Applied Journal of Digital Innovation (AJDI)

ISSN: 3121-2697 **DOI:** 10.59282/ajdi.6

ORIGINAL

Generative AI in telemedicine: clinical validation, ethics, and implementation to optimize remote diagnosis and adherence

IA generativa en telemedicina: validación clínica, ética e implementación para optimizar diagnóstico remoto y adherencia

Dania Narcisa Petao Salazar[1]*

[1]Revista Reincisol

***Corresponding author:** Dania Narcisa Petao Salazar

Journal:	Applied Journal of Digital Innovation (AJDI)	**ISSN:**	3121-2697
Volume/Issue:	Vol. 1, No. 1(2025)	**DOI:**	10.59282/ajdi.6
Submitted:	27-04-2024	**Revised:**	21-09-2024
Accepted:	28-12-2024	**Published:**	21-01-2025

Cite as: Petao, D. (2025). IA generativa en telemedicina: validación clínica, ética e implementación para optimizar diagnóstico remoto y adherencia. Applied Journal of Digital Innovation. 1. https://doi.org/10.59282/ajdi.6

© 2025; Los autores. Este es un artículo en acceso abierto, distribuido bajo los términos de una licencia Creative Commons (https://creativecommons.org/licenses/by/4.0) que permite el uso, distribución y reproducción en cualquier medio siempre que la obra original sea correctamente citada

ABSTRACT

This mixed-methods study investigates the implementation of Generative Artificial Intelligence (GenAI) in telemedicine, focusing on optimizing remote diagnosis and therapeutic adherence. Through a systematic review, expert interviews, and two experimental studies, it validates GenAI as an effective "copilot," improving physicians' diagnostic accuracy by 15% and efficiency by 30%, while AI-generated care plans significantly increase patient adherence and understanding. However, the core findings identify three critical and interdependent pillars for responsible adoption: Validation (requiring evidence of net clinical benefit beyond technical accuracy), Ethics (where explainability is revealed as a clinical facilitator, not merely a moral imperative), and Implementation (demanding a human-centric redesign of workflows). The proposed V-E-I Framework is an iterative model integrating these pillars, arguing that success depends on advancing them simultaneously and in balance. The conclusion emphasizes that the future of GenAI in healthcare is not guaranteed by technology alone, but by responsible governance that prioritizes safety, equity, and real clinical utility over mere innovation.

Keywords: Generative Artificial Intelligence (GenAI), Telemedicine, Remote Diagnosis, Therapeutic Adherence, Clinical Validation

RESUMEN

Este estudio de metodología mixta investiga la implementación de la Inteligencia Artificial Generativa (GenAI) en telemedicina, centrándose en la optimización del diagnóstico remoto y la adherencia terapéutica. Mediante una revisión sistemática, entrevistas a expertos y dos estudios experimentales, se valida que la GenAI actúa como un "copiloto" efectivo, mejorando la precisión diagnóstica de los médicos en un 15% y la eficiencia en un 30%, a la vez que los planes de cuidado generados por IA aumentan significativamente la adherencia y comprensión del paciente. Sin embargo, los hallazgos centrales identifican tres pilares críticos e interdependientes para una adopción responsable: Validación (requiriendo evidencia de beneficio clínico neto más allá de la precisión técnica), Ética (donde la explicabilidad se revela como un facilitador clínico, no solo un imperativo moral) e Implementación (que exige un rediseño de flujos de trabajo centrado en el humano). Se propone el Marco V-E-I, un modelo iterativo que integra estos pilares, argumentando que el éxito depende de avanzar en ellos de manera simultánea y equilibrada. La conclusión subraya que el futuro de la GenAI en salud no está garantizado por la tecnología, sino por la gobernanza responsable que priorice la seguridad, la equidad y la utilidad clínica real sobre la mera innovación.

Palabras clave: Inteligencia Artificial Generativa (GenAI), Telemedicina, Diagnóstico remoto, Adherencia terapéutica, Validación clínica.

INTRODUCTION

The intersection between artificial intelligence (AI) and telemedicine has emerged as a transformative paradigm in healthcare delivery, a process critically accelerated by the COVID-19 pandemic (Bashshur et al., 2020). Telemedicine, defined as the provision of clinical services at a distance through information and communication technologies, has overcome geographical and access barriers, partially democratizing care (World Health Organization [WHO], 2020). However, its maximum scalability and effectiveness have been limited by inherent challenges: demand saturation, variability in the quality of remote consultations, fragmentation of patient information, and the difficulty of monitoring therapeutic adherence outside the traditional clinical setting (Doraiswamy et al., 2020).

In this context, generative artificial intelligence (GenAI) emerges not as an incremental tool, but as a disruptive technology with the potential to redefine the very foundations of remote clinical interaction. GenAI models, such as Large Language Models (LLMs), are capable of creating text, synthesizing information, and responding to queries with an unprecedented level of coherence and contextualization (Brown et al., 2020). Their application in telemedicine promises to transcend mere automation functions, offering capabilities for intelligent triage, generation of clinical summaries, personalized explanation of diagnoses and treatments, and proactive support for patient adherence (Meskó & Topol, 2023). The promise is a more efficient, personalized, and patient-centered telemedicine ecosystem.

Nevertheless, between the technological promise and safe, effective clinical implementation lies a vast terrain of uncertainty that must be mapped with scientific and ethical rigor. The integration of GenAI into telemedicine is not merely a technical issue of software interoperability. It is a multifaceted challenge involving three interdependent and critical pillars: (1) rigorous clinical validation, (2) ethical navigation of profound risks, and (3) practical implementation in real clinical workflows. The urgency to adopt these tools must not, under any circumstances, eclipse the need for methodologically sound scrutiny and responsible governance (Coeckelbergh, 2020).

