REVIEW ARTICLE

Generative Artificial Intelligence and Large Language Model Applications for Non-communicable Disease Prevention, Screening, and Self-Management: A Scoping Review

Vyshnavi Reddy Adama [1], Hari Prakash G. [1], Tejaswini B. Darukaradhya [1], Vanishri Arun [2], Anush Nayak [3]

[1] Department of Public Health, Ramaiah School of Public Health, MS Ramaiah University of Applied Sciences, Bengaluru, India
[2] Department of Computer Science and Engineering, B.M.S. College of Engineering, Bangalore
[3] Division of optometry and department of Allied Health Sciences, MS Ramaiah University of Applied Sciences, Bengaluru, India

Corresponding author:
Dr. Hari Prakash,
Department of Public Health, Ramaiah School of Public Health, MS Ramaiah University of Applied Sciences, Bengaluru, India
Email: hariprakash607@gmail.com
https://orcid.org/0000-0002-9463-4084

Cite this article:
Adama VR, Prakash GH, Darukaradhya TB, Arun V, Nayak A. Generative Artificial Intelligence and Large Language Model Applications for Non-communicable Disease Prevention, Screening, and Self-Management: A Scoping Review. Int Rev Lifestyle Med. 2026;1(2):65-78.
https://doi.org/10.22034/irlm.2026.601721.1016

Article history
Received: 09 Aug 2026
Revised: 09 Sep 2026
Accepted: 19 Sep 2026

Abstract

Background and Objective: Non-communicable diseases (NCDs) account for approximately 75% of global deaths, and most premature NCD deaths occur in low- and middle-income countries (LMICs), where shortages of trained health workers remain a major obstacle. Generative AI and large language models (LLMs) could offer scalable support for NCD prevention, screening, and self-management, but no systematic mapping has been done of how they have been applied in NCD care. This review aimed to map the extent, range, and nature of the literature on generative AI/LLM tools for NCD prevention, screening, and self-management, and how that evidence is distributed across geographic and income settings.

Methods: This review followed Joanna Briggs Institute methodology and PRISMA-SCRA guidelines. We searched MEDLINE (via PubMed) and Scopus for studies published from 2022 onward. Two reviewers independently screened records and charted data, resolving disagreements through discussion. We synthesized the findings narratively, mapping networks that linked studies, NCD domains, and AI/LLM tools.

Results: 821 records were identified, of which 10 met the inclusion criteria. Cancer was the domain addressed most often (7/10), followed by cardiovascular disease and hypertension (3/10). Proprietary conversational tools, and most commonly ChatGPT, were evaluated and compared against other chatbots rather than against clinical standards. Accuracy and readability differed from study to study. One study was set in an LMIC; the rest were based in high-income settings or not in a specific country.

Conclusion: Evidence on generative AI/LLM applications for NCD-related patient information is still concentrated in high-income settings. Locally grounded research in LMICs is needed to test their effectiveness and contextual relevance and to guide responsible use in NCD prevention.

Keywords: artificial intelligence; large language models; generative AI; noncommunicable diseases; digital health; low- and middle-income countries; scoping review.

Copyright © 2026 International Review of Lifestyle Medicine at https://www.irlmjournal.ir. This work is licensed under a Creative Commons Attribution-Non-Commercial 4.0 International license (https://creativecommons.org/licenses/by-nc/4.0/). Non-commercial uses of the work are permitted, provided the original work is properly cited.

1. Introduction

Non-communicable diseases (NCDs), chiefly cardiovascular disease, cancer, chronic respiratory disease, and diabetes, killed at least 43 million people in 2021, which is roughly three-quarters of all deaths not related to the pandemic (1). Of these, 18 million deaths occurred before the age of 70, and 82% of that premature mortality was in low- and middle-income countries (LMICs) (1). This uneven burden sits alongside a severe and worsening shortage of trained health workers in the very same settings. Global projections point to a net shortfall of 15 million health workers by 2030, concentrated in low- and lower-middle-income countries, where population growth and rising demand linked to NCDs are outpacing the supply of trained clinical and community health personnel (2).

Digital health tools, including mobile health applications, SMS-based programs, and telemedicine, have been positioned by global health agencies as scalable mechanisms to extend NCD prevention, screening, and self-management into places where in-person clinical capacity is limited. Generative AI and large language models, which emerged in 2022, mark a qualitatively different step within this wider digital health landscape. Earlier rule-based chatbots follow fixed decision trees; LLMs instead produce open-ended, context-sensitive responses, which raises the prospect of scalable, individually tailored health communication at a cost and speed the older tools could not reach. That same capability brings a safety concern. LLMs are prone to hallucination, meaning they can generate fluent but factually unsupported or fabricated content, and this has direct implications for patient safety when the tools are used to answer health questions (3).

