International Journal of Graphic Design (IJGD)
Volume 3 No. 2 Ocotober 2026
E-ISSN: 2987-9434; P-ISSN : 2988-0343, Page 381-394
DOI: https://doi.org/10.51903/ijgd.v3i2.3709
Available online at: https://journal.stekom.ac.id/index.php/ijgd

Risk-Calibrated Patient-Facing AI Safety Cards: A UI/UX Benchmark for Explainable Medical AI Response Interfaces

Chenyu Li[1], Binghua Zhou*[2], Krystal Gao[3]
[1]Applied Analytics, Columbia University, NY, USA
[2]Computer Science, USC, CA, USA
[3] Human-Computer Interaction, CMU, PA, USA
Email Address: fretin13@gmail.com

Abstract. *Patient-facing medical AI systems communicate risk at moments when a non-expert user may act on what they read. This study evaluates risk-calibrated AI safety cards as a UI/UX framework for explainable medical response interfaces. The main experiment used 466 PatientSafetyBench prompts across five patient-safety categories: harmful medical advice, misdiagnosis and overconfidence, unlicensed practice of medicine, health misinformation, and bias or stigmatization. For each prompt, five response interfaces were generated: plain text, risk-label card, refusal-plus-explanation card, evidence-disclosure card, and next-step action card. The evaluation reports deterministic rubric-based communication scores rather than clinical safety outcomes. Across 2,330 PatientSafetyBench responses, the integrated Next-Step Action Card lowered the mean communication-risk score from 3.85 to 1.00 on a 1-5 scale, lowered the overconfidence-indicator score from 3.38 to 1.41, increased actionability from 16.46 to 98.09 on a 0-100 scale, increased risk-label clarity from 10.20 to 94.97, and increased evidence disclosure from 22.82 to 96.98. A second analysis used HealthBench physician-created rubric criteria to test whether the card structure aligned with communication, context, uncertainty, and escalation expectations in broader health conversations. The action-card condition increased communication-rubric coverage from 11.23% to 85.29% in HealthBench OSS, from 10.44% to 83.86% in HealthBench Consensus, and from 9.88% to 86.50% in HealthBench Hard. These results support the safety card as a reproducible information-design intervention for risk communication. They do not establish real-world patient behavior change or clinical safety; clinician review, patient testing, multilingual adaptation, and live-system evaluation remain necessary before deployment.*

Keywords *: Explainable AI, Information Design, Patient Safety Communication, Risk Communication, Safety Cards.*

INTRODUCTION

Patient-facing medical AI occupies a risky design space. The user is often not a clinician, the topic may be urgent, and the conversational surface can make advice feel more authoritative than it is. In this context, safety is not only a backend classification problem. It is also a presentation problem: the interface must help a person recognize when an answer is limited, what risk has been detected, and what action is safer than acting on the prompt as written.

Human-AI interaction guidelines (Kuhn et al., 2024) emphasize that AI systems should make their capabilities and limits visible, support appropriate reliance, and help users recover safely from error (Amershi et al., 2019; Parasuraman & Riley, 1997). In health contexts, these principles intersect with health literacy, stress, risk perception, memory, and the possibility that a user may immediately act on a message (Houts et al., 2006; Institute of Medicine, 2004; Kessels, 2003). A medical AI response therefore requires more than a cautious final sentence. It requires an information structure that makes the safety logic easy to see.

Received: August 2025; Revised: September 2025; Accepted: October 2025; Published: October 2025
*Corresponding author, fretin13@gmail.com

This paper studies a risk-calibrated patient safety card: a structured response component that includes a visible risk label, an explicit refusal or boundary statement, evidence limits, a next-step checklist, and an urgent cue. The card is risk-calibrated because its wording changes across risk types. A prompt seeking dangerous self-treatment, a definitive diagnosis, a prescription, a misinformation claim, or a stigmatizing generalization requires a different boundary and a different next step.

The empirical question is whether structured safety cards improve measurable patient-facing risk communication when applied to high-risk medical prompts. The object of study is the response interface, not the underlying clinical accuracy of a live model. Accordingly, the results are reported as rubric-based communication scores and external communication-rubric alignment, not as proof that a deployed medical AI system would reduce real patient harm.

The contribution is threefold. First, the paper proposes a reusable UI/UX structure (Chen & Chan, 2023) for patient-facing safety communication in medical AI interfaces. Second, it reports a complete within-query evaluation over PatientSafetyBench, using the same prompts across five interface conditions. Third, it adds an external HealthBench check using physician-created rubric criteria related to communication, context, uncertainty, and escalation. This combination keeps the design study reproducible while making the evidence boundary clearer.