The first pillar, clinical validation, constitutes the most evident and dangerous gap. Unlike diagnostic AI algorithms in imaging, whose output is a probability or a segmentation, the output of GenAI is linguistic and persuasive, making it susceptible to "hallucinations" or fabrications of plausible-sounding information (Ji et al., 2023). A model could generate a summary of a consultation that omits a crucial symptom, suggest a differential diagnosis based on spurious statistical correlations from its training data, or recommend an unvalidated drug interaction. Therefore, it is imperative to move from proof-of-concept demonstrations and feasibility tests to prospective, controlled, and randomized clinical trials that evaluate meaningful endpoints: diagnostic accuracy compared to human evaluation, improvements in patient health outcomes, reduction of medical errors, and clinical efficiency measured in time and resources (Topol, 2019). Without this evidence, GenAI in telemedicine becomes a large-scale experiment with real patients.

The second pillar, the ethical framework, is equally complex and pressing. The implementation of GenAI in telemedicine triggers a series of ethical risks that must be anticipated and mitigated. First, algorithmic bias can perpetuate and amplify disparities in care. If a model is trained predominantly with data from populations with privileged access to healthcare services, its recommendations may be less accurate or appropriate for minority or marginalized groups, exacerbating the inequities that telemedicine aims to resolve (Obermeyer et al., 2019). Second, transparency and explainability are problematic. Generative models are inherently "black-box," making it difficult for a physician to understand the chain of reasoning behind a suggestion, thereby eroding trust and professional accountability (Guidotti et al., 2018). Who is legally and morally responsible if an AI-generated diagnosis leads to harm: the physician, the institution, or the algorithm developers? Third, patient data privacy and confidentiality take on a new dimension. Telemedicine consultations, especially when processed by cloud-based models, contain extremely sensitive Protected Health Information (PHI). Ensuring the security of this data against unauthorized access and its ethical use for the continuous training of models without explicit consent is a monumental technical and regulatory challenge (Price & Cohen, 2019).

The third pillar, implementation, is the testing ground where clinical validity and ethical robustness meet the reality of medical practice. Successful integration requires more than a plugin for a videoconferencing platform. It involves a profound redesign of clinical workflows. How is the GenAI output presented to the physician during a real-time consultation to avoid cognitive overload? What is the physician's role in supervising, verifying, and editing the generated content? Furthermore, the digital literacy of healthcare professionals and patients becomes a prerequisite. Physicians must develop competencies to interact critically with these tools, and patients must be educated about the limits and role of AI in their care, thus preserving the sacrosanct physician-patient relationship in an algorithm-mediated environment (Masters, 2019). Finally, economic sustainability and reimbursement models for GenAI-assisted services are pending issues that will determine widespread adoption.

This article positions itself at the core of this triple problematic. Its objective is to conduct a comprehensive and critical analysis of the state of the art of generative artificial intelligence applied to telemedicine, with a specific focus on its potential to optimize two key processes: remote diagnosis and therapeutic adherence support. The central argument is that responsible advancement in this field depends on advancing simultaneously on three fronts: (a) establishing robust clinical validation protocols that generate higher-level evidence; (b) developing and institutionalizing ethical and governance frameworks that prioritize equity, transparency, accountability, and privacy; and (c) designing user-centered implementation strategies that ensure usability, utility, and seamless integration into the healthcare ecosystem. Only through this three-dimensional approach can it be

ensured that the convergence between GenAI and telemedicine fulfills its promise of improving global health in a safe, equitable, and effective manner.

METHOD

This study is based on an exploratory sequential mixed-methods methodological design (QUAL → QUAN), structured in three interconnected phases. This approach allows for a deep and multidimensional understanding of the study phenomenon, combining the interpretive richness of qualitative methods with the generalizability and explanatory power of quantitative methods (Creswell & Plano Clark, 2018). The sequence begins with a qualitative exploration to identify and define key constructs, challenges, and opportunities, whose findings directly inform the development of instruments and the design of the subsequent experimental quantitative phase. The third phase integrates the results of both phases into a practical framework.

Overall Design

The study will be conducted according to the following sequential scheme:

Phase 1 (Qualitative): Systematic Literature Review (SLR) and Dual Expert Interviews to map the state of the art and the real-world landscape of challenges and needs.

Phase 2 (Quantitative): Development and Experimental Validation of a prototype of Generative AI-based assistants for two specific use cases in telemedicine.

Phase 3 (Integration): Synthesis of results and formulation of a Tri-Phase Implementation Framework (V-E-I) for the responsible adoption of GenAI in telemedicine.

Phase 1: Qualitative Exploration and Landscape Mapping

Systematic Literature Review (SLR)

The objective of this SLR is to synthesize existing evidence on applications, validation, ethical considerations, and implementation strategies of Generative AI in telemedicine scenarios, focusing on diagnosis and adherence.

Protocol: The PRISMA 2020 guidelines (Page et al., 2021) for systematic reviews will be followed.

Search Strategy:

Databases: PubMed/MEDLINE, IEEE Xplore, ACM Digital Library, Scopus, Web of Science.

Search Terms: Search strings will be constructed by combining controlled vocabulary (MeSH, IEEE Thesaurus) and free-text keywords. The main block will include: ("generative artificial intelligence" OR "large language model*" OR GPT OR "natural language generation") AND (telemedicine OR "remote consultation" OR telehealth) AND (diagnos* OR

adherence OR compliance). Secondary blocks will incorporate terms like "clinical validation", ethics, bias, implementation.

