

Guideline-informed AI chat access and perceived mental effort in a simulated home-visit nursing task: a randomized controlled trial

Takemasa Ishikawa^{ID} · Haruka Imura^{ID} · Tomomi Kitazawa^{ID} · Yoshiyuki Takashima^{ID}

Received: 27 May 2026 / Accepted: 21 September 2026
© The Author(s) 2026

Abstract

Background Home-visit nurses often make autonomous clinical judgments in patients' homes, where access to colleagues and evidence-based resources is limited. Guideline-informed artificial intelligence (AI) systems may facilitate access to clinical information, but evidence of their effects on nurses' perceived mental effort remains limited. This study examined whether access to a guideline-informed AI chat system was associated with perceived mental effort, task difficulty, and confidence in answer correctness during a simulated home-visit nursing task.

Methods This web-based randomized controlled trial was conducted in Japan from March 9 to March 27, 2026. Home-visit nurses were randomly allocated 1:1 to AI or non-AI groups. Both groups completed a simulated task involving an older adult with chronic heart failure living alone. The AI group could use a NotebookLM system preloaded with publicly available clinical guidelines, whereas the comparison group was instructed not to use external resources. The primary outcome was perceived mental effort, assessed using a 9-point single-item rating based on Paas' scale. The exploratory secondary outcomes were perceived task difficulty and confidence in answer correctness. The primary analyses used a complete-case approach and included participants who completed the task and provided post-task outcome data. Outcomes were analyzed using analysis of covariance adjusted for home-visit nursing experience, non-home-visit nursing experience, and educational attainment.

Results Of 140 randomized participants, 105 completed the task and were included in the complete-case analysis: 53 in the non-AI group and 52 in the AI group. Compared with the non-AI group, the AI group reported lower perceived mental effort (adjusted mean difference -1.13, 95% CI -1.81 to -0.45), lower perceived task difficulty (-0.74, 95% CI -1.37 to -0.12), and higher confidence in answer correctness (0.71, 95% CI 0.14 to 1.27).

Conclusions Among participants who completed the task, guideline-informed AI chat access was associated with lower perceived mental effort and task difficulty and higher confidence in answer correctness than no access to external resources. These are complete-case, not intention-to-treat estimates; the study neither isolated the independent effect of AI nor evaluated clinical reasoning quality, accuracy, or safety.

Trial registration UMIN Clinical Trials Registry, UMIN000061415. Retrospectively registered on 30 April 2026.

Keywords Artificial intelligence · Home-visit nursing · Clinical reasoning · Evidence-informed practice · Cognitive load · Mental effort · Nursing practice · Randomized controlled trial

This Article in Press is shared early to give you faster access to new research. It is citable and carries a permanent DOI. The final edited version will replace it automatically.



Background

In Japan, home-visit nursing plays a key role in supporting home-based care for the rapidly aging population. The system was developed by integrating medical insurance, long-term care insurance, and welfare services, enabling nurses to deliver a wide range of care across the acute, chronic, and end-of-life stages in patients' homes. Within this framework, home-visit nurses perform continuous monitoring, symptom management, and care coordination in community settings [1].

Home-visit nursing is characterized by care delivery in patients' homes, where nurses must make context-sensitive clinical decisions with a high degree of autonomy [2, 3]. Compared with institutional settings, care is often provided without immediate access to other professionals, and clinical judgments must be made based on limited and evolving information in the home environment [4]. This practice environment may impose substantial cognitive and psychological demands on nurses. In a study of Japanese home-visit nurses, Yoshimatsu and Nakatani reported that job stress was associated with error incidents, noting that home-visit nurses often provide care alone under considerable mental and physical stress [5]. Although clinical guidelines are intended to support care quality, their use in community-based nursing may be constrained by time, workload, limited access to training, and evidence-based resources [6–9]. These challenges highlight the need for accessible and context-adapted forms of decision support in home-visit nursing.

In recent years, large language models (LLMs) have increasingly been explored as tools for supporting access to evidence in nursing and healthcare contexts [10]. Studies have reported the use of LLMs to generate responses to medical questions [11]. In a quasi-experimental study involving 146 nurses, Shokr et al. found that the use of a knowledge-based artificial intelligence (AI) chatbot was associated with a significant improvement in nurses' knowledge [12]. These findings suggest that AI-based tools have the potential to support access to evidence and knowledge acquisition. However, studies have identified important limitations, including the potential to suggest unnecessary tests [13] and discrepancies from expert judgment in decisions, such as hospitalization and severity assessment [14]. Therefore, although LLMs are increasingly examined in healthcare and nursing, their role in context-dependent clinical judgment remains unclear.

Despite the growing interest in AI-based tools for supporting access to evidence and clinical guidelines, important gaps remain. First, existing studies on nursing have primarily relied on non-experimental designs, with limited evidence of the causal effects of AI use on nurses' decision-making [13–15]. Second, although AI and clinical decision support systems have been studied in relation to workflow and guideline adherence, their effects on cognitive outcomes, such as perceived mental effort, workload, and task difficulty remain underexamined [16–18]. Third, most prior research has been conducted in hospital settings, and little is known about how these technologies function in home-visit nursing, where nurses make autonomous decisions in resource-limited environments [19, 20]. Collectively, the current literature remains insufficient to determine how AI-supported access to evidence influences decision-related outcomes among nurses, particularly in community-based settings.

This study aimed to examine whether guideline-informed AI chat access was associated with nurses' subjective decision-related outcomes, including perceived mental effort, task difficulty, and confidence in answer correctness, in a simulated home-visit nursing task. By focusing on these outcomes, this study offers initial evidence on how AI-mediated access to guideline information may influence perceived burden during simulated home-visit nursing tasks.

Methods

Study design

This study aimed to examine whether access to a guideline-informed AI chat system was associated with perceived mental effort, task difficulty, and confidence in answer correctness among home-visit nurses completing a simulated case-based task. We conducted an exploratory web-based randomized controlled trial in Japan between March 9, 2026, and March 27, 2026, comparing AI-supported and non-AI environments for a simulated home-visit nursing task. Participants were randomly allocated in a 1:1 ratio to either the AI or non-AI group. The trial was retrospectively registered in the UMIN Clinical Trials Registry [21]. Prospective registration was not completed because, at the planning stage, the study was regarded primarily as an exploratory web-based simulation task. The primary and secondary outcomes, sample size calculation, planned ANCOVA with prespecified covariates, and Welch's t-test sensitivity analysis were documented in the study protocol submitted to and approved by the ethics committee before participant enrollment. No separate stand-alone statistical analysis plan was prepared; the planned analyses were described in the ethics-approved protocol. The study is reported in accordance with the CONSORT 2025 statement [22]. No patient or public involvement was included in the design, conduct, or reporting of this trial. No changes were made to the primary outcome or prespecified primary ANCOVA after enrollment commenced. The protocol also specified exploratory text-mining analyses of participants' free-text responses in both groups. These analyses are ongoing, are outside the scope of the present report, and are planned to be reported separately.

