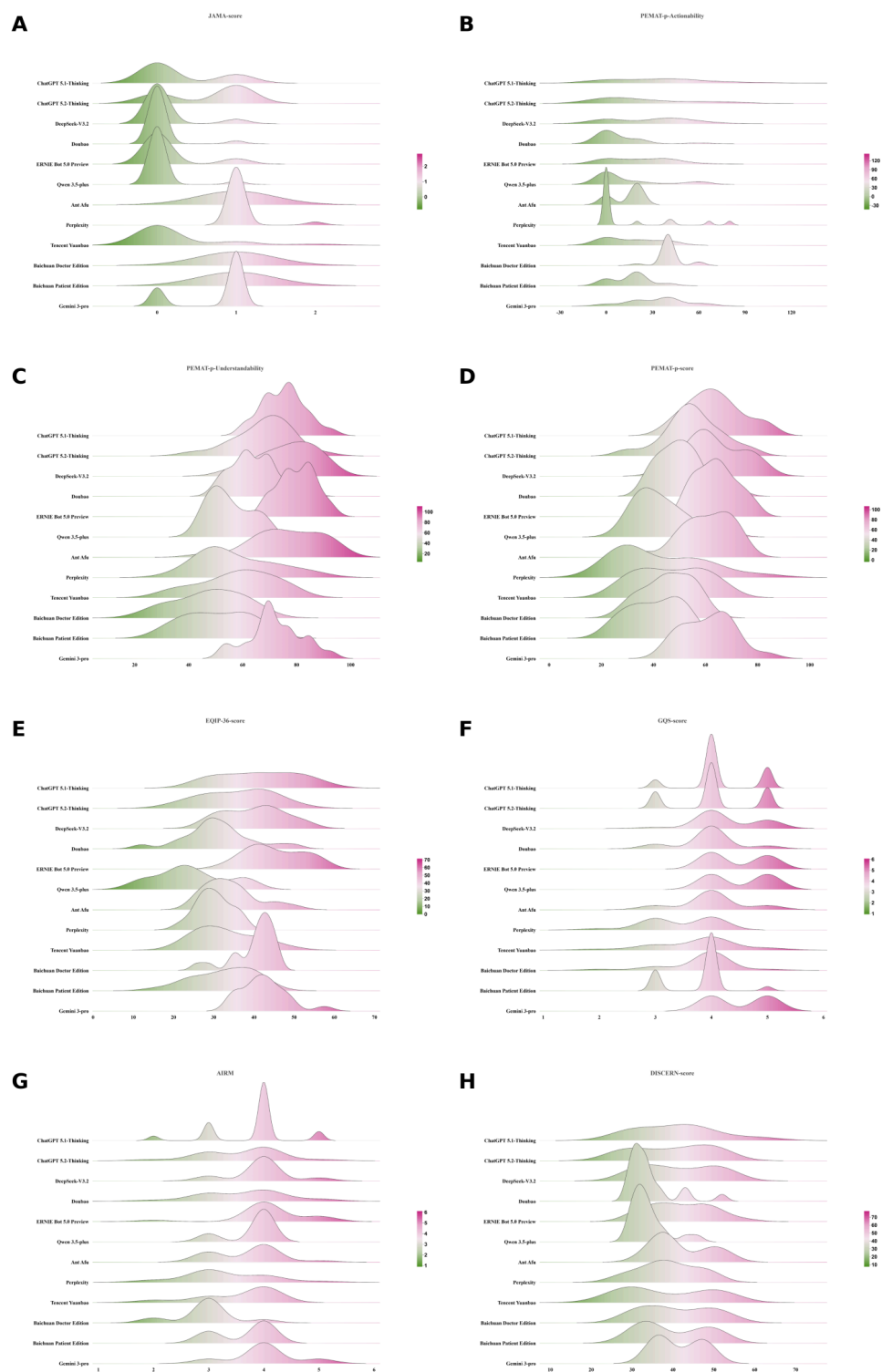


Figure 1

Distribution of actionability, information quality, and overall quality of patient education texts for hypertension with CKD generated across platforms