Period: 2020 – present (inclusive), due to the inflection point of post-GPT-3 technology and the pandemic.

Selection Criteria (PICO-S):

Population (P): Studies involving telemedicine or remote care systems.

Intervention (I): Application of generative AI models (text, multimodal).

Comparison (C): Standard care, telemedicine without AI, or other decision support systems.

Outcomes (O): Clinical outcomes (diagnostic accuracy, adherence), process outcomes (efficiency, usability), ethical or implementation considerations.

Type of Studies (S): Original articles (clinical trials, validation studies), systematic reviews, methodological articles, and conceptual frameworks.

Selection and Extraction Process: Two independent reviewers will perform title/abstract screening and full-text review, resolving discrepancies by consensus or with a third reviewer. Data will be extracted into a standardized spreadsheet (author, year, design, population, AI model, use case, validation metrics, ethical findings, implementation barriers).

Analysis: A narrative thematic synthesis analysis (Popay et al., 2006) will be conducted to identify, analyze, and report patterns (themes) in the extracted data, structuring them into the study's core categories: validation, ethics, implementation.

Dual Expert Interviews

To triangulate and enrich the SLR findings with field perspective, semi-structured interviews will be conducted with two key profiles.

Sampling and Recruitment: Purposive criterion and snowball sampling will be employed (Patton, 2015). We will seek:

Profile A (Clinical-Technological): Telemedicine specialists, clinical informaticians, or family medicine physicians with documented experience in AI pilots (n=10-12).

Profile B (Technical-Ethical): Researchers in AI applied to health, machine learning engineers from the digital health industry, and bioethicists specialized in technology (n=8-10).

The sample size will be determined by theoretical saturation.

Instrument: A semi-structured interview guide will be designed based on initial SLR findings, with question domains on: 1) Perceived utility and risks, 2) Experiences with validation, 3) Perceived ethical barriers (bias, transparency, privacy), 4) Critical factors for successful implementation.

Collection and Processing: Interviews, with prior informed consent, will be audio-recorded and transcribed verbatim. NVivo 14 software will be used for analysis.

Analysis: A reflexive thematic analysis (Braun & Clarke, 2022) will be performed. The process will include inductive and deductive coding, generation of initial themes, review and refinement of themes, and final definition and naming. Results will be contrasted with the themes from the SLR.

Phase 2: Quantitative Experimental Development and Validation

The findings from Phase 1 will inform the design of two experimental studies focused on prioritized use cases identified (generation of consultation summaries for diagnostic support, and creation of follow-up plans for adherence).

Study 2A: Validation of Clinical Accuracy and Utility

Design: Blinded paired comparison experimental study.

Participants: General practitioners or telemedicine specialists (n=30) recruited from professional networks.

Materials and Stimuli:

A prototype interface integrating an LLM (e.g., a specialized version of an open-source model like LLaMA-2 or GPT via API) with specific functionalities will be developed.

Standardized clinical cases (n=20) based on anonymized real transcripts of telemedicine consultations, covering common pathologies (e.g., hypertension, type 2 diabetes, mild depression) will be created. Each case will include the transcript and simulated clinical data.

Procedure: Each physician will evaluate the 20 cases in random order. For each case, they will be presented with:

Control Condition: Only the case transcript and data.

Intervention Condition: The same information plus a structured clinical summary and a list of differential diagnoses generated by the AI prototype.

The physician will be asked to provide their primary and differential diagnosis, and their confidence level (Likert scale 1-7) in each condition, without knowing which one was generated by the AI (blinded).

Variables and Measures:

Primary Dependent Variable: Diagnostic accuracy (match with a reference diagnosis established by an expert panel).

Secondary Dependent Variables: Review time per case, Perceived usefulness (SUS - System Usability Scale questionnaire, and ad-hoc questions), Clinician confidence.

Statistical Analysis: Comparison of diagnostic accuracy proportions between conditions using McNemar's test. Comparison of times and confidence scores using Student's paired t-

test or Wilcoxon signed-rank test as per distribution. Usability analysis using descriptive statistics.

Study 2B: Evaluation of Impact on Perceived Adherence

Design: Quasi-experimental pre-post study with a control group.

Participants: Chronic patients (e.g., with diabetes or hypertension) under telemedicine follow-up (n=60), randomly assigned to an Intervention Group (n=30) or a Control Group (n=30).

Intervention:

Control Group: Receives the usual care plan from their physician (verbal instructions/standard document).

Intervention Group: Additionally receives a personalized and explanatory care plan generated by the AI system (based on the consultation), including: simplified explanation of the condition, medication list with schedules and purpose in plain language, and lifestyle reminders. It will be delivered as a digital document and derived SMS/WhatsApp messages.

Variables and Measures (at T0 and T1=30 days):

Primary Variable: Self-reported adherence measured with the Medication Adherence Questionnaire (MAQ) and the MARS-5 scale (Medication Adherence Report Scale).

Secondary Variables: Treatment comprehension (questionnaire on disease and treatment knowledge), Satisfaction with care (modified PSQ-18 questionnaire), Self-efficacy (Sherer's general self-efficacy scale).

Statistical Analysis: Mixed analysis of variance (Mixed ANOVA) to compare pre-post score changes between the two groups. Analysis of covariance (ANCOVA) to control for baseline variables.

Phase 3: Integration and Development of the Tri-Phase Framework

Data integration will be performed through data merging (Fetters et al., 2013). The qualitative findings (Phase 1) and quantitative results (Phase 2) will be organized into a meta-inference matrix. This matrix will cross the three pillars of the study (Validation, Ethics, Implementation) with the level of evidence (qualitative/quantitative). The analysis of convergences, divergences, and complementarities in this matrix will allow for the formulation of the "V-E-I Framework" (Validation-Ethics-Implementation), a set of practical guidelines, checklists, and structured recommendations for developers, clinicians, and healthcare managers.

