scientific reports

Check for updates

OPEN

Adapting quality function deployment to translate patient feedback into prioritized technical requirements for healthcare artificial intelligence

Nora Muda✉ & Muhammad Hafiz Sulaiman

Human-centered artificial intelligence (AI) is increasingly recognized as important to advancing quality, safety, and trust in healthcare analytics, yet most validation frameworks continue to prioritize technical metrics over interpretability and stakeholder experience. Quality Function Deployment (QFD) offers a systematic method for translating customer requirements into technical design, but its application within healthcare AI remains limited. This study adapts QFD methodology to systematically align user feedback with prioritized technical requirements in healthcare AI systems. The analysis encompassed 14,938 patient reviews from 53 hospitals, from which 1,279 negative reviews were extracted for thematic analysis using large language model-driven coding (Cohen's Kappa = 0.81) and empirical factor structure, mapping multidimensional patient needs to technical specifications through a House of Quality matrix. Sensitivity analysis revealed that Granular Categorization demonstrated the highest improvement potential, achieving 21.9% advantage over LLM Coding Accuracy. This framework offers a potential approach for integrating technical validation and human-centered quality assessment, and may provide guidance for developing trustworthy, interpretable, and equitable digital medicine. While validation is limited to Malaysian private hospitals, the methodology offers a potentially scalable approach for healthcare AI development that warrants further validation. Future directions include real-world deployment across diverse populations and dynamic regulatory contexts.

Keywords Human-centered AI, Healthcare analytics, Explainable AI, Patient feedback, Quality function deployment, Continuous quality improvement, Trust, Interpretability, Clinical outcomes

Artificial intelligence (AI) is rapidly shaping the future of healthcare analytics, promising transformative advances in diagnostic support, patient engagement, and operational efficiency[1]. However, the translation of AI developments into clinically impactful systems faces considerable barriers[2], notably the AI translation gap; where robust algorithmic metrics do not automatically equate to improved patient outcomes, trust, or clinical adoption[3]. This disconnect is aggravated by most validation frameworks' emphasis on technical measures such as accuracy and precision, often at the expense of user experience, interpretability, and real-world usability[4].

Healthcare's complex operational environments require that AI systems are not only performant but also human-centered, transparent, and ethically robust[5]. Recent international frameworks, including the FUTURE-AI consensus guideline[6], WHO guidelines on AI ethics in healthcare, and the FDA's Good Machine Learning Practice principles, consistently call for approaches that achieve stakeholder alignment, support continuous feedback, mitigate bias, and elevate patient voices throughout the AI lifecycle[6–8]. Despite these imperatives, few practical methodologies exist to systematically translate diverse user requirements into actionable, prioritized technical specifications; a gap this study aims to address.

Quality Function Deployment (QFD), a proven methodology for mapping customer requirements into engineering targets, offers a potential solution. Developed in the manufacturing sector[9,10], QFD's strengths in structuring stakeholder feedback and prioritizing design have enabled significant success in industries ranging from automotive to healthcare quality management[11–13]. QFD aligns conceptually with HCAI principles: both prioritize stakeholder voice, systematic translation of qualitative requirements into quantifiable specifications,

Department of Mathematical Sciences, Faculty of Science and Technology, Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor, Malaysia. ✉email: noramuda@ukm.edu.my

and continuous improvement cycles. Yet, QFD's application to healthcare AI; where balancing technical capability and human-centered values is critical and remains underexplored.

Human-centered AI (HCAI) has emerged as a critical direction in the pursuit of transparent, safe, and usable intelligent systems in healthcare[7,14,15]. HCAI emphasizes deep integration of stakeholder feedback, rigorous interpretability, and robust ethical safeguards, moving decisively beyond purely algorithm-centric models[16,17]. Models built on stakeholder involvement demonstrate increased adoption, reduced clinical risk, and meaningful improvements to workflow and patient experience[18,19]. Advances in natural language processing (NLP) and large language models (LLMs) have substantially increased the capacity to systematically mine patient and provider feedback at scale[20–22]. Reliable coding and robust factor-analytic validation now enable nuanced user perspectives to be mapped into actionable design requirements[23].

However, the integration of feedback with existing workflows and technical prioritization frequently relies on manual, ad hoc, or expert-driven processes; lacking the systematic frameworks needed for consistent, evidence-based improvement. Responsible and explainable AI frameworks continue to strive toward trust, bias mitigation, privacy, and auditability, yet operationalizing these values remains challenging[24–28]. Implementation science and stakeholder engagement, particularly among underserved patient groups; are central to real-world adoption and equity in digital health innovation[29,30]. A crucial gap persists: despite growing recognition of HCAI principles, few published frameworks have systematically integrated QFD methodology with empirically validated, large-scale patient feedback to generate prioritized technical requirements specifically for healthcare AI analytics systems[31].

To date, prior studies either leverage QFD in narrowly defined service improvement projects or analyze user-centric AI with minimal technical specification mapping[32]. The lack of continuous feedback integration and explicit linking of technical requirements to validated constructs constrains the advancement of trustworthy, clinically applicable AI systems.

This study proposes a QFD framework for healthcare AI with initial empirical support from a Malaysian private hospital context. By adapting classic QFD to integrate large language model-driven thematic analysis, robust factor structure, and multidimensional stakeholder input, this research proposes a potentially scalable methodology for translating real-world patient reviews into prioritized technical requirements. This approach may support dynamic, evidence-based prioritization and continuous improvement, potentially helping to address the research gap for safe, equitable, and human-centered digital medicine.

Methodology

Conceptual background: healthcare AI and house of quality

For the purposes of this study, 'Healthcare AI' refers specifically to artificial intelligence systems designed for patient feedback analytics, quality assessment, and experience monitoring. This operational definition encompasses NLP-based text analysis, sentiment classification, thematic coding, and pattern recognition applied to patient-generated content. The scope excludes clinical decision support systems, diagnostic algorithms, and treatment recommendation engines, which involve distinct regulatory and validation requirements under FDA and international medical device frameworks[33,34].

The House of Quality is the primary planning matrix in QFD methodology, structured as follows: the 'left wall' contains customer requirements (Voice of Customer); the 'ceiling' lists technical requirements; the central 'relationship matrix' quantifies connections between customer needs and technical capabilities; the 'roof' (correlation matrix) identifies interdependencies among technical requirements; and the 'basement' contains priority calculations and benchmarking data[9,35]. This study adapts each HOQ component to accommodate the unique characteristics of healthcare AI development, integrating HCAI principles to ensure holistic consideration of ethical and practical dimensions[6]. The mapping of HCAI principles to QFD elements is shown in Table 1.

Research design overview

Figure 1 presents the overall research methodology flowchart, illustrating the five-phase process for developing prioritized technical requirements through the QFD framework.

Phase 1: data collection and preprocessing The study collected 14,938 patient reviews from Google Maps across 53 private hospitals in Selangor, Malaysia over a 12-month period (January–December 2023). Of these, 12,035 (81%) reviews were accompanied by comments, while 2,903 (19%) contained ratings only and were excluded from analysis. A machine learning classifier integrating natural language processing with support vector machine and logistic regression algorithms was employed to detect potentially fraudulent reviews. The classifier was trained and tested on labelled datasets from yelp.com, a platform that separates fake reviews from authentic reviews. The preprocessing pipeline included standardization, punctuation removal, numerical removal, tokenization, stop word removal, and trigram formation. The classifier achieved precision of 0.87, recall

HCAI principle	QFD component	Operationalization
Transparency	Relationship matrix	Explicit documentation of VOC-technical requirement linkages with strength ratings
Stakeholder inclusion	Voice of customer	Patient feedback serves as primary input; expert panel validates ratings
Fairness	Priority calculation	Weighted prioritization ensures high-frequency concerns receive appropriate attention
Privacy	Technical requirements	Data privacy protocols included as explicit technical requirement
Reliability	Roof matrix	Inter-requirement correlations identify synergies and trade-offs

Table 1. Mapping of HCAI principles to QFD elements.

