Review

Application of Large Language Models in Chronic Disease Care: Mixed Methods Systematic Review and Thematic Synthesis

Linghui Zhang[1*], MA; Panpan Huai[1*], PhD; Rui Xu[1], MA; Jingjing Sun[1], MA; Ruihua Jin[1], MA; Huimei Lv[2], BA

[1]Shanxi Medical University, Taiyuan, Shanxi, China

[2]The Fifth Clinical College of Shanxi Medical University, Taiyuan, Shanxi, China

*these authors contributed equally

Corresponding Author:

Huimei Lv, BA
The Fifth Clinical College of Shanxi Medical University
No. 29, Shuangta Temple Street, Yingze District
Taiyuan, Shanxi 030012
China
Phone: 86 13073585857
Email: 2823704581@qq.com

Abstract

Background: Chronic diseases account for nearly three-quarters of global deaths and demand continuous, personalized long-term management; yet traditional care models often fall short in delivering such sustained support. Large language models, with advanced conversational and analytical capabilities, present promising opportunities to address the problem by offering scalable, interactive support. However, a comprehensive synthesis of evidence across diverse study designs, which moves beyond isolated technical metrics to evaluate large language models through a structured, theory-driven lens, remains limited.

Objective: This study aimed to synthesize quantitative, qualitative, and mixed methods evidence on technical performance, application scenarios, and documented challenges of large language models in chronic disease care, and to critically evaluate these findings through a theory-driven, 3D framework informed by Orem's Self-Care Theory.

Methods: The mixed methods systematic review adhered to the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 and SWiM (Synthesis Without Meta-Analysis) guidelines. A comprehensive computer-based search was conducted across PubMed, Web of Science, Embase, Cochrane Library, CINAHL, Wiley Online Library, SpringerLink, ScienceDirect, China National Knowledge Infrastructure, Wanfang Data, VIP Database, and the China Biology Medicine Disc from inception to April 2026. Two independent researchers performed study screening, data extraction, and quality appraisal using the Mixed Methods Appraisal Tool 2018. Given significant clinical and methodological heterogeneity across the included studies, a quantitative meta-analysis was not appropriate; instead, thematic synthesis was used following Thomas and Harden's 3-step approach, with NVivo 14 (Lumivero) used for line-by-line coding and theme development, all informed by Orem's 3D framework.

Results: A total of 20 studies were included, all rated as moderate or high quality. The thematic synthesis revealed three core themes aligned with the proposed framework: (1) foundational safety, privacy, and fairness (hallucination risks and data concerns); (2) self-care enablement through perceived usefulness (patient education, decision support, and self-management); and (3) design and system integration challenges (readability mismatches and workflow gaps). Patient education and clinical decision support were the most common application scenarios. Key technical enhancements (retrieval-augmented generation [RAG] and fine-tuning) primarily strengthened the second theme, while barriers such as content readability, hallucination risks, and ethical ambiguities limited real-world readiness.

Conclusions: This review is the first to integrate Orem's Self-Care Theory into a 3D evaluative framework for large language models in chronic care, thereby moving beyond fragmented, technology-centric assessments toward a structured, nursing-informed, and theory-driven synthesis. Unlike prior reviews that primarily focused on isolated technical metrics or broad feasibility, this synthesis provides a layered, discipline-grounded evaluation distinguishing foundational safety, self-care enablement, and system integration. These findings show evidence mainly supports the intermediate layer, while safeguards

and operational integration remain deficient, guiding nursing research toward safety assurances, health-literacy-adaptive design, and implementation science for equitable, patient-centered care.

Trial Registration: PROSPERO CRD420251208327; https://www.crd.york.ac.uk/PROSPERO/view/CRD420251208327

J Med Internet Res 2026;28:e90744; doi: 10.2196/90744

Keywords: large language models; chronic disease care; mixed methods systematic review; thematic synthesis; Orem's Self-Care Theory

Introduction

Background

Chronic noncommunicable diseases refer to conditions that are not primarily caused by acute infections, characterized by long duration and typically slow progression, and requiring long-term treatment and care [1]. Major categories include cardiovascular diseases (eg, hypertension, coronary heart disease, and stroke), cancers, chronic respiratory diseases (eg, chronic obstructive pulmonary disease and asthma), diabetes mellitus, chronic kidney disease, digestive system diseases (eg, chronic hepatitis, cirrhosis, and inflammatory bowel disease), musculoskeletal disorders (eg, arthritis and fibromyalgia), and neurodegenerative conditions (eg, Parkinson disease and Alzheimer disease) [2]. According to World Health Organization statistics, chronic diseases account for nearly three-quarters of all deaths globally, with approximately 17 million people dying from a chronic disease before the age of 70 each year, 86% of which occur in low- and middle-income countries [1,3]. The Lancet has noted that the current burden of healthy life-years lost and premature mortality due to chronic diseases remains substantial [4]. With accelerating population aging and epidemiological transition, chronic disease care faces multiple challenges, including a massive patient population, diverse health needs, and prolonged management cycles [5]. Traditional chronic disease care models, largely dependent on regular outpatient follow-ups and standardized health education, struggle to deliver continuous, personalized, and dynamic interventions [6]. This limitation is particularly pronounced in regions with unevenly distributed medical resources and a shortage of specialized professionals [7]. Among major chronic diseases, conditions such as diabetes, hypertension, and nonalcoholic fatty liver disease affect tens of millions to hundreds of millions of people globally [8,9]. Conditions like hypertension and diabetes demand high continuity and personalization in care due to the necessity for ongoing monitoring, behavioral interventions, and frequent health information needs [10].

The World Health Organization's 14th General Program of Work (2025-2028) emphasizes leveraging emerging technologies like AI and strengthening digital interventions to support health literacy [11]. Integrating this model with precision care has become a significant trend [12]. Currently, information management platforms such as mobile health apps and smart wearable devices have shown initial success in chronic disease data collection and remote monitoring but still possess limitations [13]. Research by Cunningham et al [14] found that while common smart follow-up tools can perform basic data collection, their information integration capabilities are limited, making it difficult to extract personalized insights from diverse health data. Large language models (LLMs), as a groundbreaking AI technology, leverage powerful natural language understanding and generation capabilities and have rapidly developed in the health care domain [15]. LLM platforms such as OpenAI's ChatGPT, Google's Gemini, Anthropic's Claude, and DeepSeek have demonstrated intelligent support capabilities in nursing scenarios [16]. The application of LLMs has permeated various health care fields, showing significant value in education and clinical practice [17-19]. Research by Harrington et al [20] points out that LLMs can act as intelligent tutors in nursing education. Studies have also found that LLMs can serve as 24/7 information assistants, responding to general inquiries about disease symptoms, medications, and lifestyle [21]. Patients can ask questions conversationally about diet control, exercise safety, or medication side effects and receive detailed explanations [22]. Compared to the preset, static responses of traditional clinical decision support systems, this flexible, interactive method demonstrates stronger adaptability [23]. The characteristics of chronic disease care—long duration, need for continuous health education, and daily support—align well with the advantages LLMs offer [24].

However, alongside these promising capabilities, significant concerns have emerged. Studies have consistently reported that LLMs may exhibit "hallucinations"—generating inaccurate or even dangerous medical information [25,26]. In the context of chronic disease management, where patients often have complex comorbidities and rely on precise medication and dietary advice, such errors could lead to adverse health outcomes [27]. Furthermore, LLMs may misinterpret diverse, colloquial patient descriptions, particularly from individuals with lower health literacy or nonstandard language use [28]. Additional barriers include data privacy and security risks, unclear ethical and legal responsibility for AI-generated recommendations, and potential algorithmic biases that could disadvantage certain patient populations [29]. Moreover, the readability of LLM-generated content often exceeds the comprehension level of average patients, limiting its practical utility for self-care [30,31]. Given this duality, substantial potential alongside considerable risks—a systematic and critical synthesis of the evidence is urgently needed [32].

Li et al [33] conducted a mixed method review of LLMs in chronic disease management, with a primary focus on quantitative efficacy indicators. Watson et al [34] performed an integrative review on generative AI in general nursing practice but did not specifically address chronic disease care.

Existing reviews on LLMs in chronic care have primarily focused on technical performance metrics (eg, accuracy and precision) or have broadly cataloged application scenarios without a structured, theory-driven evaluation framework [32,35]. They often treat safety, usability, and clinical integration as separate issues, failing to provide a coherent model for assessing whether and how LLMs can genuinely support chronic care from a nursing science perspective [36]. To address this gap, this review introduces Orem's Self-Care Theory as an analytical lens [37,38]. This theory, which centers on patients' self-care agency and the nursing systems required to support it, offers a structured way to evaluate LLMs across three critical dimensions: foundational safety and ethics, enablement of self-care, and operational integration into care workflows [36,39]. By applying this theory-driven framework, this review aims to move beyond isolated technical assessments toward a holistic evaluation of LLMs in chronic disease care.

Objectives

This mixed methods systematic review aims to synthesize quantitative, qualitative, and mixed methods evidence on the application of LLMs in chronic disease care, and to critically evaluate this evidence through a theory-driven 3D framework informed by Orem's Self-Care Theory (safety and ethics, self-care enablement, and operational and system integration).

Methods

Study Design and Registration

This study is a mixed methods systematic review. The review protocol was registered with the International Prospective Register of Systematic Reviews (registration number: CRD420251208327). The reporting of this systematic review followed the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 statement [40]. The PRISMA 2020 expanded checklist is provided in Checklist 1.

Information Sources

This review conducted a comprehensive search across 12 electronic databases, including PubMed, Web of Science, Embase, Cochrane Library, CINAHL, Wiley Online Library, SpringerLink, ScienceDirect, China National Knowledge Infrastructure, Wanfang Data, VIP Database, and the China Biology Medicine Disc, from inception to April 2026. Both Chinese and English databases were included to ensure comprehensive coverage of the literature. In addition to the database search, we performed backward and forward citation searching of the reference lists of included studies and relevant reviews to identify additional eligible records.

Search Strategy

The search strategy was developed collaboratively by the research team (LZ, HL, PH, and JS), all of whom had received standardized training in systematic review search methodology prior to conducting the search. The team first identified key concepts based on the Sample, Phenomenon of Interest, Design, Evaluation, Research type framework and translated these concepts into search terms using Boolean operators (AND, OR, NOT). Medical Subject Headings terms and free-text words were combined to maximize sensitivity and specificity. The preliminary search strategy was then reviewed and validated by an experienced medical librarian to ensure methodological rigor and adherence to best practices in systematic review searching. Following the librarian's feedback, the final search strategies were refined and finalized by the research team.

The search used Boolean logic with AND to combine the 3 thematic domains (chronic disease, LLMs, and application context), OR to link synonymous terms within each domain, and NOT to exclude irrelevant content when necessary. This structured approach ensured comprehensive retrieval of potentially relevant studies while maintaining acceptable precision. The complete search strategies for all 12 databases, including database-specific syntax adjustments and any applied limits (eg, language restrictions to Chinese and English), are provided in Multimedia Appendix 1 for full transparency and reproducibility.