Two earlier reviews touch on this area but do not cover the specific intersection addressed here. A 2022 scoping review of AI applications for strengthening health systems in LMICs found only ten eligible studies across all types of AI application; that search predated the widespread availability of generative AI and LLM-based tools and did not focus on NCDs (4). A 2024 scoping review looked at AI-powered conversational agents, including chatbots, voicebots, and digital avatars, for the remote management of NCDs and drew on 43 included studies, but it was not limited to LMIC settings and predates the rapid growth of publicly available LLM tools since 2023 (5). Neither review maps the use, evaluation, and reported safety of generative AI/LLM-specific tools for NCD prevention, screening, or self-management in LMIC settings specifically.

Publication in this field is moving quickly, and the topic bears directly on the workforce-constrained health systems that carry the highest NCD burden. For these reasons, we undertook this scoping review to map the current evidence on generative AI/LLM tools for NCD prevention, screening, and self-management, to identify gaps in methods and reporting, and to inform future primary research, implementation planning, and policy. Because of the burden and workforce issues set out above, we paid particular attention to the geographic and income-setting distribution of the evidence rather than limiting eligibility to a single setting. The review question was: what does the existing literature show about the development, application, and evaluation of generative AI/LLM tools for NCD prevention, screening, and self-management, and how is that evidence spread across geographic and income settings?

2. Methods

2.1 Protocol and Registration

The review was carried out using the methodological framework for scoping reviews first proposed by Arksey and O'Malley (6) and later refined by Levac, Colquhoun, and O'Brien (7), and together with the Joanna Briggs Institute (JBI)

methodology for scoping reviews (8). Reporting follows the Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR) checklist (9) . We developed the protocol a priori, setting out the review question, eligibility criteria, search strategy, and data-charting plan before screening began; it is available from the corresponding author on request.

2.2 Eligibility Criteria

We defined eligibility using the Population, Concept, and Context (PCC) framework recommended for scoping reviews.

Element	Criteria
Population	Any population (general public, patients with or at risk of NCDs, community health workers, or clinicians), in any country or income setting. Country and World Bank income classification were charted as data items to enable geographic analysis, rather than applied as an inclusion filter.
Concept	Generative AI and/or LLM-based tools, including conversational agents, chatbots, symptom checkers, and clinical decision-support systems built on generative architectures, applied to NCD prevention, screening, or self-management. Non-generative, rule-based systems were excluded unless directly compared against a generative system within the same study.
Context	Any NCD domain (cardiovascular disease, diabetes, cancer, chronic respiratory disease); any care setting (community, primary care, remote or telehealth, online information-seeking); any country or income setting.

Eligible sources included primary studies of any design, study protocols, pilot or feasibility studies, implementation reports, and relevant grey literature. We excluded editorials and opinion pieces without original data, along with non-English-language records for which no translation was available. We also excluded studies of AI tools that were neither generative nor conversational, and studies dealing solely with infectious disease, maternal and child health, or mental health, unless these were framed explicitly as an NCD risk factor. No lower date limit was set, although we expected most eligible work to post-date 2022, reflecting the arrival of clinically applied generative AI.

2.3 Information Sources

We searched two bibliographic databases, MEDLINE (via PubMed) and Scopus, chosen together to cover the core biomedical literature as well as the wider multidisciplinary and informatics literature relevant to generative AI.

2.4 Search Strategy

The search combined three concept blocks: generative AI/LLM terms, NCD terms, and LMIC terms, joined by Boolean AND between blocks and OR within each block, with truncation to catch term variants. Terms in blocks 1 and 2 were limited to the title field, and block 3 was searched across title and abstract. We restricted the search to records published from 2022 onward to keep precision manageable given the size of the underlying literature.

PubMed (MEDLINE)

("large language model*"[ti] OR "LLM"[ti] OR "LLMs"[ti] OR "ChatGPT"[ti] OR "GPT-4"[ti] OR "GPT-3*"[ti] OR "generative AI"[ti] OR "generative artificial intelligence"[ti]) AND ("noncommunicable

disease*"[ti] OR "non-communicable disease*"[ti] OR "NCD"[ti] OR "NCDs"[ti] OR "diabetes"[ti] OR "cardiovascular disease*"[ti] OR "hypertension"[ti] OR "cancer"[ti] OR "chronic respiratory disease*"[ti] OR "COPD"[ti]) AND ("low- and middle-income country*"[tiab] OR "low and middle income country*"[tiab] OR "LMIC*"[tiab] OR "developing country*"[tiab] OR "resource-limited setting*"[tiab] OR "low-income country*"[tiab] OR "middle-income country*"[tiab])
Filters: 2022/01/01-present

Scopus

TITLE("large language model*" OR "LLM" OR "LLMs" OR "ChatGPT" OR "GPT-4" OR "GPT-3*" OR "generative AI" OR "generative artificial intelligence") AND TITLE("noncommunicable disease*" OR "non-communicable disease*" OR "NCD" OR "NCDs" OR "diabetes" OR "cardiovascular disease*" OR "hypertension" OR "cancer" OR "chronic respiratory disease*" OR "COPD") AND TITLE-ABS-KEY("low- and middle-income country*" OR "low and middle income country*" OR "LMIC*" OR "developing country*" OR "resource-limited setting*" OR "low-income country*" OR "middle-income country*") AND PUBYEAR > 2021 AND LIMIT-TO(LANGUAGE, "English")

The search returned 678 records from PubMed and 143 from Scopus (n = 821 in total). We kept the third concept block of LMIC-related terms in the search strings to prioritise geographically relevant literature at the retrieval stage, but we did not then apply country or income setting as a strict eligibility filter during screening; eligible studies were kept regardless of country, in line with the broadened population criterion in section 2.2. We return to this design choice, and what it means for the geographic make-up of the included sample, in the Discussion.