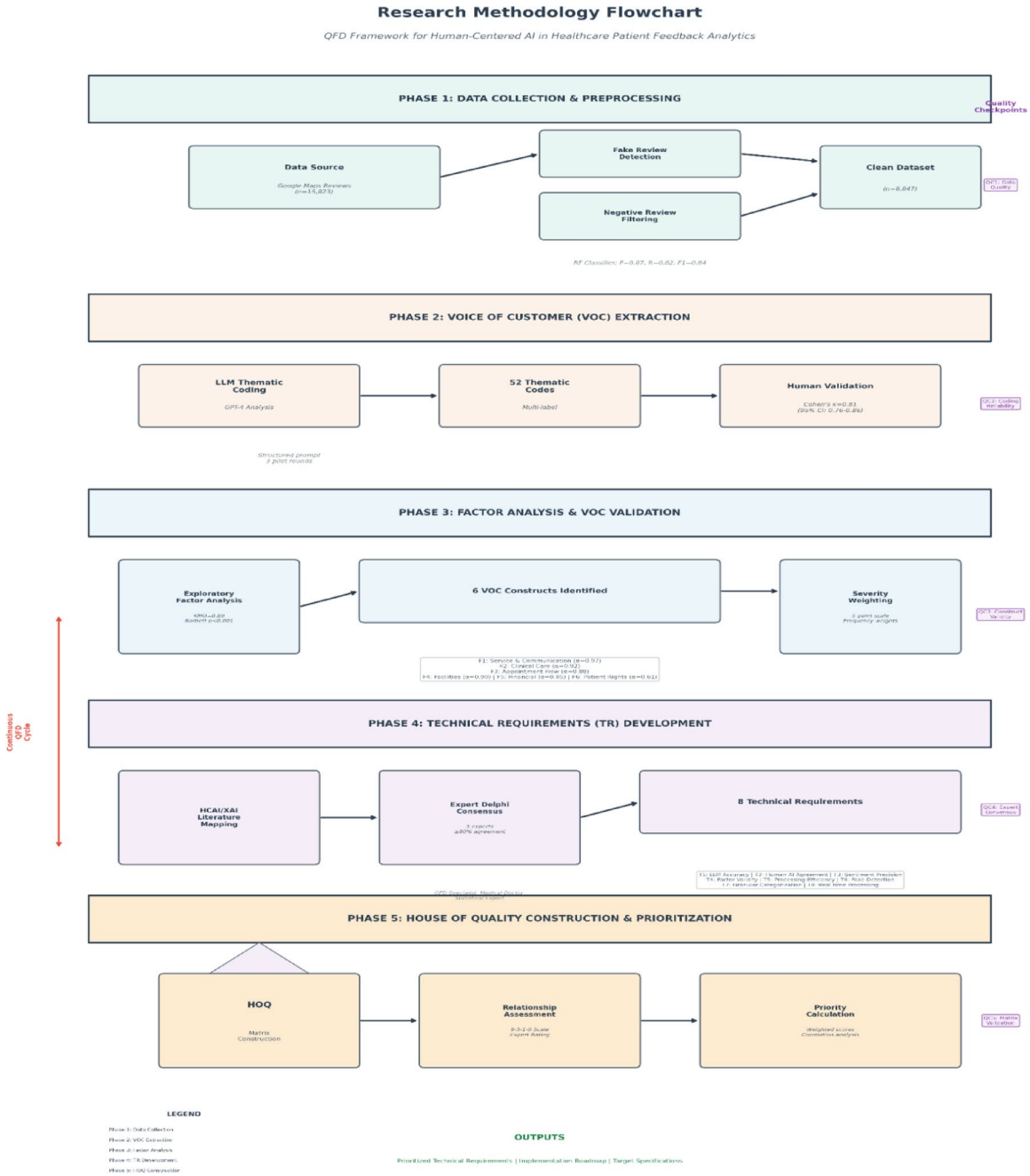

Fig. 1. Research methodology flowchart illustrating the five-phase QFD framework.

of 0.89, and accuracy of 0.88 on the validation dataset. Following fake review removal ($n = 1,121$; 9.3% of reviews with comments), sentiment classification using the GPT-4o mini model API isolated 1,279 negative reviews containing substantive quality concerns for thematic analysis. The exclusive focus on negative reviews reflects QFD methodology's emphasis on identifying improvement opportunities, which are most clearly articulated in negative feedback[36].

Phase 2: voice of customer (VOC) extraction Thematic coding employed GPT-4 with a structured prompt specifying 41 predefined thematic categories derived from healthcare quality literature. Each review was coded independently with multi-label assignment permitted. The coding schema was iteratively refined through three pilot rounds involving 100 reviews each, with expert feedback incorporated after each round. Human validation on a random sample of 200 reviews achieved Cohen's Kappa of 0.81 (95% CI: 0.76–0.86, $p < 0.001$), confirming substantial inter-rater agreement between human and AI coding. The complete list of thematic codes is provided in Supplementary Table S3.

Phase 3: factor analysis and VOC validation Exploratory factor analysis verified sampling adequacy (KMO = 0.89, Bartlett's test $p < 0.001$) and employed Varimax rotation with eigenvalue > 1 criterion for factor retention. Six latent constructs emerged representing distinct patient concern dimensions, with Cronbach's alpha

coefficients ranging from 0.61 to 0.97. Severity weighting (1–5 scale: 1 = minor inconvenience, 2 = moderate concern, 3 = significant issue, 4 = serious problem, 5 = critical failure) and frequency-based adjustments determined relative VOC importance. Strategic weights were assigned according to VOC category frequency: high-priority categories (> 300 reviews) received weight 1.5; medium-priority (100–300 reviews) received 1.2; low-priority (< 100 reviews) received 1.0[37]. Details on the factor analysis results is provided in Supplementary Table S2.

Phase 4: technical requirements development Eight technical requirements were derived through systematic mapping of HCAI principles and XAI frameworks to healthcare AI capabilities (Table 2). A modified Delphi process with three domain experts established consensus (≥ 80% agreement) on TR definitions and VOC-TR relationship strengths. The expert panel comprised a QFD and quality management specialist with experience in healthcare service design, a medical doctor, and a statistical expert specializing in psychometric validation and factor analysis. Experts independently proposed requirements based on: (a) established NLP performance standards (BLEU, ROUGE, F1 metrics); (b) Explainable AI taxonomies (LIME, SHAP interpretability frameworks); and (c) healthcare analytics literature.

Phase 5: house of quality construction The HOQ matrix integrated VOC constructs, technical requirements, and relationship assessments using standard QFD scaling: strong (9), moderate (3), weak (1), and none (0). Relationship ratings were determined through expert consensus: experts independently assigned ratings, disagreements (> 2-point differences) were discussed in facilitated sessions, and final ratings required ≥ 80% agreement.

Current performance for each VOC category was calculated as the weighted average severity of complaints:

$$t_i = \frac{\sum_{k=1}^{K} number\ of\ complaints\ at\ severity\ level\ k \times k}{total\ number\ of\ complaints\ in\ category\ i}$$

where k represents observed severity levels. The target performance (u_i) was uniformly set at 5.0, representing the excellence benchmark. The improvement ratio for each category was then computed as:

$$v_i = \frac{u_i}{t_i}$$

To derive the final priority scores for each technical requirement, the absolute weights of the VOCs were multiplied by the corresponding relationship strengths (R_{ij}) and summed across all VOCs:

$$PS_j = \sum_{i=1}^{n} (z_i \times R_{ij})$$

where PS_j is the priority score for technical requirement j, z_i is the absolute weight for VOC i, R_{ij} is the relationship strength between VOC i and technical requirement j, and n is the total number of VOC categories.