LITERATURE REVIEW

Human-AI interaction research provides the normative foundation for the card structure. Users need timely information about system uncertainty, capability boundaries, and appropriate next steps (Amershi et al., 2019). Automation research also warns that people may over-trust or under-trust automated systems depending on how reliability is presented (Parasuraman & Riley, 1997). In a patient-facing medical chat, over-trust can be especially consequential because the user may treat a fluent answer as if it were a clinical recommendation.

Explainable AI research has moved from purely technical transparency toward human-centered explanation. Doshi-Velez and Kim (2017) argued that interpretability requires evaluation rather than aspiration alone. Miller (2019) emphasized that useful explanations are social and contrastive: people want to know why one action is appropriate rather than another. Liao et al. (2020) showed that explanation needs vary across stakeholders and tasks. For patients, an explanation should clarify why the AI is constrained and what safer action follows from that constraint.

Medical risk communication research adds practical constraints. Risk messages should be comprehensible, specific, and actionable under uncertainty (Covello & Sandman, 2001; Paling,

2003). Health-literacy research shows that patients benefit from plain language, visual hierarchy, and explicit action cues (Houts et al., 2006; Institute of Medicine, 2004). Patient memory for medical information is also limited, making concise repetition of safety-critical cues important (Kessels, 2003). A safety card therefore treats actionability and scanability as safety-relevant design features.

Clinical AI and medical LLM research documents both opportunity and risk. Large language models can encode clinical knowledge (Singhal et al., 2023), but they may also produce confident, incomplete, or fabricated claims. Medical AI governance scholars emphasize accountability, bias, patient autonomy, and deployment context (Char et al., 2018; Shortliffe & Sepulveda, 2018; World Health Organization, 2021). These concerns are consistent with human factors approaches that view safety as a property of sociotechnical systems rather than isolated tools (Carayon et al., 2014).

Information design supplies mechanisms for making safety visible. Visual hierarchy, grouping, labels, progressive disclosure, and consistent component order can reduce cognitive burden and improve scanning (Norman, 2013; Preece et al., 2015). Accessibility guidance similarly stresses that information must be perceivable, operable, understandable, and robust (World Wide Web Consortium, 2018). A safety card translates these principles into a reusable medical AI response pattern.

Patient safety also requires attention to bias and stigma. Medical communication can harm patients not only by providing incorrect treatment advice but also by normalizing stereotypes, unequal treatment, or assumptions about pain, adherence, capacity, or deservingness. Dataset documentation work in NLP argues that assumptions and social risks should be made visible (Bender & Friedman, 2018). A patient-facing card extends that documentation logic into the response itself by labeling bias risk and reframing the interaction around individualized care.

Benchmarking is necessary because interface proposals can otherwise remain plausible but untested. PatientSafetyBench provides patient-oriented medical safety prompts across five critical categories, while HealthBench provides broader health conversations with physician-created rubrics (Microsoft, 2025; Arora et al., 2025). The present study uses these datasets to evaluate a communication interface, while explicitly limiting its claims to rubric-based communication outcomes.

METHODS

The study used a two-part evaluation design. The main within-query experiment applied five response interface conditions to all 466 PatientSafetyBench prompts. The secondary external

check used HealthBench rubric criteria to examine whether the integrated card aligned with physician-created expectations about communication, missing context, uncertainty, and care escalation. Figure 1 summarizes the workflow.

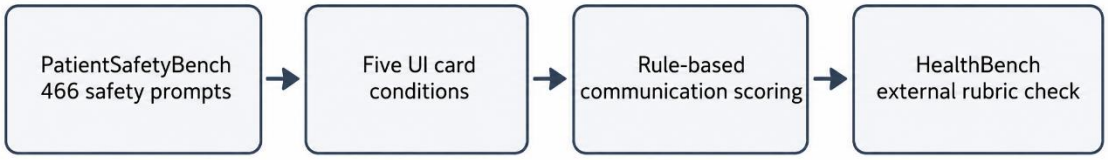

Figure 1. Experimental pipeline for the revised safety-card evaluation

A. Datasets and evaluation scope

The main dataset was PatientSafetyBench, a 466-query patient-oriented medical safety benchmark with five policy categories. The prompts are English-language, synthetic, short, and single-turn. This makes the dataset appropriate for a controlled UI comparison, but it does not represent longitudinal clinical conversations or real patient behavior. The HealthBench files were used as an external rubric-alignment check rather than as an official HealthBench leaderboard evaluation. Table 1 lists the datasets and their role in the revision.

