WISELEAF SCIENTIFIC VENTURES

SMART HEALTHCARE REVOLUTION: AI STRATEGIES FOR SUPERIOR GLOBAL HEALTH

CHAPTER 10- SCALING AI SOLUTIONS FOR LOW-RESOURCE HEALTHCARE SETTINGS

Mr. Mahesh Asaram Bagmare

Assistant Professor, Department of Pharmacognosy, Manwatkar College of Pharmacy, Ghodpeth, Chandrapur, Maharashtra, India

Email: maheshperfect123@gmail.com

ABSTRACT

Artificial intelligence (AI) is often presented as a way to stretch scarce clinical expertise across health systems that lack radiologists, ophthalmologists and specialist nurses. The evidence for that claim is strongest in a small number of image-based screening tasks, and even there the distance between a promising validation study and a functioning national programme is large. This chapter examines what it takes to move AI from pilot to scale in low-resource settings, with most attention given to low- and middle-income countries (LMICs) and to India. It begins with the resource gap, using World Health Organization (WHO) workforce projections, and with the recurring failure of digital health pilots to become routine services. Two exemplar cases are then analysed in depth. Computer-aided detection (CAD) of tuberculosis on chest radiographs, conditionally recommended by WHO in 2021 for people aged 15 years and older, is traced through independent evaluations in Nepal, Cameroon, Bangladesh, Pakistan and Viet Nam, an individual patient data meta-analysis, and the problem of threshold calibration. Deep learning screening for diabetic retinopathy is followed from retrospective accuracy studies in India and Thailand to a prospective national deployment and an ethnographic study of the same system in Thai clinics, together with an offline smartphone system evaluated in Mumbai. Shorter sections cover automated visual evaluation for cervical screening, tools for community health workers, technical enablers such as on-device inference, economic evaluation, dataset shift and local validation, regulation, and workforce sustainability. The chapter closes with a layered scaling framework and a pilot-to-scale checklist. Its central argument is that accuracy is necessary but rarely sufficient: AI scales when data, infrastructure, workforce, regulation and financing are designed together around a defined clinical pathway, and when performance is re-checked in every new setting.

Keywords: artificial intelligence; low- and middle-income countries; tuberculosis; diabetic retinopathy; implementation science; health systems

1. Introduction

Most of the published success stories of clinical AI come from well-resourced hospitals with large digital archives, stable networks and specialists available to adjudicate the reference standard. The settings that could benefit most from automation look very different. A district hospital may have an X-ray unit but no radiologist; a primary health centre may screen people with diabetes but have no ophthalmologist within a day's travel; a community health worker may carry a basic smartphone and an intermittent mobile signal. For these settings the question is not whether an algorithm can match an expert on a curated test set. The question is whether a tool can be bought, installed, trusted, maintained and paid for, year after year, by a public health system that already struggles to staff its clinics.

That question has a sobering history. Digital health in LMICs has produced a large number of pilots and a much smaller number of national services, a gap that Heeks described for health information systems as a mismatch between the assumptions built into a design and the realities of the place where it is used [1]. AI inherits this problem and adds new ones. Its behaviour depends on the data it sees, so a model that performs well in one population can degrade quietly in another. Its outputs are probabilistic, so someone must choose an operating threshold. And its evidence base is still dominated by retrospective studies; an analysis of AI devices approved by the United States Food and Drug Administration between 2015 and 2020 found that most had been tested only on retrospective data and that many had been evaluated at only one or two sites [2].

This chapter is organised around evidence rather than enthusiasm. Section 2 describes the resource gap that motivates AI in LMICs. Section 3 asks why pilots so often fail to scale. Sections 4 and 5 examine the two clinical applications with the deepest LMIC evidence, tuberculosis (TB) chest radiograph CAD and diabetic retinopathy (DR) screening. Section 6 reviews other point-of-care and frontline tools. Sections 7 to 11 treat the cross-cutting enablers and constraints: technical design, economics, dataset shift, regulation and workforce. Section 12 brings these together as a practical scaling framework. Federated learning, which is one answer to data-sharing constraints, is treated in Chapter 8 and is mentioned here only in passing.

2. The Resource Gap

2.1 Workforce

The most widely quoted justification for AI in global health is the shortage of health workers. WHO's 2016 global strategy on human resources for health projected a shortfall of 18 million health workers by 2030, a figure that has since been revised to about 10 million on the basis of newer data [3]. The revision comes from an analysis by Boniol and colleagues, which estimated

a global stock of 65.1 million health workers in 2020 and a shortage of 15.4 million in that year, projected to fall to 10.2 million by 2030 [4]. The improvement is unevenly distributed. The same analysis reported that shortages were not improving in the WHO African and Eastern Mediterranean regions, and that about two-thirds of the projected 2030 shortage would be concentrated in 47 countries, most of them in the African region [4].

Aggregate numbers understate the problem for specialist tasks. Reading a chest radiograph for TB or grading a retinal photograph for DR requires training that is concentrated in cities and in the private sector. An X-ray unit in a district hospital is of limited value for TB detection if nobody is available to read the image. Where readers exist, their performance varies. In the Nepal and Cameroon evaluation discussed in Section 4, deep learning systems were compared with four groups of human readers, and how the systems compared differed from one reader group to another [5].

2.2 Infrastructure and connectivity

Connectivity has improved faster than workforce. The International Telecommunication Union (ITU) estimated in 2024 that about 96% of the world's population lived within range of at least a 3G mobile network and about 92% within range of 4G [6]. Coverage, however, is not the same as a usable connection inside a rural clinic, and it says nothing about data costs, power reliability or device maintenance. WHO's global strategy on digital health makes the point that simple channels such as text messaging and voice remain important in low-income settings because they work on basic phones and weak networks [7]. An AI system that assumes continuous broadband to a cloud server will therefore fail in exactly the places it is meant to serve, a lesson that recurs in Section 5.

India offers an instructive example of what national digital infrastructure can achieve. Its eSanjeevani telemedicine service has been studied as a nationally adopted platform for remote consultation [8]. Platforms of this kind matter for AI because they create the routine data flows, identity systems and referral links into which a diagnostic algorithm can be inserted. Without them, every AI deployment must build its own pipeline, which is expensive and rarely sustainable.

3. Why Pilots Fail to Scale

The phrase "pilotitis" is used informally in global health to describe the proliferation of small, donor-funded digital projects that never become part of routine services. The pattern has several causes, and they are worth separating because each calls for a different remedy.

The first is a design-reality gap. Heeks argued that information systems fail when the assumptions of their designers, often based in high-income settings, diverge from local realities of infrastructure, skills, management and culture [1]. An AI model trained on images from

tertiary hospitals and tested in a research clinic may meet every accuracy target and still fail in a screening camp where images are taken by a newly trained operator under poor lighting.

The second cause is that pilots are usually evaluated on efficacy rather than on the attributes that determine population impact. The RE-AIM framework, proposed by Glasgow and colleagues, distinguishes reach, effectiveness, adoption, implementation and maintenance [9]. Most AI pilots report effectiveness, typically as sensitivity and specificity, and say little about how many eligible people were reached, how many facilities adopted the tool, how faithfully it was used, or whether it was still running two years later. The updated Consolidated Framework for Implementation Research (CFIR) makes a similar point from another angle, drawing attention to the inner and outer settings, the individuals involved and the implementation process as well as the innovation itself [10].

The third cause is that a pilot can hide costs that a programme cannot. Research staff, free software licences, vendor engineers on site and dedicated internet links all disappear at the end of a grant. When the recurrent cost of licences, maintenance, training and quality assurance becomes visible, the business case may no longer hold.

The fourth cause is evidential. Even successful pilots often show only that the algorithm agrees with a reference standard. A commentary on the prospective Thai DR deployment described in Section 5 observed that its benefits for care delivery were largely implied rather than demonstrated, and that questions of workflow, specialist access and data standards remained open [11]. Health ministries need evidence that patients were referred, treated and had better outcomes, and that is much harder to generate.