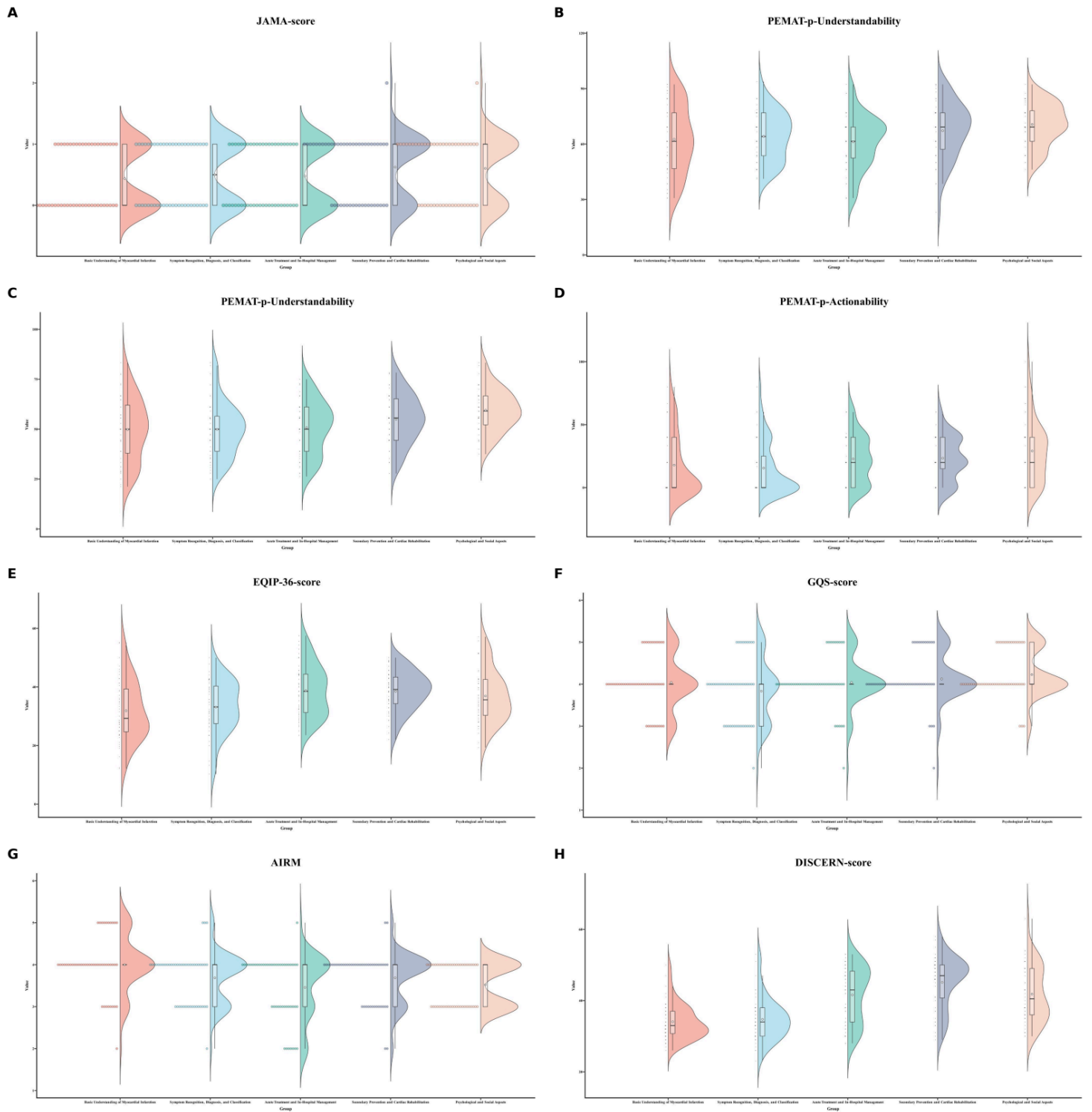
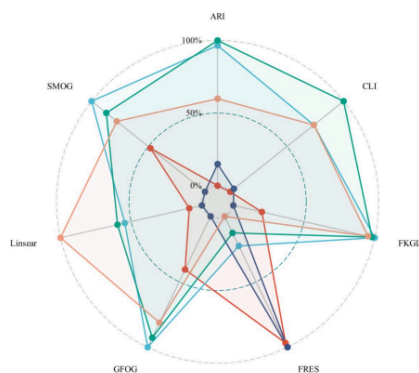


