Unveiling the potential of large language models in transforming chronic disease management: A mixed-method systematic review

Caixia Li, Yina Zhao, Yang Bai, Baoquan Zhao, Yetunde Oluwafunmilayo, Carmen W.H. Chan, Meifen Zhang, Xia Fu

Submitted to: Journal of Medical Internet Research
on: December 24, 2024

Table of Contents

Unveiling the potential of large language models in transforming chronic disease management: A mixed-method systematic review

Caixia Li[1] PhD; Yina Zhao[1] MM; Yang Bai[2] PhD; Baoquan Zhao[2] PhD; Yetunde Oluwafunmilayo[3] PhD; Carmen W.H. Chan[4] PhD; Meifen Zhang[2] PhD; Xia Fu[1] MD

[1] Eighth Affiliated Hospital of Sun Yat-sen University Shenzhen CN
[2] Sun Yat-sen University Guangzhou CN
[3] Conestoga College Kitchener CA
[4] Chinese University of Hong Kong Hong Kong HK

Corresponding Author:
Xia Fu MD

Eighth Affiliated Hospital of Sun Yat-sen University
Room 501, Administration Building
The Eighth Affiliated Hospital, Sun Yat-sen University
Shenzhen
CN

Abstract

Background: Accounting for nearly three-quarters of deaths worldwide, chronic diseases are a major global health burden. Large language models (LLMs) are advanced artificial intelligence systems, possessing transformative potential to optimise chronic disease management, yet robust evidence is lacking.

Objective: To synthesise evidence on the feasibility, opportunities, and challenges of LLMs across the disease management spectrum–from prevention to screening, diagnosis, treatment, and long-term care.

Methods: Following the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analysis) guidelines, eleven databases (Cochrane Central Register of Controlled Trials, CINAHL, Embase, IEEE Xplore, Medline via Ovid, ProQuest Health & Medicine Collection, ScienceDirect, Scopus, Web of Science Core Collection, China National Knowledge Internet, and SinoMed) were searched on 17 April 2024. Intervention and simulation studies were included if they examined LLMs in managing chronic diseases. Narrative analysis with descriptive figures were utilised to synthesise study findings. Random-effects meta-analyses were conducted to assess pooled effect estimates for LLM feasibility in chronic disease management.

Results: Twenty studies were eligible examining general-purpose (n = 17) and fine-tuned LLMs (n = 3) in managing chronic diseases, including cancer, cardiovascular diseases, and metabolic disorders. LLMs demonstrated feasibility across the chronic disease management spectrum by generating relevant, comprehensible, and accurate health recommendations (71%; 95% confidence interval [CI] = 0.59, 0.83; I2 = 88.32%) with fine-tuned LLMs having higher accurate rates compared to general-purpose LLMs (odds ratio = 2.89; 95% CI = 1.83, 4.58; I2 = 54.45%). LLMs facilitated equitable information access, increased patient awareness of ailments, preventive measures, and treatment options, and promoted self-management behaviours in lifestyle modification and symptom coping. Additionally, LLMs facilitated compassionate emotional support, social connections, and healthcare resource to improve health outcomes of chronic diseases. However, LLMs faced challenges in addressing privacy, language, and cultural issues, undertaking advanced tasks, including diagnostic, medication, and comorbidities management, and generating personalised regimens with real-time adjustments and multiple modalities.

Conclusions: LLMs demonstrated potential to transform chronic disease management at individual, social, and healthcare levels, yet their direct application in clinical settings is still in its infancy. A multifaced approach–incorporating robust data security, domain-specific model fine-tuning, multimodal data integration, and wearables–is crucial to evolve LLMs into invaluable adjuncts for healthcare professionals to transform chronic disease management. Clinical Trial: PROSPERO (CRD42024545412).

(JMIR Preprints 24/12/2024:70535)
DOI: https://doi.org/10.2196/preprints.70535

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Title

Unveiling the potential of large language models in transforming chronic disease management: A mixed-method systematic review

Authors

Caixia **Li PhD**[1], Yina **Zhao MM**[1], Yang **Bai PhD**[2], Baoquan **Zhao PhD**[3], Yetunde **Oluwafunmilayo PhD**[4], Carmen W.H. **Chan PhD**[5$], Meifen **Zhang PhD**[2$], Xia **Fu MD**[1*$]

Affiliations

[1]The Department of Nursing, The Eighth Affiliated Hospital, Sun Yat-sen University, China
[2]The School of Nursing, Sun Yat-sen University, China
[3]The School of Artificial Intelligence, Sun Yat-sen University, China
[4]The Department of Clinical Research, Conestoga College, Canada
[5]The Nethersole School of Nursing, The Chinese University of Hong Kong, China

Corresponding Author:

*Xia Fu, The Nursing Department, The Eighth Affiliated Hospital, Sun Yat-Sen University, Shenzhen, Guangdong Province, China
Email: fuxia5@mail.sysu.edu.cn
Tel: (86) 13829706026

Other author footnotes

[$]These authors jointly supervised this work and contributed equally.

Additional emails

Caixia Li: licx23@mail.sysu.edu.cn
Yina Zhao: zhaoyn33@mail.sysu.edu.cn
Yang Bai: baiy36@mail.sysu.edu.cn
Baoquan Zhao: zhaobaoquan@mail.sysu.edu.cn
Yetunde Oluwafunmilayo: tolayetunde702@link.cuhk.edu.hk
Carmen W.H. Chan: whchan@cuhk.edu.hk
Meifen Zhang: zhmfen@mail.sysu.edu.cn

Acknowledgements

The authors would like to thank librarians from the Chinese University of Hong Kong and Sun Yat-sen University for refining search strategies of this review.

Funding

This review is supported by the National Natural Science Foundation of China (72404286), the Shenzhen Medical Research Fund (A2403034), the Shenzhen Science and Technology Program (JCYJ20240813150722029), and the Futian Healthcare Research Project (FTWS026). The funders played no role in study design, data collection, analysis and interpretation of the data, or the writing of this manuscript.

CRediT authorship contribution statement

Caixia LI: Conceptualisation, Data curation, Formal analysis, Project administration, Funding acquisition, Writing – original draft, Writing – review & editing. **Yina Zhao**: Data curation, Validation, Writing – review & editing. **Yang Bai**: Conceptualisation, Validation, Writing – review & editing. **Baoquan Zhao**: Conceptualisation, Validation, Writing – review & editing. **Yetunde Oluwafunmilayo**: Conceptualisation, Validation, Writing – review & editing. **Carmen W.H. Chan**: Conceptualisation, Study design, Supervision, Writing – review & editing. **Meifen Zhang**: Conceptualisation, Study design, Supervision, Writing – review & editing. **Xia Fu**: Conceptualisation, Study design, Supervision, Writing – review & editing.

Conflicts of interest

None declared.

Data availability

The search strategies and data extracted for this review are available in the main text and supplementary information files; any additional data are available on reasonable request.

Unveiling the potential of large language models in transforming chronic disease management: A mixed-method systematic review

Abstract

Background: Accounting for nearly three-quarters of deaths worldwide, chronic diseases are a major global health burden. Large language models (LLMs) are advanced artificial intelligence systems, possessing transformative potential to optimise chronic disease management, yet robust evidence is lacking.

Objective: To synthesise evidence on the feasibility, opportunities, and challenges of LLMs across the disease management spectrum–from prevention to screening, diagnosis, treatment, and long-term care.

Methods: Following the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analysis) guidelines, eleven databases (Cochrane Central Register of Controlled Trials, CINAHL, Embase, IEEE Xplore, Medline via Ovid, ProQuest Health & Medicine Collection, ScienceDirect, Scopus, Web of Science Core Collection, China National Knowledge Internet, and SinoMed) were searched on 17 April 2024. Intervention and simulation studies were included if they examined LLMs in managing chronic diseases. Narrative analysis with descriptive figures were utilised to synthesise study findings. Random-effects meta-analyses were conducted to assess pooled effect estimates for LLM feasibility in chronic disease management.

Results: Twenty studies were eligible examining general-purpose (n = 17) and fine-tuned LLMs (n = 3) in managing chronic diseases, including cancer, cardiovascular diseases, and metabolic disorders. LLMs demonstrated feasibility across the chronic disease management spectrum by generating relevant, comprehensible, and accurate health recommendations (71%; 95% confidence interval [CI] = 0.59, 0.83; I^2 = 88.32%) with fine-tuned LLMs having higher accurate rates compared to general-purpose LLMs (odds ratio = 2.89; 95% CI = 1.83, 4.58; I^2 = 54.45%). LLMs facilitated equitable information access, increased patient awareness of ailments, preventive measures, and treatment options, and promoted self-management behaviours in lifestyle modification and symptom coping. Additionally, LLMs facilitated compassionate emotional support, social connections, and healthcare resource to improve health outcomes of chronic diseases. However, LLMs faced challenges in addressing privacy, language, and cultural issues, undertaking advanced tasks, including diagnostic, medication, and comorbidities management, and generating personalised regimens with real-time adjustments and multiple modalities.

Conclusions: LLMs demonstrated potential to transform chronic disease management at individual, social, and healthcare levels, yet their direct application in clinical settings is still in its infancy. A multifaced approach–incorporating robust data security, domain-specific model fine-tuning, multimodal data integration, and wearables–is crucial to evolve LLMs into invaluable adjuncts for healthcare professionals to transform chronic disease management.

Trial Registration: PROSPERO (CRD42024545412).

Keywords: Artificial intelligence; Chronic disease; Health management; Large language model; Systematic review

Introduction

Accounting for nearly three-quarters of deaths worldwide, chronic diseases have become a major challenge to global health [1]. These diseases, primarily cardiovascular diseases, cancers, diabetes and chronic respiratory diseases, are responsible for 41 million deaths each year globally, 41.5% of which are premature deaths in people younger than 70 years [1]. Approximately 37.2% of adults worldwide suffer from multiple chronic diseases and experience increased symptom burdens, emergency medical admissions, and healthcare expenditures [2]. The health burden of chronic diseases is further exacerbated by population ageing, urbanisation, and unhealthy lifestyles, including a lack of physical activity [3]. Projections indicate that globally, chronic diseases will cause 77.6% of disability-adjusted life years in 2050 [4]. To address this challenge, the WHO 2030 agenda adopted a global target to reduce premature mortality from chronic diseases by one-third by 2030 [5], highlighting efforts in the prevention, detection, treatment, and long-term management of chronic diseases.

Currently, healthcare systems for chronic disease management face multidimensional challenges. These systems must process and integrate large volumes of patient data, including health records, genomic data, and real-time data (e.g., glucose levels) [6, 7]. Failure to process such data may lead to fragmented information, impeding the potential for tailored treatment and holistic management of chronic diseases and ultimately compromising patient care. Successful chronic disease management also requires day-to-day persistence, with approximately 50% of patients failing to consistently follow prescribed treatment regimens, medications, diets, and physical activities, leading to disease progression and long-term complications [8, 9]. Limited access to specialised healthcare services for chronic diseases presents another challenge, especially in low-resource healthcare settings [10]. Approximately 43.3% of people worldwide cannot reach healthcare facilities by foot within an hour, and those living in rural or remote regions often face increased travel time, costs, and difficulties in accessing healthcare [11]. This disparity might result in inadequate health promotion, delayed diagnosis, and disrupted treatment of chronic diseases [12]. These challenges collectively contribute to suboptimal health outcomes, reinforcing the need for novel approaches to enhance chronic disease management.

Large language models (LLMs), such as ChatGPT, have emerged as promising solutions to address the complexities associated with chronic disease management. Leveraging advanced natural language processing and machine learning algorithms, LLMs are trained on extensive datasets with billions of parameters, which are particularly advantageous for analysing and synthesising multifaceted health data to assist in developing integrated management plans for chronic diseases [13, 14]. Additionally, LLMs perform well in answering medical questions and can provide adaptive communication to patient queries for various chronic diseases, including head and neck cancer [15], gastroesophageal reflux disease [16], and cardiovascular diseases [17]. This could inform patients with personalised health management suggestions, fostering patient engagement and adherence to chronic disease management [18]. More importantly, LLMs can be integrated with existing health applications and systems via their application programming interfaces to enhance telemedicine [19]. This could enable them to monitor patients' chronic health conditions, provide diagnosis and treatment information, and aid in follow-up care [20, 21], bridging the gap in healthcare access, especially in low-resource healthcare settings [21].

