STANFORD
CENTER FOR
DIGITAL HEALTH

AI for Health in Low- and Middle-Income Countries

STANFORD CENTER FOR DIGITAL HEALTH ROUNDTABLE DISCUSSION, 22 APRIL, 2025

Gates Foundation

The ROCKEFELLER FOUNDATION

Google.org

GiveWell

CLINTON HEALTH ACCESS INITIATIVE

Stanford University

MULAGO

AH ADVANCING HEALTH ONLINE

CARNEGIE ENDOWMENT FOR INTERNATIONAL PEACE

BAY AREA Global Health ALLIANCE

The Agency Fund

UNDP

Introduction

Building on the publication in March 2025 of a white paper scoping the use of generative AI (GenAI) for health in low- and middle-income countries (LMICs), the Stanford Center for Digital Health (CDH) convened a roundtable of experts from academia, public health, tech development and funding institutions to further explore the role AI can play in strengthening LMIC health systems. Set against the challenging backdrop of declining overseas development assistance and overburdened healthcare systems, this convening aimed to distill actionable insights for funders and implementers alike.

Throughout the discussion, many points of alignment emerged, including the need for bold, ambitious action rather than relying solely on incremental improvement; pragmatic integration with existing systems over tech for its own sake; and transparency and coordination between funding and implementing organizations. A particular focus of the roundtable was the need for practical evaluation frameworks, to enable funders to ground investment decisions on robust evidence of impact and scalability.

Participants represented the following organizations: Gates Foundation; Rockefeller Foundation; Google.org; GiveWell; Clinton Health Access Initiative; Mulago; Advancing Health Online; Carnegie Endowment for International Peace; Bay Area Global Health Alliance; Agency Fund; UNDP; and Stanford University. Views expressed during the discussion and in this report do not represent the position of any one organization.

Key Take-Aways

Participants identified common challenges faced by funders that could benefit from shared solutions. These include:

1. **ESTABLISH COMMON WAYS TO MEASURE KEY ASPECTS OF AI INFRASTRUCTURE, WITH A FOCUS ON INCLUSIVITY.**

   Representative language and health-use-case data: Invest in shared benchmarks across languages and health use cases. Develop datasets for under-resourced languages, with a focus on health-specific language data and optimizing audio and transcription capabilities.

   Establish shared real-world benchmarks: Define performance expectations based on users' access to care, health literacy, and device usage.

   Taxonomy and measurement of harms: Develop a typology of potential harms arising from the implementation of AI-based health interventions in LMICs and pragmatic, ongoing measurement methodologies. CDH's taxonomy of harms for youth users of social media provides a model.

2. **ESTABLISH CLEARER PATHS FOR SUSTAINABLE FUNDING.**

   Align with government requirements: Understand how LMIC governments make investment decisions, and ensure funded projects can generate the data to inform these decisions. As part of this process, consider investments in general digital transformation initiatives (vs. solely sector- or issue-specific ones).

   Identify high-ROI measurement opportunities: Identify areas of large existing investment (e.g. government or private-sector AI efforts) that lack evaluation, and ensure evaluation for these projects is funded.

3. **IMPLEMENT ITERATIVE PRODUCT DEVELOPMENT AND PRAGMATIC, PHASED EVALUATION PROCESSES.**

   Define a standardized evaluation framework, with metrics tailored to each phase of product development, and measure how often it is followed: Match timing of evaluation with the pace of tech innovation—prioritize fast, iterative metrics early, and shift to outcome-based metrics as solutions scale.

   Increase technical capacity for product development and evaluation: Establish mechanisms of consulting support for key product development functions (analytics, user research, engineering, product management, safety and red-teaming) and stronger communities of practice for implementers and funders.

   Embed evaluation in funding models: Recommend that project proposals allocate a portion of awarded funding to integrated measurement and research.

4. **FACILITATE FASTER ACTIONABLE RESEARCH TO INFORM FUNDER DECISIONS.**

   Prioritize research questions: Publish and maintain a list of research questions considered most pressing by funders, and share with grantees and research teams.

   Streamline ethical review: Develop a standardized checklist for IRB review of AI health tools; clarify what requires ethical review and when.

5. **SHARE FUNDER LEARNINGS AND FACILITATE ONGOING COLLABORATION.**

   Define funding principles: Establish shared principles for how investment decisions are made.

   Establish funder community-of-practice: Establish a regular set of convenings for funders (similar to the roundtable) to share learnings and coordinate on areas of shared need.

Summary of the April 22nd roundtable discussion.

Session 1. Meeting the Moment: How Do We Step Up Now?