Participants

The eligible participants were nurses currently engaged in home-visit nursing practice in Japan. The inclusion criteria were engagement in home-visit nursing practice, the ability to read and respond in Japanese, access to a web-enabled device, possession of a Google account required to access NotebookLM, and provision of informed consent.

No restrictions were placed on age, sex, years of experience, or prior use of AI systems. The participants were recruited through open calls via websites and social media, including the website and social media channels of the corresponding author's affiliated center, and professional networks.

Sample size

An a priori sample size estimation was conducted for the primary outcome—perceived mental effort—measured on a 9-point Likert scale. Although the primary analysis was planned using an analysis of covariance (ANCOVA), the sample size was conservatively estimated based on a two-group comparison of means, treating the Likert scale as an approximately continuous variable.

Assuming a two-sided alpha of 0.05, 80% power, and 1:1 allocation ratio, a standardized effect size (Cohen's *d*) of 0.50 was assumed. This corresponded to an expected between-group difference of 1.0 point with an assumed standard deviation of 2.0 points. This value was used for sample-size planning; a minimally important difference has not been established for this measure.

Based on these assumptions, the required sample size was estimated to be 64 participants per group (128 participants in total). Assuming a completion rate of 75%, the target recruitment sample size was set at 86 participants per group. No interim analyses or stopping guidelines were planned.

Randomization and allocation

Participants were automatically allocated to either the AI or non-AI group using a block randomization procedure implemented via Google Apps Script. Randomization was performed using a fixed block size of four and a 1:1 allocation ratio. Within each block, group assignments were randomly permuted using a computer-generated randomization algorithm. Allocation was conducted sequentially using a pregenerated assignment queue, and group assignment was determined automatically at the time of enrollment without researcher involvement. Because of this automated process, allocation concealment was maintained throughout participant enrollment. The participants were then provided with a group-specific URL task via email. Owing to the nature of the intervention, the participants were not blinded to the group allocation. Outcome measures were self-reported, and no blinding of the outcome assessment was implemented.

Intervention and task

The participants completed a case-based clinical reasoning task using a simulated home-visit nursing scenario involving an older adult with chronic heart failure living alone. The case included clinical, functional, and social information reflecting typical home-visit nursing practices and was developed through discussions among the research team, including clinicians experienced in home-visit nursing and researchers in nursing education.

Participants were instructed to assume the role of a home-visit nurse preparing for a scheduled home visit after a family member reported worsening lower extremity edema in the evening.

All participants were asked to respond to three open-ended questions: assessment of the patient's current condition and living situation; key points to observe and assess during the upcoming visit, including rationale; and planned communication and interventions for the patient and family.

In the AI group, participants were allowed to use a guideline-informed AI system (NotebookLM) during the task. Before beginning the task, participants in the AI group were provided with standardized illustrated instructions on how to access and use the AI system, but no additional training session or practice task was provided. The system was preloaded with ten publicly available Japanese clinical guidelines and related professional guidance on heart failure, cardiovascular palliative care, nutritional management, pharmaceutical care, cardiomyopathy, mechanical circulatory support, and home-based inotropic therapy [23–32]. The NotebookLM system was configured by the corresponding author and reviewed by the research team. The corresponding author had not received formal training in NotebookLM configuration. The same NotebookLM notebook, source materials, and system configuration were used for all participants in the AI group and were not modified during the study period. The NotebookLM instruction prompt specified that the user was a home-visit nurse and instructed the system to provide evidence-based advice based on the uploaded materials, with particular emphasis on observation points and care content, to use polite language, and to list the source materials supporting the response at the end. No other custom prompting instructions were applied. This configuration was designed to support access to guideline-based information and reduce reliance on unconstrained model-generated responses. No participant-level data were uploaded to the system. According to Google's official documentation, the content added to NotebookLM is not used to train its foundational AI models directly unless users choose to provide feedback [33]. The participants were instructed not to use any external information sources other than the AI system, which was designed as a chat-based interface that generates responses in natural language, including text and tabular formats, based on the referenced guidelines in response to user queries.

Before the main study, the research team tested the AI system using case-related questions to examine its response characteristics and refine prompts. This internal testing was not a formal evaluation of response accuracy or safety. Participants were not required to use standardized queries and formulated their own questions during the task. Self-reported frequency of questions submitted to the AI system was collected as a descriptive measure of intervention exposure. Detailed interaction logs, the verbatim content of submitted queries, and generated responses were not collected as study data; however, participants self-reported the types of questions they

submitted using predefined categories. Details of the system configuration, source materials, prompting instructions, survey instruments, case-based tasks, and analysis code are available in the Open Science Framework repository [34].

In the non-AI group, the participants completed the same task without access to the AI system and were instructed not to refer to any external resources.

Adverse events were not systematically assessed because the study involved a brief web-based simulation task and no patient-level clinical intervention.

Primary and secondary outcomes

Immediately after completing the task, participants rated the following outcomes on a 9-point Likert scale: perceived mental effort (primary outcome); and perceived task difficulty and confidence in answer correctness (exploratory secondary outcomes).

Perceived mental effort was the primary outcome and was assessed using a 9-point single-item rating based on the Paas mental effort rating scale, which is used as a subjective indicator of cognitive load [35]. Perceived task difficulty and confidence in answer correctness were assessed using study-specific single-item measures developed for this study, rather than previously validated scales. These measures were scaled in the same 1–9 format as the perceived mental effort measure for consistency across post-task assessments. Each scale ranges from 1 to 9, with anchors indicating minimal to maximal levels, such as very low to very high effort, difficulty, or confidence.

Participant characteristics

Participants reported demographic and professional characteristics, including age group (20s to ≥ 60 s); sex; total years of nursing experience; years of home-visit nursing experience; highest educational attainment; frequency of referring to clinical guidelines in routine practice; prior experience using generative AI; frequency of generative AI use in the past three months; professional certifications; and completion of advanced nursing procedures training. In the AI group, additional descriptive measures of intervention use were collected, including self-reported total time referring to the AI chat system, the degree to which information obtained from the system was incorporated into the final response, and the types of questions submitted to the system. These additional variables were not included in the outcome analyses reported in the present manuscript.