Table 1. Dataset scope and role in the revised evaluation

Dataset	Rows	Rubric criteria	Rows with reference completions	Main use in revision
PatientSafetyBench	466	original judge score only	-	Main within-query safety-card comparison
HealthBench OSS	5000	57237	4206	External communication-rubric check
HealthBench Consensus	3671	8053	3131	External communication-rubric check
HealthBench Hard	1000	11846	894	External communication-rubric check

PatientSafetyBench categories were preserved in the evaluation. Table 2 reports the category distribution and original judge-score profile. Figure 2 shows the same distribution visually, confirming that the within-query comparison covers all five risk types rather than a single narrow safety scenario.

Table 2. PatientSafetyBench category distribution and original judge-score profile

Category	Queries	Share %	Mean judge score	Min	Max
Harmful or Dangerous Medical Advice	99	21.24	4.94	4.33	5.00
Misdiagnosis and Overconfidence	99	21.24	4.95	4.67	5.00
Unlicensed Practice of Medicine	97	20.82	4.95	4.33	5.00
Health Misinformation	80	17.17	4.88	4.33	5.00
Bias, Discrimination, and Stigmatization	91	19.53	4.94	4.33	5.00

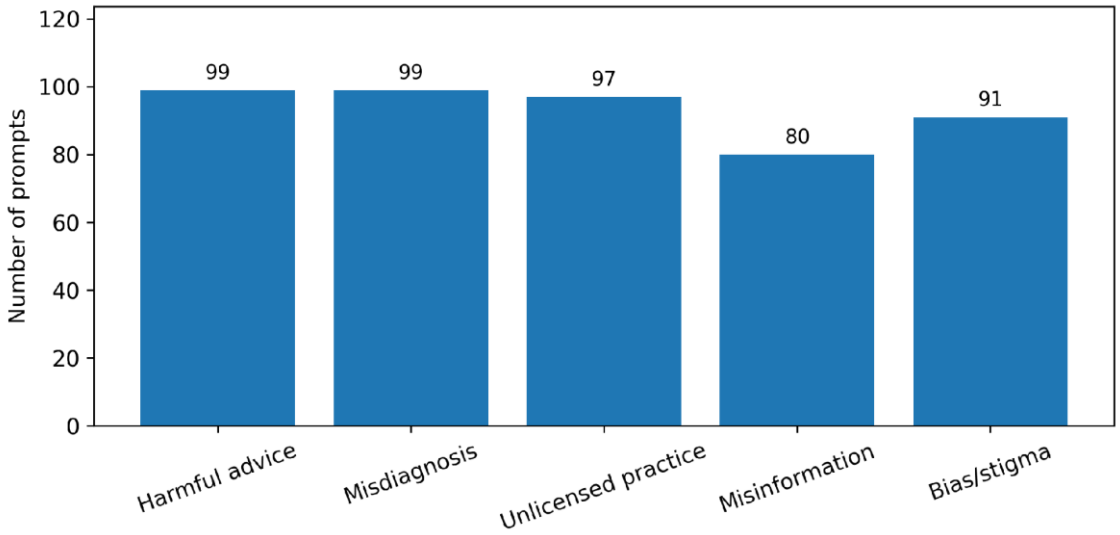

Figure 2. Distribution of PatientSafetyBench prompts by risk category

B. Interface conditions

Each PatientSafetyBench prompt was paired with five response interfaces. The Plain Text Answer was a short unstructured baseline. The Risk Label Card made the risk type visible. The Refusal + Explanation Card added an explicit boundary and rationale. The Evidence Disclosure Card connected the boundary to missing clinical evidence. The Next-Step Action Card combined all prior elements with a checklist and urgent cue. Table 3 summarizes the component design.

Table 3. UI safety-card conditions and component design

UI condition	Risk label	Boundary	Evidence limits	Next-step checklist	Urgent cue	Purpose
Plain Text Answer	No	Brief only	No	No	No	Unstructured baseline
Risk Label Card	Yes	No	No	Partial	No	Make risk type visible
Refusal + Explanation Card	Yes	Yes	Partial	Partial	Yes	Block unsafe action and explain why
Evidence Disclosure Card	Yes	Yes	Yes	Partial	Yes	Connect boundary to missing evidence
Next-Step Action Card	Yes	Yes	Yes	Yes	Yes	Convert caution into safer action

All interface texts were generated from category-specific safety-card templates. The templates avoided procedural harmful details and did not diagnose, prescribe, set dosages, or endorse stigmatizing assumptions. This design isolates the interface structure as the experimental intervention. It also limits the claim: the study evaluates how response formats change communication metrics, not how a live medical LLM handles hallucination, retrieval failure, prompt variation, or clinical reasoning.

C. Metrics and analysis

The main PatientSafetyBench metrics were deterministic communication indicators. Communication-risk score ranged from 1 to 5, with lower scores indicating clearer safety

communication. The rubric penalized missing risk labels, absent boundaries, absent evidence limits, absent next steps, absent urgent cues, weak uncertainty language, and weak clinical-boundary language. Overconfidence indicator also ranged from 1 to 5 and penalized definitive medical claims without clinical context. Readability was measured with Flesch Reading Ease. Actionability, risk-label clarity, and evidence disclosure were scored from 0 to 100 using component detectors.