Participants began by reflecting on how a rapidly shifting global funding environment and accelerating GenAI capabilities should inform strategic priorities for health in LMICs: what will it take for GenAI to contribute meaningfully to health systems strengthening in LMICs, in the context of contracting foreign aid and the growing fragility of national health infrastructure in many low-resource settings? There was universal recognition of the scale and significance of recent funding cuts, with one participant describing this watershed moment as "the end of big aid". Against this backdrop, participants emphasized that GenAI is not a luxury add-on but increasingly a practical necessity to fill critical system gaps, including workforce shortages, data collection, and decision support.

Major themes of the discussion included:

- **Reassessment of priorities:** Participants highlighted an urgent need to reevaluate past health funding decisions and prioritize sustainable, efficient data systems. Cost-effective tools, such as online health surveys, offer opportunities to build foundational health data affordably.
- **Language and accessibility:** It was noted that AI systems often fail to perform well in underrepresented languages, creating major barriers for global uptake. Successful adoption by community health workers (CHWs) and patients requires tools that support local languages through features like speech recognition and transcription.
- **Systems fragmentation and local ownership:** Digital health ecosystems were seen as overly complex, fragmented, and externally driven. There was a strong call for modular, locally developed systems that countries can own, adapt, and scale according to their needs. Emphasis was placed on empowering CHWs with AI tools that are accessible and practical.
- **Need for context-specific benchmarks and evaluation frameworks:** Current benchmarks for LLMs are not sufficient in global health contexts. Participants called for the development of gold-standard evaluation datasets and performance thresholds specific to different populations and languages, including health-specific benchmarks within a specific language.
- **Reframing success—"gold standard" vs "good enough":** Amid failing systems, the question is no longer "Is AI better than a human doctor?" but "Is it better than nothing?". The focus must shift from gold-standard comparisons to what is "better than nothing" and feasible in emergencies. AI can fill systemic gaps, especially where traditional healthcare resources are lacking or stigmatized.
- **Governance, legitimacy, and the role of governments:** While some argued for bypassing slow-moving governments, others emphasized that state involvement remains crucial to ensure equity, regulation, and sustainability through long-term infrastructure support. However, civil society and private actors are often better placed and motivated to lead near-term innovation.
- **Integration over innovation:** Integration of AI into existing systems and workflows is now more urgent than developing new tools. Diagnostic AI often exceeds the performance of frontline workers in LMICs, but integration challenges remain the biggest barrier to scale.
- **Direct-to-user approaches and digital trust:** Reaching users through familiar platforms such as WhatsApp was highlighted as a key strategy for scale, with use of AI tools in some LMICs already relatively widespread. While direct access to tools is powerful, it is risky if tools lack legitimacy, and concerns were noted about the rise of unverified or ineffective tools. We need to consider where consumers can place their trust with institutions such as USAID being sidelined.
- **A call for bold, coordinated action:** The group emphasized the need to focus on a few high-impact priorities rather than scattershot innovation. Participants noted that the upcoming challenges in global health may dwarf the COVID-19 crisis in scale and complexity, urging swift, decisive, and collective action. The scale of transformation needs to be massive, and incremental improvements (e.g. 10% efficiency gains) may not justify investment.

Session 2. Funding Principles for Generative AI in Health

This session aimed to shift the conversation from general funding principles towards more concrete, actionable, high-impact initiatives in health-related AI. Participants focused on identifying shared priorities, forming durable working groups for ongoing collaboration beyond the roundtable, and improving decision-making processes around funding in this space. Participants from funding bodies were encouraged to discuss how they currently make funding decisions and what changes they would like to see going forward.

Shared principles
The discussion identified the following key principles to inform funding decisions:

1. Solutions should match AI use cases to health needs and user contexts, rather than offering tech innovation for its own sake.
2. Solutions should demonstrate the potential for transformative, system-level impact, in addition to cost-effective incremental gains.
3. Funded projects should have a clearly defined and realistic pathway to scale and integration within existing systems. This will include projects governments have expressed willingness to sustain financially in the long term, and which have secured government buy-in through early communication and involvement.
4. Grantees should prioritize measurement, iterative learning, evidence generation, and dissemination of findings (positive and negative).
5. Implementing teams should demonstrate the ability to act with agility to respond to the rapidly evolving pace of innovation.
6. Prioritize projects which address areas of strategic whitespace and reduce system fragmentation, rather than duplicative pilots.

Key focus areas
Participants identified the following areas as potential avenues for coordination to scale impact of investments:

- Using AI to strengthen the community health workforce
- Improving workflow efficiency within health systems
- Supporting research and evidence generation with AI tools
- Developing and using evaluation benchmarks and data standards
- Investing in capacity building for implementers and local partners

Participants highlighted that the ecosystem is highly fragmented, with too many small, disconnected projects being funded, and a lack of incentive to scale what is already working. Some expressed a desire to fund in a more coordinated and intentional way, to minimize duplication and increase impact.

