

# **Reckoning with Medicine, Public Health, and Health Systems in the Age of AGI**

Clinical Evidence, Patient Safety, and the Boundaries of Population Health

Reckoning with Medicine, Public Health, and Health Systems in the Age of AGI

Clinical Evidence, Patient Safety, and the Boundaries of Population Health

Medicine / Public Health / Health Systems: Standalone Professional In-Depth Paper for External Release V1.1

Author: Zijun Fu | Produced by: Civilization Leap Research Group

Evidence cutoff: August 26, 2026

## Abstract

Once artificial intelligence enters health care, an especially common cross-level inference is that a model's greater accuracy in testing means patients will achieve better health outcomes. Medicine, clinical epidemiology, evidence-based medicine, patient safety, public health, and health-systems research show that an entire real-world chain still intervenes: whether study participants resemble actual patients; whether risk estimates are calibrated; whether diagnostic results change clinical decisions; whether treatment actually improves patient outcomes; whether errors can be detected and control taken over; whether minority groups bear disproportionate harm; whether deployment changes workflows, referrals, and resource use; whether evidence remains valid after system updates; and whether local benefits at the individual level retain a net benefit at health-system and population scales. This paper therefore begins its professional reassessment with the proposition that “model accuracy may belong to AI evaluation; clinical net benefit, patient safety, and population health do not thereby automatically belong to AI.”

This paper is a Perspective / Conceptual Review, not a clinical guideline, systematic review, regulatory opinion on drugs or medical devices, or advice on diagnosing or treating particular patients. It carries forward established evidentiary disciplines in medicine and public health: diagnostic research must distinguish sensitivity, specificity, calibration, external validity, and clinical consequences; prediction models require transparent reporting and assessment of risk of bias and applicability; AI entering real workflows requires early clinical evaluation, prospective trials, real-world evidence, ongoing monitoring, and version revalidation; patient safety requires locating errors within the full system chain; and public health requires further evidence translating accurate individual diagnosis into population surveillance, intervention effectiveness, access, equity, and health-system resilience.[1–16] These standards and methods do not establish special exclusion zones for AI. They prevent technical performance from bypassing evidentiary bridges medicine already requires.

By 2026, real-world evidence had supplied important pressures in both directions. External validation of the Epic sepsis model at a large academic medical center found a hospitalization-level AUC of only 0.63 and substantial alert burden, demonstrating that performance in development settings and widespread deployment cannot replace independent external validation.[17] In a randomized trial involving 50 physicians, access to GPT-4 assistance did not significantly improve physicians' diagnostic-reasoning scores on simulated complex cases; in an exploratory secondary analysis, the model alone scored higher on the same task, showing that a stronger model does not imply a naturally stronger human–AI team.[18] Conversely, Sweden's MASAI randomized screening trial randomized 105,934 women (19 subsequently excluded, leaving 105,915 in the analysis) and found that AI-supported breast screening increased cancer detection, reduced reading workload by approximately 44%, and did not significantly increase the false-positive rate; interval-cancer results reported subsequently in 2026 met the noninferiority criterion.[19–20] Germany's PRAIM was a prospective, multicenter, real-world, observational noninferiority implementation study covering 12 screening sites, more than 460,000 women, and 119 radiologists. The AI-supported group had a higher detection rate without a worsened recall rate, but physicians voluntarily chose whether to use AI, so selection bias and implementation differences still limit causal extrapolation.[21] Together, these findings show that medical AI cannot be closed off by the pessimistic narrative of “excellent in the laboratory, useless in practice,” or prematurely credited through the optimistic narrative of “higher accuracy automatically establishes clinical value.”

The paper further situates medical AI within public health and health systems. The 2025 global UHC monitoring report estimated that approximately 4.6 billion people still lacked adequate access to essential health services in 2023, and approximately 2.1 billion experienced financial hardship due to out-of-pocket health spending in 2022; WHO also revised its projected global shortage of health and care workers in 2030 upward to approximately 11.1 million.[25–26] AI's opportunity to reduce professional scarcity and improve access is therefore very real. Yet that opportunity becomes population health only when infrastructure, language, digital capabilities, referral pathways, treatment supply, payment, patient trust, and health-worker capabilities are jointly in place. In public health, WHO's EIOS 2.0 is used in more than 110 Member States and approximately 30 organizations and networks, and incorporates AI functions; in 2026 WHO also publicly demonstrated AI-supported community listening for cholera response.[27–28] These are facts of real-world adoption, but they do not automatically demonstrate that faster surveillance has produced lower morbidity or mortality, or more legitimate public-health interventions.

The paper ultimately argues that when technology enters health care, the first question is not how intelligent the system is, but whether patients consequently achieve better real outcomes. When it enters public health, the questions must extend to whether population health improves, who remains excluded, and whether the health system becomes more resilient. Health is an important condition of civilized life, but it is not the entirety of life's meaning. Medicine can establish which interventions are safer and more effective; it cannot define a person's complete worth through disease risk, life expectancy, or health scores. Nor does more accurate prediction give AI authority over treatment, public power, or meaning. Medicine's professional contribution is not to endorse a direction for civilization. It is to establish, at the point closest to consequences for life, "what genuinely helps, whom it helps, and under what conditions," and, where its evidence ends, hand questions over for renewed judgment by law, ethics, economics, politics, social inquiry, and deliberation on civilization's direction.

Keywords: AGI era; medical artificial intelligence; evidence-based medicine; clinical effectiveness; patient safety; real-world evidence; public health; health equity; health-system resilience

## Abstract

Artificial intelligence in health care creates a recurrent inferential error: a model performs better on a technical benchmark, therefore patients must experience better health outcomes. Medicine, clinical epidemiology, evidence-based practice, patient safety, public health, and health-systems research show why this inference does not hold without additional evidence. Between model performance and patient benefit lie calibration, external validity, workflow integration, human-AI interaction, clinical decisions, treatments, adverse events, subgroup effects, resource use, follow-up, system resilience, and population-level consequences.

This paper is a Perspective/Conceptual Review rather than a clinical guideline, systematic review, regulatory opinion, or medical advice. It distinguishes established or verifiable facts and formal materials (A), unsettled mechanisms and competing explanations (B), cross-domain structural hypotheses (C), and explicit normative commitments (F). It argues that diagnostic accuracy, regulatory authorization, or model reliability cannot by themselves establish clinical net benefit in every deployment context; public-health surveillance performance cannot by itself establish intervention effectiveness; and average performance cannot erase subgroup harm. At the same time, emerging randomized and real-world evidence in screening demonstrates that well-designed AI-supported workflows can improve clinically relevant process and screening outcomes in defined settings.

The central thesis is: model accuracy may belong to AI evaluation, but clinical net benefit, patient safety, and population health require independent medical and public-health evidence. The appropriate endpoint of medical AI is therefore not "a smarter model" but a defensible chain from technical performance to real patient outcomes, equitable population benefit, and a resilient health system. Health is a major condition of a good civilization, but it is not a total score of human worth or meaning.