The reporting of the search strategy followed the PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses literature search extension) guideline [41]. A detailed explanation of how each PRISMA-S item was addressed is provided in Checklist 2.

Eligibility Criteria

This study used the SPIDER (sample, phenomenon of interest, design, evaluation, research) framework to formulate the inclusion criteria [42]. Eligibility criteria are listed in Textbox 1.

Textbox 1. Inclusion and exclusion criteria.

Inclusion criteria

- Sample: patients with major chronic noncommunicable diseases as defined by the World Health Organization, including but not limited to cardiovascular diseases (hypertension, coronary heart disease, and stroke), diabetes mellitus (type 1 and type 2), cancers (various types), chronic respiratory diseases (chronic obstructive pulmonary disease and asthma), chronic kidney disease, digestive system diseases (chronic hepatitis, cirrhosis, inflammatory bowel disease, and celiac disease), musculoskeletal disorders (arthritis, fibromyalgia, and axial spondyloarthritis), and neurodegenerative conditions (Parkinson and Alzheimer disease), etc [43]. These disease categories were explicitly incorporated as search keywords to ensure comprehensive coverage of the chronic disease spectrum (see Multimedia Appendix 2).

- Phenomenon of interest: studies where LLMs (eg, general-purpose models like ChatGPT, Gemini, Claude, DeepSeek, Llama, or customized models developed for specific diseases) were applied in activities related to chronic disease care, including but not limited to: clinical decision support, patient health education, self-management support, symptom assessment, risk prediction, data interpretation, and emotional support.
- Design: all types of original research: quantitative (randomized controlled trials, nonrandomized controlled trials, cohort studies, case-control studies, cross-sectional studies, and quasi-experimental studies), qualitative (phenomenology, grounded theory, ethnography, and case studies), and mixed methods studies.
- Evaluation: any outcome measure related to LLM application, including model performance metrics, clinical outcomes, patient-reported outcomes, user experience, and safety indicators.
- Research type: original research published in peer-reviewed journals, with language restricted to Chinese or English.

Exclusion criteria

- Studies not involving LLM interventions.
- Studies not addressing chronic disease care.
- Publication types such as reviews, systematic reviews, meta-analyses, conference abstracts, dissertations, commentaries, letters, news reports, and book chapters.
- Articles where full text could not be obtained.
- Studies rated low quality during quality appraisal.

Selection Process and Data Collection Process

This review used the literature management software EndNote 21 (Clarivate). Two researchers (LZ and HL) independently performed study selection and data extraction. Initial screening was based on titles and abstracts; full-text screening was conducted for potentially relevant studies. For full texts that could not be accessed, attempts were made to contact corresponding authors or obtain copies through institutional library resources. Disagreements were resolved through discussion or by consulting a third researcher (PH).

Data Items

Data extraction was performed using a predesigned standardized data extraction form, including (1) basic information, (2) study design, (3) disease type or study population, (4) LLM, (5) application scenario, (6) key findings, (7) assessment tools, (8) framework dimension.

Study Risk-of-Bias Assessment

Methodological quality was assessed using the Mixed Methods Appraisal Tool version 2018 [44]. Based on the ratings, each study was assigned an overall quality rating: "High quality" (meeting all five criteria), "Moderate quality" (meeting four criteria), or "Low quality" (meeting three or fewer criteria). Two researchers (LZ and HL) independently conducted the quality appraisal. First, for each included study, the appropriate Mixed Methods Appraisal Tool category was determined based on its study design. Subsequently, both researchers independently read the full text, evaluated each of the five criteria for the respective category, and completed a predesigned quality assessment form. After completion, the researchers cross-checked their ratings. Disagreements were resolved through discussion. When there was disagreement, the decision was made through joint discussion with the third researcher (PH).

Synthesis Methods

Due to significant clinical and methodological heterogeneity among the included studies regarding study designs, intervention implementations, and outcome measures, a quantitative meta-analysis was not appropriate [45]. Therefore, this study used thematic synthesis to inductively synthesize the findings from the included literature, reporting in accordance with the Synthesis without meta-analysis guidelines [46]. The Synthesis without meta-analysis Checklist is provided in Checklist 3.

To enhance the analytical depth and theoretical explanatory power of the thematic synthesis, the recently proposed 3D evaluative framework (foundational: safety or privacy or fairness; intermediate: self-care enablement; top: design or system integration), informed by Orem's Self-Care Theory, was adopted as the theoretical lens. This framework is suitable for systematically analyzing user acceptance, trust, and the clinical integration of LLMs.

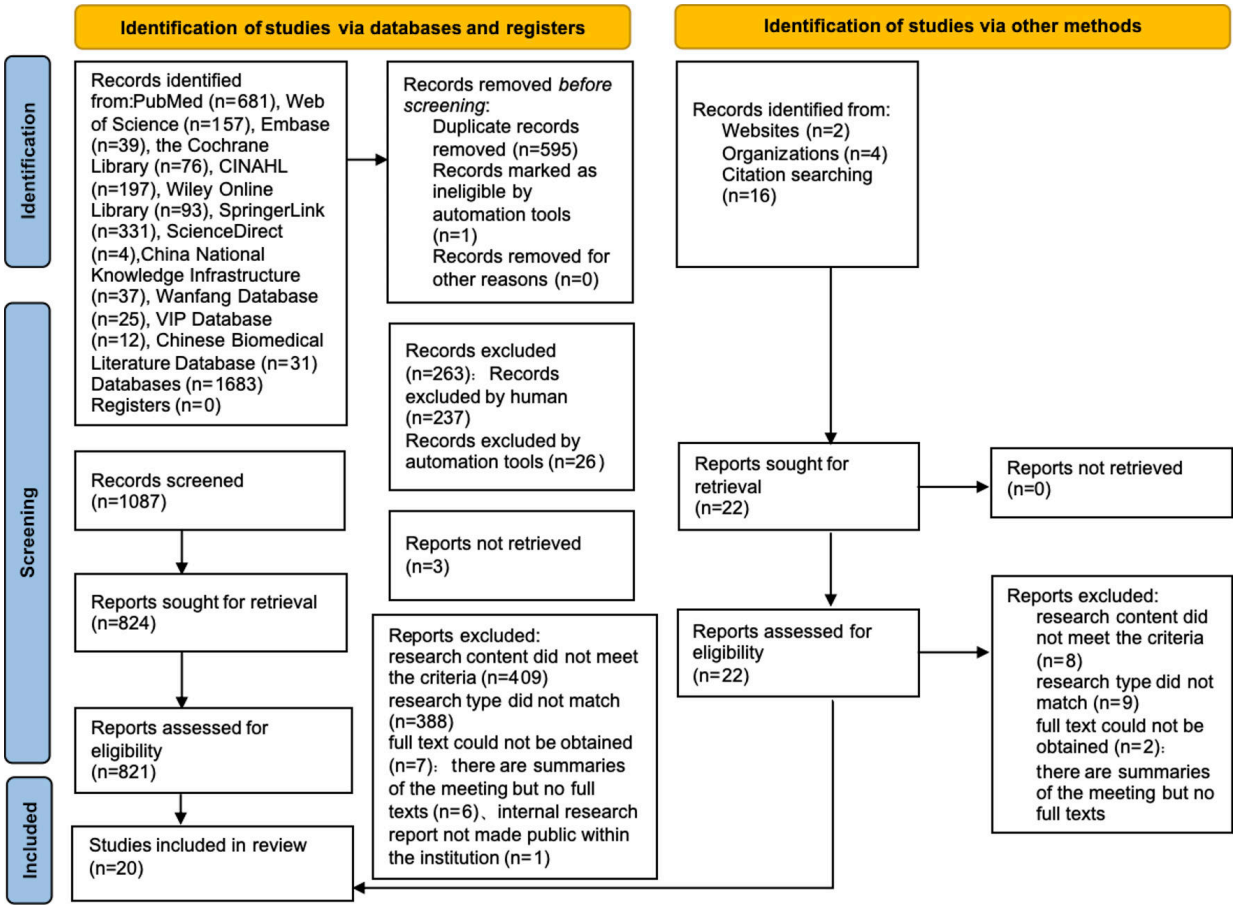
The thematic synthesis followed an iterative, 3-step process informed by Thomas and Harden approach [47]. First, line-by-line coding. Two researchers (LZ and PH) independently extracted text segments reporting findings on LLM application, efficacy, or challenges, and assigned descriptive codes using NVivo 14. Intercoder agreement was 92.3% (Cohen κ=0.88). Second, development of descriptive themes. Related codes were grouped (eg, "high accuracy in diagnosis," "effective for patient education" → "Application efficacy"). Third, generation of analytical themes. Descriptive themes were mapped onto the three dimensions of the Orem-informed framework. For example, codes related to "hallucination risks" and "data security" were interpreted as part of the foundational layer (safety and ethics); codes related to "improved self-efficacy" and "emotional support" were mapped to the intermediate layer (self-care enablement); and codes related to "workflow integration" and "cost-effectiveness" were interpreted as the top layer (design and operational integration). The complete coding table is available in Multimedia Appendix 2.

Results

Study Selection

The initial search yielded 2164 records (1959 in English and 205 in Chinese). Reports excluded (1) research content did not meet the criteria– that is, the study did not involve patients with chronic noncommunicable diseases as defined in our inclusion criteria, did not apply LLMs as the core intervention, or investigated LLMs for purposes unrelated to chronic disease care (eg, general medical education, administrative tasks, or nonclinical applications); (2) research type did not match–that is, the publication was not an original research article (eg, it was a review, systematic review, meta-analysis, conference abstract, commentary, letter, news report, or book chapter), which did not meet our design inclusion criteria; (3) full text could not be obtained even after attempts to contact corresponding authors and search through institutional library resources. Following the stepwise screening process, 20 studies were ultimately included, all published in English [48-67]. The detailed screening process is illustrated in Figure 1.

Figure 1. Literature screening process.



Study Characteristics

The 20 included studies were published in 2025 (n=17) [48-59,61,63,64,66,67] and 2026 (n=3) [60,62,65], systematically representing recent research progress on LLMs in chronic disease care. A variety of study designs were represented: cross-sectional studies (n=8) [48,50,53,54,57,62,63,67], quasi-experimental studies (n=3) [52,56,58], mixed methods studies (n=2) [50,55], randomized controlled trials (n=2) [60,64], comparative studies (n=2) [65,66], cohort study (n=1) [61], qualitative study (n=1) [58], and case study (n=1) [59].

The number of the types of literature involved in this study is shown in Figure 2. The geographical distribution of the studies was broad, involving the United States (n=5) [51,54,57,59,60], China (n=5) [50,52,53,58,62], Italy (n=2) [51,63], Turkey (n=2) [48,67], Saudi Arabia (n=2) [55,65], Spain (n=1) [49], Indonesia (n=1) [56], South Korea (n=1) [56], the Philippines (n=1) [56], Brazil (n=1) [61], Germany (n=1) [64], and Austria (n=1) [66]. The distribution of the countries included in the literature is shown in Figure 3.