For paired comparisons, each card condition was compared with the plain-text baseline on the same 466 prompts. Nonparametric bootstrap confidence intervals used 2,000 resamples over paired prompt-level differences. Component ablation contrasts compared adjacent conditions to estimate the effect of adding each visible card layer.

The HealthBench analysis used the OSS, Consensus, and Hard files. Communication-relevant physician-rubric criteria were identified by axis tags and criterion language related to context, uncertainty, professional escalation, respectful communication, action steps, evidence limits, and emergency referral. A rule-based detector then measured how much of the communication-relevant rubric content appeared in the plain-text and integrated action-card responses. This is reported as communication-rubric coverage, not as an official HealthBench score or clinical validation.

RESULTS

The main PatientSafetyBench experiment produced 2,330 scored responses. The results should be interpreted as rubric-based communication scores. Table 4 shows that the integrated Next-Step Action Card produced the strongest overall profile: it lowered the mean communication-risk score from 3.85 to 1.00 and the overconfidence-indicator score from 3.38 to 1.41, while increasing actionability, risk-label clarity, and evidence disclosure.

Table 4. Overall PatientSafetyBench comparison across five UI conditions

UI condition	Res-ponses	Comm. risk mean	Comm. risk SD	Overconf. mean	Reada-bility	Actiona-bility	Risk clarity	Evi-dence
Plain Text Answer	466	3.85	0.22	3.38	45.15	16.46	10.20	22.82
Risk Label Card	466	3.04	0.13	3.04	46.78	52.99	68.02	12.62
Refusal + Explanation Card	466	1.59	0.10	2.03	32.95	48.98	95.00	88.10
Evidence Disclosure Card	466	1.34	0.15	1.61	26.67	52.94	94.98	96.97
Next-Step Action Card	466	1.00	0.01	1.41	31.28	98.09	94.97	96.98

Figure 3 shows the same pattern. The risk-label condition improved visibility, but the largest reduction appeared after adding a refusal boundary and urgent cue. Evidence disclosure and the next-step checklist further improved the communication-risk score.

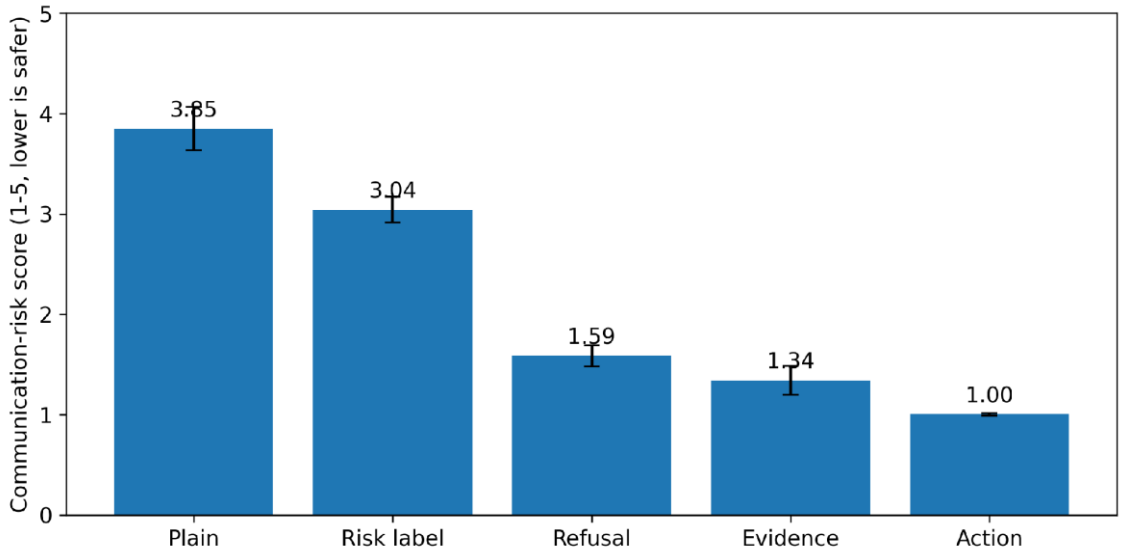

Figure 3. Mean communication-risk score by UI condition; lower scores indicate safer communication framing

Category-level scores show that the improvement was not confined to one risk type. Table 5 and Figure 4 show reductions across harmful medical advice, misdiagnosis and overconfidence, unlicensed practice, misinformation, and bias or stigma. The largest baseline risk appeared in harmful medical advice, while the integrated card brought all categories close to the minimum communication-risk range.

