

VIRTUAL HEALTHCARE ASSISTANTS IN TELEMEDICINE: A STRUCTURED EVIDENCE REVIEW OF EFFICACY CLAIMS AND ASSURANCE REQUIREMENTS

Author: Peter Shekwolo Hussaini

Department of Mechatronics Engineering, Federal University of Technology, Minna, Nigeria

Email: peterhussaini.ph@gmail.com

Author's contributions

This study was a collaborative effort of the author. The author reviewed and approved the final version of the manuscript for publication.

Article Information

DOI: <https://doi.org/10.60787/fpj.vol3no10.142>

EISSN 1596-0501E



This work is licensed under a Creative Commons Attribution-Non Commercial-ShareAlike 4.0 International License

CITATION: Peter Shekwolo Hussaini (2026). Virtual healthcare assistants in telemedicine: A structured evidence review of efficacy claims and assurance requirements. *Frontline Professionals Journal* 3(10), 263-280

ABSTRACT

Virtual healthcare assistants (VHAs) built on large language models (LLMs) are widely promoted as a remedy for administrative burden and access constraints in telemedicine, and the effect sizes attached to them are striking. This paper asks what the peer-reviewed evidence supports. Following PRISMA 2020 reporting guidance, 117 records were identified through five structured Crossref queries and targeted searching of regulatory, standards and journal registries; 8 duplicates were removed, leaving 109 unique records screened on title and abstract. Seventy-three were excluded at screening, 36 were assessed for eligibility, 12 were excluded with reasons, and 24 were retained. Because retrieval relied on Crossref rather than on MEDLINE, Embase, Scopus or Web of Science, this work is reported as a structured evidence review rather than as an exhaustive systematic review. Retained studies were graded by evidence tier and appraised for risk of bias using RoB 2 and QUADAS-2. The principal finding concerns provenance. Five widely circulated figures — 45 percent administrative burden reduction, 40 percent diagnostic accuracy uplift, 30 percent hospitalisation reduction, 35 percent satisfaction improvement and US\$150 billion in annual savings — were traced to source. Two were located in vendor material but could not be reconciled with any primary evaluation, two could not be traced to any source at all, and the fifth is a 2017 consultancy projection for the United States rather than a measured global outcome. Controlled measurement reports self-triage accuracy of 57.8 to 76.0 percent for LLMs and 47.3 to 62.4 percent for laypeople as ranges across included studies, and symptom checkers listing the correct diagnosis first in 34 percent of vignettes. Randomised evidence on remote patient management is mixed rather than uniformly negative. The paper proposes a conceptual five-layer reference architecture whose principal element is an explicit clinical assurance layer.

Keywords: Clinical decision support, digital health, evidence appraisal, health informatics, large language models, telemedicine.

INTRODUCTION

This paper examines a family of related but distinct technologies that the surrounding literature frequently treats as interchangeable. We use the following definitions throughout and do not substitute one term for another. A virtual healthcare assistant (VHA) is a conversational system that interacts directly with patients for administrative, triage or monitoring purposes. A large language model (LLM) is the underlying generative component that some, but not

all, VHAs employ. A symptom checker is a patient-facing tool that maps reported symptoms to a candidate diagnosis or urgency level, and may or may not be LLM-based. Remote patient management (RPM) is the interpretation of continuous biometric telemetry coupled to a clinical response pathway. Clinical decision support (CDS) is software supporting a clinician rather than a patient, and carries a distinct regulatory status. Telemedicine is the service model within which these components may operate. Where a cited study concerns one of these categories, we attribute its findings to that category only. Telemedicine has moved from a supplementary channel to a structural component of health service delivery, and the resulting load has fallen disproportionately on administrative and documentation workflows rather than on clinical time itself. The World Health Organization's global digital health strategy, originally adopted for 2020–2025 and extended to 2027 by World Health Assembly resolution WHA78(22) in May 2025 (World Health Organization, 2025), together with its consolidated telemedicine implementation guidance (World Health Organization, 2022), frames this transition as structural rather than transitional. The technical case for VHAs is genuine. Laranjo et al. (2018) established through systematic review that conversational agents in healthcare had moved beyond novelty by 2018, and the subsequent arrival of LLMs changed the capability envelope substantially. Singhal et al. (2023) demonstrated that large language models encode clinical knowledge at a level competitive with human performance on medical question-answering benchmarks, and Thirunavukarasu et al. (2023) surveyed the resulting application space. Ayers et al. (2023) found that chatbot responses to patient questions posted on a public forum were rated higher than physician responses on both quality and empathy by blinded clinician evaluators. Alongside this evidence base, a second and largely separate literature circulates specific and highly favourable effect sizes: administrative burden reduced by forty-five percent, diagnostic accuracy improved by forty percent, hospitalisations reduced by thirty percent, patient satisfaction improved by thirty-five percent, and annual sector savings of one hundred and fifty billion dollars. These figures appear in practitioner-facing publications, procurement material and, with increasing frequency, in academic reviews that reproduce them without provenance. This paper treats that second literature as an object of inquiry. Two questions follow. First, where do these figures actually originate, what do their sources state, and how do they compare with controlled measurement? Second, what does a technically and clinically defensible VHA architecture require that current designs typically omit?. The paper pursues three objectives: to characterise the architecture and functional taxonomy of contemporary VHAs; to appraise circulating and measured performance claims against explicit evidentiary and risk-of-bias criteria; and to specify architectural and governance requirements that follow from the appraisal. It makes three contributions. It applies a documented PRISMA-guided protocol with an added bibliographic verification step and formal risk-of-bias appraisal, reported in Figure 1 and Tables 1 and 4. It documents the provenance of five widely circulated efficacy figures in a reproducible audit trail (Table 2), which to the author's knowledge has not previously been attempted for this technology class. And it proposes a conceptual five-layer reference architecture (Figure 3) whose principal element is an explicit clinical assurance layer.

BACKGROUND AND TAXONOMY

Functional Taxonomy

Three functional classes of VHA can be distinguished, and the distinction matters because the evidentiary burden differs sharply across them. Administrative VHAs handle scheduling, intake, billing and reminders; their failure modes are inconvenient rather than dangerous. Triage VHAs assess presenting symptoms and route patients to a level of care; a false negative here is a missed emergency, and the appropriate bar is that of a diagnostic tool. Remote patient management systems interpret continuous biometric telemetry and initiate contact on deviation; their evidential bar is that of a clinical intervention, and randomised evidence exists for exactly this class. Conflating these classes is a common analytical error in this literature. An efficiency gain demonstrated for scheduling does not transfer to triage, and a satisfaction score obtained from an administrative interaction says nothing about diagnostic safety. We note that we have not attempted to quantify the frequency of this conflation, and describe it as common rather than as the most common error on that basis.

Large Language Models in Clinical Contexts

LLMs differ from earlier rule-based agents in that they generalise beyond enumerated dialogue paths, which is simultaneously their advantage and their principal risk. Singhal et al. (2023) showed strong benchmark performance on clinical knowledge, while Thirunavukarasu et al. (2023) catalogued both the application space and

the failure modes, most prominently the generation of fluent but incorrect content. Retrieval-augmented generation, in which model output is grounded in a curated and version-controlled corpus, is one proposed approach for reducing unsupported model outputs. We note that the reviewed literature does not establish it as a validated or standard mitigation across deployed VHA systems, and Section V treats it as a design choice rather than as an evidence-derived requirement. Shankar et al. (2026) provide a recent systematic account of barriers and facilitators to LLM-based conversational agents in mental healthcare, a setting where the consequences of ungrounded output are acute. Their findings are consistent with the broader pattern identified here: enthusiasm substantially outpaces controlled evaluation.