Figure 2. Types of research literature.

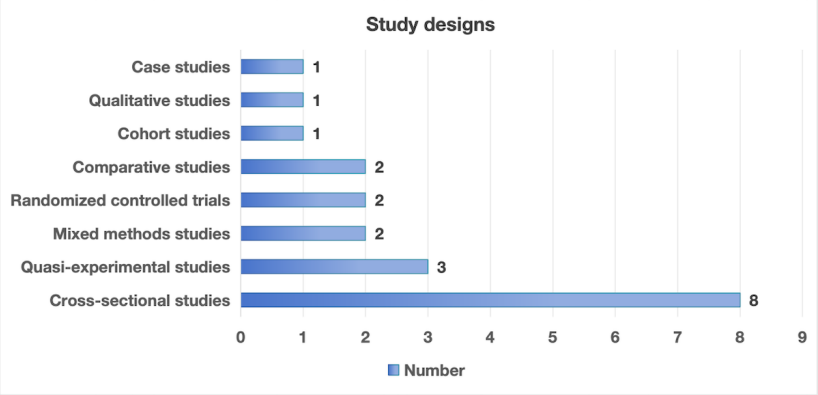
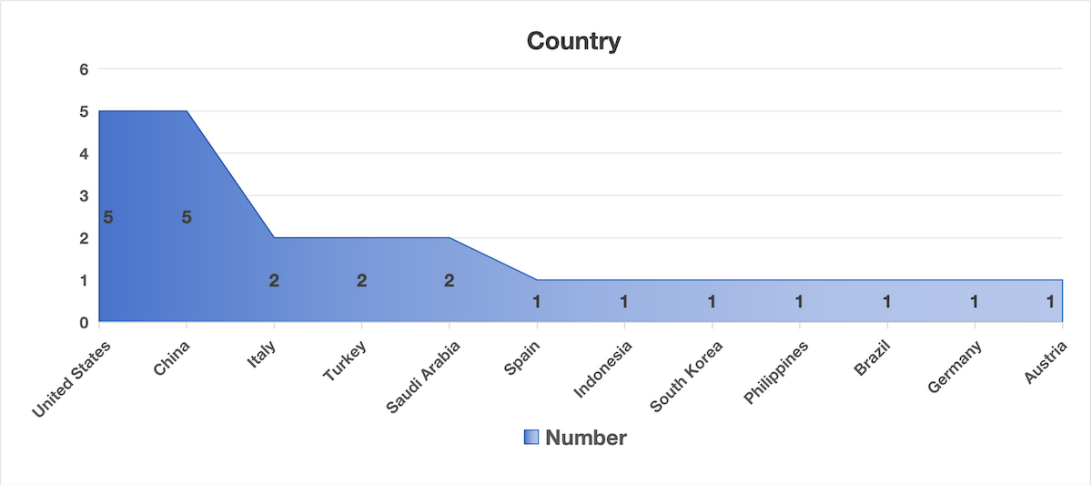


Figure 3. Distribution of the countries included in the literature.



Results of Individual Studies

The included studies covered a diverse range of chronic disease types, including diabetes and its related complications (n=7) [55-57,59,61,65,67], digestive system diseases (n=4) [51,52,60,63], cardiovascular diseases and thrombotic disorders (n=4) [54,60,64,66], as well as other conditions such as cancer [58], kidney stones [53], fibromyalgia [49], and axial spondyloarthritis [48]. The distribution of disease types included in this review is shown in Figure 4. Regarding application scenarios, the use of LLMs in chronic disease care primarily focused on patient health education (n=6) [54,55,62,63,66,67], clinical decision support (n=5) [48,51,60,64,66], and patient self-management and assessment (n=4) [49,56,61,65]. Additional studies explored scenarios like data summarization and interpretation [59] and disease prevention support [53]. The distribution of the application scenarios covered by this review is shown in Figure 5. In terms of model selection, most studies used general-purpose LLMs, with OpenAI's GPT series (including GPT-3.5, GPT-4, GPT-4o, GPT-4 Turbo, GPT-5, etc) being the most widely applied (n=17) [48-53,55,56,58,59,61-67]. Google Gemini series (including Gemini 2.0 Flash, Gemini 2.5 Pro, etc) was also evaluated in several studies (n=8) [48,50,55,57,60,63,66,67]. Furthermore, some studies explored customized or task-specific fine-tuned models, such as GutGPT for gastrointestinal diseases [52], KSrisk-GPT for identifying kidney stone risk factors [53], CARDIO for cardiovascular health education [55], and a Retrieval-Augmented Generation (RAG)-enhanced model for hepatitis C management [51]. Detailed characteristics of the included studies are presented in Table 1.

Figure 4. Distribution of disease types.

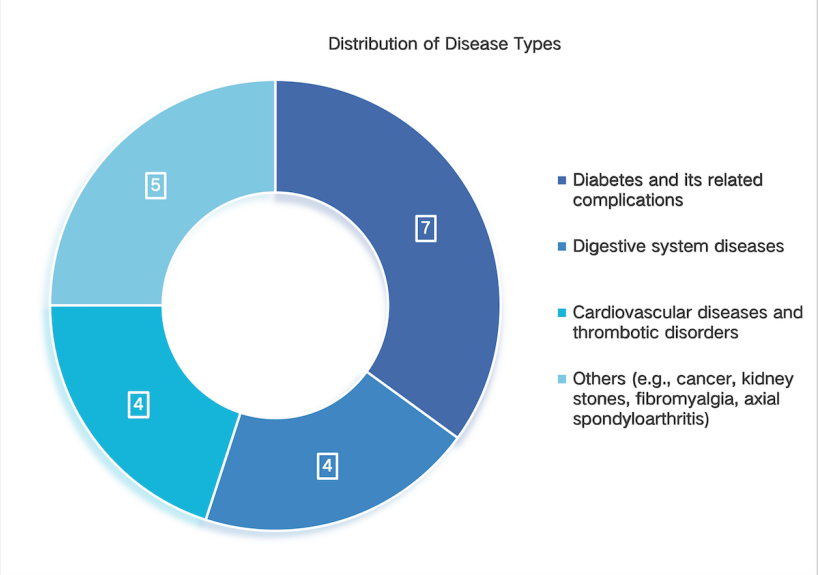


Figure 5. Distribution of application scenarios.

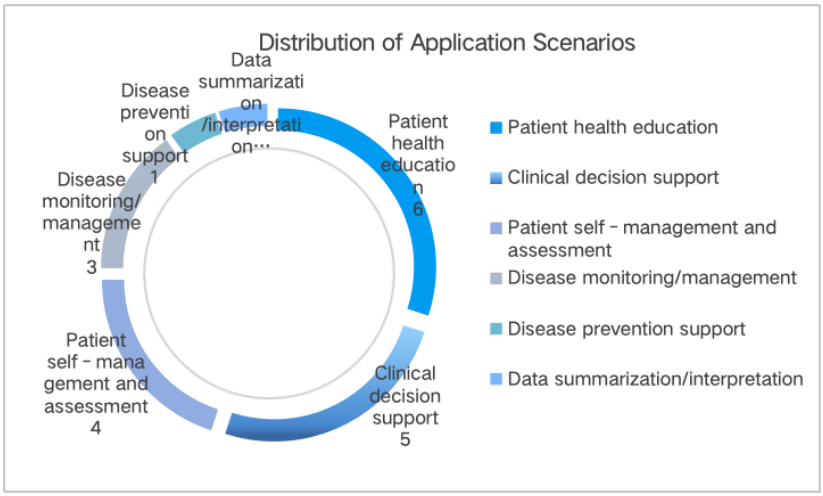


Table 1. Basic characteristics of included literature (n=20).

Author(s), Year	Nation	Study design	Disease type or study population	Large language model	Application scenario	Assessment tools	Key findings	Framework dimension
Usen et al, 2025 [48]	Turkey	Cross-sectional study	Axial spondyloarthritis	ChatGPT-3.5 or 4o, Gemini 2.0 Flash	Clinical decision support	7–point Likert scale, Flesch–Kincaid, ROUGE–L[a]	ChatGPT–4o and Gemini demonstrated superior reliability or usability; improved information accessibility.	Intermediate: clinical decision support+top: readability
Amidei et al, 2025 [49]	Spain	Mixed-methods study	Fibromyalgia, chronic pain	GPT-4	Patient self-assessment	Relevance, RMSE[b], Gwet AC2, Krippendorff alpha, revised FIQR[c]	High accuracy, comparable to experts; exhibited cross–linguistic adaptability.	Intermediate: self-assessment
Liet al, 2025	China	Cross-sectional study	Hepatitis B virus infection	ChatGPT-3.5 or 4.0, Google Gemini	Patient information support	4–point accuracy scale, Gunning Fog index, Flesch–Kincaid	ChatGPT–4.0 achieved the highest accuracy; however, excessive readability levels limited its usability.	Intermediate: accuracy+top: readability

Author(s), Year	Nation	Study design	Disease type or study population	Large language model	Application scenario	Assessment tools	Key findings	Framework dimension
Giuffrè et al, 2025 [51]	United States, Italy	Quasi-experimental study	Hepatitis C	GPT-4 Turbo	Treatment decision support	10–point Likert scale, Fleiss' Kappa, ICC[d], expert consensus	RAG–Top10[e] achieved 91.7% accuracy, reduced hallucinations, and improved guideline adherence.	Foundational: hallucination mitigation+intermediate: accuracy
Zhang et al, 2025 [52]	China	Quasi-experimental study	Gastrointestinal diseases	GutGPT	Disease diagnosis support	Expert evaluation, ROUGE or BLEU[f], public datasets	Diagnostic accuracy improved by 9.59%-22.47%; enhanced patient self–management.	Intermediate: diagnosis support+self-management
Mao et al,2025 [53]	China	Cross-sectional study	Kidney stones	GPT-4.0 (KSrisk-GPT)	Disease prevention support	Accuracy, Precision, Recall, F_1-score	KSrisk–GPT identified risk with 95.9% accuracy; improved patient cognition.	Intermediate: risk identification
Rullo et al, 2025 [54]	United States	Mixed-methods study	Cardiovascular disease and HIV infection	CARDIO (fine-tuned Llama 3.1-8B)	Patient health education	BLEU, METEOR[g], ROUGE, Kincaid, Jargon score, expert evaluation	Fine–tuning improved accuracy, readability, and professionalism; reduced jargon.	Intermediate: accuracy+top: readability
Jamil et al, 2025 [55]	Saudi Arabia	Cross-sectional study	Celiac disease, Type 1 diabetes	ChatGPT 3.5 or 4.0, Google Gemini	Patient health education	3–point accuracy or comprehensiveness scale, Flesch score, Flesch–Kincaid	ChatGPT 4.0 exhibited the best readability and consistency; overall high information accessibility.	Top: readability
Pawana et al, 2025 [56]	Indonesia, South Korea, Philippines	Quasi-experimental study	Diabetes mellitus	LLaMA 3.2, GPT-2, Phi-1, Gemma	Disease management and control	Accuracy, Precision, Recall, F_1-score, Confusion matrix	Fine–tuned LLaMA 3.2 was optimal for anomaly detection; improved monitoring capabilities.	Intermediate: disease monitoring
Kim et al, 2025 [57]	United States	Cross-sectional study	Diabetes mellitus	Llama 3.1, Gemini Pro 1.5, OpenAI o1, DeepSeek R1	Detection of disease symptoms	F_1-score, Precision, Recall, Accuracy, PHQ–[h]4 scale	Most LLMs[i] achieved >90% accuracy in symptom identification; Llama 3.1 405B performed best.	Intermediate: symptom detection
Zeng et al, 2025 [58]	China	Qualitative study	Cancer	ChatGPT, Kimichat	Patient information inquiry	Semi–structured interviews, Colaizzi's method of analysis	Physicians acknowledged management potential but concerns regarding liability and misinformation persisted.	Foundational: ethical liability+intermediate: potential
Healey et al, 2025 [59]	United States	Case study	Type 1 diabetes (CGM analysis)	GPT-4 (Data Analyst)	CGM data summarization	Accuracy or Completeness or Safety or Appropriateness ratings, Gwet's AC1	Perfect scores on 9/10 quantitative metrics; high scores (8–10/10) for accuracy, completeness, and safety in qualitative summaries.	Foundational: safety+Intermediate: data summarization
O'Sullivan et al, 2026 [60]	United States	Randomized controlled study	Inherited cardiomyopathies	AMIE (Gemini 2.0 Flash)	Clinical decision support	10–domain preference assessment, error analysis, self–reported time savings	Physicians assisted by AMIE were preferred by experts, made fewer errors (24.3% vs 13.1%), and had fewer	Intermediate: clinical decision support