2.5 Study Selection Process

Records from both databases were exported to the PRAMANA platform and de-duplicated, which removed 15 duplicates and left 806 unique records for title and abstract screening. Two reviewers screened titles and abstracts independently against the eligibility criteria, and 753 records were excluded as clearly ineligible, leaving 36 for full-text assessment. At full-text, 26 records were excluded: other reasons (n = 11), wrong population (n = 7), wrong outcome (n = 5), and wrong intervention (n = 3). Ten studies met all criteria and were carried forward for data charting and synthesis. Any disagreement at either stage was resolved by discussion. Figure 1 summarizes the selection process.

2.6 Data Charting Process

A standardized, piloted data-charting form was used to extract information from each included study. Two reviewers charted the data independently and cross-checked their entries for consistency, resolving discrepancies by re-examining the source article. When a variable was not reported in, or could not be extracted from, the available text of a study, we coded it as not reported (NR) rather than leaving it blank.

2.7 Data Items

For each included study, we charted the following: first author and year; country and setting; World Bank income classification; NCD domain; care phase; target population; generative AI/LLM tool type; underlying model or architecture; comparator; primary outcome measure(s); safety and harm outcomes; sample size or evaluation scale; study design; and funding source.

2.8 Synthesis of Results

Because we anticipated substantial heterogeneity in study designs and outcome measures, we did not undertake a formal risk-of-bias appraisal or meta-analysis, which is consistent with standard scoping review methodology. Findings were synthesized narratively and set out in tables. We also built a network analysis linking each included study to the NCD domain it addressed and to the generative AI/LLM tool(s) it evaluated, to show how the evidence clusters and to reveal which tools and domains have drawn the most, and the least, empirical attention.**3.**

Results

3.1 Study Selection

The database search identified 821 records (PubMed, n = 678; Scopus, n = 143). After 15 duplicates were removed, 806 records were screened at the title and abstract level, and 753 were excluded as ineligible. Of the 36 records assessed in full text, 26 were excluded (other reasons, n = 11; wrong population, n = 7; wrong outcome, n = 5; wrong intervention, n = 3), which left 10 studies for inclusion. Nothing was pooled quantitatively, in keeping with the descriptive, evidence-mapping aim of a scoping review. Figure 1 presents the selection process.

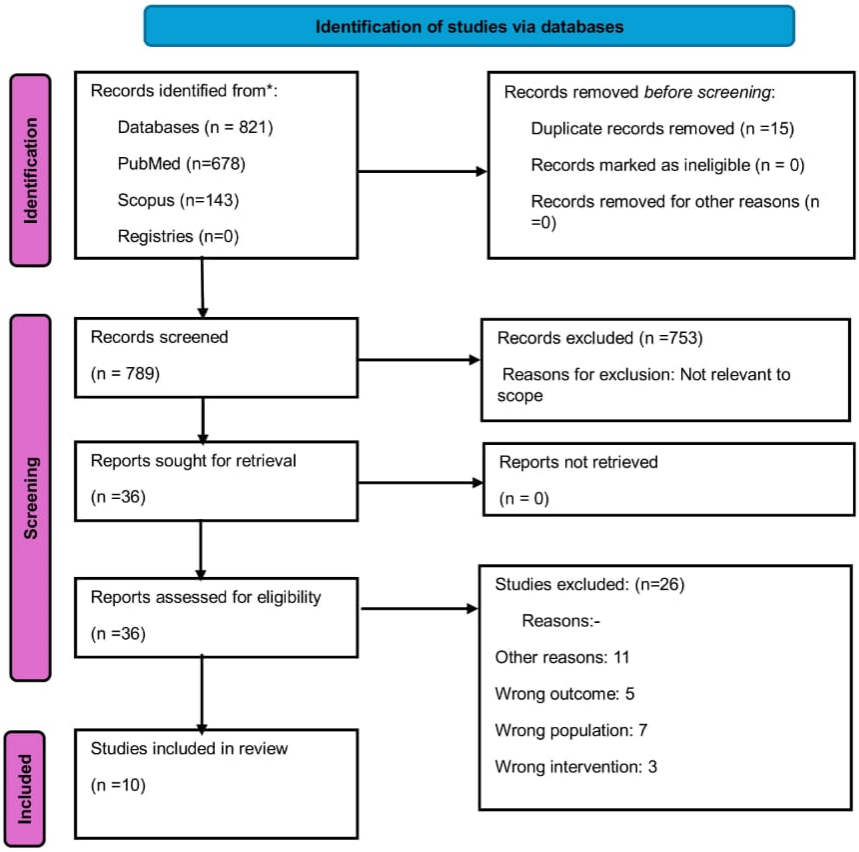


Figure 1. PRISMA-ScR flow diagram of study identification, screening, and inclusion

3.2 Characteristics of Included Studies

Ten studies published between 2024 and 2026 met the inclusion criteria (Table 1). Cancer was the NCD domain addressed most often, covering seven of the ten studies and spanning prostate cancer (10,11) , head and neck cancer (12), radionuclide-treated cancer (13), lung cancer (14), bladder cancer (15), and breast cancer (16). Three studies addressed cardiovascular disease and hypertension (17,18,19). None addressed diabetes or chronic respiratory disease as a primary domain. One study looked at idiopathic intracranial hypertension (19), a neurological condition at the edge of the review's core NCD domains, which we retained as charted.