LLMs may introduce transformative potential to enhance healthcare practice across the chronic disease management spectrum. However, several challenges have impeded their optimal integration into this domain [22, 23]. Notably, hallucination—scenarios in which LLMs generate inaccurate or misleading information—can lead to incorrect diagnoses and inappropriate treatment recommendations [24, 25]. Despite this, the transformative force of LLMs necessitates an in-depth understanding of how they can be effectively integrated into current healthcare systems to enhance chronic disease management. Given the lack of robust evidence in this area, this review was conducted to consolidate current research findings and provide a comprehensive understanding of the feasibility, opportunities, and challenges associated with the application of LLMs in chronic disease management. These insights can inform future research and practice, guiding the strategic employment of LLMs in chronic disease management to ultimately alleviate the global burden of chronic diseases.

Methods

Study Design

A mixed-method systematic review was conducted following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses statement. The protocol was registered with PROSPERO (CRD42024545412) on 21 May 2024.

Inclusion and Exclusion Criteria

This review sought evidence on the potential of LLMs to transform chronic disease management to inform future practice. The detailed inclusion criteria was formulated following the "Participant-Intervention-Comparator-Outcomes-Study design" (PICOS) framework [26]:

Population. Studies conducted among patients with chronic diseases or individuals at high risk of developing chronic diseases, such as obese individuals, were eligible for inclusion. In contrast, studies that focused on health conditions other than chronic diseases, such as plastic surgery and acute appendicitis, were excluded. Chronic diseases were defined as long-lasting conditions that primarily included cardiovascular diseases, cancers, chronic respiratory diseases, and diabetes [1]. Given that LLMs have not been widely used in clinical settings, studies employing simulated patient profiles and scenarios to examine LLMs in chronic disease management were also considered eligible.

Intervention. Studies were included if they examined LLMs in managing chronic diseases from prevention to screening, diagnosis, treatment, or follow-up care. LLMs are defined as deep learning models trained on large datasets to comprehend and generate human language text content, which include, but are not limited to, ChatGPT, Bard, BERT, and Llama [27]. Those focusing on general artificial intelligence, algorithm-based chatbots, or expert systems were excluded.

Comparator. There were no restrictions on the comparators, including standard comparisons or no comparators.

Outcomes. The study outcomes pertain to the feasibility (e.g., accuracy and relevance of responses), opportunities, and challenges (e.g., privacy issues) of LLMs in managing chronic diseases. This may include the potential benefits of LLMs in enhancing health knowledge, attitudes, and self-care behaviours in chronic disease management.

Study design. Interventions, simulations, and case studies that tested LLMs were eligible, including proof-of-concept, feasibility, and experimental studies. Conference abstracts, commentaries, editorials, and review articles were excluded.

Search Strategy

Eleven databases, including Cochrane Central Register of Controlled Trials, CINAHL, Embase, IEEE Xplore, Medline via Ovid, ProQuest Health & Medicine Collection, ScienceDirect, Scopus, Web of Science Core Collection, China National Knowledge Internet, and SinoMed, were searched on 17 April 2024. By conducting an initial search in Medline, thirty-seven search terms were developed about LLMs and outcomes of interest, including ‘large language model’, ‘generative pretrained transformer’, ‘ChatGPT’, and ‘self-care’. The title, abstract, and subject-heading fields were searched to identify studies. Truncations, adjacency searches, and Boolean operators were applied to ensure the comprehensive retrieval of relevant literature. Two medical librarians refined the search strategy by reviewing detailed search records in Medline. A manual search of the included studies was performed to identify additional relevant studies. There were no restrictions on publication language, date, or type. Supplementary material Table 1 provides the complete search strategy for each database.

Study Screening

The search results were exported to Covidence (Veritas Health Innovation, Melbourne, Australia) to remove duplicates. Two authors screened the titles and abstracts, followed by a full-text review. Exclusion decisions made during full-text screening were also documented. Discrepancies were resolved through discussions with a third reviewer.

Data Extraction

Data were extracted using a pilot-tested and standardised data extraction form. The form encompasses the study authors, country of origin, study design, chronic health conditions, characteristics of LLMs (including the name of LLMs, scenarios, prompts, and process of LLMs in generating responses), and the primary outcomes of the applicability, opportunities, and challenges of LLMs in chronic disease management. One reviewer independently extracted the data, which were proofed by a second author and agreed upon by all the authors.

Data Synthesis

A narrative synthesis was conducted in which the characteristics, feasibility, opportunities, and challenges of LLMs in chronic disease management were described as reported in the included studies. A descriptive figure was used to visualise the results. To synthesize feasibility outcomes of LLMs, meta-analyses were conducted using the STATA 18.0 (StataCorp LLC, College Station, the United States of America). Pooled accurate rates, odds ratios (ORs) comparing accurate rates of fine-tuned LLMs and general-purpose LLMs, and effect size for readability scores with 95% confidence intervals (CIs) were calculated. The statistical significance level was set at $P < 0.05$. Heterogeneity was assessed using I^2 statistics (I^2 of 25, 50, and 75% indicating low, moderate, and high heterogeneity, respectively) and Q statistics ($P < 0.10$ indicating statistically significant heterogeneity) [28]. Due to high heterogeneity among included studies, the random-effects Dersimonian-Laird model was applied in the analyses. Sensitivity analysis was conducted using the leave-one-out approach to evaluate robustness of the pooled analyses.

Methodology Quality Assessment

A rating rubric was used to assess the quality of the methodology used in simulation-based articles. The rubric contained 16 items: study design, sample size, simulation development and implementation, and study instruments [29]. Each rubric item was graded on a scale of 0–4, and the total scores were transferred into percentage scores by averaging the total number of appraisal questions eligible for the study [29]. In quasi-experimental studies, the Risk of Bias in Non-Randomized Studies of Interventions (ROBINS-I) tool was used to assess the risk of bias due to confounding, participant selection, classification of interventions, deviations from intended interventions, missing data, outcome measurements, and selection of the reported outcome [30]. Each domain and overall methodology were rated as having low, moderate, serious, or critical risk of bias. Two reviewers independently appraised the study quality, and disagreements were resolved by discussion.

Results

Study Selection

The database search yielded 8,391 records. After removing 1,017 duplicates, 7,374 titles and abstracts were screened. Among the 180 full-text articles retrieved, 163 were excluded, primarily due to their irrelevance to LLMs or chronic disease management, leaving 17 eligible studies. An additional 607 records were identified by manually searching the reference lists of the 17 eligible studies, resulting in 20 studies included in this review (Figure 1).

Characteristics of the Included Studies

Published between 2023 and 2024, the included studies mainly originated from developed countries (12 [60%] of 20), including the United States of America (n = 6) [31-36], Australia (n = 2) [37, 38], Canada (n = 2) [39, 40], Singapore (n = 1) [41], and South Korea (n = 1) [42]. Most studies (15 [75%] of 20) used simulation-driven or proof-of-concept designs that did not involve human subjects (Table 1). Only three quasi-experimental studies involved real-world clinical implementation of LLMs among patients, and their sample sizes ranged from 24 to 72 [18, 43, 44]. The included studies utilised LLMs, such as ChatGPT, DocsGPT, Google Bard, and Bing Chat, to manage a spectrum of chronic illnesses, including cancer, cardiovascular diseases, metabolic disorders, respiratory diseases, musculoskeletal disorders, mental health disorders, and substance use disorders. The findings of this review are shown in Table 1 and visualised in Figure 2.

Table 1 Characteristics of the included studies (n = 20)

Author, year, and country	Study design	Chronic health conditions	Characteristics of LLMs	Outcome assessment	Study outcomes
AI-Anezi [43], Saudi Arabia	Quasi-experimental study	Cancer, diabetes, and kidney failure	a) LLM(s): ChatGPT 3.5. b) Scenarios and prompts: Participant inquiries. c) Process: Participants chatted with ChatGPT as a virtual coach for chronic disease management. Participants used ChatGPT for at least 15 min daily at home, lasting 2 weeks.	a) Methods: semistructured interviews. b) Assessors: N.S.	<u>**Opportunities**</u> a) Enhancing awareness of updated information on chronic disease management (e.g., diet and physical activities). b) Reducing reliance on healthcare specialists and offering scalable support. c) Offering free access that reduced disparities in accessing health information in chronic diseases, especially among those from rural areas. d) Motivating health goals and promoting health behaviours regarding diets, meditation, sleep, and exercise. e) Adapting to their preferences and needs during communication. f) Linking to online communities to facilitate social support. <u>**Challenges**</u> a) Lack of physical examinations in monitoring comorbidities and potential for inaccurate diagnoses. b) Lack of empathy for patients. c) Ineffective in analysing complex conditions regarding chronic diseases. d) Concerns about privacy and security for personal health data.
Alanezi et al. [44], Saudi Arabia	Quasi-experimental study	Chronic mental health conditions, including anxiety, depression, and behaviour disorders	a) LLM(s): ChatGPT 3.5. b) Scenarios and prompts: participant inquiries. c) Process: Participants chatted with ChatGPT to seek support for managing their mental health issues for at least 15 min per day at home, lasting 2 weeks.	a) Methods: Semistructured interviews. b) Assessors: N.S.	<u>**Opportunities**</u> a) Improving mental health literacy and enhancing mental health symptoms management through psychoeducation. b) Provide a non-judgemental channel to express their concerns and offer compassionate responses. c) Setting achievable health goals and developing plans to improve mental health. d) Recommending mental health resources and suggesting ways to seek medical support. e) Facilitating self-assessment and self-care practices regarding mental health symptoms. f) Providing cognitive behaviour therapy techniques and facilitating psychotherapeutic exercises to manage negative thoughts. g) Providing crisis intervention support for those experiencing acute distress. <u>**Challenges**</u> a) Potential ethical and legal concerns in data privacy, confidentiality, and biases. b) Accuracy and reliability issues. c) Lacking capabilities in assessing mental health conditions. d) Not fully equipped to understand and address cultural and linguistic diversities.
Alanezi. [18], Saudi Arabia	Quasi-experimental study	Cancer	a) LLM(s): ChatGPT 3.5. b) Scenarios and prompts: Participant inquiries. c) Process: Participants chatted with ChatGPT	a) Methods: Focus group interviews. b) Assessors: A researcher.	<u>**Opportunities**</u> a) Improving health knowledge about cancer, its treatment and management of side effects without language barriers. b) Facilitating self-management behaviours by motivating, reminding,

			to seek information about cancer and its prevention and management measures, lasting 2 weeks.		monitoring, and tracking. c) Connecting with social resources and support for cancer management. d) Enhancing emotional and peer support. **Challenges** a) Privacy and reliability concerns. b) Lack of personalisation.
Aliyeva et al. [31], USA	Simulation study	Severe hearing loss	a) LLM(s): ChatGPT 4.0. b) Scenarios and prompts: Frequently asked postoperative questions (n = 5) by patients and their families gathered by experienced otolaryngologists over three months. c) Process: Five postoperative questions were posed to ChatGPT 4.0, focusing on symptoms after cochlear implant surgery, caring for the implant site, hearing, activities, and management after implantation.	a) Methods: Survey to evaluate accuracy, response time, clarity, understandability, and relevance of the responses. b) Assessors: Five specialists in otolaryngology.	**Feasibility** a) Accuracy: Aligned with guidelines for cochlear implant postoperative care, ChatGPT 4.0 responses were 100% accurate. b) Response time: Questions were answered within seconds, suggesting rapid response efficiency. c) Clarity and understandability: The average clarity and understandability score reached 98%. d) Relevance: The relevance of responses averaged 92%.
Choo et al. [42], South Korea	A simulation study	Stage IV, recurrent, synchronous colorectal cancers	a) LLM(s): ChatGPT™. b) Scenarios and prompts: Stage IV and recurrent colorectal cancer cases discussed in the multidisciplinary tumour board at a tertiary institution, followed by inquiry treatment options. c) Process: Patient information, including demographics and clinical history, was entered into the ChatGPT™ to generate treatment recommendations.	a) Methods: Concordance rates between ChatGPT™ and the treatment recommendations made by a multidisciplinary tumour board. b) Assessors: N.S.	**Feasibility** a) Concordance rates: Approximately 73.3% of the cases were concordant with the first treatment recommendation by ChatGPT™. The oncological management recommendation concordance rate between ChatGPT™ and the multidisciplinary team was 86.7%.
Dergaa et al. [45], Qatar	A simulation study	Mental health	a) LLM(s): ChatGPT. b) Scenarios and prompts: Three imaginary patient scenarios represented different mental health problems. Prompts focused on mental health and sleep management. c) Process: ChatGPT was engaged as a virtual psychiatric provider to interact with imaginary patients following a structured format, including providing an overview of the patient's mental health problems and detailed treatment recommendations.	a) Methods: Qualitative assessment. b) Assessors: academic healthcare professors.	**Opportunities** a) ChatGPT provides quick responses and simulates empathy. b) ChatGPT recommended non-pharmacological interventions as a first-line option for a simple patient scenario (a male college student), aligning with current guidelines and clinical standards of care. **Challenges** a) ChatGPT cannot interact with users and ask for further clarification. b) Could not recommend customised assessment and treatment plans for complex patient scenarios (e.g., a female with systemic lupus erythematosus).
Dergaa et al. [46], Qatar	A simulation study	Arterial hypertension, osteoarthritis, anxiety and stress-related issues, diabetes, asthma	a) LLM(s): ChatGPT 4.0. b) Scenarios and prompts: Five hypothetical patient profiles with different health conditions focusing on cardiovascular health, musculoskeletal strength, mental	a) Methods: qualitative assessment. b) Assessors: a panel of academic professors and doctors in	**Feasibility** a) ChatGPT 4.0 could create general safety-conscious exercise programmes, which adhered to FITT (Frequency, intensity, time, and type) principles, RPE (rate of perceived exertion) guidelines, and research evidence. **Challenges**