Author(s), Year	Nation	Study design	Disease type or study population	Large language model	Application scenario	Assessment tools	Key findings	Framework dimension
							omissions (37.4% vs 17.8%).	
Furtado et al, 2025 [61]	Brazil	Cohort study	Type 2 diabetes	GPT-3.5-based MarIA	Patient management support	Engagement metrics, expert qualitative assessment, user questionnaires	Personalization increased engagement by 26% and quadrupled message length; no hallucinations but general advice required caution.	Foundational: no hallucinations +intermediate: personalization
Zhang et al, 2026 [62]	China	Cross-sectional study	Skin cancer	Doubao, DeepSeek, Wenxin Yiyan, Tongyi Qianwen, GPT-5	Patient health education	c–PEMAT–P[j], GQS[k], 7 readability metrics	GPT–5 had the highest GQS score; readability varied significantly among models, showing weak correlation with quality.	Top: readability
Bertani et al, 2025 [63]	Italy	Cross-sectional study	Celiac disease	ChatGPT-4, Claude 3.7, Gemini 2.0	Patient health education	5–point accuracy or clarity scale, readability metrics, misinformation detection	Gemini was optimal for accuracy, clarity, and readability; however, all models exhibited 13%-24% misinformation.	Foundational: misinformation
Carl et al, 2025 [64]	Germany	Randomized controlled study	Urological tumors	UroBot (GPT-4o+RAG) vs. ChatGPT	Clinical decision support	4–domain assessment, preference Likert scale, Krippendorff alpha	UroBot was superior in correctness of recommendations (73% vs 50%), source attribution (74% vs 30%), and verifiability (84% vs 35%); physicians trusted it more.	Foundational: source attribution+intermediate: trust
Alredaini et al, 2026 [65]	Saudi Arabia	Comparative study	Type 2 diabetes (blood glucose prediction)	GPT-4.1, MiniGPT, LLaMA-1B, LLaMA-7B, traditional or deep learning models	Patient disease prediction	MAE[l], RMSE, MAPE[m], R^2[n], SHAP[o], GPT explanation	GPT–4.1 performed best at 30/60 minutes; LLaMA–7B at 90 minutes; LLMs outperformed other models.	Intermediate: glucose prediction
Vladic et al, 2025 [66]	Austria	Comparative study	Venous thromboembolism	Le Chat Pixtral Large, DeepSeek-R1, ChatGPT-4.5	Patient health education, Clinical decision support	10–point adequacy scale, identifiability scale, potential harm assessment	LLMs provided superior patient education compared to experts; DeepSeek–R1 outperformed experts in clinical decision–making; physicians could not distinguish.	Intermediate: patient education+clinical decision
Yigit Yalcin et al, 2025 [67]	Turkey	Cross-sectional study	Gestational diabetes	ChatGPT-4o, DeepSeek R-1, Gemini 2.5 Pro, Grok 3.0	Patient health education	mDISCERN[p], GQS, readability metrics, TTR[q]	Grok and Gemini scored highest on mDISCERN or GQS; DeepSeek had the best readability, but all FRES scores were <60.	Top: readability

[a]ROUGE–L: Recall-Oriented Understudy for Gisting Evaluation – Longest Common Subsequence.
[b]RMSE: root mean square error.
[c]FIQR: Revised Fibromyalgia Impact Questionnaire.

[d]ICC: Intraclass Correlation Coefficient.
[e]RAG: retrieval-augmented generation.
[f]BLEU: Bilingual Evaluation Understudy.
[g]METEOR: Metric for Evaluation of Translation with Explicit ORdering.
[h]PHQ: Patient Health Questionnaire.
[i]LLM: large language model.
[j]c–PEMAT–P: Chinese Version of the Patient Education Materials Assessment Tool for Printable Materials
[k]GQS: Global Quality Score.
[l]MAE: mean absolute error.
[m]MAPE: mean absolute percentage error.
[n]R^2: coefficient of determination.
[o]SHAP: Shapley Additive Explanations.
[p]mDISCERN: Modified DISCERN.
[q]TTR: type-token ratio.

Reporting Biases

All 20 included studies met the inclusion criteria and underwent rigorous quality assessment. All were rated as moderate or high quality; no studies were rated as low quality. The detailed results of the methodological quality appraisal are shown in Table 2.

Table 2. Methodological quality assessment of included studies using the mixed methods appraisal.

Author(s), Year	Study design (MMAT[a] category)	S1[b]	S2	Q1[c]	Q2	Q3	Q4	Q5	Overall quality
Usen et al, 2025 [48]	Quantitative descriptive (4)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	High
Amidei et al, 2025 [49]	Mixed methods (5)	Yes	Yes	Yes	Yes	Yes	Yes	Unclear	Moderate
Li et al, 2025 [50]	Quantitative descriptive (4)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	High
Giuffrè et al, 2025 [51]	Quantitative nonrandomized (3)	Yes	Yes	Unclear	Yes	Yes	Yes	Yes	Moderate
Zhang et al, 2025 [52]	Quantitative nonrandomized (3)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	High
Mao et al,2025 [53]	Quantitative descriptive (4)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	High
Rullo et al, 2025 [54]	Mixed methods (5)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	High
Jamil et al, 2025 [55]	Quantitative descriptive (4)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	High
Pawana et al, 2025 [56]	Quantitative nonrandomized (3)	Yes	Yes	Unclear	Yes	Yes	Yes	Yes	Moderate
Kim et al, 2025 [57]	Quantitative descriptive (4)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	High
Zeng et al, 2025 [58]	Qualitative research (1)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	High
Healey et al, 2025 [59]	Mixed methods (5)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	High
O'Sullivan et al, 2026 [60]	Randomized controlled trial (2)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	High
Furtado et al, 2025 [61]	Quantitative nonrandomized (3)	Yes	Yes	Yes	Yes	Yes	No	Yes	Moderate
Zhang et al, 2026 [62]	Quantitative descriptive (4)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	High
Bertani et al, 2025 [63]	Quantitative descriptive (4)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	High
Carl et al, 2025 [64]	Randomized controlled trial (2)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	High
Alredaini et al, 2026 [65]	Quantitative nonrandomized (3)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	High
Vladic et al, 2025 [66]	Quantitative nonrandomized (3)	Yes	Yes	Unclear	Yes	Yes	Yes	Yes	Moderate
Yigit Yalcin et al, 2025 [67]	Quantitative descriptive (4)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	High

[a]MMAT: Mixed Methods Appraisal Tool.
[b]S: screening questions.
[c]Q: quality criteria.

Results of Syntheses

The thematic synthesis, informed by the Orem-based 3D evaluative framework, generated three analytical themes corresponding to the framework's core layers.

1. Theme 1: foundational layer: safety, privacy, and fairness as prerequisites. This theme captures the nonnegotiable baseline requirements for LLM deployment. Across studies, primary challenges included hallucination risks (ie, generation of factually incorrect information), data privacy and security concerns, and ambiguous ethical responsibility for AI-generated recommendations. For instance, Bertani et al [63] found that all tested models exhibited 13%-24% misinformation when providing patient education for celiac disease. From the perspective of the 3D framework, these issues represent failures in the foundational layer. Without robust safeguards against hallucinations and breaches of privacy, LLMs cannot be considered safe for integration into chronic care, regardless of their performance in other areas [68].
2. Theme 2: intermediate layer: self-care agency enablement through perceived usefulness and trust. This theme reflects the LLM's ability to meet self-care

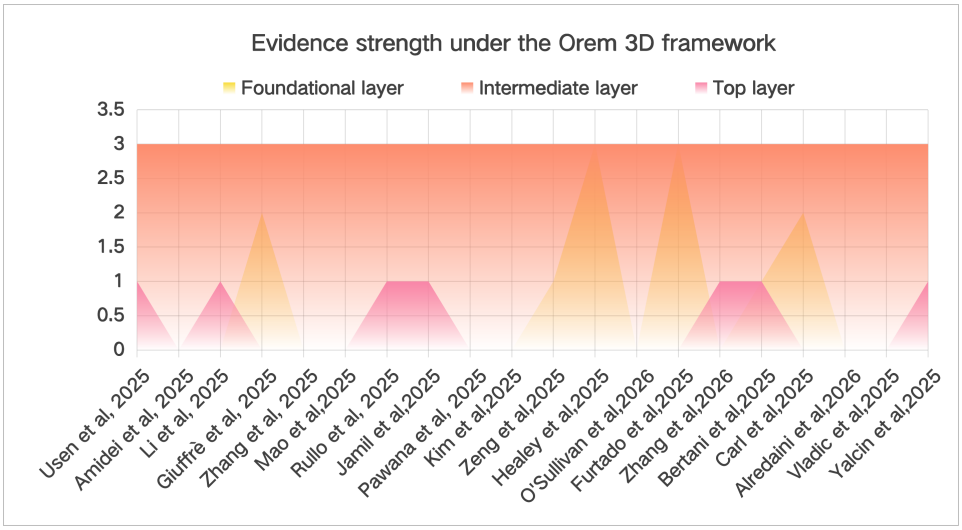
needs and enhance patient capabilities. Patient health education (n=6) emerged as the most common application [54,55,62,63,66,67]. For instance, Jamil et al [55] found ChatGPT-4.0 superior in readability and consistency for celiac disease and type 1 diabetes education. Clinical decision support (n=5) represented another core domain [48,51,60,64,66]. O'Sullivan et al.'s [60] randomized controlled trial showed that physicians assisted by AMIE (Google Research) made fewer errors. Patient self-management and assessment (n=4) was an emerging domain [49,56,61,65]. Amidei et al [49] found GPT-4 comparable to experts in pain self-assessment for fibromyalgia. From the framework's perspective, these scenarios demonstrate the LLM's perceived usefulness in providing cognitive support, thereby contributing to the patient's self-care agency [39]. Three key drivers of enhanced efficacy were identified in this layer: (a) foundation model iteration (eg, GPT-4 series outperformed GPT-3.5); (b) RAG, which improved accuracy from 36.6% to 91.7% in one hepatitis C study; (c) specialized fine-tuning, which improved diagnostic accuracy by 9.6%-–22.5% for GutGPT [51,61,64]. Within the framework, RAG and finding primarily enhance perceived usefulness and trust, thereby solidifying the LLM's role in enabling self-care.