Patient education and health information seeking was the dominant care phase, reported in nine of ten studies (10-16,19); screening-related evaluation was reported in two studies (14,17). No study addressed self-

management support as a primary focus. All ten evaluated tools were intended for patients, caregivers, or the general public. Clinicians were an explicit target user in only one study, and none involved community health workers as intended end users (18), and no study involved community health workers as intended end users.

Table 1. Characteristics of Included Studies (n = 10)

Study (Ref.), Year	Country / Income Classification	NCD Domain	Care Phase	Target Population	Study Design
Luo et al. (10), 2025	China (upper-middle-income); South Korea (high-income) and Mixed contexts	Cancer, Prostate	Patient education	Patients (older adults); general public	Cross-sectional comparative study
Belge Bilgin et al. (11), 2024	USA (Mayo Clinic), High-income	Cancer, Prostate	Patient education	Patients	Cross-sectional expert-rated study
Wei et al. (12), 2024	USA, High-income	Cancer, Head and neck	Information seeking	General public (online information seekers)	Cross-sectional comparative evaluation
Şan et al. (13), 2024	Not Reported	Cancer, Radionuclide treatment	Information seeking / self-education	Cancer patients and caregivers	Cross-sectional study
Nguyen et al. (14), 2025	USA, High-income	Cancer, Lung	Prevention, screening, terminology	General public; patients	Cross-sectional assessment
Patel et al. (15), 2025	United Kingdom, High-income	Cancer, Bladder	Diagnosis / treatment information	Patients (general public)	Comparative cross-sectional evaluation
Haver et al. (16), 2024	USA, High-income	Cancer, Breast	Prevention, screening	General public; patients	Retrospective exploratory study
Wang et al.(17), 2026	China (multi-center hospital), Upper-middle-income (LMIC)	CVD, Hypertension	Screening, patient education	Patients with suspected hypertension; general public	Two-phase retrospective + prospective validation
Lee et al. (18) 2024	USA, High-income	Hypertension	Patient education	Patients; clinicians	Cross-sectional study
Dihan et al. (19), 2025	USA, High-income	Idiopathic intracranial hypertension	Patient education	Patients	Cross-sectional study

3.3 Generative AI/LLM Tool Characteristics

All ten included studies evaluated proprietary, general-purpose conversational LLMs rather than purpose-built clinical tools. The only architectural modification reported was in Wang et al., who layered a custom Retrieval-Augmented Generation pipeline over foundational models (17). ChatGPT, across versions 3.5 to 4o, was evaluated in all ten studies, most frequently benchmarked against Google Gemini or Bard (11,13,18,19,14,15), DeepSeek (10,17,14,15), Claude (14,15), Grok (15), and Perplexity (14). Comparisons were

mostly between models rather than between a model and a clinical standard; only three studies tested LLM output against a non-AI reference, namely practicing physicians (17), a standard web search engine (12), or existing published patient education materials (19).

Table 2. Generative AI/LLM Tool Characteristics and Comparators

Study (Ref.)	Tool Type	Underlying Model(s)	Comparator	Evaluation Scale
Luo et al. (10)	Chatbot	ChatGPT-4.0, DeepSeek-R1	Inter-model comparison	NR
Belge Bilgin et al. (11)	Chatbot	ChatGPT-4, Google Bard	Inter-model comparison	8 expert raters (4 radiologists, 4 oncologists)
Wei et al. (12)	Chatbot	ChatGPT-4	Top Google search result	49 questions; 2 independent reviewers
Şan et al. (13)	Chatbot	GPT-4, Bard	Inter-model comparison	1 expert rater (8 years' experience)
Nguyen et al. (14)	Chatbot	GPT-o3-mini, DeepSeek R1, Perplexity, Claude 3.5 Sonnet, Gemini 2.0 Pro	Claude 3.5 Sonnet as reference model	40 questions × 5 LLMs = 200 responses; 3 expert radiologists (600 ratings)
Patel et al. (15)	Chatbot	ChatGPT-4o, Gemini 2.0 Flash, Grok 3, Claude Sonnet 3.7, DeepSeek R1	Inter-model comparison	10 questions; 5 chatbots; 50 answers; 2 urologist reviewers
Haver et al. (16)	Chatbot	ChatGPT (GPT-3.5)	Original vs simplified chatbot output	25 questions; 75 original + 225 simplified responses
Wang et al. (17)	Custom RAG-enhanced chatbot pipeline	ChatGPT-4o, ChatGPT-4o Mini, DeepSeek-V3 (+ Cascade architecture)	Baseline (unenhanced) LLMs; 3 practicing physicians	6 LLM configurations (Phase 1 benchmarking); 107 patients (Phase 2 external validation)
Lee et al. (18)	Chatbot	ChatGPT-3.5, Gemini 1.0	Inter-model comparison vs ACC reference answers	52 questions
Dihan et al. (19)	Chatbot	ChatGPT-3.5, ChatGPT-4, Google Bard	Original published patient education materials	20 source materials; LLM-generated/rewritten versions

