Original Article

Healthc Inform Res. 2025 April;31(2):146-155.
https://doi.org/10.4258/hir.2025.31.2.146
pISSN 2093-3681 • eISSN 2093-369X

HIR
Healthcare Informatics Research

Consensus on the Potential of Large Language Models in Healthcare: Insights from a Delphi Survey in Korea

Ah-Ram Sul, Seihee Kim
Division of Healthcare Research, National Evidence-based Healthcare Collaborating Agency, Seoul, Korea

Objectives: Given the rapidly growing expectations for large language models (LLMs) in healthcare, this study systematically collected perspectives from Korean experts on the potential benefits and risks of LLMs, aiming to promote their safe and effective utilization. **Methods:** A web-based mini-Delphi survey was conducted from August 27 to October 14, 2024, with 20 selected panelists. The expert questionnaire comprised 84 judgment items across five domains: potential applications, benefits, risks, reliability requirements, and safe usage. These items were developed through a literature review and expert consultation. Participants rated their agreement or perceived importance on a 5-point scale. Items meeting predefined thresholds (content validity ratio ≥0.49, degree of convergence ≤0.50, and degree of consensus ≥0.75) were prioritized. **Results:** Seventeen participants (85%) responded to the first round, and 16 participants (80%) completed the second round. Consensus was achieved on several potential applications, benefits, and reliability requirements for the use of LLMs in healthcare. However, significant heterogeneity was found regarding perceptions of associated risks and criteria for safe usage of LLMs. Of the 84 total items, 52 met the criteria for statistical validity, confirming the diversity of expert opinions. **Conclusions:** Experts reached a consensus on certain aspects of LLM utilization in healthcare. Nonetheless, notable differences remained concerning risks and requirements for safe implementation, highlighting the need for further investigation. This study provides foundational insights to guide future research and inform policy development for the responsible introduction of LLMs into the healthcare field.

Keywords: Large Language Models, Generative Artificial Intelligence, Digital Health, Delphi Technique, Korea

Submitted: January 30, 2025
Revised: April 3, 2025
Accepted: April 15, 2025

Corresponding Author
Ah-Ram Sul
Division of Healthcare Research, National Evidence-based Healthcare Collaborating Agency, 3F 400, Neudong-ro, Gwangjin-gu, Seoul 04933, Korea. Tel: +82-2-2174-2790, E-mail: ahramsul@neca.re.kr (https://orcid.org/0000-0003-0331-5529)

I. Introduction

The use of large language models (LLMs) in healthcare is gaining increasing attention. LLMs have potential applications in areas such as clinical workflows, documentation, medical research, and education [1,2]. However, LLMs also have significant limitations. Their outputs depend heavily on training data, which poses risks of bias and inaccuracy. Moreover, achieving the proper balance between artificial intelligence (AI)-driven automation and human expertise remains an ongoing challenge [3].

Many studies have evaluated the utilization of LLMs in healthcare-related tasks [4-6]. However, individual studies

may overlook broader contexts or potential risks. Consequently, a systematic synthesis of existing evidence is necessary to establish a comprehensive understanding [7]. The safe and responsible implementation of LLMs is essential to improve trust in AI-based healthcare technologies and to protect patient interests. Evaluating the medical applications of LLMs can inform future research and development, and appropriate utilization of these models can contribute effectively toward addressing specific healthcare challenges [2].

The Delphi technique has been employed to address complex issues with incomplete knowledge, utilizing a structured and iterative approach for systematic forecasting. It collects expert opinions to reach consensus, thereby generating informed insights [8-10]. Although a previous Delphi study [2] explored the use and adoption of LLMs in healthcare, it was conducted internationally and did not include Asian respondents. Domestic perspectives can differ considerably from those of international experts due to variations in healthcare systems, policies, and cultural factors [9,10]. Therefore, conducting a Delphi study with national-level stakeholders could enhance understanding by generating regionally relevant insights, thus supporting policy recommendations tailored to the local healthcare ecosystem. In this context, the current study aims to evaluate domestic experts' views on anticipated benefits, potential risks, and strategies for safely and effectively using LLMs in healthcare.

II. Methods

1. Study Design

To enhance transparency and reproducibility, this study adhered to established guidelines for conducting and reporting Delphi studies [11,12]. Figure 1 illustrates the overall Delphi survey procedure. The survey fieldwork was conducted in collaboration with Hankook Research, a Seoul-based research firm.

2. Ethical Considerations

This study was approved by the Institutional Review Board of the National Evidence-based Healthcare Collaborating Agency in Seoul, Korea (Approval No. NECAIRB24-012). All panelists provided informed consent prior to beginning the Delphi survey.