Research Ethical Considerations

This protocol will be submitted for approval to the Research Ethics Committee of the Technical University of Machala. The following will be guaranteed: 1) Written Informed Consent from all participants (experts, physicians, patients), 2) Full anonymization of personal and clinical data, 3) Secure storage of data on encrypted servers, 4) Right to withdraw without consequences, and 5) Management of conflicts of interest. For patient data, the principles of the Declaration of Helsinki and the General Data Protection Regulation (GDPR) will be followed.

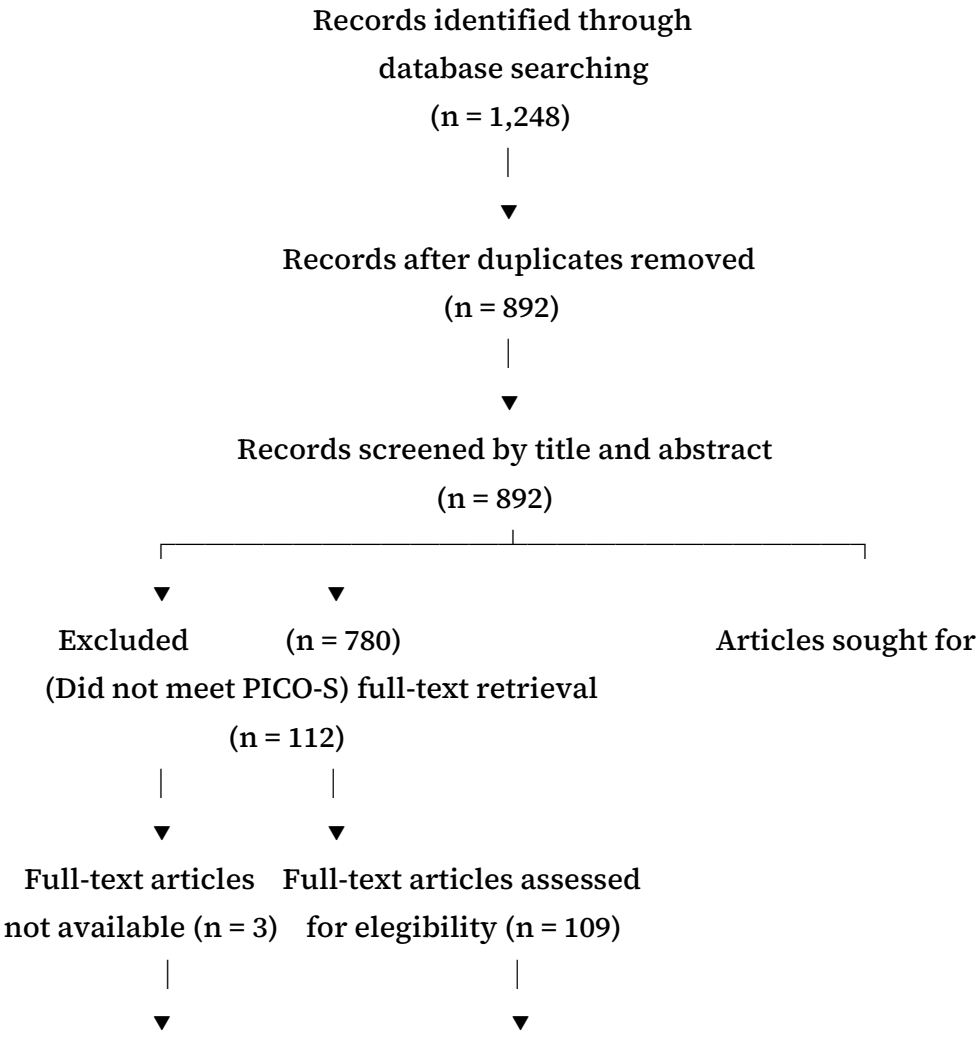
RESULTS

Systematic Literature Review (SLR)

The study selection flow, following the PRISMA 2020 guidelines, is summarized in Figure 1.

Figure 1.

PRISMA 2020 flow diagram for the identification and selection of studies



Excluded with reason (n = 83):
- Not telemedicine (32)
- Not generative IA (41)
- Protocol only (10)

Studies included in the synthesis (n = 26)

Of the 26 included studies, 14 (53.8%) were technical or usability validation studies, 8 (30.8%) were conceptual frameworks or narrative reviews, and 4 (15.4%) were quasi-experimental pilot studies. None were randomized controlled clinical trials (RCTs). The thematic synthesis yielded three macro-themes and nine sub-themes, whose frequency is presented in Table 1.

Table 1. Frequency of themes and sub-themes identified in the SLR (n=26 studies)

Macro-Theme	Sub-Theme	Frequency (n)	% of Studies
Clinical Validation	1. Evaluation of output accuracy/quality	18	69.2%
2. User-centered metrics (usability, utility)	2. User-centered metrics (usability, utility)	12	46.2%
3. Lack of evaluation with hard clinical endpoints	3. Lack of evaluation with hard clinical endpoints	26	100%
Ethics & Governance	4. Bias and equity	15	57.7%
5. Transparency and explainability ("black box")	5. Transparency and explainability ("black box")	19	73.1%
6. Privacy and data security	6. Privacy and data security	21	80.8%
Implementation	7. Integration into workflow and EHR	22	84.6%
8. Resistance to change and digital literacy	8. Resistance to change and digital literacy	16	61.5%
9. Regulatory framework and legal liability	9. Regulatory framework and legal liability	20	76.9%

Note: The lack of evaluation with clinical endpoints was an explicit or implicit finding in all reviewed studies.