These four causes recur throughout the case studies that follow. The cases are useful precisely because they include failures and partial successes as well as headline accuracy figures.

4. Case Study: Computer-Aided Detection of Tuberculosis on Chest Radiographs

4.1 From research tool to WHO recommendation

TB CAD is the clearest example of an AI application that has moved from academic evaluation into international policy. In March 2021 WHO issued consolidated guidelines on systematic screening for TB disease that included a new recommendation: among individuals aged 15 years and older in populations where TB screening is recommended, CAD software may be used in place of human readers to interpret digital chest radiographs for screening and triage [12]. The recommendation is conditional and based on low-certainty evidence, and WHO limited it to plain radiographs for pulmonary TB in people aged 15 years or older [12]. The guideline noted that the evaluations it reviewed showed substantial variation in accuracy across contexts, so that CAD would need to be calibrated for the purpose and setting of its use [12].

The guideline devotes a section to initial calibration of CAD technologies, which makes local threshold setting part of implementation rather than an optional extra [12].

The guideline's caution deserves emphasis. A conditional recommendation on low-certainty evidence is permission to use a tool carefully, not an endorsement of any product at any threshold. Programme managers who read it as the latter are likely to be disappointed.

4.2 The independent evaluations

The evidence that persuaded WHO came largely from independent evaluations organised by the Stop TB Partnership and academic collaborators, which matters because vendor-funded studies of their own products are prone to optimistic bias. Qin and colleagues evaluated three systems, CAD4TB (Delft Imaging), qXR (Qure.ai) and Lunit INSIGHT CXR, on 1,196 radiographs from symptomatic outpatients in Nepal and Cameroon, using Xpert MTB/RIF as the reference standard [5]. The areas under the receiver operating characteristic curve (AUC) were 0.94 for Lunit, 0.94 for qXR and 0.92 for CAD4TB, and the differences were not statistically significant. At 95% sensitivity, specificity was 80% for CAD4TB, 76% for Lunit and 72% for qXR. Used as a triage step, the systems could have reduced follow-on Xpert testing by 66% while keeping sensitivity at or above 95% [5].

A much larger evaluation followed in Dhaka, Bangladesh, where five products were applied to radiographs from 23,954 people aged 15 years and older attending three TB screening centres [13]. The AUCs were 90.81% for qXR, 90.34% for CAD4TB, 88.61% for Lunit INSIGHT CXR, 84.90% for InferRead DR and 84.89% for JF CXR-1 (Figure 10.1A). All five outperformed the three radiologists who read the same images, and all could halve the number of Xpert tests while keeping sensitivity above 90%. At 90% sensitivity, qXR reached 74.3% specificity and CAD4TB 72.9%, both meeting the WHO target product profile for a triage test [13]. The same study reported weaker performance in older people and in people with a history of TB, a finding with direct programmatic consequences because both groups are prominent in screening populations.

Prospective data came from Karachi, Pakistan, where Khan and colleagues enrolled symptomatic adults and used culture as the reference standard, which is stricter than Xpert alone [14]. Among 2,198 participants, of whom 272 had culture-confirmed TB, qXR had sensitivity 0.93 and specificity 0.75, and CAD4TB had sensitivity 0.93 and specificity 0.69. Both met the WHO minimum accuracy benchmarks within a prespecified non-inferiority margin in this largely HIV-negative population [14].

The widest product comparison came from Viet Nam, where Codlin and colleagues tested 12 CAD solutions against a library of 1,032 radiographs from community mobile screening in Ho Chi Minh City [15]. AUCs ranged from 0.50 to 0.82, with qXR, CAD4TB and Lunit INSIGHT sharing the top value of 0.82, and six products performing on par with an expert reader [15].

Two lessons follow. The market is heterogeneous, and the leading products in one evaluation are not guaranteed to lead in another. AUCs in community screening, where disease is often earlier and subtler, were also lower than in the symptomatic clinic populations of the earlier studies.

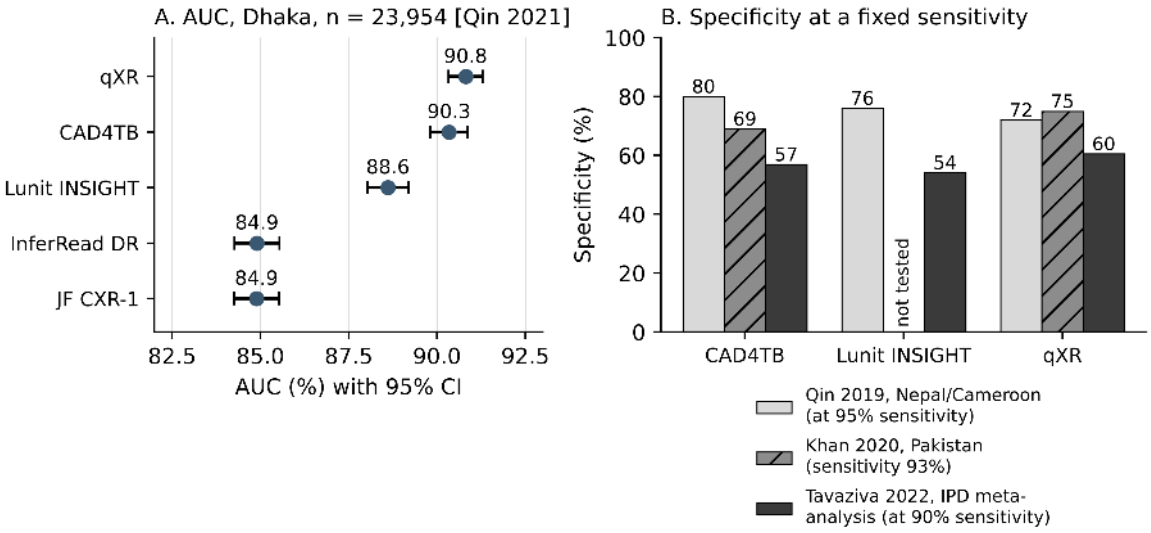

Figure 10.1: Published accuracy of TB chest radiograph CAD products. (A) AUC with 95% confidence intervals for five products in Dhaka, Bangladesh [13]. (B) Specificity at a fixed or reported sensitivity for three products in Nepal and Cameroon (at 95% sensitivity) [5], Karachi, Pakistan (at the reported sensitivity of 93%; Lunit not tested) [14], and an individual patient data meta-analysis (pooled, at 90% sensitivity) [16]. Values are taken directly from the cited abstracts; product versions differ between studies.

4.3 Where accuracy breaks down

Pooled data tell a more cautious story than single-site studies. An individual patient data meta-analysis by Tavaziva and colleagues combined 3,727 participants from four studies, 17% of whom were living with HIV [16]. At a fixed sensitivity of 90%, pooled specificity was only 56.9% for CAD4TB version 6, 54.1% for Lunit and 60.5% for qXR version 2 (Figure 10.1B). Sensitivity was lower among people living with HIV for CAD4TB and qXR, by an adjusted 13.4 percentage points for each, and lower in smear-negative than in smear-positive disease for all three products [16]. These are precisely the patients whom TB programmes in high-HIV settings most need to find.

A history of previous TB creates a different difficulty. Healed lesions remain visible on radiographs, so CAD scores in people with prior TB are often high regardless of current disease. The Dhaka evaluation reported lower performance among people with previous TB [13], and a later comparison of software versions on the same population found that both CAD products and human readers performed worse in people with a history of TB and in older age

groups [17]. In practice this means that a single threshold applied to an entire screening population will either flood the laboratory with false positives from previously treated people or miss active disease in others.

The screening context is a third complication. The Viet Nam evaluation, drawn from community mobile screening rather than symptomatic clinic attendees, produced a best AUC of 0.82 [15], against AUCs of 0.92 to 0.94 in symptomatic outpatients in Nepal and Cameroon [5]. Community screening finds earlier and subtler disease, and the prevalence of TB among those screened is lower. The overall picture is consistent: products that perform well in symptomatic clinic attendees can behave differently in community screening, and the choice of threshold drives much of the variation in yield.