Keywords: AGI era; medical AI; evidence-based medicine; clinical effectiveness; patient safety; real-world evidence; public health; health equity; health-system resilience

## Research Positioning, Scope, and Argumentative Discipline

This is a professional in-depth paper on medicine, public health, and health systems. It independently assumes responsibility for disciplinary facts concerning clinical effectiveness, clinical net benefit, diagnostic and treatment safety, external validity, real-world evidence, patient heterogeneity, interfaces of clinical responsibility, epidemiology and public-health surveillance, intervention effectiveness, health equity, health resources, and health-system resilience. It is not an expansion of the life-sciences paper: life sciences may study mechanisms of life and disease, individuation, and continuity of life, whereas clinical medicine must determine whether a diagnosis, intervention, or workflow actually benefits patients, and public health must determine whether an intervention improves health in real populations. The original life-sciences paper explicitly states that it is not a clinical guideline and that biological facts must not automatically generate rights, personhood, or political standing. This is one stopping point that establishes the present paper's independent remit.

Nor does this paper decide model architectures, benchmark performance, formal guarantees, or system security on behalf of AI / computer science; legal liability, informed consent, privacy, or the legal force of rights on

behalf of law and ethics; prices, payment, or resource allocation on behalf of economics; public-health powers or emergency authority on behalf of political science; or civilization's ultimate purposes on behalf of philosophy or deliberation on civilization's direction. Medicine is responsible for establishing health consequences. Once a question concerns rights, coercion, distribution, and ultimate values, another handoff is required.

**Table 1. A / B / C / F Argumentative Status and Medical Boundaries**

| Status                                          | Typical content in this paper                                                                                                                                 | What it can support                                                                                                        | What it cannot automatically establish                                                            |
|-------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------|
| A   Relatively stable facts / formal materials  | Clinical trials, external validation, real-world studies, epidemiological data, patient-safety reports, formal guidelines, and regulatory materials           | Factual judgments bounded by patients, settings, time, versions, and endpoints                                             | Net benefit for all populations, legal liability, or civilization's direction                     |
| B   Frontier questions / competing explanations | Long-term human-AI collaboration, changes in skills, continuously learning models, patient-facing LLMs, and organizational adaptation                         | Comparing competing mechanisms and developing further research                                                             | Established, universal medical conclusions                                                        |
| C   Cross-domain structural judgments           | Model → workflow → behavior → clinical outcomes → resources → trust / access → population health; shared models → correlated errors → systemic risk           | Refutable cross-domain hypotheses and stress tests                                                                         | An entire civilizational chain presented as though already established by a single clinical trial |
| F   Normative commitments                       | Accuracy must not replace patient safety; people's basic standing is not determined by disease risk or health scores; health is not an overall score for life | Explicitly explaining why this paper requires real outcomes, attention to heterogeneity, and the possibility of opting out | Substitutes for clinical facts, existing law, public authorization, or individual value choices   |

## The Age of AGI Does Not Await a Fully Realized AGI

"The age of AGI" describes the actual conditions in which advanced AI, generative AI, and agentic systems have begun entering medical knowledge, documentation, diagnostic support, imaging, screening, patient communication, public-health intelligence, and health-system management. As of 2026, the FDA continues to update its list of AI-enabled medical devices. Devices on the list have met the relevant premarket requirements through their respective applicable pathways, while the FDA explicitly states that the list is not comprehensive.[5] WHO likewise describes actual AI applications in diagnosis, clinical care, drug development, disease surveillance, outbreak response, and health-system management.[1–2,29] These conditions suffice for the present paper; future AGI, superintelligence, or machine subjectivity need not be assumed.

## Contents

Introduction | Medicine Truly Begins Once a Model Crosses the Consulting-Room Threshold

1 | Established Disciplinary Inheritance: What Medicine, Public Health, and Health Systems Already Know

2 | Objects of Study: From Model Outputs to Consequences for Real Patients and Populations

3 | Evidence Cannot Skip Levels: From Technical Performance to Clinical Net Benefit

4 | Diagnosis: Why Accuracy Is Not Yet Diagnostic Safety

5 | Treatment and Clinical Decisions: Correct Advice Does Not Necessarily Benefit Patients

6 | Human–AI Collaboration: Why a Stronger Model Does Not Necessarily Make a Stronger Team

7 | Positive Evidence: Under What Conditions Has AI Already Improved Real Clinical Workflows?

8 | The Real World: External Validity, Drift, Versions, and Continuous Monitoring

9 | Patient Heterogeneity and Health Equity: Average Performance Cannot Obscure Who Is Harmed

10 | Generative AI and Direct Patient Use: New Routes to Health Information and New Evidence Gaps

11 | Public Health: Earlier Detection Does Not Establish Effective Intervention

12 | Access and Health Resources: Can Reduced Professional Scarcity Truly Become Broadly Shared Benefit?

13 | Health-System Resilience: Shared Models, Correlated Errors, and Dependence on Digital Infrastructure

14 | Interfaces of Clinical Responsibility: Who Must Reenter, and at Which Points?

15 | Medicine and Civilization's Direction: Health Matters, but Is Not an Overall Score for Life

16 | Interdisciplinary Handoffs: Who Reenters When Evidence Stops?

17 | What Facts Could Change This Paper? The Strongest Objections and a Refutable Research Agenda

Conclusion | The True Endpoint Is Better Real Health Outcomes, Not a Smarter Model

Further Reading and Entry Points for Shared Research

Appendix A | Register of Source–Claim Boundaries

Appendix B | Checklist for Evidentiary Transitions in Medical AI

Statement on the Use of Generative Artificial Intelligence

References

Version Note

## Introduction | Medicine Truly Begins Once a Model Crosses the Consulting-Room Threshold

The most dangerous misunderstanding between AI and medicine is not portraying an ordinarily performing model as an “all-purpose doctor.” It is moving directly from a genuinely excellent model on a bounded task to the conclusion that real patients will necessarily benefit. The former is easy to recognize; the latter is harder because its first step may be entirely correct: models can indeed achieve high performance in certain imaging, screening, prediction, or knowledge tasks. The real break occurs at the next step: whose workflow the model enters, which judgments it changes, which tests or treatments it triggers, whom it misses, how many additional people it recalls, which resources it consumes, whether errors are detected, whether patients can refuse it, and whether the evidence still holds after system updates.

Medicine already has a complete set of methods for addressing this distance from metrics to consequences. WHO's 2023 regulatory considerations for AI in health place risk–benefit assessment, intended use, external validation, data quality, continuous learning, human intervention, and lifecycle monitoring within one framework. Its guidance on large models for health further notes that models may generate inaccurate, incomplete, or false information and create risks involving automation bias, deskilling, privacy, and equity.[1–2] This is not special distrust of AI. Medicine already disallows substituting an intermediate metric for patient outcomes.

Patient safety makes the point more directly. WHO's Global Patient Safety Report 2024 estimates that diagnostic errors account for approximately 16% of preventable patient harm worldwide and notes that history

taking, examination, test selection and interpretation, communication, and other stages may all contribute to diagnostic harm.[4] AI entering these stages may help detect errors or introduce new sources of error. The issue cannot be reduced to a model's overall accuracy, because medical harm occurs along an entire chain of action.

*Model accuracy may belong to AI evaluation; clinical net benefit, patient safety, and population health do not thereby automatically belong to AI.*

This is also why the paper must stand independently of life sciences, AI science, and ethics and law. Life sciences tell us about mechanisms of disease and life; AI science tells us what systems can do; ethics and law tell us which rights and responsibilities must be respected. Medicine and public health still need independently to establish whether real patients and populations become better or worse off as a result.