Figure 2

Distribution of actionability, information quality, and overall quality of patient education texts for hypertension with CKD across thematic modules

A



B

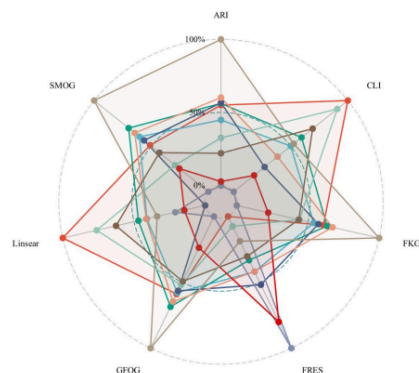


Figure 3

Readability of patient education texts for hypertension with CKD across thematic modules and platforms

	PEMAT-p-Understandability	PEMAT-p-Actionability	PEMAT-p-score	EQIP-36-score	GQS-score	AIRM	DISCERN- (Q1-Q8) score	DISCERN-(Q9-Q15) score	DISCERN-score	JAMA-score	words	sentences	characters	letters	syllables	complex_words	ARI	CLI	FKGL	FRES	GFOG	Linear	SMOG
PEMAT-p-Understandability	1	0.20	0.89	0.36	0.24	0.22	0.23	0.13	0.20	-0.18	0.33	0.39	0.31	0.33	0.32	0.30	-0.21	-0.04	-0.24	0.13	-0.25	-0.00	-0.27
PEMAT-p-Actionability	0.20	1	0.57	0.51	0.22	-0.22	0.16	0.39	0.36	0.14	0.37	0.39	0.37	0.38	0.38	0.37	-0.03	-0.00	-0.05	0.07	-0.07	0.05	-0.04
PEMAT-p-score	0.89	0.57	1	0.49	0.31	0.08	0.24	0.25	0.29	-0.10	0.41	0.47	0.40	0.41	0.40	0.39	-0.19	0.00	-0.23	0.13	-0.25	0.03	-0.25
EQIP-36-score	0.36	0.51	0.49	1	0.31	-0.02	0.38	0.58	0.60	0.13	0.52	0.54	0.52	0.53	0.53	0.52	-0.10	-0.06	-0.13	0.09	-0.14	0.10	-0.08
GQS-score	0.24	0.22	0.31	0.31	1	0.21	0.25	0.17	0.25	-0.15	0.13	0.17	0.11	0.13	0.12	0.11	-0.28	0.01	-0.22	0.13	-0.24	-0.05	-0.21
AIRM	0.22	-0.22	0.08	-0.02	0.21	1	0.00	-0.15	-0.09	-0.25	-0.08	-0.04	-0.10	-0.09	-0.09	-0.09	-0.26	-0.12	-0.19	0.11	-0.15	-0.08	-0.15
DISCERN- (Q1-Q8) score	0.23	0.16	0.24	0.38	0.25	0.00	1	0.31	0.73	0.51	0.51	0.47	0.52	0.51	0.51	0.50	0.14	0.04	-0.01	0.01	-0.03	0.13	0.01
DISCERN-(Q9-Q15) score	0.13	0.39	0.25	0.58	0.17	-0.15	0.31	1	0.86	0.13	0.35	0.34	0.37	0.37	0.37	0.37	0.09	0.10	0.06	-0.07	0.01	0.22	0.05
DISCERN-score	0.20	0.36	0.29	0.60	0.25	-0.09	0.73	0.86	1	0.34	0.52	0.48	0.53	0.53	0.53	0.53	0.14	0.09	0.05	-0.05	0.00	0.22	0.05
JAMA-score	-0.18	0.14	-0.10	0.13	-0.15	-0.25	0.51	0.13	0.34	1	0.31	0.20	0.33	0.31	0.31	0.32	0.47	0.11	0.27	-0.11	0.24	0.08	0.32