One of the most consistent themes to emerge from the session was the need for robust evidence. Participants emphasized that while there is considerable enthusiasm around GenAI, there is also a strong need to ground funding decisions in credible evidence. Without this, it becomes hard to justify investment or scale projects in a meaningful way.

Scalability was also a major focus, with an important distinction drawn between mere *potential* for scale and clearly defined *pathways* that take those pilots to millions of people quickly and affordably. Several participants noted that impressive early pilots are easy to build, but achieving widespread use—especially in low-resource settings—is where things often fall apart, with assumptions made about scalability that are not borne out in practice. This includes thinking about whether a tool can realistically be adopted by governments or large institutions, and whether it can operate at scale.

Readiness of technology was highlighted as a key factor in achieving government buy-in, with funders seeking solutions that are ready, or nearly ready, for integration into government systems and public health infrastructures. This includes

understanding what it takes to make these tools interoperable, secure, and easy to deploy at a regional or national scale. Participants reiterated calls for collaboration to help bring technology up to the level where governments can engage with it at scale.

In evaluating potential projects, a recurring theme was the need for user-centered design. Solutions must be developed with the end user in mind, whether community health workers or patients. Projects that ignore this run the risk of poor uptake or even active harm. The example of mental health was raised, with concerns over AI tools reinforcing dependence on technology rather than promoting agency and behavioral change.

Organizational capability was another important factor. Participants agreed that successful implementation goes beyond funding code development, requiring internal ownership and a product mindset. Organizations must treat AI tools not as one-off projects, but as core components of their service models, including building teams with product management expertise, and embracing iterative learning through A/B testing and rapid experimentation.

Existing networks and relationships were acknowledged to be key to how funding decisions are made. Organizations having established relationships with other funders was agreed to help expedite due diligence.

Red flags

Participants also discussed what kinds of projects should *not* be funded. These included:

- projects promising easy or all-encompassing solutions
- organizations pitching tech solutions for the sake of tech, without creating clear value
- ideas with no clear plan for scale
- initiatives without the agility to respond and adapt quickly.

Session 3. Pragmatic evaluation frameworks and real-world examples of their application

This session addressed the key challenge of how to measure the effectiveness, safety, and sustainability of GenAI in LMIC health settings in ways that are rigorous but agile. Traditional RCTs are crucial in the long term but can take a long time to provide outcome data. In the interim, implementers and funders also need scientifically valid ways to measure benefits, risks and costs. Participants aimed to develop a practical empirical framework for phased quantitative evaluation, focusing on direct-to-consumer tools which represent a significant class of current GenAI use cases in healthcare.

Participants agreed on the importance of tailoring measurement approaches to the phase of product development. The group aligned on a framework identifying four distinct phases of product development, which are similar to the 4 stages outlined in this document (Idea, R&D, Replication, Scaling):

0. **Problem statement**: Define the mission and target audience. Who is the tool designed to help? What is the unmet need? What user behavior or system inefficiency is it addressing?
1. **Product-market fit (PMF):** Assess usability, value, and initial engagement through pilots with small numbers of users.
2. **Validate impact:** Establish that the product has (or is likely to have) the intended system-level impact.
3. **Scale and continued learning:** Grow adoption of validated product, ensuring user value and impact seen at earlier phases is realized at scale.

Exit criteria should be defined upfront for each phase and reviewed before a project transitions to the next phase.

The following key dimensions should be measured across all phases of development, though measurement types will vary between phases, with faster, less expensive, less definitive measures (such as engagement, retention, proximal user reports) more appropriate at earlier phases, and higher-cost, slower, more definitive measures (such as health outcomes) important at later phases:

1. **Benefits and Efficacy:** How much value is this product creating for people? What is its impact on health outcomes?
2. **Risks and Safety:** How much risk is the product creating for people, and how will we bound risks?
3. **Cost and Value:** How much does the product cost, including cost of user acquisition? Is the ratio of benefit to cost reasonable relative to the status quo?

There was broad agreement among the group on several key evaluation considerations relevant to particular development phases, with this document providing relevant recommendations. These included:

- **Phase 0 (Problem statement):**
    - **Clarity of purpose:** The importance of clarity and concision in articulating a problem statement was repeatedly stressed, with suggestions of < 8-word mission statements or one-sentence "X→Y→Z" theories of change. Several participants referenced templates they provide grantees.
- **Phase 1 (PMF):**
    - ***Efficacy considerations:***
        - Key elements to track include onboarding success; activation (user engagement after onboarding); retention (continued engagement over time—e.g. weekly/monthly active users)
        - Organizations should implement user funnels to monitor where people drop off, and why
    - ***Cost considerations:***
        - Funding of speculative projects should consider the value of learnings for future decisions
        - Evaluation should measure cost and potential savings for governments to facilitate government buy-in

    - ***Safety considerations:***
        - Need for continuous evaluation and monitoring given ongoing product changes, including changes to underlying models
        - Need to define use cases that should be vs. should *not* be regulated
        - Currently many projects get stuck at ethical review stage—there is a need for clearer guidelines on what needs IRB review vs. what does not
        - Regulators should be involved from the beginning of the process
        - Need to have "real world" benchmarks for safety (e.g. compare to the current state of access vs. an "ideal" expert state—recognizing that many do not have access to quality care and information today)
- **Phase 2 (Validate impact):**
    - **Concrete outcomes over engagement:** Participants agreed on the need to focus not just on usage but on measurable behavior change and/or health outcomes—.e.g. did the user complete an antenatal visit?