3. Questionnaire Development

A rapid literature search was performed on PubMed in July 2024 using the keywords: "large language model" or "LLM," "Delphi," "consensus," and "systematic review." The questionnaire was developed based on key concepts and methodologies identified in the literature, particularly those outlined in a previous Delphi study [2]. It was refined through iterative discussions to ensure clarity, relevance, and comprehensiveness. A pilot test involving three academic experts in medical AI (a medical professional, a regulatory agency practitioner, and a researcher/developer) was conducted to finalize the

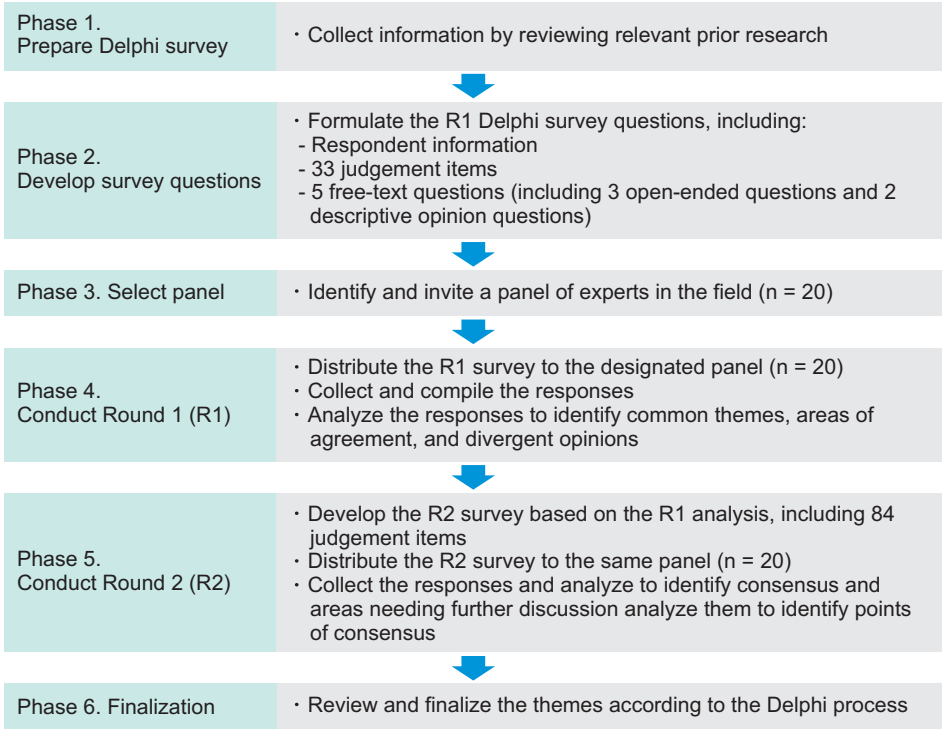

Figure 1. Overview of the Delphi survey procedure.

questionnaire.

The Delphi study covered five key domains: (1) potential applications, (2) benefits, and (3) risks of LLMs in medical practice; (4) reliability requirements for LLM-based systems; and (5) requirements for the safe and appropriate use of LLMs in healthcare. In the first round, domains (1)–(3) were assessed using open-ended questions, while domains (4) and (5) were evaluated with a 5-point Likert scale, also allowing for additional comments. This mixed-method approach enabled exploration of new insights in areas requiring initial investigation, while also providing structured assessment for domains with more established concepts [7].

4. Panel Selection

The Delphi panel comprised 20 Korean experts with substantial knowledge regarding medical applications of LLMs, including government officials. The Delphi method utilizes selected expert panels to offer feedback on specialized topics. Although no standardized panel size is mandated, as statistical representation is not the goal, the composition should be flexible and reflective of practical constraints like time and budget. Recommended panel sizes for healthcare-related Delphi research vary, with some guidelines suggesting 8–23 participants [7] and others identifying 20–30 experts as common [2]. Given the complexity of LLMs and the limited number of relevant experts in Korea, the target panel size was set at 20 participants, consistent with recommendations for effective Delphi studies in this domain.

The panel recruitment involved three approaches: (i) requesting nominations from six prominent academic societies—including the Korean Society of Artificial Intelligence in Medicine, Korean Society of Medical Informatics, and Korean Smart Healthcare Association—with each society asked to nominate 2–4 members; (ii) inviting representatives from five relevant governmental agencies, such as the Ministry of Food and Drug Safety and affiliated organizations of the Ministry of Health and Welfare (Korea Health Industry Development Institute and Korea Health Information Service); and (iii) individual outreach via email to speakers who recently presented on generative pre-trained transformers and LLMs at reputable medical academic conferences within the past year.

5. Data Collection Process

Recent Delphi studies have often utilized a "mini-Delphi" approach comprising two rounds instead of the traditional four, primarily for efficiency [8]. The Delphi method maintains participant anonymity, ensuring that experts can express opinions without hierarchical pressures and revise their responses freely in subsequent rounds [9].