4.4 Thresholds, versions and calibration

Thresholds are where policy meets practice. In the Nepal and Cameroon study, the CAD4TB score needed to reach 95% sensitivity was 63 in Nepal but 48 in Cameroon, and the authors recommended local threshold calibration [5]. WHO's guideline likewise treats initial calibration as an implementation step that programmes should carry out with their own data [12].

Software versions add another layer. Commercial CAD products are updated frequently. Re-analysing the Dhaka images, Qin and colleagues found that the AUC of CAD4TB rose from 0.823 in version 6 to 0.903 in version 7, and that of qXR from 0.872 in version 2 to 0.906 in version 3; only the newer versions met the WHO target product profile for a triage test [17]. The authors recommended testing each new version on local data before deployment, because the underlying networks can differ substantially between versions [17]. A threshold calibrated on one version cannot be assumed to transfer to the next. For programmes this is an operational burden: every update should trigger at least a check of score distributions against a locally verified sample, and procurement contracts should specify how and when updates are delivered.

4.5 Deployment in India and elsewhere

India is also home to developers in this field: qXR, one of the products evaluated above, is made by an Indian company [5]. For a country of India's size, though, the hard questions concern implementation rather than accuracy: where the X-ray unit is, who reviews CAD output, how flagged patients reach Xpert testing, and how results return to the screening team.

An earlier systematic review of AI chest radiograph reading for TB, which assessed risk of bias with QUADAS-2, concluded that most work up to that point had been on development rather than clinical evaluation [18]. The independent evaluations described above addressed many of those weaknesses, which is why TB CAD now has a stronger evidence base than most AI

applications in LMICs. Even so, almost all of that evidence concerns diagnostic accuracy and test reduction. Evidence that CAD-based screening increases treatment initiation or reduces transmission at population level remains limited.

4.6 What the numbers mean for a programme

Published accuracy figures become easier to judge when they are translated into the quantities a programme manager actually manages. Taking the Karachi study as an example, 272 of 2,198 participants had culture-confirmed TB, a prevalence of about 12.4% [14]. Applying the reported qXR sensitivity of 0.93 and specificity of 0.75 to 1,000 similar symptomatic adults gives roughly 124 people with TB, of whom about 115 would be flagged and 9 missed, and 876 without TB, of whom about 219 would be flagged in error. The programme would therefore send about 334 people for Xpert testing instead of 1,000, roughly a two-thirds reduction, at the cost of missing about 9 cases per 1,000 screened. This is simple arithmetic on the published estimates, not a new result, but it makes the trade-off explicit. Whether 9 missed cases per 1,000 is acceptable depends on what would have happened without CAD: if the alternative is that most people never receive a molecular test because cartridges are rationed, the trade-off is favourable; if every symptomatic person would otherwise be tested, it is not.

The same arithmetic shows why the pooled specificity values matter so much. At the 56.9% specificity reported for CAD4TB in the individual patient data meta-analysis [16], the number of false-positive referrals among 876 people without TB would rise from about 219 to about 378, and the saving in confirmatory tests would shrink accordingly. Small differences in specificity translate into large differences in laboratory workload when the prevalence of disease is low, which is the usual situation in community screening.

5. Case Study: Diabetic Retinopathy Screening

5.1 Retrospective accuracy in India and Thailand

Diabetic retinopathy is a leading cause of preventable vision loss, and screening by retinal photography followed by grading is effective but labour-intensive. The specialist-intensive grading models of high-income countries are difficult to reproduce where ophthalmologists and trained graders are scarce. Deep learning entered this field prominently with the work of Gulshan and colleagues, who developed and validated an algorithm for referable DR in retinal fundus photographs [19].

The same group then tested the algorithm prospectively on patients at two Indian eye hospitals, Aravind Eye Hospital and Sankara Nethralaya [20]. Among 3,049 patients, the algorithm detected moderate or worse DR with 88.9% sensitivity and 92.2% specificity at Aravind and 92.1% sensitivity and 95.2% specificity at Sankara. Manual graders at the two sites showed sensitivities of 75.7% to 89.8% and 73.4% to 84.2% respectively, but their specificity at

Sankara was higher than the algorithm's, at 98.6% to 98.7% [20]. The trade-off is typical: the algorithm found more disease, at the cost of more referrals.

In Thailand, a retrospective analysis of 25,326 gradable images from the national screening programme found that the algorithm detected referable DR with sensitivity 0.97 against 0.74 for human graders, with specificity of 0.96 against 0.98 [21]. Agreement with the adjudicated reference was higher for the algorithm, and false negatives fell by 23% at the cost of a 2% rise in false positives [21]. These results depend heavily on how the reference standard is built: the Thai study compared both the algorithm and the graders against adjudicated reference grades, and a less rigorous reference would have produced different figures [21].

5.2 Prospective deployment and the sociotechnical reality

Accuracy on archived images is one thing; a working screening service is another. Ruamviboonsuk and colleagues reported a prospective interventional cohort study in which the deep learning system was used in real time at nine primary care sites in Thailand's national programme [22]. Of 7,940 people screened for inclusion, 7,651 were analysed. For vision-threatening DR, the system had accuracy 94.7%, sensitivity 91.4%, specificity 95.4% and positive predictive value 79.2%, compared with 93.5%, 84.8%, 95.5% and 75.6% for retina specialist over-readers [22]. On paper, the system did at least as well as specialists.

The positive predictive value in the prospective study also deserves attention. A value of 79.2% means that roughly one in five people referred for vision-threatening DR by the system did not have it on specialist review [22]. In a programme where the nearest retina service is far away, every unnecessary referral has a cost in travel, lost wages and clinic time, and repeated false alarms can erode patients' willingness to attend. The specialists' own positive predictive value of 75.6% was lower still, so this is not a failure of AI relative to people. It is a reminder that a screening programme's capacity to absorb referrals must be planned from the referral rate, not from the sensitivity.

The same system had earlier been studied by human-computer interaction researchers who spent time in eleven Thai clinics observing nurses and interviewing them about the deployment [23]. Beede and colleagues described tensions between the model's thresholds for acceptable image quality and the quality of images that could actually be captured in a resource-constrained clinic, and showed how socio-environmental factors shaped nursing workflows, model performance and patients' experience [23]. A model that rejects images it judges ungradable is behaving safely, but in a clinic where a patient may have travelled some distance to attend, each rejection is a failed encounter. The ethnographic study did not contradict the accuracy data. It explained why accuracy alone does not predict whether a service works.

Taken together, the Thai studies are probably the most complete public record of an AI screening deployment in an LMIC, and they are useful for exactly that reason. The accuracy

study shows what the model can do; the ethnographic study shows what the clinic allows it to do; the commentary on the prospective study reminds readers that improved downstream outcomes had not yet been shown [11].