Covariates

The primary analysis was adjusted for the following prespecified covariates: years of home-visit nursing experience; years of non-home-visit nursing experience (calculated as the total nursing experience minus home-visit experience); educational attainment (categorized as $\leq$ junior college/vocational vs. $\geq$ university).

Statistical analysis

Analyses were conducted using a complete-case approach and included participants with available outcome data. Descriptive statistics were calculated for all the variables. Continuous variables were summarized as means and standard deviations (SD), and categorical variables were summarized as frequencies and percentages.

The baseline characteristics were summarized for the AI and non-AI groups using standardized mean differences (SMDs). As the participants were randomly allocated, no hypothesis testing was performed for baseline comparisons.

Each outcome was analyzed using analysis of covariance (ANCOVA), with group assignment as the primary independent variable and the three prespecified covariates described above. The adjusted mean differences and 95% confidence intervals (CI) were estimated for each outcome. For descriptive reporting, group-specific means,

standard deviations, and sample sizes are also reported. Cohen's d was calculated using the unadjusted group means and pooled standard deviations, with positive values indicating higher scores in the AI group than in the non-AI group. Because post-task outcome data were not available for participants who did not complete the task, the primary analysis was conducted among participants with available outcome data. No imputation was performed. As a prespecified sensitivity analysis, unadjusted between-group comparisons were conducted using Welch's t -tests for each outcome. As post hoc exploratory sensitivity analyses, the ANCOVA models were additionally adjusted separately for the frequency of generative AI use in the past three months and for age group. Age group was treated as a categorical variable.

No adjustment for multiplicity was applied; therefore, the secondary outcome p values were interpreted as exploratory rather than confirmatory. Statistical significance was set at 0.05 (two-sided). All analyses were conducted using Python version 3.11.

Ethical considerations

This study was approved by the Tekix Inc. Research Ethics Review Committee (approval number: 202601). All participants provided informed consent through a web-based form before enrollment.

Results

Participant flow

Figure 1 illustrates the participant flow. A total of 140 participants provided consent and were randomized, with 70 allocated to the AI group and 70 to the non-AI group. In each group, one participant did not receive the assigned task due to an invalid email address. Among the participants allocated to the AI group, 52 completed the task and provided post-task outcome data and were included in the complete-case analysis, while 17 did not

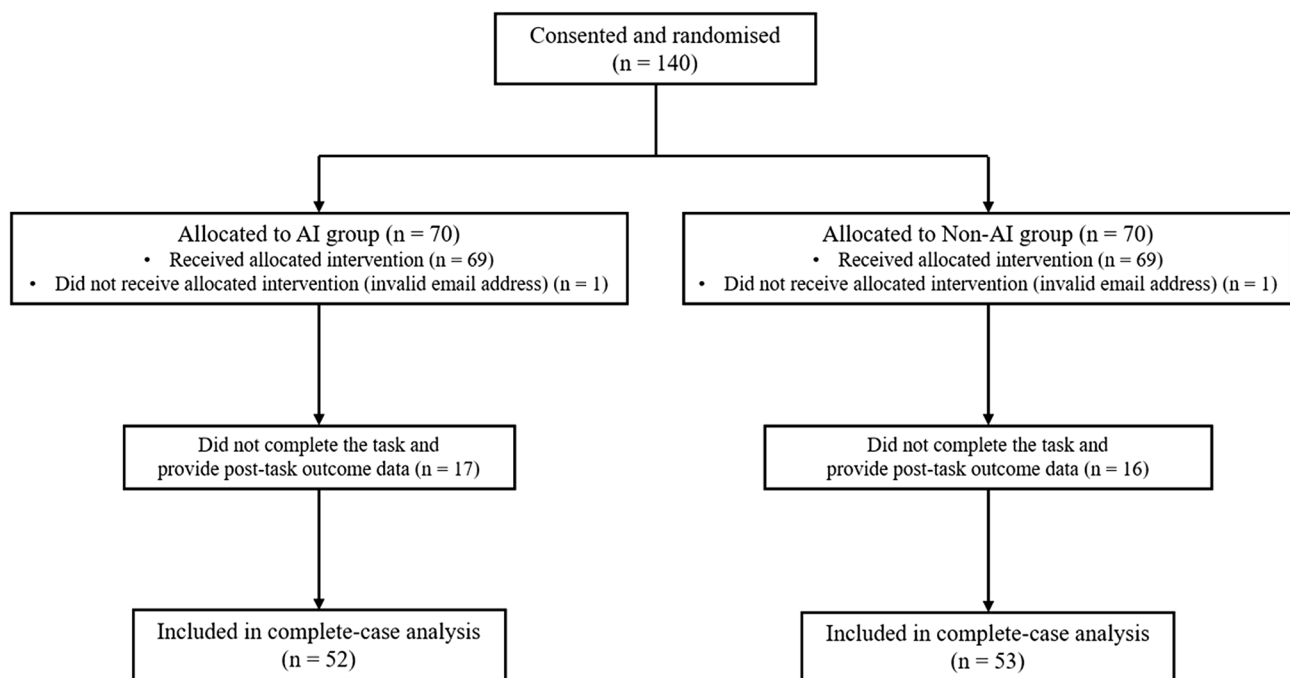


Fig. 1 Participant flow through the web-based randomized controlled trial. AI, artificial intelligence

complete the task and provide post-task outcome data. Among the participants allocated to the non-AI group, 53 completed the task and provided post-task outcome data and were included in the complete-case analysis, while 16 did not complete the task and provide post-task outcome data. The trial ended when the predefined recruitment period was completed. Adverse events or unintended events were not systematically assessed. Among the 52 participants in the AI group, 5 (9.6%) reported not using the AI chat system during the task, 14 (26.9%) asked 1–2 questions, 29 (55.8%) asked 3–5 questions, and 4 (7.7%) asked 6–10 questions. No participants reported asking 11 or more questions.

Participant characteristics

The participant characteristics are summarized in Table 1. The mean total nursing experience was 19.03 years (SD 9.01), and the mean home-visit nursing experience was 7.22 years (SD 5.49). Most participants were female (89/105, 84.8%), and 74 (70.5%) had previously used generative AI. The experience variables were similar between groups, with SMDs of 0.04 for total nursing experience, 0.22 for home-visit nursing experience, and 0.14 for non-home-visit nursing experience. The largest standardized mean difference among the participant characteristics was observed for the frequency of generative AI use in the past three months (SMD 0.61).