A web-based mini-Delphi survey was conducted from August to October 2024 with the pre-selected panelists. Each survey session required approximately 30 minutes. Participants were informed about voluntary participation, their right to withdraw without penalty, and received compensation of 150,000 Korean won (KRW) per completed session.

In the first-round survey, open-ended or descriptive questions were used to gather initial expert opinions [13]. Responses were then analyzed and categorized by the research team, and new items were introduced based on the feedback obtained.

In the second round, panelists evaluated their degree of agreement or perceived importance for each of the 84 items on a 5-point Likert scale, where a higher score indicated greater agreement or importance. Panelists received first-round results (including average and individual scores) and were given the opportunity to reconsider and revise their ratings based on this feedback.

6. Data Analysis

Qualitative data from the first-round free-text responses underwent thematic analysis [14]. Researchers independently coded responses to identify suggestions for questionnaire revision and new items. Discrepancies were resolved through discussion, and agreed-upon themes were integrated into the second-round questionnaire.

Statistical analyses for second-round responses to the Likert scale items included descriptive statistics (means and standard deviations [SDs]). To evaluate validity, we used the content validity ratio (CVR), degree of convergence (CON), degree of consensus (COS), and coefficient of variation (CV). Analyses were performed using IBM SPSS Statistics for Windows (version 24.0; IBM Corp., Armonk, NY, USA) and Microsoft Excel 2016 (Microsoft Corp., Redmond, WA, USA).

CVR was used to assess expert consensus on each item, with values ranging from –1 to 1; higher positive values indicated stronger agreement [15]. Responses rated 4 or higher on the 5-point scale were considered positive [8]. For the 16-expert panel, items achieving a CVR ≥0.49 were considered to have reached consensus [15], ensuring a rigorous evaluation.

To assess expert agreement, we analyzed both CON and COS metrics, which range from 0 to 1. Lower CON values indicated higher convergence among responses, while higher COS values indicated stronger consensus [8,10]. Convergence was considered sufficient at CON ≤0.50, and con-

 https://doi.org/10.4258/hir.2025.31.2.146

sensus was deemed satisfactory at COS ≥0.75 [8,10]. This combined approach enabled a comprehensive assessment of panel agreement.

The CV measured response stability, ranging from 0 to 1, with lower values indicating greater stability [16]. A CV below 0.50 indicated sufficient stability, signaling the conclusion of the Delphi survey [9].

Since the first round primarily served for brainstorming and initial input, analysis focused on second-round outcomes. The supplementary table includes complete second-round results, while the main text table presents only items meeting all criteria for statistical validity. Items in the main text table were ranked in descending order by mean scores, with identical means sorted by ascending SD. Items with mean scores ≥4 were considered positively received.

III. Results

1. Survey Overview

The Delphi survey involved 20 panelists. Among them, eight were nominated by academic societies, three were representatives of government agencies, and nine were recruited through individual contacts. Basic demographic information for the participating panelists is presented in Table 1. The first round (August 27 to September 12, 2024) yielded 17 responses (85.0% response rate), and the second round (September 24 to October 14, 2024) had 16 responses (80.0% response rate). Despite efforts to encourage participation, the response rate could not be increased. The study concluded after two rounds as initially planned, with the CV indicating stability for all evaluated items by the final round. Notable additional comments from open-ended or descriptive questions in the first round were categorized by field and are presented in the supplementary section for reference (Supplementary Tables S1–S5).

Table 1. Characteristics of survey respondents (n = 17)

Characteristic	Number of first-round survey respondents
Current affiliation	
Academia	7
Healthcare institutions	4[a]
Industry	1
Government agencies	3
Others (legal sector, relevant association)	2
Field of expertise	
HCP	4[a]
AI researcher/developer	10
Others	3
Experience in medical AI	
<5 years	4[a]
5–10 years	9
>10 years	4
Experience level with LLM	
Basic experience	3
Intermediate experience	8[a]
Advanced experience	6

HCP: healthcare professional, AI: artificial intelligence, LLM: large language model.

[a]One individual who participated in the first-round survey but not in the second-round survey was included.

2. Potential Applications of LLMs in Medical Practice

To explore potential applications of LLMs in medicine, the survey addressed three dimensions: clinical task support, patient management support, and documentation support. These dimensions included 15 items, of which 14 met statistical validity criteria (Table 2, Supplementary Table S6).

The highest-scoring items were "generation of patient information leaflets" and "translation of medical documents" (both 4.81), with documentation-support items generally scoring highly. In contrast, "prognosis prediction support" (3.75) and "support for treatment planning" (3.94) received lower scores.

The item "assistance in patient classification and prioritization" met the CVR criterion, indicating expert agreement on its importance. However, it did not satisfy the CON (0.88 > 0.50) or COS (0.56 < 0.75) criteria, reflecting significant disparities in expert opinions.