The strategic weight (w_i) was assigned according to the frequency of each VOC category, with values of 1.5, 1.2, and 1.0. These specific weights were chosen to reflect the relative frequency and strategic importance of each VOC category, following QFD best practices and recent literature recommendations[37]. A weight of 1.5 was assigned to high-priority VOC categories, defined as those mentioned in more than 300 reviews, to emphasize their widespread impact and ensure that system improvements addressing these needs are prioritized. Medium-priority VOC categories, mentioned in 100–300 reviews, were assigned a weight of 1.2, recognizing their moderate but significant importance. Low-priority VOC categories, mentioned in fewer than 100 reviews, received a baseline weight of 1.0, ensuring all needs are considered but that less frequent issues do not overshadow more common concerns. The use of these weights is consistent with QFD practice, where stepwise increments between 1.0 and 1.5 are commonly used to distinguish between tiers of customer importance or market impact, and the values are chosen to be large enough to affect prioritization while maintaining a balanced scheme. This approach also provides simplicity and transparency for stakeholder communication and future replication.

The absolute weight for each VOC was calculated as:

$$z_i = s_i \times v_i \times w_i$$

Technical requirement	Literature source	Standard metric	XAI/HCAI link
LLM coding accuracy	Khanbhai et al.[21]	F1-score, Cohen's Kappa	Reliability principle
Human-AI agreement	Chen et al.[16]	Inter-rater reliability	Human oversight
Sentiment analysis precision	Feizollah et al.[38]	Precision, Recall	Interpretability
Factor analysis validity	Cronbach and Meehl[39]	Cronbach's α, KMO	Construct validity
Data processing efficiency	Van Der Vegt et al.[19]	Throughput, latency	Usability
Fake review detection	Kim and Kwak[40]	Precision, recall	Data integrity
Granular categorization	Hake et al.[41]	Category coverage	Transparency
Real-time processing	Feng et al.[42]	Response time	Actionability

Table 2. Technical requirements provenance and literature mapping.

Sensitivity analysis confirmed that priority rankings remained stable under ± 20% weight variations.

Quality checkpoints (QC1-QC5) were embedded throughout the methodology to ensure data quality, coding reliability, construct validity, expert consensus, and matrix validation. The QFD process was designed for continuity, with quarterly re-collection of VOC data and re-evaluation of technical requirements to ensure ongoing alignment with evolving user needs and healthcare standards.

House of quality framework

Figure 2 illustrates the House of Quality (HOQ) framework adapted for healthcare AI system design. The HOQ structure comprises six interconnected components that systematically translate patient feedback into prioritized technical requirements.

Left wall (voice of customer) The left wall captures the Voice of Customer (VOC) derived from factor analysis. Six validated constructs emerged: Service and Communication Effectiveness (F1; α = 0.97), Clinical Care and Patient Experience (F2; α = 0.92), Appointment and Patient Flow (F3; α = 0.89), Facilities and Amenities Quality (F4; α = 0.90), Financial and Insurance Management (F5; α = 0.88), and Patient Rights and Accessibility (F6;

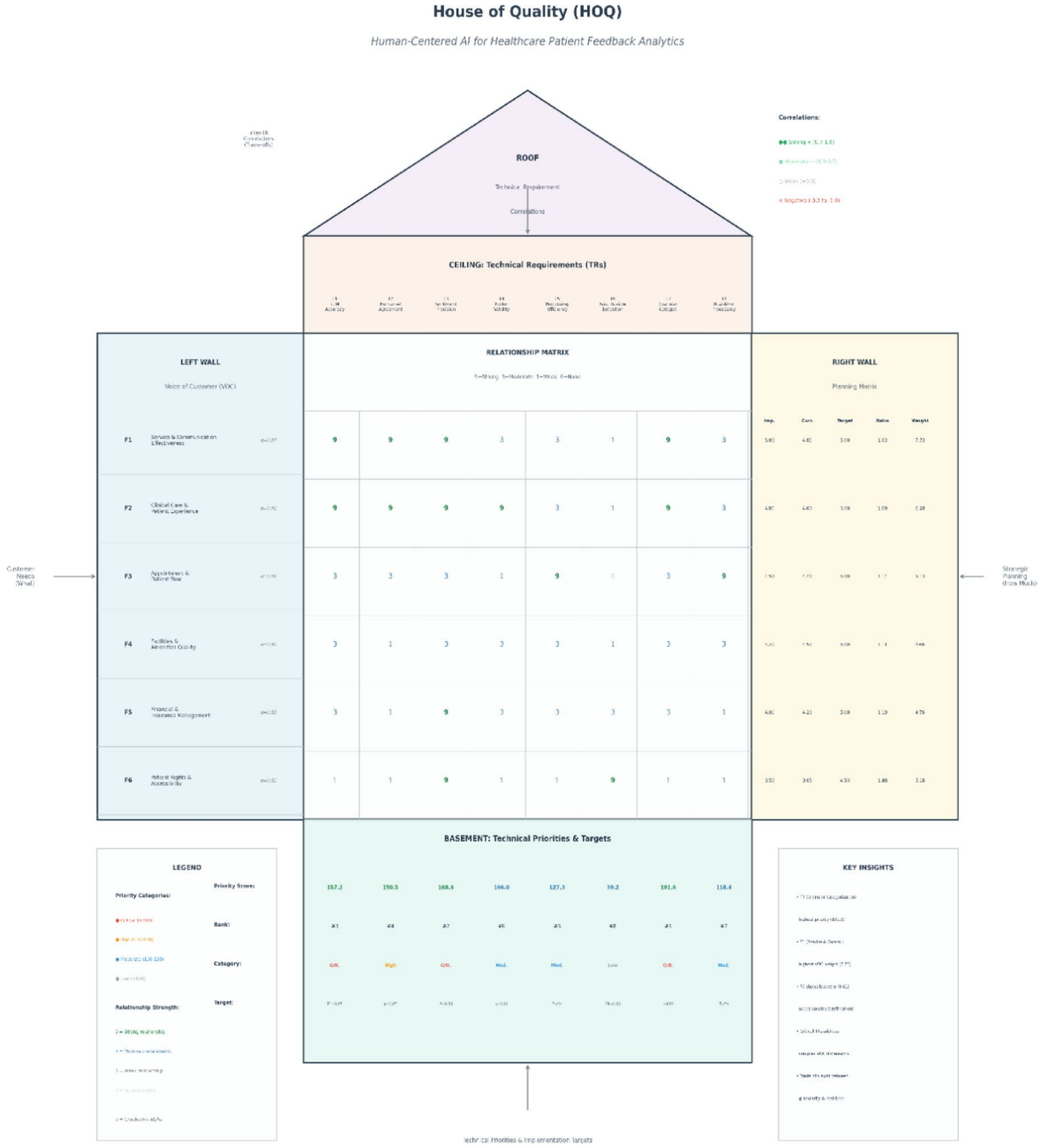

Fig. 2. House of Quality (HOQ) framework for Human-Centered AI system design in healthcare patient feedback analytics.

α = 0.61). These constructs represent the "Whats"—what patients value and expect from healthcare services that AI systems must effectively capture and analyze.

Ceiling (technical requirements) The ceiling houses eight technical requirements (TRs) derived from HCAI principles and XAI literature: LLM Coding Accuracy (T1), Human-AI Agreement Reliability (T2), Sentiment Analysis Precision (T3), Factor Analysis Validity (T4), Data Processing Efficiency (T5), Fake Review Detection (T6), Granular Categorization Coverage (T7), and Real-time Feedback Processing (T8). These represent the "Hows"—engineering characteristics that determine how effectively the AI system addresses patient needs.