3.4 Outcomes, Safety Findings, and Key Results

Accuracy and readability were the outcome domains reported most consistently. Reported error or inaccuracy rates varied widely: one study rated 8.2% of responses incorrect (12); another found 16.6% to 29.1% incorrect or partially correct, depending on the model compared (11); a third reported that overall inaccuracy was rare, at 5.5% (18); and one recorded zero hallucinations across 50 evaluated answers, with perfect inter-rater agreement (15). When an LLM was compared with a non-AI reference, results diverged by domain. An enhanced ChatGPT-4o model reached 95.3% accuracy in blood pressure classification, ahead of three practicing physicians at 74.8% to 84.1% (17) , whereas for cancer-related information a standard web search was rated significantly higher in quality than ChatGPT-4 output (4.2 versus 3.6, $p = 0.005$) (12).

Readability came up repeatedly as a barrier to real-world use. Source materials and LLM output alike often exceeded recommended public health literacy reading levels. Only one study's best-performing model reached the target sixth-grade level without losing quality (19), and in another study explicit simplification prompting moved ChatGPT output from a thirteenth-grade to an 8.9-grade reading level, though only 8% of the simplified responses met the specific sixth-grade target requested (16)[6].

Table 3. Outcomes and Safety/Harm Findings

Study (Ref.)	Primary Outcome Measure(s)	Safety/Harm Outcome	Key Finding (Summary)
Luo et al.(10)	Quality (DISCERN), understandability/actionability (PEMAT-P), readability	NR	ChatGPT-4.0 and DeepSeek-R1 produced comparable good-quality prostate cancer information with no significant inter-model difference
Belge Bilgin et al. (11)	Accuracy, conciseness, readability, misleading-information detection	16.6% (ChatGPT-4) vs 29.1% (Bard) responses incorrect/partially correct	ChatGPT-4 more accurate than Bard; Bard more readable but more prone to misleading information
Wei et al. (12)	Accuracy, completeness, reliability, readability	8.2% (4/49) of ChatGPT responses incorrect	Google search results rated significantly higher quality than ChatGPT-4; errors concentrated in ICD-10 coding content
Şan et al. (13)	Reliability (DISCERN), readability (FRE, FKRGL)	NR (DISCERN quantified: GPT-4 46.8-48.5; Bard 49.4-53.4; no explicit hallucination metric)	Bard outperformed GPT-4 on reliability and readability across all four treatment types; both exceeded recommended public reading level; Bard cited references and recommended physician consultation more often
Nguyen et al. (14)	Accuracy, consistency, reliability	Incorrect-response rate lowest for GPT-o3-mini (2.5%); refusal rate highest for GPT-o3-mini (8.3%)	GPT-o3-mini highest accuracy (75.8%), then Gemini 2.0 Pro (74.2%), DeepSeek R1 and Perplexity (72.5% each), Claude 3.5 Sonnet (70.0%, reference); DeepSeek R1 showed strongest inter-rater agreement
Patel et al. (15)	Readability, quality/reliability (DISCERN)	Zero hallucinations detected (0/50)	ChatGPT-4o best for readability; Grok 3 highest DISCERN score; Claude 3.7 Sonnet least readable
Haver et al. (16)	Readability (Flesch Reading Ease and grade level), clinical appropriateness	92% (23/25) of simplified responses judged clinically appropriate	All 25 simplified responses reached adequate reading ease (versus 2/25 originals), but only 2/25 (8%) reached the specific sixth-grade target; original responses averaged grade 13
Wang et al. (17)	Accuracy (BP classification, risk stratification, clinical decision-making), comprehensiveness, credibility	NR	Cascade-enhanced ChatGPT-4o reached 95.3% BP classification, 90.7% risk stratification, and 92.5% clinical decision-making accuracy in external validation, exceeding 3 physicians (74.8% to 84.1%; all $p<0.001$); baseline ChatGPT-4o BP classification accuracy rose from 62.5% to 87.0% with Cascade enhancement

Study (Ref.)	Primary Outcome Measure(s)	Safety/Harm Outcome	Key Finding (Summary)
Lee et al. (18)	Accuracy, readability, word count	5.5% overall inaccuracy rate (both models combined)	Both chatbots highly accurate; Gemini responses longer, ChatGPT responses at a higher reading grade level
Dihan et al. (19)	Readability, quality (DISCERN), understandability/actionability (PEMAT), accuracy (misinformation scale)	No misinformation detected (Likert = 1) across 120 AI-generated materials	Only ChatGPT-4 reached the target sixth-grade reading level without loss of quality or understandability; none of the newly generated materials met the actionability threshold on PEMAT