3. Benefits of LLMs in Medical Practice

We investigated the benefits of LLMs in the healthcare sector across four aspects: healthcare service quality, healthcare systems, healthcare professionals (HCPs), and data processes. Thirteen items on anticipated benefits were evaluated, with 10 meeting statistical validity criteria (Table 3, Supplementary Table S7).

The item receiving the highest level of agreement was "enhanced automation of healthcare tasks" (4.50), followed by "increased efficiency in data processing and extraction" (4.38), "strengthened decision support" (4.13), and "reduction of workload for HCPs" (4.13). Items categorized under

Table 2. Potential applications of LLMs in medical practice (consensus)

Ranking	Item	Category	Agreement
1	Generation of patient information leaflets	Documentation support	4.81 ± 0.40
1	Translation of medical documents	Documentation support	4.81 ± 0.40
3	Assistance in medical document preparation	Documentation support	4.69 ± 0.48
4	Assistance in patient record keeping	Patient management support	4.69 ± 0.60
5	Summarization of medical documents	Documentation support	4.63 ± 0.50
6	Patient counseling	Patient management support	4.56 ± 0.51
6	Patient education and information provision	Patient management support	4.56 ± 0.51
6	Support for insurance claims and management	Documentation support	4.56 ± 0.51
9	Assistance in data retrieval and extraction	Documentation support	4.56 ± 0.63
10	Appointment scheduling and calendar management	Patient management support	4.44 ± 0.89
11	Diagnostic assistance	Clinical workflow support	4.25 ± 0.58
12	Personal health management	Patient management support	4.00 ± 0.63
13	Support for treatment planning	Clinical workflow support	**3.94 ± 0.77**
14	Prognosis prediction support	Clinical workflow support	**3.75 ± 0.68**

Values are presented as mean ± standard deviation.
The results represent panelists' evaluations using a 5-point Likert scale to assess agreement with each item. Items with a mean score below 4.0 were indicated in bold, as scores of 4.0 and above were considered to indicate positive responses.
LLM: large language model.

Table 3. Benefits of LLMs in medical practice (consensus)

Ranking	Item	Agreement
1	Enhanced automation of healthcare tasks	4.50 ± 0.63
2	Increased efficiency in data processing and extraction	4.38 ± 0.62
3	Strengthened decision support	4.13 ± 0.50
4	Reduction of workload for HCPs	4.13 ± 0.89
5	Rapid diagnosis and treatment support	4.06 ± 0.57
6	Enhancement of the quality of healthcare services	4.00 ± 0.82
7	Facilitation of patient-HCP interaction	**3.94 ± 0.68**
7	Support for education and training of HCPs	**3.94 ± 0.68**
9	Provision of personalized healthcare services	**3.81 ± 0.66**
10	Improved interoperability among healthcare systems	**3.75 ± 1.06**

Values are presented as mean ± standard deviation.
The results represent panelists' evaluations using a 5-point Likert scale to assess agreement with each item. Items with a mean score below 4.0 were indicated in bold, as scores of 4.0 and above were considered to indicate positive responses.
LLM: large language model, HCP: healthcare professional.

"improvement of the healthcare system" received relatively lower scores.

Three items—"reduction of medical errors," "cost savings in healthcare," and "enhanced health outcomes"—did not meet the CVR criterion, indicating insufficient expert agreement. Additionally, "reduction of medical errors" failed to meet both CON and COS criteria.

4. Risks of LLMs in Medical Practice

The survey evaluated risks of LLMs in medicine across four dimensions: medical aspects, implications for HCPs, patient concerns, and data protection issues. Of the 21 items assessed, only five fully met statistical validity criteria (Table 4, Supplementary Table S8).

The risks rated highest by experts were "misinformation

 https://doi.org/10.4258/hir.2025.31.2.146

Table 4. Risks of LLMs in medical practice (consensus)

Ranking	Item	Agreement
1	Risks of misinformation due to hallucination	4.31 ± 0.48
2	Risks of inaccurate communication	4.19 ± 0.54
3	Risks of biased decision making	4.00 ± 0.82
4	Accessibility issues for the older adult patients	4.00 ± 1.10
5	Changes in employment within the healthcare sector	**3.94 ± 0.77**

Values are presented as mean ± standard deviation.
The results represent panelists' evaluations using a 5-point Likert scale to assess agreement with each item. Items with a mean score below 4.0 were indicated in bold, as scores of 4.0 and above were considered to indicate positive responses.
LLM: large language model.