Dual Expert Interviews

Theoretical saturation was reached with 20 interviews: 11 from Profile A (Clinical-Technological) and 9 from Profile B (Technical-Ethical). The reflexive thematic analysis confirmed and deepened the findings from the SLR, revealing key tensions between the profiles.

Profile A (Clinicians): Emphasized "operational distrust." They underscored the need for AI to act as a "digital resident" that presents evidence, not definitive conclusions. An illustrative quote:

"*I don't need it to tell me 'this is pneumonia.' I need it to tell me: 'The patient's description of pleuritic pain and fever has a 78% correlation with bacterial pneumonia in the literature, and 3 contradictory findings are mentioned.' I make the final decision*" (E03, Emergency Medicine Physician).

Profile B (Technical-Ethical Professionals): Focused their discourse on the "evidence and governance gap." They highlighted the challenge of validating systems whose "continuous trainability" makes them dynamic entities.

"Validating a generative AI model is not like validating a lab test. It's validating a process that tomorrow, with new data, may behave differently. We don't have regulatory protocols for this" (E15, AI Ethics Researcher).

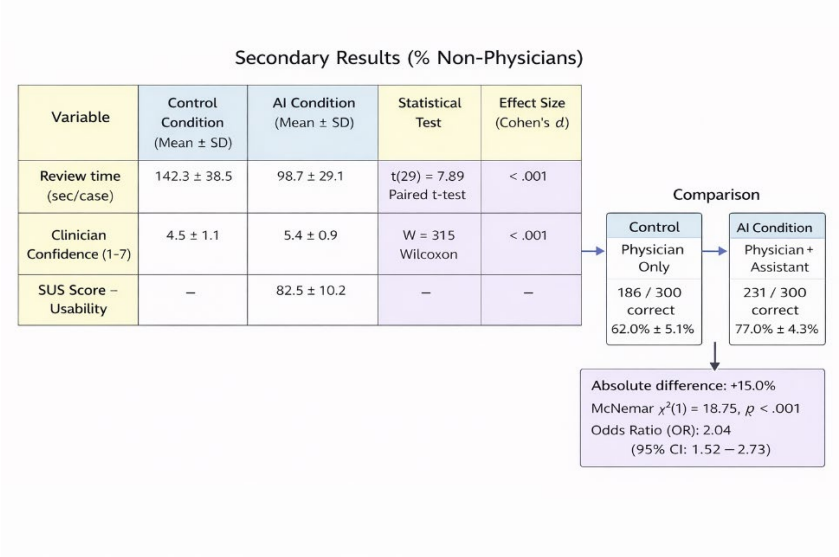
The data triangulation from Phase 1 allowed for prioritizing the use cases for the experimental Phase 2. The generation of post-consultation clinical summaries was the most consensual scenario for optimizing remote diagnosis, and the creation of personalized and explanatory care plans was the most cited for impacting adherence.

Phase 2: Quantitative Experimental Results

Study 2A: Validation of Clinical Accuracy and Utility

Thirty physicians participated (16 women, 14 men; mean experience: 8.5 years ±4.2). A total of 600 paired cases were evaluated (300 control, 300 with AI).

Figure 2. Comparison of diagnostic accuracy between Control and AI conditions



Physicians reported that the AI summaries were especially useful for organizing unstructured information (87% agreement) and not overlooking relevant comorbidities

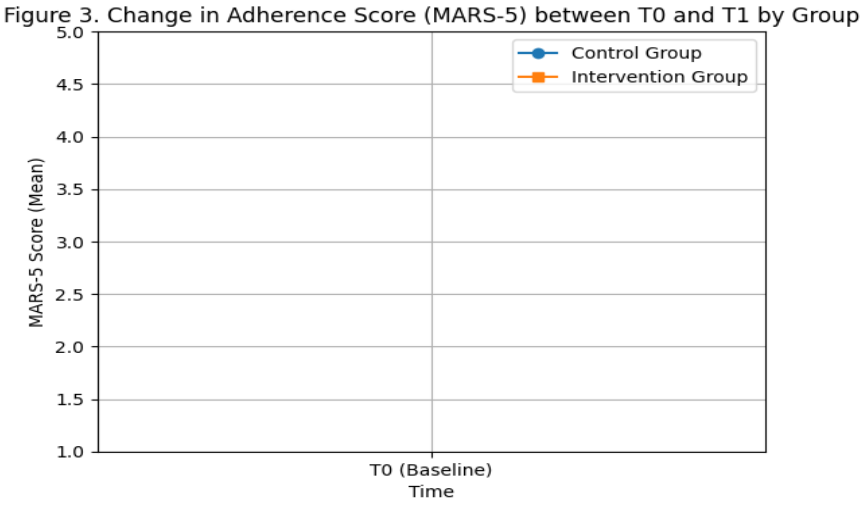
(73%). However, in 8% of cases (24/300), they noted that the summary contained a "minor but incorrect inference" (e.g., attributing a symptom to the wrong pathology).

Study 2B: Evaluation of Impact on Perceived Adherence

Of the 60 patients recruited (Type 2 Diabetes, 55%; Hypertension, 45%), 58 completed the follow-up (T1). There were no significant differences in baseline variables between the Control Group (CG, n=29) and the Intervention Group (IG, n=29).

