Global Learning Agenda on Clinical Decision Support Systems

PATH

Table of Contents

Acronyms

AI
artificial intelligence

ASHA
Accredited Social Health Activist

ASR
automated speech recognition

CDSS
clinical decision support system

CHEWA
Community Health Extension Worker Assistant

CoP
community of practice

EU
European Union

GHD
Global Health Delivery Project at Harvard

GMLP
Good Machine Learning Practice

FDA
U.S. Food and Drug Administration

LMIC
low-and middle-income country

LLM
large language model

MHRA
Medicines and Healthcare products Regulatory Agency

ML
machine learning

SAHPRA
South African Health Products Regulatory Authority

SLM
small language model

STT
speech-to-text

TTS
text-to-speech

WHO
World Health Organization

Introduction

The integration of artificial intelligence (AI) technologies into clinical practice stands as a defining challenge and opportunity for global health in the coming decade. While use cases such as clinical documentation and diagnostic support are becoming increasingly common in high-resource settings, where over half of clinicians now report using AI tools, much less is known about frontline clinicians' experiences, barriers, and real-world needs in low- and middle-income countries (LMICs).

In LMIC contexts, large language model (LLM)-enabled clinical decision support systems (CDSSs) offer the potential to extend high-quality decision support to settings with limited human and technical resources. Yet they also carry significant risks, including language bias, inequitable access, limited local capacity for oversight, and the possibility of recommendations that are misaligned with local medical guidelines that undermine patient safety or trust.

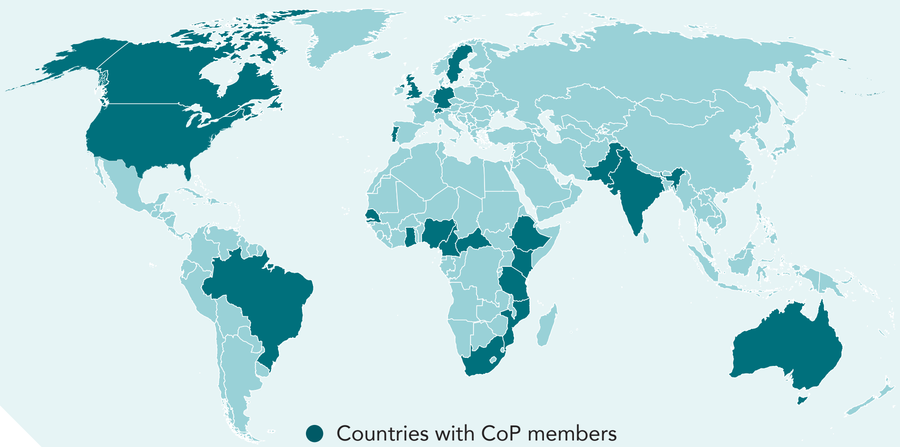
Over the past two years, rapid advances in generative AI have outpaced the establishment of evidence, governance, and capacity frameworks required for responsible use. Health systems and policymakers are now faced with urgent questions:

- How can AI tools be evaluated for clinical safety and contextual relevance?
- What data infrastructures and regulatory mechanisms are needed for equitable deployment?
- And how can local researchers and health institutions lead, not just participate, in AI innovation?

These questions form the basis of this CDSS Learning Agenda, which aims to generate collective insight and actionable evidence to guide safe, effective, and equitable integration of AI into health systems. The agenda identifies six thematic priorities—localization and language equity; evaluation and real-world evidence; voice-enabled and multimodal tools; capacity and local ownership; governance and trust; and infrastructure and enabling environments—that collectively define the ecosystem required for sustainable AI adoption in LMICs.

Purpose & Origin

Global Representation of CoP Members



The CDSS Learning Agenda originated through the LLMs for Health Equity community of practice (CoP), convened by PATH and partners with support from the Bill & Melinda Gates Foundation. The CoP brings together researchers, implementers, and policymakers from more than 60 institutions across Africa, Asia, and Latin America to share experiences, evidence, and challenges in applying large language models to clinical and public health use cases.

Over the past year, the CoP has functioned as a collaborative learning platform for understanding how LLM-based tools are being developed and tested in LMIC settings. Through structured learning sessions, member-led presentations, and collaborative evaluation clinics, participants identified common evidence gaps and areas of inquiry critical to responsible innovation.

To capture and synthesize these insights, PATH and partner institutions developed a shared learning agenda to document the work CoP members have undertaken to date. This includes describing their LLM-enabled tools, how these tools fit into the real-world workflows and needs of frontline health workers in LMICs, and the types of evidence generated through their implementation and research.

The process combined narrative submissions from CoP members, desk reviews of emerging literature, and expert consultations with evaluation and AI ethics specialists. The outputs were synthesized into this manuscript using a consortium author model that ensures equal credit to all contributing organizations.

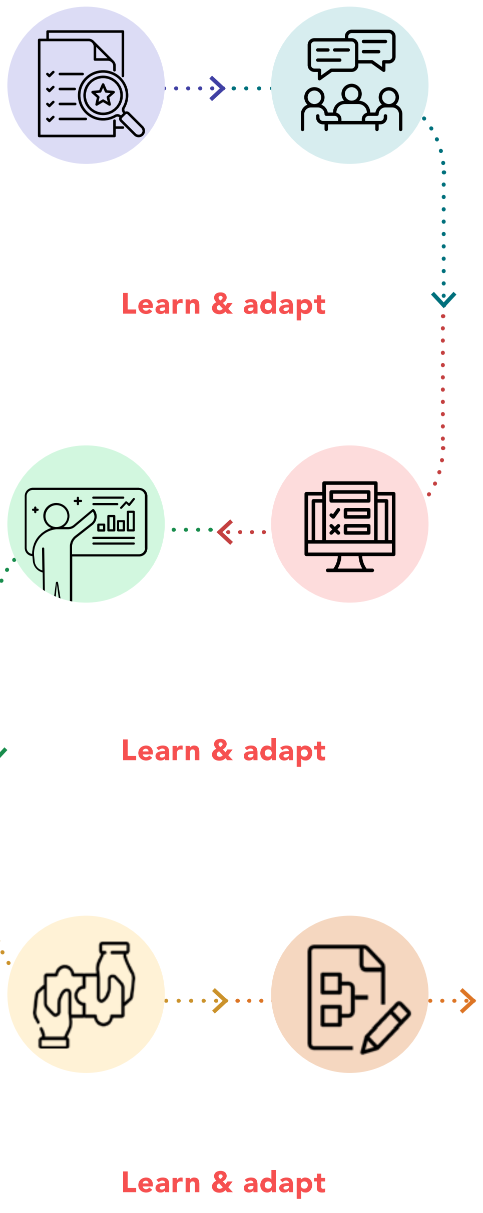
The purpose of this learning agenda is threefold:

1. To synthesize existing evidence on the design, implementation, and evaluation of LLM-enabled CDSSs in LMICs.
2. To identify critical learning questions and knowledge gaps that must be addressed to ensure equitable and responsible scale-up of these tools.
3. To provide a roadmap for collaborative research and policy engagement, supporting governments, donors, and implementers to make informed, evidence-driven decisions about AI for health.

By grounding this agenda in real-world experiences and lessons from across the CoP, the project moves beyond abstract debate toward actionable, locally informed guidance for the global health community.

Methodology

Co-Creation Process for the LLMs for Health Equity Learning Agenda



1. Desk review of global evidence

Conducted a comprehensive review of published and grey literature on LLMs and CDSSs, identifying common challenges and emerging innovations across low- and middle-income contexts.

2. Community of Practice (CoP) collaborative validation

Preliminary findings were presented and co-reviewed during the CoP's launch meeting, using breakout discussions and live annotation to refine and contextualize the initial themes.

3. Monthly learning sessions and case sharing

Ongoing CoP learning sessions featured innovators and researchers presenting cases that deepened understanding of each theme, surfacing lessons, challenges, and opportunities for collaboration.

4. Community-wide survey

An online survey captured member feedback to validate the themes and identify additional evidence gaps related to data, the health care workforce, and equitable access.

5. Evidence synthesis and thematic coding

PATH and partners systematically reviewed CoP materials, coding reports, tools, and frameworks by theme to identify converging evidence and persistent gaps.

6. Collaborative manuscript development

Members contributed short case narratives that were synthesized into a collective manuscript through iterative drafting and review, ensuring balanced representation and shared ownership.

The development of this learning agenda was grounded in both academic evidence and community knowledge. From the outset, the goal was to ensure that the identified themes reflected not only published research on CDSSs but also the practical experience of implementers and researchers across LMICs.

We began with a comprehensive desk review of global and regional evidence on LLMs and CDSSs. This included peer-reviewed studies, implementation reports, and grey literature on emerging AI applications in health. The review highlighted recurring challenges such as the lack of representative datasets in under-resourced languages, the technical difficulties of integrating AI tools into existing clinical workflows and digital systems, and unresolved governance questions. It also surfaced promising innovations, including early work on voice-enabled tools, community-driven data efforts, and multimodal approaches to overcome literacy barriers. This initial synthesis provided a strong evidence base but also revealed a need to validate and contextualize these insights through practitioner experience.

The preliminary findings were shared at the LLMs for Health Equity CoP kick-off session in September 2024, which convened participants representing ministries of health, academic institutions, and implementing organizations. Rather than a one-way presentation, this session used breakout discussions and collaborative annotation to co-create the agenda. Participants reflected on the draft themes, shared real-world examples, and highlighted context-specific challenges that literature alone could not capture. Their contributions underscored the importance of grounding this agenda in the realities of LMICs, where digital health systems, infrastructure, and workforce readiness vary widely.