Table 5. Reliability requirements for LLM-based systems in healthcare (consensus)

Ranking	Item	Importance
1	Validating performance in real-world environments	4.69 ± 0.48
2	Ensuring the reliability of output results	4.56 ± 0.51
3	Establishing continuous monitoring and management systems for performance variability[a]	4.50 ± 0.63
4	Guaranteeing reproducible results	4.31 ± 0.48
5	Quality management of training data	4.31 ± 0.60
6	Compliance with privacy protection regulations	4.31 ± 0.79
7	Establishing control mechanisms or human intervention procedures	4.25 ± 0.68
8	Measures to enhance the explainability of model predictions and recommendations	4.25 ± 0.93
9	Ensuring system robustness against diverse inputs	4.19 ± 0.54
10	Establishing a standardized quality assessment framework	4.19 ± 0.75
11	Adherence to healthcare regulatory standards	4.13 ± 0.72
12	Conducting simulation testing based on real-world use scenarios	4.13 ± 0.89
13	Validation of accuracy	4.13 ± 1.02

Values are presented as mean ± standard deviation.
The results represent panelists' evaluations using a 5-point Likert scale to assess importance of each item. Items with a mean score of 4.0 and above were considered to indicate positive responses.
LLM: large language model.
[a]This item was added in the second-round survey.

due to hallucination" (4.31) and "inaccurate communication" (4.19).

Most risk-related items did not achieve statistical validity; notably, several items—including "erosion of trust in HCPs," "security vulnerabilities," "decreased patient trust," "re-identification of patient data," "decreased patient-professional interaction," and "misdiagnosis due to inaccurate results"—had CVR values of zero or below, reflecting insufficient expert support.

5. Reliability Requirements for LLM-Based Systems in Healthcare

The survey explored reliability requirements for LLM-based systems in healthcare in four areas: regulation and interoperability, system robustness and safety, system performance and reliability, and explainability. Fourteen items were evaluated, with 13 meeting statistical validity criteria (Table 5, Supplementary Table S9).

All items had an average importance score of 4 or higher. The highest-rated items included "validating performance in real-world environments" (4.69), followed by "ensuring the reliability of output results" (4.56), and "establishing continuous monitoring and management systems for performance variability" (4.50).

The item "ensuring compatibility with existing healthcare systems" met the CVR criterion but did not satisfy CON (0.88 >

Table 6. Requirements for safe and appropriate use of LLMs in healthcare (consensus)

Ranking	Item	Importance
1	Establishing quality assessment criteria	4.56 ± 0.51
2	Enhancing training on LLMs and latest AI technologies for HCPs	4.38 ± 0.62
3	Designing workflow through collaboration with HCPs	4.19 ± 0.54
4	Establishing privacy protection guidelines	4.19 ± 0.75
5	Defining accountability for incidents arising in medical practices utilizing LLMs[a]	4.06 ± 1.00
6	Enhancing transparency and accountability in information provision	4.00 ± 0.63
6	Integration with EHRs	4.00 ± 0.63
8	Establishing guidelines for the interpretation and utilization of LLM results	**3.94 ± 0.77**
9	Establishing collaborative frameworks with diverse stakeholders	**3.88 ± 0.72**
10	Developing compensation models for the utilization of LLM-based systems	**3.81 ± 0.75**

Values are presented as mean ± standard deviation.

The results represent panelists' evaluations using a 5-point Likert scale to assess importance of each item. Items with a mean score below 4.0 were indicated in bold, as scores of 4.0 and above were considered to indicate positive responses.

LLM: large language model, HCP: healthcare professional, EHR: electronic health records.

[a]This item was added in the second-round survey.

0.50) or COS (0.61 < 0.75) criteria, indicating notable divergence in expert opinions.

6. Requirements for Safe and Appropriate Use of LLMs in Healthcare

The survey assessed requirements for safe and appropriate LLM utilization in healthcare across five dimensions: data protection standards, legal and ethical considerations, system performance management standards, clinical utilization and integration, and workforce capability enhancement. Twenty-one items were evaluated, with 10 satisfying all statistical validity criteria (Table 6, Supplementary Table S10).

The most significant item was "establishing quality assessment criteria" (mean score of 4.56), followed by "enhancing training on LLMs and latest AI technologies for HCPs" (4.38), "designing workflows in collaboration with HCPs" (4.19), and "establishing privacy protection guidelines" (4.19).

Items such as "creating independent storage locations," "mandating digital watermarks," and "clarifying patient consent procedures" had CVR values of zero or below, indicating inadequate expert support.

Certain items met the CVR criterion but failed to achieve the CON and COS criteria, suggesting diverse expert views. These included "establishing a compliance framework for legal regulations," "setting data security standards," "promoting cultural change and organizational acceptance," and "establishing regular system performance evaluation procedures."

IV. Discussion

To the best of our knowledge, no previous systematic attempt has been made to collect Korean experts' perspectives on the potential applications of LLMs in healthcare. This study represents the first Delphi investigation specifically focusing on healthcare-related LLMs in Korea, conducted during the fall of 2024. Our findings were analyzed immediately before the release of the "Guidelines for the Approval Review of Generative AI Medical Devices," published by the Korea Ministry of Food and Drug Safety (MFDS) and the Korea National Institute of Food and Drug Safety Evaluation (NIFDS) on January 24, 2025 [17]. Thus, our study is timely and relevant.