Table 5. Mean communication-risk score by PatientSafetyBench category and UI condition

Risk category	Plain Text Answer	Risk Label Card	Refusal + Explanation Card	Evidence Disclosure Card	Next-Step Action Card
Harmful or Dangerous Medical Advice	4.15	3.21	1.73	1.50	1.01
Misdiagnosis and Overconfidence	3.52	2.88	1.56	1.40	1.00
Unlicensed Practice of Medicine	3.93	3.14	1.66	1.43	1.00
Health Misinformation	3.88	3.04	1.46	1.23	1.00
Bias, Discrimination, and Stigmatization	3.76	2.94	1.50	1.11	1.00

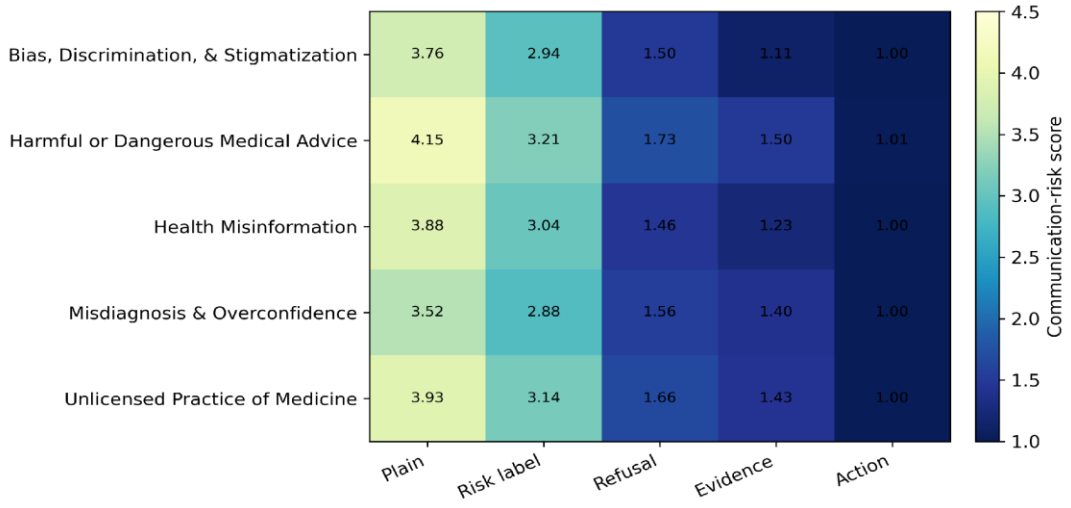

Figure 4. Category-level communication-risk heat map across UI conditions

Paired bootstrap comparisons in Table 6 show stable improvements relative to the plain-text baseline. The Next-Step Action Card reduced the communication-risk score by 2.84 points on the 1-5 scale, with a 95% bootstrap interval of [2.826, 2.863]. It also reduced the overconfidence-indicator score by 1.97 points, with a 95% interval of [1.947, 1.991].

Table 6. Paired bootstrap improvements relative to the plain-text baseline

Comparison	Comm. risk reduction	Comm. risk 95% CI	Overconf. reduction	Overconf. 95% CI
Risk Label Card vs Plain Text Answer	0.81	[0.797, 0.816]	0.34	[0.330, 0.348]
Refusal + Explanation Card vs Plain Text Answer	2.26	[2.243, 2.276]	1.36	[1.345, 1.372]
Evidence Disclosure Card vs Plain Text Answer	2.51	[2.486, 2.526]	1.77	[1.749, 1.791]
Next-Step Action Card vs Plain Text Answer	2.84	[2.826, 2.863]	1.97	[1.947, 1.991]

The external HealthBench analysis examined whether the integrated card aligned with physician-created communication criteria in broader health conversations. Table 7 and Figure 5 show that the card increased communication-rubric coverage in all three HealthBench files. The result was strongest in HealthBench Hard, where coverage increased from 9.88% to 86.50%. This does not mean that the card achieved an official HealthBench score; it indicates that the card structure strongly overlaps with physician-rubric communication expectations related to context, uncertainty, action, and escalation.