		and reduced pulmonary function	health, diabetes, and respiratory health management, respectively. Prompts focused on generating a 30-day fitness program table based on the scenarios. c) Process: Five scenarios with patient profiles (e.g., sex, age, height, basal metabolic rate, and medication intake) were created to test the ability of ChatGPT 4.0 to prescribe a 30-day fitness program.	exercise science or medicine.	a) Prescribed plans did not vary. ChatGPT 4.0 prioritised excessive safety over effectiveness of training and might fall short of providing necessary stimuli for health improvement. b) ChatGPT 4.0 lacked preliminary patient assessment and was unable to monitor physiological response and adjust personalised physical exercise regimens in real-time.
Franco D'Souza et al. [47], India	A simulation study	Psychiatric disorders	a) LLM(s): ChatGPT 3.5. b) Scenarios and prompts: 100 clinical case vignettes from a book representing different psychiatric illnesses, which were used to converse with ChatGPT 3.5 to generate responses.	a) Methods: Grading. b) Assessors: two experts with clinical experience in psychiatry.	**Feasibility** a) By assessing coverage of a standard answer, the responses of ChatGPT 3.5 in 61, 31, and 8 of 100 cases received 'Grade A', 'Grade B', and 'Grade C' ratings, respectively. b) ChatGPT performed well in generating management strategies followed by diagnosis for psychiatric conditions.
Kianian et al. [32], USA	A simulation study	Glaucoma	a) LLM(s): ChatGPT. b) Scenarios and prompts: Seven prompts were given to ChatGPT to generate patient handouts regarding surgery management of glaucoma with simple language and to include references.	a) Methods: readability was assessed using FKRE, FKGL, GFI, and SMOG. Quality was assessed using DISCERN instrument. b) Assessors: two authors.	**Feasibility** a) Readability: The health information generated by ChatGPT was easily readable compared to webpages (9th-grade reading level vs. 11th-grade reading level). Readability did not differ significantly. b) Quality: ChatGPT scored the quality of health resources with high precision (r = 0.725, *P* < 0.001).
Lim et al. [41] Singapore	A simulation study	Colorectal cancer	a) LLM(s): Contextualised ChatGPT 4.0. The ChatGPT 4.0 model was contextualised based on guidelines for colorectal cancer screening and surveillance, which were processed into a knowledge base. The process included splitting each article into textual chunks and embedding each chunk based on similarities between words to allow fast searching by the model. b) Scenarios and prompts: 62 hypothetical patient scenarios with or without risks of colorectal cancer. c) Process: The contextualised ChatGPT 4.0 model was instructed to recommend colonoscopy screening intervals in each simulated patient scenario.	a) Methods: scoring approach (accuracy) and content comparison (hallucination). b) Assessors: three gastroenterology fellows under the supervision of two senior gastroenterologists.	**Feasibility** a) Accuracy: Compared with the standard ChatGPT 4.0 model, the contextualised model performed better in recommending correct screening intervals overall (79% vs. 50.5%, *P* < 0.01) and in each patient risk category, including screening (87.7% vs. 52.6%, p <0.01), surveillance (63.2% vs. 40.2%, *P* < 0.01), post-cancer surveillance (100% vs. 67.7%), and others (100% vs. 70.8%). **Challenges** a) The contextualised ChatGPT 4·0 failed to identify a high-risk feature in one response and experienced hallucination in two responses. By contrast, the standard ChatGPT 4.0 failed to identify high-risk features in 12 responses and hallucinated a high-risk feature in 13 responses.
Mondal et al. [48], India	A simulation study	A set of 20 lifestyle-related chronic diseases, including obesity, diabetes, cardiovascular	a) LLM(s): ChatGPT 3.5. b) Scenarios and prompts: 20 cases mimicking individuals seeking information related to lifestyle-related diseases. c) Process: 20 cases were formulated into four questions and presented to ChatGPT	a) Methods: Readability was assessed using the FKRE and FKGL. Qualitative ratings	**Feasibility** a) Readability: The mean FKRE score was 27.8 (5.74), suggesting the generated text was understandable by college students. b) Accuracy and applicability: The average accuracy and applicability scores of the ChatGPT responses were 1.83 (SD 0.37) and 1.9 (0.21), respectively, significantly higher than the hypothesized median score of

		health, and mental health	for generating relevant answers.	were used to assess the accuracy and applicability of responses. b) Assessors: two academic primary care physicians.	1.5.
Papastratis et al. [49], Greece	A simulation study	Noncommunicable diseases (e.g., obesity, diabetes, and cardiovascular diseases)	a) LLM(s): ChatGPT 3.5 and ChatGPT 4. b) Scenarios and prompts: 15 user profiles (e.g., physical characteristics and medical conditions) for patients with noncommunicable diseases, including obese adults and adults with cardiovascular diseases. c) Process: To generate weekly meal plans per day for patients with non-communicable diseases by interacting with their profiles, including body mass index, basal metabolic rate, and types of non-communicable diseases. Information including meal types (including breakfast, morning snack, lunch, afternoon snack, dinner, and supper), calories, total protein, total fat, and total number of vegetables were outputted.	a) Methods: Accuracy and variability of the meal plans were assessed based on guidelines and unique numbers of meals. b) Assessors: N.S.	**Feasibility** a) Accuracy: Compared with a knowledge-based recommender (91%), ChatGPT 3.5 (81.53%) and ChatGPT 4.0 (81.62%) had lower nutrient accuracy rate for users with non-communicable diseases overall. By inputting personalised target energy intake, the nutrient accuracy rate was improved to 86% in ChatGPT 4.0; and the average caloric difference was 17% and 3% in ChatGPT 3.5 and ChatGPT 4.0, suggesting that the recommended energy intake was comparable to the user's suggested energy intake. b) Variability: The average meal variety was highest in ChatGPT 3.4 (6.58), followed by ChatGPT 4.0 (6.40), and the knowledge-based recommender (4.89).
Pradhan et al. [33], USA	A simulation study	Liver cirrhosis	a) LLM(s): ChatGPT 4.0, DocsGPT, Google Bard, and Bing Chat. b) Scenarios and prompts: direct inquiries for generating patient educational materials on cirrhosis. c) Process: 1-page patient education sheet instructing patients about cirrhosis.	a) Methods: Readability and grade level were assessed using FKRE, SMOG, and FKGL. b) Assessors: 14 patients/caregivers and 8 transplant hepatologists.	**Feasibility** a) Readability: Compared with human-derived health educational materials, LLM-generated materials had higher FKRE scores in, indicating they were easier to comprehend. The FKGLs ranged from 5.5 to 7.9 in LLM-generated materials, indicating they were comprehendible by people with an eighth-grade educational level. Based on the SMOG, LLM-generated materials were expected to be comprehensible by people with high school educational levels or above (ranging from 9.4 to 12.3), except that in Google Bard-derived educational material. b) Actionability: No significant difference in the scores for actionability between the human-derived and LLM-derived educational materials ($p > 0.05$). Only human-derived health educational material met the actionable score cut-off of ≥70%. c) Accuracy: The health educational materials generated by humans and LLMs (Bing Chat, ChatGPT 4.0, and Google Bard) were considered to contain 76% to 99% accurate information, as assessed by hepatologists.
Puerto Nino et al. [39], Canada	A simulation study	Benign prostate enlargement	a) LLM(s): ChatGPT 4.0+. b) Scenarios and prompts: 88 benign prostate enlargement-centric queries were formulated based on the patient	a) Methods: Performance metrics and general quality score using a 5-point	**Feasibility** a) Performance: Precision score ranged from 0.50 to 1, with an overall performance score of 0.66. Recall score ranged from 0.9 to 1.0, with an overall performance score of 0.97. b) General quality score: A median general quality score of 4 was obtained.

			frequently asked questions form on the European Association of Urology and American Urological Association websites. c) Process: 88 queries related to symptoms, diagnoses, complications, and treatment options of benign prostate enlargement were fed to ChatGPT independently to generate responses.	Likert scale. b) Assessors: Two examiners.	
Seth et al. [37] Australia	A simulation study	Carpal tunnel syndrome	a) LLM(s): ChatGPT (no version number). b) Scenarios and prompts: six questions regarding the diagnosis and management of carpal tunnel syndrome. c) Process: six inquiries with common clinical scenarios of carpal tunnel syndrome were inputted into ChatGPT to generate management strategies.	a) Methods: Efficacy and performance of ChatGPT were assessed using a 5-point Likert scale. b) Assessors: Two experienced plastic surgeons.	**Feasibility** a) Accuracy: ChatGPT recommended concise treatment options and considered personal factors in recommending treatment choices for carpal tunnel syndrome._ChatGPT accurately evaluated the efficacy of surgical and nonsurgical treatments for carpal tunnel syndrome in short- and long-term outcomes. ChatGPT accurately identified the diagnosis and recommended further investigation (e.g., physical examination and nerve conduction) and management strategies in an easy-to-understand manner. ChatGPT correctly identified the deterioration of the symptoms and recommended seeking appropriate medical attention. **Challenges** a) There were erroneous references and only three of the five references could be found in the literature. b) Although it retrieved level I evidence, the depth and detail of the information (e.g., statistics of recurrence rates) were not sufficient. c) ChatGPT cited two seminal and three nonexistent ‘recent’ studies to support its answer.
Singer et al. [34] USA	A simulation study	Eye care	a) LLM(s): *Aeyeconsult* powered by ChatGPT 4·0. b) Scenarios and prompts: 260 questions from OphthoQuestions.com. c) Process: Based on textbook source material, *Aeyeconsult* was developed by integrating Lang Chian to extract and split texts of source materials and user queries into chunks with unique values and stored into a vector store with Pinecone. Via comparison with chunks of user queries, the 10 most similar chunks of texts were identified to serve as context to generate a natural language response. Only the first response was recorded as an answer.	a) Methods: The correct rates of the answers given by ChatGPT 4·0. b) Assessors: N.S.	**Feasibility** a) Rates of correct responses: *Aeyeconsult* performed more accurately than ChatGPT 4.0 (83.4% vs. 69.2%, p = 0.0118) regarding the 260 questions from OphthoQuestions. ChatGPT 4.0 performed worst in Retina and Vitreous (37.5%). *Aeyeconsult* had the lowest accuracy rate (68.1%) in clinical optics than other categories (e.g., cornea, fundamentals, general medicine, glaucoma, and paediatrics), whereas it still outperformed ChatGPT 4.0 (45.5%) in this category. b) No answers: *Aeyeconsult* had fewer no answers than ChatGPT 4.0 (5 vs. 18). c) Multiple answers: *Aeyeconsul*t had fewer multiple answers than ChatGPT 4.0 (0 vs. 7). d) Consistency of answers: *Aeyeconsult* had complete consistency for questions initially answered correctly. In contrast, ChatGPT gave different responses to 3 of the 13 questions initially answered correctly. For questions initially answered incorrectly (n = 13), *Aeyeconsult* provided different answers for 8 of the questions (61.5%) and ChatGPT 4.0 provided different responses for all of the 13 questions (100%) over 10 attempts.