To expand participation, the CoP team circulated an online survey inviting all members to rank priority themes and identify additional evidence gaps. Responses from CoP members emphasized the need for locally relevant datasets, secure and equitable data-sharing mechanisms, and greater focus on workforce capacity, including developing digital skills among frontline health workers, strengthening clinicians’ ability to safely use and interpret AI-enabled tools, and building institutional capacity for evaluation, oversight, and responsible deployment. These findings informed the refinement of the thematic structure and validated the community’s ownership of the process.

Following this consultative phase, the CoP embarked on a year of collective exploration and evidence-building. Monthly learning sessions featured case presentations from innovators, academic researchers, and implementers across Africa, Asia, and Latin America. Each session focused on one of the emerging themes and combined short presentations with open discussions to identify common challenges, collaboration opportunities, and practical lessons.

To ensure systematic knowledge capture, PATH and partner institutions collected and triangulated a wide range of materials shared through the CoP, including reports, study protocols, evaluation frameworks, and white papers. These were reviewed and coded by theme to surface shared learning and evidence gaps.

Building on this evidence base, the CoP adopted a consortium author approach for manuscript development. Members submitted short narratives describing their work, research, or implementation experiences aligned with one or more learning themes. These contributions were harmonized through an iterative synthesis process and reviewed collectively to ensure balanced representation and accuracy.

The resulting agenda is not a static document, but a living framework that evolves as new insights emerge. Each learning theme integrates global evidence with real-world examples, providing both high-level guidance and practical pathways for research, policy, and implementation.

Global Learning Themes

Drawing on the participatory process described above, the synthesis of literature, community discussions, and member submissions revealed six interrelated learning themes that together define the current state and future direction of LLM-enabled CDSSs. These themes emerged not as isolated technical topics, but as interconnected areas where evidence, practice, and policy intersect. They capture both the progress achieved by early adopters and the persistent challenges that continue to shape implementation in LMICs.

1. Localization & Language Equity

Achieving language equity is essential to ensure that LLM-enabled CDSSs are safe, trustworthy, and relevant across diverse contexts. Progress depends on expanding multilingual datasets, advancing inclusive model training, and embedding language governance mechanisms that respect cultural nuance. Strengthening these foundations will allow global and local actors to create digital tools that truly reflect the voices and needs of the populations they serve.

2. Voice-Enabled & Multimodal Tools

Voice and multimodal capabilities offer a path toward more inclusive, human-centered digital health systems. By leveraging speech recognition, conversational interfaces, and image-based inputs, LLMs can make CDSSs accessible to users with low literacy, simple devices, or limited connectivity. Continued experimentation and co-design are needed to refine these tools for use in frontline and low-resource environments.

3. Evaluation & Real-World Evidence

Robust evaluation remains a cornerstone of responsible LLM adoption. Shared frameworks, common metrics, and real-world studies are essential to assess accuracy, safety, and cost-effectiveness in diverse clinical and operational contexts. Building comparable evidence across countries and organizations will strengthen collective learning, accountability, and confidence in the use of LLM-enabled CDSSs.

4. Governance, Trust & Transparency

Trust in LLMs and broader AI systems depends on strong governance and transparency. Defining ethical standards, regulatory pathways, and accountability mechanisms that balance innovation with protection will help ensure safe, equitable deployment with oversight that reflects public interest and local values.

5. Capacity & Local Ownership

Sustainable use of LLM-enabled CDSSs requires leadership and expertise within countries. Strengthening institutional capacity, research partnerships, and digital health governance empowers ministries, universities, and implementers to adapt and regulate tools in alignment with national priorities. Investing in people and systems today will determine who shapes the next generation of responsible digital innovation.

6. Infrastructure & Enabling Environments

Reliable infrastructure and supportive ecosystems are essential for equitable LLM deployment. Investments in connectivity, data systems, and sustainable computing, supported by open standards and cross-sector collaboration, will enable countries to build resilient, future-ready health systems.

Each theme represents a collective effort to translate diverse experiences into shared understanding. Through case narratives, desk research, and collaborative analysis, the learning agenda highlights where the global community is generating evidence and where new inquiry is needed to guide responsible and equitable scale-up. The following sections are organized to make this connection clear.

Each thematic section begins with a rationale that explains why the topic is important for the safe and effective use of CDSSs. This is followed by a synthesis of current evidence, which integrates published research with practitioner insights and field experience. Implementation Spotlights, drawn directly from member contributions, illustrate how these issues play out in real-world contexts, offering grounded examples of both innovation and constraint. Finally, each section concludes with a set of learning questions and proposed activities, outlining how the CDSS community can work collectively to fill critical gaps in evidence, strengthen implementation capacity, and inform policy and investment decisions moving forward.

Together, these themes form the backbone of this Learning Agenda. They provide a roadmap not only for advancing technical understanding but also for building the institutional, ethical, and infrastructural foundations required to ensure that AI-enabled clinical decision support contributes meaningfully to stronger and more equitable health systems.

1. Localization & Language Equity

Rationale

Localization and language equity sit at the core of equitable digital transformation in health. CDSSs powered by LLMs depend on linguistic precision and cultural relevance to function safely. Yet, most existing LLMs are trained primarily on high-resource languages, limiting their effectiveness and accuracy in LMIC contexts.

In multilingual health systems where community health workers and clinicians routinely work across multiple languages, dialects, and literacy levels, failures in localization can result in miscommunication, mistrust, or even clinical harm. As demonstrated by early pilots in India, errors in translation or speech recognition can change the meaning of clinical symptoms or advice, directly affecting patient outcomes. Ensuring language equity is therefore not a peripheral concern, but a clinical imperative for patient safety and trust.

Localization also extends beyond translation. It encompasses data governance, cultural adaptation, and community participation in data collection, annotation, and evaluation. Embedding these principles into CDSS design is essential to ensure that LLMs are not only technically sound but also socially and linguistically inclusive.

Evidence

Language inequity remains one of the most significant and under-addressed barriers to the safe and effective use of LLM-enabled CDSSs in LMICs. The challenge is particularly acute in India, where more than 780 languages and 19,500 dialects are spoken, yet less than 1 percent of digitized content is available in these local languages [1]. Similar patterns exist across South Asia and Sub-Saharan Africa, where LLMs trained primarily on English datasets perform poorly in real-world clinical contexts. Benchmarking studies show that state-of-the-art models such as GPT-4o perform 12–20 percentage points lower on medical reasoning tasks in African and South Asian languages compared to English [2].

Technical challenges compound these disparities. Automatic speech recognition (ASR) and text-to-speech (TTS) systems often misinterpret colloquial or mixed-language inputs, for example, "Hinglish," "Tanglish," or hybrid vernacular phrases, especially in noisy clinical environments. Many frontline health workers use "vernacularized" medical terms that blend English and local language (for instance, *blood pressure low hai*), which LLMs trained on standardized corpora cannot reliably interpret. These mismatches can distort meaning in high-stakes interactions such as antenatal counseling, medication adherence, or emergency triage.

Emerging strategies show promise but remain under-tested. Selective pre-translation, where only key segments of prompts are translated before inference, has improved accuracy and cost-efficiency in low-resource settings [3]. In India, fine-tuning multilingual transformer models on region-specific medical corpora in Hindi, Telugu, and Tamil has improved domain accuracy by up to 15 percent in small-scale pilots [4]. However, sustainability depends on developing open, validated, and ethically governed datasets.

Community-led initiatives such as Masakhane, Karya, and Mozilla Common Voice are expanding language coverage, though medical terminology remains limited [5]. CoP discussions emphasized that progress on language equity will require coordinated investment in localized dataset creation, interoperable annotation standards, and mechanisms that protect data sovereignty while enabling responsible data sharing.

Research Questions

- How do selective pre-translation, adaptive tokenization, and low-resource fine-tuning strategies influence the accuracy, safety, and usability of LLM-enabled CDSSs across diverse languages and dialects?
- What models of community-driven dataset creation and validation best balance open data sharing, data quality, and linguistic data sovereignty in low- and middle-income settings?
- How can national digital health strategies and procurement frameworks incorporate localization standards and multilingual performance requirements for CDSS tools?
- What types of infrastructure investments (data, compute, connectivity) are most critical for enabling language equity in LMIC CDSS adoption?

Learning Activities for the Global Community

- Conduct implementation research comparing selective pre-translation, direct inference, and low-resource fine-tuning across multiple underrepresented languages.
- Establish multicountry partnerships for the creation of open, annotated, and validated clinical language corpora, leveraging community annotation groups and medical professionals.
- Develop and test standardized methods for evaluating linguistic accuracy, cultural appropriateness, and clinical safety in multilingual CDSS tools.
- Convene governments, funders, research institutions, and AI developers to co-develop shared annotation standards, data-sharing protocols, and ethical safeguards for multilingual health data.

IMPLEMENTATION SPOTLIGHT

Localization and Clinical Validation of SMARThealth GPT in India

Contributed by The George Institute for Global Health, University of Oxford and partners

The SMARThealth GPT initiative integrates a LLM chatbot into the SMARThealth Pregnancy platform, aiming to support more than one million Accredited Social Health Activists (ASHAs) who provide care to approximately 25 million pregnant women across rural India. The chatbot is designed to deliver real-time, guideline-based, and gender-equitable responses in Hindi and Telugu to assist with maternal health questions ASHAs receive during their field work. The chatbot is trained on guidelines related to maternal health conditions such as anemia, hypertension, and gestational diabetes [5].

The team started by conducting a structured clinical validation of SMARThealth GPT. Using a multi-criteria Likert-based framework, clinician reviewers in India assessed LLM responses for accuracy, completeness, appropriateness, and bias in English [7]. This process revealed that while model performance remained high in accuracy, many responses required additional information to align with regional knowledge and practices. Further user testing in Rohtak and Jhajjar districts in Haryana and Siddipet district in Telangana underscored the importance of linguistic and contextual adaptation. During the user testing phase, ASHAs were observed while they used the chatbot in a simulated field setting. The community health workers struggled while using the chatbot because local dialects or phonetic variations were misunderstood by the speech-to-text (STT) system. For example, a Telugu-speaking health worker asked about vomiting during pregnancy using the term *vanthulu*, which the STT system misheard as *panthulu*, producing an unrelated response about vaginal discharge. The incident caused confusion and disrupted the conversation flow. [6].