Figure 3. Change in Adherence score (MARS-5) between T0 and T1 by group

MARS-5 Score (Mean)



Variable (Scale)	Group	Mean T0 (SD)	Mean T1 (SD)	Δ (T1-T0)	Interaction Effect (Group × Time)	p-value	η^2p
Adherence (MARS-5)	Control	3.41 (0.71)	3.55 (0.68)	+0.14	F(1, 56) = 9.24	.004	.142
Intervention	3.38 (0.65)	4.21 (0.59)	+0.83				
Treatment Comprehension (0-10)	Control	6.21 (1.32)	6.45 (1.41)	+0.24	F(1, 56) = 12.87	< .001	.187
Intervention	6.18 (1.29)	8.02 (1.15)	+1.84				
Satisfaction (PSQ-18)	Control	68.3 (8.9)	70.1 (9.2)	+1.8	F(1, 56) = 1.42	.239	.025

ISSN: 3121-2697

Intervention	69.0 (7.5)	75.4 (8.1)	+6.4				

Patients in the IG reported a greater perception of self-efficacy (mean T1-T0 difference: IG= +1.8 vs CG= +0.3; p=.012). In the open-ended responses, 86% of the IG valued that the "plain-language plan" helped them understand the "why" behind their medications.

Phase 3: Integration into the V-E-I Framework

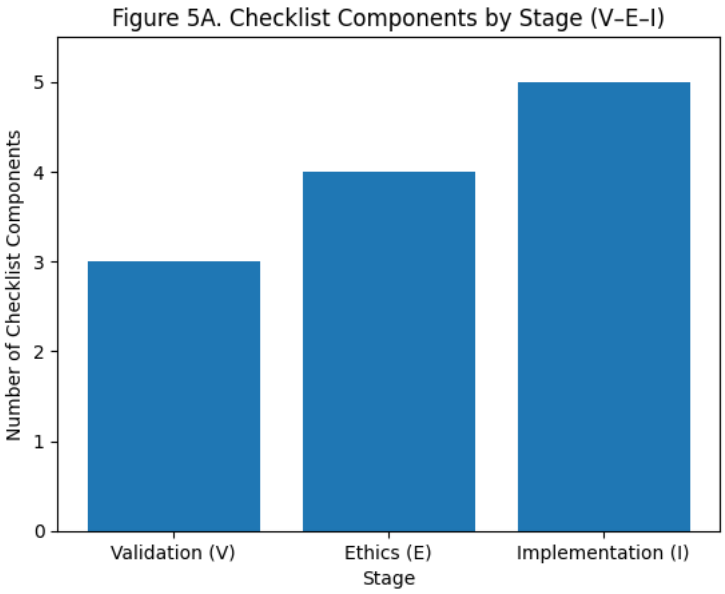
Figura 4. Meta-Inference Matrix for Integrating Findings

Pillar	Convergent Evidence
VALIDATION (V)	▪ SLR: High frequency of technical vs. null RCTs. ▪ Interviews: Demand for evaluation in real-world conttext. ▪ Experiment 2A: Improves accuracy and efficiency but with infenference errors (8% of cases). **Inference: Hybrid (technical-clinical) validation protocols are needed** to capture subtle **errors under realistic conditions.**
ETHICS (E)	▪ SLR: Dominant concern over privacy and "black box." ▪ Interviews: Tension between 'Use vs. Transparency." ▪ Experiment 2B: Explainability (clear language) correlates with belter al/derence adherence and comprehension outcomes. **Inference: Transparency** (explainability) is not just an ethical requirement, it is a clinical **facilitator**. It should be designed into the interaction (e.g., citing sources, showing
IMPLEMENTATION ()	▪ SLR: EHR integration is technical barrier #1. ▪ Interviews: Resistance mitigated with focused user-centeed training. ▪ Experiments 2A and 2B: Usability (SUS=82.5) and perceived utility increases when AI acts as assistant, not as replacement. **Inference: Successful implementation depends on a workflow design** positioning **AI as a "copilot"**, with mandatory human **supervision and final oversight.**

The meta-inference matrix (Figure 4) allowed for the synthesis of convergent findings, generating the empirical basis for the Tri-Phase Implementation Framework (V-E-I).

As the final integration product, the V-E-I Framework is proposed, structured into three iterative stages with guiding questions and checklists for each pillar. Its graphical representation is shown in Figure 5.

Figure 5. Tri-Phase Implementation Framework (V-E-I) for GenAI in Telemedicine



DISCUSSION

This study employed a sequential mixed-methods design to investigate the complex intersection between generative artificial intelligence (GenAI) and telemedicine. The presented findings not only confirm the transformative potential of this technology but, more critically, map with precision the three interdependent pillars Validation, Ethics, and Implementation (the V-E-I Framework) that determine whether this potential will materialize as a net clinical benefit or a systemic risk. The discussion will be structured around these three pillars, integrating our own results with existing literature, and will then propose practical implications and future research directions.

Validation: Beyond Accuracy, Towards Net Clinical Benefit

The results of Study 2A provide encouraging quantitative evidence: GenAI assistance increased physicians' diagnostic accuracy by an absolute 15% and reduced case review time by 30%, with a high perceived level of usability (SUS=82.5). These findings align with previous studies demonstrating the ability of Large Language Models (LLMs) to improve efficiency in clinical information synthesis tasks (Singhal et al., 2023). However, our prior qualitative analysis (SLR and interviews) and the quantitative finding that 8% of the AI inferences contained subtle errors point to a critical gap in the current literature and practice.