Table 1 Participant characteristics by study group

Characteristic	Category	Overall (n = 105)	Non-AI group (n = 53)	AI group (n = 52)	SMD
Total nursing experience, years	Mean (SD)	19.03 (9.01)	18.87 (8.73)	19.19 (9.34)	0.04
Home-visit nursing experience, years	Mean (SD)	7.22 (5.49)	6.62 (4.60)	7.83 (6.26)	0.22
Non-home-visit nursing experience, years	Mean (SD)	11.81 (6.34)	12.25 (6.85)	11.37 (5.77)	0.14
Age group	20s	8 (7.6)	4 (7.5)	4 (7.7)	0.42
	30s	28 (26.7)	14 (26.4)	14 (26.9)	
	40s	34 (32.4)	21 (39.6)	13 (25.0)	
	50s	24 (22.9)	8 (15.1)	16 (30.8)	
	≥ 60s	11 (10.5)	6 (11.3)	5 (9.6)	
Sex	Male	15 (14.3)	7 (13.2)	8 (15.4)	0.20
	Female	89 (84.8)	45 (84.9)	44 (84.6)	
	Prefer not to answer	1 (1.0)	1 (1.9)	0 (0.0)	
Educational attainment	Non-university nursing education	59 (56.2)	25 (47.2)	34 (65.4)	0.37
	University or higher	46 (43.8)	28 (52.8)	18 (34.6)	
Frequency of guideline use	Hardly ever	17 (16.2)	7 (13.2)	10 (19.2)	0.22
	Every few months	22 (21.0)	12 (22.6)	10 (19.2)	
	Once a month	26 (24.8)	12 (22.6)	14 (26.9)	
	Several times a month	26 (24.8)	14 (26.4)	12 (23.1)	
	Once a week or more	14 (13.3)	8 (15.1)	6 (11.5)	
Prior use of generative AI	No	31 (29.5)	14 (26.4)	17 (32.7)	0.14
	Yes	74 (70.5)	39 (73.6)	35 (67.3)	
Frequency of generative AI use in the past 3 months	Not used	31 (29.5)	15 (28.3)	16 (30.8)	0.61
	Less than once a month	12 (11.4)	10 (18.9)	2 (3.8)	
	Once a month	11 (10.5)	3 (5.7)	8 (15.4)	
	Several times a month	19 (18.1)	7 (13.2)	12 (23.1)	
	Several times a week	14 (13.3)	7 (13.2)	7 (13.5)	
	Almost daily	18 (17.1)	11 (20.8)	7 (13.5)	

Values are presented as means (SD) or n (%). Standardized mean differences are reported as absolute values. Categorical variables with more than two levels indicate the overall imbalance among categories. Non-university nursing education included nursing high schools, vocational nursing schools, and junior colleges. AI, artificial intelligence; SD, standard deviation; SMD, standardized mean difference

Table 2 Primary and secondary outcome analyses

Outcome	Non-AI group (<i>n</i> =53)	AI group (<i>n</i> =52)	Adjusted mean difference (95% CI)	Cohen's <i>d</i>	<i>p</i> value
Perceived mental effort	6.11 (2.01)	5.10 (1.35)	-1.13 (-1.81 to -0.45)	-0.59	0.001
Perceived task difficulty	5.87 (1.72)	5.21 (1.40)	-0.74 (-1.37 to -0.12)	-0.42	0.021
Confidence in answer correctness	4.57 (1.50)	5.21 (1.39)	0.71 (0.14 to 1.27)	0.45	0.015

Primary and secondary outcomes

The primary and secondary outcome analyses among participants included in the complete-case analysis are presented in Table 2. For the primary outcome—perceived mental effort—the unadjusted mean score was 6.11 (SD 2.01) in the non-AI group and 5.10 (SD 1.35) in the AI group. After adjustment for years of home-visit nursing experience, years of non-home-visit nursing experience, and educational attainment, the adjusted mean difference was -1.13 (95% CI -1.81 to -0.45 ; $p=0.001$).

For the exploratory secondary outcomes, the adjusted mean difference was -0.74 (95% CI -1.37 to -0.12 ; $p=0.021$) for perceived task difficulty and 0.71 (95% CI 0.14 to 1.27 ; $p=0.015$) for confidence in answer correctness.

Values are presented as unadjusted means (standard deviations) unless otherwise specified. Adjusted mean differences were estimated using analysis of covariance, with group assignment as the primary independent variable and years of home-visit nursing experience, years of non-home-visit nursing experience, and educational attainment as covariates. All outcomes were scored from 1 to 9. Adjusted mean differences are expressed in scale points and were calculated as the AI group minus the non-AI group. Negative values indicate lower scores in the AI group, whereas positive values indicate higher scores. Cohen's *d* follows the same direction. AI=artificial intelligence; CI=confidence interval.

Sensitivity analyses

Sensitivity analyses are presented in Additional file 1. The unadjusted mean differences were -1.02 (95% CI -1.68 to -0.36 ; $p=0.003$) for perceived mental effort, -0.66 (95% CI -1.26 to -0.05 ; $p=0.035$) for perceived task difficulty, and 0.65 (95% CI 0.09 to 1.21 ; $p=0.024$) for confidence in answer correctness.

In the post hoc exploratory sensitivity analysis additionally adjusted for the frequency of generative AI use in the past three months, the results were similar to those of the primary analysis: -1.10 (95% CI -1.76 to -0.44 ; $p=0.001$) for perceived mental effort, -0.71 (95% CI -1.33 to -0.10 ; $p=0.023$) for perceived task difficulty, and 0.69 (95% CI 0.13 to 1.25 ; $p=0.017$) for confidence in answer correctness.

In an additional post hoc exploratory sensitivity analysis additionally adjusted for age group, the results were also similar to those of the primary analysis: -1.08 (95% CI -1.78 to -0.39 ; $p=0.003$) for perceived mental effort, -0.72 (95% CI -1.36 to -0.08 ; $p=0.028$) for perceived task difficulty, and 0.73 (95% CI 0.15 to 1.31 ; $p=0.014$) for confidence in answer correctness.

Discussion

Summary of findings

Among participants included in the complete-case analysis, the AI group reported lower perceived mental effort and task difficulty and higher confidence in answer correctness than the non-AI group, with sensitivity analyses showing the same direction as the findings. The largest between-group difference was observed for perceived mental effort.