## **1 | Established Disciplinary Inheritance: What Medicine, Public Health, and Health Systems Already Know**

### **1.1 Evidence-Based Medicine: Evidence Centers on Clinical Questions and Patient Outcomes, Not Technical Novelty**

Medical evidence has never ranked higher simply for being newer. Randomized trials, diagnostic-accuracy studies, cohorts, case-control studies, real-world data, mechanistic research, experience, and patient preferences answer different questions. AI's involvement does not invalidate this basic structure. SPIRIT-AI and CONSORT-AI require trials of AI interventions to specify versions, inputs and outputs, human-AI interaction, error analyses, and implementation settings; DECIDE-AI specifically addresses human factors, workflows, and generalizability in the transition from offline performance to early real clinical use. These are reporting guidelines, not guarantees that study findings are valid.[10–13]

Topol early identified complementarity between AI and clinical professionals as a potential direction for “high-performance medicine.” Kelly and colleagues, focusing on the conditions for achieving clinical impact, emphasized barriers involving robust evaluation, generalizability, workflow integration, regulation, and ongoing monitoring.[33–34] This is the disciplinary inheritance the paper takes forward: AI's clinical value must be established within real health-care systems, not advanced on credit against technical potential or a single benchmark result.

### **1.2 Clinical Epidemiology: Distinguishing Discrimination, Calibration, External Validity, and Decision Consequences**

TRIPOD+AI requires transparent reporting of clinical prediction-model development and evaluation. PROBAST+AI separates model development from model evaluation and assesses quality, bias, and applicability across participants / data, predictors, outcomes, and analysis. STARD-AI adds reporting requirements for AI diagnostic-accuracy studies concerning dataset practices, AI index tests, algorithmic bias, fairness, and generalization.[8–10] Their shared function is not to produce an “overall evidence score for medical AI,” but to show readers precisely where a performance finding holds.

High discrimination does not mean reliable risk estimation. Van Calster and colleagues stress that prediction models can be seriously miscalibrated despite good discrimination, thereby misleading clinical decisions; external validation and model updating when necessary are therefore indispensable.[15] Decision curve analysis, developed by Vickers and Elkin, further incorporates thresholds and clinical consequences into model assessment, reminding us that accuracy metrics do not directly answer whether using a model is worthwhile.[16] “Clinical net benefit” has a broader meaning in this paper: overall clinical outcomes after patient-relevant benefits are weighed against harms, erroneous interventions, opportunity costs, and system burdens. It is not identical to the technical net-benefit measure in decision curve analysis.

### **1.3 Patient Safety: Errors Are System Events, Not Merely Component Events**

WHO's Global Patient Safety Action Plan 2021–2030 incorporates safety culture, organizational reliability, clinical-process safety, patient and family engagement, health-worker education, information and research, and collaboration into systemic action.[3] Bates and colleagues' scoping review of AI and patient safety likewise notes that AI may reduce diagnostic, medication, and monitoring errors, but can also cause harm through erroneous

recommendations, automation bias, integration failures, and new workflow risks.[31] Patient safety therefore asks not only whether a model can err, but how errors are prevented, detected, amplified, or remedied.

1.4 Public Health: The Object of Study Expands from Patients to Populations

Public health addresses morbidity, mortality, exposure, transmission, prevention, screening, surveillance, intervention, equity, and determinants of health. An algorithm may improve an individual prediction, but public health must still ask whether the target population is actually covered; how false positives and false negatives are distributed; whether interventions are delivered; whether resources suffice; how behavior and trust change; and whether morbidity and mortality ultimately improve. Earlier signal detection by a surveillance system is an important intermediate outcome, not proof of an effective public-health intervention.

1.5 Health Systems: Service Access, Workforce, Infrastructure, and Resilience Determine Whether Technology Becomes Health

WHO’s 2025 UHC monitoring indicates that approximately 4.6 billion people worldwide still lacked adequate access to essential health services in 2023, while approximately 2.1 billion experienced financial hardship from health spending in 2022.[25] WHO Executive Board materials on the health workforce in 2025 revised the projected global shortage of health and care workers in 2030 upward to approximately 11.1 million.[26] These figures are not evidence that AI will necessarily solve the problems. They explain why reducing professional scarcity, documentation burden, and barriers to knowledge has enormous potential value. If technology only improves local efficiency at well-resourced institutions without expanding actual access to services, the civilizational dividend at health-system level has not yet materialized.

2 | Objects of Study: From Model Outputs to Consequences for Real Patients and Populations

“Medical AI” is not a single object. Model weights, an FDA-authorized medical device, a hospital-deployed service, a physician workflow, a chatbot used directly by patients, and a national surveillance platform are entirely different evidentiary objects. The FDA’s list of AI-enabled medical devices establishes only that specified devices obtained US market authorization through the relevant pathways; it does not establish clinical net benefit across all deployment environments or patient groups.[5] Likewise, a general-purpose large model not regulated as a medical device can influence patient behavior without thereby acquiring the same evidentiary status as a medical device.

Table 2. Six Levels of Medical AI and the Stopping Points of Evidence

| Object level                | What is actually studied                                                    | Principal evidence                                                     | What does not automatically follow                    |
|-----------------------------|-----------------------------------------------------------------------------|------------------------------------------------------------------------|-------------------------------------------------------|
| Model / algorithm           | Prediction, classification, or generation using defined inputs and datasets | Performance, calibration, robustness, and bias                         | Real clinical net benefit or authority to treat       |
| AI medical product / device | A product with an intended use, version, and functions                      | Regulatory documentation, validation, and risk management              | Effectiveness across all institutions and populations |
| Deployed service            | End-to-end configuration in a particular hospital or platform               | Local validation, service monitoring, failures, and versions           | Inherently safe organizational workflows              |
| Clinical workflow           | A decision chain formed jointly by physicians, nurses, patients, and AI     | Human factors, takeover, time, training, and responsibility interfaces | Final adjudication of legal liability and rights      |
| Patient pathway             | Diagnosis, investigations, treatment, follow-up, and actual outcomes        | Clinical trials, cohorts, and real-world evidence                      | Automatic effectiveness at population scale           |

| Object level               | What is actually studied                                       | Principal evidence                                                                         | What does not automatically follow         |
|----------------------------|----------------------------------------------------------------|--------------------------------------------------------------------------------------------|--------------------------------------------|
| Health system / population | Resources, access, equity, surveillance, resilience, and trust | Epidemiology, implementation science, health-services research, and public-health research | Civilization's ultimate ordering of values |

## 2.1 The Complete Chain of Action

A minimal clinical action chain includes at least the following: who frames the clinical question; which histories, signs, images, and laboratory tests are collected; what inputs AI receives; whether its outputs are probabilities, classifications, textual recommendations, or rankings; who sees them; whether diagnosis, investigations, prescriptions, referrals, or treatment change; whether patients understand AI's involvement; who can override the system; how errors are detected; whether actual treatment is delivered; what patient outcomes occur; and whether new results feed back into future models. A change at any link may change the clinical effects of “the same model.”