3. Theme 3: top layer: design adaptability and system integration. This final theme concerns LLM's usability and its seamless integration into real-world care workflows. Multiple studies reported that the reading difficulty of LLM-generated content often exceeded patient comprehension levels (eg, Flesch-Kincaid grade levels >10) [48,50,54]. This readability mismatch represents a barrier to perceived ease of use, a key component of design adaptability, particularly for patients with lower health literacy [30]. Furthermore, some studies evaluated LLMs in controlled, siloed settings, with little evidence on their integration into existing nursing workflows, impact on nursing workload, or long-term cost-effectiveness [48-55,67]. These gaps highlight the underdeveloped nature of the top layer of the evaluation framework, limiting clinical adoption [69].

Evidence Distribution Across the 3D Framework

To visually demonstrate the coverage, intensity, and gap situation of the evidence based on Orem's 3D framework, this review has constructed an area chart (Figures 4 and 6). The x-axis represents individual studies; the y-axis represents the base layer (safety or privacy or equity: illusion risk, data privacy, ethical responsibility, algorithm fairness, etc), the middle layer (self-care ability: patient education, clinical decision support, self-management or monitoring, etc), and the top layer (design and system integration: readability matching, workflow integration, cost-effectiveness, etc).

Figure 6. The area chart of evidence strength under the Orem 3D framework [48-67].



In the foundational layer, hallucination risk and ethical responsibility show moderate evidence, but data privacy and algorithmic fairness are largely unassessed. The intermediate layer has the strongest evidence, especially in patient education, clinical decision support, and self-management, confirming that LLMs primarily enable self-care under controlled conditions. The top layer shows critical gaps: readability matching is weak, while workflow integration and cost-effectiveness are almost entirely absent. Thus, the evidence is unbalanced—strong in the intermediate layer, weak in foundational and top layers. Future research should prioritize foundational safety or fairness validation and top-layer integration and economic evaluations.

Discussion

Summary of Main Findings

This mixed methods systematic review synthesized 20 studies on LLMs in chronic disease care and evaluated them through a theory-driven, 3D framework informed by Orem's

Self-Care Theory [48-54,59,61-64,70]. Three main findings emerged. First, foundational safety, privacy, and fairness issues remain persistent barriers. Hallucinations, data security concerns, and potential algorithmic biases were consistently reported across studies, indicating that current LLMs do not yet meet the baseline requirements for safe deployment in chronic care. Second, LLMs demonstrate clear potential to enable patient self-care agency within the intermediate layer. Evidence was strongest for patient health education, clinical decision support, and self-management assistance, with technical enhancements (RAG, fine-tuning, and model iteration) further improving perceived usefulness and trust. Third, the top layer of the framework—design adaptability and system integration—is critically underdeveloped. Readability mismatches between LLM-generated content and patient health literacy levels were ubiquitous, yet evidence on workflow integration, nursing workload, cost-effectiveness, and long-term sustainability was almost entirely absent.

In summary, current LLM applications in chronic disease care have proven their value primarily within controlled technical validation settings (intermediate layer), but lack foundational safety assurances and operational integration. This review provides a structured, theory-informed roadmap for future research to address these critical gaps.

Comparison With Related Studies

The findings of this review both complement and extend prior systematic evidence. Previous work has established the feasibility of LLMs in chronic disease management, yet several important limitations remain. Li et al [33] conducted a mixed methods review focusing primarily on quantitative feasibility metrics. While their work provided a valuable overview, it did not integrate established nursing theories to ground the analysis in core principles of self-care. Consequently, their evaluation remained at the level of technical performance without a structured, multidimensional framework. Watson et al [34] performed an integrative review of generative AI in general nursing practice. Although comprehensive in scope, their review did not specifically address the unique demands of chronic disease care, nor did it offer a theory-driven approach to assess how LLMs might support patient self-care agencies or integrate into existing nursing workflows.

This review addresses these gaps directly. By adopting Orem's Self-Care Theory as an analytical lens and operationalizing it through a 3D evaluative framework—foundational safety and ethics, self-care enablement, and operational integration—move beyond isolated performance metrics to provide a structured, clinically meaningful assessment of LLM readiness for chronic care. Moreover, this review focuses on high-burden chronic conditions, and the inclusion of both quantitative and qualitative evidence allows for a more granular, human-centric critique of current applications, revealing not only where LLMs show promise but also where critical gaps persist.

Evidence Distribution and Study Characteristics

This systematic review of the mixed-method systematically collated and evaluated 20 studies, revealing the current application status, core efficacy, and challenges of LLMs in chronic disease care [48-54,59,61-64,70].

According to World Health Organization statistics, chronic diseases are the leading cause of death globally, accounting for approximately 41 million deaths annually, or 74% of all deaths [71]. Among these, diabetes, cardiovascular diseases, cancer, and chronic respiratory diseases constitute the bulk of this burden, affecting the quality of life and health expectancy of billions worldwide [72]. Given that chronic disease care is characterized by frequent daily monitoring, high dependence on behavioral interventions, and continuous health information needs, accessibility and powerful information integration and personalized generation capabilities of LLMs offer potential solutions to these persistent, dynamic care challenges [73]. In terms of disease spectrum, LLM applications have covered various chronic conditions, including diabetes, cardiovascular disease, digestive system diseases, cancer, and chronic pain, demonstrating their adaptability across diverse disease contexts [50-52,55,57-59,61-63,66,67]. Diabetes and its related complications were the most studied area (n=7) [55-57,59,61,65,67], which is consistent with its status as one of the world's major chronic diseases with complex management demands reliant on continuous monitoring and behavioral intervention [27,74]. However, despite cardiovascular and chronic respiratory diseases being major sources of global disease burden, the number of relevant studies was relatively low, suggesting a need for increased focus on these high-burden conditions in future research [75,76].

Regarding study design, the included literature predominantly consisted of quantitative descriptive studies (n=8) [48,50,53,55,57,62,63,67] and nonrandomized quantitative studies (n=5) [52,56,58,65,66], with fewer randomized controlled trials [60,64], mixed methods [49,55], and qualitative studies [58]. This distribution reflects that the current field is still in a preliminary stage, primarily focused on technical validation and efficacy exploration [77]. Qualitative and mixed methods of research hold unique value in understanding user (both patient and health care provider) acceptance, usage experiences, and potential socio-cultural barriers to new technologies [78]. This is particularly crucial during the early integration of novel technologies like LLMs into complex clinical environments, where in-depth exploration of nurse-patient interaction, trust-building, and ethical dilemmas from a "human-centric" perspective is essential to guide technical optimization and practical translation [79-81]. Future studies should incorporate more such designs to complement quantitative insights with deeper contextual understanding.

Geographically, the included studies originated mainly from countries such as Turkey, the United States, Italy, China, and Germany, indicating a shared global interest in LLM applications [48,50-53,55,57-60,62,63,67]. However, this also highlights that current evidence predominantly

comes from high-income countries or regions, with underrepresentation from low- and middle-income countries or medically underserved areas [82,83]. Given that factors like health literacy levels, health care resource accessibility, and cultural beliefs can significantly influence the effectiveness of LLM applications, future research should conduct localized studies across diverse health care settings to promote global health equity in technology deployment [84,85].

Drivers of Intermediate-Layer Performance

This study identified three drivers of improved technical performance: foundation model iteration, RAG, and specialized fine-tuning. The research indicates RAG enhances clinicians' perceived usefulness by improving accuracy and source verifiability [86-88]. Fine-tuning enhances users' perceived usefulness by tailoring outputs to specific specialty knowledge needs [89]. However, the finding by Li et al [50] that ChatGPT-4.0 had the highest accuracy but excessive readability levels illustrates that high technical performance within the intermediate layer does not guarantee success in the "top layer" (design adaptability). Thus, future technical development should simultaneously address both cognitive accuracy (intermediate layer) and user-centric design (top layer) [90-92].

Documented Challenges and the Foundational Layer: Safety, Privacy, Fairness

Primary challenges consistently reported across studies align with the "foundational layer" of framework [50,61,66]. Viewed through the lens of Orem's theory, these are not merely technical glitches but fundamental threats to the nursing system's ability to provide safe and ethical care [93]. Hallucinations, for instance, could actively undermine a patient's self-care agency by providing dangerous misinformation [94]. Data privacy breaches violate the core ethical principle of nonmaleficence [95-97]. These foundational issues should be resolved before LLMs can reliably support any form of compensated or supportive nursing system [98-100].

Future Research Directions

Based on the evidence gaps identified in this review, future research should prioritize the neglected "top layer" of framework: (1) clinical effectiveness trials measuring patient-centered outcomes (eg, disease control, quality of life) to assess the real-world value of LLMs beyond technical performance [101] [102]; (2) implementation science studies that evaluate the integration of LLMs into existing nursing workflows, including their impact on nursing workload, cost-effectiveness, and scalability [103,104]; (3) readability-adaptive interfaces that can tailor output to individual patient health literacy levels, directly addressing the "design adaptability" sub-dimension [105,106]; (4) safety assurance systems incorporating real-time hallucination detection and verification mechanisms to secure the "foundational layer" [107]; (5) standardized evaluation frameworks developed through interdisciplinary consensus to capture all three dimensions of LLM performance.

Limitations

Although this mixed methods systematic review strictly adhered to evidence-based research methodologies, several limitations should be acknowledged: (1) the search was limited to published Chinese and English language literature, excluding gray literature, commentaries, books, and news reports, which may introduce publication bias. In addition, although attempts have been made to contact the authors, there are still some articles whose full texts could not be fully retrieved during the search process; (2) despite strict adherence to methodological protocols, the possibility of subjective bias during the synthesis process cannot be entirely ruled out; (3) the study populations originated from various countries with differences in economic status, national policies, and cultural backgrounds, which may affect the comprehensiveness and cross-cultural applicability of the conclusions.