Relationship matrix (center) The central matrix quantifies the strength of relationships between each VOC construct and technical requirement. For instance, Service and Communication Effectiveness (F1) exhibits strong relationships (score = 9) with LLM Coding Accuracy (T1), Human-AI Agreement (T2), Sentiment Precision (T3), and Granular Categorization (T7), reflecting the complexity of communication-related feedback that demands high analytical precision. Conversely, Patient Rights and Accessibility (F6) show strong relationships primarily with Sentiment Precision (T3) and Fake Review Detection (T6), indicating that rights-related concerns require accurate sentiment interpretation and authentic feedback verification.

Right wall (planning matrix) The planning matrix contains strategic planning elements: importance ratings reflecting the relative weight of each VOC construct, current performance assessments, target performance goals, improvement ratios, and absolute weights.

Roof (correlation matrix) The triangular roof depicts inter-correlations among technical requirements, identifying synergies and trade-offs. Strong positive correlations ($r > 0.78$) exist between LLM Coding Accuracy and Granular Categorization, suggesting that improvements in one capability reinforce the other. Notably, negative correlations were identified between Granular Categorization and Real-time Processing ($r = -0.25$, $p < 0.05$) and between LLM Coding Accuracy and Data Processing Efficiency ($r = -0.18$, $p < 0.10$), indicating resource allocation trade-offs that require explicit consideration during system design.

Basement (technical priorities) The basement consolidates priority scores. Granular Categorization (T7) achieved the highest priority score (191.58), followed by Sentiment Analysis Precision (T3; 168.89) and LLM Coding Accuracy (T1; 157.13). Technical requirements were categorized as Critical (score > 150), High (130–150), Moderate (100–130), or Low (< 100), with corresponding target specifications established for implementation guidance.

Ethical considerations

Data analyzed in this study were obtained from publicly accessible online review platforms where patients voluntarily share feedback. No personally identifiable information was collected or stored. Given the use of publicly available data without direct participant interaction, institutional review board approval was not required under Malaysian research ethics guidelines. The study adhered to platform terms of service for data collection.

Results

Patient feedback analysis and thematic structure

Analysis of 1,279 negative patient reviews yielded 41 granular thematic codes consolidated into ten primary quality dimensions. Service Quality and Professionalism led at 511 reviews (39.9%), followed by Communication Issues at 506 reviews (39.6%) and Waiting Time at 382 reviews (29.9%). This hierarchy suggests a predominant focus on interpersonal and operational aspects rather than clinical outcomes alone. Multiple thematic mentions occurred in 48.7% of reviews, indicating interconnected patient experience dimensions. The complete frequency distribution of thematic codes is provided in Supplementary Table S1.

Methodological reliability achieved Cohen's Kappa of 0.81 between AI and expert coders (95% CI: 0.76–0.86, $p < 0.001$), exceeding the threshold for substantial agreement and supporting the reliability of the LLM-based thematic coding approach.

Empirical factor structure and strategic prioritization

Exploratory factor analysis consolidated the 41 thematic codes into six latent constructs representing distinct patient concern dimensions. Table 3 presents the Customer Needs (Voice of Customer) Planning Matrix, integrating factor structure with strategic prioritization elements.

Customer need (VOC)	Importance*	Current performance†	Target performance†	Improvement ratio	Absolute weight‡	Priority rank
Service & communication effectiveness	5	4.85	5.0	1.03	7.73	1
Clinical care & patient experience	4	4.60	5.0	1.09	5.65	2
Appointment & patient flow	4	4.45	5.0	1.12	5.39	3
Financial & insurance management	3	4.40	5.0	1.14	3.41	4
Facilities & amenities quality	3	4.50	5.0	1.11	3.33	5
Patient rights & accessibility	2	3.05	4.5	1.48	2.36	6

Table 3. Customer needs (voice of customer) planning matrix. *Importance: Scale 1–5 (5 = most critical); †Performance: Scale 1–5 (5 = excellent performance); ‡Absolute Weight = Importance (s_i) × Improvement Ratio (v_i) × Strategic Weight (w_i).

Service and Communication Effectiveness demonstrated highest complexity with highest absolute weight (7.73), reflecting patient emphasis on interpersonal healthcare dimensions. Clinical Care and Patient Experience ranked second (5.65), followed by Appointment and Patient Flow (5.39). The performance gap between highest (97% of target) and lowest (68% of target) performing dimensions indicates heterogeneity in organizational capability that may warrant targeted intervention.

Patient Rights and Accessibility showed lowest current performance (3.05) with substantial improvement ratio (1.48) yet received lowest absolute weight (2.36) due to lower importance rating. Notably, this construct demonstrated lowest internal consistency ($\alpha = 0.61$). Inter-item correlation analysis suggests this construct may be multidimensional: "Patient Rights" items (privacy, informed consent) showed weak correlations ($r = 0.21$–0.34) with "Accessibility" items (physical access, language services), suggesting these may represent conceptually distinct dimensions requiring separate technical approaches.

VOC-technical requirements relationships

Table 4 presents the House of Quality relationship matrix quantifying the strength of associations between the six customer need constructs and eight technical requirements. Relationship strengths were assigned using a 9–3–1–0 scale based on expert consensus, where 9 indicates strong direct impact, 3 moderate supporting relationship, 1 weak indirect influence, and 0 no significant link.

Service and Communication Effectiveness (F1) and Clinical Care and Patient Experience (F2) demonstrated the highest row totals (46 and 49 respectively), indicating broad technical dependencies across the AI system architecture. These constructs exhibited strong relationships (score = 9) with LLM Coding Accuracy (T1), Human-AI Agreement (T2), Sentiment Precision (T3), and Granular Categorization (T7), reflecting the complexity of communication-related and clinical feedback that demands high analytical precision.

Conversely, Patient Rights and Accessibility showed the lowest row total (19), with strong relationship primarily concentrated on Fake Review Detection (T6; score = 9), indicating that rights-related concerns require authentic feedback verification more than analytical sophistication.

Column totals reveal that Granular Categorization (T7) maintains the highest aggregate relationship strength (34), followed by Sentiment Precision (T3; 32) and LLM Coding Accuracy (T1; 28), providing initial indication of technical priority before weight adjustments.

Technical requirements prioritization

Priority scores were calculated by summing the products of relationship strengths and VOC absolute weights across all customer need dimensions. Table 5 presents the final priority rankings with strategic categorization.

Priority scores demonstrated pronounced hierarchical structure with nearly five-fold variation (191.58 to 39.21). Granular Categorization emerged as dominant, achieving 13.43% advantage over Sentiment Analysis Precision (168.89). The Critical tier encompasses three requirements (T7, T3, T1) that collectively address the analytical complexity demanded by high-weight VOC constructs. Human-AI Agreement Reliability (150.46) positioned at the Critical-High boundary, reflecting its importance for trustworthy human-centered AI deployment.

Figure 3 displays the priority ranking of technical requirements derived from the QFD analysis, showing clear stratification with priority scores ranging from 191.58 to 39.21. This nearly five-fold variation provides quantitative justification for differential resource allocation strategies. Granular Categorization ranks highest (191.58), achieving a 13.4% lead over Sentiment Analysis Precision (168.89), underscoring the essential role of detailed classification in qualitative patient feedback analysis. The Critical capability cluster—comprising Granular Categorization, Sentiment Analysis Precision, and LLM Coding Accuracy (scores 157–192)—reflects complementary functions: structural organization of feedback themes, emotional context capture, and reliable automated text processing.

Human-AI Agreement Reliability ranked fourth (150.46), positions at the Critical-High boundary, highlighting the importance of validation and human oversight in automated healthcare quality assessment. This aligns with human-centered AI design principles emphasizing trustworthy human–machine collaboration.