## 2.2 Regulatory Authorization, Technical Performance, Clinical Benefit, and Public-Health Effectiveness Require Separate Accounts

FDA authorization can establish that a particular product meets applicable premarket regulatory requirements. Technical evaluations can establish a particular version's performance on particular data. Clinical research can establish the effects of a deployment on patient outcomes or workflows. Public-health research goes on to examine population coverage, resources, health equity, and population outcomes. These four support one another, but no automatic conversion exists between them. Conflating them is one of the most common forms of evidentiary overreach in medical AI.

## 3 | Evidence Cannot Skip Levels: From Technical Performance to Clinical Net Benefit

Medical AI needs a causal chain, not an “evidence pyramid” that automatically grades models. Different technologies and uses may require different study designs, but at minimum the questions are whether a model reliably predicts the right outcomes for the right people using the right inputs; whether those predictions actually change clinical behavior; whether the changes improve patient-relevant outcomes; whether deployment errors and burdens offset benefits; and whether effects persist across institutions, groups, and time.

**Table 3. Evidentiary Transitions from Model Performance to Clinical and Population Outcomes**

| Evidence stage                       | Core question                                                           | Typical evidence                                                      | Stopping point                                                              |
|--------------------------------------|-------------------------------------------------------------------------|-----------------------------------------------------------------------|-----------------------------------------------------------------------------|
| Offline / retrospective performance  | Does the model discriminate or predict on defined data?                 | AUC, sensitivity / specificity, error, and calibration                | Does not establish effectiveness in clinical use                            |
| External validation                  | Is performance maintained across institutions, time periods, or groups? | Independent cohorts, calibration, and subgroup performance            | Does not establish workflow or human-factors safety                         |
| Early real-world clinical evaluation | How is the system used in actual workflows?                             | DECIDE-AI-type studies, human factors, takeover, and errors           | Cannot replace large-scale comparative outcomes                             |
| Comparative clinical research        | Is an AI-supported workflow better than standard care?                  | Randomized trials, quasi-experiments, and patient / clinical outcomes | Does not automatically establish long-term or cross-regional generalization |

| Evidence stage       | Core question                                                                | Typical evidence                                      | Stopping point                                            |
|----------------------|------------------------------------------------------------------------------|-------------------------------------------------------|-----------------------------------------------------------|
| Real-world evidence  | Effects, rare risks, and heterogeneity following deployment at scale         | EHRs, registries, claims, device data, and monitoring | Observational associations require control of confounding |
| Lifecycle monitoring | Does the evidence remain valid after changes in versions, data, or practice? | Drift, incidents, recalibration, and version review   | Past effectiveness cannot confer permanent authorization  |

The FDA's medical-device real-world evidence guidance updated in 2025 states that its central purpose is to assess whether real-world data are of sufficient quality to generate RWE usable in regulatory decisions; it does not alter the evidentiary standards applicable to the products themselves.[6] Real-world evidence is therefore not a substitute that makes poor-quality data acceptable, nor does data volume permit bypassing causality and bias. NICE's evidence standards framework for digital health technologies likewise separates technical performance, economic impact, deployment, and lifecycle requirements, and incorporated adaptive AI and data-driven technologies in 2022.[30]

*The real evidentiary advance in medical AI is not a larger model: the closer a claim comes to consequences for actual patients, the more the relevant evidence must independently enter the assessment.*

## 4 | Diagnosis: Why Accuracy Is Not Yet Diagnostic Safety

### 4.1 Diagnosis Is Not a Single Classification

The US National Academies' 2015 *Improving Diagnosis in Health Care* describes diagnosis as a process unfolding over time that involves collaboration between patients and professionals, information gathering, and interpretation.[32] WHO's 2024 patient-safety report similarly includes history taking, physical examination, test selection and interpretation, communication, and follow-up within pathways to diagnostic error.[4] Thus, even accurate identification of an image does not necessarily complete a diagnosis; even a correct differential diagnosis does not ensure timely testing and treatment for the patient.

### 4.2 Failed External Validation Can Make an Already Deployed Model an Object of Research Again

Wong and colleagues independently externally validated the Epic Sepsis Model in 27,697 patients and 38,455 hospitalizations, finding a hospitalization-level AUC of 0.63 and noting low sensitivity and substantial alert burden at the threshold studied. Conducted at one academic medical center, the study cannot be extrapolated to all hospitals, but it demonstrates that developer-reported performance plus widespread deployment cannot replace independent local validation.[17] Differences in subsequent results across hospitals further show that setting, baseline prevalence, workflows, and implementation can change a model's value.

### 4.3 Discrimination, Calibration, Thresholds, and Consequences Must Be Examined Together

A risk model may rank high-risk patients first while systematically overestimating or underestimating actual risk. A diagnostic model may increase sensitivity while generating many unnecessary investigations. A screening system may detect more lesions primarily by identifying lesions of limited clinical significance. STARD-AI, TRIPOD+AI, PROBAST+AI, calibration research, and decision curve analysis address these problems through reporting, bias / applicability, reliability of risk estimates, and decision consequences, respectively.[8–10,15–16] They are not interchangeable.

## 5 | Treatment and Clinical Decisions: Correct Advice Does Not Necessarily Benefit Patients

Moving from diagnosis to treatment raises the evidentiary threshold further. AI may recommend drugs, dosages, tests, surgery, referrals, or follow-up, but recommendations themselves are not treatment effects. The actual benefits of medicines and medical interventions depend on patient characteristics, contraindications, comorbidities, interactions, adherence, medical resources, and subsequent monitoring. The more fluently a model

restates guidelines, the more easily it creates an illusion that the answer already exists. Clinical decisions, however, often arise precisely where evidence is incomplete, patient preferences conflict, and risks require balancing.

SPIRIT-AI and CONSORT-AI require AI-intervention trials to describe inputs, outputs, interaction, errors, and implementation environments, and DECIDE-AI emphasizes real workflows and human factors, fundamentally because AI is not a pill detached from the clinical system but part of a complex intervention.[10–13] The 2025 FUTURE-AI international consensus translates principles including fairness, universality, traceability, usability, robustness, and explainability into lifecycle practice recommendations. It is a multistakeholder consensus guideline, not clinical data establishing a particular AI product's effectiveness.[14]

## **5.1 Clinical Net Benefit Is Not a Single Metric**

This paper does not construct a new “overall clinical net-benefit score.” Clinical net benefit is a structure of professional questions: do patient-relevant benefits actually increase relative to standard care; are additional harms, overdiagnosis, missed diagnoses, unnecessary treatment, delays, alert fatigue, and opportunity costs acceptable; do different patient groups experience similar or explicably different outcomes; and can benefits persist? Individual studies should still select appropriate patient outcomes and analytical methods.

## **5.2 Authority for High-Risk Action Must Be Separated from the Scope of Evidence**

The AI / computer-science paper's “medical agent” stress test already states that diagnostic performance does not automatically confer authority to prescribe, cancel appointments, make insurance decisions, or allocate resources. The present paper adds the medical perspective: even if AI recommendations have established clinical value, law, ethics, and institutional arrangements must reenter to determine how treatment authority is assigned, how patients consent, and who bears legal liability. Medicine can only specify which professional review and clinical takeover are necessary for safety; it cannot substitute professional authority for public power.