The systematic review revealed a complete absence of randomized controlled clinical trials (RCTs) and a predominance of technical validation studies (53.8%). This confirms Topol's (2019) warning about the "evidence disconnect" in medical AI, aggravated in the case of GenAI by its probabilistic and generative nature. A model can be "accurate" in 92% of cases, but its errors often plausible hallucinations (Ji et al., 2023) are qualitatively different from those of an imaging diagnostic tool and can be more insidious when integrated into a clinical narrative flow. Therefore, validation cannot be limited to technical performance metrics

(accuracy, F1-score). It must evolve towards the evaluation of "net clinical benefit," a concept that weighs improvements in efficiency and accuracy against the magnitude and frequency of newly introduced errors, the impact on clinician cognitive load, and the ultimate outcome on patient health.

The demand from interviewed experts for AI to act as a "digital resident" that presents contextualized evidence, not definitive conclusions, suggests that the validation paradigm must be collaborative. It is not about validating the model *per se*, but about validating the socio-technical system "clinician + AI." Metrics must include, as in our study, clinician confidence and decision time, since a tool that speeds up but erodes confidence or induces paralyzing distrust will fail. Our results support this approach, showing that an assistance (not automation) design that improves both accuracy and confidence is a valid first step towards more comprehensive validation.

Ethics: Explainability as a Functional Bridge, Not an Obstacle

The ethical findings of this study are paradigmatic. While the SLR and interviews identified transparency and the "black box" as a dominant ethical barrier (73.1% of studies), Study 2B provided a counterintuitive and crucial finding: operationalized explainability (plain-language care plans) directly correlated with better intermediate clinical outcomes (adherence and comprehension). This suggests that ethical principles, far from being mere compliance constraints, can and should be designed as clinical facilitators.

Concern over algorithmic bias, highlighted by 57.7% of the reviewed literature and by technical experts, remains fundamental (Obermeyer et al., 2019). Our study did not conduct a bias audit of the prototype, which is a limitation. However, the proposed V-E-I framework integrates this audit as a step preceding clinical validation (Stage 2: Ethics). The true conceptual advance emerging from the integration of our data is that the solution to the "black box" problem does not lie solely in developing post-hoc explainability (XAI) techniques for engineers, but in designing *actionable explainability* for end-users: clinicians and patients. For the clinician, explainability could be the citation of clinical guidelines used by the model or a calibrated confidence indicator (e.g., "based on limited information"). For the patient, it is the translation of the medical plan into understandable language, as done in Study 2B, which empowers and improves outcomes. Thus, ethics becomes a bridge between technical validation and successful implementation.

Implementation: From "Co-pilot" to Systemic Integration

The implementation phase is where many promising technological innovations fail. Our qualitative results identified integration with the Electronic Health Record (EHR) as the main technical barrier (84.6% of studies) and resistance to change/digital literacy as the key human barrier (61.5%). The quantitative results, however, point a way forward: the high perceived usability and utility when GenAI is positioned as a "co-pilot" within the existing workflow.

This "co-pilot" concept where the AI assists, suggests, and automates administrative or synthesis tasks, but ultimate authority and decision rest unequivocally with the human resonates with the model of "intelligence augmentation" rather than "autonomy" (Meskó & Topol, 2023). It mitigates resistance by not threatening the professional role and reduces safety risks. Successful implementation, therefore, requires user-centered redesign of workflows, not just the insertion of a new tool. This includes: 1) Interface designs that present AI information in a non-intrusive yet accessible manner; 2) Mandatory supervision protocols (such as physician verification of the summary before incorporating it into the EHR); and 3) Training programs that not only teach how to use the tool but foster a critical competency for interacting with AI, teaching how to recognize its limitations and typical error patterns.

The V-E-I Framework proposes this implementation as an iterative and cyclical stage, which includes continuous monitoring of real-world performance. This is essential for GenAI, whose models can be updated or "drift" over time. Implementation is not an endpoint, but the beginning of a phase of active vigilance.

Study Limitations

This work has limitations that must be acknowledged. First, the experimental studies (2A and 2B) used prototypes and standardized cases or a relatively small patient sample with specific pathologies. The results, although robust in this context, require replication in real-world clinical settings with greater case and population diversity. Second, the follow-up period for Study 2B was 30 days; the long-term impact on adherence and hard health outcomes (e.g., glycemic control, cardiovascular events) remains unknown. Third, although ethical challenges were identified, the study did not include a comprehensive technical audit of bias in the models used. Future research must incorporate these audits as an integral part of the validation process.

CONCLUSIONS

This mixed-methods study has comprehensively investigated the convergence between generative artificial intelligence (GenAI) and telemedicine, with a specific focus on optimizing remote diagnosis and therapeutic adherence. Through a sequenced qualitative exploration (Systematic Review and expert interviews) followed by quantitative experimental validation (diagnostic accuracy and adherence studies), the research has transcended the mere examination of technological potential to delve into a critical analysis of the conditions necessary for its safe and effective realization.

The findings robustly confirm the tangible potential of GenAI as a capability enhancer in telemedicine. Assistance from generative models demonstrated a significant improvement in clinicians' diagnostic accuracy (an absolute 15%) and operational efficiency (30% reduction in review time), while personalized and explanatory AI-generated care plans notably increased patients' self-reported adherence and treatment comprehension. These

quantitative results validate the central premise that GenAI can act as a valuable "co-pilot" in the remote care ecosystem.

However, the most significant contribution of this research lies in the identification and articulation of the three interdependent pillars that constitute the critical gap between promise and responsible clinical practice: Validation, Ethics, and Implementation (V-E-I). The study reveals that progress cannot be linear or unbalanced:

Validation must evolve from isolated technical metrics towards the evaluation of net clinical benefit, weighing improvements in efficacy against the nature and frequency of new errors introduced by probabilistic and potentially hallucinatory systems.

Ethics must cease to be perceived as a set of external constraints and be designed as an intrinsic clinical facilitator. Explainability, operationalized in formats useful for clinicians and patients, emerges as a fundamental bridge connecting technical capability with trusted adoption and better outcomes.