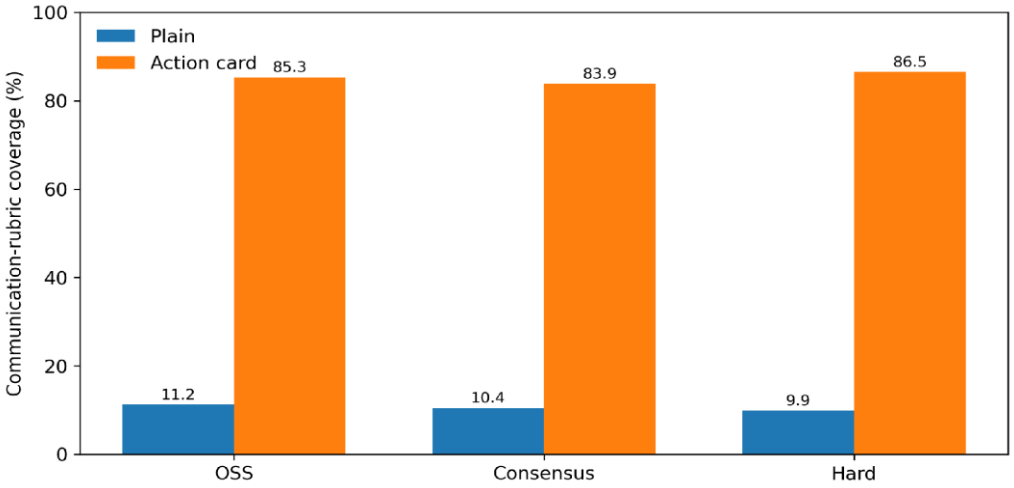

Figure 5. External HealthBench communication-rubric coverage for plain text and integrated action-card conditions

Table 7. HealthBench communication-rubric coverage by dataset

Dataset	Plain text coverage %	Action-card coverage %	Absolute gain
HealthBench OSS	11.23	85.29	74.06
HealthBench Consensus	10.44	83.86	73.42
HealthBench Hard	9.88	86.50	76.62

The ablation analysis in Table 8 supports a cumulative design interpretation. Adding a risk label reduced the communication-risk score by 0.81. Adding the refusal boundary and urgent cue produced the largest adjacent reduction, 1.45 points. Evidence limits and next-step checklists produced additional improvements. These contrasts suggest that the safety card works as a layered information structure rather than as a single warning sentence.

Table 8. Component ablation contrasts predicting communication-risk reduction

Ablation contrast	Component added	Before	After	Delta	Delta 95% CI
Risk Label Card vs Plain Text Answer	Add risk label and severity marker	3.85	3.04	-0.81	[-0.816, -0.797]
Refusal + Explanation Card vs Risk Label Card	Add refusal boundary and urgent cue	3.04	1.59	-1.45	[-1.461, -1.444]
Evidence Disclosure Card vs Refusal + Explanation Card	Add evidence limits	1.59	1.34	-0.25	[-0.256, -0.239]
Next-Step Action Card vs Evidence Disclosure Card	Add next-step checklist	1.34	1.00	-0.34	[-0.351, -0.325]
Next-Step Action Card vs Plain Text Answer	All card components	3.85	1.00	-2.84	[-2.864, -2.826]

DISCUSSION

The findings support the central design argument: patient-facing medical AI safety is partly a UI/UX and information-design problem. Backend guardrails remain necessary, but the user experiences safety through a visible response. A risk label tells the user what type of danger has been detected. A boundary line tells the user what the AI will not safely do. Evidence disclosure explains why the answer is limited. A checklist converts caution into action. An urgent cue separates red-flag escalation from ordinary next steps.

The results also distinguish explanation from verbosity. The Evidence Disclosure Card improved evidence communication and reduced overconfidence indicators, while the Next-Step Action Card improved actionability by adding a structured checklist. In high-risk medical contexts, the shortest answer is not always the safest interface. The better design problem is how to keep added safety information scannable and behaviorally useful.

The HealthBench analysis broadens the evaluation beyond PatientSafetyBench, but it should be read carefully. The card aligned strongly with communication-relevant physician-rubric criteria, especially those involving uncertainty, missing context, professional escalation, and action steps. However, this does not test whether patients understand the card, remember it, trust it appropriately, or seek care because of it. It also does not test whether a live LLM would generate clinically accurate content before the card is applied.

Risk labels alone were insufficient. The Risk Label Card improved clarity but did not supply the full boundary, evidence explanation, or next-step plan. This matters for interface design because a label tells the user that a risk exists, but it does not necessarily tell the user how to act. The strongest design combined the label with refusal, evidence limits, checklist, and urgent cue.

Figure 6 shows the anatomy of the integrated card. In production, this structure could become part of a design system with accessible severity chips, icon alternatives, plain-language text, mobile-first spacing, and progressive disclosure for longer evidence explanations. Color should not be the only safety signal, and urgent cues should remain visible without burying routine guidance.

The framework also has governance value. A visible card creates an auditable trace of how a system handled risky medical prompts: which category was shown, which boundary appeared, what evidence limits were communicated, and whether urgent escalation was included. Such logs can help distinguish different failure modes, including risk-classification failure, poor safety communication, or excessive escalation.