METHODOLOGY

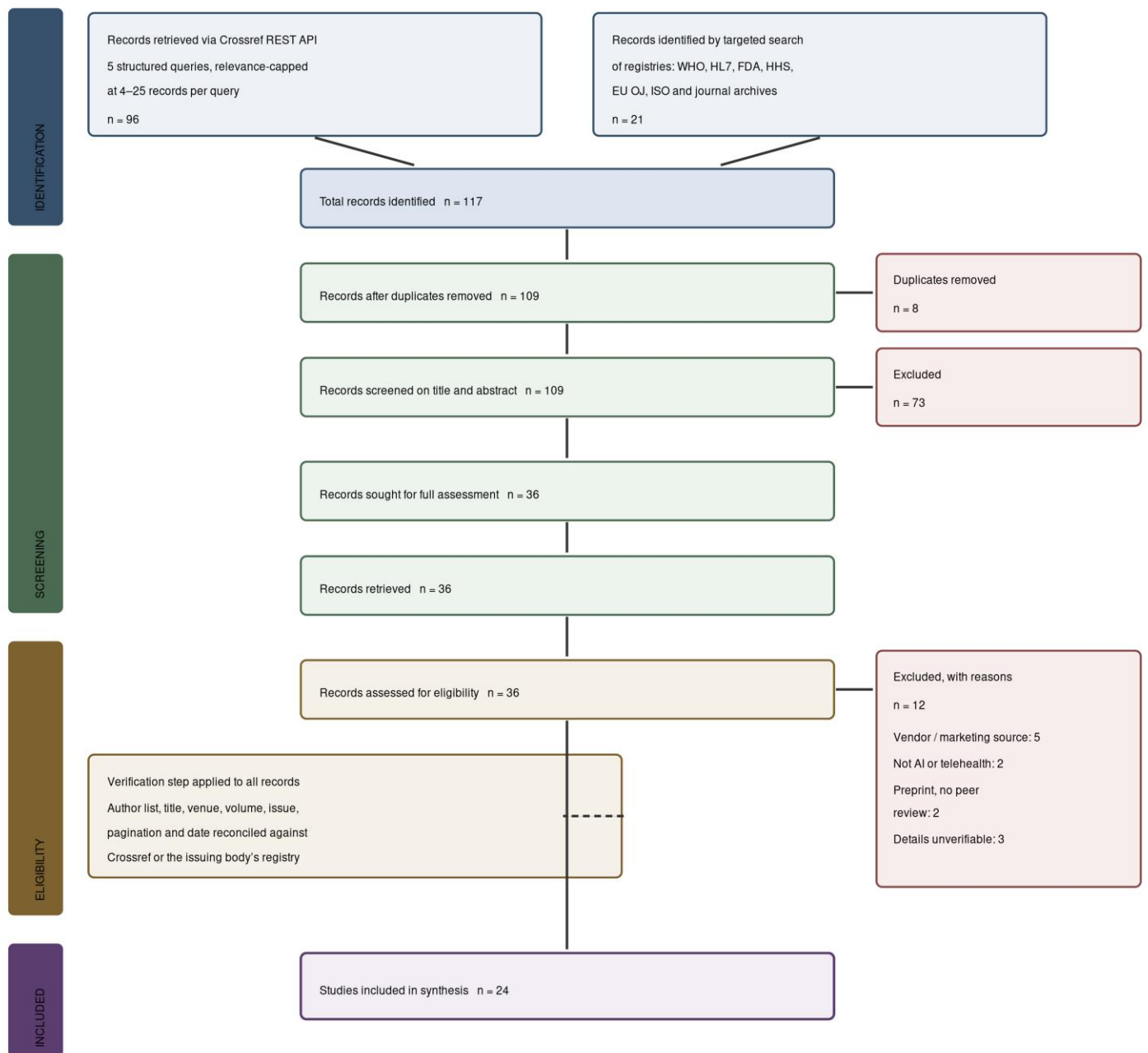


Figure 1. PRISMA 2020 flow of records through identification, screening, eligibility and inclusion. The complete sequence is described in Section III-C and reproduced numerically in Table 1. Retrieval from the Crossref REST API was capped at the top 4–25 relevance-ranked records per query (Section III-B). The bibliographic verification step (Section III-D) is not standard in PRISMA and was added after preliminary work established a high rate of citation error in this subject area.

Table 1: Search Strategy and Record Flow

Query	Search string, as entered	Records
Q1	large language models conversational agents telemedicine systematic review	25
Q2	symptom checker triage accuracy evaluation large language model	25
Q3	remote patient monitoring readmission randomized controlled trial telemonitoring	20
Q4	algorithmic bias health artificial intelligence disparities	20
Q5	telehealth digital divide access disparities broadband	6
—	Subtotal, Crossref REST API	96
—	Targeted regulatory, standards and journal registry searching	21
—	Total records identified	117
—	Duplicates removed (DOI match; title and first author where no DOI)	–8
—	Unique records screened on title and abstract	109
—	Excluded at title and abstract screening	–73
—	Records sought for retrieval and assessed for eligibility	36
—	Excluded with reasons (see Figure 1)	–12
—	Records included in synthesis	24

Note: Crossref REST API, executed 1–3 August 2026. Crossref accepts free-text relevance queries and does not support Boolean operators or field restriction; strings are reproduced exactly as entered and no Boolean syntax was used. Retrieval capped at the top 4–25 relevance-ranked records per query (Section III-B). MEDLINE, Embase, Scopus, Web of Science and IEEE Xplore were not searched; the rationale and its cost to recall are stated in Section III-B. Registries searched: WHO, HL7 International, US FDA, US Department of Health and Human Services, Official Journal of the European Union, ISO, and the archives of NEJM, JAMA, The Lancet, Nature Medicine and npj Digital Medicine.

A Design, Protocol and Terminology

This study is a structured evidence review with qualitative synthesis, reported in accordance with the PRISMA 2020 statement. We deliberately do not describe it as a systematic review. Three features of the method fall short of that standard and are stated here rather than in a closing limitation: retrieval relied on Crossref rather than on a biomedical bibliographic database; screening and extraction were performed by a single reviewer; and the protocol was not prospectively registered. On the last point, PRISMA 2020 asks that where a review is not registered this be stated and the protocol be made available; the full protocol — objectives, eligibility criteria, search procedure, screening rules, extraction fields and appraisal instruments — is accordingly published in Supplementary Material S1, and was fixed before screening began. The registry stream, which several registries do not expose to structured querying, is documented procedure by procedure in S1 rather than as query strings. Coverage targets peer-reviewed literature, regulatory instruments and technical standards published between 2015 and August 2026, with the lower bound set to capture the foundational symptom-checker audit literature that predates the LLM era and remains the reference point for triage safety.

B. Search Strategy and Database Selection

Two identification streams were used. The first executed five structured queries against the Crossref REST API; queries are reproduced verbatim, as entered, in Table 1. Crossref accepts free-text relevance queries and does not

support Boolean operators or field restriction in the manner of MEDLINE or Embase, so no Boolean syntax was used and none is reported. Because Crossref returns relevance-ranked results against a very large loosely matched candidate set, retrieval was capped at the top four to twenty-five records per query. MEDLINE, Embase, Scopus, Web of Science and IEEE Xplore were not searched. Crossref was selected because it indexes across publishers and disciplines without subscription dependency, which suited a review spanning clinical, informatics and regulatory literature; the cost is materially lower recall than a biomedical database search would achieve, particularly for clinical trial literature indexed with MeSH terms that Crossref does not carry. This is the principal methodological limitation of the review and is the reason the work is reported as a structured rather than systematic review. Extending the search to MEDLINE and Embase is identified in Section VI-C as the first priority for any subsequent version of this work. The second stream comprised targeted searching of regulatory, standards and journal registries: the World Health Organization, HL7 International, the United States Food and Drug Administration, the United States Department of Health and Human Services, the Official Journal of the European Union, the International Organization for Standardization, and the archives of the New England Journal of Medicine, JAMA, The Lancet, Nature Medicine and npj Digital Medicine. Searches were executed between 1 and 3 August 2026, with provenance tracing for Table 2 conducted on 20 August 2026.