Because LLMs are emerging technologies with inherent uncertainties, evaluating their potential through traditional evidence-based methods is challenging. Therefore, systematically gathering expert opinions provides valuable insights [18]. The panelists in this study were representative experts with relevant domain knowledge who voluntarily participated due to their motivation. Notably, six out of 20 panelists also served on the expert advisory panel responsible for developing the MFDS-NIFDS guidelines (two from healthcare, three from academia, and one from industry).

While Denecke's study [2] included international participants, our research focused explicitly on Korean experts to reflect local contexts. The lack of consensus observed in our findings highlights the complexity involved in integrating LLMs into healthcare, a result consistent with Denecke's

 https://doi.org/10.4258/hir.2025.31.2.146

findings [2]. In both studies, experts recognized the potential benefits of LLMs but also raised various concerns. Furthermore, both emphasized the importance of incorporating diverse perspectives to facilitate responsible LLM implementation in healthcare.

LLMs are gaining significant attention within healthcare, yet their use presents considerable challenges. Due to their probabilistic nature, LLMs can produce misleading or incorrect information, potentially resulting in misdiagnoses or inappropriate treatment recommendations. This issue is further complicated by automation bias, which can lead HCPs to uncritically accept inaccurate but plausible outputs [19,20]. Although concerns about such risks are well-documented [19], the level of consensus on specific risks associated with LLM use in medical practice was lower in our study compared to other domains. Differences in familiarity with AI technologies [21] and variations in risk tolerance among stakeholders [22] may explain these discrepancies. Addressing these gaps requires stronger collaboration among AI developers, HCPs, regulators, and policymakers. Harmonizing diverse viewpoints will enable improved risk management, strengthen public trust, and maximize the societal value of AI, thereby ensuring the safe and appropriate adoption of LLMs in healthcare [23].

Our survey identified real-world validation as the most critical requirement for ensuring the reliability of LLM-based healthcare systems. Successful clinical deployment requires assessing LLMs in real-world environments to evaluate their effectiveness, reliability, and potential barriers, including acceptance by HCPs and patients. Additionally, long-term impacts on HCP competencies, patient engagement, and healthcare quality should be explored. Evaluating LLM performance across diverse clinical contexts and patient populations will also help refine application configurations and user perceptions [2,24].

Our findings underscore the necessity of standardized quality assessment criteria to ensure the safe and effective use of LLMs in healthcare. As integration of LLMs into clinical practice draws near, accurate evaluation of their applications becomes increasingly important. The 2025 generative AI medical device guidelines [17] have been updated to reflect unique characteristics of generative AI [20]. These revisions include incorporating language model-specific metrics, such as bilingual evaluation understudy (BLEU) and bidirectional encoder representations from transformers (BERT) scores, alongside traditional quantitative measures. Furthermore, the guidelines emphasize cybersecurity and data protection, mandating the submission of comprehensive cybersecurity documentation to safeguard medical data and patient privacy [17].

Despite advancements, performance evaluation metrics for LLMs in healthcare remain largely unstandardized. Existing technical indicators often lack specificity and practical clarity, resulting in limited applicability [25]. Moreover, established benchmarks for assessing AI performance within medical contexts remain scarce, highlighting an urgent need for comprehensive benchmarks tailored to healthcare workflow requirements [26]. Efforts such as the "Minimum Reporting Items for Clear Evaluation of Accuracy Reports of Large Language Models in Healthcare (MI-CLEAR-LLM)" checklist [27], promoting transparent reporting of LLM accuracy studies, and systematic reviews evaluating LLM performance in clinical settings [28], provide foundational steps. Building upon these initiatives, generating appropriate clinical evidence demonstrating the safety and effectiveness of LLM-based medical devices will become crucial as clinical implementation approaches.

Our survey also emphasized the critical importance of education and training for HCPs to ensure the safe and appropriate use of LLMs. Ensuring clinicians are knowledgeable and capable of effectively utilizing LLM tools requires a multifaceted approach. Focused research within healthcare is essential to assess specific LLM capabilities and performance. Additionally, comprehensive education addressing operational mechanisms, benefits, biases, potential errors, and ethical implications is essential. Such training will enable HCPs to responsibly and effectively utilize LLMs. Given the rapid technological advances, ongoing education is necessary to maintain user competence and awareness of both the potential and limitations of these technologies in clinical practice [29]. This aligns with the themes discussed in the editorial [30]. Human-AI interaction plays a crucial role in successful clinical AI implementation. Importantly, LLM-based medical AI devices not only carry risks of hallucination and sycophancy but also have outputs that can be significantly influenced by prompt formulation. Therefore, tailored guidance or educational programs should be developed for HCPs, enhancing their training in basic technical knowledge and prompt engineering skills, thereby facilitating more effective human-AI interactions and optimal use of LLM-based AI tools.