3.5 Network Analysis of Studies, NCD Domains, and Tools

To show how the evidence base clusters, we built a network linking each included study to the NCD domain it addressed and to the generative AI/LLM tool(s) it evaluated (Figure 2). The network held 24 nodes (10 studies, 8 NCD domains, 6 distinct tools) joined by 34 edges. ChatGPT was the most connected node, with a degree of 10, since every included study evaluated it, followed by Gemini/Bard (degree 6), DeepSeek (degree 4), Claude (degree 2), and Grok and Perplexity (degree 1 each). Among the NCD domains, hypertension/cardiovascular disease and prostate cancer were each linked to two studies, while the other six domains, idiopathic intracranial hypertension, head and neck cancer, radionuclide-treated cancer, lung cancer, bladder cancer, and breast cancer, were each linked to a single study. At the study level, Nguyen et al. and Patel et al. shared the highest degree, at 6 each, reflecting their five-way tool comparison designs.

This structure points to a shallow rather than a deep evidence base. The apparent breadth of ten studies collapses, on closer inspection, into a small set of proprietary tools, dominated by a single vendor's product, spread across a scattered and largely non-overlapping group of single-study NCD domains. Apart from the prostate cancer and hypertension pairs, no two studies addressed the same disease and tool combination as an independent replication, so most of the findings in Table 3 currently rest on one study each.

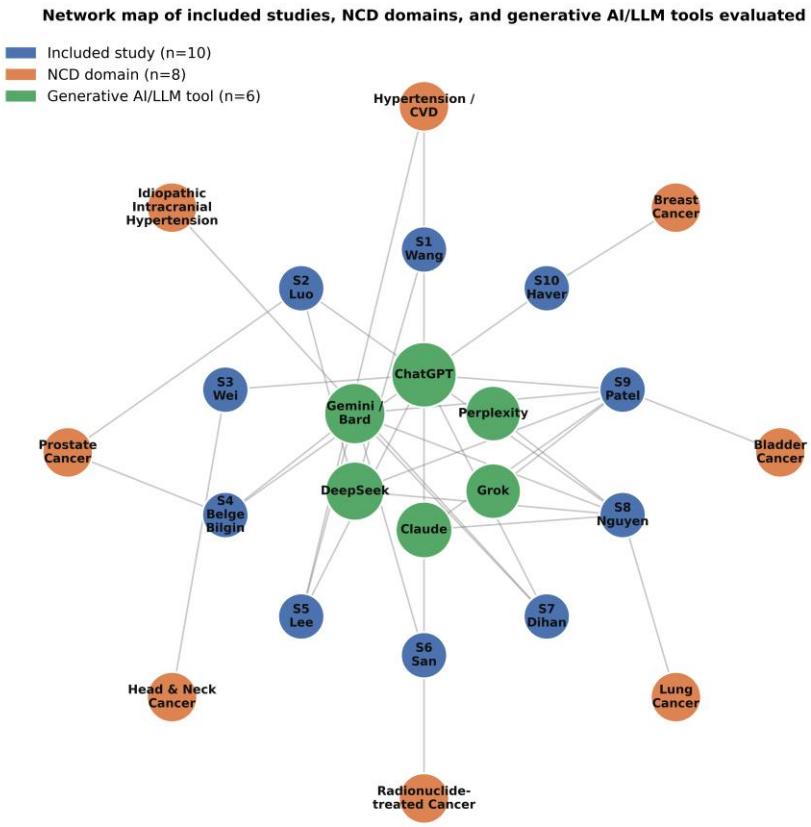


Figure 2. Network map of included studies, NCD domains, and generative AI/LLM tools evaluated

3.6 Geographic and Income-Setting Distribution of Included Studies

A further, important finding concerns how the included evidence is distributed by geography and income setting. Of the ten studies, one was conducted entirely in a setting classifiable as an LMIC, China, an upper-middle-income country (17), and one more was multinational, spanning China and South Korea, an upper-middle-income and a high-income country, respectively (10) . Seven were conducted only in high-income settings: the United States (six studies) (11,12,16,18,19,14) and the United Kingdom (one study) (15). The country of study was not reported or not extractable for the remaining study (13). Taken together, at most two of the ten studies had any LMIC link, and only one was conducted entirely in an LMIC setting. No study reported barriers or facilitators to deployment in resource-limited settings, and none reported ethical, regulatory, or data-governance considerations specific to such settings; both fields were coded as not reported across the whole dataset.

4. Discussion

This scoping review mapped ten studies evaluating generative AI and large language model tools for information related to non-communicable disease prevention, screening, or self-management. The picture that emerges is of a young, narrow, and geographically skewed evidence base, not the broad foundation that the scale of proposed deployment would seem to demand. Three findings stand out and are taken in turn: the evidence clusters around a single commercial tool, it is markedly uneven across NCD domains and outcome types, and it almost entirely omits low- and middle-income countries, the settings most burdened by NCDs and most constrained in health workforce.