The project demonstrated that linguistic validation should not be separated from clinical evaluation, but is an essential dimension of it. As a future direction, the team is considering iterative fine-tuning of multilingual transformer models on local language corpora, creation of clinical dialogue datasets verified by local experts, and incorporation of local language term dictionaries. The team also advocated for the establishment of a Task Force on Language Equity in Digital Health to standardize annotation protocols, promote interoperability, and mobilize investment in open, ethically sourced language resources [10]. Beyond technical improvement, the experience emphasized that trust in AI tools among community health workers depends on linguistic familiarity, cultural sensitivity, and clarity.

The SMARThealth GPT initiative demonstrated that localization and clinical validation are mutually reinforcing. Through iterative testing and refinement, the chatbot achieved measurable improvements in response accuracy and user comprehension. The case provides compelling evidence that embedding linguistic validation within model design and evaluation processes is vital to ensuring safety, equity, and real-world usability of LLM-enabled clinical decision support systems in multilingual health systems.

Photo credit: The George Institute for Global Health

2. Voice-Enabled & Multimodal Tools

Rationale

In many LMIC health systems, limited digital literacy, scarce devices, and unreliable internet connectivity continue to constrain the reach of text-based CDSSs. Community and frontline health workers frequently operate in environments where smartphones, stable networks, and typing proficiency cannot be assumed. In such contexts, voice-enabled and multimodal interfaces represent a critical innovation for equity. By allowing users to communicate naturally through speech, audio, and images, these tools can bridge literacy gaps, expand access in local languages, and extend digital decision support to settings that have historically been excluded from the benefits of digital health transformation.

Voice and multimodal tools also align with human-centered design principles by meeting users where they are, leveraging familiar interaction modes to reduce cognitive load and improve adoption. For health workers operating in multilingual or low-connectivity environments, these systems have the potential to improve efficiency and clinical accuracy while enhancing inclusivity for populations with limited literacy.

However, these approaches introduce new technical, operational, and ethical challenges. ASR systems often underperform in African and South Asian languages, particularly across diverse accents and dialects. Bandwidth and processing requirements for multimodal tools can exceed what is available in most public health settings, necessitating offline or hybrid deployments. In addition, privacy, data security, and consent considerations remain underdeveloped for voice data, where inadvertent capture of sensitive information is a growing risk.

Understanding whether these tools can be implemented safely, effectively, and sustainably in resource-constrained contexts is therefore essential. Evidence is still limited on how multimodal CDSSs perform in real-world health systems, how they influence clinical workflows and decision quality, and what infrastructure, policy, and training investments are required to scale them equitably.

Evidence

Emerging research demonstrates that voice-enabled and multimodal CDSSs can expand access to digital health tools while improving efficiency and usability in LMIC contexts. These solutions lower entry barriers for frontline workers by reducing reliance on text-based inputs, enabling natural speech interactions, and allowing decision support through basic devices.

Early implementations illustrate both the potential and the complexity of deploying these tools at scale. In Nigeria, Viamo's Community Health Extension Worker Assistant (CHEWA) allows community health extension workers to dial a toll-free number from feature phones to receive real-time, LLM-generated clinical guidance. This approach effectively bypasses barriers related to connectivity and device ownership, highlighting how voice-based tools can reach last-mile health workers. Yet, it also underscores persistent concerns about clinical reliability, latency, and the need for robust safeguards to ensure the accuracy of time-sensitive recommendations.

Other innovations are exploring multimodality to enhance both provider and patient engagement. In Rwanda, an ambient listening pilot is testing whether passive audio capture of community health worker consultations in Kinyarwanda can strengthen diagnostic reasoning and referral decisions, an early example of combining speech recognition, translation, and LLM reasoning in a low-resource language. Similarly, Dimagi's experiments in Kenya and Senegal integrate asynchronous voice interfaces and culturally resonant personas into chatbot-based behavior change interventions, showing promise for improved comprehension and sustained engagement.

A recent systematic review of AI-based speech recognition for clinical documentation synthesizes global findings across high-income settings and confirms that such systems can improve documentation efficiency and accuracy but continue to face challenges with error rates, domain adaptation, and workflow integration [14]. This underscores that many of the issues identified in LMIC deployments, such as the need for domain-specific tuning, interoperability, and strong human oversight, mirror those observed globally, reinforcing the importance of localized, context-driven evaluation.

Taken together, these initiatives suggest that voice-enabled and multimodal systems can extend the reach of CDSSs into populations previously excluded from digital innovation. However, the path to scale remains constrained by several persistent challenges. ASR systems continue to struggle with accented speech and code-switching across African and South Asian languages. Limited offline functionality and high bandwidth requirements restrict rural deployment. Ethical and governance frameworks for voice data, particularly around ownership, consent, and storage, are still underdeveloped. Most critically, longitudinal evidence on safety, usability, and patient outcomes remains scarce, underscoring the need for sustained evaluation to ensure that these technologies can scale responsibly and equitably.

Research Questions

- Do voice-enabled and multimodal CDSSs sustainably improve care quality, workflow efficiency, and access in low-literacy or low-connectivity communities?
- What design and deployment models best enable offline, hybrid, or low-bandwidth use of voice-enabled CDSS in LMIC health systems?
- How should health systems measure the human and ethical dimensions of multimodal CDSS, such as clinician well-being, trust, data ownership, and informed consent for voice data?

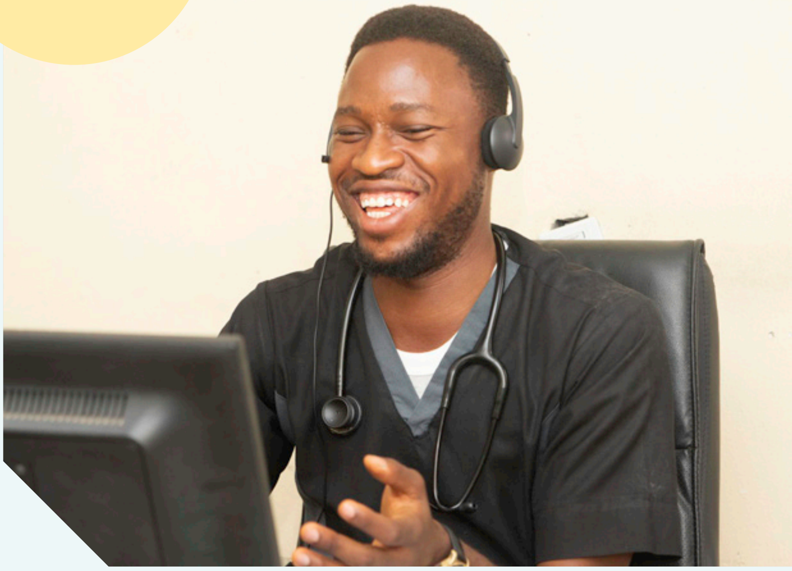
Learning Activities for the Global Community

- Conduct collaborative implementation studies and operational pilots across diverse LMIC contexts to evaluate the usability, safety, and sustainability of voice-enabled and multimodal CDSSs. Comparative evaluations should assess outcomes such as workflow efficiency, documentation quality, and access for low-literacy health workers.
- Launch coordinated, multicountry trials to test offline and hybrid deployment models under varying infrastructure conditions. Use standardized evaluation frameworks and common metrics to enable cross-site learning and evidence synthesis on performance, data privacy, and patient safety.
- Integrate participatory design and qualitative inquiry to examine how these tools influence clinician well-being, trust, and patient communication. Develop shared ethical guidelines for voice data governance, consent, and cultural adaptation, particularly for marginalized and multilingual populations.
- Convene policy dialogues and evidence-sharing workshops with ministries of health, implementers, and funders to translate emerging findings into guidance for procurement, scale-up, and governance of multimodal CDSSs.

IMPLEMENTATION SPOTLIGHT

Intron Health's Speech-to-Text Deployment in Nigeria

Contributed by Intron Health



Between 2024 and 2025, Intron Health implemented a medical STT solution across eight tertiary hospitals in Nigeria, including Aminu Kano and Barau Dikko Teaching Hospitals. The initiative aimed to evaluate whether voice-enabled documentation could improve efficiency, accuracy, and clinician well-being in environments characterized by high patient volumes and limited administrative support.

The implementation involved 34 clinicians across internal medicine, radiology, obstetrics, and emergency departments, producing over 6,400 transcribed encounters. Quantitative data showed that documentation time per patient decreased by 62 percent (from 8.3 to 3.2 minutes), while same-day report completion rose from 11 percent to 89 percent [11]. Facilities recorded a 20 percent increase in daily patient throughput, largely attributed to reduced documentation delays and shorter case backlogs. Average time to finalize diagnostic reports dropped from 18 hours to under 6 hours, and clinicians reported saving 6 to 12 hours per week on paperwork.

Intron's locally trained ASR models, built using the AfriSpeech-200 dataset, which encompasses over 500 African accents and 20 languages, achieved a word error rate of below 8 percent in controlled environments and below 15 percent in live clinical use [12,13]. Clinicians described dictation as two to five times faster than typing, with 78 percent reporting richer, more complete notes and 94 percent recommending the tool to their peers. Notably, radiologists highlighted a significant reduction in repetitive strain injuries and transcription fatigue, while hospital administrators reported a decline in patient complaints about delayed records.