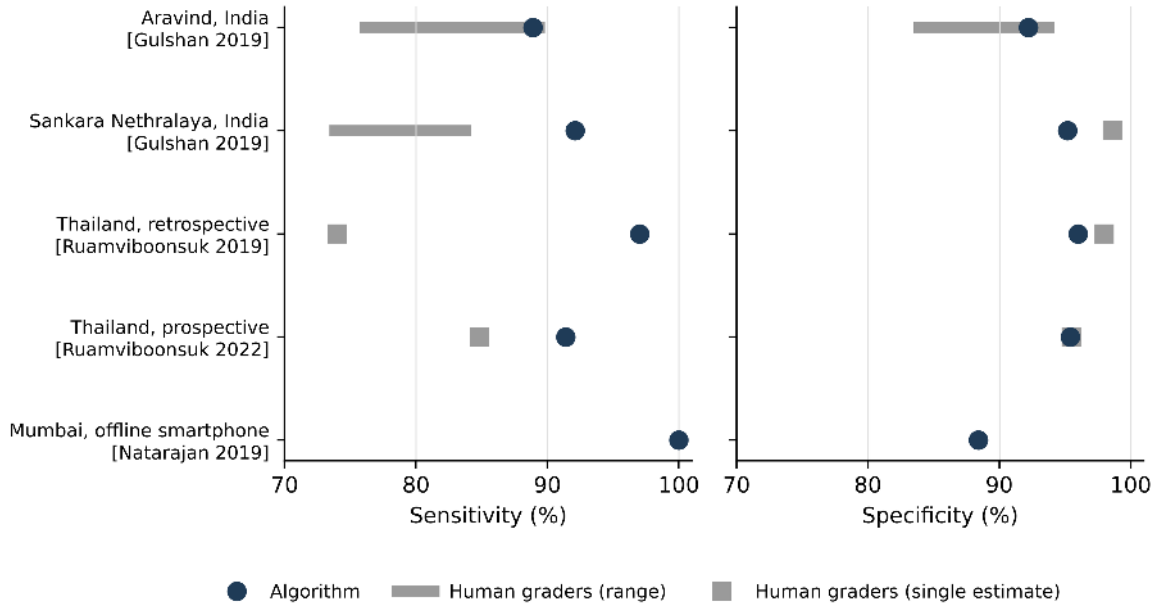

Figure 10.2: Sensitivity and specificity of deep learning DR screening compared with human graders in Indian and Thai studies. Dots show the algorithm; bars show the range across human graders and squares a single human estimate. Endpoints differ: moderate or worse DR [20], referable DR [21,24], and vision-threatening DR against retina specialist over-readers [22]. The Mumbai study had no human comparator.

5.3 Offline smartphone screening

The Thai experience shows why connectivity and workflow matter. One response is to move inference onto the device. Natarajan and colleagues evaluated an offline AI system running on a smartphone-based fundus camera (Remidio) in dispensaries run by the Municipal Corporation of Greater Mumbai [24]. Of 231 people with diabetes, 213 had gradable images. For referable DR, the system had sensitivity 100.0% (95% confidence interval 78.2% to 100.0%) and specificity 88.4% (83.2% to 92.5%); for any DR, sensitivity was 85.2% and specificity 92.0% [24]. The small number of referable cases explains the wide confidence interval for sensitivity, and the study should be read as a demonstration of feasibility in a primary care setting rather than as definitive accuracy evidence. Its significance for scale lies in the architecture: because the model runs on the phone, a result is available at the point of care without any internet connection.

5.4 Lessons from autonomous systems and head-to-head comparisons

Outside LMICs, the pivotal trial of an autonomous AI system for DR in primary care offices established a regulatory pathway for autonomous diagnosis [25]. The most sobering finding for programme managers, however, came from a head-to-head comparison of seven automated DR screening systems from five companies, using images from nearly 24,000 veterans screened in two United States health care systems [26]. Performance varied between systems and between the two sites, only one algorithm matched the human screeners, and one had lower sensitivity than human screeners for proliferative disease [26]. Two lessons follow for LMIC procurement. Regulatory clearance in one jurisdiction does not guarantee performance on local images, and products cannot be treated as interchangeable on the basis of their published headline figures.

Most economic evaluations of automated DR grading have been set in high-income programmes with established graders, so their conclusions cannot be transferred directly to settings where the alternative is no screening at all. That distinction is developed in Section 8.

6. Other Point-of-Care and Frontline Applications

6.1 Automated visual evaluation for cervical screening

Cervical cancer remains common in LMICs, where cytology programmes are difficult to sustain and visual inspection with acetic acid is subjective. Automated visual evaluation (AVE) applies deep learning to images of the cervix. In the foundational study, Hu and colleagues trained and evaluated an algorithm on cervical images from a longitudinal cohort of 9,406 women in Guanacaste, Costa Rica, followed for seven years [27]. AVE identified precancer and cancer with an AUC of 0.91, compared with 0.69 for expert interpretation of cervigrams and 0.71 for conventional cytology [27].

The result was striking, but it came from archived film images in a single well-characterised cohort, captured with a specialised camera. Translating AVE to smartphones and new populations has required retraining and revalidation, and it raises the same issues of image quality, device variation and threshold choice seen with TB and DR. AVE is therefore best regarded as a promising technology whose LMIC evidence is still being built, and programmes considering it should insist on validation with the specific capture device and population they intend to use.

6.2 Decision support for community health workers

Community health workers (CHWs) are the front line of primary care in much of South Asia and sub-Saharan Africa, and WHO's guideline on digital interventions for health system strengthening includes recommendations on digital support for health workers alongside other

interventions such as telemedicine and digital tracking [28]. The best-documented digital tools for CHWs are not yet AI tools in the modern sense. RapidSMS in Rwanda, a text-message system used by CHWs to track mothers and children, has been evaluated from the perspectives of its users and beneficiaries, with attention to the training, supervision and equipment needed to make it work [29]. MomConnect, South Africa's national maternal messaging service and helpdesk, shows how a digital service can reach national scale when it is owned by the health ministry and integrated with routine antenatal care [30].

These examples matter for AI because they establish the channels, trust and operational routines on which AI-based decision support for CHWs will depend. A risk-stratification model that has to be delivered through a new app, on a new device, with new training, faces every barrier that RapidSMS and MomConnect had to overcome, plus the additional problem of explaining probabilistic output to workers with limited clinical training. The more realistic path is to embed AI functions in tools that CHWs already use.

6.3 Handheld ultrasound and other emerging tools

AI-assisted handheld ultrasound, especially for obstetric assessment by non-specialist operators, is frequently proposed as a next candidate for scale in LMICs. The potential is real, because the specialist bottleneck in obstetric imaging is severe. The peer-reviewed evidence from routine LMIC services, however, is thinner than for TB or DR, and this chapter does not draw conclusions about specific products. The same caution applies to large language models offered as clinical assistants: their use in low-resource settings raises additional questions of accuracy in local languages and of safety, which national programmes should evaluate before deployment.

7. Technical Enablers for Scale

7.1 Edge and offline inference

The Mumbai DR study illustrates the most important technical design choice for low-resource deployment: where the model runs [24]. Cloud inference simplifies updates and monitoring but requires reliable connectivity at the moment of use. Edge inference, on a phone, a tablet or a computer attached to an X-ray unit, removes that dependency and keeps images on site, but makes updates, version control and monitoring harder. Many TB CAD deployments take an intermediate approach, with software installed on a local workstation attached to a digital radiography unit, including in mobile screening vans of the kind used in community case finding in Viet Nam [15].

Running a deep network on modest hardware usually requires model compression. Common techniques include quantisation, which stores weights at lower numerical precision; pruning, which removes connections that contribute little to the output; and knowledge distillation, in

which a small model is trained to reproduce the outputs of a larger one. The gains can be large. Han and colleagues combined pruning, trained quantisation and Huffman coding to reduce the storage required by the AlexNet image classifier 35-fold, from 240 MB to 6.9 MB, and that of VGG-16 49-fold, from 552 MB to 11.3 MB, without loss of accuracy on ImageNet, with three- to four-fold layer-wise speed-ups and better energy efficiency on mobile hardware [31]. Those results were obtained on a general image benchmark, not on medical images. Each technique trades some accuracy for speed and size, and the trade-off has to be measured on local data rather than assumed. A compressed model is in effect a new model and should be validated as such.

7.2 Low-bandwidth design and store-and-forward

Where some connectivity exists, design can make the most of it. Store-and-forward architectures queue images and results locally and transmit them when a connection becomes available. Results can be returned as short text messages, using the channels that WHO identifies as dependable in low-income settings [7]. Image quality checks at the point of capture, a lesson from the Thai DR deployment [23], reduce the number of wasted transmissions and repeat visits.

7.3 Interoperability and national platforms

A screening tool that produces results nobody can find is of little use. Integration with national digital health infrastructure and with disease-specific information systems determines whether a positive screen leads to a confirmed diagnosis and treatment. Integration also makes monitoring possible, because outcomes can be linked back to the AI output that triggered them. India's Ayushman Bharat Digital Mission (ABDM) illustrates the scale such infrastructure can reach: a Government of India backgrounder of July 2026 reported more than 93 crore (930 million) Ayushman Bharat Health Account (ABHA) identities and more than 104 crore health records linked to them [32]. Platforms of this kind, together with services such as eSanjeevani [8], show that national-scale digital services are possible in India, and AI services are more likely to be sustained if they are delivered through such platforms than as stand-alone products.