Customer needs (voice of customer)	T1	T2*	T3	T4	T5	T6	T7	T8	Total
Service & communication effectiveness	**9**	**9**	**9**	3	3	1	**9**	3	**46**
Clinical care & patient experience	**9**	**9**	**9**	**9**	3	0	**9**	1	**49**
Facilities & amenities quality	3	1	3	3	**9**	0	3	**9**	**31**
Appointment & patient flow	3	3	1	3	**9**	0	**9**	**9**	**37**
Financial & insurance management	1	1	**9**	1	3	3	3	3	**24**
Patient rights & accessibility	3	3	1	1	1	**9**	1	0	**19**
Total	**28**	**26**	**32**	**20**	**28**	**13**	**34**	**25**	

Table 4. House of quality relationship matrix between customer needs and technical requirements. **Technical Requirements: T1: LLM Coding Accuracy, T2: Human-AI Agreement Reliability, T3: Sentiment Analysis Precision, T4: Factor Analysis Validity, T5: Data Processing Efficiency, T6: Fake Review Detection, T7: Granular Categorization, T8: Real-time Feedback Processing. **Relationship Strength: **9** = Strong, 3 = Moderate, 1 = Weak, 0 = None. Significant values are in bold.

Rank	Technical requirement	Priority score	Category	Strategic focus
1	Granular categorization	191.58	Critical	Detailed patient feedback classification
2	Sentiment analysis precision	168.89	Critical	Emotional context understanding
3	LLM coding accuracy	157.13	Critical	Core AI reliability
4	Human-AI agreement reliability	150.46	High	Validation and quality assurance
5	Data processing efficiency	131.28	High	Operational performance
6	Real-time feedback processing	117.61	Moderate	Timely response capability
7	Factor analysis validity	106.01	Moderate	Statistical robustness
8	Fake review detection	39.21	Low	Data authenticity

Table 5. Technical requirements priority ranking based on house of quality analysis. Priority Categories: Critical (>150): Immediate investment required—highest impact on customer satisfaction. High (130–150): Important for system reliability and operational excellence. Moderate (100–130): Valuable enhancements for competitive advantage. Low (<100): Specialized functions with limited but important applications.

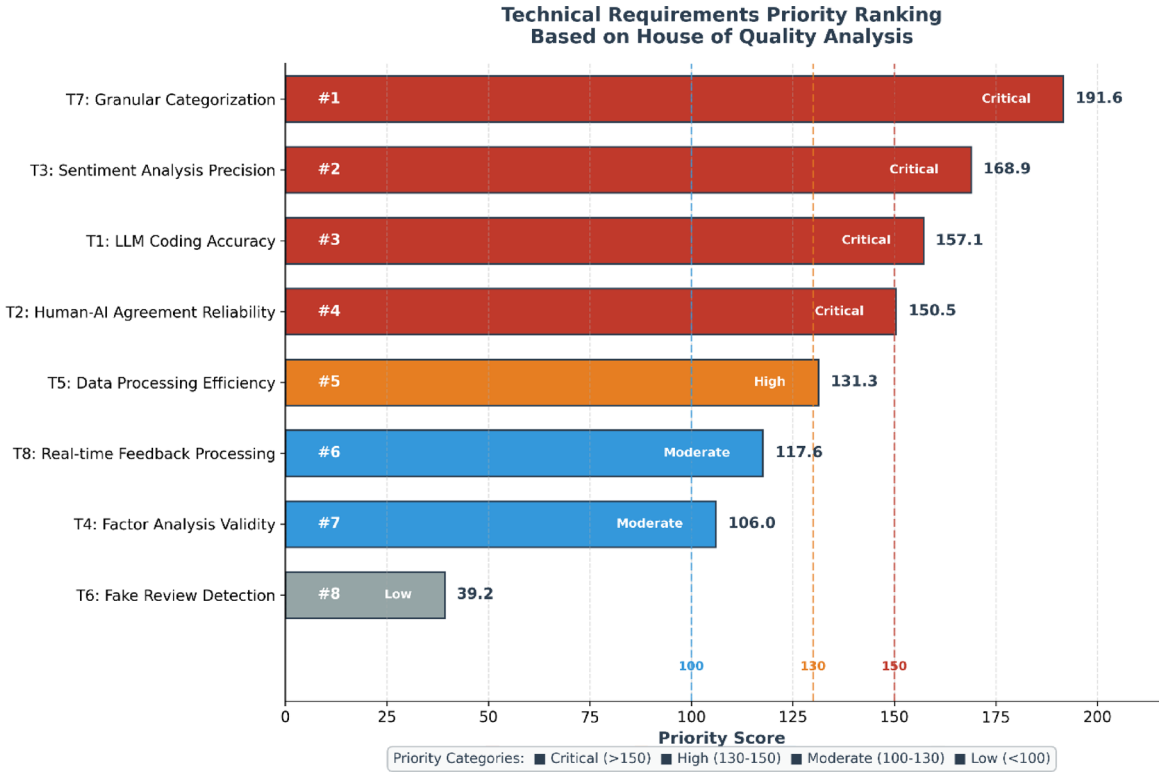

Fig. 3. Technical requirements priority scores ranked by weighted Voice of Customer importance.

The High tier (Data Processing Efficiency, 131.28) and Moderate tier (Real-time Processing, Factor Analysis Validity) represent operational and analytical enhancements respectively.

Fake Review Detection occupies the Low tier (39.21), reflecting its specialized function in data authenticity rather than direct patient satisfaction impact. This priority distribution supports a phased implementation strategy: Phase 1 (0–6 months) targeting Critical requirements with 60% of resources, Phase 2 (6–12 months) addressing High/Moderate requirements with 25%, and Phase 3 (12–18 months) completing Low-priority capabilities with 15%.

Correlation matrix analysis

The technical requirements correlation matrix ("roof" of HOQ) revealed predominantly positive correlations with notable trade-offs requiring management attention. Figure 4 presents the correlation heatmap visualization.

The strongest positive correlation exists between LLM Coding Accuracy and Granular Categorization ($r = 0.78$, $p < 0.001$), reflecting fundamental interdependence where improvements in coding accuracy reinforce categorization coverage. Human-AI Agreement and LLM Coding Accuracy formed a reliability cluster ($r = 0.72$, $p < 0.001$). Data Processing Efficiency and Real-time Feedback Processing demonstrated strong operational synergy ($r = 0.72$, $p < 0.001$), suggesting shared infrastructure investments yield dual benefits.