Qualitative interviews further emphasized the human-centered benefits of voice documentation. Clinicians reported that the system allowed them to spend up to 30 percent more time in direct patient interaction, and several noted improved clarity in referral communication and continuity of care. In facilities with consistent usage, radiology and laboratory backlogs were cleared within two months of deployment, creating measurable operational efficiencies and contributing to improved patient satisfaction.

However, the pilot also revealed persistent challenges. Model performance fluctuated in noisy wards, and specialized medical terminology sometimes required manual correction. Offline functionality remained limited, particularly in rural hospitals without stable connectivity. Data governance and patient consent procedures required further strengthening, highlighting the need for policy-aligned ethical frameworks to ensure responsible scale-up.

Overall, Intron Health's experience demonstrates the feasibility and health system value of localized, voice-enabled CDSSs in LMIC contexts. The pilot provided some of the first quantitative evidence that medical speech interfaces can simultaneously enhance documentation quality, reduce clinician workload, and improve service delivery efficiency. At the same time, it reinforced that responsible deployment requires ongoing investment in local language data, domain-specific model tuning, and sustainable governance to balance innovation with patient safety and data integrity.

Photo credit: Intron Health

3. Evaluation & Real-World Evidence

Rationale

Evaluation is the foundation of safe, equitable, and sustainable adoption of CDSSs. Without robust evidence, health systems risk deploying tools that are ineffective, unsafe, or misaligned with the realities of clinical care in LMICs.

To date, most evaluations of LLM-enabled CDSSs have relied on text-based evaluation metrics such as BLEU and ROUGE, which measure word overlap between human reference outputs and model-generated outputs. While these metrics can capture certain linguistic similarities, they do not assess the dimensions that matter most for health systems, such as whether recommendations are clinically appropriate, safe for patients, usable for frontline health workers, or trusted by providers and communities. Responsible innovation requires evaluation frameworks that move beyond surface-level text similarity to ensure that AI tools used in health care are safe, effective, and equitable.

Evaluation is particularly urgent in LMIC contexts for three reasons.

1. **High stakes for patient safety.** Many LMIC health systems already face constrained capacity, limited diagnostic resources, and fragile referral pathways. Deploying untested AI tools could amplify risks rather than address them.
2. **Contextual variation.** A "one-size-fits-all" evaluation approach is not viable. Local health worker roles, workflows, and cultural contexts profoundly shape how CDSS tools are used and must be reflected in evaluation design.
3. **Policy and procurement needs.** Donors, governments, and regulators increasingly demand clear evidence before endorsing, funding, or scaling digital health tools. Without trusted evaluation frameworks, promising innovations risk stalling before reaching scale.

Evidence

Across the Community of Practice CoP, evaluation has emerged as a cornerstone of responsible innovation for LLM-enabled CDSS. Over the past year, the field has begun moving from simulated benchmarks to prospective, real-world evaluations within LMIC health systems. This shift reflects a growing recognition that technical metrics such as BLEU or ROUGE capture only narrow aspects of model performance, whereas health systems require evidence on safety, appropriateness, usability, equity, and trust [16].

The broader research and policy environment is now reinforcing the need to move beyond algorithmic accuracy toward clinical validity, communication quality, and contextual relevance [16]. Systematic reviews confirm that randomized or prospective AI trials in LMICs remain exceedingly rare, and that evaluation frameworks often lack standardized reporting [15, 17, 18]. In parallel, regulatory guidance from World Health Organization (WHO) *Ethics and Governance of AI for Health* and the joint *Good Machine Learning Practice (GMLP)* framework developed by the U.S. Food and Drug Administration (FDA), Health Canada, and the United Kingdom's Medicines and Healthcare products Regulatory Agency (MHRA) calls for lifecycle oversight, encompassing transparency, generalizability, data quality, and post-market monitoring [19, 20].

Together, these trends signal a decisive turn toward evidence-driven accountability. The CoP's ongoing studies, spanning evaluation frameworks, randomized and observational trials, and implementation research, are helping to operationalize these global expectations by generating the first prospective, context-grounded data on the safety, effectiveness, and usability of LLM-enabled CDSSs in LMIC settings.

Research Questions

- How can we establish a standardized yet adaptable framework to evaluate the safety, effectiveness, and usability of LLM-enabled CDSSs across diverse health system contexts?
- How can real-world evaluation methods better capture contextual performance and identify when models become misaligned with local clinical needs?
- How can cost-effectiveness evidence be generated and used to inform investment, procurement, and scale-up decisions for AI-enabled decision support tools?
- What participatory evaluation approaches can strengthen trust, accountability, and local capacity to generate and use evidence on AI in health care?

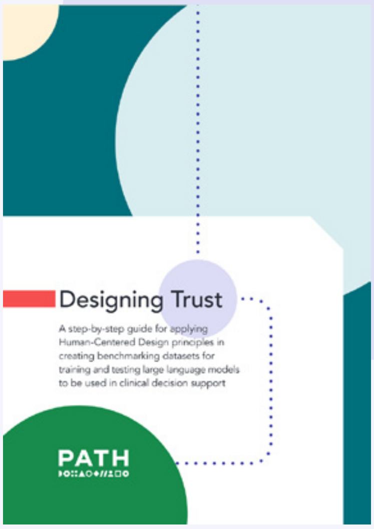
Learning Activities for the Global Community

- Develop a global CDSS evaluation framework and metrics library that defines shared indicators for clinical validity, safety, usability, and equity, and is adaptable to local implementation contexts.
- Pilot multicountry evaluation clinics and shared vignette libraries to test standardized methods across regions, using locally developed case vignettes and scoring rubrics to assess contextual performance, safety, and communication quality.
- Generate and synthesize economic evidence through comparative cost-effectiveness studies and modeling tools that quantify efficiency gains, cost per decision, and return on investment, producing practical guidance for implementers, donors, and policymakers.
- Strengthen evaluation capacity and evidence translation by developing training programs for implementers and regulators, and hosting evidence-sharing forums that connect evaluation results to product improvement, procurement, and policy adaptation.

IMPLEMENTATION SPOTLIGHT

University of Birmingham: Evaluation Clinics and the Playbook Framework

Contributed by the University of Birmingham and PATH



Designing Trust

A step-by-step guide for applying Human-Centered Design principles in creating benchmarking datasets for training and testing large language models to be used in clinical decision support

PATH

The University of Birmingham (UoB), in collaboration with PATH and CoP members, led a multi-country effort to develop and test a shared methodology for evaluating LLM-enabled clinical decision support systems (CDSS). Building on UoB's long-standing work on scientific and regulatory standards for digital health, the team created a structured Evaluation Playbook that offers a four-domain assessment framework: clinical validity, communication quality, equity and inclusivity, and usability and workflow integration.

To operationalize the framework, UoB hosted a series of Evaluation Clinics between late 2024 and mid-2025 with implementers working in India, Kenya, Nigeria, Rwanda, South Africa, and Brazil. Each clinic combined a brief pre-meeting to understand project needs with a structured assessment using locally developed clinical vignettes. These vignettes proved essential, grounding evaluations in the realities of frontline care and allowing reviewers to examine not only factual accuracy but also tone, contextual appropriateness, and adherence to national guidelines.

Across the clinics, participants observed a recurring pattern: most tools performed well on core clinical content, yet struggled with communication nuances such as empathy, phrasing, and clear guidance on next steps. Maternal health and antimicrobial stewardship teams in particular noted gaps where LLMs failed to fully capture local danger signs or integrate context-specific reasoning. These insights underscored the value of multidisciplinary reviewer panels, bringing together clinicians, data scientists, and implementers, and reinforced the need for continued refinement of the evaluation process.

The clinics also helped strengthen UoB's LLM Evaluation Rubric, which many teams incorporated into their internal QA processes. The rubric's structured scoring across reasoning quality, safety, communication, and potential harm gave implementers a practical pathway for improving model outputs and informing product roadmaps.

Across countries, the Evaluation Clinics demonstrated that a participatory, multi-country process is both feasible and essential for responsible AI deployment. The Playbook and associated tools are now being used by CoP members to harmonize evaluation approaches, support procurement and governance decisions, and embed quality assurance into scale-up planning. By anchoring evaluation in local context and shared standards, this work is helping define practical, trustworthy methods for assessing LLM-enabled CDSS in low- and middle-income settings.

IMPLEMENTATION SPOTLIGHT

Penda Health (Kenya): Real-World Validation and Cost-Effectiveness of LLM-Assisted Triage

Contributed by Penda Health, Kenya

Penda Health conducted one of the first real-world clinical trials of an LLM-assisted triage system in its network of primary-care clinics across Nairobi and Kiambu Counties. The trial evaluated the impact of the tool on clinical accuracy, workflow efficiency, and cost-effectiveness compared to standard clinician practice.

Across 1,469 outpatient encounters, AI-assisted triage recommendations were compared against clinician decisions and national outpatient guidelines. Quantitative analysis demonstrated 89 percent concordance for routine cases and 76 percent for complex presentations. Discrepancies were largely due to incomplete clinical context in data entry. Cost modeling estimated a 15 percent reduction in per-visit triage time and a projected 8 percent decrease in operational costs per consultation, primarily through shorter encounters and improved documentation efficiency.

Clinician interviews revealed a balance of enthusiasm and caution: while many reported improved efficiency and confidence in standardized documentation, others highlighted concerns around automation bias and overreliance on AI-generated outputs. These findings prompted the inclusion of supervisory review mechanisms and iterative user training to strengthen accountability and reinforce clinician oversight.

The project demonstrated that LMICs can be early and responsible adopters of AI in health care, leveraging high patient volumes and urgent decision-support needs to pilot safe, scalable innovation. The collaboration produced two manuscripts, one quantitative and one focused on user experience, submitted to preprint servers to inform future policy and implementation decisions. Clinician feedback underscored that success depends on pairing technical innovation with trust-building, transparency, and careful rollout strategies.