Conclusion

Applying a novel, Orem-informed 3D framework, this mixed methods systematic review concludes that the current evidence on LLMs in chronic disease care primarily supports the intermediate layer—that is, self-care enablement through useful and trustworthy tools. Within this layer, LLMs demonstrate promising technical performance in areas such as patient education, clinical decision support, and self-management assistance under controlled conditions. However, persistent challenges related to the foundational layer (safety, privacy, and fairness) remain, and evidence for the top layer (design adaptability and system integration) is still limited. These gaps may constrain the readiness of LLMs for routine clinical implementation at present. Future research would benefit from prioritizing clinical effectiveness trials, implementation science, and patient-centered design to address these shortcomings.

Funding

This research did not receive any specific funding from public, commercial or nonprofit sectors.

Data Availability

All data supporting the findings of this systematic review are presented within the manuscript and its supplementary materials. No additional datasets were generated or analyzed for this study.

Authors' Contributions

LZ and PH contributed equally to this work and share first authorship. HL is the primary corresponding author, and RJ is the co-corresponding author.
Data curation: LZ (lead), HL (equal), PH (supporting)
Formal analysis: LZ (lead), HL (equal), PH (supporting), RX (supporting)
Funding acquisition: HL. Investigation: JS
Methodology: LZ (lead), HL (equal), PH (supporting), JS (supporting)
Project administration: HL (lead), LZ (equal), RJ (supporting)
Resources: HL
Supervision: RJ
Validation: RJ
Visualization: HL (lead), RJ (supporting)
Writing - original draft: LZ (lead), HL (equal), PH (supporting)
Writing - review & editing: LZ (lead), HL (supporting), PH (supporting)

Conflicts of Interest

None declared.

Multimedia Appendix 1

The complete search formula for all databases.
[DOCX File (Microsoft Word File), 20 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Subject comprehensive coding table.
[DOCX File (Microsoft Word File), 16 KB-Multimedia Appendix 2]

Checklist 1

PRISMA 2020 expanded checklist.
[DOCX File (Microsoft Word File), 98 KB-Checklist 1]

Checklist 2

PRISMA-S checklist.
[DOCX File (Microsoft Word File), 17 KB-Checklist 2]

Checklist 3

SWiM checklist.
[DOCX File (Microsoft Word File), 19 KB-Checklist 3]

References

1. Zhang T, Jiang H, Xu X, Zhao Z, Zhou M. Non-communicable disease burden in China, 1990-2023: evidence from the Global Burden of Disease Study 2023. Chin Med J (Engl). Jan 5, 2026;139(1):48-57. [doi: 10.1097/CM9.0000000000003898] [Medline: 41199464]
2. Armocida B, Klepp KI, Onder G, et al. Advancing Europe's non-communicable diseases agenda through cross-national collaboration: translating WHO-Europe findings into actionable strategies. Lancet Reg Health Eur. Aug 2025;55:101361. [doi: 10.1016/j.lanepe.2025.101361] [Medline: 40950942]
3. NCD Countdown 2030 collaborators. NCD Countdown 2030: pathways to achieving Sustainable Development Goal target 3.4. Lancet. Sep 2020;396(10255):918-934. [doi: 10.1016/S0140-6736(20)31761-X] [Medline: 32891217]
4. GBD 2023 Disease and Injury and Risk Factor Collaborators. Burden of 375 diseases and injuries, risk-attributable burden of 88 risk factors, and healthy life expectancy in 204 countries and territories, including 660 subnational locations, 1990-2023: a systematic analysis for the Global Burden of Disease Study 2023. Lancet. Oct 18, 2025;406(10513):1873-1922. [doi: 10.1016/S0140-6736(25)01637-X] [Medline: 41092926]
5. Feng G, Weng F, Lu W, et al. Artificial intelligence in chronic disease management for aging populations: a systematic review of machine learning and NLP applications. Int J Gen Med. 2025;18:3105-3115. [doi: 10.2147/IJGM.S516247] [Medline: 40529344]
6. Endalamaw A, Zewdie A, Wolka E, Assefa Y. Care models for individuals with chronic multimorbidity: lessons for low- and middle-income countries. BMC Health Serv Res. Aug 5, 2024;24(1):895. [doi: 10.1186/s12913-024-11351-y] [Medline: 39103802]
7. Struss L, Afia Gyinae Wilberforce P, Opoku D, et al. Barriers and facilitators to the implementation and scale-up of a mHealth integrated care program for diabetes and hypertension in Ghana: a qualitative study of the Akoma Pa program. BMC Health Serv Res. Feb 21, 2026;26(1):304. [doi: 10.1186/s12913-026-14175-0] [Medline: 41723421]

8. Noubiap JJ, Nansseu JR, Nyaga UF, et al. Worldwide trends in metabolic syndrome from 2000 to 2023: a systematic review and modelling analysis. Nat Commun. Dec 6, 2025;17(1):573. [doi: 10.1038/s41467-025-67268-5] [Medline: 41350289]
9. Zhang H, Chen QF, Lip GYH, et al. Burden of metabolic diseases, 1990-2023, with forecasts to 2030 for the Asia-Pacific region. Metabolism. Jun 2026;179:156575. [doi: 10.1016/j.metabol.2026.156575] [Medline: 41763388]
10. Bae YS, Park S, Noh C, et al. A comprehensive digital medicine platform for hypertension and diabetes care in primary care: a real-world feasibility test. Digit Health. 2025;11:20552076251344375. [doi: 10.1177/20552076251344375] [Medline: 40416073]
11. Wamala-Andersson S, Uitto M, Diop-Christensen A, et al. Understanding digital health literacy as a structural determinant of health and public health capability. BMC Glob Public Health. Jan 14, 2026;4(1):7. [doi: 10.1186/s44263-025-00236-9] [Medline: 41535922]
12. Liang S, Kennedy E, Gale N, et al. Implementation of precision medicine in treating non-communicable diseases: a systematic review. J Transl Med. Oct 27, 2025;23(1):1174. [doi: 10.1186/s12967-025-07201-y] [Medline: 41146138]
13. Dave M, Patel N. Artificial intelligence in healthcare and education. Br Dent J. May 2023;234(10):761-764. [doi: 10.1038/s41415-023-5845-2] [Medline: 37237212]
14. Cunningham JW, Abraham WT, Bhatt AS, et al. Artificial intelligence in cardiovascular clinical trials. J Am Coll Cardiol. Nov 12, 2024;84(20):2051-2062. [doi: 10.1016/j.jacc.2024.08.069] [Medline: 39505413]
15. Yip HF, Li Z, Zhang L, Lyu A. Large language models in integrative medicine: progress, challenges, and opportunities. J Evid Based Med. Jun 2025;18(2):e70031. [doi: 10.1111/jebm.70031] [Medline: 40384541]
16. Sozen Yanik I, Sahin Hazir D, Bilgin Avsar D. Cross-lingual performance of large language models in maxillofacial prosthodontics: a comparative evaluation. BMC Oral Health. Oct 17, 2025;25(1):1630. [doi: 10.1186/s12903-025-07035-6] [Medline: 41107888]
17. Kianian R, Sun D, Crowell EL, Tsui E. The use of large language models to generate education materials about uveitis. Ophthalmol Retina. Feb 2024;8(2):195-201. [doi: 10.1016/j.oret.2023.09.008] [Medline: 37716431]
18. Li R, Wu T. Delving into the practical applications and pitfalls of large language models in medical education: narrative review. Adv Med Educ Pract. 2025;16:625-636. [doi: 10.2147/AMEP.S497020] [Medline: 40271151]
19. Zheng T. Comparative analysis of AI tools for disseminating ADA 2025 Diabetes Care Standards: implications for cardiovascular physicians. J Diabetes. Mar 2025;17(3):e70072. [doi: 10.1111/1753-0407.70072] [Medline: 40051065]
20. Harrington J, Booth RG, Jackson KT. Large language models in nursing education: concept analysis. JMIR Nurs. Aug 22, 2025;8:e77948. [doi: 10.2196/77948] [Medline: 40845300]
21. Huo B, Boyle A, Marfo N, et al. Large language models for chatbot health advice studies: a systematic review. JAMA Netw Open. Feb 3, 2025;8(2):e2457879. [doi: 10.1001/jamanetworkopen.2024.57879] [Medline: 39903463]
22. Jeong H, Han SS, Yu Y, Kim S, Jeon KJ. How well do large language model-based chatbots perform in oral and maxillofacial radiology? Dentomaxillofac Radiol. Sep 1, 2024;53(6):390-395. [doi: 10.1093/dmfr/twae021] [Medline: 38848473]
23. Chen D, Alnassar SA, Avison KE, Huang RS, Raman S. Large language model applications for health information extraction in oncology: scoping review. JMIR Cancer. Mar 28, 2025;11:e65984. [doi: 10.2196/65984] [Medline: 40153782]
24. Serugunda HM, Jianquan O, Kasujja Namatovu H, et al. Using large language models for chronic disease management tasks: scoping review. JMIR Med Inform. Sep 29, 2025;13:e66905. [doi: 10.2196/66905] [Medline: 41021927]
25. Agrawal M, Chen IY, Gulamali F, Joshi S. The evaluation illusion of large language models in medicine. NPJ Digit Med. Oct 7, 2025;8(1):600. [doi: 10.1038/s41746-025-01963-x] [Medline: 41057566]
26. Shan G, Chen X, Wang C, et al. Comparing diagnostic accuracy of clinical professionals and large language models: systematic review and meta-analysis. JMIR Med Inform. Apr 25, 2025;13:e64963. [doi: 10.2196/64963] [Medline: 40279517]
27. Mondal A, Naskar A, Roy Choudhury B, et al. Evaluating the performance and safety of large language models in generating type 2 diabetes mellitus management plans: a comparative study with physicians using real patient records. Cureus. Mar 2025;17(3):e80737. [doi: 10.7759/cureus.80737] [Medline: 40248538]
28. You Q, Zhou L, Ma Y, et al. Comparison of ChatGPT-3.5, ChatGPT-4.0 and DeepSeek in generating dietary plans for patients with chronic kidney disease: a focus on nutritional accuracy and dietary inflammation. Nutrition. Feb 2026;142:112957. [doi: 10.1016/j.nut.2025.112957] [Medline: 41135437]
29. Yu E, Chu X, Zhang W, et al. Large language models in medicine: applications, challenges, and future directions. Int J Med Sci. 2025;22(11):2792-2801. [doi: 10.7150/ijms.111780] [Medline: 40520893]