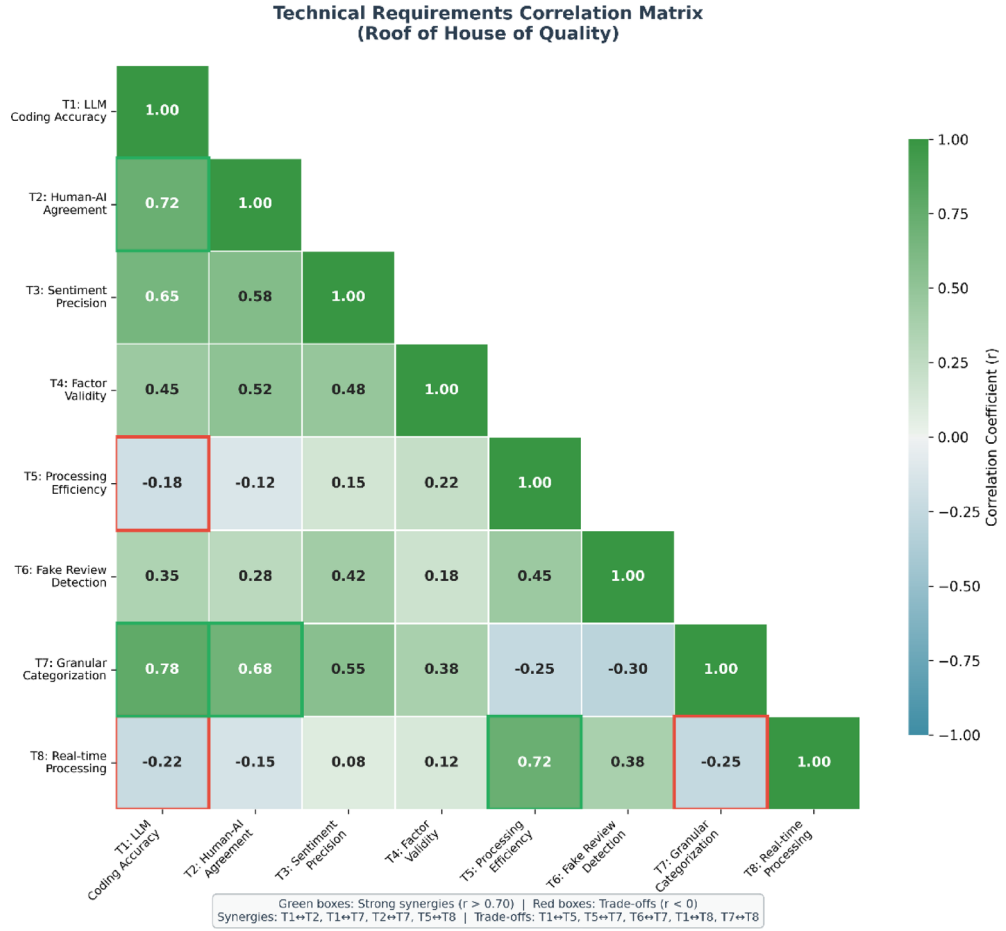

Fig. 4. Technical requirements correlation matrix (Roof of House of Quality).

Critically, negative correlations identify trade-offs requiring explicit management. LLM Coding Accuracy and Fake Review Detection exhibited a moderate negative correlation ($r = -0.30$, $p < 0.05$), indicating resource competition between analytical depth and authentication processing. Granular Categorization and Real-time Processing demonstrated a similar inverse relationship ($r = -0.25$, $p < 0.05$), reflecting the latency costs associated with computational complexity for fine-grained classification. A weaker negative correlation emerged between LLM Coding Accuracy and Data Processing Efficiency ($r = -0.18$, $p < 0.10$), suggesting computational overhead inherent in sophisticated language models.

These trade-offs require explicit consideration in implementation planning, particularly for real-time hospital environments where processing latency constraints may limit achievable granularity. System architects must balance analytical sophistication against operational responsiveness based on deployment context requirements.

Sensitivity analysis

Sensitivity analysis examined the impact of ± 20% variation in VOC weights on priority rankings. Results confirmed that the top three technical requirements (Granular Categorization, Sentiment Analysis Precision, LLM Coding Accuracy) maintained their Critical classification across all weight perturbations, demonstrating robust prioritization.

Additionally, analysis of 20% improvement potential in top three technical capabilities revealed: Granular Categorization offered highest improvement potential (38.32 points), representing 21.9% advantage over LLM Coding Accuracy (31.43 points). Sentiment Analysis Precision occupied intermediate position (33.78 points). Cumulative improvement potential (103.53 points) exceeded simple additive effects (97.89 points), indicating synergistic relationships supporting integrated rather than sequential development strategies.

Figure 5 presents the complete House of Quality matrix integrating patient-derived VOC constructs with technical requirements for the healthcare AI system. The matrix reveals a pronounced hierarchical structure in technical priorities, with nearly five-fold variation between highest (Granular Categorization, 191.58) and lowest (Fake Review Detection, 39.21) priority scores. Three technical requirements—Granular Categorization, Sentiment Analysis Precision, and LLM Coding Accuracy—emerged in the Critical category (> 150), collectively addressing the analytical complexity demanded by high-weight VOC constructs such as Service and Communication Effectiveness (absolute weight = 7.73) and Clinical Care and Patient Experience (absolute weight = 5.65). The relationship matrix demonstrates that complex, well-measured constructs require

VOC / Technical Requirements	T1	T2	T3	T4	T5	T6	T7	T8	Importance	Current Perf.	Target Perf.	Imp. Ratio	Abs. Weight	Row Total
	LLM Coding Accuracy	Human-AI Agreement	Sentiment Precision	Factor Analysis Validity	Data Processing Efficiency	Fake Review Detection	Granular Categorization	Real-time Processing						
Service & Communication Effectiveness	9	9	9	3	3	1	9	3	5	4.85	5	1.03	7.73	46
Clinical Care & Patient Experience	9	9	9	9	3	0	9	1	4	4.6	5	1.09	5.65	49
Appointment & Patient Flow	3	3	1	3	9	0	9	9	4	4.45	5	1.12	5.39	37
Facilities & Amenities Quality	3	1	3	3	9	0	3	9	3	4.5	5	1.11	3.33	31
Financial & Insurance Management	1	1	9	1	3	3	3	3	3	4.4	5	1.14	3.41	24
Patient Rights & Accessibility	3	3	1	1	1	9	1	0	2	3.05	4.5	1.48	2.36	19
COLUMN TOTALS	28	26	32	20	28	13	34	25						
PRIORITY SCORES	157.13	150.46	168.89	106.01	131.28	39.21	191.58	117.61						
PRIORITY RANK	3	4	2	7	5	8	1	6						
CATEGORY	Critical	High	Critical	Moderate	High	Low	Critical	Moderate						
LEGEND:	9 = Strong	3 = Moderate	1/0 = Weak/None											

Fig. 5. House of quality—healthcare AI patient feedback analytic.

proportionally sophisticated analytical capabilities, with F1 and F2 exhibiting strong relationships (score = 9) across multiple technical requirements. Human-AI Agreement Reliability positioned at the Critical-High boundary (150.46), reflecting the importance of validation infrastructure in trustworthy AI deployment. These findings provide quantitative justification for a phased implementation strategy prioritizing the Critical capability cluster in initial development phases.

Discussion

This study extends prior research advocating for human-centered and explainable AI in healthcare by integrating QFD methodology with empirically validated patient feedback constructs. The QFD-factor analysis integration responds to calls for methodologies incorporating human-centered evaluation dimensions[43] beyond technical performance alone[43,44]. By anchoring technical prioritization in validated patient-derived constructs, the proposed approach may help address the "AI translation gap" between research prototypes and clinically deployable systems[3].

The observed correspondence between factor complexity and technical prioritization aligns with established psychometric principles. Complex, well-measured constructs such as Service and Communication Effectiveness ($\alpha = 0.97$) appear to require proportionally sophisticated analytical capabilities, a finding consistent with stratified approaches to feedback analysis reported in patient-centered NLP research[21,38]. Conversely, lower reliability domains including Patient Rights and Accessibility ($\alpha = 0.61$) emerged as potential opportunities for foundational measurement refinement rather than immediate technological investment. Inter-item correlation analysis suggested that this construct may be multidimensional, with "Patient Rights" items showing weak correlations with "Accessibility" items, indicating possible need for construct disaggregation before targeted technical development.

Granular Categorization ranking as highest priority (191.58) offers preliminary empirical support for resource allocation decisions, consistent with literature emphasizing patient-experience-driven quality improvements[41,45]. The hierarchical technical requirements architecture revealed through HOQ analysis suggests potential value in integrated, systems-thinking approaches for medical AI deployment[19,46]. The Critical capability cluster—comprising Granular Categorization, Sentiment Analysis Precision, LLM Coding Accuracy, and Human-AI Agreement Reliability; demonstrated strong positive inter-correlations, suggesting that these capabilities may function interdependently in supporting robust patient feedback analysis systems.