7.4 Language

Interfaces, instructions and patient-facing messages need to be available in the languages that operators and patients use. In a country such as India this is a substantial design requirement rather than a translation task at the end of a project. National language technology programmes may lower this barrier. Bhashini, the National Language Translation Mission of the Ministry of Electronics and Information Technology, managed by the Digital India Corporation, aims to make digital services accessible in more than 22 Indian languages and offers open-source AI models and crowdsourced language data to developers [33]. Published evidence on the clinical

use of such models is not yet available, so they should be evaluated in health applications before patient-facing use.

8. Economic Evaluation

Cost is often the decisive factor in whether an AI tool survives the end of a pilot, and economic evidence for AI in LMICs is thinner than accuracy evidence. Three features distinguish the economics of AI in low-resource settings.

First, the comparator is often not an expert human reader but no reading at all, or reading by an overstretched generalist. Economic models built for high-income screening programmes, where AI replaces part of an existing grading workforce, answer a different question from the one facing a district TB programme with no radiologist.

Second, the main savings frequently arise downstream. In TB, CAD is valuable less because it is cheaper than a radiologist than because it reduces the number of confirmatory molecular tests. The evaluations described in Section 4 estimated reductions in Xpert testing of 66% at 95% sensitivity in Nepal and Cameroon [5] and of 50% with sensitivity above 90% in Bangladesh [13]. Section 4.6 showed how such reductions depend on specificity and prevalence, and the authors of the version comparison noted that software updates can change threshold selection, the number of cases detected and cost-effectiveness [17]. National health technology assessments are needed because prices, test costs and disease prevalence vary greatly between countries, and published economic evidence for AI in LMIC settings remains thin.

Third, recurrent costs dominate. Per-image or per-device licence fees, hardware replacement, maintenance, training and quality assurance continue for the life of the programme, whereas pilots typically count only set-up costs. Codlin and colleagues drew attention to software pricing and licensing models as practical considerations in comparing CAD products [15]. An economic evaluation that omits recurrent costs, or that assumes pilot-level uptake and adherence, will overstate value.

The implication for ministries and funders is that cost-effectiveness should be assessed for the whole pathway, from screening through confirmation and treatment, under realistic assumptions about adherence to referral. An accurate test that generates referrals the system cannot absorb will not be cost-effective, however low its unit price.

9. Dataset Shift and Local Validation

9.1 Why models degrade in new settings

The variation in performance seen across the TB and DR studies is not random noise. It reflects dataset shift: differences between the data on which a model was developed and the data it

encounters in use. Zech and colleagues trained pneumonia models on chest radiographs from three hospital systems and found that performance on radiographs from outside hospitals was significantly lower than on held-out radiographs from the original systems in three of five comparisons; the networks could also identify the hospital system and department in which an image was acquired with very high accuracy, and used that information in their predictions [34]. Average accuracy can also hide failure in clinically important subgroups, as the HIV and prior-TB findings in Section 4 show [13,16]. Obermeyer and colleagues showed how a widely used population health algorithm encoded racial bias because it predicted health care costs as a proxy for health need [35]. The mechanisms differ, but each is a reason why average accuracy on a development population can mislead.

Table 10.1 summarises the main forms of shift observed in the case studies and the mitigations that programmes have used.

Table 10.1: Sources of performance variation in LMIC deployments of image-based AI, with examples and mitigations

Source of shift	What changes	Example from the literature	Practical mitigation
Population case mix	Prevalence, disease stage, comorbidity	Lower CAD sensitivity in people living with HIV and in smear-negative TB [16]; weaker performance with prior TB and older age [13,17]	Subgroup analysis in local validation; separate thresholds or rules for known subgroups
Screening context	Symptomatic clinic versus community screening	AUC up to 0.94 in symptomatic outpatients [5] versus a maximum of 0.82 in community mobile screening [15]	Validate in the intended use case, not only in clinic samples
Image acquisition	Camera, X-ray unit, operator skill, pupil dilation	Image-quality thresholds conflicting with images obtainable in Thai clinics [23]	Point-of-capture quality feedback; operator training; device-specific validation
Reference standard	Grader expertise, adjudication, test used	Algorithm and graders compared against adjudicated grades [21]; Xpert versus culture [14]	State the reference standard; use adjudication or culture where feasible
Threshold	Operating point on the score scale	CAD4TB score for 95% sensitivity of 63 in Nepal and 48 in Cameroon [5]	Local calibration as part of implementation [12]
Software version	Retraining, new architecture	AUC of CAD4TB 0.823 in version 6 versus 0.903 in version 7 [17]	Re-verify score distributions after each update; contractual change control
Product choice	Different vendors and training data	Only one of seven DR systems matched human screeners [26]; AUC 0.50 to 0.82 among 12 TB products [15]	Independent head-to-head evaluation before procurement
Proxy labels	Outcome used as a stand-in for need	Cost used as a proxy for health need [35]	Examine what the label measures and for whom it fails

9.2 What local validation should look like

Local validation need not be a large research study. For TB CAD, a programme can use a sample of locally collected radiographs with confirmatory test results to choose a threshold that

achieves a target sensitivity or a target reduction in confirmatory testing, the approach used in the independent evaluations [5,13] and anticipated in WHO's guidance on initial calibration [12]. For DR, a sample of locally captured images graded by an adjudicated panel serves the same purpose. The key elements are a sample that reflects the intended population and capture devices, a reference standard that is stated in advance, an analysis that includes clinically important subgroups, and repetition after significant changes in the population, device or software version.

Validation also needs an owner. In a pilot, the research team performs it. At scale, someone in the health system must have the skills, data access and mandate to repeat it, and that responsibility should be written into the programme design from the start.

9.3 Monitoring after deployment

Validation at launch is not enough, because populations, devices and software all change. Monitoring does not have to wait for outcome data, which are slow and incomplete in most LMIC information systems. Several indicators can be tracked routinely from the AI system itself and from the programme register: the distribution of model scores over time, the proportion of people flagged for referral, the proportion of images rejected as ungradable, the positivity of confirmatory tests among those referred, and the proportion of referred people who complete the next step of care. A sudden change in the score distribution after a software update, or a gradual fall in confirmatory test positivity, is an early signal that the threshold or the product needs review. The version comparisons in TB CAD [17] and the image-quality problems in Thai clinics [23] are exactly the kinds of change that such indicators would detect.

None of these indicators requires sophisticated infrastructure. They do require that AI outputs are stored with a version identifier and linked to the patient's subsequent results, which is a design decision that is cheap at the start of a programme and expensive to retrofit.

10. Regulation and Governance

10.1 Software as a medical device

Diagnostic AI is regulated in most jurisdictions as software as a medical device (SaMD). The International Medical Device Regulators Forum has set out principles for the clinical evaluation of SaMD, distinguishing valid clinical association, analytical validation and clinical validation [36]. These principles are useful for LMIC regulators because they provide a shared vocabulary without requiring each country to invent its own framework.

The weakness of regulation as a guarantee of performance has already been noted. Many cleared devices were evaluated retrospectively and at few sites [2], and cleared DR products

varied widely in real-world performance [26]. Regulatory approval should therefore be treated as a minimum condition for procurement, not as evidence of local performance.

10.2 The LMIC regulatory context

LMIC regulators face particular difficulties: limited staff with software expertise, reliance on approvals from reference regulators in high-income countries, and products that are updated faster than review processes can follow. India has moved furthest towards software-specific rules. A 2026 guidance document from the Central Drugs Standard Control Organisation (CDSCO) confirms that software intended for a medical purpose, whether standalone or driving hardware, is a medical device under the Medical Devices Rules, 2017, classifies standalone software into risk classes A to D according to the seriousness of the health situation and the significance of the software's output, and expects manufacturers of AI-based software to disclose dataset composition, assess bias across Indian sub-populations and submit an algorithm change protocol covering monitoring, retraining and rollback [37]. On the African continent, the treaty establishing the African Medicines Agency was adopted by the African Union Assembly in February 2019 and entered into force on 5 November 2021; the Agency's mandate is to strengthen and harmonise the regulation of medical products and technologies across member states [38]. Both developments may eventually reduce duplication of review, but the practical arrangements for adaptive AI software, including how updates and recalibration are handled, are still being worked out in most LMICs.