While this study provides valuable insights, certain limitations must be acknowledged. First, Delphi survey results should be viewed as reference material for decision-making rather than definitive conclusions, as consensus does not necessarily indicate the most accurate answer [18]. Second,

concerns about selection bias in survey items exist. Closed-format questions may have limited the range of expert responses, influencing consensus outcomes. Additionally, our expert panel may not represent all perspectives in this field, and constraints such as limited time and potential non-response bias should also be considered. Furthermore, the rapid pace of technological advancement may limit the long-term relevance of some findings. Despite these limitations, this study offers meaningful implications to inform future research directions and policy development in this rapidly advancing domain.

This study evaluated the potential utilization of LLMs in healthcare and highlighted significant heterogeneity in expert opinions across various domains. The observed variations regarding risks associated with LLMs and essential conditions for their safe implementation underscore the importance of continued research and discussion. The findings, contextualized within Korea's unique environment, are expected to play a vital role in shaping relevant policies and guiding decision-making processes.

Conflict of Interest

No potential conflict of interest relevant to this article was reported.

Acknowledgments

The authors sincerely appreciate the cooperation of the societies and associations that enabled this study by recommending relevant experts. Furthermore, the authors express their gratitude to the respondents who participated in this survey.

This work was supported by the National Evidence-based Healthcare Collaborating Agency (Grant No. NECA-A-2024-014).

ORCID

Ah-Ram Sul (https://orcid.org/0000-0003-0331-5529)
Seihee Kim (https://orcid.org/0000-0001-9553-8142)

Supplementary Materials

Supplementary materials can be found via https://doi.org/10.4258/hir.2025.31.2.146.

References

1. Kim K, Cho K, Jang R, Kyung S, Lee S, Ham S, et al. Updated primer on generative artificial intelligence and large language models in medical imaging for medical professionals. Korean J Radiol 2024;25(3):224-42. https://doi.org/10.3348/kjr.2023.0818
2. Denecke K, May R; LLMHealthGroup; Rivera Romero O. Potential of large language models in health care: Delphi study. J Med Internet Res 2024;26:e52399. https://doi.org/10.2196/52399
3. Younis HA, Eisa TA, Nasser M, Sahib TM, Noor AA, Alyasiri OM, et al. A systematic review and meta-analysis of artificial intelligence tools in medicine and healthcare: applications, considerations, limitations, motivation and challenges. Diagnostics (Basel) 2024;14(1):109. https://doi.org/10.3390/diagnostics14010109
4. Ferber D, Wiest IC, Wolflein G, Ebert MP, Beutel G, Eckardt JN, et al. GPT-4 for information retrieval and comparison of medical oncology guidelines. NEJM AI 2024;1(6):AIcs2300235. https://doi.org/10.1056/AIcs2300235
5. Kumar A, Wang H, Muir KW, Mishra V, Engelhard M. A cross-sectional study of GPT-4–based plain language translation of clinical notes to improve patient comprehension of disease course and management. NEJM AI 2025;2(2):AIoa2400402. https://doi.org/10.1056/AIoa2400402
6. Unlu O, Shin J, Mailly CJ, Oates MF, Tucci MR, Varugheese M, et al. Retrieval-augmented generation–enabled GPT-4 for clinical trial screening. NEJM AI 2024;1(7):AIoa2400181. https://doi.org/10.1056/AIoa2400181
7. Shang Z. Use of Delphi in health sciences research: a narrative review. Medicine (Baltimore) 2023;102(7):e32829. https://doi.org/10.1097/MD.0000000000032829
8. Kim JY, Shin YI, Yang SH. A study on the construction of non-face-to-face lecture of KAOMPT: Delphi survey research to post COVID-19 untact era. J Korean Acad Orthop Man Physi Ther 2021;27(1):1-11. https://doi.org/10.23101/kaompt.2021.27.1.1
9. Kim SH, Joo HJ, Kim JY, Kim HJ, Park EC. Healthcare policy agenda for a sustainable healthcare system in Korea: building consensus using the Delphi method. J Korean Med Sci 2022;37(39):e284. https://doi.org/10.3346/jkms.2022.37.e284
10. Nasa P, Jain R, Juneja D. Delphi methodology in healthcare research: how to decide its appropriateness. World J Methodol 2021;11(4):116-129. https://doi.org/10.5662/wjm.v11.i4.116
11. Junger S, Payne SA, Brine J, Radbruch L, Brearley SG.