The dominance of one vendor's product, present in every included study, mirrors the wider pattern that Anisha and colleagues reported in their 2024 review of AI conversational agents for NCD management, where a small number of commercially available tools saw rapid clinical and public uptake ahead of rigorous, setting-specific evaluation (5). This is not a minor methodological point. The safety findings in Table 3 run from zero hallucinations in one study to a 29.1% incorrect-or-partially-correct response rate in another (15,11), which fits the general caution Azamfirei and colleagues raised about the tendency of LLMs to produce fluent but factually unsupported content (3), Together these results suggest that accuracy is highly dependent on task and domain rather than a fixed property of a given model or model family. A tool that seems acceptable for one NCD domain or question type cannot be assumed acceptable for another without domain-specific evaluation, a point this review illustrates directly through its own contrast between head and neck cancer and hypertension information, where the same underlying model family performed very differently (12,17).

The clustering of studies in cancer-related patient education, together with the complete absence of any study addressing diabetes or chronic respiratory disease as a primary domain, is another imbalance worth noting. It plausibly reflects how much easier it is to generate publicly answerable, non-individualized patient education content for cancer topics than to support the longitudinal, individualized self-management that diabetes and chronic respiratory disease usually demand. That kind of support is harder to evaluate with a single-session question-and-answer design and is arguably more clinically consequential, since it turns on sustained behavior change rather than one-off information.

The most consequential finding, though, is how nearly absent low- and middle-income settings are from the eligible evidence, even though these settings carry the larger share of the NCD burden and workforce constraint described in the Introduction. At most two of the ten studies had any LMIC link, and only one was conducted entirely in an LMIC setting (17). This is striking given that generative AI tools are widely proposed as a scalable answer to exactly the health workforce shortages concentrated in low- and lower-middle-income countries (2), and that the World Health Organization's own Global Strategy on Digital Health frames digital health, AI-enabled tools included, as a means of advancing universal health coverage in precisely these settings (20). The gap extends rather than resolves the concern raised in 2022 by Ciecierski-Holmes and colleagues, whose scoping review of AI for strengthening LMIC health systems found only ten eligible studies across all AI application types combined and concluded that LMIC-specific evidence lagged well behind the claims being made for AI in these settings (4). More than two years after the release of the LLM tools evaluated here, that lag does not appear to have narrowed for the generative AI subset of this literature. The risk is not simply an absence of evidence but a live problem of generalizability. Yang and colleagues showed empirically that a clinical AI model trained and validated in a high-income country can perform substantially worse when applied without adaptation in a low- and middle-income country hospital, and needs explicit local recalibration before it can be deployed safely (21). There is no reason to think conversational LLMs, trained mostly on high-income, English-language health information, would be exempt from the same problem when applied to LMIC populations, languages, and health-literacy contexts, yet no included study tested this assumption directly.

Several limitations should temper how these findings are read. The search covered only two databases, MEDLINE and Scopus, and the current sample of ten studies may under-represent LMIC-based research indexed in

regionally focused, non-English-language, or grey literature; adding further databases is advisable before this pattern is treated as definitive. Relatedly, the search strings kept a third concept block of LMIC-related terms as a relevance signal even though country was not applied as a strict eligibility filter during screening. Because that block was not validated against a parallel search without it, we cannot tell from the present data whether it inadvertently narrowed retrieval of non-LMIC or unclassified literature or had little effect. Re-running the search without this block and comparing the record sets is advisable before the geographic distribution reported in section 3.6 is read as a precise estimate rather than an indicative pattern. One included study addressed a neurological condition, idiopathic intracranial hypertension (19), whose fit within the review's NCD scope warrants confirmation. Finally, publication in this field is rapid enough that the present search will date quickly relative to publication lag, so a living-review or periodic-update approach is advisable as the field develops.

Taken together, these findings suggest that before generative AI and LLM tools are built into NCD prevention or care pathways in LMICs, rather than merely proposed for that role, locally grounded evaluation is needed. Such evaluation should test the tools against the specific languages, health literacy levels, disease priorities, and health system contexts of the settings where they would actually be used, instead of relying on evidence generated in high-income settings and assuming it will carry over.

5. Conclusion

Generative AI and large language model tools for NCD prevention, screening, and self-management have been evaluated in a small but growing body of work that is concentrated in cancer-related patient education, dominated by a handful of proprietary conversational tools, and drawn almost entirely from high-income settings. This concentration sits in direct tension with the reason most often given for deploying these tools, namely their potential to extend NCD-related services in workforce-constrained low- and middle-income settings. Closing that evidence gap through locally grounded, LMIC-based evaluation should be a priority for future research.

Author Contributions (CRediT)

Vyshnavi A: Conceptualization, Investigation, Data curation, Formal analysis, Writing-original draft, Writing-review & editing; Vaishnavi Arun: Investigation, Data curation, Formal analysis, Writing-original draft; Anush Nayak: Visualization, Writing-review & editing; Hari Prakash G.: Formal analysis, Writing-review & editing; Tejaswini B. D.: Supervision, Writing-review & editing; all authors: Final approval of the manuscript.

Funding: The authors received no external funding for this study.