C. Record Flow

The complete record sequence is as follows and is shown in Figure 1. Ninety-six records were retrieved from Crossref and twenty-one from registry searching, giving 117 records identified. Eight duplicates were removed, all arising where a record retrieved by a Crossref query was also located in a journal archive during registry searching; duplicates were identified by DOI, and where no DOI was present, by exact title and first-author match. This left 109 unique records screened on title and abstract. Seventy-three were excluded at screening, leaving 36 records sought for retrieval and assessed for eligibility. Twelve were excluded with reasons — vendor or marketing source (5), not AI or telehealth (2), preprint without peer review (2), and bibliographic details unverifiable (3) — leaving 24 records included in the synthesis.

D. Inclusion Criteria and Bibliographic Verification

Records were included where they reported empirical evaluation of conversational, triage or remote-monitoring systems in a health context; or constituted a binding regulatory instrument or technical standard governing such systems; and where bibliographic details could be reconciled against an external record. Records were excluded where the source was vendor or marketing material, where the subject was neither AI nor telehealth, where the record was a preprint without peer review, or where bibliographic details could not be reconciled. A verification step not standard in PRISMA was applied to every record surviving eligibility assessment. Author list, title, venue, volume, issue, pagination and publication date were reconciled against the Crossref record or the issuing institution's own registry. Because this review has a single author, the verification was not independent in the sense of a second reviewer; it was independent of the citing source, in that every bibliographic detail was checked against the publisher or issuing body rather than accepted from the secondary source in which the citation was encountered. This distinction is stated explicitly because the term is otherwise ambiguous. Three records were excluded on this ground (Figure 1).

E. Evidence Grading and Risk of Bias

Each retained source was assigned an evidence tier. Tier A denotes a randomised controlled trial or a systematic review of such trials. Tier B denotes a controlled or audit evaluation with a defined reference standard. Tier C denotes a regulatory instrument, technical standard or guidance document from a recognised authority. Tier D denotes vendor, consultancy or media material; Tier D sources are excluded from the retained evidence corpus and appear only in Table 2 as the object of provenance appraisal. Evidence tiering classifies study design and is not a substitute for risk-of-bias appraisal, so a formal appraisal was performed separately and is reported in Table 4. The three randomised trials were appraised using the Cochrane RoB 2 tool across its five domains. The two diagnostic accuracy studies were appraised using QUADAS-2 across its four domains of patient selection, index test, reference standard, and flow and timing, with applicability judged separately. A third instrument, ROBINS-I, was considered and not applied. ROBINS-I is designed for non-randomised studies of interventions, and the non-randomised records

in this corpus are not intervention studies: Obermeyer et al. (2019) is a retrospective analysis of a deployed algorithm's outputs, Omiye et al. (2023) elicits model responses under probing conditions, Iasiello et al. (2023) is a cross-sectional survey of reported access, and Ayers et al. (2023) is a blinded comparative rating of written responses. None involves an intervention assigned to participants, so the ROBINS-I domains concerning confounding and deviation from intended intervention do not apply. These four records are instead appraised narratively in Table 4 against the threats most relevant to their designs. Two limitations of the appraisal are recorded here rather than left implicit. It was performed by a single reviewer. And it was conducted from the published trial and study reports and their design descriptions as indexed, rather than from independently obtained full texts for every domain; domain judgements therefore reflect documented design features and reported conduct, and should be read as an appraisal of what the published record states rather than as a verification of underlying trial data. We note that Kopka et al. (2025) report their own QUADAS-2 appraisal, and our judgements for that study are consistent with theirs.

F. Limitations

Beyond the design limitations stated in Section III-A, four further limitations bear on interpretation. First, the relevance cap on Crossref retrieval means the search prioritises precision over recall. Second, the randomised evidence available for this technology class concerns remote patient management in heart failure specifically, and its extension to conversational and triage functions is an inference rather than a finding. Third, our search identified a small number of controlled evaluations of LLM triage performance, so the comparison in Section IV-B rests on a narrow empirical base. Fourth, Figure 2 reports quantities of two different kinds; they are presented in separate panels with separate axes for that reason, and no numerical comparison between panels is intended or supported.

RESULTS

Table 2: Provenance Audit of Five Circulating Efficacy Claims

Claim	Figure	Source located, with URL or DOI	Accessed	What that source actually states	Verification result
Administrative burden reduction	45%	Attributed to a McKinsey report on generative AI avatars in clinics; that report could not be located. A 45% figure appears in vendor material: Appinventiv, appinventiv.com/blog/ai-in-healthcare-administration/	20 Aug 2026	Vendor material reports operational efficiency gains of up to 45%, not administrative burden reduction. The nearest identifiable McKinsey position concerns the automation potential of administrative tasks, not realised reduction	Not verified. Attributed source not locatable; the figure conflates operational efficiency with administrative burden
Diagnostic accuracy uplift	40%	No source located	20 Aug 2026	—	Not verified. Not interpretable as stated: no baseline, task or population specified
Hospitalisation reduction	30%	No source located. Possible transposition from TIM-HF2 (Koehler <i>et al.</i> , 2018),	20 Aug 2026	TIM-HF2 reports HR 0.70 for all-cause mortality, a 30% relative reduction in	Not verified. Randomised evidence is mixed (Section

		doi.org/10.1016/S0140-6736(18)31880-4		mortality, not in hospitalisation	IV-C); no single reduction figure is supportable
Patient satisfaction improvement	35%	Healthcare virtual-assistant vendor site: DocVA, docva.com/virtual-health-assistant-workflow-optimization/	20 Aug 2026	Asserts a 35% increase in patient satisfaction scores from unspecified survey data; no sample, instrument or comparator stated	Not verified. No underlying study identified
Annual sector saving	US\$150 bn	Accenture (2017), clinical AI analysis, accenture.com (report PDF; full URL in reference list)	20 Aug 2026	Projects US\$150 bn in annual savings for the US healthcare economy by 2026, across use cases including robot-assisted surgery (US\$40 bn), virtual nursing assistants (US\$20 bn) and administrative workflow (US\$18 bn)	Traced. Is a 2017 projection, US-specific, target year 2026. Not a measured outcome and not global. Compare Sahni et al. (2023): 5–10% of US health spending, approximately US\$200–360 bn/yr

Note. Provenance tracing conducted 20 August 2026. Sources (Accenture, 2017; Appinventiv, n.d.; DocVA, n.d.) are Tier D under Section III-E and are cited here solely to document the origin of the claims audited; they are not part of the retained evidence corpus and are not used to support any finding. Full URLs and retrieval dates for all three appear in the reference list. Sahni et al. (2023) is a National Bureau of Economic Research working paper and is cited as the preferred economic estimate.