# **6 | Human-AI Collaboration: Why a Stronger Model Does Not Necessarily Make a Stronger Team**

Real medical-AI deployment often takes the form of humans plus AI, so comparing models with physicians in isolation may miss the most important system under study. Goh and colleagues' 2024 randomized clinical trial enrolled 50 physicians in internal medicine, family medicine, and emergency medicine and compared “physicians + GPT-4 + conventional resources” with “physicians + conventional resources” on six complex clinical cases. Diagnostic-reasoning scores did not differ significantly (76% versus 74%; adjusted difference, 2 percentage points; 95% CI, –4 to 8;  $P=.60$ ). In an exploratory secondary analysis, the model alone scored 16 percentage points higher than the conventional-resources group on the same cases (95% CI, 2 to 30;  $P=.03$ ).[18] This latter comparison was not the primary randomized comparison and cannot carry the same evidentiary weight as the primary outcome. Moreover, the study used simulated cases and had a small sample, so it cannot establish that LLMs provide no benefit in real clinical practice. What it does support is that connecting physicians to a stronger model does not automatically translate the model's full capabilities into human-AI team capabilities.

## **6.1 Human-AI Collaboration Requires at Least Four Types of Evidence**

First, can professionals understand when to trust and when to question? Second, do AI outputs change decisions in the right direction? Third, do workflows give people sufficient time, authority, and alternative pathways? Fourth, does long-term use alter human skills, attention, and responsibility? WHO's LMM guidance explicitly identifies automation bias, deskilling, misinformation, and changes in interpersonal relationships as risks. [2] Many of these remain category B questions. One cannot declare in advance that AI necessarily deskills physicians, nor infer from short-term efficiency that long-term professional capabilities remain unaffected.

## **6.2 Documentation Automation: Reduced Burden Has Real Value but Is Not Evidence of Patient Outcomes**

Duggan and colleagues' 2025 study of early clinical use of ambient AI documentation tools observed reductions in documentation time and after-hours charting burden. Olson and colleagues' 2025 multi-institutional before-and-after study also reported a reduced proportion of clinicians experiencing burnout after using ambient AI scribes.[35]

–36] These studies provide positive real-world evidence that AI can free clinical time, but their principal endpoints concern documentation burden, time, and professionals' experience, and some designs were nonrandomized. They do not directly establish greater diagnostic accuracy, lower patient mortality, or inevitable improvements in clinician–patient relationships.

## **7 | Positive Evidence: Under What Conditions Has AI Already Improved Real Clinical Workflows?**

### **7.1 MASAI: From Randomized Screening to Interval-Cancer Outcomes**

Medical AI does have genuinely positive evidence. Sweden's MASAI randomized screening study randomized 105,934 women to an AI-supported workflow or standard double reading; 19 were subsequently excluded, leaving 105,915 in the analysis. In the 2025 report, cancer detection rates were 6.4 per 1,000 in the AI group and 5.0 per 1,000 in controls; false-positive rates did not significantly increase, while reading workload fell by 44.2%.<sup>[19]</sup> The subsequent interval-cancer study published in 2026 reported interval-cancer rates of 1.55 per 1,000 in the AI group and 1.76 per 1,000 in controls, meeting the prespecified noninferiority criterion.<sup>[20]</sup> This advances the evidence beyond increased detection toward an important downstream outcome for screening safety.

Even so, MASAI does not establish equivalent results across every country, device, screening age, or AI-only reading arrangement. Nor can screening accuracy alone resolve mortality, long-term overdiagnosis, cost-effectiveness, or equity across populations. A high-quality positive trial does not close the question; it defines more precisely which facts have been established.

### **7.2 PRAIM: Large-Scale Real-World Implementation Provides Another Kind of Evidence**

Germany's PRAIM was a prospective, multicenter, real-world, observational noninferiority implementation study embedded in the German population-based breast-screening program. Its abstract reports that 463,094 women underwent screening during the study period; the analytical sample described as the main study population in the results comprised 461,818 women and 119 radiologists across 12 screening sites. Cancer detection rates were 6.7 per 1,000 in the AI group and 5.7 per 1,000 in controls, without a worsened recall rate in the AI group.<sup>[21]</sup> Its importance lies in scale, real clinical settings, and multicenter implementation. However, physicians voluntarily chose whether to use AI, and allocation was not randomized, so selection bias and implementation differences must remain explicit. Randomized trials and real-world studies do not replace each other; they supply complementary evidence on causal effects and implementation beyond the trial setting.

### **7.3 Colonoscopy: Higher Detection Still Requires Asking What Is Being Detected**

Makar and colleagues' 2025 meta-analysis of 28 randomized trials involving 23,861 participants found that AI-assisted colonoscopy increased adenoma detection and reduced adenoma miss rates. However, additional detection primarily involved diminutive lesions; advanced-adenoma detection did not differ significantly, while non-neoplastic resection rates increased.<sup>[37]</sup> This illustrates an important medical discipline: improved intermediate outcomes represent genuine progress, but the closer one moves toward long-term patient benefit, the more additional evidence is required. “Detecting more” cannot automatically mean “better overall health.”

## **8 | The Real World: External Validity, Drift, Versions, and Continuous Monitoring**

Health-care systems change: disease patterns, epidemics, testing methods, clinical guidelines, population structures, equipment, EHR fields, referral logic, and physician behavior all evolve. AI systems also change: models, data, prompts, retrieval sources, thresholds, interfaces, and suppliers may be updated. Validation that held in 2025 does not automatically cover a new version in 2027.

### **8.1 Real-World Evidence Is Not Merely Watching What Happens After Launch**

The FDA's 2025 final RWE guidance treats the relevance and reliability of real-world data as central to whether they can generate evidence suitable for regulatory use, and explicitly recognizes sources including EHRs, registries,

claims, devices, and patient-generated data.[6] Researchers must therefore continue addressing missingness, measurement error, selection, confounding, outcome definitions, and traceability. Large-scale data do not eliminate bias; they can make its amplification more consistent.

## **8.2 Version Changes Must Trigger Reassessment of the Evidentiary Object**

For updatable models, change is not an abstract “performance upgrade”; it alters the object to which evidence applies. A change in input data, model, threshold, output format, alert logic, workflow, or intended population may require revalidation. FDA regulatory practice for AI devices and clinical decision support already treats intended use, transparency, software functions, and the lifecycle as important boundaries.[5,7] This paper does not present US regulatory pathways as a globally uniform standard. It uses them as a verifiable example of how established regulatory practice separates software functions, medical uses, and evidentiary status.

## **8.3 Real-World Monitoring Must Make Declining Effectiveness Visible**

A health-care system that records only how often AI is used, without recording where it errs, which groups experience failure, whether alerts are ignored, or how clinical behavior changes, cannot become a genuine learning system. Continuous monitoring should include versions, patient composition, key performance measures, clinical outcomes, major incidents, coverage and refusal, human takeover, and remediation. These data must also respect privacy, minimum necessity, and patient rights; monitoring cannot turn all aspects of patients' lives into an unlimited data source.