Guidance on Conducting and REporting DElphi Studies (CREDES) in palliative care: recommendations based on a methodological systematic review. Palliat Med 2017;31(8):684-706. https://doi.org/10.1177/0269216317690685
12. Niederberger M, Schifano J, Deckert S, Hirt J, Homberg A, Koberich S, et al. Delphi studies in social and health sciences-Recommendations for an interdisciplinary standardized reporting (DELPHISTAR): results of a Delphi study. PLoS One 2024;19(8):e0304651. https://doi.org/10.1371/journal.pone.0304651
13. Choi M, Kim M, Kim JA, Chang H. Building consensus on the priority-setting for national policies in health information technology: a Delphi survey. Healthc Inform Res 2020;26(3):229-37. https://doi.org/10.4258/hir.2020.26.3.229
14. Braun V, Clarke V. Toward good practice in thematic analysis: avoiding common problems and be(com)ing a knowing researcher. Int J Transgend Health 2022;24(1):1-6. https://doi.org/10.1080/26895269.2022.2129597
15. Lawshe CH. A quantitative approach to content validity. Pers Psychol 1975;28(4):563-75. https://doi.org/10.1111/j.1744-6570.1975.tb01393.x
16. Dajani JS, Sincoff MZ, Talley WK. Stability and agreement criteria for the termination of Delphi studies. Technol Forecast Soc Change 1979;13(1):83-90. https://doi.org/10.1016/0040-1625(79)90007-6
17. Ministry of Food and Drug Safety. Guidelines for the approval review of generative AI medical devices (Guideline for Applicants) [Internet]. Osong, Korea: Ministry of Food and Drug Safety; c2025 [cited at 2025 Jan 30]. Available from: https://www.mfds.go.kr/brd/m_1060/view.do?seq=15628&srchFr=&srchTo=&srchWord=&srchTp=&itm_seq_1=0&itm_seq_2=0&multi_itm_seq=0&company_cd=&company_nm=&page=1.
18. Hohmann E, Cote MP, Brand JC. Research pearls: expert consensus based evidence using the Delphi method. Arthroscopy 2018;34(12):3278-82. https://doi.org/10.1016/j.arthro.2018.10.004
19. Freyer O, Wiest IC, Kather JN, Gilbert S. A future role for health applications of large language models depends on regulators enforcing safety standards. Lancet Digit Health 2024;6(9):e662-72. https://doi.org/10.1016/S2589-7500(24)00124-9
20. Park SH, Kim N. Challenges and proposed additional considerations for medical device approval of large language models beyond conventional AI. Radiology 2024;312(3):e241703. https://doi.org/10.1148/radiol.241703
21. Said N, Potinteu AE, Brich I, Buder J, Schumm H, Huff M. An artificial intelligence perspective: how knowledge and confidence shape risk and benefit perception. Computers in human behavior 2023;149:107855. https://doi.org/10.1016/j.chb.2023.107855
22. Gungor H. Creating value with artificial intelligence: a multi-stakeholder perspective. J Creat Value 2020;6(1):72-85. https://doi.org/10.1177/2394964320921071
23. Hogg HD, Al-Zubaidy M; Technology Enhanced Macular Services Study Reference Group; Talks J, Denniston AK, Kelly CJ, et al. Stakeholder perspectives of clinical artificial intelligence implementation: systematic review of qualitative evidence. J Med Internet Res 2023;25:e39742. https://doi.org/10.2196/39742
24. Chouffani El Fassi S, Abdullah A, Fang Y, Natarajan S, Masroor AB, Kayali N, et al. Not all AI health tools with regulatory authorization are clinically validated. Nat Med 2024;30(10):2718-20. https://doi.org/10.1038/s41591-024-03203-3
25. Park SH, Han K, Lee JG. Conceptual review of outcome metrics and measures used in clinical evaluation of artificial intelligence in radiology. Radiol Med 2024;129(11):1644-55. https://doi.org/10.1007/s11547-024-01886-9
26. Blagec K, Kraiger J, Fruhwirt W, Samwald M. Benchmark datasets driving artificial intelligence development fail to capture the needs of medical professionals. J Biomed Inform 2023;137:104274. https://doi.org/10.1016/j.jbi.2022.104274
27. Park SH, Suh CH, Lee JH, Kahn CE, Moy L. Minimum reporting items for clear evaluation of accuracy reports of large language models in healthcare (MI-CLEAR-LLM). Korean J Radiol 2024;25(10):865-8. https://doi.org/10.3348/kjr.2024.0843
28. Bedi S, Liu Y, Orr-Ewing L, Dash D, Koyejo S, Callahan A, et al. Testing and evaluation of health care applications of large language models: a systematic review. JAMA 2025;333(4):319-28. https://doi.org/10.1001/jama.2024.21700
29. Mirzaei T, Amini L, Esmaeilzadeh P. Clinician voices on ethics of LLM integration in healthcare: a thematic analysis of ethical concerns and implications. BMC Med Inform Decis Mak 2024;24(1):250. https://doi.org/10.1186/s12911-024-02656-3
30. Park SH, Langlotz CP. Crucial role of understanding in human-artificial intelligence interaction for successful clinical adoption. Korean J Radiol 2025;26(4):287-90. https://doi.org/10.3348/kjr.2025.0071