From a cost perspective, the LLM intervention proved highly efficient. Each model call costs less than USD 0.01 in this retrospective analysis, demonstrating the potential for low-cost, high-impact AI tools that enhance decision-making in resource-constrained settings. The economic modeling reinforced that with continued investments in safety, usability, and clinician engagement, LLM-based CDSSs can contribute meaningfully to cost-effective health system strengthening in LMICs.

Penda Health's evaluation provides early, quantitative evidence that AI-assisted clinical decision tools can improve efficiency and lower costs when deployed responsibly. The study highlights that the sustainable scale-up of AI in health care systems depends not only on clinical effectiveness, but also on demonstrable economic value and trust among end users.

Photo: credit: PATH

IMPLEMENTATION SPOTLIGHT

Intelehealth (India): Controlled Evaluation of an LLM use case for Telemedicine

Contributed by Intelehealth, India

Intelehealth is conducting a matched case-control study on integrating an AI-powered diagnostic assistant within the Arogya Sampada telemedicine program serving rural and tribal populations in India. The study investigates whether large language models can improve diagnostic accuracy, consultation efficiency, and treatment quality in low-resource telemedicine contexts where clinicians face high patient loads, incomplete health histories, and limited diagnostics.

Using a matched-pairs within-case randomized design, 20 clinicians were organized into 10 pairs matched on experience, qualification, and setting. Each clinician alternated between AI-assisted and unassisted cases, enabling within-case comparison. In the assisted arm, the model produced ranked differentials and treatment suggestions. Ground-truth diagnoses were independently verified and used to evaluate the influence of AI on diagnostic accuracy and the quality of treatment plan.

Key evaluation parameters included diagnostic accuracy and time to diagnosis. The quality of the LLM-generated differential diagnoses was assessed on appropriateness, comprehensiveness, and top-5 ranking performance, and treatment plans were evaluated for completeness and appropriateness. Investigations, medical advice, and referrals were rated for clinical relevance on validated 5-point scales, drawing on prior methods for differential diagnosis tools and clinical reasoning evaluation [22, 23]. Treatment quality incorporated the Medication Appropriateness Index to assess prescribing appropriateness [21].

With AI assistance, diagnostic accuracy improved by 12 percent, and mean consultation time decreased by 47% (~2 minutes 25 seconds). The correct diagnosis appeared within the top three suggestions in 90.5 percent of cases, and expert reviewers rated 90.8 percent of AI-generated differential lists as very appropriate. In terms of comprehensiveness, the true or closely related diagnosis appeared in 92 percent of cases. Treatment plans in the AI-assisted arm were more complete, and Medication Appropriateness Index scores were lower, indicating more appropriate medication choices.

However, the non-AI arm performed better on recommending tests, medical advice, and referrals. Qualitative and quantitative evidence suggested that some model outputs reflected ideal clinical standards rather than feasible options in Indian rural settings. The authors note the need to retrain the model with locally relevant data and to encode awareness of referral networks, infrastructure constraints, and context-specific care pathways. Clinicians viewed the AI as a safe, efficient "cognitive co-pilot" that enhanced reasoning and streamlined consultations, reporting no clinically inappropriate or unsafe outputs; however, they noted its limited contextual awareness and cited integration, workflow, and training gaps as barriers to wider adoption. Lower overall accuracy across both arms likely reflects the controlled vignette design, which removed opportunities for patient interaction and contextual cues that typically inform teleconsultations, as well as limited familiarity with local disease patterns and health-system norms.

Intelehealth's study provides early, practice-relevant evidence that AI assistance can enhance diagnostic accuracy, efficiency, and prescribing quality in LMIC telemedicine while also revealing context gaps that require localization and system-aware tuning. The mixed-methods approach shows how rigorous evaluation can surface both gains and limitations, informing responsible adaptation of LLM-enabled decision support in low-resource settings.

4. Governance, Trust & Transparency

Rationale

Trust and governance are essential for the safe and sustainable adoption of LLM-enabled CDSSs. While model performance is advancing rapidly, the absence of robust oversight frameworks creates risks that unsafe or biased systems may be deployed prematurely. Traditional digital health policies often do not extend to AI-specific risks such as hallucinations, model drift, or opaque training data.

Regulatory authorities in high-income countries have begun to fill this gap. The European Union (EU) AI Act classifies health-related AI as high-risk, requiring developers to document training data, ensure human oversight, and maintain post-market monitoring systems. In 2021, the FDA, MHRA, and Health Canada jointly released the GMLP principles, which outline expectations for data quality, transparency, generalizability, and lifecycle monitoring of AI-enabled medical technologies. Subsequent guidance released in 2023 introduced the concept of predetermined change control plans (PCCPs), providing a regulatory pathway for managing updates to adaptive AI systems. The World Health Organization has likewise emphasized lifecycle evaluation, explainability, and accountability in its 2023 guidance on AI ethics and governance in health.

In LMICs, regulatory frameworks are often limited or absent. This creates a dual challenge: health systems risk either falling behind global adoption curves or adopting tools without adequate safeguards. Without locally adapted regulatory models, ministries of health cannot easily evaluate whether LLM-enabled CDSSs meet acceptable thresholds for safety, performance, and equity.

Evidence

In Africa, the South African Health Products Regulatory Authority (SAHPRA) released its first guidance on AI/Machine Learning (ML)-enabled medical devices in 2025. The document requires manufacturers to classify AI tools within South Africa's risk-based medical device categories and to provide technical files detailing algorithms, validation results, and cybersecurity protections. Clinical validation must demonstrate safety and benefit in local populations, including evidence on generalizability where models were trained on external datasets. Developers are also required to address risk management, transparency, and explainability, as well as compliance with South Africa's data protection law (POPIA). SAHPRA further mandates post-market surveillance to monitor for drift and unsafe outputs. It identified particular challenges with generative AI-enabled medical devices, citing the unpredictability of outputs, broad intended use, and the need for predefined change control plans for adaptive algorithms.

Evidence from the CoP illustrates how governance issues manifest in practice. In Rwanda, pilots using voice-enabled LLMs raised questions about accountability when AI-generated recommendations conflicted with national referral protocols. In Nigeria, the CHEWA hotline prompted debates over data sovereignty as clinical interactions were stored on cloud-based systems managed by external partners. CoP members also observed that while global regulators and donors are advancing AI pilots, few governance frameworks are being shared or adapted locally, leaving ministries of health and implementers without clear models to guide responsible adoption.

Research Questions

- What regulatory models are most effective for classifying and evaluating LLM-enabled CDSS in LMIC health systems, and how can global frameworks like GMLP or the EU AI Act be adapted to local contexts?
- How can accountability be assigned when AI-generated recommendations conflict with national clinical guidelines or lead to patient harm?
- What standards and methods can ensure explainability and transparency of LLM outputs so that frontline health workers and patients can trust and safely act on recommendations?
- What governance mechanisms are needed to safeguard data sovereignty in contexts where AI tools depend on cloud-based storage and external infrastructure?
- How can post-market surveillance systems be designed to detect model drift, unsafe outputs, or bias in real-world LMIC deployments?

Learning Activities

- Support pilot projects with embedded regulatory partnerships, where LMIC regulators co-design evaluation and approval processes for LLM-enabled CDSSs.
- Convene cross-country working groups to draft minimum transparency and explainability standards for AI-enabled CDSSs, tailored to LMIC needs.
- Fund and test sovereign data governance models, including locally hosted data storage or hybrid approaches that balance accessibility with sovereignty.
- Design and test post-market monitoring frameworks, including clinician reporting systems and automated audit tools, that LMICs can adopt for ongoing oversight.

5. Capacity & Local Ownership

Rationale

The integration of AI technologies into clinical practice stands as both a defining challenge and a transformative opportunity for global health in the coming decade. While applications such as clinical documentation and diagnostic support are rapidly becoming normalized in high-resource settings, with over half of clinicians in the United States now reporting some use of AI, much less is known about the perceptions, barriers, and real-world needs of frontline clinicians in LMICs.

The sustainability and equity of LLM-enabled CDSSs depend not only on technical accuracy but also on the capacity of local actors to design, adapt, and govern these tools. Without meaningful local ownership, AI risks reinforcing historical imbalances in global health, where externally developed innovations are introduced without sufficient investment in the local skills, institutions, and infrastructure required for long-term success.

Building capacity for responsible AI adoption extends far beyond end-user training. It includes cultivating regulatory and evaluation expertise within ministries of health, ensuring that local universities, developers, and research institutions play central roles in model design and dataset curation, and empowering health workers to integrate AI tools safely and confidently into their workflows. Institutional mechanisms, such as national AI hubs, ethics review boards, and interoperable data systems, are equally critical to enable sustained local stewardship and accountability.

True local ownership also requires alignment with national digital health strategies and integration into existing systems, not parallel, donor-driven pilots. For AI to meaningfully strengthen health systems, LMICs must be positioned not merely as adopters but as co-creators, leading the development, testing, and governance of CDSSs within their contexts. Strengthening this local capacity is thus both a technical necessity and an equity imperative, ensuring that the global expansion of AI in health advances sustainability, sovereignty, and shared benefit.

Evidence

Emerging evidence underscores that the integration of AI into clinical workflows is outpacing the readiness of many health systems to manage it safely and sustainably. Across LMICs, digital infrastructure, workforce preparedness, and institutional mechanisms for governance remain uneven, constraining both adoption and scale.

Global research points to the magnitude of this readiness gap. Large-scale surveys of clinicians in Asia and Africa show that while interest in AI tools is growing rapidly, many health workers lack the infrastructure and institutional support required for safe use. Fewer than half of respondents in one multicountry study reported regular access to a computer, and over 90 percent relied primarily on personal smartphones for clinical information. Less than one-fifth used electronic health records, and nearly half continued to depend on paper-based documentation, underscoring that even the most advanced CDSS tools will remain inaccessible without offline-first design and tailored capacity development.