Spallek et al. [38], Australia	A simulation study	Mental health and substance use disorders	a) LLM(s): ChatGPT 4.0 pro. b) Scenarios and prompts: direct user queries and four factsheets related to mental health and substance use were chosen from educational portals. c) Process: queries from Positive Choices and Cracks in the Ice were used to prompt ChatGPT 4.0 to capture real-world communication.	a) Methods: readability (using Sydney Health Literacy Lab), quality, and following guidelines were assessed. b) Assessors: Study authors.	**Feasibility** a) Readability: The ChatGPT 4·0 outputs generated a desirable low text complexity rating, ranging from 24% to 33%, indicating fewer uncommon words, medical jargon, and acronyms. The reading levels were higher in ChatGPT 4·0-generated materials using direct user queries (grade 13.9, SD 1.52) and two types of engineered prompts (grade 13.1, SD 3.20; grade 12.9, SD 1.20) than that in expert-developed factsheets (grade 12.2, SD 1.44). b) Adherence to communication guidelines: ChatGPT 4.0 responses to direct user queries and simple prompts had lower average adherence to communication guidelines, with 23% (5/22) of the outputs having at least one stigmatising phrase. ChatGPT 4.0 outputs responding to engineered prompts were more likely to have a cautionary tone and disclaimers than direct queries and simple prompts. c) Quality of advice: ChatGPT 4.0 outputs featured a high level of accuracy without hallucinations and could tailor target audiences, whereas it lacks breadth and depth of expertise compared to human experts.
Willms & Liu [40], Canada	An autoethnographic case study	Disease prevention by increasing physical activity	a) LLM(s): ChatGPT 3.0 combined with Pathverse, a no-code app. b) Scenarios and prompts: prompts were developed based on topics including parental support, active attitudes, and self-monitoring of physical activity. The prompts also considered instruction, context information, and output indicators (e.g., word count). Generated contents were added to Pathverse. c) Process: Based on the multi-process action control framework, ChatGPT 3.0 was used to generate just-in-time adaptive interventions addressing the intention-behaviour gap to support parents in helping their children (8–12 years old) be physically active.	a) Methods: Acceptability, relevance, and tone of responses were assessed based on field notes and discussions. Future recommendations were also suggested. b) Assessors: two researchers.	**Feasibility** a) Accuracy and relevance: ChatGPT had acceptable accuracy and relevance in responding to prompts, whereas it might provide false academic references. b) Tone of responses: acceptable for research purposes and matched the prompts given (in a fun and positive voice).
Yang et al. (2024), USA	A case study	Diet management for preventing chronic illnesses	a) LLM(s): *ChatDiet* based on ChatGPT 3.5 Turbo. b) Scenarios and prompts: users' inquiries and individual-specific information, including personal food preferences, dietary history, health records, and physiological signals gathered from wearable devices. Synthetic participants (n = 100). c) Process: *ChatDiet* included Orchestrator to interact with personal (e.g., causal discovery) and population models (e.g.,	a) Methods: Quantitative causal graphs and qualitative analyses of the outputs were implemented. b) Assessors: researchers.	**Feasibility** **a)** Effectiveness: The recommendation effectiveness ratio of personalised food recommendations ranged from 85% to 95% for heart rate variability, sleep quality, and duration. **b)** Personalisation: *ChatDiets* can adapt to personal needs by considering an individual's unique nutritional nuances based on personal nutrition effects. **c)** Interactivity: *ChatDiets* demonstrated interactivity within the recommendation process by initiating follow-up questions and offering alternative options when users lacked interest in the suggested foods. **Challenges** a) Food suggestions were confined to the factors in the dataset.

			food nutrition list loading) to extract information related to diet based on users' inquiries. The aggregated information was sent to ChatGPT 3.5 Turbo to integrate with its internal knowledge and realise its interactions with the users. The tasks involved retrieving (filtering and retrieving relevant information based on users' inquiries from the individual and population models), transcribing (converting data into textual information), and promoting engineering to instruct ChatGPT 3.5 to provide food recommendations adhering to inputs.		Nonsensical recommendations and hallucinations occurred in some cases.
Yeo et al. [36], USA	A simulation study	Liver cirrhosis and hepatocellular carcinoma	a) LLM(s): ChatGPT Dec 15 version. b) Scenarios and prompts: 73 questions and 91 questions were selected for hepatocellular carcinoma and cirrhosis posted by well-recognised professional societies and institutions and posts by patient support groups on Facebook. c) Process: 164 questions were entered into ChatGPT twice, and both responses were recorded and assessed.	a) Methods: Grading and qualitative assessment. b) Assessors: two transplant hepatologist reviewers.	**Feasibility** a) Accuracy: ChatGPT had 79.1% and 74.0% accurate rate in knowledge of cirrhosis and hepatocellular carcinoma. 76.9% accuracy rate (20/26 quality measures) in the knowledge regarding cirrhosis management. 50% accuracy rate in knowledge of hepatocellular carcinoma screening, compared to 28.8% to 45.4% responded by physicians. **Opportunities** **a)** Acknowledged potential emotional responses and provided actional suggestions and motivational responses to treating and managing hepatocellular carcinoma. **b)** Psychological and practical emotional support given to patient caregivers. **Challenges** a) Failed to identify correct cut-offs for and window time for specific conditions (e.g., liver transplantation). b) Failed to identify age cut-off, screening tests, and surveillance eligibility of hepatocellular carcinoma.

Abbreviations: FKGL, Flesch–Kincaid Grade Level; FKRE, Flesch-Kincaid reading ease score; GFI, Gunning Fog index; LLM, large language model; N.S., not specified; SD, standard deviation; SMOG, simple measure of gobbledygook.

Characteristics of LLMs

Notably, general-purpose LLMs, such as ChatGPT, were applied in most of the included studies (17 [85%] of 20) (Table 1), although their direct deployment often lacked the required specificity for chronic diseases. Three studies fine-tuned LLMs through retrieval-augmented generation, which combines LLMs with access to an external knowledge base through a retrieval mechanism [34, 35, 41]. For example, Lim et al. [41] employed the Open AI API to fine-tune ChatGPT 4.0 by embedding colorectal cancer screening guidelines, decomposing them into manageable textual chunks, and leveraging semantic encoding to facilitate accurate retrieval by the LLMs to generate context-aware recommendations. Similarly, to develop *Aeyeconsult* (an ophthalmology chatbot), Singer et al. [34] fine-tuned ChatGPT 4, which applied LangChina and Pinecone to process ophthalmology textbooks into chunks and embeddings stored in a vector store to allow precise information retrieval. Furthermore, *ChatDiet* by Yang et al. [35] illustrates an innovative integration of GPT-3·5 Turbo with a population model (a layer of nutrition and dietary guidelines) to deliver nutrition recommendations. Orchestrator enabled *ChatDiet* to retrieve information from the population model in response to user queries and perform a prompt engineering function to instruct LLMs effectively. Through retrieval-augmented generation, LLMs were fine-tuned to enhance their applicability in chronic disease management.

Prompt Engineering and Wearable Devices Interacted with LLMs

Prompt engineering comprised of instructions, scenarios, queries, and output indicators in includes studies. Instructions regarding the roles (e.g., physician assistants) and tasks (e.g., generating weekly meal plans) of LLMs were reported in two studies [41, 49]. Real-world clinical cases [42] and imaginary patient scenarios [41, 45, 46, 48, 49] with medical profiles were created to simulate healthcare management in chronic diseases, including diabetes, obesity, cardiovascular diseases, and mental health issues. The queries included patient real-time queries [18, 43, 44] and selected common queries from patients and their families [31, 34, 36, 38, 39]. These scenarios and queries interacted with the LLMs to generate responses related to the symptoms, complications, diagnoses, treatment, and management of chronic diseases. Output indicators, including word count, references, literacy levels, tone, format, and principles for generating management plans (e.g., frequency, intensity, time, and type of exercises and diversity of meals) were applied to enhance the applicability of LLM recommendations [32, 37, 38, 40, 46].

Most importantly, Yang et al. [35] explored the integration of ChatGPT with wearable devices for monitoring physical activity levels, sleep patterns, and electrodermal activity to update patient health profiles, allowing ChatGPT to dynamically adjust food recommendations. The integration allows collecting patient data in real-time, providing a dynamic and responsive healthcare management platform.

Feasibility, Opportunities, and Challenges of LLMs Across the Chronic Disease Management Spectrum

LLMs were engaged in a spectrum of roles encompassing prevention, screening, diagnosis, treatment, and long-term care of chronic diseases. Two studies used LLMs to generate suggestions for increasing physical activity and generating nutrition-oriented food recommendations, contributing to the prevention of chronic diseases [35, 40]. Lim et al. [41] integrated LLMs to recommend screening and surveillance intervals for colorectal cancer and streamlining efforts for early detection of chronic diseases. Three studies focused on treating chronic diseases by applying LLMs to generate treatment recommendations and support postoperative care [31, 32, 42]. In 14 studies, LLMs acted as virtual health coaches, offered mental health support, managed symptoms, and generated diet and exercise plans to assist in long-term care of chronic diseases [18, 33, 34, 36-39, 43-49]. The feasibility, opportunities, and challenges inherent in these roles are delineated below.

Feasibility of LLMs in managing chronic diseases

Relevance and accuracy

The feasibility of LLMs in managing chronic diseases, including the relevance, accuracy, reliability, readability, and actionability of their responses, was assessed by patients, caregivers, researchers, and healthcare specialists using interviews [18, 43, 44], content comparisons [42], grading [36, 47], and measurements (e.g., the Flesch–Kincaid Grade level) [32, 33, 48].

Averaging 92%, LLMs demonstrated ability to generate relevant recommendations, showing high pertinence to the patients’ concerns [31, 40]. LLMs also had acceptable accuracy in identifying diagnoses and deterioration of symptoms, recommending investigation and treatment options, and generating health educational materials for several chronic diseases (including carpal tunnel syndrome, liver cirrhosis and mental health) [37, 38], with rates ranging from 76% to 99% as validated by healthcare experts [33]. LLMs also demonstrated high concordance rates with various guidelines, including those for postoperative care of hearing loss (100%) [31], multidisciplinary tumour board recommendations (86.7%) [42], as well as creating general exercise programs that adhere to rate of perceived exertion guidelines and research evidence [46]. Two studies further corroborated these findings, indicating that ChatGPT exceeds hypothesized median scores for accuracy in addressing queries related to chronic diseases (e.g., obesity and diabetes) [48], albeit with occasional issues regarding the accuracy of cited references [40].

In contrast, Yeo et al. [36] revealed mixed performance of LLMs, identifying 50% mixed or incorrect responses in screening, diagnosing, and managing hepatocellular carcinoma. In particular, LLMs failed to correctly identify eligibility and screening tests for hepatocellular carcinoma based on patient characteristics (e.g., age) and failed to determine cut-offs for specific conditions, such as liver transplantation [36]. Similarly, in colorectal cancer screening, ChatGPT 4.0 experienced hallucinations in identifying high-risk features of colorectal cancer and recommended incorrect screening intervals in 51% cases [41]. LLMs might experience inaccuracy in performing complex cancer screening tasks.

Further comparative analyses revealed that fine-tuned LLMs outperformed general-purpose LLMs in terms of response accuracy. For example, fine-tuned ChatGPT 4·0 resulted in few hallucinations and significantly outperformed its predecessor in correctly recommending colorectal cancer screening intervals (79% vs. 50.5%) [41] and responding to ophthalmological questions (83.4% vs. 69.2%) [34].

The random-effects meta-analysis of the above studies [31, 34, 36, 41, 42] showed a pooled accurate rate of 71% (95% CI = 0.59, 0.83; I^2 = 88.32%; P < 0.001) (Figure 3a). Using leave-one-out approach, sensitivity analysis showed that the removal of individual studies on the pooled accurate rate was not statistically significant (Supplementary material Figure 1). Compared to general-purpose LLMs, fine-tuned LLMs had higher rate of accurate responses (OR = 2.89; 95% CI = 1.83, 4.58; I^2 = 54.45%; P < 0.001) (Figure 3b) [34, 41].