30. Li H, Lin Y, Lv L. Performance evaluation of mainstream large language models in autoimmune hepatitis patient education: a comparative study of readability, quality, and reliability. Front Public Health. 2026;14:1805848. [doi: 10.3389/fpubh.2026.1805848] [Medline: 41938975]
31. Rust P, Frings J, Meister S, Fehring L. Evaluation of a large language model to simplify discharge summaries and provide cardiological lifestyle recommendations. Commun Med (Lond). May 29, 2025;5(1):208. [doi: 10.1038/s43856-025-00927-2] [Medline: 40442348]
32. Triantafyllidis A, Segkouli S, Kokkas S, et al. Large language models for cardiovascular disease, cancer, and mental disorders: a review of systematic reviews. Healthcare (Basel). Dec 24, 2025;14(1):45. [doi: 10.3390/healthcare14010045] [Medline: 41516976]
33. Li C, Zhao Y, Bai Y, et al. Unveiling the potential of large language models in transforming chronic disease management: mixed methods systematic review. J Med Internet Res. Apr 16, 2025;27:e70535. [doi: 10.2196/70535] [Medline: 40239198]
34. Watson AL, Bond C, Aveyard H, Smith GD, Jackson D. Generative AI at the bedside: an integrative review of applications and implications in clinical nursing practice. J Clin Nurs. Nov 26, 2025. [doi: 10.1111/jocn.70151] [Medline: 41293898]
35. Mallinar N, Heydari AA, Liu X, et al. A scalable framework for evaluating health language models. NPJ Digit Med. Feb 27, 2026;9(1):437. [doi: 10.1038/s41746-026-02492-x] [Medline: 41760912]
36. Sblendorio E, Dentamaro V, Lo Cascio A, Germini F, Piredda M, Cicolini G. Integrating human expertise & automated methods for a dynamic and multi-parametric evaluation of large language models' feasibility in clinical decision-making. Int J Med Inform. Aug 2024;188:105501. [doi: 10.1016/j.ijmedinf.2024.105501] [Medline: 38810498]
37. Nasiri M, Jafari Z, Rakhshan M, et al. Application of Orem's theory-based caring programs among chronically ill adults: a systematic review and dose-response meta-analysis. Int Nurs Rev. Mar 2023;70(1):59-77. [doi: 10.1111/inr.12808] [Medline: 36418147]
38. Hartweg DL, Metcalfe SA. Orem's Self-Care Deficit Nursing Theory: relevance and need for refinement. Nurs Sci Q. Jan 2022;35(1):70-76. [doi: 10.1177/08943184211051369] [Medline: 34939484]
39. Iovino P, Uchmanowicz I, Vellone E. Self-care: an effective strategy to manage chronic diseases. Adv Clin Exp Med. Aug 2024;33(8):767-771. [doi: 10.17219/acem/191102] [Medline: 39194160]
40. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. Mar 29, 2021;372:n71. [doi: 10.1136/bmj.n71] [Medline: 33782057]
41. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. Syst Rev. Jan 26, 2021;10(1):39. [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
42. Yousefi F, Dehnavieh R, Laberge M, et al. Opportunities, challenges, and requirements for artificial intelligence (AI) implementation in Primary Health Care (PHC): a systematic review. BMC Prim Care. Jun 9, 2025;26(1):196. [doi: 10.1186/s12875-025-02785-2] [Medline: 40490689]
43. Milton K, Salvo D, Gomersall SR. The Political Declaration of the Fourth United Nations High-Level Meeting on Noncommunicable Diseases and Mental Health-Implications for Physical Activity Promotion. J Phys Act Health. Feb 1, 2026;23(2):141-142. [doi: 10.1123/jpah.2025-0895] [Medline: 41468210]
44. Hong QN, Gonzalez-Reyes A, Pluye P. Improving the usefulness of a tool for appraising the quality of qualitative, quantitative and mixed methods studies, the Mixed Methods Appraisal Tool (MMAT). J Eval Clin Pract. Jun 2018;24(3):459-467. [doi: 10.1111/jep.12884] [Medline: 29464873]
45. Lensen S. When to pool data in a meta-analysis (and when not to)? Fertil Steril. Jun 2023;119(6):902-903. [doi: 10.1016/j.fertnstert.2023.03.015] [Medline: 36948444]
46. Campbell M, McKenzie JE, Sowden A, et al. Synthesis without meta-analysis (SWiM) in systematic reviews: reporting guideline. BMJ. Jan 16, 2020;368:l6890. [doi: 10.1136/bmj.l6890] [Medline: 31948937]
47. Thomas J, Harden A. Methods for the thematic synthesis of qualitative research in systematic reviews. BMC Med Res Methodol. Jul 10, 2008;8:45. [doi: 10.1186/1471-2288-8-45] [Medline: 18616818]
48. Usen A, Kuculmez O. Evaluation of the performance of large language models in the management of axial spondyloarthropathy: analysis of EULAR 2022 Recommendations. Diagnostics (Basel). Jun 7, 2025;15(12):12. [doi: 10.3390/diagnostics15121455] [Medline: 40564776]
49. Amidei J, Nieto R, Kaltenbrunner A, Ferreira De Sá JG, Serrat M, Albajes K. Exploring the capacity of large language models to assess the chronic pain experience: algorithm development and validation. J Med Internet Res. Mar 31, 2025;27:e65903. [doi: 10.2196/65903] [Medline: 40163858]
50. Li Y, Huang CK, Hu Y, Zhou XD, He C, Zhong JW. Exploring the performance of large language models on hepatitis B infection-related questions: a comparative study. World J Gastroenterol. 2025;31(3):101092. [doi: 10.3748/wjg.v31.i3.101092] [Medline: 39839898]

51. Giuffrè M, Pugliese N, Kresevic S, et al. From guidelines to real-time conversation: expert-validated retrieval-augmented and fine-tuned GPT-4 for hepatitis C management. Liver Int. Oct 2025;45(10):e70349. [doi: 10.1111/liv.70349] [Medline: 40960299]
52. Zhang RY, Qiang PP, Hao YX, et al. GutGPT: a multidimensional knowledge-enhanced large language model for gastrointestinal medicine. J Biomed Inform. Sep 2025;169:104885. [doi: 10.1016/j.jbi.2025.104885] [Medline: 40720988]
53. Mao C, Li J, Pang PCI, Zhu Q, Chen R. Identifying kidney stone risk factors through patient experiences with a large language model: text analysis and empirical study. J Med Internet Res. May 22, 2025;27:e66365. [doi: 10.2196/66365] [Medline: 40403294]
54. Rullo R, Maatouk A, Huang T, et al. Interdisciplinary development and fine-tuning of CARDIO, a large language model for cardiovascular health education in HIV care: tutorial. J Med Internet Res. Sep 12, 2025;27:e77053. [doi: 10.2196/77053] [Medline: 40794856]
55. Jamil SF, Alshathri NN, Alsalamah SS, et al. Leveraging large language models to inform paediatric chronic condition care: a cross-sectional study. BMJ Paediatr Open. Aug 14, 2025;9(1):e003742. [doi: 10.1136/bmjpo-2025-003742] [Medline: 40813141]
56. Pawana I, Astillo PV, You I. Lightweight LLM-based anomaly detection framework for securing IoTMD enabled diabetes management control systems. IEEE J Biomed Health Inform. Jun 9, 2025;PP. [doi: 10.1109/JBHI.2025.3577604] [Medline: 40489281]
57. Kim J, Ma SP, Chen ML, et al. Optimizing large language models for detecting symptoms of depression/anxiety in chronic diseases patient communications. NPJ Digit Med. Sep 30, 2025;8(1):580. [doi: 10.1038/s41746-025-01969-5] [Medline: 41028413]
58. Zeng L, Li Q, Zuo Y, Zhang Y, Li Z. Perceptions and attitudes of Chinese oncologists toward endorsing AI-driven chatbots for health information seeking among patients with cancer: phenomenological qualitative study. J Med Internet Res. Jul 23, 2025;27:e71418. [doi: 10.2196/71418] [Medline: 40699917]
59. Healey E, Tan ALM, Flint KL, Ruiz JL, Kohane I. A case study on using a large language model to analyze continuous glucose monitoring data. Sci Rep. Jan 7, 2025;15(1):1143. [doi: 10.1038/s41598-024-84003-0] [Medline: 39774031]
60. O'Sullivan JW, Palepu A, Saab K, et al. A large language model for complex cardiology care. Nat Med. Feb 2026;32(2):616-623. [doi: 10.1038/s41591-025-04190-9] [Medline: 41652123]
61. Furtado V, Araujo J, Furtado ES, et al. Assessing the user experience of an LLM-based conversational assistant in diabetes mellitus care. J Healthc Inform Res. Mar 2026;10(1):116-153. [doi: 10.1007/s41666-025-00217-5] [Medline: 41658406]
62. Zhang Y, Wang L, Zhang W, Lan W. Decoupled quality and readability in skin cancer education from large language models. Front Public Health. 2026;14:1777577. [doi: 10.3389/fpubh.2026.1777577] [Medline: 41799487]
63. Bertin L, Branchi F, Ciacci C, et al. Efficacy of large language models in providing evidence-based patient education for celiac disease: a comparative analysis. Nutrients. Dec 6, 2025;17(24):3828. [doi: 10.3390/nu17243828] [Medline: 41470773]
64. Carl N, Hetz MJ, Wies C, et al. Enhancing clinicians' trust in large language models via transparent source attribution: a randomized controlled evaluation in uro-oncology. Eur J Cancer. Jan 17, 2026;233:116168. [doi: 10.1016/j.ejca.2025.116168] [Medline: 41401634]
65. Alredaini R, Abulkhair M, Almisbahi H. Interpretable glucose forecasting for type 2 diabetes across traditional, deep, and large language models. Sci Rep. Dec 16, 2025;16(1):2421. [doi: 10.1038/s41598-025-32373-4] [Medline: 41402531]
66. Vladic N, Nopp S, Pabinger I, et al. Large language models vs thrombosis experts: a comparative study on patient education and clinical decision-making in venous thromboembolism. Journal of Thrombosis and Haemostasis. Mar 2026;24(3):943-954. [doi: 10.1016/j.jtha.2025.09.004]
67. Yigit Yalcın B, Mutlu U, Ok AM, et al. Multidimensional assessment of large language model responses to patient questions on gestational diabetes mellitus. Sci Rep. Dec 12, 2025;15(1):43758. [doi: 10.1038/s41598-025-27235-y] [Medline: 41387473]
68. Modi ND, Alex CA, Awaty AA, et al. Cross-sectional evaluation of medical disinformation safeguards in consumer-facing large language model platforms. JMIR Infodemiology. Apr 20, 2026;6:e89831. [doi: 10.2196/89831] [Medline: 42008624]
69. Campbell L, Tiase VL. Bridging the gap between potential and practice: an integrative review of generative artificial intelligence in nursing. Comput Inform Nurs. Jul 1, 2026;44(7):e01481. [doi: 10.1097/CIN.0000000000001481] [Medline: 41662629]