At the systems level, the evidence base reveals persistent gaps in training, governance, and local participation. Clinicians often lack exposure to AI tools or mechanisms to evaluate their reliability, while ministries of health face limited capacity to regulate, procure, or evaluate AI technologies for safety, cost-effectiveness, and equity. Although 80 percent of clinicians in recent surveys express interest in learning about AI applications, concerns about accuracy, patient trust, and professional autonomy persist.

Field experience reflects similar trends. In Nigeria, the CHEWA hotline pilot showed that community health extension workers could use a voice-enabled LLM effectively only after structured training and continuous supervision. In Rwanda, limited participation by local universities in dataset development highlighted the need for stronger academic–policy collaboration to sustain innovation beyond donor cycles. Meanwhile, initiatives such as Masakhane and Karya demonstrate how locally led AI and data efforts can strengthen regional technical expertise while embedding equity principles in data generation.

Complementary evidence from ongoing implementation research further demonstrates that local leadership and institutional readiness are critical for real-world evaluation. Recent multicountry pilots of LLM-enabled CDSSs, including work in Kenya and Rwanda, emphasize that rigorous evaluation depends not only on technical design but also on the availability of trained clinical reviewers, local governance mechanisms, and ministry oversight. These experiences demonstrate that the success of AI trials and implementations ultimately depends on sustained local capacity to operationalize, interpret, and act on the findings.

Overall, the evidence reinforces that strengthening local ownership must be a central pillar of responsible AI for health. Building not only technical but also institutional capacity, encompassing regulation, data governance, and long-term stewardship, will determine whether LLM-enabled CDSSs contribute to sustainable, equitable health system transformation.

Research Questions

- What models of capacity strengthening are most effective for enabling LMIC ministries of health, regulators, and local research institutions to evaluate, procure, and govern LLM-enabled CDSSs safely and equitably?
- What training interventions best foster trust, proficiency, and sustained use of AI-enabled CDSSs among clinicians and frontline health workers, particularly in low-connectivity or low-digital-literacy contexts?
- How can donor and implementer partnerships be structured to ensure local ownership of data, tools, and research agendas, rather than externally driven priorities?
- How do patient and community perspectives on AI-mediated care shape trust, communication, and the quality of clinician–patient relationships, and how can these insights guide the responsible integration of digital tools into routine practice?

Learning Activities

- Conduct multicountry studies assessing national and subnational capacity to evaluate, procure, and govern AI tools. Develop and validate AI readiness and governance frameworks co-designed with ministries of health, regulatory bodies, and local research institutions.
- Implement and evaluate training and mentorship programs that build clinician trust and proficiency with AI-enabled CDSS. Use mixed methods approaches to measure impacts on confidence, adoption, and decision quality across different cadres and digital literacy levels.
- Study and document models of donor–implementer collaboration that promote local ownership of data, tools, and research agendas. Develop guidance for equitable authorship, funding structures, and data stewardship practices that center LMIC institutions.
- Conduct qualitative and participatory research exploring patient and community perspectives on AI-mediated care. Use findings to co-design communication strategies and training materials that safeguard empathy, trust, and human connection in increasingly digital clinical environments.

IMPLEMENTATION SPOTLIGHT

Building Capacity for Responsible AI Adoption

Contributed by the Global Health Delivery Project at Harvard

The integration of AI technologies into clinical practice represents a critical inflection point for global health in the coming decade. Evidence suggests that frontline providers in LMICs are not lacking enthusiasm, but rather access to deployment frameworks and institutional support structures.

To examine this dynamic, the Global Health Delivery Project at Harvard (GHD) conducted a global survey to explore how clinicians in its Better Evidence Network perceive, access, and apply AI tools in their work. The survey engaged over 800 respondents from 90 countries, with 70 percent based in Africa and Asia, regions where demand for digital health solutions is intensifying rapidly, despite inconsistent deployment infrastructure.

The data reveal a clear pattern: infrastructure constraints exist, but they do not reflect provider reluctance or lack of eagerness. Fewer than half of clinicians reported having access to a workplace computer, and only one in four had a dedicated device for clinical use. Yet over 90 percent already use smartphones for professional tasks, demonstrating adaptive capacity and digital fluency. While fewer than 20 percent reported using electronic health records and nearly half still rely on paper-based documentation, these findings underscore not a resistance to technology but rather the absence of systematically deployed digital infrastructure.

Most striking are the findings, which challenge assumptions about LMIC provider eagerness. Eighty percent of respondents expressed strong interest in learning how to integrate AI tools into care delivery, a level of enthusiasm that exceeds adoption rates in many high-resource settings. While only one-third currently feel confident deploying these tools, the gap is attributable to a lack of training and institutional guidance rather than skepticism about AI's clinical value. Commonly cited barriers include inadequate training, uncertainty about data accuracy, limited institutional support, and concerns about patient trust. These barriers are fundamentally operational. Qualitative responses consistently emphasized a desire for concrete deployment guidance, contextually tailored tools, and trust-building protocols, indicating that providers are actively seeking implementation frameworks.

The analysis also reveals a significant knowledge gap at the institutional level. Many ministries of health and regulatory bodies lack the technical capacity to evaluate AI tools for safety, efficacy, equity, and cost-effectiveness. Consequently, procurement and evaluation decisions are frequently delegated to external partners, creating dependency relationships that may not align with local health priorities. This capacity deficit represents a critical bottleneck: frontline providers are ready and motivated, but lack the institutional architecture necessary to translate interest into systematic deployment.

GHD's findings indicate that the sustainable integration of AI-enabled clinical decision support depends less on changing provider attitudes, which are already favorable, and more on establishing the institutional mechanisms, technical training pathways, and governance frameworks that providers are actively seeking. The evidence suggests that frontline clinicians in LMICs represent an underutilized asset: a digitally fluent, highly motivated workforce ready to adopt AI tools once clear deployment pathways and institutional support systems are established.

Photo: The Global Health Delivery Project

6. Infrastructure & Enabling Environments

Rationale

The effectiveness of LLM-enabled CDSSs depends on far more than the algorithms themselves. Reliable internet connectivity, access to computing resources, device availability, and secure data systems determine whether AI tools can move beyond pilot settings into routine clinical practice. In many LMICs, health workers operate in environments where connectivity is inconsistent, devices are limited, and health information systems are fragmented, conditions that constrain the real-world impact of digital tools designed for decision support.

An enabling environment also extends beyond physical infrastructure to include policies, financing mechanisms, and digital health strategies that sustain long-term integration. Even when pilot projects demonstrate feasibility, their durability depends on whether they align with national priorities, leverage existing data systems, and are supported by ongoing investment and governance. By focusing on both infrastructure and enabling ecosystems, stakeholders can identify the practical and policy conditions that enable AI innovations to transition from isolated pilots to scalable, sustainable implementations within diverse health system contexts.

Evidence

Infrastructure and enabling environments consistently emerged as priorities across community consultations and case studies. Participants emphasized that connectivity, device access, and integration into existing health systems remain among the most significant barriers to adoption. Without stable infrastructure, even well-designed CDSSs cannot be deployed effectively or evaluated in real-world settings, limiting their ability to improve decision-making at the point of care.

Examples from the ecosystem illustrate both innovation and constraint. In Kenya, the *Clinical Reasoning Challenge*, organized by Zindi, simulated the realities faced by nurses in rural hospitals, making clinical decisions with limited connectivity, scarce resources, and minimal specialist support. Using more than 400 authentic vignettes, the challenge demonstrated the potential of AI-enabled reasoning while underscoring the infrastructural limitations that impede real-world deployment in such settings.

In Nigeria, the Viamo CHEWA offered an alternative model by integrating GPT-4 into a toll-free, voice-based platform accessible on basic feature phones. This design aligned with Nigeria's national clinical guidelines and included safeguards such as human review and monitoring dashboards. The pilot showed that CDSSs could reach community health workers in low-connectivity environments and that iterative improvements, including context memory, simplified navigation, and feedback loops, enhanced usability and safety. Yet, the system's reliance on external cloud infrastructure and donor-supported hosting highlighted ongoing challenges of local sustainability, affordability, and integration into national health systems.

Beyond connectivity and devices, computational infrastructure and energy sustainability are increasingly critical considerations. LLMs consume vast amounts of energy, both visible and hidden, during training and inference. Emerging strategies seek to improve efficiency by developing small language models (SLMs) or combining LLMs with prompt engineering and TinyML for device-level use. These approaches raise a central question: What are we trying to solve using LLMs or SLMs, and at what computational cost?

Most LLMs used in health care applications rely on retrieval-augmented generation, where the model retrieves information from external resources to generate contextually relevant outputs. However, when an LLM functions as a high-throughput retrieval engine, processing millions of queries per second, its energy demands can far exceed those of traditional database-driven systems, without guaranteeing higher accuracy. While commercial providers have introduced token-based pricing mechanisms to manage usage costs, open-source models such as LLaMA offer alternatives that can be deployed offline and at lower energy expense.

A second structural bottleneck lies in computational acceleration. Architectures like CUDA have positioned a handful of companies as global leaders in training and inference for large transformer-based models, raising questions about dependency, affordability, and access to high-performance computing in LMICs. In the health sector, data needs extend far beyond natural language processing alone, encompassing imaging, biomedical informatics, and epidemiological modeling. Consequently, the next generation of AI-integrated health systems will depend not on universal reliance on LLMs but on identifying where LLMs add unique value and combining them with lower-computing-cost models and statistical frameworks. This hybrid approach is essential to building sustainable computing infrastructure and enabling the future of energy-efficient, AI-integrated health care systems.