# **9 | Patient Heterogeneity and Health Equity: Average Performance Cannot Obscure Who Is Harmed**

## **9.1 Average Accuracy Can Coexist with Structural Inequity**

Obermeyer and colleagues' 2019 analysis of a widely used population-health management algorithm found that it used health-care costs as a proxy for health needs. Black patients were actually sicker at the same risk score; correcting the bias would increase the proportion of Black patients receiving additional care from 17.7% to 46.5%. [22] The central problem was not simply that the model “saw race,” but that an apparently effective proxy encoded existing disparities in health-care access as health needs. Fairness in medical AI therefore begins with problem definition and outcome selection.

## **9.2 Local Fairness Corrections Do Not Guarantee Fairness Elsewhere**

Yang and colleagues' 2024 study examined fairness and generalization across medical-imaging tasks in radiology, dermatology, and ophthalmology and six global chest-radiograph datasets. Fairness corrections effective within local data distributions did not guarantee optimal fairness in new external testing environments.[24] Fairness therefore cannot be permanently settled by a one-time disparity metric on a development dataset. Patient heterogeneity, equipment, disease patterns, baseline prevalence, and health-system differences all need to enter external validation.

## **9.3 LLM Fairness Requires Multidimensional and Participatory Evaluation**

Pfohl and colleagues' 2024 work with Med-PaLM 2 and EquityMedQA developed multidimensional, participatory human evaluation methods, showing that fairness-related harms in long-form medical answers require evaluators with different professional and lived experiences to make them visible. The authors also explicitly note that these tools are insufficient to establish equitable health outcomes after deployment as a whole. [23] Model fairness evaluation is thus a necessary condition, not a population-level equity outcome.

*Better average performance does not automatically cancel minority groups' higher error rates, delays, excessive interventions, or exclusion from services.*

## **10 | Generative AI and Direct Patient Use: New Routes to Health Information and New Evidence Gaps**

Generative AI changes one of medicine's most traditional entry points: patients can obtain highly coherent, personalized, sustained conversational explanations about health without first entering a hospital or search engine. WHO's LMM guidance identifies direct patient use as one of five major application categories while explicitly identifying false, inaccurate, biased, or incomplete information, privacy, automation bias, and manipulation risks.[2] Neither the opportunities nor the risks are fully covered by the traditional question of whether a medical device is approved, because many general-purpose models enter everyday life outside the status of regulated medical devices.

### **10.1 Health Information Is Neither Diagnosis nor Treatment**

Direct AI use can reduce language and knowledge barriers, help patients prepare questions for appointments, understand terminology, organize medical histories, and recognize signs that warrant medical attention. However, once a system undertakes risk stratification for patients, recommends stopping medication, delays care seeking, or changes treatment, it enters a higher-consequence level. The same output carries different evidentiary requirements in health education, triage guidance, and treatment recommendations.

### **10.2 Greater Direct Access Makes Redesigning Professional Relationships More Necessary**

AI can make health knowledge continuously available, but patients often need more than answers: physical examination, laboratory testing, treatment resources, long-term care, shared decisions with families, and someone to assume responsibility. If generative AI becomes the first entry point, research should examine when it should proactively escalate to clinicians, how it should express uncertainty, how to prevent fluent language from concealing insufficient evidence, and whether it actually reduces rather than increases delays and misinformation among disadvantaged groups.

## **11 | Public Health: Earlier Detection Does Not Establish Effective Intervention**

### **11.1 AI Has Already Entered Public-Health Intelligence**

WHO launched EIOS 2.0 in 2025 for early detection of public-health threats from open sources, adding new AI functions. At launch, the system was already used in more than 110 Member States and approximately 30 organizations and networks; WHO's current official EIOS page further records use by 120 countries and 30 organizations and networks by the end of 2025.[27,38] This establishes that AI is no longer a hypothetical public-health tool but has entered real global health-intelligence infrastructure. The scale of adoption alone, however, does not establish improved public-health outcomes.

### **11.2 Detection Still Has to Be Followed by Verification, Interpretation, and Intervention**

EIOS helps experts identify relevant content more rapidly from open sources such as websites and social media, after which verification and assessment remain necessary.[27] Public-health errors may involve missing real threats, treating noise as an outbreak, misreading geographical or group-level signals, excessive intervention, or damage to trust. Faster AI signal detection does not establish epidemiological causation, much less the legitimacy of a coercive measure.

### **11.3 AI-Supported Community Listening: From More Information to Better-Targeted Action**

WHO's 2026 EPI-WIN cholera event publicly demonstrated AI-supported community listening: analysis of hotlines, social media, radio, surveys, and frontline reports to identify outbreak reports, concerns, rumors, service gaps, and barriers to seeking care, informing targeted decisions on WASH, risk communication, vaccination, and rapid response.[28] This is category A material documenting a real application pathway, but an event and its cases do not constitute randomized evidence of effects. The actual public-health questions remain whether action becomes earlier, better targeted, and more equitable, and whether morbidity, mortality, and service disruption decline.

## **12 | Access and Health Resources: Can Reduced Professional Scarcity Truly Become Broadly Shared Benefit?**

One of medical AI's most promising possibilities is reducing the marginal cost of knowledge and professional services. Enormous global service gaps and health-worker shortages make remote support, automated documentation, imaging assistance, training, and patient education practically significant.[25–26,29] Yet many conditions still separate “a model can answer” from “a person actually receives health care”: devices, connectivity, language, payment, testing, medicines, surgery, nursing, referral, and transport.

### **12.1 Lowering the Cost of One Judgment Does Not Necessarily Lower Total System Costs**

Greater AI screening sensitivity may generate more subsequent investigations. Time saved through reduced documentation burden may permit treating more patients or be filled again by new productivity requirements. Easier access to health advice may reduce low-value visits or increase unnecessary consultations. The increased non-neoplastic resections in Makar and colleagues' colonoscopy meta-analysis provide a concrete example of stronger detection also increasing resource use.[37] Resource savings must therefore be established across the complete health-service chain, not inferred from a model.

### **12.2 Broadly Shared Benefit Requires Examining Who Still Cannot Get In**

Health equity concerns more than differences in model accuracy across groups. It includes basic digital access, language, disability access, payment, geography, health workforce, trust, and avenues of redress. WHO's global health strategy places health equity, primary health care, UHC, and resilience within one strategy, and includes responsible AI, digital health systems, and cybersecurity among future capabilities.[29] This is a baseline for direction and governance, not evidence that a particular AI system has improved UHC.

## **13 | Health-System Resilience: Shared Models, Correlated Errors, and Dependence on Digital Infrastructure**

When a hospital has only one AI tool, an error may remain local. When many hospitals share a model, cloud platform, imaging algorithm, identity interface, and supplier, errors may become correlated events. Health-care systems have traditionally reduced single-point failure through professional redundancy, review, referral, and multiple technological pathways. AI standardization may improve quality and access while also reducing cognitive and infrastructural redundancy.

### **13.1 Correlated Errors Deserve Greater System-Level Attention Than Individual Errors**

If a model bias affects millions of patients simultaneously, the risk structure of traditional occasional individual misdiagnosis changes. A shared model may also cause different institutions to fail simultaneously for the same subgroup, anomalous input, or updated version. Health-system safety assessment should therefore include the radius of failure, shared dependencies, backup workflows, human takeover, and interinstitutional incident sharing, rather than examining only average failure rates per device.