words	0.33	0.37	0.41	0.52	0.13	-0.08	0.51	0.35	0.52	0.31	1	0.94	0.99	1.00	0.99	0.97	0.03	-0.03	-0.09	0.06	-0.11	0.09	-0.04
sentences	0.39	0.39	0.47	0.54	0.17	-0.04	0.47	0.34	0.48	0.20	0.94	1	0.94	0.94	0.94	0.93	-0.19	0.03	-0.29	0.09	-0.29	0.16	-0.29
characters	0.31	0.37	0.40	0.52	0.11	-0.10	0.52	0.37	0.53	0.33	0.99	0.94	1	1.00	1.00	0.98	0.08	0.04	-0.05	-0.01	-0.06	0.14	-0.01
letters	0.33	0.38	0.41	0.53	0.13	-0.09	0.51	0.37	0.53	0.31	1.00	0.94	1.00	1	1.00	0.99	0.05	0.05	-0.05	-0.01	-0.07	0.15	-0.02
syllables	0.32	0.38	0.40	0.53	0.12	-0.09	0.51	0.37	0.53	0.31	0.99	0.94	1.00	1.00	1	0.99	0.05	0.05	-0.05	-0.02	-0.07	0.16	-0.02
complex_words	0.30	0.37	0.39	0.52	0.11	-0.09	0.50	0.37	0.53	0.32	0.97	0.93	0.98	0.99	0.99	1	0.09	0.15	0.01	-0.12	0.01	0.25	0.04
ARI	-0.21	-0.03	-0.19	-0.10	-0.28	-0.26	0.14	0.09	0.14	0.47	0.03	-0.19	0.08	0.05	0.05	0.09	1	0.44	0.87	-0.63	0.82	0.26	0.83
CLI	-0.04	-0.00	0.00	-0.06	0.01	-0.12	0.04	0.10	0.09	0.11	-0.03	0.03	0.04	0.05	0.05	0.15	0.44	1	0.52	-0.86	0.53	0.71	0.35
FKGL	-0.24	-0.05	-0.23	-0.13	-0.22	-0.19	-0.01	0.06	0.05	0.27	-0.09	-0.29	-0.05	-0.05	-0.05	0.01	0.87	0.52	1	-0.79	0.96	0.40	0.93
FRES	0.13	0.07	0.13	0.09	0.13	0.11	0.01	-0.07	-0.05	-0.11	0.06	0.09	-0.01	-0.01	-0.02	-0.12	-0.63	-0.86	-0.79	1	-0.81	-0.71	-0.63
GFOG	-0.25	-0.07	-0.25	-0.14	-0.24	-0.15	-0.03	0.01	0.00	0.24	-0.11	-0.29	-0.06	-0.07	-0.07	0.01	0.82	0.53	0.96	-0.81	1	0.45	0.94
Linear	-0.00	0.05	0.03	0.10	-0.05	-0.08	0.13	0.22	0.22	0.08	0.09	0.16	0.14	0.15	0.16	0.25	0.26	0.71	0.40	-0.71	0.45	1	0.28
SMOG	-0.27	-0.04	-0.25	-0.08	-0.21	-0.15	0.01	0.05	0.05	0.32	-0.04	-0.29	-0.01	-0.02	-0.02	0.04	0.83	0.35	0.93	-0.63	0.94	0.28	1

Figure 4

Correlation analysis of quality, readability, and actionability of patient education texts for hypertension with CKD generated by cross-platform LLMs