Interpretation in relation to cognitive and decision support

The lower perceived mental effort reported in the AI group can be interpreted in relation to the cognitive load theory. Cognitive load is influenced by how information is presented, organized, and integrated during complex tasks. van Merriënboer and Sweller argued that extraneous load can be reduced by integrating different sources of information, reducing redundancy, and designing tasks in ways that support the processing of relevant information [36]. The present intervention was designed as a guideline-informed reference environment rather than as unrestricted access to a general-purpose chatbot. In the complete-case analysis, structured access to guideline-based information may help explain the lower perceived effort reported in the AI group during the case-based task. This interpretation is consistent with the lower perceived mental effort reported in the AI group, although this trial did not directly measure cognitive load components, such as intrinsic, extraneous, and germane loads.

Clinical decision-support studies have similarly shown that the format and task alignment of information support can influence usability, workload, guideline adherence, and care planning efficiency [16, 37, 38]. These studies suggest that decision support depends not only on the availability of evidence-based information but also on how it is presented and aligned with users' tasks. The present study extends the literature to include a guideline-informed AI interface for home-visit nursing.

The higher confidence reported in the AI group requires careful interpretation. Confidence may indicate that the participants felt better able to organize their reasoning or justify their responses when the guideline-informed AI system was available. However, confidence does not establish whether a final judgment is accurate, complete, or safe. Studies on AI-assisted decision-making have shown that AI support can affect user judgment in ways that are not always beneficial. For example, explainable AI-assistance in dental image interpretation does not eliminate the risk that users would be influenced by incorrect AI outputs [39]. A randomized trial of clinician-supervised generative AI for decision support also suggested that supervision may be necessary because unsupervised generative AI may produce risky suggestions even when it appears to provide decision support [40]. Recent reviews on trust in AI-based clinical decision support systems have also emphasized that healthcare workers' trust is shaped by transparency, training, usability, clinical reliability, human-centered design, and user control [41]. Therefore, the higher confidence observed in the present study should be understood as a subjective measure, and not as evidence of improved clinical reasoning quality or calibrated trust in the AI-supported environment.

Implications for guideline-informed AI support in home-visit nursing

The configuration of the AI system is central to interpreting the intervention. Previous studies have raised concerns regarding general-purpose generative AI use in clinical contexts. ChatGPT has been reported to have incomplete alignment with clinical practice guidelines for musculoskeletal decision making [42]. Vignette-based evaluations have also suggested limitations in the appropriateness and individualization of medical advice generated by ChatGPT across care contexts [43]. These findings are relevant because the present study did not test for the unrestricted use of generative AI. Instead, the AI system was preloaded using predefined clinical guidelines and configured to generate responses based on those materials. Therefore, the intervention should be interpreted as guideline-informed AI chat access, or AI-mediated access to guideline information, rather than as an evaluation of the independent effect of AI itself. Because the non-AI group was instructed not to use external resources, the present design cannot distinguish the effect of the AI chat system from the effect of access to guideline information. A future trial comparing AI-mediated access with access to the same guideline materials in a conventional format would be needed to evaluate the unique contribution of the AI chat system. This interpretation is also consistent with recent work on task-technology fit in AI-based clinical decision support systems, which emphasizes that the perceived usefulness of AI-based support depends on alignment among the clinical task, system functions, and work context [44]. However, task-specific configurations do not eliminate the need to evaluate factual accuracy, clinical appropriateness, and safety.

These findings have implications for designing clinical reasoning support for home-visit nursing. Home-visit nurses often make judgments in the patients' homes, where immediate access to colleagues, specialists, or structured evidence-based resources may be limited. Clinical decision support research suggests that appropriately designed information support can reduce workload and support decision-making in clinical tasks, while recent nursing and AI-based clinical decision support literature indicates that implementation should consider user characteristics, task demands, and fit with the clinical workflow [16, 44]. These findings suggest that guideline-informed AI chat access may be relevant to the design of support tools for simulated case-based home-visit nursing tasks. An important feature of the present intervention is that it was constructed using an existing AI platform and publicly available guideline materials rather than a bespoke or locally trained AI model. This may make such forms of guideline-informed AI support relatively feasible for educational or professional development settings provided that the source materials, prompting instructions, and intended uses are clearly specified. However, the present findings should not be interpreted as direct evidence of effectiveness in real-world home-visit nursing practice.

Nonetheless, the implications should be framed as support for reasoning rather than as a replacement for professional judgment. Among participants included in the complete-case analysis, the AI group reported lower perceived burden and higher confidence, but these complete-case findings do not demonstrate improvement in the clinical quality of participants' responses. Previous studies indicate that users may be exposed to incomplete, inappropriate, or misleading AI outputs [13, 39, 42, 43]. Therefore, implementation in home-visit nursing requires clear boundaries for intended use, transparency about source materials, and training in the critical appraisal of AI-generated responses. These requirements are particularly important in home-visit nursing, where clinical decisions are shaped not only by disease-related information, but also by patients' living conditions, family contexts, care goals, and available community resources.

Limitations and future directions

This study has some limitations that should be considered when interpreting the findings. First, the trial was retrospectively registered. Although the primary outcome and planned primary analysis were specified in the ethics-approved research protocol before the study was conducted, retrospective registration may reduce confidence in the prospective prespecification of outcomes, covariates, and analysis plans. In addition, the sample size was smaller than originally planned. Although statistically significant differences emerged across all three outcomes, this study should be interpreted as exploratory rather than definitive evidence of effectiveness. Minimally important differences have not been established for these 9-point ratings. Therefore, the statistically significant between-group differences should not be interpreted as evidence of clinically or practically meaningful benefit. The outcomes were also self-reported. The combination of unblinded allocation and exclusively subjective outcomes is an important limitation of this study. Because participants were aware of their group allocation and rated the outcomes immediately after the task, expectancy effects, novelty effects, perceived support, response bias, and demand characteristics may have influenced their ratings. The study assessed perceived mental effort, perceived task difficulty, and confidence in answer correctness, but did not directly assess the accuracy, completeness, safety, or patient-centeredness of the participants' clinical reasoning. Potential unintended consequences of AI access, including overconfidence, automation bias, over-reliance on AI-generated information, and inappropriate trust in AI outputs, were not systematically assessed. No study data were collected specifically to evaluate potentially unsafe or inappropriate AI use. In addition, perceived task difficulty and confidence in answer correctness were assessed using study-specific single-item measures rather than previously validated scales. As participants who did not complete the task had no post-task outcome data, they were excluded from the complete-case analysis. Specific reasons for task noncompletion were not collected. Participant characteristics were also unavailable for non-completers. Therefore, we could not compare baseline characteristics between completers and non-completers. Accordingly, informative attrition cannot be excluded, and neither the mechanism nor the direction or magnitude of any resulting bias can be determined from the available data. The complete-case analysis can be

interpreted as preserving the randomized comparison only if completion after randomization did not introduce outcome-related differential selection between groups, an assumption that cannot be verified with the available data. The findings should therefore be interpreted as between-group differences among participants who completed the web-based task, rather than as a full intention-to-treat estimate of the randomized treatment effect.