### **13.2 Cybersecurity and the Ability to Exit Belong within Health-System Evidence**

WHO's 2023 regulatory considerations and its global health strategy for 2025–2028 both place cybersecurity, data governance, and resilience among the practical boundaries of digital health and AI.[1,29] If deep integration leaves a hospital unable to continue diagnosis and treatment during an AI service interruption, even high technical performance may create new vulnerability. In health care, the ability to exit does not end with a patient pressing an off switch: after refusal or failure, viable human pathways must remain, essential records must remain accessible, prescriptions and referrals must continue, and emergency services must not be interrupted.

### **13.3 Resilience Does Not Require Every System to Run Permanently in Duplicate**

Full redundancy has real costs, and not every low-risk function needs the same backup. A reasonable approach tiers requirements according to consequences, difficulty of substitution, recovery time, and scale of dependence. A temporary AI documentation-tool failure plainly differs from interruption of ICU alerts or national public-health

surveillance. This paper requires resilience to enter evidence and design; it does not make “more redundancy is always better” an absolute rule.

## 14 | Interfaces of Clinical Responsibility: Who Must Reenter, and at Which Points?

Responsibility in medicine is first a relationship within professional practice: who takes the history, interprets results, decides on further investigations, orders treatment, monitors adverse reactions, communicates with patients, and reassesses deteriorating conditions. AI adds model developers, product manufacturers, hospital procurement, IT, and data teams to the responsibility chain, but additional nodes cannot leave clinical consequences without an accountable party.

**Table 4. Professional Handoffs across the Complete Responsibility Chain of Medical AI**

| Node                       | What medicine must answer first                                                | Principal responsible participants                | Further handoff                                                                |
|----------------------------|--------------------------------------------------------------------------------|---------------------------------------------------|--------------------------------------------------------------------------------|
| Data / inputs              | Is the information sufficient, and are its provenance and quality appropriate? | Clinical team + data / laboratory professionals   | Privacy and legality of data use → law                                         |
| Model outputs              | Performance, calibration, uncertainty, and subgroup limitations                | AI / computer science + clinical disciplines      | System guarantees → AI; evidentiary strength → epistemology                    |
| Clinical interpretation    | Does the output fit the case, and is review or further testing needed?         | Physicians / relevant professionals               | Treatment authority and professional standards → law / professional regulation |
| Patient decisions          | Have risks, choices, and alternatives been adequately communicated?            | Patients + clinicians                             | Consent and surrogate decision-making → ethics / law                           |
| Treatment / implementation | Is the intervention safely delivered for the indicated use?                    | Clinical organizations and professionals          | Resources and payment → economics / institutions                               |
| Follow-up / remediation    | Are errors detected, corrected, remedied, and learned from?                    | Health-care institutions + patient-safety systems | Compensation / legal liability → law                                           |
| Population monitoring      | Do effects and harms change at population scale?                               | Public-health / health-services research          | Coercion and public power → political science / law                            |

This table neither assigns proportions of legal liability nor makes physicians the ultimate bearers of all AI errors. It blocks one form of responsibility evasion: developers say it is only a tool, hospitals say they procured a compliant product, and physicians say the system recommended it, leaving patients to bear the full consequences while no one can explain why. Law must reenter to determine legal attribution, but the health-care system must first ensure that the factual chain and the chain of professional decisions can be reconstructed.

## 15 | Medicine and Civilization's Direction: Health Matters, but Is Not an Overall Score for Life

The paper on civilization's direction sets out a progression of goals: “freedom from survival insecurity → happiness in life → a more meaningful life.” Medicine most directly supports safety of life, basic health, and reduced avoidable suffering, and contributes to happiness through health, capabilities, and conditions of care. But

medicine cannot inflate its professional objective into civilization's overall objective. Illness, disability, aging, or shorter life expectancy does not confer lesser personal worth, less political standing, or less meaning in life.

### 15.1 Health Optimization Cannot Become Society's Single Objective Function

AI can predict risks, prompt exercise, and optimize treatment and public-health resources. But if minimizing health risk expands into a uniform management standard for all lifestyles, relationships, occupations, and freedoms, medical metrics have exceeded their proper object. Medical facts can explain risks; they do not automatically confer authority to choose how people should live.

### 15.2 Efficiency Cannot Entirely Absorb Care

Health care's value encompasses more than improved biological indicators: explanation, reassurance, shared decision-making, companionship, care, and confronting uncertainty and death also matter. AI can free physicians from documentation, reduce information burdens, and extend professional capabilities. If these efficiencies ultimately deepen clinical relationships, they expand human capabilities. If efficiency merely assigns more patients to each physician and relegates people to final confirmation interfaces, technological progress may coexist with thinner relationships. Existing ambient AI documentation research demonstrates the practical possibility of the former, but patient experience and long-term clinical outcomes still require further validation.[35–36]

*Health is an important condition of civilized life, but it is not the entirety of life's meaning.*

## 16 | Interdisciplinary Handoffs: Who Reenters When Evidence Stops?

**Table 5. Strict Handoffs between the Medicine / Public-Health Paper and Adjacent Disciplines**

| Receiving discipline                 | What it can answer first                                                                      | Where the medical paper stops                                                               |
|--------------------------------------|-----------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------|
| Life sciences                        | Disease mechanisms, life processes, and biological individuation                              | Cannot substitute for clinical effectiveness, risk-benefit assessment, or population health |
| AI / computer science                | Model performance, system guarantees, versions, permissions, and technical exit               | Cannot substitute for patient net benefit or public authorization                           |
| Epistemology / philosophy of science | Evidence, causation, models, and uncertainty                                                  | Cannot produce facts for particular diseases / interventions in their place                 |
| Cognitive science                    | Automation bias, judgment, learning, and mechanisms of human–AI collaboration                 | Cannot substitute for clinical outcomes or determinations of responsibility                 |
| Ethics / law / political philosophy  | Consent, privacy, rights, responsibility, personhood, and normative boundaries                | Cannot substitute for clinical risk-benefit data                                            |
| Political science                    | Public-health powers, national health capacity, and emergency governance                      | Cannot substitute for epidemiological effects or clinical facts                             |
| Economics / political economy        | Payment, costs, ownership, and allocation of health resources                                 | Cannot substitute for treatment effectiveness or patient safety                             |
| Sociology / anthropology             | Clinician–patient relationships, trust, illness experiences, cultures of care, and inequality | Cannot substitute for clinical trials or epidemiology                                       |
| Systems science                      | Cascades, resilience, correlated failures, and complex-system structures                      | Cannot substitute for specific medical facts                                                |

| Receiving discipline                         | What it can answer first                                                                   | Where the medical paper stops                                                 |
|----------------------------------------------|--------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------|
| Education                                    | Medical education, capability formation, and professional training in the AI era           | Cannot substitute for patients' clinical outcomes                             |
| International relations / strategic security | Cross-border health security and strategic interaction in international crises             | Cannot substitute for outbreak facts, international law, or clinical judgment |
| Ecology / Earth systems                      | Shared material conditions of environment and health, and infrastructure carrying capacity | Ecological facts cannot directly define treatment or health-related rights    |

## 16.1 Real Causation Extends Outward, but Evidentiary Standards Do Not Merge

AI-assisted diagnosis may alter patient flows and consequently health resources; resource changes affect access and inequality; public-health data influence government action; and health systems' dependence on a small number of clouds and models may become a national-security issue. Real causal chains can extend outward throughout, but each step requires new evidence from the appropriate discipline. Medicine's proximity to consequences for life does not give it superior authority over all civilizational questions.