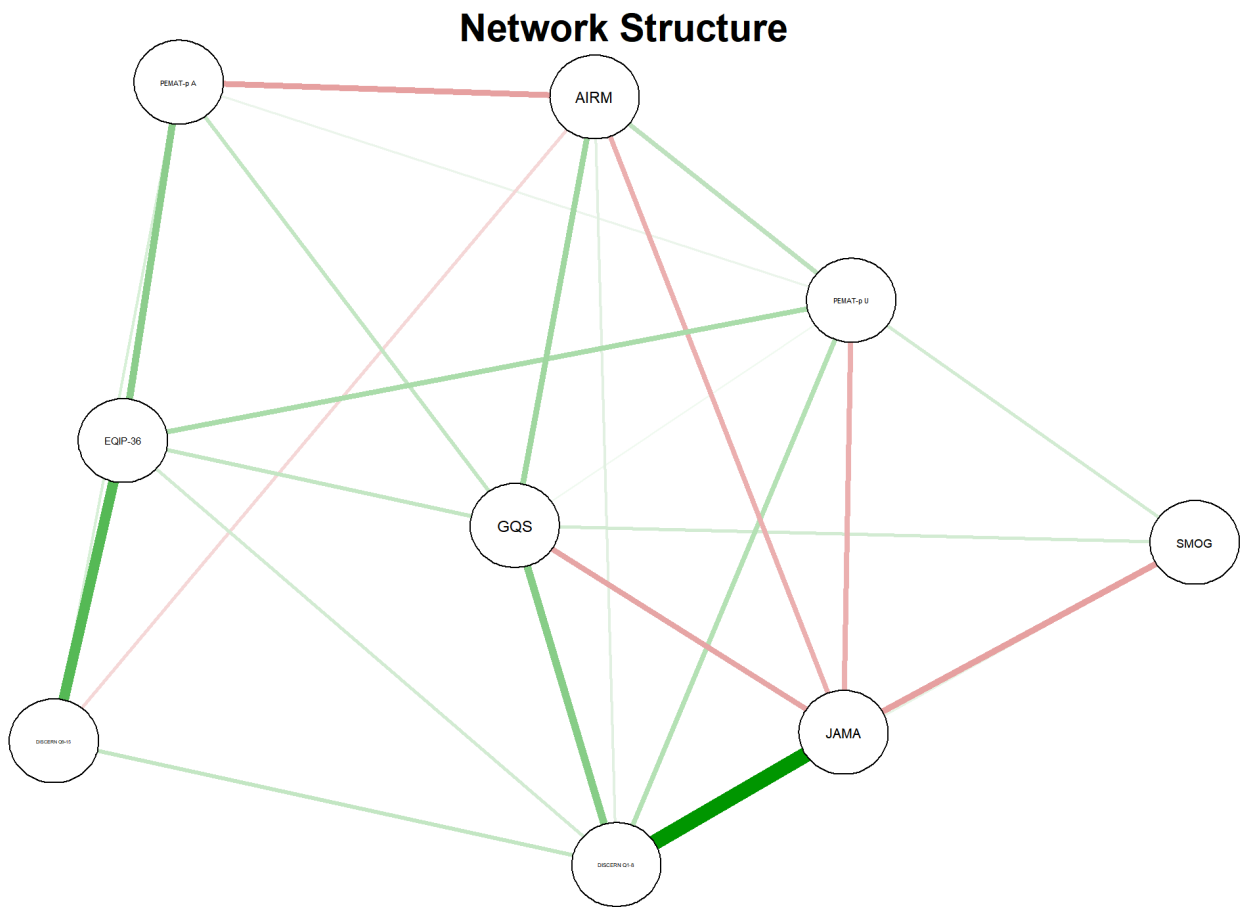


Figure 5

Network structure of quality, readability, and actionability metrics in LLM-generated patient education texts.

Note:Nodes represent the evaluation metrics included in the analysis, and edges represent regularized partial correlations between nodes. Green edges indicate positive correlations, red edges indicate negative correlations, and thicker edges indicate stronger associations. Abbreviations: PEMAT-P-U, PEMAT-P understandability; PEMAT-P-A, PEMAT-P actionability; EQIP-36, Ensuring Quality Information for Patients; DISCERN-R, DISCERN (Q1-Q8); DISCERN-T, DISCERN (Q9-Q15); GQS, Global Quality Score; AIRM, Artificial Intelligence Response Measurement; JAMA, JAMA benchmark criteria.