Conflicts of Interest: The authors declare no conflicts of interest.

Ethics: Not applicable.

References

1. World Health Organization. Noncommunicable diseases [Internet]. Geneva: WHO; 2025 [cited 2026 Aug]. Available from: https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases

2. Liu JX, Goryakin Y, Maeda A, Bruckner T, Scheffler R. Global health workforce labour market projections for 2030. Hum Resour Health. 2017; 15:11. doi:10.1186/s12960-017-0187-2

3. Azamfirei R, Kudchadkar SR, Fackler J. Large language models and the perils of their hallucinations. Crit Care. 2023; 27:120. doi:10.1186/s13054-023-04393-x

4. Ciecierski-Holmes T, Singh R, Axt M, Brenner S, Barteit S. Artificial intelligence for strengthening healthcare systems in low- and middle-income countries: a systematic scoping review. NPJ Digit Med. 2022; 5:162. doi:10.1038/s41746-022-00700-y

5. Anisha SA, Sen A, Bain C. Evaluating the potential and pitfalls of AI-powered conversational agents as humanlike virtual health carers in the remote management of noncommunicable diseases: scoping review. J Med Internet Res. 2024;26: e56114. doi:10.2196/56114

6. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. Int J Soc Res Methodol. 2005;8(1):19-32. doi:10.1080/1364557032000119616

7. Levac D, Colquhoun H, O'Brien KK. Scoping studies: advancing the methodology. Implement Sci. 2010; 5:69. doi:10.1186/1748-5908-5-69

8. Peters MDJ, Godfrey C, McInerney P, Munn Z, Tricco AC, Khalil H. Chapter 11: scoping reviews (2020 version). In: Aromataris E, Munn Z, editors. JBI Manual for Evidence Synthesis. Adelaide: JBI; 2020. doi:10.46658/JBIMES-20-12

9. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med. 2018;169(7):467-473. doi:10.7326/M18-0850

10. Luo Z, Lin C, Kim TH, Shin YS, Ahn ST. Quality and readability analysis of artificial intelligence-generated medical information related to prostate cancer: a cross-sectional study of ChatGPT and DeepSeek. World J Mens Health. 2025. doi:10.5534/wjmh.250144

11. Belge Bilgin G, Bilgin C, Childs DS, Orme JJ, Burkett BJ, Packard AT, et al. Performance of ChatGPT-4 and Bard chatbots in responding to common patient questions on prostate cancer 177Lu-PSMA-617 therapy. Front Oncol. 2024. doi:10.3389/fonc.2024.1386718

12. Wei K, Fritz C, Rajasekaran K. Answering head and neck cancer questions: an assessment of ChatGPT responses. Am J Otolaryngol. 2024. doi: 10.1016/j.amjoto.2023.104085

13. Şan H, Bayrakcı Ö, Çağdaş B, Serdengeçti M, Alagöz E. Reliability and readability analysis of ChatGPT-4 and Google Bard as a patient information source for the most commonly applied radionuclide treatments in cancer patients. Rev Esp Med Nucl Imagen Mol (Engl Ed). 2024. doi: 10.1016/j.remnie.2024.500021

14. Nguyen D, Rahsepar AA, Chalian H, Hassani C, Kim GHJ, Bedayat A. Revisiting LLMs and lung cancer questions: how AI responds to common lung cancer questions two years later. Acad Radiol. 2025. doi: 10.1016/j.acra.2025.08.056

15. Patel K, Radcliffe R. Evaluating the readability and quality of bladder cancer information from AI chatbots: a comparative study between ChatGPT, Google Gemini, Grok, Claude and DeepSeek. J Clin Med. 2025. doi:10.3390/jcm14217804

16. Haver HL, Gupta AK, Ambinder EB, Bahl M, Oluyemi ET, Jeudy J, et al. Evaluating the use of ChatGPT to accurately simplify patient-centred information about breast cancer prevention and screening. Radiol Imaging Cancer. 2024. doi:10.1148/rycan.230086

17. Wang Y, Tan W, Cheng S, Peng C, Peng J, Qin F, et al. Large language model agent for managing patients with suspected hypertension. Hypertension. 2026. doi:10.1161/HYPERTENSIONAHA.125.25305

18. Lee TJ, Campbell DJ, Patel S, Hossain A, Radfar N, Siddiqui E, et al. Unlocking health literacy: the ultimate guide to hypertension education from ChatGPT versus Google Gemini. Cureus. 2024. doi:10.7759/cureus.59898

19. Dihan QA, Brown AD, Zaldivar AT, Chauhan MZ, Eleiwa TK, Hassan AK, et al. Advancing patient education in idiopathic intracranial hypertension: the promise of large language models. Neurol Clin Pract. 2025. doi:10.1212/CPJ.0000000000200366

20. World Health Organization. Global strategy on digital health 2020-2025. Geneva: WHO; 2021.

21. Yang J, Dung NT, Thach PN, Phong NT, Phu VD, Phu KD, et al. Generalizability assessment of AI models across hospitals in low-, middle-, and high-income countries. Nat Commun. 2024; 15:8270. doi:10.1038/s41467-024-52618-6