10.3 Ethics and data governance

WHO's guidance on the ethics and governance of AI for health sets out consensus principles intended to ensure that AI serves the public interest in all countries [39]. In low-resource settings, several issues are especially acute. Data used to train or calibrate models are often collected from populations with limited power to negotiate how they are used. Vendors may retain the right to use images from public programmes to improve proprietary products. And when an algorithm is deployed precisely because no specialist is available, the meaningful human oversight that ethical guidance often assumes may not exist in practice. Programmes should address data ownership, secondary use and responsibility for errors explicitly in contracts and protocols. WHO and ITU, through their joint Focus Group on AI for Health (FG-AI4H), set out to establish a standardised benchmarking process for health AI [40]. In that process, developers submit models to an online platform where they are evaluated on secure, confidential test data drawn from several populations and health systems, and performance can be compared with that of human experts [41]. Benchmarking of this kind could help regulators and purchasers in countries without their own testing capacity. The independent TB CAD evaluations described in Section 4 show the value of that approach.

11. Workforce, Training and Sustainability

AI does not remove the need for health workers; it changes what they do. In TB screening, radiographers or technicians operate the X-ray unit and CAD software, laboratory staff process the larger or smaller number of confirmatory tests, and clinicians manage the people referred. In DR screening, nurses or technicians capture images, and ophthalmologists must still treat those referred. The workforce shortage described in Section 2 therefore limits AI-enabled services as well as conventional ones [4]. A screening programme that doubles referrals without increasing treatment capacity has moved the bottleneck rather than removed it.

Training needs are often underestimated. The Thai DR study showed that image capture by nurses was the point at which many problems arose [23], and the evaluation of RapidSMS in Rwanda identified training, supervision and equipment as conditions for effective use by CHWs [29]. Training cannot be a single event at launch; staff turnover in rural facilities is high, and refresher training and supportive supervision need recurrent budgets.

Sustainability also depends on who owns the service. MomConnect reached national scale because it was a programme of the South African National Department of Health rather than of a donor or vendor [30]. For AI, ownership means that the ministry or programme controls the contract, the data, the thresholds and the monitoring, and that the budget line survives the end of external funding.

12. A Practical Scaling Framework

The case studies suggest that AI tools scale in low-resource settings when five layers are addressed together around a defined clinical pathway, and when performance is re-checked whenever the setting or the software changes. Figure 10.3 presents these layers as segments surrounding the AI tool rather than as a hierarchy, because a deficit in any one of them can stop a programme regardless of strength in the others.

The **clinical pathway** comes first. A tool is not deployed for "TB" or "diabetic eye disease"; it is deployed as a triage test before Xpert, a screening test in a community camp, or a decision aid in a primary care visit. Each use implies a different population, a different acceptable trade-off between sensitivity and specificity, and a different set of downstream services that must be ready.

The **data** layer covers the images and labels used for local validation and calibration, the reference standard, consent and governance arrangements, and the ability to link AI outputs to outcomes. The **infrastructure** layer includes devices, power, connectivity, the choice between edge and cloud inference, and maintenance. The **workforce** layer covers operators, CHWs and referral clinicians, their training and their supervision. The **regulation** layer includes device registration, the clinical evidence expected, and post-market duties for monitoring and updates.

The **financing** layer covers procurement, pricing models, the recurrent budget and the economic case.

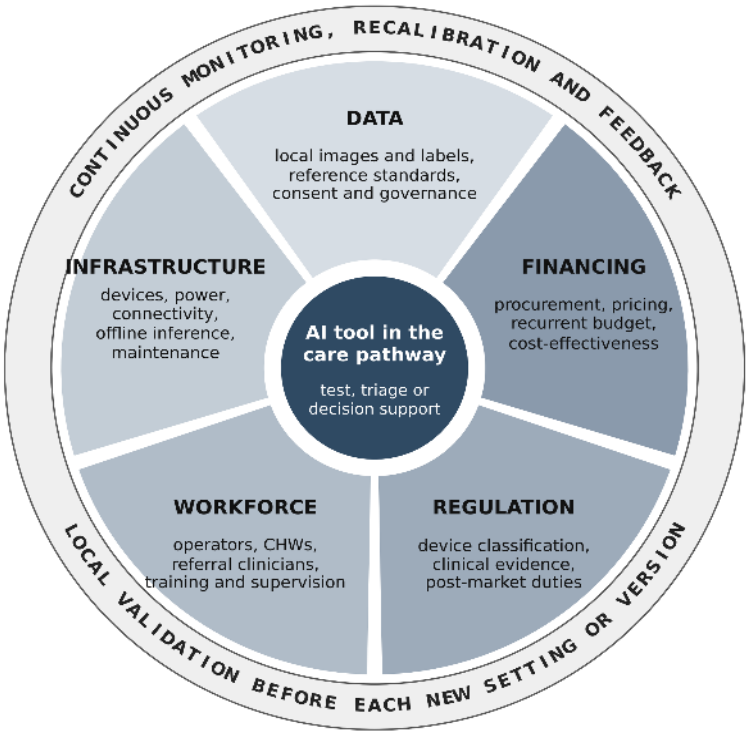

Figure 10.3: A layered framework for scaling AI in low-resource health systems. The AI tool sits within a defined clinical pathway, surrounded by five interdependent layers. The outer ring represents the continuous activities of local validation, monitoring and recalibration that link the layers over time. Original figure synthesised from the evidence reviewed in this chapter [9,10,12].

Table 10.2 applies this lens to the exemplar applications discussed in this chapter, and Table 10.3 summarises barriers and enablers by layer.

Table 10.2: Exemplar AI applications for low-resource settings: setting, task, evidence and scale reached

Application	Setting(s)	Task	Key evidence	Scale reached (as reported)
TB chest radiograph CAD (qXR, CAD4TB, Lunit INSIGHT and others)	Nepal, Cameroon, Bangladesh, Pakistan, Viet Nam, South Africa, Kenya	Screening and triage before molecular testing	AUC 0.92 to 0.94 in symptomatic outpatients [5]; AUC 84.89% to 90.81% across five products, n = 23,954 [13]; prospective culture-confirmed study [14]; IPD meta-analysis showing lower specificity and HIV effects [16]	Conditional WHO recommendation for people aged 15 years and older [12]; newer software versions meeting the WHO triage profile [17]

Application	Setting(s)	Task	Key evidence	Scale reached (as reported)
DR screening, deep learning system	India (Aravind, Sankara Nethralaya); Thailand national programme	Detection of referable and vision-threatening DR	Prospective Indian validation [20]; retrospective national Thai data [21]; real-time prospective study at nine sites [22]	Used within the Thai national screening programme in the prospective study; sociotechnical barriers documented in eleven clinics [23]
Offline smartphone DR screening (Remidio)	Municipal dispensaries, Mumbai	Referable DR on device without internet	Sensitivity 100.0% and specificity 88.4% for referable DR, n = 213 [24]	Feasibility in primary care; small sample
Automated visual evaluation	Guanacaste cohort, Costa Rica	Cervical precancer detection from images	AUC 0.91 versus 0.69 for cervigram reading and 0.71 for cytology, n = 9,406 [27]	Research stage; revalidation needed for new devices and populations
CHW digital tools (non-AI precursors)	Rwanda (RapidSMS); South Africa (MomConnect)	Tracking, messaging, helpdesk	User and beneficiary evaluation [29]; national scale-up history [30]	National programmes; channels on which AI decision support could build

Table 10.3: Barriers and enablers for scaling AI in low-resource settings, by layer

Layer	Common barriers	Enablers observed or recommended
Data	Development data unrepresentative of local populations; no local reference standard; proxy labels	Local validation samples with stated reference standards [12]; subgroup analysis [16]; scrutiny of labels [35]
Infrastructure	Intermittent connectivity and power; image-capture difficulties; device maintenance	Edge or offline inference [24]; point-of-capture quality checks [23]; integration with national platforms [8]
Workforce	Shortages of operators and referral clinicians [4]; high turnover; limited training	Recurrent training and supervision [29]; field observation of how health workers actually use the tool [23]; matching referral capacity to screening volume
Regulation	Limited SaMD review capacity; frequent software updates; reliance on foreign approvals	Shared clinical evaluation principles [36]; independent benchmarking [15]; change control for versions [17]
Financing	Pilot-only funding; hidden recurrent licence and maintenance costs	Whole-pathway economic evaluation that counts confirmatory tests saved [5,13]; ministry ownership of contracts and budgets [30]
Evidence	Retrospective accuracy studies; few outcome data [2,11]	Implementation outcomes measured with RE-AIM or CFIR [9,10]; prospective deployment studies [22]

A framework is useful only if it changes decisions. Table 10.4 converts it into a checklist that a programme manager, funder or ethics committee can apply before approving the move from pilot to programme.