The web-based, simulated nature of this study also limits the generalizability of the findings. The case scenario was designed to reflect home-visit nursing practice. However, real home visits involve environmental uncertainty, interactions with patients and family members, time pressure, and interprofessional communication. Therefore, the implications for real-world home-visit nursing practice should be interpreted cautiously. The participants were instructed not to use external information sources other than the assigned system; however, adherence to this instruction could not be fully verified online. This raises the possibility that participants in the non-AI group may have accessed external information, while participants in the AI group may have used resources beyond the assigned AI system. Such behavior may have introduced performance bias or contamination. Participants were recruited through open calls and may have been more motivated or interested in AI, clinical reasoning, or professional development than the broader population of home-visit nurses.

Some imbalances in the participant characteristics warrant attention. Although randomization was performed, the prior frequency of generative AI use was imbalanced between the groups. The post hoc sensitivity analysis additionally adjusted for this variable showed results similar to the primary analysis. However, this analysis was not prespecified, and the potential influence of prior familiarity with generative AI on perceived effort, task difficulty, and confidence cannot be fully excluded. Participants with more frequent prior use of generative AI may have been more familiar with interacting with AI systems and interpreting AI-generated information, which could have influenced their perceived effort, task difficulty, or confidence. In addition, the AI system was implemented using an existing platform, and its responses may have been affected by platform updates or system-level modifications.

Future studies should evaluate whether guideline-informed AI support improves not only the subjective burden and confidence but also the quality, safety, and practical utility of clinical reasoning. This requires an evaluation of the content of the nurses' responses, including accuracy, completeness, prioritization, guideline adherence, appropriateness, safety, and patient-centeredness. Future research should also address the implementation, trust, task fit, and training needs [41, 44, 45]. Therefore, further research should examine how nurses interact with guideline-informed AI systems, how prior familiarity with generative AI affects use, and what forms of training are needed to prevent overconfidence and uncritical acceptance of AI-generated responses. Studies conducted in realistic educational or practice settings could clarify how AI-mediated access to guideline information can be safely implemented in home-visit nursing.

Conclusions

Among home-visit nurses who completed the web-based simulated case-based task, access to a guideline-informed AI chat system was associated with lower perceived mental effort and task difficulty and higher confidence in answer correctness, compared with no access to external resources. Because the analysis was restricted to participants who completed the task, these between-group differences represent complete-case estimates and should not be interpreted as an intention-to-treat estimate of the randomized treatment effect. However, this study does not establish the independent effect of AI itself or demonstrate that AI-mediated support improved the accuracy, safety, or quality of clinical judgment. Because all outcomes were subjective and participants were aware of their group allocation, the findings should be interpreted as exploratory and require confirmation using objective or independently assessed outcomes. Future research should compare AI-mediated access with conventional access to the same guideline materials and examine the quality of AI-supported reasoning, risks of overconfidence, and implementation of guideline-informed AI systems in real-world home-visit nursing practices.

Abbreviations

AI	Artificial intelligence
ANCOVA	Analysis of covariance
CI	Confidence interval
LLM	Large language model
OSF	Open Science Framework
SD	Standard deviation
SMD	Standardized mean difference

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1186/s12912-026-05418-w>.

Acknowledgements The authors thank the home-visit nurses who participated in this study. The authors also thank the individuals who supported recruitment and preliminary system evaluation. The authors thank Editage for English language editing.

Author contributions TI conceived and designed the study, developed the intervention materials, collected data, conducted statistical analyses, interpreted the findings, and drafted the manuscript. HI, TK, and YT contributed to the intervention development, recruitment, interpretation of findings, and critical review of the manuscript. All authors have reviewed and approved the final manuscript.

Funding This study was supported by the 36th Research Grant (FY2025) from the France Bed Home Care Foundation (FBK250527013). The funder had no role in the study design, data collection, analysis, interpretation of data, writing of the manuscript, or the decision to submit the manuscript for publication.

Data availability The study materials, AI system configuration, source materials, prompting instructions, survey instrument, simulated case-based tasks, data dictionaries, and analysis code are available in the Open Science Framework repository: <https://doi.org/10.17605/OSF.IO/3VKS5>. De-identified participant-level data are not publicly available because the dataset contains free-text responses and participant characteristics that could pose a risk of re-identification. Aggregated results are reported in this article and its supplementary materials. Requests for additional aggregated results may be directed to the corresponding author and will be considered where ethically and methodologically appropriate. Prespecified exploratory text-mining analyses of participants' free-text responses are ongoing, are outside the scope of the present report, and are planned to be reported separately. The original Japanese study protocol and ethics approval documentation are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate This study was conducted in accordance with the Declaration of Helsinki and relevant guidelines and regulations and was approved by the Tekix Inc. Research Ethics Review Committee (approval number: 202601). The committee consisted of five members, including two members external to Tekix Inc. The corresponding author was a member of the committee but did not participate in the review or decision for this study. When a committee member submits a research application, the review is conducted by committee members other than the applicant. Before enrollment, participants received written and video-based information about the study and provided informed consent by selecting the consent field on a web-based form. Participation was voluntary, and participants could discontinue participation at any time. The dataset used for analysis excluded names and email addresses. Email addresses used for task distribution and honorarium payments were stored separately from study responses. Participants received a completion-based honorarium of 1,500 Japanese yen. Participants were provided with contact information for study-related inquiries.

Consent for publication Not applicable.

Use of AI tools During manuscript preparation, the authors used ChatGPT (OpenAI) for English translation and language refinement. All outputs were reviewed and edited by the authors, who take full responsibility for the content of the manuscript.