Reliability

The reliability of LLM performance remained a major concern. Although fine-tuned LLMs, such as *Aeyeconsult*, demonstrated improved response reliability compared to general-purpose LLMs (e.g., ChatGPT 4.0), these models still experienced missing, multiple, or contradictory answers [34]. These inconsistencies often stemmed from failures within the underlying LLM architecture, information retrieval processes, or the synthesis of information from multiple sources [34]. Studies also noted instances of hallucination, citation of non-existent sources, and insufficient depth of information [34, 37]. Qualitative feedback from patients further emphasised the unreliability of LLM-generated information, citing concerns about outdated data, biases in training datasets, and LLMs' self-acknowledged limitations in verifying factual accuracy [18, 44].

Readability

The readability of LLM responses was consistently rated high across the included studies (67% to 98%) [31-33, 38, 48]. LLMs outperformed webpages [32] and human-derived materials [33] in comprehension, in which the healthcare recommendations regarding cirrhosis, obesity, diabetes, and cardiovascular diseases were easily understood by individuals with high school or above educational levels [33, 48]. The random-effects meta-analysis of two included studies [32, 48] indicated that the Flesch-Kincaid Grade Level score and the Flesch-Kincaid Reading Ease Score were 12.04 (95% CI = 7.18, 16.90; I^2 = 87.97%; P < 0.001) (Figure 3c) and 42.49 (95% CI = 12.71, 72.26; I^2 = 89.51%; P = 0.01) (Figure 3d), respectively.

Actionability

The actionability of LLM responses remains a grey area. While Pradhan and colleagues reported no significant differences between LLM-derived and human-derived cirrhosis management materials concerning actionability, only human-derived content met the actionable score threshold of ≥70% [33]. LLM responses might lack depth and detail for practical application

[37].

Opportunities of LLMs in managing chronic diseases

Increasing knowledge and awareness

Using internet-enabled devices, LLMs provide equal and free access to chronic disease information, especially for patients from rural areas [18, 43]. This utility enhances patient knowledge and awareness about ailments, preventive measures, symptoms, and the management of chronic diseases, including cancer, diabetes, and kidney failures [18, 43, 44]. This also helped dispel misperceptions about lifestyle modification (e.g., diet and smoking cession) and chemotherapy in cancer management [18]. As noted from participants, this benefit is particularly pronounced in comparison to traditional search engines, which require navigating multiple websites for consolidated information [43].

Promoting self-management behaviours

By motivating health goals and developing achievable plans, LLMs promote patient self-management behaviours, including diets, smoking cessation, physical activities, sleep, and meditation [43, 44]. LLMs also informed patients with non-pharmacological techniques, such as relaxation, sleep hygiene practices, and stress-reduction techniques, to cope with symptoms of chronic diseases, including insomnia, fatigue, nausea, and pain [18, 44]. However, a major limitation is the current inability of LLMs to store and manage long-term data for behaviour change. It is suggested to integrate LLMs with eHealth systems, wearables, and health management applications for continuous monitoring and tracking of their health conditions (e.g., blood glucose level), facilitating personalised care plans, and setting reminders for health behaviours (e.g., taking medication) [43].

Enhancing emotional, social, and healthcare support

LLMs provided a non-judgmental space for emotional expression and offered compassionate responses to enhance patient emotional well-being [44]. For example, Yeo et al. [36] highlighted ChatGPT’s psychological and practical support to patients following a diagnosis of hepatocellular carcinoma. LLMs also helped patients practice cognitive behavioural therapy techniques, which involved identifying negative thoughts and replacing them with more balanced thoughts, to reframe their emotions positively [44]. However, some critical concerns persist, such as lack of capabilities in assessing mental health conditions [44] and patient perceived lack of deep understanding and personalised empathy compared to human healthcare professionals [43].

At social level, LLMs demonstrated capability to guide patients on accessing hotlines, counsellors, and online supporting groups, such as cancer support groups, which are important to connect with other patients, attain peer support, and access to updated treatment information regarding chronic diseases [18, 43, 44].

At healthcare level, LLMs improved healthcare support by linking health resources [44] and providing scalable support that is accessible to a large number of patients simultaneously [43].

Many patients reported difficulties in securing appointments with specialists for their chronic diseases and were dissatisfied with the limited time they had with specialists to understand their conditions, preventive measures, and treatment procedures [43]. LLMs alleviated this issue by offering comprehensive information on various chronic conditions, thus decreasing the need for frequent specialist consultations, helping patients become more self-reliant and better informed about their health, thereby reducing the strain on the healthcare system and improving overall patient outcomes [43, 44].

Challenges of LLMs in managing chronic diseases

Potential privacy, language, and cultural issues

Privacy and security concerns are paramount for patients when using LLMs for chronic disease management [18, 43, 44]. Patients reported the absence of data protection guidelines and lack of anonymous features of LLMs because most tools required registration. Patients were not confident about sharing their personal health data and feared potential misuse. Additionally, although LLMs can manage basic linguistic tasks, they often fail to grasp dialectal subtleties [43, 44]. Spallek et al. [38] reported that 23% of LLM outputs had at least one stigmatising phrase. This inadequacy is evident in the context of traditional medicine, where culturally rooted concepts are not effectively understood, potentially leading to misinformation [43].

Incompetence in tackling advanced tasks in chronic disease management

LLMs are not mature enough to address advanced chronic disease management tasks. Some LLMs cannot interpret complex diagnostic reports that include nontext inputs, including radiology images, blood tests, and other medical documents [43, 44]. LLMs rely on patients' self-reported symptoms to diagnose diseases, and they cannot initiate interactive dialogues, which precludes them from probing hidden symptoms, clarifying patient conditions, and identifying appropriate management plans [38, 43-45]. LLMs are also unreliable as they provide simplified recommendations and frequently acknowledge potential inaccuracies in addressing complex comorbidities of chronic diseases and recommending effective medicine based on patient data [18, 43, 44].

Gaps in generating personalised chronic disease management regimens

Regarding content and format, LLMs could not generate personalised chronic disease management regimens. For example, LLMs was unable to monitor individuals' physiological response, failed to adjust physical exercise regimens in real time, and could not generate customised treatment plans in complex disease scenarios (e.g., systematic lupus erythematosus) [18, 46]. By contrast, Yang and colleagues integrated ChatGPT with wearable devices to monitor patient physical activity, sleep patterns, and electrodermal activity to update health profiles in real time, allowing ChatGPT to adjust food recommendations dynamically [35]. Integration LLMs with wearables might address this challenge by providing a dynamic and responsive platform for chronic disease management. Moreover, LLMs might also struggle to transform text-based information into multimodal formats (e.g., images and video), influencing the effective delivery of information tailored to patient preferences [18].

Methodology Quality Assessment Results

Supplementary material Table 2 and Table 3 present the quality assessment results. Three quasi-experimental studies likely suffered from a serious risk of bias due to potential covariates and potential deviations from intended interventions (e.g., accessing healthcare information from other online resources) [18, 43, 44]. Seventeen simulation and case studies attained methodology quality scores ranging from 66.7% to 89.6%, which were influenced primarily by the lack of valid instruments for measuring the feasibility of LLMs [31, 47, 49]; inadequate reporting of qualitative data collection, coding, and analysis processes [40, 45, 46]; and having small samples of patient scenarios mimicking chronic disease healthcare seeking, ranging from 5 to 30 in ten studies [31-33, 37, 38, 42, 45, 46, 48, 49].

Discussion

This systematic review included 20 studies to synthesise evidence on the feasibility, opportunities, and challenges of LLMs in transforming chronic disease management. Findings suggested that LLMs can feasibly recommend relevant, comprehensible, and accurate health information (71%; 95% CI = 0.59, 0.83; I^2 = 88.32%; P < 0.001). They enhanced equitable information access, patient awareness, and self-management behaviours as well as provided emotional support, social connections, and healthcare resource linkage, collectively contributed to improving chronic disease outcomes. Nevertheless, LLMs faced challenges in addressing privacy, language, and cultural issues, undertaking advanced diagnostic and medication recommendation tasks, and generating personalised regimens with real-time adjustments and multiple modalities. These insights are pivotal for healthcare professionals in harnessing the transformative potential of LLMs in chronic disease management.

Feasibility, encompassing relevance, accuracy, and reliability, is the premise for LLM applications into chronic disease management. Consistent with previous literature [27], LLMs exhibit the capacity to generate relevant responses tailored to the concerns of patients with chronic diseases [31, 40]. This adaptability is attributed to their advanced natural language processing abilities, which enable them to align closely with patient inquiries and medical contexts [13, 14]. However, LLMs present a mixed profile regarding accuracy across different tasks of chronic disease management with a pooled accurate rate of 71%. Specifically, LLMs have shown acceptable accuracy (76% to 99%) in generating health educational materials, while their accuracy is particularly concerning when applied to cancer screening tasks [36, 41]. To enhance accuracy, several studies have fine-tuned LLMs using retrieval-augmented generation to combine LLMs with contextual knowledge bases [34, 35, 41]. Compared to general-purpose models, fine-tuned LLMs exhibit greater accuracy and adhere more closely to medical guidelines [41]. Despite these advancements, issues such as hallucination, contradictory responses, and the citation of non-existent sources persist [37]. Furthermore, the tendency of LLMs to provide simplified recommendations, acknowledge potential inaccuracies, and advise users to consult healthcare professionals diminishes their reliability [18, 43, 44]. Therefore, integrating LLMs into chronic disease management necessitates ongoing development, clinical validation, and collaboration with healthcare professionals to optimise their accuracy and reliability.

This review confirms that LLMs provide multifaceted opportunities for chronic disease management at individual, social, and healthcare system levels. At individual level, LLMs provide equitable access to health information about chronic diseases, particularly benefiting patients residing in rural areas who may have limited access to health information and suffer from low health literacy [18, 43]. Consistent with this review, previous literature also highlighted that LLMs facilitate health communication through telehealth, minimize geography, travel, and financial challenges, and provide a solution to disparities in healthcare information access among rural communities [21]. Such accessibility enhances patient knowledge and awareness of preventive measures, diagnoses, symptoms, treatment, and management strategies for various chronic diseases, including cancer, diabetes, and kidney failure [18, 43, 44]. In line with findings of this review, previous reviews also confirmed that LLMs play significant roles in patient health education by generating health education materials and providing multifaced recommendations covering medical information, lifestyle recommendations, and perioperative care instructions [50, 51]. The advanced algorithms and extensive dataset training of LLMs might enable natural language interactions, real-time feedback, and personalised health information, effectively addressing patient queries and improving patient knowledge about chronic diseases [15]. These characteristics make LLMs more advantageous for health education than traditional algorithm-based applications, which typically rely on tedious checkbox questionnaires, offer constrained responses, and require backend processing from healthcare professionals [52, 53].

Additionally, this review reveals a burgeoning interest in leveraging LLMs to enhance behaviour change interventions. The included studies demonstrate that LLMs positively influence patient adherence to recommended healthy behaviours, including balanced diets, regular exercise, and smoking cessation [43]. LLMs also showed promise in assisting patients with disease monitoring and self-management of physical (e.g., fatigue and pain) and emotional symptoms (e.g., fear and anxiety) by recommending practical tips and psychotherapeutic exercises such as guided imagery [18, 44]. The findings are consistent with previous research indicating the effectiveness of LLMs in promoting health behaviour change by enhancing health knowledge, debunking health myths, and providing motivational support [50, 54]. However, a critical limitation in the current literature is the insufficient integration of established behaviour change theories within the LLM-based interventions. This methodological gap limits the efficiency of intervention contents and hinders the identification of causal mechanisms. To optimize LLM efficacy in facilitating behaviour change, it is suggested to integrate theoretical models to precisely tailor content and maximize the likelihood of sustained behaviour changes [40].

While LLMs showed promise in enhancing emotional support for patients with chronic diseases, their capacity for genuine empathy remains controversial. Several included studies reported positive patient experience with LLM-provided emotional support [18, 36]; however, it might lack the empathy and personalised support compared to human healthcare professional interactions [43, 45]. In line with previous literature [23], existing algorithms of LLMs, while capable of emulating empathetic responses and offering practical advice, may not fully grasp the complexities of human emotion. Therefore, LLMs should currently be considered as supplementary tools, rather than replacements, for human-centred emotional care in chronic diseases management.