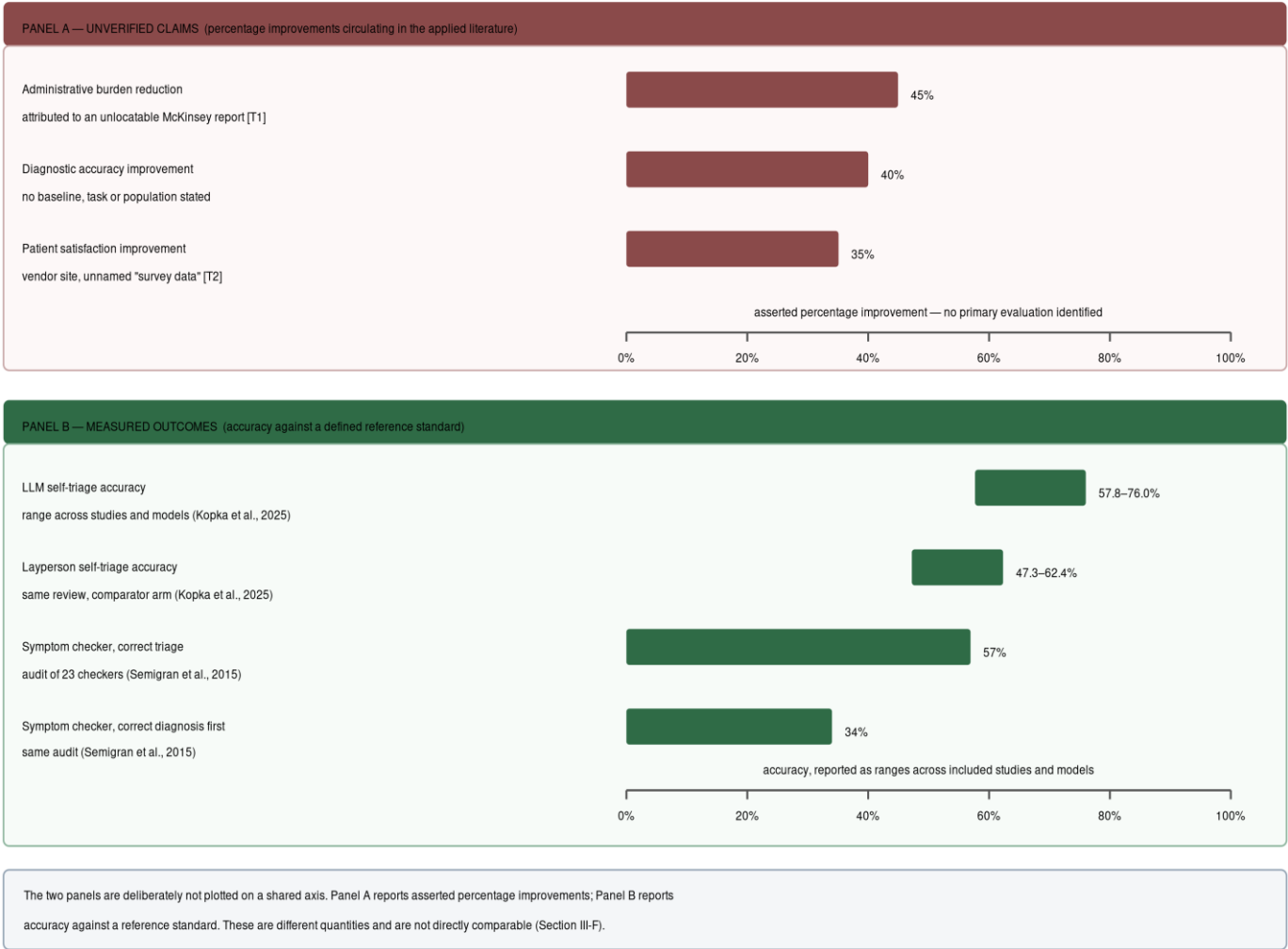


Figure 2. Circulating claims and measured outcomes, shown in separate panels with separate axes. Panel A reports asserted percentage improvements traced in the provenance audit of Table 2; T1 and T2 denote the vendor sources (Appinventiv, n.d.; DocVA, n.d.), respectively. Panel B reports accuracy against a defined reference standard, given as ranges across the studies and models included in the cited reviews. The two panels are deliberately not plotted on a common axis because they report different quantities; no numerical comparison between panels is intended (Section III-F).

A. Provenance of the Circulating Figures

Table 2 reports the provenance audit for the five headline figures, giving for each the claim as encountered, the source to which it could be traced, what that source actually states, and the verification outcome. The audit was conducted on 20 August 2026. Four findings follow. The US\$150 billion figure is traceable and is the clearest case. It originates in Accenture's 2017 analysis of clinical AI applications, which projected US\$150 billion in annual savings for the United States healthcare economy by 2026, disaggregated across use cases including robot-assisted surgery (US\$40 billion), virtual nursing assistants (US\$20 billion) and administrative workflow assistance (US\$18 billion) (Accenture, 2017). Three features of the figure are lost in circulation: it is a projection rather than a measured outcome, it is specific to the United States rather than to a global industry, and its target year is 2026 rather than the 2025 commonly cited. A more rigorous economic estimate exists and should be preferred: Sahni et al. (2023) estimate that wider AI adoption could yield savings of five to ten percent of United States healthcare spending, approximately US\$200–360 billion annually in 2019 dollars, from use cases attainable within five years. The forty-five percent administrative reduction is attributed in the circulating literature to a McKinsey report on generative

AI avatars in clinics. No such report could be located. The nearest identifiable McKinsey position concerns the share of healthcare administrative tasks that could in principle be automated, which is an automation-potential estimate rather than a realised reduction; a forty-five percent figure does appear in vendor material, but describes operational efficiency gains rather than administrative burden (Appinventiv, n.d.). The claim therefore conflates two different quantities and attributes them to a report that could not be found. The thirty-five percent satisfaction improvement traces to healthcare virtual-assistant vendor sites citing unnamed survey data, with no sample, instrument or comparator stated (DocVA, n.d.). The forty percent diagnostic accuracy improvement could not be traced to any source, and is in any case not interpretable as stated: no baseline, task or population is specified, and a percentage improvement without a denominator is not a proposition that can be evaluated. The thirty percent hospitalisation reduction merits separate treatment and is discussed in Section IV-C. We note here one plausible route by which it may have entered circulation: TIM-HF2 reported a hazard ratio of 0.70 for all-cause mortality, which is a thirty percent relative reduction — but in mortality, not hospitalisation. We advance this as a hypothesis about how a real trial result may have been transposed onto a different outcome, not as an established account of the figure's origin.

B. Measured Performance

Table 3 and Figure 2, Panel B, set out the measured evidence. Kopka et al. (2025), in a controlled evaluation against a defined reference standard, report self-triage accuracy of 57.8 to 76.0 percent for large language models, 11.5 to 90.0 percent for online symptom-assessment applications, and 47.3 to 62.4 percent for laypeople. These are ranges observed across the different studies, models and evaluation conditions included in that review, not pooled point estimates, and the underlying studies differ in design, model, population and task. Two observations follow, both stated with the caution the underlying heterogeneity requires. LLMs showed higher reported self-triage accuracy than laypeople across the studies included in the review, which is the relevant comparison for a patient-facing self-triage tool. And reported self-triage accuracy reached 76 percent in the included studies — the upper value observed, not a theoretical performance limit — on a task whose failure mode is a missed emergency.

The foundational audit by Semigran et al. (2015), covering twenty-three symptom checkers against standardised vignettes, remains the reference point: the correct diagnosis was listed first in thirty-four percent of cases and correct triage advice was given in fifty-seven percent. Chen et al. (2026), in a systematic review and meta-analysis searching seven databases and screening 10,398 records to include 50 studies across 25 large language models published between January 2020 and September 2025, report the relative diagnostic accuracy of LLMs against healthcare professionals — the ratio of LLM accuracy to clinician accuracy, where unity denotes parity — as 0.89 (95% CI 0.79–1.00) for top-1 diagnosis, 0.91 (0.83–1.00) for top-3, 1.04 (0.89–1.22) for top-5 and 1.17 (0.87–1.57) for top-10. The pattern is that LLMs underperform healthcare professionals when a single diagnosis is required and approach or exceed them only as the permitted differential widens, with confidence intervals crossing unity at every level beyond top-3. That is a materially more qualified result than a summary statement about collaborative superiority would convey. On the collaborative question specifically, Wang et al. (2026), in a separate systematic review and meta-analysis of human–LLM collaboration in clinical medicine, report a positive but statistically imprecise effect on diagnostic and interpretation accuracy (risk ratio 1.59, 95% CI 0.08–32.74, non-significant) alongside a small but statistically significant improvement in composite diagnostic and management scores of 4.88 percentage points (95% CI 0.65–9.12). The collaborative case is therefore directionally supported but weakly estimated, and the width of the accuracy interval should temper any claim that collaboration reliably exceeds either party alone. On response quality rather than accuracy, Ayers et al. (2023) provide the most favourable controlled result in the corpus: blinded clinician evaluators preferred chatbot responses to physician responses on both quality and empathy. The finding is robust within its design but concerns asynchronous written responses to forum questions, not triage decisions under time pressure with incomplete information, and should not be generalised beyond that.