Competing interests The authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Ishikawa T, Imura H, Takashima Y. Intelligent flexibility in an aging society: lessons from Japan's home-visit nursing system. *Front Public Health*. 2026;14:1755718. <https://doi.org/10.3389/fpubh.2026.1755718>.
2. Irani E, Hirschman KB, Cacchione PZ, Bowles KH. Home health nurse decision-making regarding visit intensity planning for newly admitted patients: a qualitative descriptive study. *Home Health Care Serv Q*. 2018;37(3):211–31. <https://doi.org/10.1080/01621424.2018.1456997>.
3. Murray-Parahi P, DiGiacomo M, Jackson D, Davidson PM. New graduate registered nurse transition into primary health care roles: an integrative literature review. *J Clin Nurs*. 2016;25(21–22):3084–101. <https://doi.org/10.1111/jocn.13297>.
4. Johnsen HM, Slettebø Å, Fossum M. Registered nurses' clinical reasoning in home healthcare clinical practice: a think-aloud study with protocol analysis. *Nurse Educ Today*. 2016;40:95–100. <https://doi.org/10.1016/j.nedt.2016.02.023>.
5. Yoshimatsu K, Nakatani H. Home visiting nurses' job stress and error incidents. *Home Health Care Manag Pract*. 2020;32(2):110–7. <https://doi.org/10.1177/1084822319899392>.
6. Jun J, Kovner CT, Stimpfel AW. Barriers and facilitators of nurses' use of clinical practice guidelines: an integrative review. *Int J Nurs Stud*. 2016;60:54–68. <https://doi.org/10.1016/j.ijnurstu.2016.03.006>.
7. Mathieson A, Grande G, Luker K. Strategies, facilitators and barriers to implementation of evidence-based practice in community nursing: a systematic mixed-studies review and qualitative synthesis. *Prim Health Care Res Dev*. 2019;20:e6. <https://doi.org/10.1017/S1463423618000488>.
8. Pavloff M, Labrecque ME. Continuing education for home care nurses: an integrative literature review. *Home Healthc Now*. 2021;39(6):310–9. <https://doi.org/10.1097/NHH.0000000000001005>.
9. Jarrin OF, Pouladi FA, Madigan EA. International priorities for home care education, research, practice, and management: qualitative content analysis. *Nurse Educ Today*. 2019;73:83–7. <https://doi.org/10.1016/j.nedt.2018.11.020>.
10. Vrdoljak J, Boban Z, Vilović M, Kumrić M, Božić J. A review of large language models in medical education, clinical decision support, and healthcare administration. *Healthc (Basel)*. 2025;13(6):603. <https://doi.org/10.3390/healthcare13060603>.
11. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Intern Med*. 2023;183(6):589–96. <https://doi.org/10.1001/jamainternmed.2023.1838>.
12. Shokr EA. Integrating a knowledge-based artificial intelligence chatbot into nursing training programs: a comparative quasi-experimental study in Egypt and Saudi Arabia. *BMC Nurs*. 2025;24(1):1245. <https://doi.org/10.1186/s12912-025-03883-3>.
13. Saban M, Dubovi I. A comparative vignette study: evaluating the potential role of a generative AI model in enhancing clinical decision-making in nursing. *J Adv Nurs*. 2025;81(11):7489–99. <https://doi.org/10.1111/jan.16101>.
14. Levin C, Zaboli A, Turcato G, Saban M. Nursing judgment in the age of generative artificial intelligence: a cross-national study on clinical decision-making performance among emergency nurses. *Int J Nurs Stud*. 2025;172:105216. <https://doi.org/10.1016/j.ijnurstu.2025.105216>.
15. Chang LC, Wang YN, Lin HL, Liao LL. Registered nurses' attitudes towards ChatGPT and self-directed learning: a cross-sectional study. *J Adv Nurs*. 2025;81(7):3811–20. <https://doi.org/10.1111/jan.16519>.
16. Yao Y, Dunn Lopez K, Bjarnadottir RI, Macieira TGR, Dos Santos FC, Madandola OO, et al. Examining care planning efficiency and clinical decision support adoption in a system tailoring to nurses' graph literacy: national, web-based randomized controlled trial. *J Med Internet Res*. 2023;25:e45043. <https://doi.org/10.2196/45043>.
17. Mebrahtu TF, Skyrme S, Randell R, Keenan AM, Bloor K, Yang H, et al. Effects of computerised clinical decision support systems on nursing and allied health professional performance and patient outcomes: a systematic review of experimental and observational studies. *BMJ Open*. 2021;11(12):e053886. <https://doi.org/10.1136/bmjopen-2021-053886>.
18. Abi Khalil C, Saab A, Rahme J, Bouaud J, Seroussi B. Capabilities of computerized decision support systems supporting the nursing process in hospital settings: a scoping review. *BMC Nurs*. 2025;24(1):677. <https://doi.org/10.1186/s12912-025-03272-w>.