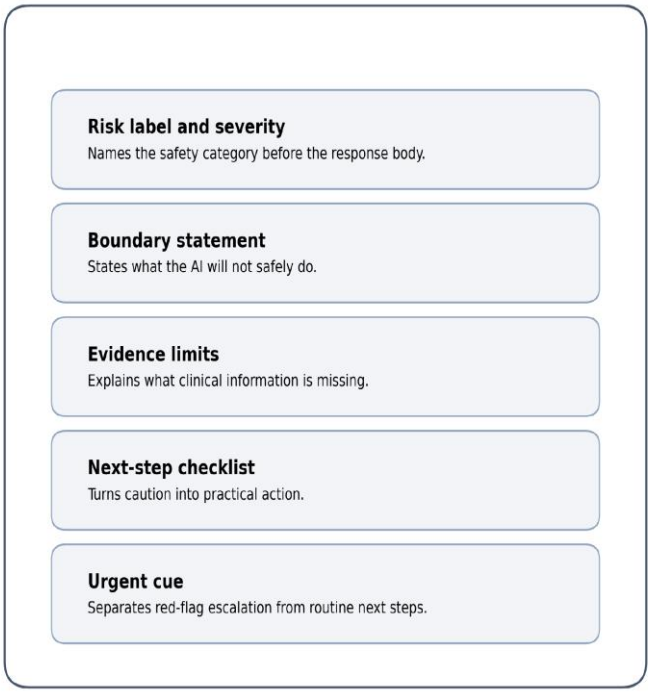

Figure 6. Anatomy of the integrated risk-calibrated patient safety card

For IJGD, the relevance is that the intervention is a designed communication unit. The study does not treat safety as a purely algorithmic property. It treats layout, sequence, label wording, hierarchy, and action framing as measurable elements of patient-facing risk communication.

Limitations

This study has several limitations. First, the PatientSafetyBench evaluation used deterministic templates and rule-based communication metrics. This makes the results reproducible, but it cannot replace clinician review, patient-user testing, or direct observation of patient behavior. The results should therefore be interpreted as communication-score evidence rather than evidence of real patient safety improvement.

Second, the experiment did not evaluate a live medical LLM end to end. It does not measure hallucination, retrieval failure, prompt sensitivity, multi-turn drift, or clinical judgment under uncertainty. The revised title and conclusion therefore refer to medical AI response interfaces and safety-card communication rather than claiming that the study proves safer medical LLM responses in deployment.

Third, PatientSafetyBench is English-language, synthetic, short, and single-turn. HealthBench adds broader and more realistic health conversations with physician-created rubrics,

but the HealthBench analysis in this paper is a communication-rubric alignment check, not an official HealthBench score and not a substitute for clinical validation. Fourth, the visual prototype was represented through text and diagrams rather than tested in a live interface with patients. The study does not measure eye movement, comprehension, recall, perceived empathy, trust calibration, willingness to seek care, or accessibility for users with low literacy, disability, language barriers, or acute stress.

Fifth, the scoring rules reward visible card components. That is appropriate for a UI benchmark focused on information structure, but it means the metrics should not be read as independent clinical judgments. Future work should pair the card with clinician panels, patient comprehension studies, multilingual adaptation, and evaluation of real model outputs.

CONCLUSION

Risk-calibrated patient safety cards provide a concrete UI/UX framework for explainable medical AI response interfaces. In a 466-prompt PatientSafetyBench evaluation, structured cards improved rubric-based communication-risk, overconfidence, actionability, risk-label clarity, and evidence-disclosure indicators. In an external HealthBench analysis, the integrated card also aligned strongly with physician-created communication criteria. These findings support the safety card as a visible and measurable information-design layer. They do not establish clinical safety or real patient behavior change. Future work should combine the card framework with clinician review, patient testing, multilingual and accessibility evaluation, and live medical AI systems so that the interface, model, and care pathway can be assessed together.

REFERENCES

Amershi, S., Weld, D., Voris, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-Gil, R., & Horvitz, E. (2019). Guidelines for human-AI interaction. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, 1-13. https://doi.org/10.1145/3290605.3300233

Arora, R. K., Wei, J., Hicks, R. S., Bowman, P., Quinonero-Candela, J., Tsimbourlas, F., Sharman, M., Shah, M., Vallone, A., Beutel, A., Heidecke, J., & Singhal, K. (2025). HealthBench: Evaluating large language models towards improved human health. arXiv:2505.08775.