At social level, LLMs may guide patients in supporting their networks and facilitating peer and social support. Previous studies corroborate this finding, indicating that LLMs can support patients by connecting patients with relevant resources [15, 55]. At healthcare level, the findings of this review suggest that LLMs can deliver scalable health support [43], link medical resources, and recommend crisis interventions [44] to improve healthcare support and reduce patients' reliance on face-to-face consultations in healthcare facilities. These findings align with previous literature, suggesting that the broader implementation of LLMs could alleviate the burden on healthcare systems [24, 56]. Nonetheless, the positive effects of LLMs have only been tested among a small cohort of patients [18, 43, 44], and rigorous randomised controlled trials are needed to validate these outcomes.

Despite this, several challenges exist in the application of LLMs for managing chronic diseases. First, uncertainty remains around data privacy and sharing conflicts when using LLMs for chronic disease management. LLMs use personal data to provide accurate and customised recommendations; however, this requirement often conflicts with stringent privacy protocols. Due to a lack of robust data protection guidelines and the inability to register anonymously, patients may fear data leakage or misuse and hesitate to share personal health information [43, 44]. Studies have integrated wearable devices with LLMs to collect real-time data, such as sleep patterns [35], which complicates issues related to data encryption and security during transmission [56].

Second, the ethical implications of using LLMs to diagnose chronic diseases should also be clarified. Although studies have shown that LLMs can accurately diagnose conditions such as psychiatric disorders [47] and carpal tunnel syndrome [37], their comprehensive diagnostic capabilities remain limited. As noted in included studies [38, 43-45], some LLMs rely on patients' self-reported symptoms and general medical knowledge and cannot perform physical examinations or interpret complex diagnostic reports exactly, such as radiology images and blood tests. This may lead to inaccurate or incomplete diagnoses [36], delay necessary treatments, and jeopardise patient safety. Therefore, ethical standards should be carefully considered before integrating LLMs into healthcare systems for performing diagnostic tasks.

Another challenge is that LLM recommendations cannot achieve adequate personalisation regarding information content and modality. Despite the potential of LLMs in personalised healthcare based on their natural language processing capabilities [21, 23], significant gaps remain in achieving genuine personalisation. For complex conditions, such as cancer [18] and systemic lupus [46], LLMs tend to provide generic and broad advice and struggle to make real-time adjustments to personalised regimens (e.g., exercises) [45]. Moreover, some LLMs (e.g., ChatGPT 3.0) primarily operate in text-based formats and do not readily generate multimodal information to effectively deliver information tailored to patient preferences [18]. Their inability to process and generate multimodal information, such as images, videos, and other nontextual medical data, limits their efficacy in personalised chronic disease management. Additional interdisciplinary collaborations between clinical experts and artificial intelligence professionals, coupled with the conduct of rigorous clinical trials, are imperative to ensure effective integration of LLMs into chronic disease management frameworks.

Limitations

First, although this review searched 11 databases, our search strategy did not include keywords representing various chronic diseases and might have missed relevant studies. Second, heterogeneity was observed in the pooled analyses of LLM feasibility outcomes, including accurate rates and readability, which could stem from variations in the aetiologies of chronic diseases and types of LLMs. The heterogeneity might undermine the confidence in the synthesised findings. Third, most studies included in this review used simulations, and their results may not be generalised to real-world clinical settings, especially when considering the small sample of patient scenarios. Patient experiences with LLMs in chronic disease management should be explored using rigorous randomised controlled trials.

Implications for Practice, Research, and Policymaking

Implications for Practice

LLMs have high potential to support healthcare professionals in optimising chronic disease management by enhancing patient awareness and self-management behaviours as well as providing emotional support, social connections, and healthcare resource linkage. Considering that only three included studies involved real-world clinical implementation of LLMs among patients with small sample sizes, the application and performance of LLMs in chronic disease management require further empirical inquiry. Critically, empathic communication, a cornerstone of healthcare practice, is currently irreplaceable by LLMs. An integrative strategy that leverages the strengths of LLMs while preserving the irreplaceable human touch should be implemented in future practice for chronic disease management.

Implications for Research

A multifaceted approach is needed to enhance the transformative potential of LLMs in chronic disease management. First, robust data protection measures must be implemented to safeguard patient information. Advanced anonymisation techniques and end-to-end encryption protocols should be used to protect patient privacy and secure data transmission while interacting with LLMs. Second, to enhance the diagnostic accuracy of LLMs for chronic diseases, these models should be fine-tuned with domain-specific knowledge to ensure they align with the most current medical standards and practices. Ideally, multimodal LLMs capable of integrating diverse data modalities, such as radiology images and laboratory data, should be developed to improve diagnostic precision. Third, integrating wearables with LLMs may increase the provision of dynamic and personalised health recommendations. Wearable devices can continuously monitor vital physiological parameters such as heart rate, blood pressure, glucose levels, and physical activity. These real-time data can be fed into LLMs to adjust treatment plans timeously and offer highly personalised healthcare interventions.

Implications for Policymaking

In rural and remote settings, where access to chronic disease healthcare might be hindered by geographical, infrastructural, and socioeconomic barriers, policymakers should strategically integrate LLMs into healthcare systems. This integration can bridge critical gaps, ensuring equitable access to health information, resources, and online support communities. This could help foster proactive health awareness, empower informed health decisions, and facilitate timely interventions for underserved populations, thereby reducing systemic disparities in chronic disease management.

Conclusions

This review, encompassing 20 publications, synthesised evidence on the transformative potential of LLMs in chronic disease management. LLMs demonstrated feasibility in generating relevant, comprehensible, and accurate health information, although their reliability and actionability remain debated. They presented opportunities in (1) enhancing patient knowledge and awareness of aliments, prevention, symptoms, and management of chronic diseases, (2) facilitating self-management behaviours in lifestyle modification and symptom coping, and (3) linking supporting groups and healthcare resources to enhance emotional, social, and healthcare support.

However, their real-world clinical implementation among patients remains nascent, highlighting further empirical inquiry. For future research, a multifaced approach–incorporating robust data security, domain-specific model fine-tuning, multimodal data integration, and wearables–is crucial to evolve LLMs into invaluable adjuncts for healthcare professionals to transform chronic disease management. Strategic integration of LLMs in rural and remote areas is also suggested to bridge service access gaps and reduce disparities in chronic disease outcomes.

Conflicts of Interest

None declared.

Abbreviations

LLM: Large Language model

Data availability

The search strategies and data extracted for this review are available in the main text and supplementary information files; any additional data are available on reasonable request.

References

1. World Health Organization. Noncommunicable diseases key facts. 2023 Sep. URL: https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases [accessed 2024-11-23]
2. Chowdhury SR, Chandra Das D, Sunna TC, Beyene J, Hossain A. Global and regional prevalence of multimorbidity in the adult population in community settings: a systematic review and meta-analysis. EClinicalMedicine 2023 Feb 16:57:101860 [doi: 10.1016/j.eclinm.2023.101860] [Medline: 36864977]
3. Cordova R, Viallon V, Fontvieille E, Peruchet-Noray L, Jansana A, Wagner KH, et al. Consumption of ultra-processed foods and risk of multimorbidity of cancer and cardiometabolic diseases: a multinational cohort study. Lancet Reg Health Eur. 2023 Nov 14:35:100771 [doi: 10.1016/j.lanepe.2023.100771] [Medline: 38115963]
4. GBD 2021 Forecasting Collaborators. Burden of disease scenarios for 204 countries and territories, 2022-2050: a forecasting analysis for the Global Burden of Disease Study 2021. Lancet 2024 May 18;403(10440):2204-2256 [doi: 10.1016/s0140-6736(24)00685-8] [Medline: 38762325]
5. United Nations. Transforming our world: the 2030 agenda for sustainable development. 2015. URL: https://sustainabledevelopment.un.org/post2015/transformingourworld/publication [accessed 2024-11-29]
6. Badr Y, Abdul Kader L, Shamayleh A. The Use of Big Data in Personalized Healthcare to Reduce Inventory Waste and Optimize Patient Treatment. J Pers Med 2024 Apr 3;14(4):383 [doi: 10.3390/jpm14040383] [Medline: 38673011]
7. Stefanicka-Wojtas D, Kurpas D. Personalised Medicine-Implementation to the Healthcare System in Europe (Focus Group Discussions). J Pers Med 2023 Feb 21;13(3):380 [doi: 10.3390/jpm13030380] [Medline: 36983562]
8. Burnier M. The role of adherence in patients with chronic diseases. European Journal of Internal Medicine 2024 Jan:119:1-5 [doi: 10.1016/j.ejim.2023.07.008] [Medline: 37479633]
9. Treatment adherence: can fixed-dose combinations help? Lancet Diabetes Endocrinol 2015 Feb;3(2):91 [doi: 10.1016/s2213-8587(15)70011-2] [Medline: 25618290]
10. Bello AK, Okpechi IG, Levin A, Ye F, Damster S, Arruebo S, et al. An update on the global disparities in kidney disease burden and care across world countries and regions. Lancet Glob Health 2020 Dec;26(12):1835-1838 [doi: 10.1016/s2214-109x(23)00570-3] [Medline: 38365413]
11. Weiss DJ, Nelson A, Vargas-Ruiz CA, Gligorić K, Bavadekar S, Gabrilovich E, et al. Global maps of travel time to healthcare facilities. Nature Medicine 2020 Dec;26(12):1835-1838 [doi: 10.1038/s41591-020-1059-1] [Medline: 32989313]
12. Lyons J, Akbari A, Abrams KR, Azcoaga Lorenzo A, Ba Dhafari T, Chess J, et al. Trajectories in chronic disease accrual and mortality across the lifespan in Wales, UK (2005-2019), by area deprivation profile: linked electronic health records cohort study on 965,905 individuals. Lancet Reg Health Eur 2023 Jul 17:32:100687 [doi:

10.1016/j.lanepe.2023.100687] [Medlline: 37520147]

13. Cinquin O. ChIP-GPT: a managed large language model for robust data extraction from biomedical database records. Brief Bioinform 2024 Jan 22;25(2):bbad535 [doi: 10.1093/bib/bbad535] [Medline: 38314912]
14. Yang X, Chen A, PourNejatian N, Shin HC, Smith KE, Parisien C, et al. A large language model for electronic health records. NPJ Digit Med 2022 Dec 26;5(1):194 [doi: 10.1038/s41746-022-00742-2] [Medline: 36572766]
15. Zhu L, Anand A, Gevorkyan G, McGee LA, Rwigema JC, Rong Y, et al. Testing and Validation of a Custom Trained Large Language Model for HN Patients with Guardrails. International Journal of Radiation Oncology*Biology*Physics. 2024;118(5):e52-e3.
16. Henson JB, Glissen Brown JR, Lee JP, Patel A, Leiman DA. Evaluation of the Potential Utility of an Artificial Intelligence Chatbot in Gastroesophageal Reflux Disease Management. Am J Gastroenterol 2023 Dec 1;118(12):2276-2279 [doi: 10.14309/ajg.0000000000002397] [Medline: 37410934]
17. Lautrup AD, Hyrup T, Schneider-Kamp A, Dahl M, Lindholt JS, Schneider-Kamp P. Heart-to-heart with ChatGPT: the impact of patients consulting AI for cardiovascular health advice. Open Heart 2023 Nov;10(2):e002455 [doi: 10.1136/openhrt-2023-002455] [Medline: 37945282]
18. Alanezi F. Examining the role of ChatGPT in promoting health behaviors and lifestyle changes among cancer patients. Nutr Health 2024 Apr 3:2601060241244563 [doi: 10.1177/02601060241244563] [Medline: 38567408]
19. Sievert M, Aubreville M, Mueller SK, Eckstein M, Breininger K, Iro H, et al. Diagnosis of malignancy in oropharyngeal confocal laser endomicroscopy using GPT 4.0 with vision. Eur Arch Otorhinolaryngol 2024 Apr;281(4):2115-2122 [doi: 10.1007/s00405-024-08476-5] [Medline: 38329525]
20. Liu S, McCoy AB, Wright AP, Carew B, Genkins JZ, Huang SS, et al. Leveraging Large Language Models for Generating Responses to Patient Messages. medRxiv 2023 Jul 16:2023.07.14.23292669 [doi: 10.1101/2023.07.14.23292669] [Medline: 37503263]
21. Wang X, Sanders HM, Liu Y, Seang K, Tran BX, Atanasov AG, et al. ChatGPT: promise and challenges for deployment in low- and middle-income countries. Lancet Reg Health West Pac. 2023 Sep 15:41:100905 [doi: 10.1016/j.lanwpc.2023.100905] [Medline: 37731897]
22. Wu X, Duan R, Ni J. Unveiling security, privacy, and ethical concerns of ChatGPT. Journal of Information and Intelligence. 2024 Jun:246:104264 [doi: 10.1016/j.actpsy.2024.104264] [Medline: 38626597]
23. Clusmann J, Kolbinger FR, Muti HS, Carrero ZI, Eckardt J-N, Laleh NG, et al. The future landscape of large language models in medicine. Communications Medicine 2023 Oct 10;3(1):141 [doi: 10.1038/s43856-023-00370-1] [Medline: 37816837]
24. Karabacak M, Margetis K. Embracing Large Language Models for Medical Applications: Opportunities and Challenges. Cureus. 2023 May 21;15(5):e39305 [doi: 10.7759/cureus.39305] [Medline: 37378099]
25. Chen S, Guevara M, Moningi S, Hoebers F, Elhalawani H, Kann BH, et al. The effect of using a large language model to respond to patient messages. Lancet Digit Health 2024 Jun;6(6):e379-e381 [doi: 10.1016/s2589-7500(24)00060-8] [Medline: 38664108]