C. Remote Patient Management: Mixed Randomised Evidence

Remote patient management is the one VHA-adjacent function for which large randomised trials exist, and the evidence is mixed rather than uniformly negative. Chaudhry et al. (2010), in the Tele-HF trial of 1,653 patients, found

that telemonitoring after heart-failure hospitalisation did not reduce readmission or death, the trial's primary outcome. Ong et al. (2016), in BEAT-HF, likewise found no reduction in 180-day readmission. Koehler et al. (2018), in TIM-HF2, found benefit. Reporting that result in full rather than in summary: among patients in NYHA class II or III with a heart-failure admission in the previous twelve months, structured remote patient management reduced the percentage of days lost to unplanned cardiovascular hospital admission or all-cause death over twelve months from 6.64 percent (95% CI 6.19–7.13) under usual care to 4.88 percent (4.55–5.23), a ratio of 0.80 (95% CI 0.65–1.00; $p = 0.0460$). All-cause mortality showed a hazard ratio of 0.70, and the effect was driven predominantly by a reduction in cardiovascular deaths. The primary outcome estimate is statistically borderline, with a confidence interval reaching unity and a p -value close to the conventional threshold, and should be read accordingly. The correct synthesis is therefore not that randomised evidence contradicts a benefit from remote monitoring. It is that the evidence is heterogeneous, and that the trial reporting benefit differed from those reporting none in delivering monitoring through a structured, physician-supervised management programme with defined escalation pathways rather than through instrumentation alone. Any single headline reduction figure attributed to monitoring technology per se is unsupported by this literature, in either direction. This finding has a direct architectural consequence, developed in Section V-B.

D. Risk of Bias

Table 4 reports the risk-of-bias appraisal. Among the randomised trials, Tele-HF and BEAT-HF were judged at low risk of bias overall, with some concerns arising only from the unavoidable absence of participant and personnel blinding in a behavioural intervention. TIM-HF2 was judged to have some concerns overall, arising from the unmasked design, the borderline primary-outcome estimate, and the composite nature of that outcome. Among the diagnostic accuracy studies, Semigran et al. (2015) carries a high applicability concern under QUADAS-2 because standardised vignettes do not reproduce the presentation characteristics of real patients, and Kopka et al. (2025) carries some concerns for patient selection for the same reason. These judgements do not overturn any finding reported above, but they bound the confidence attachable to each.

E. Bias, Subgroup Performance and Access: Four Distinct Evidence Types

The literature bearing on equity in this domain contains at least four distinct kinds of evidence, which are frequently merged. They are set out separately here because they support different conclusions and none of them substitutes for the others. The first is evidence of algorithmic bias in deployed health systems. Obermeyer et al. (2019) demonstrated that a widely used commercial health-risk algorithm systematically under-referred Black patients because it used healthcare cost as a proxy for healthcare need. This is direct, real-world evidence — but it concerns a risk-stratification algorithm, not a virtual healthcare assistant. The second is evidence that large language models generate race-based medical claims. Omiye et al. (2023) showed that contemporary models reproduce discredited race-based content when queried directly. This is direct evidence about model outputs under adversarial or probing conditions, not about performance in a deployed clinical workflow. The third is evidence of unequal telehealth access. Iasiello et al. (2023) found materially lower patient-reported telehealth access among Black patients in an oncology population. This concerns access to the service channel, not the performance of any system operating within it. The fourth is direct evidence of subgroup performance of virtual healthcare assistants in clinical deployment — that is, disaggregated accuracy or safety by demographic group for a fielded VHA. Our search identified no such study. This is the most consequential gap in the present evidence base: the equity question most relevant to VHA deployment is, at the time of writing, empirically open. What the first three evidence types establish is a well-grounded prior that subgroup performance should not be assumed uniform, and a corresponding obligation to measure it. They do not establish that VHAs do in fact perform worse for particular groups, and this paper does not claim that they do.

Proposed Reference Architecture

Table 3: Measured Outcomes by Evidence Tier

Outcome	Value	Source (tier)
LLM self-triage accuracy, range across included studies	57.8–76.0%	Kopka et al. (2025) — B

Symptom-assessment application self-triage, range	11.5–90.0%	Kopka et al. (2025) — B
Layperson self-triage accuracy, range	47.3–62.4%	Kopka et al. (2025) — B
Symptom checker, correct diagnosis listed first	34%	Semigran et al. (2015) — B
Symptom checker, correct triage advice	57%	Semigran et al. (2015) — B
LLM vs. clinician relative diagnostic accuracy, top-1	0.89 (95% CI 0.79–1.00)	Chen et al. (2026) — B
LLM vs. clinician relative diagnostic accuracy, top-3	0.91 (0.83–1.00)	Chen et al. (2026) — B
LLM vs. clinician relative diagnostic accuracy, top-5	1.04 (0.89–1.22)	Chen et al. (2026) — B
LLM vs. clinician relative diagnostic accuracy, top-10	1.17 (0.87–1.57)	Chen et al. (2026) — B
Human–LLM collaboration, diagnostic / interpretation accuracy	RR 1.59 (95% CI 0.08–32.74), non-significant	Wang et al. (2026) — B
Human–LLM collaboration, composite diagnostic and management score	+4.88 percentage points (0.65–9.12)	Wang et al. (2026) — B
Chatbot vs. physician response quality and empathy	chatbot preferred, blinded raters	Ayers et al. (2023) — B
Tele-HF, readmission or death (primary outcome)	no significant reduction	Chaudhry et al. (2010) — A
BEAT-HF, 180-day readmission	no significant reduction	Ong et al. (2016) — A
TIM-HF2, % days lost to unplanned CV admission or death	4.88% vs 6.64%; ratio 0.80 (0.65–1.00), p = 0.046	Koehler et al. (2018) — A
TIM-HF2, all-cause mortality	HR 0.70	Koehler et al. (2018) — A
Health-risk algorithm, referral by race	systematic under-referral of Black patients	Obermeyer et al. (2019) — B
LLM race-based medical claims when probed	Reproduced	Omiye et al. (2023) — B
Patient-reported telehealth access by race	materially lower among Black patients	Iasiello et al. (2023) — B
Subgroup performance of a deployed VHA	no study identified	no study identified — gap

Note. Evidence tiers per Section III-E: (A) randomised controlled trial or systematic review of trials; (B) controlled or audit evaluation against a defined reference standard. Accuracy figures from (Kopka *et al.*, 2025) are ranges across the studies, models and conditions included in that review, not pooled estimates. Risk-of-bias appraisal for these studies is reported in Table 4.

Table 4: Risk-of-Bias and Methodological Appraisal (RoB 2, QUADAS-2 and Narrative Appraisal)

Study and tool	Domain judgements	Overall
Chaudhry et al. (2010), Tele-HF — RoB 2	Randomisation: low. Deviations from intended intervention: some concerns (unmasked behavioural intervention). Missing outcome data: low. Outcome	Low

	measurement: low (objective endpoints). Selective reporting: low	
Ong et al. (2016), BEAT-HF — RoB 2	Randomisation: low. Deviations: some concerns (unmasked). Missing outcome data: low. Outcome measurement: low. Selective reporting: low	Low
Koehler et al. (2018), TIM-HF2 — RoB 2	Randomisation: low. Deviations: some concerns (unmasked, co-intervention within management programme). Missing outcome data: low. Outcome measurement: some concerns (composite primary outcome). Selective reporting: low	Some concerns
Semigran et al. (2015) — QUADAS-2	Patient selection: high concern (standardised vignettes, not real patients). Index test: low. Reference standard: low. Flow and timing: low. Applicability: high concern	High applicability concern
Kopka et al. (2025) — QUADAS-2	Patient selection: some concerns (vignette-based). Index test: low. Reference standard: low. Flow and timing: low. Applicability: some concerns	Some concerns
Ayers et al. (2023) — narrative appraisal	Blinded comparative rating of written responses. Rater blinding: low concern. Selection of question set: some concerns (single public forum). Construct validity: some concerns (rated empathy in asynchronous text is not therapeutic presence). Generalisability to triage: high concern	Some concerns
Obermeyer et al. (2019) — narrative appraisal	Retrospective analysis of a deployed algorithm. Data completeness: low concern. Outcome definition: low concern. Applicability to VHAs: high concern (risk-stratification algorithm, not a conversational system)	Low, applicability high concern
Omiye et al. (2023) — narrative appraisal	Model probing under adversarial framing. Prompt specification: low concern. Reproducibility across model versions: some concerns. Applicability to deployed workflow: high concern	Some concerns
Iasiello et al. (2023) — narrative appraisal	Cross-sectional patient-reported survey. Sampling frame: some concerns (single oncology population). Self-report measurement: some concerns. Applicability beyond access: high concern	Some concerns

Note. RoB 2 applied to randomised trials, QUADAS-2 to diagnostic accuracy studies. ROBINS-I was considered and not applied; the rationale is given in Section III-E. The four non-randomised records are appraised narratively against design-appropriate threats. Appraisal performed by a single reviewer, from the published record rather than from independently obtained full texts for every domain; see Section III-E. Kopka et al. (2025) report their own QUADAS-2 appraisal; the judgements above are consistent with theirs. RoB 2 domain 2 judgements of 'some concerns' for all three trials reflect the impossibility of masking a behavioural monitoring intervention rather than any deficiency of trial conduct.