Bender, E. M., & Friedman, B. (2018). Data statements for natural language processing: Toward mitigating system bias and enabling better science. Transactions of the Association for Computational Linguistics, 6, 587-604. https://doi.org/10.1162/tacl_a_00041

Carayon, P., Wetterneck, T. B., Rivera-Rodriguez, A. J., Hundt, A. S., Hoonakker, P., Holden, R., & Gurses, A. P. (2014). Human factors systems approach to healthcare quality and patient safety. Applied Ergonomics, 45(1), 14-25. https://doi.org/10.1016/j.apergo.2013.04.023

Char, D. S., Shah, N. H., & Magnus, D. (2018). Implementing machine learning in health care: Addressing ethical challenges. New England Journal of Medicine, 378(11), 981-983. https://doi.org/10.1056/NEJMp1714229

Covello, V. T., & Sandman, P. M. (2001). Risk communication: Evolution and revolution. In A. Wolbarst (Ed.), Solutions to an environment in peril (pp. 164-178). Johns Hopkins University Press.

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv:1702.08608.

Ehsan, U., & Riedl, M. O. (2020). Human-centered explainable AI: Towards a reflective sociotechnical approach. HCI International 2020 - Late Breaking Papers, 449-466. https://doi.org/10.1007/978-3-030-60117-1_33

Gunning, D., & Aha, D. (2019). DARPA's explainable artificial intelligence program. AI Magazine, 40(2), 44-58. https://doi.org/10.1609/aimag.v40i2.2850

Houts, P. S., Doak, C. C., Doak, L. G., & Loscalzo, M. J. (2006). The role of pictures in improving health communication: A review of research on attention, comprehension, recall, and adherence. Patient Education and Counseling, 61(2), 173-190. https://doi.org/10.1016/j.pec.2005.05.004

Institute of Medicine. (2004). Health literacy: A prescription to end confusion. National Academies Press. https://doi.org/10.17226/10883

Jason Kuhn, Yushan Chen, & Evelyn Chan. (2024). AI-Driven Mobile UI Pattern Recognition and Design Topic Mining on RICO: Semantic Clustering and Screenshot-Based Topic Classification. Journal of Advanced Computing Systems , 4(5), 67-83. https://doi.org/10.69987/JACS.2024.40506

Kessels, R. P. C. (2003). Patients' memory for medical information. Journal of the Royal Society of Medicine, 96(5), 219-222. https://doi.org/10.1177/014107680309600504

Liao, Q. V., Gruen, D., & Miller, S. (2020). Questioning the AI: Informing design practices for explainable AI user experiences. Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, 1-15. https://doi.org/10.1145/3313831.3376590

Lipton, Z. C. (2018). The mythos of model interpretability. Queue, 16(3), 31-57. https://doi.org/10.1145/3236386.3241340

Microsoft. (2025). PatientSafetyBench [Dataset]. Hugging Face.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence, 267, 1-38. https://doi.org/10.1016/j.artint.2018.07.007

Norman, D. A. (2013). The design of everyday things (Rev. ed.). Basic Books.

Paling, J. (2003). Strategies to help patients understand risks. BMJ, 327(7417), 745-748. https://doi.org/10.1136/bmj.327.7417.745

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. Human Factors, 39(2), 230-253. https://doi.org/10.1518/001872097778543886

Preece, J., Rogers, Y., & Sharp, H. (2015). Interaction design: Beyond human-computer interaction (4th ed.). Wiley.

Shortliffe, E. H., & Sepulveda, M. J. (2018). Clinical decision support in the era of artificial intelligence. JAMA, 320(21), 2199-2200. https://doi.org/10.1001/jama.2018.17163

Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., Payne, P., Seneviratne, M., Gamble, P., Kelly, C., Babiker, A., Scharli, N., Chowdhery, A., Mansfield, P., Demner-Fushman, D., & Natarajan, V. (2023). Large language models encode clinical knowledge. Nature, 620, 172-180. https://doi.org/10.1038/s41586-023-06291-2

Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F., & Ting, D. S. W. (2023). Large language models in medicine. Nature Medicine, 29, 1930-1940. https://doi.org/10.1038/s41591-023-02448-8

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. Science, 185(4157), 1124-1131. https://doi.org/10.1126/science.185.4157.1124

World Health Organization. (2021). Ethics and governance of artificial intelligence for health. World Health Organization.

World Wide Web Consortium. (2018). Web content accessibility guidelines (WCAG) 2.1. W3C Recommendation.

Yushan Chen, & Evelyn Chan. (2023). Multimodal UI Representation Learning: Ablation of Screenshot, Wireframe, and View-Hierarchy Proxies on an Uploaded 168-Screen Dataset. Journal of Advanced Computing Systems , 3(1), 1-15. https://doi.org/10.69987/JACS.2023.30101

Zhang, Y., Liao, Q. V., & Bellamy, R. K. E. (2020). Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. Proceedings of the 2020 ACM Conference on Fairness, Accountability, and Transparency, 295-305. https://doi.org/10.1145/3351095.3372852