70. NCD Countdown 2030 collaborators. NCD Countdown 2030: worldwide trends in non-communicable disease mortality and progress towards Sustainable Development Goal target 3.4. Lancet. Sep 22, 2018;392(10152):1072-1088. [doi: 10.1016/S0140-6736(18)31992-5] [Medline: 30264707]
71. Freihat O, Sipos D, Aamir M, Kovacs A. Global burden and future projections of non-communicable diseases (2000-2050): progress toward SDG 3.4 and disparities across regions and risk factors. PLoS One. 2025;20(12):e0336036. [doi: 10.1371/journal.pone.0336036] [Medline: 41370213]
72. Unger Z, Soffer S, Efros O, Chan L, Klang E, Nadkarni GN. Clinical applications and limitations of large language models in nephrology: a systematic review. Clin Kidney J. Sep 2025;18(9):sfaf243. [doi: 10.1093/ckj/sfaf243] [Medline: 41018275]
73. Osonuga A, Olawade DB, Gore M, et al. Generative artificial intelligence in predictive analysis of diabetes and its complications: a narrative review. Ann Transl Med. Oct 31, 2025;13(5):59. [doi: 10.21037/atm-25-62] [Medline: 41211117]
74. Nigro M, Behring GE, Aliverti A, et al. Accuracy, comprehensiveness and understandability of AI-generated answers to questions from people with COPD: the AIR-COPD Study. Respir Res. Dec 16, 2025;27(1):19. [doi: 10.1186/s12931-025-03438-9] [Medline: 41402777]
75. Shi H, Liang S, Wang Z, Lv Q, Zhang Q, Li M. Non-targeted analysis of odor components and hazardous volatiles in children's raincoats. Ecotoxicol Environ Saf. May 2025;296:118220. [doi: 10.1016/j.ecoenv.2025.118220] [Medline: 40253877]
76. Pariente B, Varennes O, Burgun A, Azizi M, Amar L, Tsopra R. Empowering patients and clinicians: LLMs in hypertension care, a scoping review. Hypertension. Jul 2026;83(7):e27004. [doi: 10.1161/HYPERTENSIONAHA.126.27004] [Medline: 42021742]
77. Yang X, Xiao Y, Liu D, et al. Factors influencing adoption of large language models in health care: multicenter cross-sectional mixed methods observational study. J Med Internet Res. Dec 11, 2025;27:e84918. [doi: 10.2196/84918] [Medline: 41380031]
78. Yıldız E. Optimisation drift and substitution risk in artificial intelligence-supported personalised mental health nursing: a critical synthesis on therapeutic presence and care biography. J Psychiatr Ment Health Nurs. Aug 2026;33(4):641-647. [doi: 10.1111/jpm.70137] [Medline: 42080623]
79. Moëll B, Sand Aronsson F. Harm reduction strategies for thoughtful use of large language models in the medical domain: perspectives for patients and clinicians. J Med Internet Res. Jul 25, 2025;27:e75849. [doi: 10.2196/75849] [Medline: 40712151]
80. Kang R, Xuan Z, Tong L, Wang Y, Jin S, Xiao Q. Nurse researchers' experiences and perceptions of generative AI: qualitative semistructured interview study. J Med Internet Res. Aug 25, 2025;27:e65523. [doi: 10.2196/65523] [Medline: 40853413]
81. Chen H, Zeng D, Qin Y, et al. Large language models and global health equity: a roadmap for equitable adoption in LMICs. Lancet Reg Health West Pac. Oct 2025;63:101707. [doi: 10.1016/j.lanwpc.2025.101707] [Medline: 41158957]
82. Yu L, Darmstadt GL, Ward V, Wong RJ, Stevenson DK, Maric I. Large language models for maternal and neonatal health care in low- and middle-income countries. J Pediatr. Jun 2026;293:115037. [doi: 10.1016/j.jpeds.2026.115037] [Medline: 41692226]
83. Adedinsewo DA, Onietan D, Morales-Lara AC, et al. Contextual challenges in implementing artificial intelligence for healthcare in low-resource environments: insights from the SPEC-AI Nigeria trial. Front Cardiovasc Med. 2025;12:1516088. [doi: 10.3389/fcvm.2025.1516088] [Medline: 40134980]
84. Strika Z, Petkovic K, Likic R, Batenburg R. Bridging healthcare gaps: a scoping review on the role of artificial intelligence, deep learning, and large language models in alleviating problems in medical deserts. Postgrad Med J. Dec 23, 2024;101(1191):4-16. [doi: 10.1093/postmj/qgae122] [Medline: 39323384]
85. Alu FF, Oluwadare S. An auditable and source-verified framework for clinical AI decision support: integrating retrieval-augmented generation with data provenance. Front Artif Intell. 2026;9:1737532. [doi: 10.3389/frai.2026.1737532] [Medline: 41716615]
86. Aguzzi G, Magnini M, Farahmand A, Ferretti S, Pengo MF, Montagna S. RAG-Enhanced open SLMs for hypertension management chatbots. J Med Syst. Nov 13, 2025;49(1):159. [doi: 10.1007/s10916-025-02297-7] [Medline: 41231304]
87. Masanneck L, Epping PZ, Meuth SG, Pawlitzki M. Evaluating web retrieval-assisted large language models with and without whitelisting for evidence-based neurology: comparative study. J Med Internet Res. Oct 29, 2025;27:e79379. [doi: 10.2196/79379] [Medline: 41159599]
88. Singhal K, Tu T, Gottweis J, et al. Toward expert-level medical question answering with large language models. Nat Med. Mar 2025;31(3):943-950. [doi: 10.1038/s41591-024-03423-7]

89. Wang Z, Li X, Ma C, Zhang Z. Challenges of using generative AI for patient education in chronic heart failure: an evaluation of content quality, readability, and actionability in cross-platform LLM-generated texts. Front Public Health. 2026;14:1801829. [doi: 10.3389/fpubh.2026.1801829]
90. Shafau F, Wahl C. Evaluating the readability of AI-generated patient information on chronic diseases. Chronic Dis Transl Med. Dec 2025;11(4):316-317. [doi: 10.1002/cdt3.70020] [Medline: 41341736]
91. Tarapore R, Gupta S, Means KR Jr, Giladi AM. Artificial intelligence can answer postoperative questions about distal radius fractures—but can patients understand the answers? J Hand Surg Glob Online. Nov 2025;7(6):100822. [doi: 10.1016/j.jhsg.2025.100822] [Medline: 41356624]
92. Fareed M, Fatima M, Uddin J, Ahmed A, Sattar MA. A systematic review of ethical considerations of large language models in healthcare and medicine. Front Digit Health. 2025;7:1653631. [doi: 10.3389/fdgth.2025.1653631] [Medline: 41019285]
93. Pal A, Wangmo T, Bharadia T, et al. Generative AI/LLMs for plain language medical information for patients, caregivers and general public: opportunities, risks and ethics. Patient Prefer Adher. 2025;19:2227-2249. [doi: 10.2147/PPA.S527922] [Medline: 40771655]
94. Zhou S, Liu X, Xu Z, et al. Mitigating ethical issues for large language models in oncology: a systematic review. JCO Clin Cancer Inform. Sep 2025;9(9):e2500076. [doi: 10.1200/CCI-25-00076] [Medline: 40991877]
95. Haltaufderheide J, Ranisch R. The ethics of ChatGPT in medicine and healthcare: a systematic review on large language models (LLMs). NPJ Digit Med. Jul 8, 2024;7(1):183. [doi: 10.1038/s41746-024-01157-x] [Medline: 38977771]
96. Shojaeinia M, Hosseini A, Naderi M, et al. A comprehensive overview: deep learning approaches to central serous chorioretinopathy diagnosis. BMC Ophthalmol. 2025;25(1):549. [doi: 10.1186/s12886-025-04372-6]
97. Topaz M, Peltonen LM, Michalowski M, et al. The overlooked dark side of generative AI in nursing: an international think tank's perspective. J of Nursing Scholarship. Jul 2025;57(4):559-562. [doi: 10.1111/jnu.70016]
98. Workum JD, van de Sande D, Gommers D, van Genderen ME. Bridging the gap: a practical step-by-step approach to warrant safe implementation of large language models in healthcare. Front Artif Intell. 2025;8:1504805. [doi: 10.3389/frai.2025.1504805] [Medline: 39931218]
99. Mohammad M, Jimenez-Solem E, Hejmadi M, Pihl A. Self-regulating the use of large language models in clinical practice: a risk-stratified approach. BMJ Health Care Inform. May 6, 2026;33(1):e101921. [doi: 10.1136/bmjhci-2025-101921] [Medline: 42091170]
100. Li S, Li Y, Zhou S, et al. A community-codesigned LLM-powered chatbot for primary care: a randomized controlled trial. Nat Health. 2026;1(2):238-250. [doi: 10.1038/s44360-025-00021-w] [Medline: 41659358]
101. Chen SF, Alyakin A, Seas A, et al. LLM-assisted systematic review of large language models in clinical medicine. Nat Med. Mar 2026;32(3):1152-1159. [doi: 10.1038/s41591-026-04229-5] [Medline: 41776077]
102. Chen RJ, Wu MS, Tsai LW, Chang SS, Shen Hsiao ST, Lo YS. Integrating a large language model to streamline nursing handover documentation across multiple hospitals in Taiwan: development and implementation study. J Med Internet Res. Mar 12, 2026;28:e81604. [doi: 10.2196/81604] [Medline: 41819121]
103. Michalowski M, Topaz M, Peltonen LM. An AI-enabled nursing future with no documentation burden: a vision for a new reality. J Adv Nurs. Jan 2026;82(1):907-912. [doi: 10.1111/jan.16911] [Medline: 40129115]
104. Tran H, Yao Z, Jang WS, et al. MedReadCtrl: personalizing medical text generation with readability-controlled instruction learning. medRxiv. Preprint posted online on Jul 11, 2025. [doi: 10.1101/2025.07.09.25331239] [Medline: 40672473]
105. Tilton AK, Caplan BE, Cole BJ. Generative AI in consumer health: leveraging large language models for health literacy and clinical safety with a digital health framework. Front Digit Health. 2025;7:1616488. [doi: 10.3389/fdgth.2025.1616488] [Medline: 40933812]
106. Hakim JB, Painter JL, Ramcharran D, et al. The need for guardrails with large language models in pharmacovigilance and other medical safety critical settings. Sci Rep. Jul 31, 2025;15(1):27886. [doi: 10.1038/s41598-025-09138-0] [Medline: 40738919]
107. Tam TYC, Sivarajkumar S, Kapoor S, et al. A framework for human evaluation of large language models in healthcare derived from literature review. NPJ Digit Med. Sep 28, 2024;7(1):258. [doi: 10.1038/s41746-024-01258-7] [Medline: 39333376]

Abbreviations

LLM: large language model
PRISMA: Preferred Reporting Items for Systematic reviews
PRISMA-S: Preferred Reporting Items for Systematic reviews and Meta-Analyses literature search extension
RAG: retrieval-augmented generation

SPIDER: sample, phenomenon of interest, design, evaluation, research
SWiM: Synthesis Without Meta-Analysis

Edited by Stefano Brini; peer-reviewed by James Plaisimond, Wang Guimeng, Xianying Lu; submitted 02.Jan.2026; final revised version received 08.Jul.2026; accepted 09.Jul.2026; published 11.Aug.2026

Please cite as:
Zhang L, Huai P, Xu R, Sun J, Jin R, Lv H
Application of Large Language Models in Chronic Disease Care: Mixed Methods Systematic Review and Thematic Synthesis
J Med Internet Res 2026;28:e90744
URL: https://www.jmir.org/2026/1/e90744
doi: 10.2196/90744