26. Amir-Behghadami M, Janati A. Population, Intervention, Comparison, Outcomes and Study (PICOS) design as a framework to formulate eligibility criteria in systematic reviews. Emergency Medicine Journal. 2020 Jun;37(6):387 [doi: 10.1136/emermed-2020-209567] [Medline: 32253195]
27. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature 2023 Aug;620(7972):172-180 [doi: 10.1038/s41586-023-06291-2] [Medlline: 37438534]
28. Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. Bmj 2003 Sep 6;327(7414):557-60 [doi: 10.1136/bmj.327.7414.557] [Medline: 12958120]
29. Fey MK, Gloe D, Mariani B. Assessing the Quality of Simulation-Based Research Articles: A Rating Rubric. Clinical Simulation In Nursing. 2015 Dec;11(12):496-504 [doi: 10.1016/j.ecns.2015.10.005]
30. Sterne JAC, Hernán MA, Reeves BC, Savović J, Berkman ND, Viswanathan M, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ 2016 Oct 12:355:i4919 [doi: 10.1136/bmj.i4919] [Medline: 27733354]
31. Aliyeva A, Sari E, Alaskarov E, Nasirov R. Enhancing Postoperative Cochlear Implant Care With ChatGPT-4: A Study on Artificial Intelligence (AI)-Assisted Patient Education and Support. Cureus. 2024 Feb 9;16(2):e53897 [doi: 10.7759/cureus.53897] [Medline: 38465158]
32. Kianian R, Sun D, Giaconi J. Can ChatGPT Aid Clinicians in Educating Patients on the Surgical Management of Glaucoma? J Glaucoma 2024 Feb 1;33(2):94-100 [doi: 10.1097/ijg.0000000000002338] [Medline: 38031276]
33. Pradhan F, Fiedler A, Samson K, Olivera-Martinez M, Manatsathit W, Peeraphatdit T. Artificial intelligence compared with human-derived patient educational materials on cirrhosis. Hepatol Commun 2024 Feb 14;8(3):e0367 [doi: 10.1097/hc9.0000000000000367] [Medline: 38358382]
34. Singer MB, Fu JJ, Chow J, Teng CC. Development and Evaluation of Aeyeconsult: A Novel Ophthalmology Chatbot Leveraging Verified Textbook Knowledge and GPT-4. J Surg Educ 2024 Mar;81(3):438-443 [doi: 10.1016/j.jsurg.2023.11.019] [Medline: 38135548]
35. Yang Z, Khatibi E, Nagesh N, Abbasian M, Azimi I, Jain R, et al. ChatDiet: Empowering personalized nutrition-oriented food recommender chatbots through an LLM-augmented framework. Smart Health 2024 June;32:100465. [doi: 10.1016/j.smhl.2024.100465]
36. Yeo YH, Samaan JS, Ng WH, Ting PS, Trivedi H, Vipani A, et al. Assessing the performance of ChatGPT in answering questions regarding cirrhosis and hepatocellular carcinoma. Clin Mol Hepatol 2023 Jul;29(3):721-732 [doi: 10.3350/cmh.2023.0089] [Medline: 36946005]
37. Seth I, Xie Y, Rodwell A, Gracias D, Bulloch G, Hunter-Smith DJ, et al. Exploring the Role of a Large Language Model on Carpal Tunnel Syndrome Management: An Observation Study of ChatGPT. J Hand Surg Am 2023 Oct;48(10):1025-1033 [doi: 10.1016/j.jhsa.2023.07.003] [Medline: 37530687]
38. Spallek S, Birrell L, Kershaw S, Devine EK, Thornton L. Can we use ChatGPT for Mental

Health and Substance Use Education? Examining Its Quality and Potential Harms. JMIR Med Educ. 2023 Nov 30:9:e51243 [doi: 10.2196/51243] [Medline: 38032714]

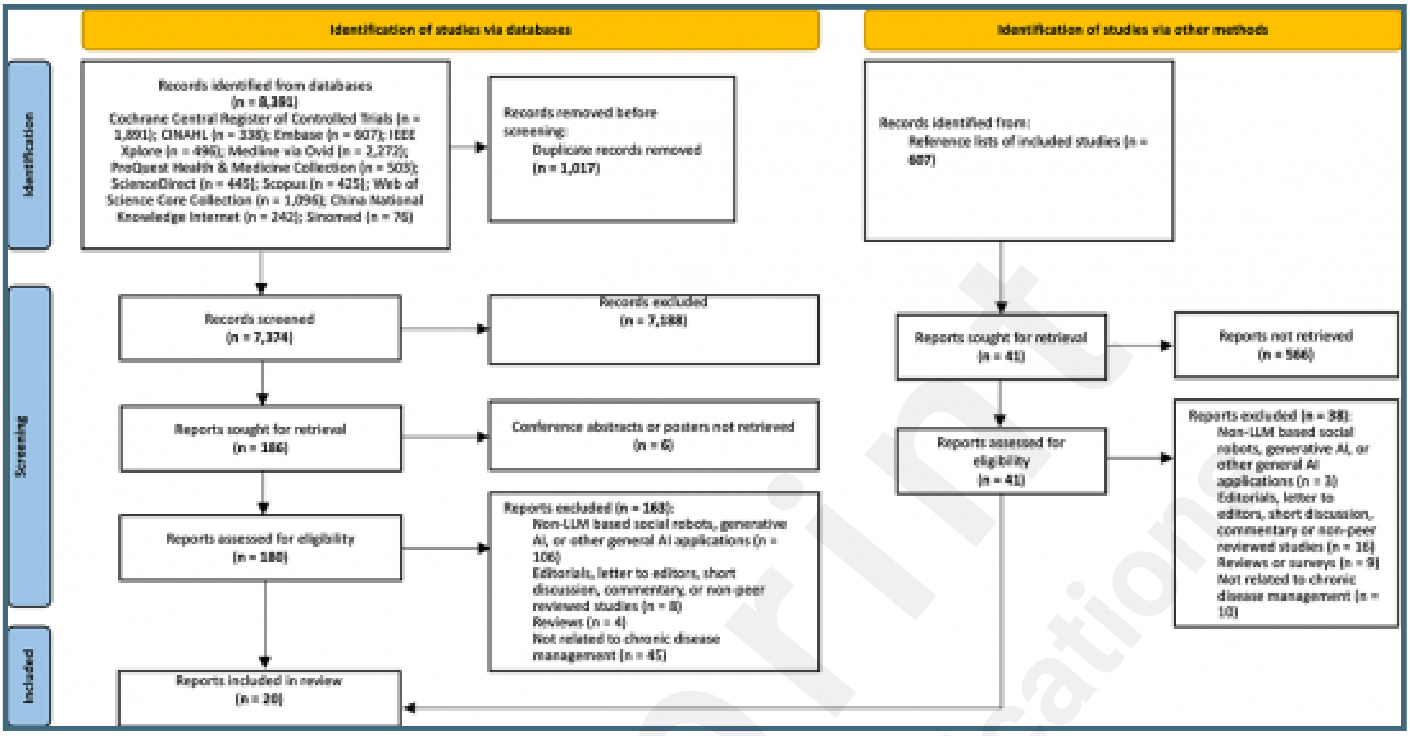
39. Puerto Nino AK, Garcia Perez V, Secco S, De Nunzio C, Lombardo R, Tikkinen KAO, et al. Can ChatGPT provide high-quality patient information on male lower urinary tract symptoms suggestive of benign prostate enlargement? Prostate Cancer Prostatic Dis. 2024 Jun 13 [doi: 10.1038/s41391-024-00847-7] [Medline: 38871841]
40. Willms A, Liu S. Exploring the Feasibility of Using ChatGPT to Create Just-in-Time Adaptive Physical Activity mHealth Intervention Content: Case Study. JMIR Med Educ. 2024 Feb 29:10:e51426 [doi: 10.2196/51426] [Medline: 38421689]
41. Lim DYZ, Tan YB, Koh JTE, Tung JYM, Sng GGR, Tan DMY, et al. ChatGPT on guidelines: Providing contextual knowledge to GPT allows it to provide advice on appropriate colonoscopy intervals. J Gastroenterol Hepatol 2024 Jan;39(1):81-106 [doi: 10.1111/jgh.16375] [Medline: 37855067]
42. Choo JM, Ryu HS, Kim JS, Cheong JY, Baek SJ, Kwak JM, et al. Conversational artificial intelligence (chatGPT™) in the management of complex colorectal cancer patients: early experience. ANZ J Surg 2024 Mar;94(3):356-361 [doi: 10.1111/ans.18749] [PMID: 37905713]
43. Al-Anezi FM. Exploring the use of ChatGPT as a virtual health coach for chronic disease management. Learning Health Systems 2024 Jan 11;8(3):e10406 [doi: 10.1002/lrh2.10406] [Medline: 39036525]
44. Alanezi FA-O. Assessing the Effectiveness of ChatGPT in Delivering Mental Health Support: A Qualitative Study. J Multidiscip Healthc 2024 Jan 31:17:461-471[doi: 10.2147/JMDH.S447368] [Medline: 38314011]
45. Dergaa I, Fekih-Romdhane F, Hallit S, Loch AA, Glenn JM, Fessi MS, et al. ChatGPT is not ready yet for use in providing mental health assessment and interventions. Front Psychiatry 2024 Jan 4:14:1277756 [doi: 10.3389/fpsyt.2023.1277756] [Medline: 38239905]
46. Dergaa I, Saad HB, El Omri A, Glenn JM, Clark CCT, Washif JA, et al. Using artificial intelligence for exercise prescription in personalised health promotion: A critical evaluation of OpenAI's GPT-4 model. Biol Sport. 2024 Mar;41(2):221-241 [doi: 10.5114/biolsport.2024.133661] [Medline: 38524814]
47. Franco D'Souza R, Amanullah S, Mathew M, Surapaneni KM. Appraising the performance of ChatGPT in psychiatry using 100 clinical case vignettes. Asian J Psychiatr 2023 Nov:89:103770 [doi: 10.1016/j.ajp.2023.103770] [Medline: 37812998]
48. Mondal H, Dash I, Mondal S, Behera JK. ChatGPT in Answering Queries Related to Lifestyle-Related Diseases and Disorders. Cureus 2023 Nov 5;15(11):e48296 [doi: 10.7759/cureus.48296] [Medline: 38058315]
49. Papastratis I, Stergioulas A, Konstantinidis D, Daras P, Dimitropoulos K. Can ChatGPT provide appropriate meal plans for NCD patients? Nutrition 2024 May:121:112291 [doi: 10.1016/j.nut.2023.112291] [Medline: 38359704]
50. Busch F, Hoffmann L, Rueger C, van Dijk EHC, Kader R, Ortiz-Prado E, et al. Systematic Review of Large Language Models for Patient Care: Current Applications and Challenges. medRxiv [doi: 10.1101/2024.03.04.24303733]