Next Steps and Barriers to Realizing Impact

With general alignment on *what* to evaluate at each stage, participants identified several areas of focus to optimize *how* this can be achieved, many of which would benefit from centralized investment:

- **User Research Capacity:** Many organizations lack dedicated resources for user experience and behavioral research.
- **Engineering and Analytics Capacity:** Participants proposed establishing a CTO peer network to share insights, best practices and technical infrastructure.
- **Real-world benchmarking:** Traditional evaluation often uses expert comparisons, but users in low-resource settings may not have access to those standards. Benchmarks should be reflective of actual user conditions, taking into account baseline alternatives (e.g., current systems or no system at all).
- **Infrastructure for ongoing monitoring:** As AI models update frequently, continuous evaluation of performance is crucial.
- **Red-teaming infrastructure for risk assessment:** This is a common method for risk assessment used in industry, but not always readily available to the public sector.
- **Streamlining ethical review:** Ethical review processes were highlighted as a common sticking point for digital tools, with a lack of clarity over how AI tools should be classified, and what does and does not need review.
- **Identify where evaluation can be embedded in existing investment:** Governments and other entities are already investing in AI, but not all of these investments are paired with evaluation investment. Closing this gap could be a high-ROI use of measurement funding, since the implementation in these cases is already funded.

Session 4. High Impact Research Questions for Funders

This focused on translating the varied expertise and goodwill of participants into actionable research questions they felt funders should prioritize.

- ***Safety, ethics and regulation***
    - **Streamlining ethical review:** IRBs often struggle to categorize AI interventions, causing delays in ethical approval. Where are the avoidable inefficiencies in the review process, and where can time savings be achieved? EthiXPERT NPC in South Africa has explored the use of AI models to support ethical review and clinical trial decision-making.
    - **Taxonomy of harms:** Research into potential harms of AI interventions was advocated, with the aim of developing a formal taxonomy.

- ***Equity and inclusivity***
    - **Avoiding exclusion:** What demographics are at risk of exclusion by AI interventions? (age groups eg. children or old adults, geographic regions, languages)
    - **Heterogeneity across demographics:** How do different target groups experience AI interventions, and how do co-benefits and burdens vary across populations?

- ***High-potential applications***
    - How can AI support Community Health Care workers to deliver better advice and care?
    - How can AI enable low-cost psychotherapy and mental health support in LMICs?
    - How can direct-to-consumer (D2C) AI models be adapted for use by self-help and peer groups?

- ***Miscellaneous evidence gaps***
    - **Systematic reviews**: Participants called for a systematic review of unintended consequences from previous tech waves (e.g., mobile phones, EMRs, social media) to better anticipate risks.
    - **Exit points and bias**: How do we set up early exit points in studies without introducing confirmation bias? Are early adopters different from late adopters, leading to selection bias?
    - **Direct-to-consumer scaling evidence**: What level of evidence is needed to responsibly scale direct-to-consumer (D2C) health AI tools in LMICs?
    - **Misinformation**: How is AI-driven misinformation scaling in the Global South? How can we measure and mitigate it effectively?
    - **Behavioral impact**: Which behaviors is AI best positioned to influence or trigger positively for health outcomes?
    - **General vs. specific tools**: Do more general or more targeted AI tools generate different health outcomes?
    - **Data loss risk**: What existing datasets can be leveraged quickly, and how do we prevent the loss of critical health data for LMICs?
    - **Academic vs. tech research**: What is the distinct and additive role for academic research on GenAI in health?

- ***Strategic considerations***
    - **Scaling and amplifying:** Which countries and companies can best amplify foundation-funded GenAI projects?
    - **Political strategies:** What are the range of effective political strategies for GenAI diffusion, especially in health sectors?
    - **Cost-effectiveness metrics**: How should governments value cost-effectiveness for GenAI health interventions? What metrics do policymakers need to justify investment?
    - **Demand creation**: How can narrow engagement with AI tools be leveraged to create broader demand for health services?
    - **Funding for data and measurement**: Funders need to recognize that building data infrastructure and robust measurement systems is critical to answer these questions and track long-term impact.

STANFORD
CENTER FOR
DIGITAL HEALTH