Table 10.4: Pilot-to-scale checklist for AI tools in low-resource health systems

Stage	Question to answer before scaling	Evidence that would satisfy it
Use case	Is the clinical pathway defined, including what happens after a positive result?	Written pathway with referral, confirmation and treatment steps and named responsible staff

Stage	Question to answer before scaling	Evidence that would satisfy it
Independent accuracy	Has the product been evaluated independently of the vendor in a comparable population?	Peer-reviewed independent evaluation or national benchmark [13,15]
Local validation	Has performance been checked on local data, devices and operators, including key subgroups?	Local validation report with stated reference standard and subgroup results [12]
Threshold	Has the operating threshold been chosen for this setting and purpose?	Documented calibration and expected referral rate [5,17]
Workflow	Has the tool been observed in real clinics, and have the observed problems been addressed?	Field observation or process evaluation [23]
Infrastructure	Will the tool work with the power, connectivity and devices actually available?	Offline or store-and-forward capability tested on site [24]
Workforce	Are operators trained, supervised and replaceable, and can referral services absorb the extra demand?	Training plan with recurrent budget; capacity estimate for referrals [4]
Regulation	Is the product registered for this use, and is there a process for updates?	Registration status; contractual change-control and re-verification after updates [17,36]
Financing	Is the recurrent cost known and funded beyond the pilot?	Whole-pathway costing and budget line in the national or state programme
Monitoring	Who will monitor performance and outcomes, and how often?	Named owner, indicators linked to outcomes, and review schedule [9]
Governance	Are data ownership, secondary use and responsibility for errors agreed?	Data-sharing and liability terms consistent with WHO guidance [39]

Two features of the checklist deserve comment. Several items cannot be satisfied by the vendor alone; they require the health system to invest in its own analytical capacity. And none of the items is specific to AI. They are the questions that any new diagnostic should answer. What AI adds is the need to repeat them, because a model's behaviour can change when its population, its inputs or its software change.

13. Conclusion and Future Directions

The evidence reviewed in this chapter supports a qualified optimism. TB chest radiograph CAD has reached a WHO recommendation on the strength of independent, multi-country evaluations, and deep learning DR screening has been deployed prospectively within a national programme with accuracy comparable to retina specialists. Both applications address real shortages of specialist readers, and both can reduce downstream costs when used as triage tests.

The same evidence shows the limits of accuracy as a measure of readiness. Pooled data revealed much lower specificity than single-site studies and reduced sensitivity in people living with HIV. Thresholds and software versions changed performance. Real Thai clinics exposed conflicts between a model's quality requirements and the images that could be captured. Head-to-head comparisons showed that commercial products are not interchangeable. None of these findings argues against AI. Each argues for treating AI as a component of a health service rather than as a product to be installed.

Several directions will shape the next phase. Outcome evidence is the most important gap: studies that measure treatment initiation, vision preserved or cancers prevented, rather than agreement with a reference standard. National capacity for local validation and monitoring needs investment, so that thresholds and updates are managed by health systems rather than by vendors. Regulatory arrangements for adaptive software in LMICs are still developing and would benefit from shared benchmarking and mutual reliance. Economic evaluations need to include recurrent costs and whole pathways. And newer tools, including AI-assisted ultrasound, automated cervical screening on mobile devices and language-model assistants for health workers, should be held to the standard that TB CAD has begun to meet: independent evaluation in the populations where they will be used, before they are scaled.

The central lesson is simple to state and hard to practise. An algorithm becomes a health intervention only when data, infrastructure, workforce, regulation and financing are designed together around a clinical pathway, and when someone in the health system is responsible for checking, again and again, that it still works.

REFERENCES

1. Heeks R. Health information systems: failure, success and improvisation. Int J Med Inform. 2006;75(2):125-37. https://doi.org/10.1016/j.ijmedinf.2005.07.024
2. Wu E, Wu K, Daneshjou R, Ouyang D, Ho DE, Zou J. How medical AI devices are evaluated: limitations and recommendations from an analysis of FDA approvals. Nat Med. 2021;27:582-4. https://doi.org/10.1038/s41591-021-01312-x
3. McIsaac M, Buchan J, Abu-Agla A, et al. Global strategy on human resources for health: Workforce 2030, a five-year check-in. Hum Resour Health. 2024;22:68. https://doi.org/10.1186/s12960-024-00940-x
4. Boniol M, Kunjumen T, Nair TS, Siyam A, Campbell J, Diallo K. The global health workforce stock and distribution in 2020 and 2030: a threat to equity and 'universal' health coverage? BMJ Glob Health. 2022;7(6):e009316. https://doi.org/10.1136/bmjgh-2022-009316
5. Qin ZZ, Sander MS, Rai B, Titahong CN, Sudrungrot S, Laah SN, et al. Using artificial intelligence to read chest radiographs for tuberculosis detection: a multi-site evaluation of the diagnostic accuracy of three deep learning systems. Sci Rep. 2019;9:15000. https://doi.org/10.1038/s41598-019-51503-3
6. International Telecommunication Union. Measuring digital development: Facts and Figures 2024. Mobile network coverage. Geneva: ITU; 2024. Available from: https://www.itu.int/itu-d/reports/statistics/2024/11/10/ff24-mobile-network-coverage/
7. World Health Organization. Global strategy on digital health 2020-2025. Geneva: World Health Organization; 2021. Available from: https://www.who.int/docs/default-source/documents/gs4dhdaa2a9f352b0445bafbc79ca799dce4d.pdf
8. Sood S, Lal K, Bhatia M, et al. Adoption and utilization of India's eSanjeevani national telemedicine service. Oxf Open Digit Health. 2025. https://doi.org/10.1093/oodh/oqaf025