51. Preiksaitis C, Ashenburg N, Bunney G, Chu A, Kabeer R, Riley F, et al. The Role of Large Language Models in Transforming Emergency Medicine: Scoping Review JMIR Med Inform 2024 May 10:12:e53787 [doi: 10.2196/53787] [Medline: 38728687]
52. Andrew A. Potential applications and implications of large language models in primary care. Fam Med Community Health 2024 Jan 30;12(Suppl 1):e002602 [doi: 10.1136/fmch-2023-002602] [Medline: 38290759]
53. Giebel GD, Abels C, Plescher F, Speckemeier C, Schrader NF, Börchers K, et al. Problems and Barriers Related to the Use of mHealth Apps From the Perspective of Patients: Focus Group and Interview Study. J Med Internet Res 2024 Apr 23:26:e49982 [doi: 10.2196/49982] [Medline: 38652508]
54. Bak M, Chin J. The potential and limitations of large language models in identification of the states of motivations for facilitating health behavior change. Journal of the American Medical Informatics Association 2024 Sep 1;31(9):2047-2053 [doi: 10.1093/jamia/ocae057] [Medline: 38527272]
55. Bushuven S, Bentele M, Bentele S, Gerber B, Bansbach J, Ganter J, et al. "ChatGPT, Can You Help Me Save My Child's Life?" - Diagnostic Accuracy and Supportive Capabilities to Lay Rescuers by ChatGPT in Prehospital Basic Life Support and Paediatric Advanced Life Support Cases - An In-silico Analysis. J Med Syst 2023 Nov 21;47(1):123 [doi: 10.1007/s10916-023-02019-x] [Medline: 37987870]
56. Montazeri M, Galavi Z, Ahmadian L. What are the applications of ChatGPT in healthcare: Gain or loss? Health Sci Rep 2024 Feb 14;7(2):e1878 [doi: 10.1002/hsr2.1878] [Medline: 38361810]

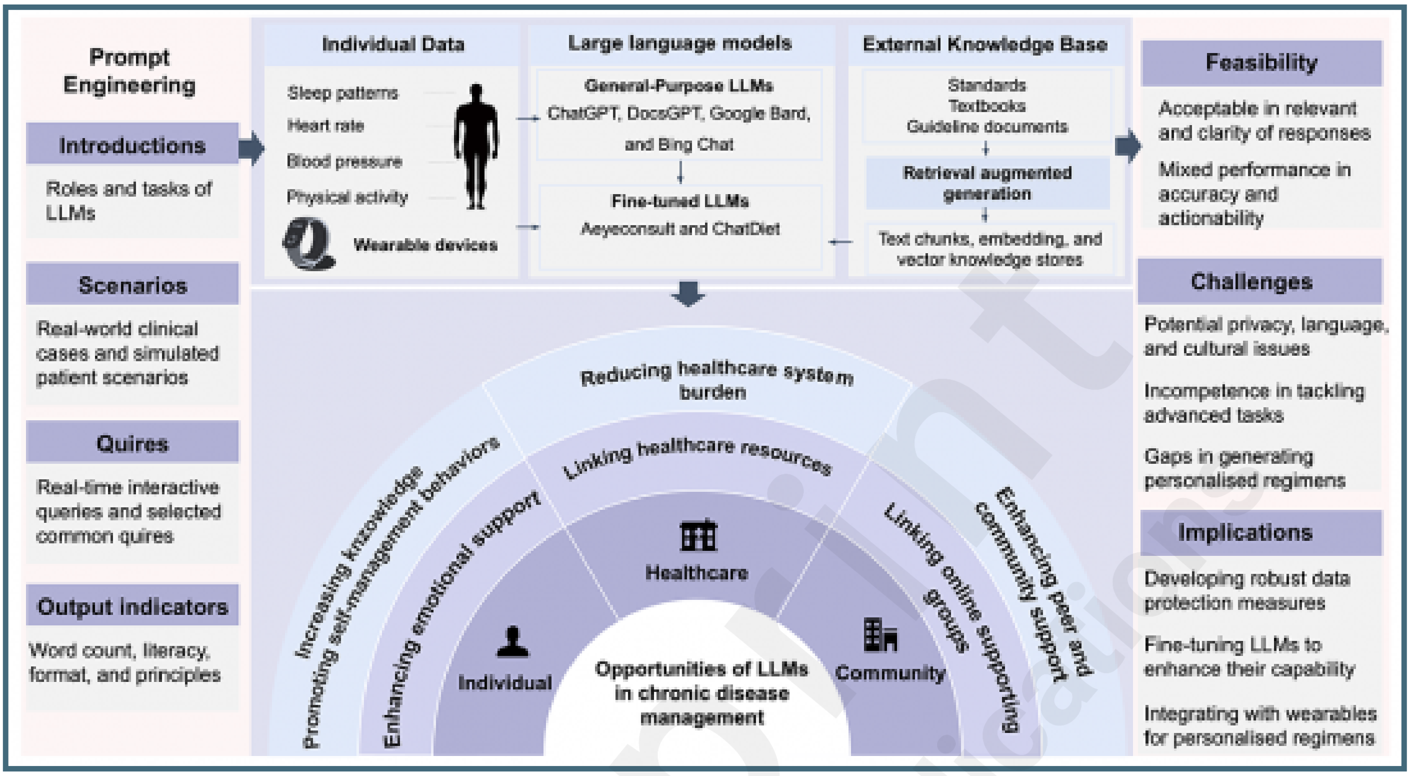
Supplementary Files

Figures

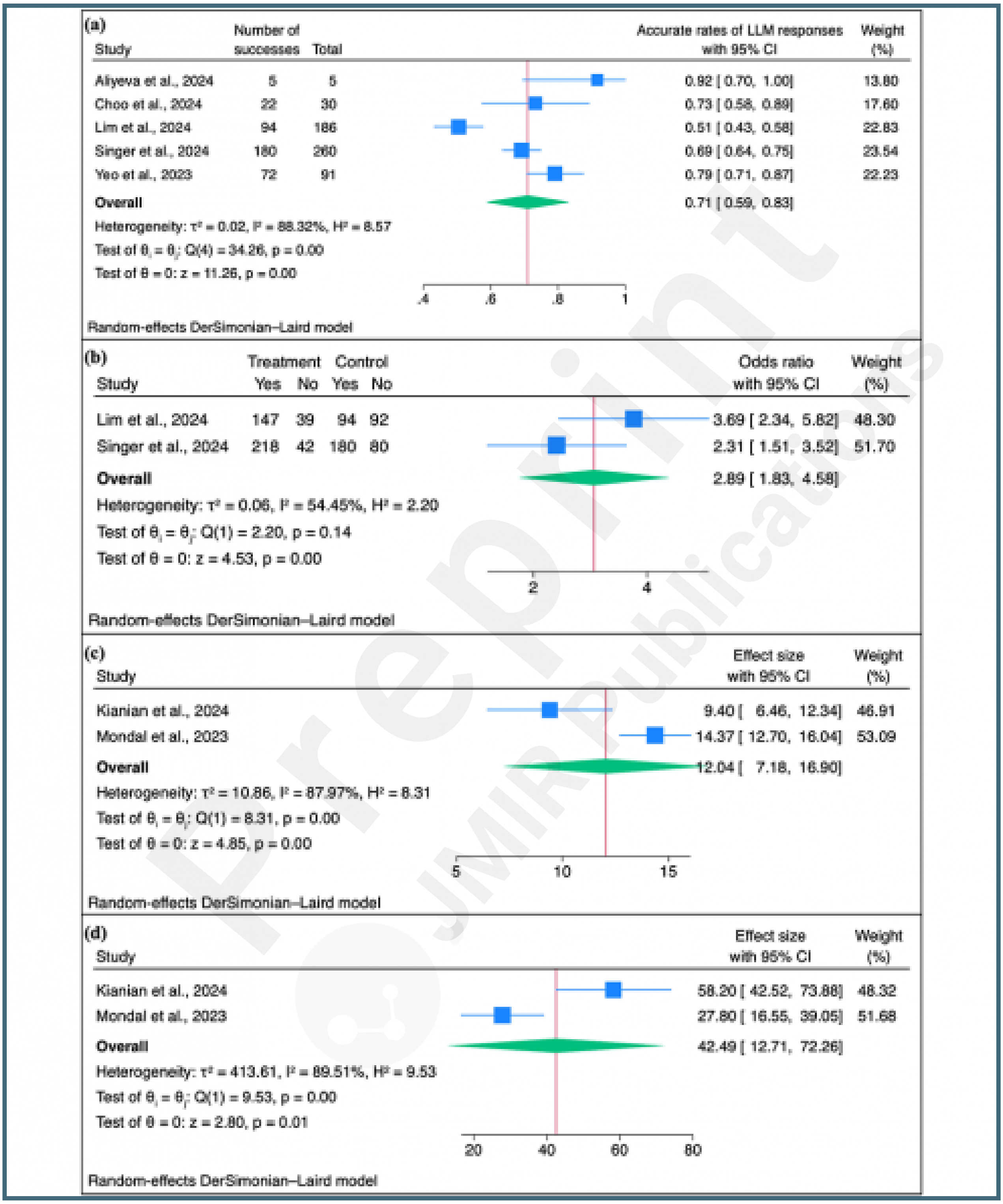
The PRISMA flowchart. AI, artificial intelligence; LLM, large language model; PRISMA, the Preferred Reporting Items for Systematic Reviews and Meta-analyses.


Identification of studies via databases
Identification of studies via other methods
Identification
Screening
Included
Records identified from databases (n = 8,391) Cochrane Central Register of Controlled Trials (n = 1,891); CINAHL (n = 338); Embase (n = 607); IEEE Xplore (n = 496); Medline via Ovid (n = 2,272); ProQuest Health & Medicine Collection (n = 503); ScienceDirect (n = 445); Scopus (n = 425); Web of Science Core Collection (n = 1,096); China National Knowledge Internet (n = 242); Sinomed (n = 76)
Records removed before screening: Duplicate records removed (n = 1,017)
Records screened (n = 7,374)
Records excluded (n = 7,188)
Reports sought for retrieval (n = 186)
Conference abstracts or posters not retrieved (n = 6)
Reports assessed for eligibility (n = 180)
Reports excluded (n = 163): Non-LLM based social robots, generative AI, or other general AI applications (n = 106) Editorials, letter to editors, short discussion, commentary, or non-peer reviewed studies (n = 8) Reviews (n = 4) Not related to chronic disease management (n = 45)
Reports included in review (n = 20)
Records identified from: Reference lists of included studies (n = 607)
Reports sought for retrieval (n = 41)
Reports not retrieved (n = 566)
Reports assessed for eligibility (n = 41)
Reports excluded (n = 38): Non-LLM based social robots, generative AI, or other general AI applications (n = 3) Editorials, letter to editors, short discussion, commentary or non-peer reviewed studies (n = 16) Reviews or surveys (n = 9) Not related to chronic disease management (n = 10)

Characteristics, feasibility, opportunities, and challenges of large language models (LLMs) in chronic disease management.


Prompt Engineering
Introductions
Roles and tasks of LLMs
Scenarios
Real-world clinical cases and simulated patient scenarios
Quires
Real-time interactive queries and selected common quires
Output indicators
Word count, literacy, format, and principles
Individual Data
Sleep patterns
Heart rate
Blood pressure
Physical activity
Wearable devices
Large language models
General-Purpose LLMs
ChatGPT, DocsGPT, Google Bard, and Bing Chat
Fine-tuned LLMs
Aeyeconsult and ChatDiet
External Knowledge Base
Standards
Textbooks
Guideline documents
Retrieval augmented generation
Text chunks, embedding, and vector knowledge stores
Reducing healthcare system burden
Linking healthcare resources
Healthcare
Increasing knowledge
Promoting self-management behaviors
Enhancing emotional support
Individual
Opportunities of LLMs in chronic disease management
Community
Linking online supporting groups
Enhancing peer and community support
Feasibility
Acceptable in relevant and clarity of responses
Mixed performance in accuracy and actionability
Challenges
Potential privacy, language, and cultural issues
Incompetence in tackling advanced tasks
Gaps in generating personalised regimens
Implications
Developing robust data protection measures
Fine-tuning LLMs to enhance their capability
Integrating with wearables for personalised regimens

Forest plot: pooled accurate rate of LLMs (a), pooled accurate rate of fine-tuned LLMs compared to general-purpose LLMs (b), and pooled effect sizes of the Flesch-Kincaid Grade Level score (c) and the Flesch-Kincaid Reading Ease Score (d). LLM, large language model; 95% CI, 95% confidence interval.



Multimedia Appendixes

Supplementary material.
URL: http://asset.jmir.pub/assets/62ef4a979f11dffa53e9d5dfd5575a6a.docx