Taken together, these experiences and insights highlight that the long-term impact of AI-enabled CDSSs depends as much on the infrastructure and energy ecosystems that support them as on model design. Equitable deployment in LMICs will require investments in digital public infrastructure, localized hosting capacity, and sustainable computing strategies, supported by enabling policies that ensure AI systems remain affordable, resilient, and environmentally responsible.

Research Questions

- What are the most cost-effective and sustainable approaches to providing computing resources for CDSSs in low-resource environments, including the role of shared national or regional infrastructure and low-power or hybrid architectures (e.g., SLMs, TinyML, edge computing)?
- How do different connectivity models, including offline, low-bandwidth, and mobile-first designs, influence the usability, safety, and equity of AI-enabled CDSSs across diverse LMIC health systems?
- What enabling policies, procurement models, and financing mechanisms best support the long-term sustainability and scale-up of CDSSs beyond donor-funded pilots, ensuring national ownership and interoperability with existing digital health systems?
- How can health systems balance performance, energy efficiency, and data sovereignty when determining the optimal mix between LLM-based and lower-computing-cost models for clinical decision support?
- How can longitudinal analyses of AI adoption track the evolution of infrastructure, digital health systems, and institutional readiness over time, and what indicators can help measure progress toward sustainable, system-wide integration?

Learning Activities

- Comparative cost-effectiveness studies assessing the performance, energy efficiency, and operational costs of different CDSS computing models (e.g., LLMs, SLMs, TinyML, and hybrid edge–cloud systems) across varying resource environments.
- Connectivity and deployment pilots testing offline and low-bandwidth CDSS prototypes in rural and peri-urban health facilities to evaluate usability, safety, and continuity of care under constrained infrastructure conditions.
- Policy and financing dialogues bringing together ministries of health, donors, and regulators to document and share enabling policies, procurement strategies, and blended financing models that promote sustainable, country-owned scale-up of CDSSs.
- Technical workshops and infrastructure audits to co-develop national or regional blueprints for low-energy, data-sovereign AI architectures, focusing on optimizing power consumption, data localization, and cybersecurity standards.
- Longitudinal implementation research tracking AI adoption and infrastructure readiness over time, integrating indicators from national digital health strategies and WHO benchmarks to assess systemic progress toward sustainability and integration.

Charting the Path Forward

The six themes outlined in this Learning Agenda reflect the collective insights, evidence, and priorities identified through the development of the CDSS community's first shared framework for learning and collaboration. Together, they highlight not only where progress has been made but also where critical knowledge gaps remain in realizing the safe, equitable, and sustainable use of LLM–enabled CDSSs.

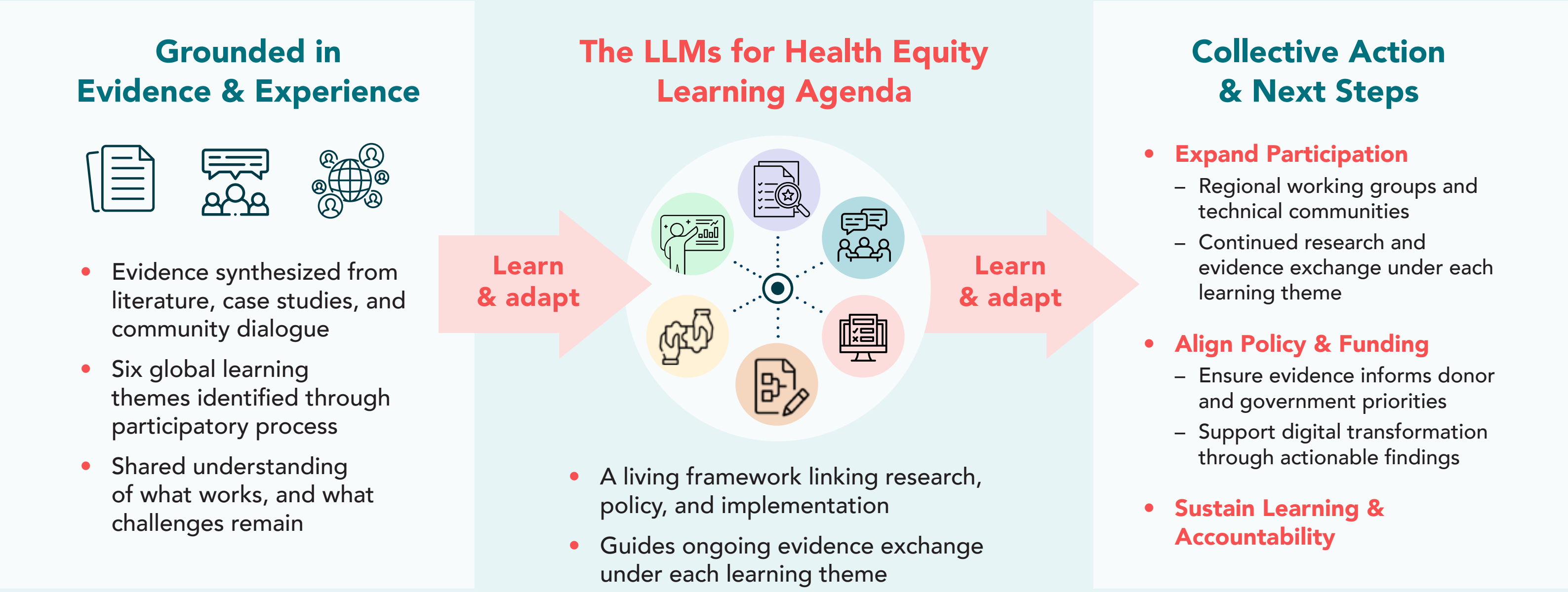
The following Learning Agenda Roadmap translates these thematic insights into concrete next steps. It organizes key research questions, learning activities, and intended outcomes across the six priority areas to guide collective action throughout 2026 and 2027. Designed as both a strategic and operational tool, the roadmap aims to:

- Align research and implementation efforts across partners and geographies;
- Promote coordinated evidence generation to inform policy and scale; and
- Foster an inclusive learning ecosystem that strengthens capacity, governance, and innovation across LMIC health systems.

By advancing these shared priorities, the roadmap provides a practical blueprint for collaboration, linking research, practice, and policy to accelerate the responsible adoption of AI and ensure that CDSS innovations serve the needs of all health workers and communities.

From Collective Learning to Collective Action

The Learning Agenda marks both a reflection of progress and a committment to future action—turning shared evidence into systems change for equitable, responsible use of LLMs in health.



Grounded in Evidence & Experience

- Evidence synthesized from literature, case studies, and community dialogue
- Six global learning themes identified through participatory process
- Shared understanding of what works, and what challenges remain

Learn & adapt

The LLMs for Health Equity Learning Agenda

- A living framework linking research, policy, and implementation
- Guides ongoing evidence exchange under each learning theme

Learn & adapt

Collective Action & Next Steps

- **Expand Participation**
  - Regional working groups and technical communities
  - Continued research and evidence exchange under each learning theme
- **Align Policy & Funding**
  - Ensure evidence informs donor and government priorities
  - Support digital transformation through actionable findings
- **Sustain Learning & Accountability**

Continuous learning, shared accountability, and collective leadership will ensure LLM-enabled CDSS serve as tools for trust, access, and quality for all.

CDSS Global Learning Agenda

2025–2027 Roadmap for Action

Core Learning Questions	Collaborative Learning Activities	Intended Outcomes
1. Localization & Language Equity		
• How can LLM-enabled CDSSs be safely localized across diverse languages and literacy levels? • What governance and data-sharing models best support equitable, community-led language datasets? • How can localization standards be embedded in national digital health strategies?	• Comparative studies of selective pre-translation and fine-tuning strategies for under-resourced languages. • Co-develop open, annotated medical corpora through community-based annotation networks. • Develop evaluation methods for linguistic accuracy, cultural relevance, and clinical safety. • Convene global partners to establish shared multilingual standards and data governance principles.	• Safer, more accurate multilingual CDSS. • Locally owned datasets and annotation ecosystems. • Global standards for equitable language inclusion in health AI.
2. Evaluation & Real-World Evidence		
• How can standardized yet adaptable frameworks measure safety, usability, and equity in LMIC contexts? • What evaluation designs generate robust, comparable real-world evidence across countries? • How can cost-effectiveness data inform procurement and policy decisions?	• Develop a global CDSS Evaluation Framework and metrics library adaptable to LMIC settings. • Implement evaluation clinics using locally generated clinical vignettes. • Conduct multicountry trials and cost-effectiveness studies on workflow impact and patient outcomes. • Build local evaluator and regulatory capacity through shared training and peer-learning exchanges.	• Strong evidence base for clinical and economic effectiveness of CDSSs. • Harmonized frameworks for evaluation and regulation. • Improved policy confidence and uptake of AI-enabled health tools.
3. Voice-Enabled & Multimodal Tools		
• Do voice-enabled and multimodal CDSSs improve access and care quality in low-literacy and low-connectivity settings? • What deployment models enable offline or hybrid use? • How can privacy and consent be safeguarded in voice-based systems?	• Comparative pilots of voice, text, and hybrid CDSSs with harmonized safety and usability metrics. • Build open benchmarks for ASR and TTS in African and South Asian languages. • Fund local data-collection initiatives (e.g., Karya, Common Voice) for medical ASR corpora. • Develop ethical toolkits and policy frameworks for voice-data governance and responsible deployment.	• Expanded access to CDSSs for frontline workers with limited literacy or connectivity. • Evidence-based models for safe and equitable voice-AI integration. • Strengthened ethical and regulatory safeguards for multimodal data.