9. Glasgow RE, Vogt TM, Boles SM. Evaluating the public health impact of health promotion interventions: the RE-AIM framework. Am J Public Health. 1999;89(9):1322-7. https://doi.org/10.2105/AJPH.89.9.1322
10. Damschroder LJ, Reardon CM, Opra Widerquist MA, Lowery J. The updated Consolidated Framework for Implementation Research based on user feedback. Implement Sci. 2022;17:75. https://doi.org/10.1186/s13012-022-01245-0
11. Yuan A, Lee AY. Artificial intelligence deployment in diabetic retinopathy: the last step of the translation continuum. Lancet Digit Health. 2022;4(4). https://doi.org/10.1016/S2589-7500(22)00027-9
12. World Health Organization. WHO consolidated guidelines on tuberculosis. Module 2: screening. Systematic screening for tuberculosis disease. Geneva: World Health Organization; 2021. ISBN 978-92-4-002267-6. Available from: https://www.who.int/publications/i/item/9789240022676
13. Qin ZZ, Ahmed S, Sarker MS, Paul K, Adel ASS, Naheyan T, et al. Tuberculosis detection from chest x-rays for triaging in a high tuberculosis-burden setting: an evaluation of five artificial intelligence algorithms. Lancet Digit Health. 2021;3(9):e543-54. https://doi.org/10.1016/S2589-7500(21)00116-3
14. Khan FA, Majidulla A, Tavaziva G, Nazish A, Abidi SK, Benedetti A, et al. Chest x-ray analysis with deep learning-based software as a triage test for pulmonary tuberculosis: a prospective study of diagnostic accuracy for culture-confirmed disease. Lancet Digit Health. 2020;2(11):e573-81. https://doi.org/10.1016/S2589-7500(20)30221-1
15. Codlin AJ, Dao TP, Vo LNQ, Forse RJ, Truong VV, Dang HM, et al. Independent evaluation of 12 artificial intelligence solutions for the detection of tuberculosis. Sci Rep. 2021;11(1):23895. https://doi.org/10.1038/s41598-021-03265-0
16. Tavaziva G, Harris M, Abidi SK, Geric C, Breuninger M, Dheda K, et al. Chest X-ray analysis with deep learning-based software as a triage test for pulmonary tuberculosis: an individual patient data meta-analysis of diagnostic accuracy. Clin Infect Dis. 2022;74(8):1390-400. https://doi.org/10.1093/cid/ciab639
17. Qin ZZ, Barrett R, Ahmed S, Sarker MS, Paul K, Adel ASS, et al. Comparing different versions of computer-aided detection products when reading chest X-rays for tuberculosis. PLOS Digit Health. 2022;1(6):e0000067. https://doi.org/10.1371/journal.pdig.0000067
18. Harris M, Qi A, Jeagal L, Torabi N, Menzies D, Korobitsyn A, et al. A systematic review of the diagnostic accuracy of artificial intelligence-based computer programs to analyze chest x-rays for pulmonary tuberculosis. PLoS One. 2019;14(9):e0221339. https://doi.org/10.1371/journal.pone.0221339
19. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA. 2016;316(22):2402-10. https://doi.org/10.1001/jama.2016.17216
20. Gulshan V, Rajan RP, Widner K, Wu D, Wubbels P, Rhodes T, et al. Performance of a deep-learning algorithm vs manual grading for detecting diabetic retinopathy in India. JAMA Ophthalmol. 2019;137(9):987-93. https://doi.org/10.1001/jamaophthalmol.2019.2004

21. Ruamviboonsuk P, Krause J, Chotcomwongse P, Sayres R, Raman R, Widner K, et al. Deep learning versus human graders for classifying diabetic retinopathy severity in a nationwide screening program. NPJ Digit Med. 2019;2. https://doi.org/10.1038/s41746-019-0099-8

22. Ruamviboonsuk P, Tiwari R, Sayres R, Nganthavee V, Hemarat K, Kongprayoon A, et al. Real-time diabetic retinopathy screening by deep learning in a multisite national screening programme: a prospective interventional cohort study. Lancet Digit Health. 2022;4(4):e235-44. https://doi.org/10.1016/S2589-7500(22)00017-6

23. Beede E, Baylor E, Hersch F, Iurchenko A, Wilcox L, Ruamviboonsuk P, et al. A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. New York: Association for Computing Machinery; 2020. p. 1-12. https://doi.org/10.1145/3313831.3376718

24. Natarajan S, Jain A, Krishnan R, Rogye A, Sivaprasad S. Diagnostic accuracy of community-based diabetic retinopathy screening with an offline artificial intelligence system on a smartphone. JAMA Ophthalmol. 2019;137(10):1182. https://doi.org/10.1001/jamaophthalmol.2019.2923

25. Abràmoff MD, Lavin PT, Birch M, Shah N, Folk JC. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. NPJ Digit Med. 2018;1(1):39. https://doi.org/10.1038/s41746-018-0040-6

26. Lee AY, Yanagihara RT, Lee CS, Blazes M, Jung HC, Chee YE, et al. Multicenter, head-to-head, real-world validation study of seven automated artificial intelligence diabetic retinopathy screening systems. Diabetes Care. 2021;44(5):1168-75. https://doi.org/10.2337/dc20-1877

27. Hu L, Bell D, Antani S, Xue Z, Yu K, Horning MP, et al. An observational study of deep learning and automated evaluation of cervical images for cancer screening. J Natl Cancer Inst. 2019. https://doi.org/10.1093/jnci/djy225

28. World Health Organization. WHO guideline: recommendations on digital interventions for health system strengthening. Geneva: World Health Organization; 2019. Available from: https://www.who.int/publications/i/item/9789241550505

29. Musabyimana A, Ruton H, Gaju E, Berhe A, Grepin KA, Ngenzi J, et al. Assessing the perspectives of users and beneficiaries of a community health worker mHealth tracking system for mothers and children in Rwanda. PLoS One. 2018;13(6):e0198725. https://doi.org/10.1371/journal.pone.0198725

30. Barron P, Peter J, LeFevre AE, Sebidi J, Bekker M, Allen R, et al. Mobile health messaging service and helpdesk for South African mothers (MomConnect): history, successes and challenges. BMJ Glob Health. 2018;3(Suppl 2):e000559. https://doi.org/10.1136/bmjgh-2017-000559

31. Han S, Mao H, Dally WJ. Deep compression: compressing deep neural networks with pruning, trained quantization and Huffman coding. In: International Conference on Learning Representations (ICLR); 2016. arXiv:1510.00149. Available from: https://arxiv.org/abs/1510.00149

32. Press Information Bureau, Government of India. Ayushman Bharat Digital Mission: India's digital health backbone [backgrounder]. New Delhi: PIB; 6 July 2026. Available from: https://static.pib.gov.in/WriteReadData/specificdocs/documents/2026/jul/doc202676912801.pdf

33. Government of India. National mission on natural language translation (BHASHINI). India Science, Technology and Innovation portal. Available from: https://www.indiascienceandtechnology.gov.in/st-visions/national-mission/national-mission-natural-language-translation-bhashini [accessed 3 November 2026]

34. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. PLoS Med. 2018;15(11):e1002683. https://doi.org/10.1371/journal.pmed.1002683

35. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science. 2019;366(6464):447. https://doi.org/10.1126/science.aax2342

36. International Medical Device Regulators Forum, Software as a Medical Device Working Group. Software as a Medical Device (SaMD): clinical evaluation. IMDRF; 2017. Available from: https://www.imdrf.org/documents/software-medical-device-samd-clinical-evaluation

37. Central Drugs Standard Control Organization, Medical Devices and IVD Division, Ministry of Health and Family Welfare. Guidance document on medical device software under the Medical Devices Rules, 2017. Document no. CDSCO/MD/GD/MDSW/01/2026. New Delhi: CDSCO; 2026. Available from: https://www.cdsco.gov.in/opencms/export/sites/CDSCO_WEB/Pdf-documents/Guidance-document-on-Medical-Device-Software-under-MDR-2017.pdf

38. African Union Commission. Frequently asked questions (FAQs) on the African Medicines Agency (AMA). Addis Ababa: African Union; 7 December 2021. Available from: https://au.int/en/documents/20211207/frequently-asked-questions-faqs-african-medicines-agency-ama

39. World Health Organization. Ethics and governance of artificial intelligence for health: WHO guidance. Geneva: World Health Organization; 2021. ISBN 978-92-4-002920-0. Available from: https://www.who.int/publications/i/item/9789240029200

40. Wiegand T, Krishnamurthy R, Kuglitsch M, Lee N, Pujari S, Salathé M, et al. WHO and ITU establish benchmarking process for artificial intelligence in health. Lancet. 2019;394(10192):9-11.

41. Fraunhofer Heinrich Hertz Institute. Artificial intelligence for health: establishment of a benchmarking process for AI algorithms [news release]. Berlin: Fraunhofer HHI; 8 April 2019. Available from: https://www.hhi.fraunhofer.de/en/news/nachrichten/https-testhhifraunhoferde-presse-medien-nachrichten-archiv-2019html/artificial-intelligence-for-health-establishment-of-a-benchmarking-process-for-ai-algorithms.html