The correlation matrix identified trade-offs that warrant consideration during implementation planning. The negative correlation between Granular Categorization and Real-time Processing ($r = -0.25$, $p < 0.05$) reflects potential tension between analytical depth and operational responsiveness. Healthcare organizations may need to calibrate this trade-off based on deployment context: real-time clinical environments might prioritize processing speed, while quality improvement initiatives might favor classification granularity.

Context dependency of priority weights

The high weighting of Service Quality and Professionalism (39.9%) likely reflects the private healthcare context of this study, where customer service may function as a market differentiator. In public healthcare systems with different resource constraints and patient expectations, priority weights could shift substantially. Clinical Care dimensions might receive higher relative priority in resource-constrained settings where service amenities are secondary to access and treatment quality. Similarly, cultural factors may influence patient expectations and complaint patterns, potentially altering VOC construct weights across geographic contexts.

The methodology presented appears generalizable across healthcare settings; however, the specific priority weights derived in this study are context-dependent and would require local validation before implementation. Healthcare organizations considering adoption of this framework should conduct site-specific VOC collection and factor analysis to ensure technical priorities reflect their patient population's concerns[47,48].

Practical implications and implementation roadmap

The priority analysis suggests a phased implementation strategy that may align with resource constraints typical of healthcare organizations. Phase 1 (0–6 months) could focus on Critical capabilities—Granular Categorization, Sentiment Analysis Precision, and LLM Coding Accuracy—commanding approximately 60% of technical

development resources. These capabilities may form an analytical foundation that could support accurate interpretation of patient feedback complexity. Phase 2 (6–12 months) would address High and Moderate priority capabilities including Human-AI Agreement Reliability and Data Processing Efficiency, allocating approximately 25% of resources to validation infrastructure and operational performance. Phase 3 (12–18 months) would develop specialized functions such as Fake Review Detection and Factor Analysis Validity with the remaining 15% of resources.

Suggested target specifications derived from current performance benchmarks and literature standards include: LLM Coding Accuracy F1-score ≥ 0.85 (current baseline: 0.81), Sentiment Analysis Precision ≥ 0.88 (current baseline: 0.83), and Granular Categorization coverage ≥ 95% of identified themes (current baseline: 89%). These targets represent potentially achievable improvements within the proposed timeline based on documented performance trajectories in healthcare NLP applications. Quarterly review cycles incorporating emerging patient concerns and technology advances could support continuous QFD refinement, potentially operationalizing the framework's iterative design philosophy.

Limitations

Several limitations should be acknowledged. First, empirical validation is based on Malaysian private hospitals, potentially limiting generalizability to other healthcare systems or cultural contexts. Patient experience factors may vary across settings, requiring localized validation before implementation.

Second, technical requirements prioritization relies on expert consensus and may not fully capture emerging AI capabilities. Third, sensitivity analysis demonstrates theoretical improvement potential; actual implementation outcomes may vary based on organizational capacity and change management effectiveness.

Fourth, the fake review classifier, while achieving acceptable accuracy (accuracy = 0.88), may introduce bias through false positives (legitimate reviews removed) or false negatives (fake reviews retained). Fifth, exclusive use of negative reviews may not capture improvement opportunities identified through positive feedback patterns.

Sixth, expert panel composition (primarily academic and clinical backgrounds) may introduce sampling bias in relationship ratings; inclusion of patient representatives and frontline staff could yield different priority weightings.

Finally, web scraping of patient reviews, while using publicly available data, raises ethical considerations regarding consent and data use that warrant continued attention as regulatory frameworks evolve. Seventh, the improvement potentials calculated (e.g., 21.9% advantage for Granular Categorization) represent theoretical projections based on sensitivity analysis rather than observed outcomes from implemented systems. Actual improvements in deployed systems may differ substantially. Eighth, the expert panel comprised only three domain experts. While consensus thresholds (≥ 80%) were applied, a larger and more diverse panel might yield different relationship ratings and priority scores. Ninth, this study presents a methodological framework without real-world implementation or outcome measurement. The practical effectiveness of the prioritization scheme in improving actual healthcare AI systems remains to be demonstrated. Future research should focus on real-world deployment studies with attention to cross-cultural adaptability.

Conclusion

This study presents an adapted QFD framework for prioritizing technical requirements in human-centered healthcare AI systems. The potential contributions include systematic integration of empirically-derived patient feedback with QFD methodology; preliminary evidence suggesting that construct complexity may correspond with technical priority requirements; identification of potential synergistic capability clusters that may support integrated development strategies; and proposed implementation guidance for healthcare organizations that warrants further validation.

The findings should be interpreted within the study's limitations. Validation was conducted within a single Malaysian private hospital context, which may limit generalizability to other healthcare settings. Nevertheless, the methodology offers a potentially scalable approach that may be applicable across diverse healthcare contexts with appropriate local calibration. Future research should focus on real-world deployment studies to assess practical effectiveness, cross-cultural validation to examine transferability of priority weights, and integration with emerging AI capabilities including federated learning approaches for privacy-preserving analytics.

Data availability

Data analyzed in this study were obtained from publicly available online hospital review platforms. No new datasets were generated. Processed data supporting the findings are available from the corresponding author upon reasonable request.

Received: 29 October 2025; Accepted: 13 January 2026
Published online: 19 January 2026

References

1. Topol, E. J. High-performance medicine: The convergence of human and artificial intelligence. *Nat. Med.* **25**(1), 44–56 (2019).
2. Shuaib, A. Transforming healthcare with AI: Promises, pitfalls, and pathways forward. *Int. J. Gen. Med.* 1765–1771 (2024).
3. Overgaard, S. M. et al. Implementing quality management systems to close the AI translation gap and facilitate safe, ethical, and effective health AI solutions. *NPJ Digit. Med.* **6**(1), 218 (2023).
4. Cross, J. L., Choma, M. A. & Onofrey, J. A. Bias in medical AI: Implications for clinical decision-making. *PLOS Digit. Health.* **3**(11), e0000651. https://doi.org/10.1371/journal.pdig.0000651 (2024).
5. Göktaş, P. & Grzybowski, A. Shaping the future of healthcare: Ethical clinical challenges and pathways to trustworthy AI. *J. Clin. Med.* **14**(5), 1605. https://doi.org/10.3390/jcm14051605 (2025).