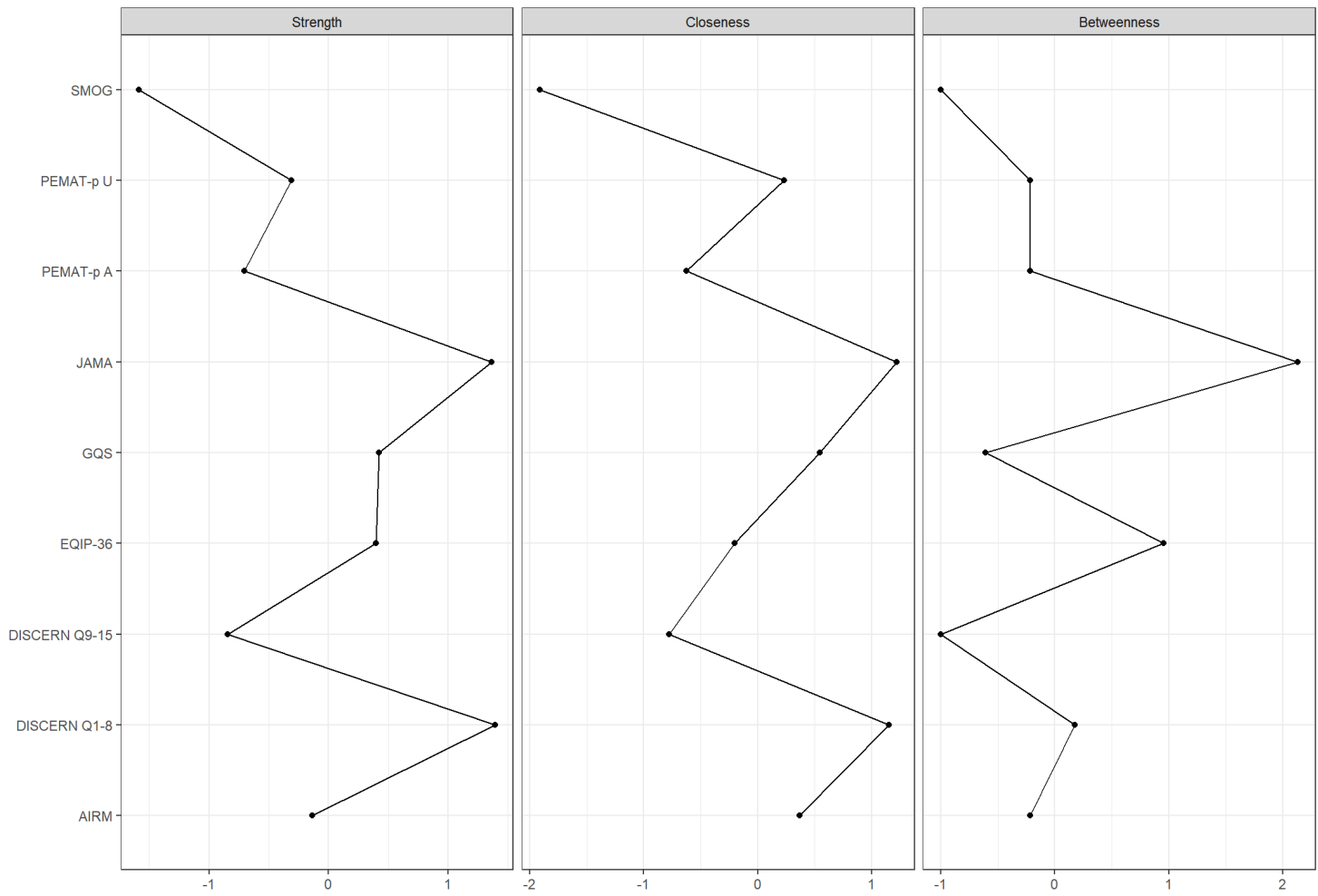


Figure 6

Centrality analysis of the network of quality, readability, and actionability metrics in LLM-generated patient education texts