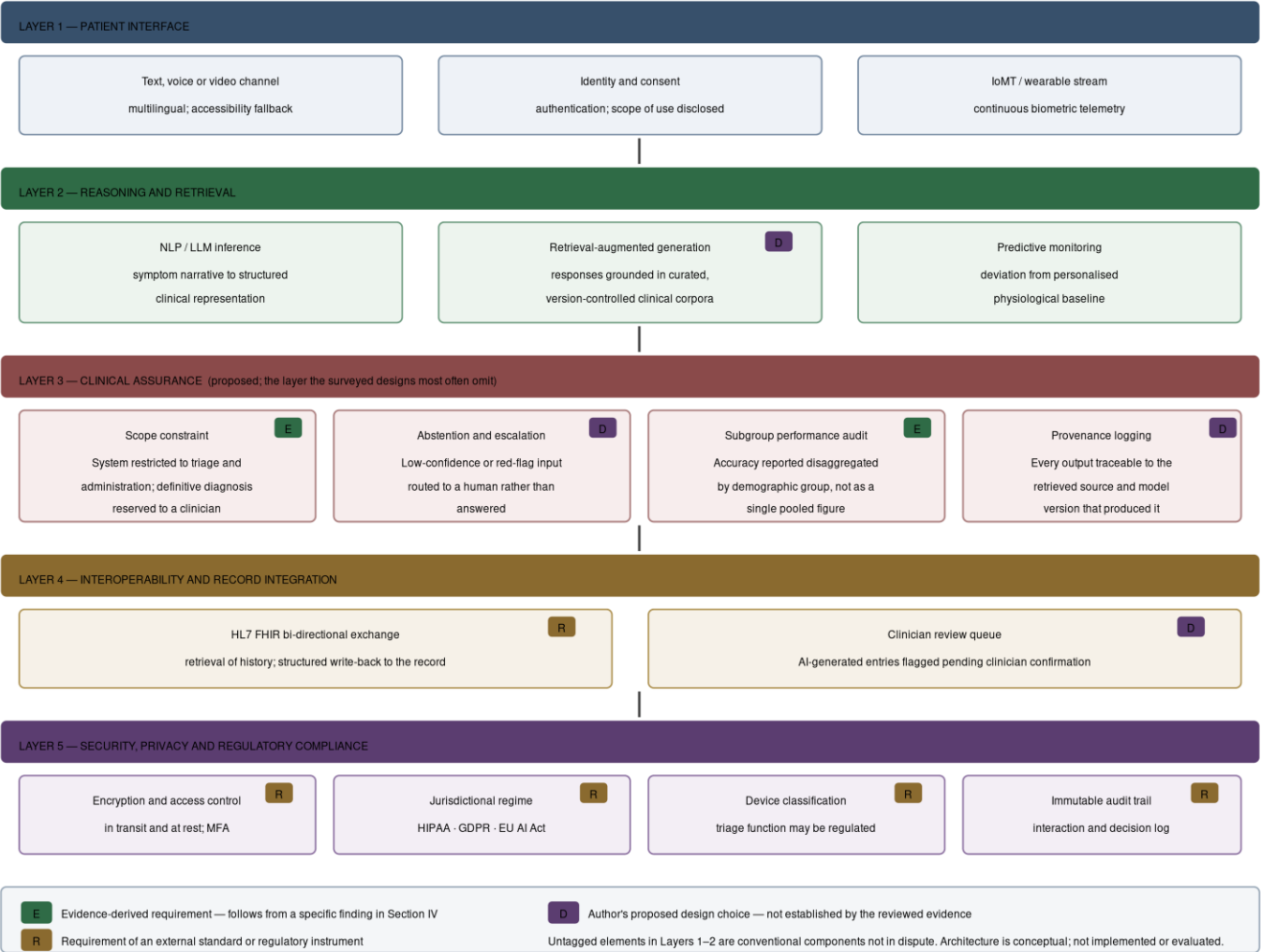


Figure 3. Proposed five-layer reference architecture for a clinically deployable virtual healthcare assistant. The architecture is conceptual and derived from the reviewed literature and cited regulatory instruments; it has not been implemented or experimentally evaluated. Elements are annotated E where they follow from a specific finding in Section IV, R where they follow from an external standard or regulatory instrument, and D where they represent the author's proposed design judgement not established by the reviewed evidence.

A. A Five-Layer Structure

Figure 3 sets out a proposed reference architecture in five layers. It is a conceptual framework derived from the reviewed literature and from the regulatory instruments cited; it has not been implemented, deployed or experimentally evaluated, and no claim of demonstrated effectiveness is made for it. Elements are annotated in the figure according to whether they follow from a specific finding in Section IV, from an external standard or regulatory instrument, or from the author's design judgement.

Layer 1 is the patient interface: multilingual text, voice or video channels with accessibility fallback, explicit identity and consent handling, and ingestion of wearable telemetry. Layer 2 is reasoning and retrieval: language understanding that converts symptom narrative into structured clinical representation, retrieval grounding, and predictive monitoring against personalised physiological baselines. Layer 4 is interoperability, using HL7 FHIR (HL7 International, 2023) for bi-directional exchange with the electronic health record, with AI-generated entries written to a clinician review queue rather than committed directly. Layer 5 is security, privacy and regulatory compliance, spanning encryption and access control, the applicable jurisdictional regime (European Parliament & Council of the

European Union, 2016, 2024; U.S. Department of Health and Human Services, n.d.), device classification (U.S. Food and Drug Administration, 2026), and immutable audit logging.

B. Layer 3: Clinical Assurance

The principal conceptual contribution of this review is the proposed Layer 3 clinical assurance framework. It sits between reasoning and record integration, and it is the layer that the architectures encountered in this review most often omit. It comprises four mechanisms, of which two follow from findings in Section IV and two are proposed design choices; Figure 3 marks the distinction. Scope constraint is evidence-derived. Given reported self-triage accuracy reaching 76 percent in the included studies (Kopka *et al.*, 2025), and relative diagnostic accuracy below parity with clinicians for single-diagnosis tasks (Chen *et al.*, 2026), a triage system should be architecturally rather than contractually restricted to triage and administrative functions, with definitive diagnosis reserved to a clinician. Subgroup performance audit is likewise evidence-derived, following from Section IV-E: because the equity question is empirically open and the prior against uniformity is well grounded (Obermeyer *et al.*, 2019; Omiye *et al.*, 2023), accuracy should be reported disaggregated by demographic group rather than as a single pooled figure, so that the question becomes answerable. Abstention and escalation is a proposed design choice rather than an evidence-derived requirement. The reviewed literature does not evaluate abstention mechanisms in deployed VHAs. The argument for it is inferential: a system operating at the reported accuracy levels that answers every query is, on the face of it, less safe than one that declines on low-confidence or red-flag input and routes it to a human. We regard this as well-motivated but untested. Provenance logging is similarly a design choice, motivated by the hallucination risk inherent to generative models (Thirunavukarasu *et al.*, 2023) and by the audit expectations of the regulatory instruments in Layer 5, rather than by any evaluation of provenance mechanisms in this corpus. The remote-monitoring evidence (Chaudhry *et al.*, 2010; Koehler *et al.*, 2018; Ong *et al.*, 2016) motivates a further requirement that cuts across layers. Because the trial reporting benefit delivered monitoring through a structured management programme with defined escalation, whereas the trials reporting no benefit did not, a monitoring VHA that generates alerts without a specified, staffed and auditable escalation route has weaker grounds to expect the TIM-HF2 result than the Tele-HF and BEAT-HF results. The response pathway is plausibly part of the intervention rather than an implementation detail, though we note that no trial has isolated the response pathway as an independent variable.