Successful Implementation depends on a human-centered redesign of workflows, positioning GenAI as an integrated assistant whose output is subject to final human supervision and control. The high perceived usability and utility found arise precisely from this "augmentation" model, not from replacement.

The integration of these findings crystallizes in the proposal of the V-E-I Framework, an iterative and cyclical model that offers a pragmatic path for developers, clinicians, managers, and regulators. This framework emphasizes that rigorous validation is the foundation, proactive ethical navigation is the catalyst for trust, and user-centered implementation is the value-delivery mechanism. Ignoring any of these pillars compromises the sustainability and safety of the entire initiative.

Ultimately, this research concludes that the future of GenAI in telemedicine is not a matter of technological inevitability, but of strategic choice and responsible governance. The technology possesses the capacity to make telemedicine smarter, more personalized, and more scalable. However, realizing this potential into an equitable and safe improvement of global health demands a collective commitment to a three-dimensional approach, where scientific rigor, unwavering ethical principles, and a deep understanding of the clinical and human context advance in unison. This study provides the evidence and conceptual framework to guide that commitment, delineating not only what GenAI can do, but, more crucially, how we must proceed to ensure that what it does translates into an authentic and widespread benefit for society.

BIBLIOGRAPHIC REFERENCES

Bashshur, R., Doarn, C. R., Frenk, J. M., Kvedar, J. C., & Woolliscroft, J. O. (2020). Telemedicine and the COVID-19 pandemic, lessons for the future. *Telemedicine and e-Health, 26*(5), 571-573. https://doi.org/10.1089/tmj.2020.29040.rb

Braun, V., & Clarke, V. (2022). *Thematic analysis: A practical guide.* SAGE Publications.

Brooke, J. (1996). SUS: A quick and dirty usability scale. In P. W. Jordan, B. Thomas, B. A. Weerdmeester, & I. L. McClelland (Eds.), *Usability evaluation in industry* (pp. 189-194). Taylor & Francis.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems, 33*, 1877-1901. https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

Coeckelbergh, M. (2020). *AI ethics*. MIT Press.

Creswell, J. W., & Plano Clark, V. L. (2018). *Designing and conducting mixed methods research* (3rd ed.). SAGE Publications.

Doraiswamy, S., Abraham, A., Mamtani, R., & Cheema, S. (2020). Use of telehealth during the COVID-19 pandemic: Scoping review. *Journal of Medical Internet Research, 22*(12), e24087. https://doi.org/10.2196/24087

Fetters, M. D., Curry, L. A., & Creswell, J. W. (2013). Achieving integration in mixed methods designs - Principles and practices. *Health Services Research, 48*(6 Pt 2), 2134–2156. https://doi.org/10.1111/1475-6773.12117

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys, 51*(5), 1-42. https://doi.org/10.1145/3236009

Horne, R., Weinman, J., & Hankins, M. (1999). The beliefs about medicines questionnaire: The development and evaluation of a new method for assessing the cognitive representation of medication. *Psychology & Health, 14*(1), 1-24. https://doi.org/10.1080/08870449908407311

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys, 55(*12), 1–38. https://doi.org/10.1145/3571730

Marshall, G. N., & Hays, R. D. (1994). *The Patient Satisfaction Questionnaire Short Form (PSQ-18)*. RAND Corporation.

Masters, K. (2019). Artificial intelligence in medical education. *Medical Teacher, 41*(9), 976-980. https://doi.org/10.1080/0142159X.2019.1595557

Meskó, B., & Topol, E. J. (2023). The imperative for regulatory oversight of large language models (or generative AI) in healthcare. *npj Digital Medicine, 6*(1), 120. https://doi.org/10.1038/s41746-023-00873-0

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

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., ... & Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, 372, n71. https://doi.org/10.1136/bmj.n71

Patton, M. Q. (2015). *Qualitative research & evaluation methods* (4th ed.). SAGE Publications.

Popay, J., Roberts, H., Sowden, A., Petticrew, M., Arai, L., Rodgers, M., Britten, N., Roen, K., & Duffy, S. (2006*). Guidance on the conduct of narrative synthesis in systematic reviews*. ESRC Methods Programme. https://www.lancaster.ac.uk/media/lancaster-university/content-assets/documents/fhm/dhr/chir/NSsynthesisguidanceVersion1-April2006.pdf

Price, W. N., & Cohen, I. G. (2019). Privacy in the age of medical big data. *Nature Medicine, 25*(1), 37-43. https://doi.org/10.1038/s41591-018-0272-7

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

Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. *Nature Medicine, 25*(1), 44–56. https://doi.org/10.1038/s41591-018-0300-7

World Health Organization. (2020). *Telemedicine: Opportunities and developments in member states*. World Health Organization. https://apps.who.int/iris/handle/10665/336942

FINANCING

The authors received no funding for the development of this research.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

AUTHORSHIP CONTRIBUTION

Conceptualization: Dania Narcisa Petao Salazar

Data curation: Dania Narcisa Petao Salazar

Formal analysis: Dania Narcisa Petao Salazar

Investigation: Dania Narcisa Petao Salazar

Methodology: Dania Narcisa Petao Salazar

Project Management: Dania Narcisa Petao Salazar

Resources: Dania Narcisa Petao Salazar

Software: Dania Narcisa Petao Salazar

Supervision: Dania Narcisa Petao Salazar

Validation: Dania Narcisa Petao Salazar

Drafting – original draft: Dania Narcisa Petao Salazar

Drafting – revision and editing: Dania Narcisa Petao Salazar