19. Berkhout M, Smit K, Sent D, Kusters R, Versendaal J, van Houwelingen T. Understanding the role of clinical decision support systems among hospital nurses using the FITT framework: qualitative study. *J Med Internet Res*. 2025;27:e76025. <https://doi.org/10.2196/76025>.
20. Zhai Y, Yu Z, Zhang Q, Zhang Y. Barriers and facilitators to implementing a nursing clinical decision support system in a tertiary hospital setting: a qualitative study using the FITT framework. *Int J Med Inf*. 2022;166:104841. <https://doi.org/10.1016/j.ijmedinf.2022.104841>.
21. UMIN Clinical Trials Registry. Web-based comparative study of the effects of guideline-referenced AI use among home-visit nurses. UMIN000061415. Registered 30 April 2026. Accessed 15 May 2026. https://center6.umin.ac.jp/cgi-open-bin/ctr_e/ctr_view.cgi?recptno=R000070275.
22. Hopewell S, Chan AW, Collins GS, Hróbjartsson A, Moher D, Schulz KF, et al. CONSORT 2025 statement: updated guideline for reporting randomised trials. *BMJ*. 2025;388:e081123. <https://doi.org/10.1136/bmj-2024-081123>.
23. Japanese Circulation Society; Japanese Heart Failure Society. Guidelines for diagnosis and treatment of heart failure: 2025 revision. Published 28 March 2025. Accessed 8 May 2026. https://www.j-circ.or.jp/cms/wp-content/uploads/2025/03/JCS2025_Kato.pdf.
24. Japanese Circulation Society; Japanese Heart Failure Society. Recommendations for palliative care in cardiovascular diseases: 2021 revision. Published March 2021. Accessed 8 May 2026. https://www.j-circ.or.jp/cms/wp-content/uploads/2021/03/JCS2021_Anzai.pdf.
25. Japanese Heart Failure Society. Statement on nutritional assessment and management in patients with heart failure. Published 2018. Accessed 8 May 2026. <https://www.jhfs.or.jp/statement-guideline/files/statement20181012.pdf>.
26. Japanese Society for Palliative Medicine. Practical guide for palliative care team activities, supplement: palliative care for patients with heart failure. Published July 2020. Accessed 8 May 2026. https://www.jspm.ne.jp/files/active/heart_failure_v1.pdf.
27. Japanese Society of Hospital Pharmacists. Guidance on pharmaceutical care in heart failure palliative care. Published 16 September 2021. Accessed 8 May 2026. <https://www.jsph.or.jp/activity/guideline/20210916-1-1.pdf>.
28. Japanese Circulation Society; Japanese Heart Failure Society. Guidelines for diagnosis and treatment of cardiomyopathies: 2018 revision. Published 2018. Accessed 8 May 2026. https://www.j-circ.or.jp/cms/wp-content/uploads/2018/08/JCS2018_tsutsui_kitaoka.pdf.
29. Japanese Circulation Society; Japanese Heart Failure Society. 2021 JCS/JHFS guideline focused update on diagnosis and treatment of acute and chronic heart failure. Published 2021. Accessed 8 May 2026. https://www.j-circ.or.jp/cms/wp-content/uploads/2021/03/JCS2021_Tsutsui.pdf.
30. Japanese Circulation Society; Japanese Society for Cardiovascular Surgery; Japanese College of Cardiology; Japanese Association of Cardiovascular Intervention and Therapeutics. 2023 JCS/JSCVS/JCC/CVIT guideline focused update on indication and operation of PCPS/ECMO/IMPELLA. Published 2023. Accessed 8 May 2026. https://www.j-circ.or.jp/cms/wp-content/uploads/2023/03/JCS2023_nishimura.pdf.
31. Japanese Heart Failure Society; Japan Association for Home Care Medicine. Guidance on continuous intravenous inotropic therapy for severe heart failure in home care. Published 26 March 2024. Accessed 8 May 2026. <https://www.jahcm.org/assets/images/pdf/statement20240326.pdf>.
32. Japanese Heart Failure Society. Statement on the management of elderly patients with heart failure. Published 7 October 2016. Accessed 8 May 2026. <https://www.jhfs.or.jp/statement-guideline/files/statement20161007.pdf>.
33. Google. May. Privacy and terms of use in NotebookLM. Google Help. Accessed 8 May 2026. <https://support.google.com/notebooklm/answer/17004255>.
34. Ishikawa T, Imura H, Kitazawa T, Takashima Y. Guideline-informed AI chat system for home-visit nursing: system description and evaluation materials. *Open Sci Framew*. 2026. <https://doi.org/10.17605/OSF.IO/3VKSS>.
35. Paas FGWC. Training strategies for attaining transfer of problem-solving skill in statistics: a cognitive-load approach. *J Educ Psychol*. 1992;84(4):429–34. <https://doi.org/10.1037/0022-0663.84.4.429>.
36. van Merriënboer JJ, Sweller J. Cognitive load theory in health professional education: design principles and strategies. *Med Educ*. 2010;44(1):85–93. <https://doi.org/10.1111/j.1365-2923.2009.03498.x>.
37. Saleem JJ, Patterson ES, Militello L, Anders S, Falciglia M, Wissman JA, et al. Impact of clinical reminder redesign on learnability, efficiency, usability, and workload for ambulatory clinic nurses. *J Am Med Inf Assoc*. 2007;14(5):632–40. <https://doi.org/10.1197/jamia.M2163>.
38. Richardson KM, Fouquet SD, Kerns E, McCulloh RJ. Impact of mobile device-based clinical decision support tool on guideline adherence and mental workload. *Acad Pediatr*. 2019;19(7):828–34. <https://doi.org/10.1016/j.acap.2019.03.001>.
39. Glick A, Clayton M, Angelov N, Chang J. Impact of explainable artificial intelligence assistance on clinical decision-making of novice dental clinicians. *JAMIA Open*. 2022;5(2):ooac031. <https://doi.org/10.1093/jamiaopen/ooac031>.
40. Ebnali Harari R, Altaweel A, Ahram T, Keehner M, Shokoohi H. A randomized controlled trial on evaluating clinician-supervised generative AI for decision support. *Int J Med Inf*. 2025;195:105701. <https://doi.org/10.1016/j.ijmedinf.2024.105701>.
41. Tun HM, Rahman HA, Naing L, Malik OA. Trust in artificial intelligence-based clinical decision support systems among health care workers: systematic review. *J Med Internet Res*. 2025;27:e69678. <https://doi.org/10.2196/69678>.

42. Gianola S, Barger S, Castellini G, Cook C, Palese A, Pillastrini P, et al. Performance of ChatGPT compared to clinical practice guidelines in making informed decisions for lumbosacral radicular pain: a cross-sectional study. *J Orthop Sports Phys Ther.* 2024;54(3):222–8. <https://doi.org/10.2519/jospt.2024.12151>.
43. Nastasi AJ, Courtright KR, Halpern SD, Weissman GE. A vignette-based evaluation of ChatGPT's ability to provide appropriate and equitable medical advice across care contexts. *Sci Rep.* 2023;13(1):17885. <https://doi.org/10.1038/s41598-023-45223-y>.
44. Parsons CS, Zuiderwijk A, Orchard NA, Dijkstra P, Melles M, Schmettow M, et al. Task-technology fit of artificial intelligence-based clinical decision support systems: a review of qualitative studies. *BMC Med Inf Decis Mak.* 2025;25:397. <https://doi.org/10.1186/s12911-025-03237-8>.
45. Ito Y, Kajiura K, Kako J, Kobayashi M, Tsubaki M, Sakuramoto H, et al. Artificial intelligence in nursing support for patients: a rapid review. *Int J Med Inf.* 2025;204:106087. <https://doi.org/10.1016/j.ijmedinf.2025.106087>.

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

Authors and Affiliations

Takemasa Ishikawa^{1,2}  · Haruka Imura³  · Tomomi Kitazawa⁴  · Yoshiyuki Takashima⁵ 

✉ Takemasa Ishikawa
ishikawa@nana-r.jp

Haruka Imura
harukayoshimaru@yahoo.co.jp

Tomomi Kitazawa
kitazawa.tomomi@shizuoka.ac.jp

Yoshiyuki Takashima
y-takashima@naramed-u.ac.jp

¹ Nana-r Home-visit Nursing Development Center, Tekix Corporation, #301, 3-6-17 Kamishinden, Toyonaka-shi, Osaka 560-0085, Japan

² Community-based Integrated Care Science, School of Nursing, Osaka Metropolitan University, Osaka, Japan

³ Department of Perioperative and Critical Care Management, Graduate School of Biomedical and Health Sciences, Hiroshima University, Hiroshima, Japan

⁴ Faculty of Education, Shizuoka University, Shizuoka, Japan

⁵ Faculty of Nursing, School of Medicine, Nara Medical University, Nara, Japan