C. Regulatory Position

A VHA performing triage may meet the definition of a regulated medical device in several jurisdictions, and classification turns on the function performed rather than on how the product is marketed. In the United States, the Food and Drug Administration issued updated final guidance on Clinical Decision Support Software on 6 January 2026, replacing its September 2022 predecessor (U.S. Food and Drug Administration, 2026). The 2026 guidance clarifies which CDS functions fall outside the device definition under section 520(o)(1)(E) of the Federal Food, Drug, and Cosmetic Act as amended by the 21st Century Cures Act. Legal commentary on the revision has emphasised a narrowing of the regulated perimeter relative to the 2022 version; we report that characterisation as commentary rather than as our own reading, and note that any developer who relied on the 2022 guidance should assess classification against the 2026 text directly. We note that the guidance addresses clinician-facing CDS specifically; a patient-facing triage VHA is a different object, and its classification does not follow automatically from the CDS framework. In the European Union, certain AI systems used in healthcare may fall within the high-risk categories of the AI Act (European Parliament & Council of the European Union, 2024), depending on their intended purpose and their relationship to regulated products or services. This is a conditional classification rather than a categorical one, and it is not the case that every conversational health application falls into the same category. The World Health Organization's guidance on the ethics and governance of AI for health (World Health Organization, 2021) and its subsequent regulatory considerations (World Health Organization, 2023) establish expectations of transparency, accountability and post-deployment surveillance that Layer 3 is intended to operationalise. At organisational level, ISO/IEC 42001 (International Organization for Standardization, 2023) specifies requirements for an AI management system; we suggest that the Layer 3 mechanisms could be documented within such a system as one route to evidencing them, while noting that this is an application we propose rather than a requirement the standard imposes.

DISCUSSION

A. The Limiting Factor Is Evidentiary

The pattern across Section IV is consistent. Where controlled evaluation exists, it shows capability that is real, useful and bounded: higher reported self-triage accuracy for LLMs than for laypeople across included studies (Kopka *et al.*, 2025); chatbot responses preferred to physician responses on quality and empathy in an asynchronous written setting (Ayers *et al.*, 2023); remote patient management reducing days lost to hospitalisation or death when delivered through a structured programme (Koehler *et al.*, 2018). Where controlled evaluation does not exist, the literature has substituted vendor-reported figures and reproduced them until they acquired the appearance of evidence, as Table 2 documents. The constraint on this field is therefore not model capability. It is that the questions with the highest clinical stakes — whether deployment reduces missed emergencies, whether it changes outcomes, whether it performs equitably across groups — are the questions least often subjected to controlled measurement. Chen *et al.* (2026) point toward the productive framing: independent LLM performance falls below clinician performance for single-diagnosis tasks. The collaborative configuration is where attention is now turning, though on present evidence its advantage is small and imprecisely estimated (Wang *et al.*, 2026).

B. Limitations of Current Systems

Three limitations persist independently of the evidentiary problem. Generative models produce fluent but incorrect content, and retrieval grounding is a proposed rather than a demonstrated remedy at deployment scale (Thirunavukarasu *et al.*, 2023). Access to telehealth is unequally distributed: Iasiello *et al.* (2023) document materially lower patient-reported access among Black patients in an oncology population, and while it is plausible that access barriers and unmet need are correlated, the cited evidence establishes the access differential rather than that correlation, and we do not assert the stronger claim. And current systems lack the capacity for clinical judgement in settings — severe psychological distress, palliative care, the delivery of a catastrophic diagnosis — where the interaction is itself the intervention. The Ayers *et al.* finding on empathy (Ayers *et al.*, 2023) does not contradict this: rated empathy in written asynchronous responses is a different construct from therapeutic presence.

C. Research Agenda

Five priorities follow. First, and in the author's assessment the highest-value single study the field could conduct, no randomised trial in the retained corpus evaluates an LLM-based triage VHA against standard care on a clinical safety endpoint. Second, subgroup-disaggregated performance reporting for deployed VHAs is absent, which makes the equity question unanswerable with present data. Third, the response-pathway dependency suggested by the monitoring literature (Chaudhry *et al.*, 2010; Koehler *et al.*, 2018; Ong *et al.*, 2016) has not been isolated experimentally, for monitoring or for conversational systems. Fourth, the economic claims that dominate procurement discussion rest on projection rather than measurement, and a properly designed cost-effectiveness analysis anchored on the Sahni *et al.* framework (Sahni *et al.*, 2023) would be a substantial contribution. Fifth, extending the search of this review to MEDLINE, Embase and Scopus would establish whether the present corpus is representative.

CONCLUSION

Virtual healthcare assistants have demonstrated capability that is genuine and, in specific configurations, clinically meaningful. Large language models encode substantial clinical knowledge (Singhal *et al.*, 2023), show higher reported self-triage accuracy than laypeople across the studies reviewed (Kopka *et al.*, 2025), produce responses that blinded clinicians rate above physician responses on quality and empathy in asynchronous settings (Ayers *et al.*, 2023), and support remote management programmes that reduce days lost to hospitalisation or death when coupled to structured clinical response (Koehler *et al.*, 2018). The figures currently driving adoption are not these. Four of the five headline claims examined here could not be traced to any primary evaluation, and the fifth is a 2017 consultancy projection for the United States rather than a measured global outcome (Table 2). Measured performance is more modest and more conditional: reported self-triage accuracy reaching 76 percent in the included studies, symptom-checker triage accuracy near sixty percent, relative diagnostic accuracy below parity with clinicians for single-diagnosis tasks, and remote-monitoring benefit contingent on delivery model rather than

on instrumentation. On equity, the evidence supports a narrower statement than is often made. Algorithmic bias has been demonstrated in a deployed risk-stratification algorithm (Obermeyer *et al.*, 2019), race-based content has been elicited from large language models (Omiye *et al.*, 2023), and unequal telehealth access has been documented (Iasiello *et al.*, 2023). Direct evidence of subgroup performance for deployed virtual healthcare assistants was not identified in this review. The appropriate conclusion is that uniform performance across groups should not be assumed and must be measured, not that unequal VHA performance has been established. The implication is not that this technology should be resisted. It is that the field should hold itself to the evidentiary standard applied to other clinical interventions: reporting disaggregated rather than pooled accuracy, designing systems that abstain rather than answer at the edge of their competence, treating the clinical response pathway as part of the intervention, and subjecting triage functions to controlled evaluation proportionate to their risk. The proposed five-layer architecture, and its clinical assurance layer in particular, is offered as one conceptual route to giving that standard implementable form, and requires empirical evaluation before it can be recommended for adoption rather than for testing.

Declarations

A. Data Availability

No new primary data were generated by this study. All analysed records are published sources listed in the References. The five Crossref queries reported verbatim in Table 1, executed between 1 and 3 August 2026, reproduce the search stream; the record flow is given in Section III-C and Figure 1; the provenance audit is given in Table 2 and the risk-of-bias appraisal in Table 4. The review protocol, the study-level extraction table, the registry search procedure and the screening log are provided as Supplementary Material S1 accompanying this submission.

B. Author Contributions

P. S. Hussaini is the sole author and is responsible for conceptualisation, methodology, investigation, formal analysis, data curation, visualisation, and writing of the original draft and its revision. The author screened all records, performed the bibliographic verification described in Section III-D and the risk-of-bias appraisal described in Section III-E, and approved the final manuscript.

C. Funding and Competing Interests

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors. The author declares no competing financial or non-financial interests.

D. Acknowledgment

The author thanks the Department of Mechatronics Engineering, Federal University of Technology, Minna, for institutional support, and the anonymous reviewer whose detailed appraisal substantially improved the methodology, provenance reporting and evidentiary calibration of this manuscript.