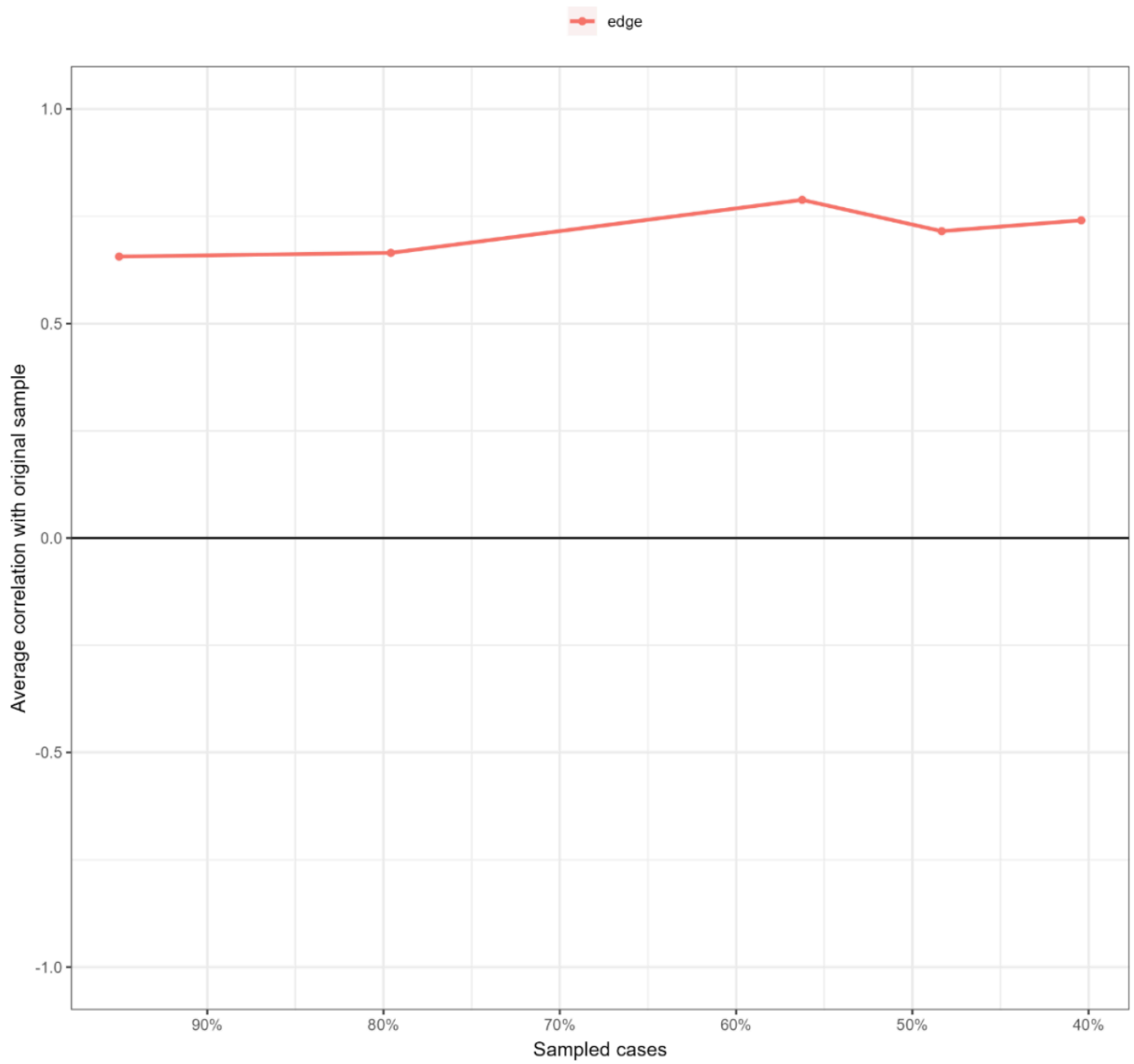


Figure 7

Stability of centrality

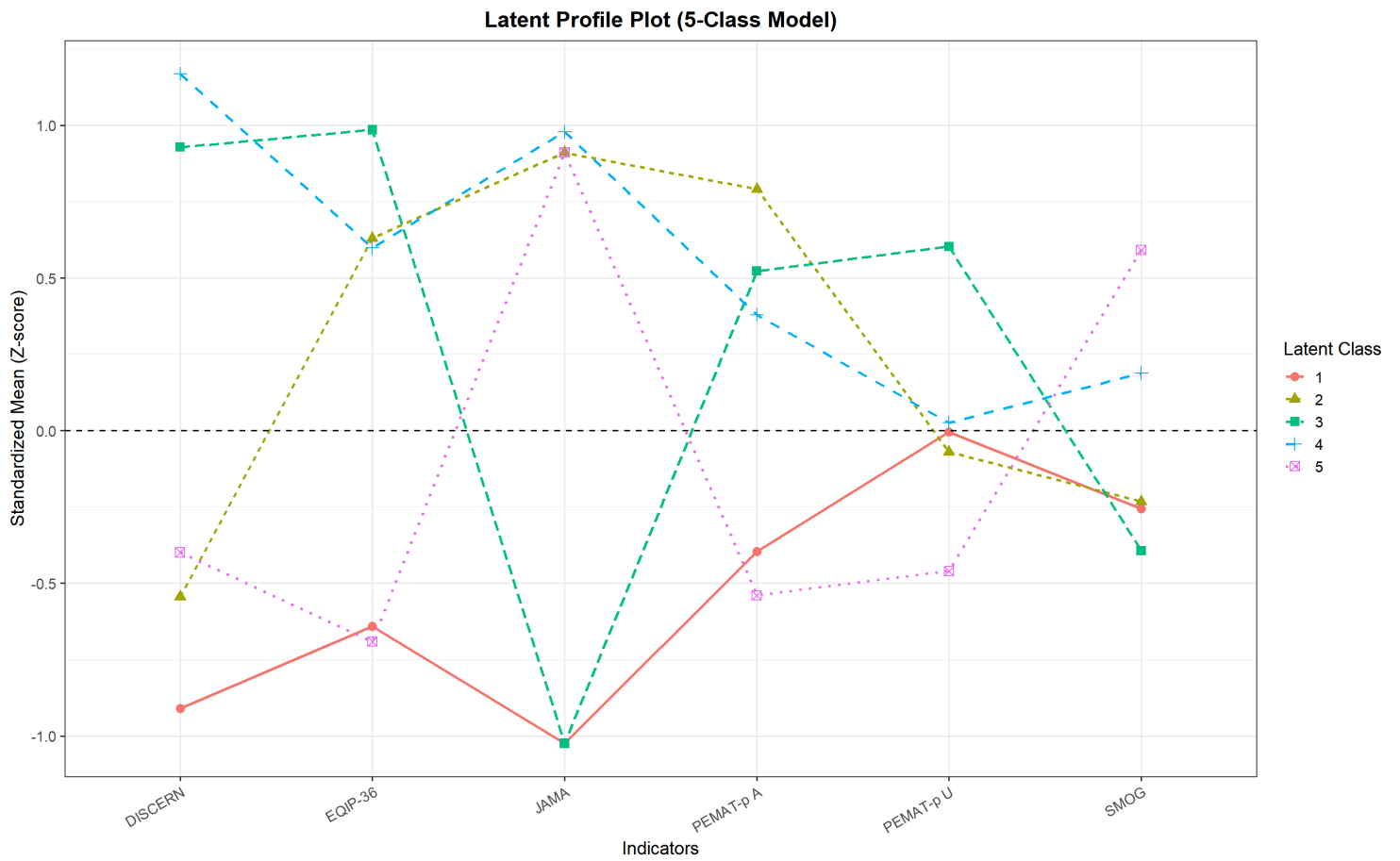


Figure 8

Profile plot of latent class characteristics

Supplementary Files

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

- [A1.xlsx](#)