Core Learning Questions	Collaborative Learning Activities	Intended Outcomes
4. Governance, Trust & Transparency		
• What governance models ensure accountability, transparency, and ethical oversight for LLM-enabled CDSSs? • How can explainability standards strengthen clinician and patient trust? • What mechanisms enable safe post-market monitoring and data sovereignty?	• Co-develop AI governance frameworks with LMIC regulators and ministries of health. • Create cross-country explainability and transparency standards for CDSS outputs. • Pilot sovereign data-hosting and post-market monitoring systems using clinician and user reporting. • Facilitate policy dialogues aligning donors, governments, and industry on responsible AI use.	• Trusted, transparent governance frameworks. • Improved clinician and patient confidence in AI-enabled health systems. • Strengthened post-market accountability and data sovereignty.
5. Capacity & Local Ownership		
• What models best build capacity to design, evaluate, and govern CDSS sin LMICs? • How can training interventions foster trust and proficiency among clinicians and policymakers? • How can partnerships ensure local data and research ownership? • How do patient perspectives inform AI adoption and trust?	• Establish regional training hubs and mentorship networks linking LMIC researchers with global developers. • Implement longitudinal studies tracking clinician proficiency and confidence. • Develop hybrid learning programs combining digital and analog formats for different health worker cadres. • Conduct qualitative research on patient and community perspectives to guide trust-building strategies.	• Strengthened institutional and technical capacity. • Increased clinician confidence and responsible AI use. • Locally led innovation, data ownership, and sustained adoption beyond donor cycles.
6. Infrastructure & Enabling Environments		
• What infrastructure investments and enabling policies ensure sustainable CDSS deployment? • How can computing and energy-efficient models (SLMs, TinyML) expand access? • How does AI adoption evolve as health systems digitize? • Which financing mechanisms best support scale-up and national integration?	• Conduct comparative analyses of compute, energy, and cost efficiency across CDSS models. • Pilot offline and low-bandwidth systems integrated with existing national data platforms. • Map digital infrastructure readiness and monitor AI adoption longitudinally. • Develop policy and financing frameworks embedding CDSS into digital health strategies.	• Resilient and sustainable infrastructure for AI in health. Reduced energy costs and environmental footprint • National ownership and alignment with digital transformation strategies.

Conclusion and Next Steps

The development of this Learning Agenda represents a collective milestone in advancing equitable, evidence-driven approaches to LLM-enabled CDSSs. Drawing from the experiences of implementers, researchers, policymakers, and funders across LMICs, it brings together a growing body of knowledge on what it takes to design, evaluate, and scale AI tools responsibly in real-world health systems.

The six global learning themes—localization and language equity, evaluation and real-world evidence, voice-enabled and multimodal tools, governance and trust, capacity and local ownership, and infrastructure and enabling environments—capture both the progress achieved and the challenges that remain. Together, they underscore that responsible AI in health care is not solely a technical endeavor, but a systemic one requiring coordination across data ecosystems, regulatory frameworks, workforce capabilities, and financing models.

Moving forward, this Learning Agenda provides a foundation for collective action and continuous learning. Its roadmap outlines a two-year plan for generating shared evidence, building local capacity, and informing global standards that can guide future CDSS implementation. In practice, this will involve:

- Expanding participation through regional working groups and technical communities that advance research and evaluation under each theme;
- Aligning with policy and donor priorities to ensure evidence generated is actionable and supports national digital transformation goals; and
- Sustaining learning and accountability through periodic synthesis reports, case documentation, and shared progress tracking across partners.

Ultimately, this Learning Agenda is both a reflection and a commitment, a reflection of what has been learned through early experimentation and collaboration, and a commitment to ensuring that the next generation of CDSSs is equitable, context-aware, and grounded in the realities of LMIC health systems. By fostering shared inquiry and partnership, the global community can turn emerging evidence into lasting systems change, where AI serves as an enabler of quality, trust, and access for all.

Contributions

The CDSS Learning Agenda was developed through a collaborative effort of the **LLMs for Health Equity CoP**, convened by **PATH** with funding from the **Gates Foundation**. This collective initiative reflects the experiences, priorities, and insights of a global community of researchers, implementers, and innovators advancing the responsible and equitable use of LLMs for health.

The CoP was co-led by Katie Thompson and Brian Kangethe, with coordination and support from Bilal Mateen, Mira Emmanuel-Fabula, Avery Wilson, Tanya Lalwani, and Jacqueline Deelstra of PATH. Their leadership and collaboration were instrumental in shaping the structure, content, and learning priorities presented in this manuscript.

We are deeply grateful to all CoP members for their engagement in shaping the scope, themes, and direction of this Learning Agenda. Special thanks are extended to those who provided detailed case narratives of their work, which directly informed the manuscript (denoted with an asterisk).

LLMs for Health Equity CoP Contributors

Andrew Bredenkamp – CLEAR Global

Ankita Sharma — University of Oxford, The George Institute for Global Health *

Dylan Green – Cooper/Smith

Sarah Morris – Audere Africa

Dr. Nirmal Ravi – EHA Clinics / eHealth Africa

Naina Ahuja – UNICEF

Prithviraj Pramanik – The George Institute for Global Health, India *

Huiqi Yvonne Lu – University of Oxford *

Anna Carter – Google

Tobi Olatunji – Intron Health *

Steven Wanyee – IntelliSOFT Consulting Limited

Habiba Issa Muller – GIZ OHDAA

Tyler R. Smith – Cooper/Smith

Vaishnavi Menon – University of Birmingham

Matthias Rüger – The Global Surgery Foundation

Rebecca Weintraub – Global Health Delivery Project, Harvard University *

Dr. Neha Verma – Intelehealth

Madior Gueye – Polytechnic Institute of Dakar (ESP/UCAD)

Elizabeth Geoffroy – Dimagi

Julie Rosenberg – Global Health Delivery Project, Harvard *

Gayatri Jayal – Dimagi

Hugo Manuel Paz Morales – Munai

Javier Elkin – ICRC

Puja Goswami – Intelehealth *

Aditya Naskar – Intelehealth *

Rainer Tan – Unisanté / SwissTPH

Sid Ravinutala – IDinsight

Joseph Ross – Ariadne Labs, Harvard School of Public Health and Brigham and Women's Hospital

Merrick Schaefer – The World Bank

Amrita Mahale – ARMMAN

Nneka Mobisson – mDoc Healthcare

Aparna Kumar – The Carter Center

Robert Korom – Penda Health *

Sarah Kiptinness – Penda Health *

Yasmina Al Ghadban – University of Oxford *

Prof. Jane Hirst – University of Oxford / Imperial College London *

Dr. Praveen Devarsetty – The George Institute for Global Health *

Stella Wanjiru – PATH

References

1. Government of India. (2024). *Language in India: 2024 status report on digital inclusion*. Ministry of Electronics and Information Technology.

2. Alhanai, T., et al. (2025). Benchmarking GPT-4o for medical reasoning across African and South Asian languages. *Nature Digital Medicine*.

3. Zhang, Y., et al. (2025). Selective pre-translation for low-resource LLM deployment in global health. *Journal of AI for Health Systems*.

4. Kumar, R., et al. (2024). Fine-tuning multilingual transformers for clinical NLP in Indian languages. *IEEE Transactions on AI in Healthcare*.

5. SMARThealth GPT Team. (2025). Developing SMARThealth Pregnancy (SHP) GPT: Addressing gender bias in LLMs for community health workers in rural India. *Proceedings of the International Conference on Gender and Technology 2025* (forthcoming).

6. Das, N. (2025). Inclusive AI design in action – Co-creating solutions with community health workers (Part 1) [version 1; not peer reviewed]. *Gates Open Research*, 9, 31. https://doi.org/10.21955/gatesopenres.1117211.1

7. University of Oxford & The George Institute for Global Health. (2025). *Clinical validation of SMARThealth GPT: Multilingual evaluation for community health workers in rural India*.

8. Transforming healthcare education: Harnessing large language models for frontline health worker capacity building using retrieval-augmented generation. (2023). *NeurIPS Workshop Proceedings*. https://doi.org/10.1101/2023.12.15.23300009

9. Kumar, R., et al. (2024). Fine-tuning multilingual transformer models for medical NLP in low-resource languages. *Journal of Biomedical Informatics*, 151, 104635. https://doi.org/10.1016/j.jbi.2024.104635

10. SMARThealth GPT Consortium. (2025). *Proposal for a task force on language equity in digital health*.

11. Intron Health. (2025). *Voice in health impact report* (Internal report, September 2025).

12. Olatunji, T., Afonja, T., et al. (2025). PATH LLM CDSS voice-enabled and multimodal tools study. *PATH and Intron Health Joint Publication*.

13. Olatunji, T., Afonja, T., et al. (2023). AfriSpeech-200: Pan-African accented speech dataset for clinical and general domain ASR. *Transactions of the Association for Computational Linguistics*.

14. Ng, J. J. W., et al. (2025). Evaluating the performance of AI-based transcription systems. *PubMed Central (PMC)*.

15. Mateen, H., et al. (2025). Generative AI in health: Evidence from early LMIC implementations. *BMJ Global Health*.

16. University of Birmingham. (2025). *Evaluation clinics progress report v0.1*.

17. Han, J., et al. (2024). Systematic review of AI trials in low-resource settings. *The Lancet Digital Health*.

18. Martindale, S., et al. (2024). Adherence to CONSORT-AI reporting standards in AI health research. *Nature Medicine*.

19. World Health Organization. (2023). *Ethics and governance of artificial intelligence for health*. WHO.

20. U.S. Food and Drug Administration, Health Canada, & Medicines and Healthcare products Regulatory Agency (MHRA). (2025). *Good machine learning practice (GMLP) for medical devices*.

21. Hanlon, J. T., Schmader, K. E., Samsa, G. P., et al. (1992). A method for assessing drug therapy appropriateness. *Journal of Clinical Epidemiology*, 45(10), 1045–1051. https://doi.org/10.1016/0895-4356(92)90144-C