6. Lekadir, K. et al. FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare. *BMJ* **388**, e081554 (2025).
7. Shneiderman, B. *Human-Centered AI* (Oxford University Press, 2022). https://doi.org/10.1093/oso/9780192845290.001.0001.
8. You, J. G., Hernandez-Boussard, T., Pfeffer, M. A., Landman, A. & Mishuris, R. G. Clinical trials informed framework for real world clinical implementation and deployment of artificial intelligence applications. *NPJ Digit. Med.* **8**(1), 107 (2025).
9. Hauser, J. R. & Clausing, D. The house of quality. *Harv. Bus. Rev.* **66**(3), 63–73 (1988).
10. Karsak, E. E. Fuzzy multiple objective programming framework to prioritize design requirements in quality function deployment. *Comput. Ind. Eng.* **47**, 149. https://doi.org/10.1016/j.cie.2004.06.001 (2004).
11. Soewardi, H. & Afgani, M. K. A. innovative design of ergonomic wheelchair for disabled people. In *IOP Conference Series: Materials Science and Engineering* vol. 598 012033 (2019). https://doi.org/10.1088/1757-899X/598/1/012033.
12. Gremyr, I. & Raharjo, H. Quality function deployment in healthcare: A literature review and case study. *Int. J. Health Care Qual. Assur.* **26**(2), 135–146. https://doi.org/10.1108/09526861311297343 (2013).
13. Mahajan, A., Sharma, S. & Soni, A. Application of quality function deployment in healthcare services: A review. *Int. J. Res. Appl. Sci. Eng. Technol.* **8**(7), 1890–1895. https://doi.org/10.22214/ijraset.2020.30707 (2020).
14. Schmager, S., Pappas, I. O., & Vassilakopoulou, P. Understanding human-centred AI: A review of its defining elements and a research agenda. *Behav. Inf. Technol.* 1–40 (2025).
15. Ahmad, K. et al. Requirements engineering framework for human-centered artificial intelligence software systems. *Appl. Soft Comput.* **143**, 110455. https://doi.org/10.1016/j.asoc.2023.110455 (2023).
16. Chen, H., Gomez, C., Huang, C. M. & Unberath, M. Explainable medical imaging AI needs human-centered design: Guidelines and evidence from a systematic review. *NPJ Digit. Med.* **5**(1), 156 (2022).
17. Chen, Y., Clayton, E. W., Novak, L. L., Anders, S. & Malin, B. Human-centered design to address biases in artificial intelligence. *J. Med. Internet Res.* https://doi.org/10.2196/43251 (2023).
18. Lee, J. W. et al. Development of AI-generated medical responses using the ChatGPT for cancer patients. *Comput. Methods Programs Biomed.* **254**, 108302 (2024).
19. Van Der Vegt, A. H. et al. Implementation frameworks for end-to-end clinical AI: Derivation of the SALIENT framework. *J. Am. Med. Inform. Assoc.* **30**(9), 1503–1515 (2023).
20. Wang, L. et al. Human-centered design and evaluation of AI-empowered clinical decision support systems: A systematic review. *Front. Comput. Sci.* **5**, 1187299 (2023).
21. Khanbhai, M. et al. Applying natural language processing and machine learning techniques to patient experience feedback: A systematic review. *BMJ Health Care Inform.* **28**(1), e100262 (2021).
22. Maity, S. & Saikia, M. J. Large language models in healthcare and medical applications: A review. *Bioengineering* **12**(6), 631 (2025).
23. Sulaiman, M. H., Muda, N. & Abdul Razak, F. Analyzing patient complaints in web-based reviews of private hospitals in Selangor, Malaysia, using large language model-assisted. *JMIR Formative Res.* **9**, e69075 (2025).
24. Ferdowsi, M., Hasan, M. M. & Habib, W. Responsible AI for cardiovascular disease detection: Towards a privacy-preserving and interpretable model. *Comput. Methods Programs Biomed.* **254**, 108289 (2024).
25. Tun, H. M., Rahman, H. A., Naing, L. & Malik, O. A. Trust in artificial intelligence-based clinical decision support systems among health care workers: Systematic review. *J. Med. Internet Res.* **27**, e69678 (2025).
26. Noor, A. A., Manzoor, A., Mazhar Qureshi, M. D., Qureshi, M. A. & Rashwan, W. Unveiling explainable AI in healthcare: Current trends, challenges, and future directions. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **15**(2), e70018 (2025).
27. Gertych, A. & Faust, O. AI explainability and bias propagation in medical decision support. *Comput. Methods Programs Biomed.* **257**, 108465 (2024).
28. Loh, H. W. et al. Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011–2022). *Comput. Methods Programs Biomed.* **226**, 107161 (2022).
29. Alharthi, S. A. mHealth applications in Saudi Arabia: Current features and future opportunities. *Healthcare* **13**(12), 1392. https://doi.org/10.3390/healthcare13121392 (2025).
30. Lopez, I., Velasquez, D. E., Chen, J. H. & Rodriguez, J. A. Operationalizing machine-assisted translation in healthcare. *npj Digit. Med.* **8**(1), 584 (2025).
31. El Arab, R. A. et al. Bridging the gap: From AI success in clinical trials to real-world healthcare implementation—A narrative review. *Healthcare* **13**(7), 701 (2025).
32. Crubezy, M., Douay, C., Michel, P. & Haesebaert, J. Using patient comments from a standardised experience survey to investigate their perceptions and prioritise improvement actions: A thematic and syntactic analysis. *BMC Health Serv. Res.* https://doi.org/10.1186/s12913-023-09953-z (2023).
33. Vidal, D. E., Loufek, B., Kim, Y. H. & Vidal, N. Y. Navigating US regulation of artificial intelligence in medicine—A primer for physicians. *Mayo Clin. Proc. Digit. Health* **1**(1), 31–39 (2023).
34. U. S. Food & Drug Administration (FDA). Clinical decision support software-Guidance for industry and food and drug administration staff (2022).
35. Akao, Y. *Quality Function Deployment: Integrating Customer Requirements into Product Design* (Productivity Press, 2004).
36. Hasibuan, A. et al. Service quality improvement by using the quality function deployment (QFD) method at the government general hospital. *J. Phys. Conf. Ser.* **1363**(1), 012095. https://doi.org/10.1088/1742-6596/1363/1/012095 (2019).
37. Ginting, R., Silalahi, R. & Marunduri, M. A. The conceptual integration of quality function deployment and value engineering for product development: A case study of water dispenser. *Int. J. Technol.* **16**(1), 124–135 (2025).
38. Feizollah, A. et al. The use of natural language processing to interpret unstructured patient feedback on health services: Scoping review. *J. Med. Internet. Res.* **27**, e72853. https://doi.org/10.2196/72853 (2025).
39. Cronbach, L. J. & Meehl, P. E. Construct validity in psychological tests. *Psychol. Bull.* **52**(4), 281–302. https://doi.org/10.1037/h0040957 (1955).
40. Kim, S. & Kwak, M. Customer complaint analysis via review-based control charts and dynamic importance-performance analysis. *Appl. Sci.* **13**(10), 5991. https://doi.org/10.3390/app13105991 (2023).
41. Hake, P., Rehse, J.-R. & Fettke, P. Toward automated support of complaint handling processes: An application in the medical technology industry. *J. Data Semant.* **10**, 41. https://doi.org/10.1007/s13740-021-00124-z (2021).
42. Feng, J. et al. Clinical artificial intelligence quality improvement: Towards continual monitoring and updating of AI algorithms in healthcare. *NPJ Digit. Med.* **5**(1), 66 (2022).
43. Donabedian, A. The quality of care: How can it be assessed?. *JAMA* **260**(12), 1743–1748 (1988).
44. Kim, J., Maathuis, H. & Sent, D. Human-centered evaluation of explainable AI applications: A systematic review. *Front. Artif. Intell.* **7**, 1456486 (2024).
45. Beheshtinia, M. A., Fathi, M., Ghobakhloo, M. & Mubarak, M. F. Enhancing hospital services: Achieving high quality under resource constraints. *Health Serv. Insights.* **18**, 11786329251331312. https://doi.org/10.1177/11786329251331311 (2025).
46. Hanefeld, J., Powell-Jackson, T. & Balabanova, D. Understanding and measuring quality of care: dealing with complexity. *Bull World Health Organ.* **95**(5), 368–374. https://doi.org/10.2471/BLT.16.179309 (2017).
47. Vats, K. Navigating the digital landscape: Embracing innovation, addressing challenges, and prioritizing patient-centric care. *Cureus* https://doi.org/10.7759/cureus.58352 (2024).
48. Rahim, A. I. A., Ibrahim, M. I., Musa, K. I., Chua, S.-L. & Yaacob, N. M. Patient satisfaction and hospital quality of care evaluation in Malaysia using SERVQUAL and facebook. *Healthcare* **9**(10), 1369. https://doi.org/10.3390/healthcare9101369 (2021).

Author contributions

Nora designed the study and developed the artificial intelligence models. Muhammad Hafiz collected and processed the data. Nora and Muhammad Hafiz performed the statistical analyses. Nora wrote the main manuscript text and prepared all figures. All authors reviewed and approved the final version.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-026-36550-x.

Correspondence and requests for materials should be addressed to N.M.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.