REFERENCES

- Accenture. (2017). Artificial intelligence in healthcare. <https://www.accenture.com/content/dam/accenture/final/a-com-migration/manual/r3/pdf/pdf-49/Accenture-health-artificial-intelligence-j.pdf>
- Appinventiv. (n.d.). AI in healthcare administration. Retrieved August 20, 2026, from <https://appinventiv.com/blog/ai-in-healthcare-administration/>
- Ayers, J. W., Poliak, A., Dredze, M., Leas, E. C., Zhu, Z., Kelley, J. B., Faix, D. J., Goodman, A. M., Longhurst, C. A., Hogarth, M., & Smith, D. M. (2023). Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Internal Medicine*, 183(6), 589. <https://doi.org/10.1001/jamainternmed.2023.1838>

- Chaudhry, S. I., Mattera, J. A., Curtis, J. P., Spertus, J. A., Herrin, J., Lin, Z., Phillips, C. O., Hodshon, B. V., Cooper, L. S., & Krumholz, H. M. (2010). Telemonitoring in patients with heart failure. *New England Journal of Medicine*, 363(24), 2301–2309. <https://doi.org/10.1056/NEJMoa1010029>
- Chen, M., Wu, Y., Ma, J., Jia, X., Gao, C., Zhao, F., & Qiao, Y. (2026). Independent and collaborative performance of large language models and healthcare professionals in diagnosis and triage. *npj Digital Medicine*, 9(1). <https://doi.org/10.1038/s41746-026-02409-8>
- DocVA. (n.d.). Virtual health assistant workflow optimization. Retrieved August 20, 2026, from <https://docva.com/virtual-health-assistant-workflow-optimization/>
- European Parliament & Council of the European Union. (2016). Regulation (EU) 2016/679 (General Data Protection Regulation). Official Journal of the European Union, L 119, 1–88. <http://data.europa.eu/eli/reg/2016/679/oj>
- European Parliament & Council of the European Union. (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union, L series. <http://data.europa.eu/eli/reg/2024/1689/oj>
- HL7 International. (2023). HL7 FHIR Release 5. Health Level Seven International. <https://hl7.org/fhir/>
- Iasiello, J. A., Rajan, A., Zervos, E., Parikh, A. A., & Snyder, R. A. (2023). Racial differences in patient-reported access to telehealth: An important and unmeasured social determinant of health. *JCO Oncology Practice*, 19(12), 1215–1223. <https://doi.org/10.1200/OP.23.00006>
- International Organization for Standardization. (2023). Information technology — Artificial intelligence — Management system (ISO/IEC Standard No. 42001:2023). <https://www.iso.org/standard/81230.html>
- Koehler, F., Koehler, K., Deckwart, O., Prescher, S., Wegscheider, K., Kirwan, B.-A., Winkler, S., Vettorazzi, E., Bruch, L., Oeff, M., Zugck, C., Doerr, G., Naegel, H., Störk, S., Butter, C., Sechtem, U., Angermann, C., Gola, G., Prondzinsky, R., ...
- Stangl, K. (2018). Efficacy of telemedical interventional management in patients with heart failure (TIM-HF2): A randomised, controlled, parallel-group, unmasked trial. *The Lancet*, 392(10152), 1047–1057. [https://doi.org/10.1016/S0140-6736\(18\)31880-4](https://doi.org/10.1016/S0140-6736(18)31880-4)
- Kopka, M., von Kalckreuth, N., & Feufel, M. A. (2025). Accuracy of online symptom assessment applications, large language models, and laypeople for self-triage decisions. *npj Digital Medicine*, 8(1). <https://doi.org/10.1038/s41746-025-01566-6>
- Laranjo, L., Dunn, A. G., Tong, H. L., Kocaballi, A. B., Chen, J., Bashir, R., Surian, D., Gallego, B., Magrabi, F., Lau, A. Y. S., & Coiera, E. (2018). Conversational agents in healthcare: A systematic review. *Journal of the American Medical Informatics Association*, 25(9), 1248–1258. <https://doi.org/10.1093/jamia/ocy072>
- Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464), 447–453. <https://doi.org/10.1126/science.aax2342>
- Omiye, J. A., Lester, J. C., Spichak, S., Rotemberg, V., & Daneshjou, R. (2023). Large language models propagate race-based medicine. *npj Digital Medicine*, 6(1). <https://doi.org/10.1038/s41746-023-00939-z>

- Ong, M. K., Romano, P. S., Edgington, S., Aronow, H. U., Auerbach, A. D., Black, J. T., De Marco, T., Escarce, J. J., Evangelista, L. S., Hanna, B., Ganiats, T. G., Greenberg, B. H., Greenfield, S., Kaplan, S. H., Kimchi, A., Liu, H., Lombardo, D., Mangione, C. M., Sadeghi, B., ... Fonarow, G. C. (2016). Effectiveness of remote patient monitoring after discharge of hospitalized patients with heart failure: The Better Effectiveness After Transition-Heart Failure (BEAT-HF) randomized clinical trial. *JAMA Internal Medicine*, 176(3), 310. <https://doi.org/10.1001/jamainternmed.2015.7712>
- Sahni, N., Stein, G., Zimmel, R., & Cutler, D. M. (2023). The potential impact of artificial intelligence on healthcare spending (Working Paper No. 30857). National Bureau of Economic Research. <https://doi.org/10.3386/w30857>
- Semigran, H. L., Linder, J. A., Gidengil, C., & Mehrotra, A. (2015). Evaluation of symptom checkers for self diagnosis and triage: Audit study. *BMJ*, 351, h3480. <https://doi.org/10.1136/bmj.h3480>
- Shankar, R., Lim, A., & Xu, Q. (2026). Barriers and facilitators to the use of large language model-based conversational agents in mental healthcare: A systematic review. *Healthcare*, 14(10), 1267. <https://doi.org/10.3390/healthcare14101267>
- Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., Payne, P., Seneviratne, M., Gamble, P., Kelly, C., Babiker, A., Schärli, N., Chowdhery, A., Mansfield, P., Demner-Fushman, D., ... Natarajan, V. (2023). Large language models encode clinical knowledge. *Nature*, 620(7972), 172–180. <https://doi.org/10.1038/s41586-023-06291-2>
- Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F., & Ting, D. S. W. (2023). Large language models in medicine. *Nature Medicine*, 29(8), 1930–1940. <https://doi.org/10.1038/s41591-023-02448-8>
- U.S. Department of Health and Human Services. (n.d.). HIPAA Security Rule, 45 C.F.R. Parts 160 and 164, Subparts A and C. Retrieved August 3, 2026, from <https://www.hhs.gov/hipaa>
- U.S. Food and Drug Administration. (2026). Clinical decision support software: Guidance for industry and Food and Drug Administration staff (Final guidance, January 6, 2026, superseding the guidance of September 28, 2022). <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/clinical-decision-support-software>
- Wang, G., Zhang, K., Jiang, J., Wang, C., Bi, H., Liang, H., Qi, Z., Huang, Y., Li, Y., & Yang, X. (2026). Human-large language model collaboration in clinical medicine: A systematic review and meta-analysis. *npj Digital Medicine*, 9(1). <https://doi.org/10.1038/s41746-026-02382-2>
- World Health Organization. (2021). Ethics and governance of artificial intelligence for health: WHO guidance. World Health Organization. ISBN 978-92-4-002920-0. <https://www.who.int/publications/i/item/9789240029200>
- World Health Organization. (2022). Consolidated telemedicine implementation guide. World Health Organization. ISBN 978-92-4-005918-4. <https://www.who.int/publications/i/item/9789240059184>
- World Health Organization. (2023). Regulatory considerations on artificial intelligence for health. World Health Organization. ISBN 978-92-4-007887-1. <https://www.who.int/publications/i/item/9789240078871>
- World Health Organization. (2025). Global strategy on digital health 2020–2027 (extended from 2020–2025 by World Health Assembly resolution WHA78(22), May 23, 2025). <https://www.who.int/publications/b/81640>