## 17 | What Facts Could Change This Paper? The Strongest Objections and a Refutable Research Agenda

### 17.1 Objection One: AI Will Eventually Surpass Physicians across the Board, So These Intermediate Evidence Gates Are Temporary

If future AI consistently outperforms current health-care systems in real open environments, across diseases, populations, institutions, and long timescales, this paper must of course rewrite the division of labor between humans and AI. Yet a stronger system does not remove patient heterogeneity, treatment consequences, rights, resources, or public-health externalities. The scope of human review is most likely to change, rather than the principle that clinical outcomes still require evidence.

### 17.2 Objection Two: RCTs Are Too Slow and AI Updates Too Quickly for Traditional Medical Evidence to Remain Applicable

This is partly correct. Not every AI change requires another large RCT; real-world evidence, quasi-experiments, continuous monitoring, staged evaluation, and predefined update plans may better suit dynamic systems.[6,10–14] But traditional RCTs being inapplicable in some circumstances does not mean technical performance can replace clinical evidence. The faster updates occur, the more necessary it becomes to bind a particular version, change, and outcome to a particular kind of evidence.

### 17.3 Objection Three: Large-Scale Real-World Adoption Is Itself the Best Evidence

Adoption can establish demand, usability, and organizational choice, but not automatically clinical net benefit. A widely used sepsis model may still perform poorly in external validation, and large observational studies may still be affected by selection and confounding.[17,21] Truly strong real-world evidence requires comparisons, outcomes, time, and bias control, rather than download counts or numbers of hospitals.

### 17.4 Objection Four: AI Only Assists, So Physicians Can Simply Bear All Final Responsibility

If physicians have sufficient information, time, capabilities, authority, and genuine alternatives, professionals can certainly bear a substantial part of clinical responsibility. But where system outputs become defaults, time is compressed, physicians cannot understand version changes, or institutions offer no alternative pathway, a physician's signature may be merely formal control. Responsibility should be reassigned according to actual control, foreseeability, professional duties, and institutional arrangements, with legal conclusions determined by law.

## 17.5 Objection Five: Fairness Requires Only Meeting Group-Metric Targets

Work by Obermeyer, Pfohl, Yang, and their colleagues collectively shows that fairness problems can arise from proxies, problem definitions, output content, data distributions, and external environments, and that local corrections may fail in new settings.[22–24] If methods emerge that reliably predict deployment equity outcomes across regions, this paper should reduce its emphasis on setting-by-setting revalidation. Current evidence does not support one-time “fairness certification.”

## 17.6 Facts That Could Change or Invalidate This Paper

- If longitudinal studies show that medical-AI model performance reliably predicts patient net benefit in the vast majority of high-consequence settings without additional clinical evidence, the paper's central argument about a broken bridge between performance and clinical consequences should contract substantially.
- If general-purpose medical AI demonstrates stable external validity across institutions, populations, times, and diseases, with deployment changes having minimal effects, the paper should reduce its demands for local revalidation and version reassessment.
- If large-scale human-AI collaboration studies show that AI reliably improves clinical judgment, professional capabilities, and patient outcomes over the long term without significant automation bias or skill loss, the corresponding category B risk judgments must narrow.
- If public-health research shows that AI-enhanced surveillance reliably translates into lower morbidity and mortality and greater equity, the paper should strengthen its positive assessment of population-health effects.
- If further preliminary disciplinary research establishes that clinical and population-health facts amount only to a few cases, and that life sciences, AI science, and law and ethics can fully assume responsibility for them without overreach, the paper's standing as an independent work should be reconsidered or even withdrawn.

## 17.7 A Research Agenda Ready to Pursue

- External validation across institutions, populations, and time: report not only AUC but also calibration, subgroups, thresholds, and clinical consequences.
- Trials of actual human-AI teams: compare models alone, professionals alone, different human-AI interfaces, and different training approaches, rather than only models against physicians.
- Patient-outcome trials: extend the links from accuracy, detection rates, and workload to mortality, complications, quality of life, overdiagnosis, and long-term follow-up.
- Version and drift research: establish empirical rules for when updates require revalidation and which changes can bridge to previous evidence.
- Direct patient use of generative AI: investigate triage escalation, delayed care seeking, erroneous advice, health literacy, disadvantaged groups, and long-term behavioral consequences.
- Health equity: connect fairness metrics in development data to actual access, treatment, outcomes, and redress, and report negative findings.
- Public health: compare whether AI-assisted surveillance actually shortens verification-to-response time, improves intervention targeting, and reduces population harm.
- Health-system resilience: stress-test shared-supplier failures, cyberattacks, correlated model errors, and fallback pathways to human operation.

## Conclusion | The True Endpoint Is Better Real Health Outcomes, Not a Smarter Model

The age of AGI will continually tempt medicine toward a shortcut: systems can now perform tasks previously reserved for professionals, so technology is solving medical problems. The facts are more complex and more hopeful. AI has indeed produced measurable positive results in some screening, detection, knowledge, and documentation settings. It has also exposed new problems, including failed external validation, incomplete translation into human-AI collaboration, group bias, and system dependence. Both sets of facts must be preserved.

Medicine's professional value emerges precisely here: it asks whether real patients benefit, rather than whether a machine resembles a doctor. Public health asks whether population health actually improves, rather than whether more signals are generated. Health systems ask whether services are more accessible, workers better

able to care, and systems better able to keep operating through errors and crises, rather than whether digitization has increased.

When technology enters health care, the first question is not how intelligent the system is, but whether patients consequently achieve better real outcomes.

Only once this question is adequately answered can civilization ask how everyone can genuinely possess these new health capabilities, without reducing human lives once again to a table of risks.

## Further Reading and Entry Points for Shared Research

### Upward | Why Do Clinical Facts Ultimately Become Questions about Civilization?

Read [Reckoning with the Disciplines in the Age of AGI: Disciplinary Boundaries, Interactions, and a Shared Research Agenda V2.4](#) to explore why disciplines must retain evidentiary independence and why health care, economics, education, public power, relationships, and meaning reconnect through real causal chains. This paper does not require prior acceptance of Civilization Leap theory; any clinical counterexample has the standing to revise the overall judgment in return.

Read [Opportunities, Directions, Methods, and Goal Realization for Humanity's Civilization Leap in the Age of AGI: Why Change Is Necessary, Where We Can Go, and How to Get There V1.1](#) to explore how health enters civilization's direction of "freedom from survival insecurity, happiness in life, and a more meaningful life." That work develops normative commitments and direction; it is not a source of this paper's clinical evidence.

### Downward | How Can Judgments Become Real Validation?

Begin with clinical trials, diagnostic studies, real-world research, patient safety, implementation science, epidemiology, health-services research, and public-health field research. For machine-side capabilities, permissions, logs, stopping, and exit, then turn to engineering research on [AGI-COS](#), a candidate engineering-governance framework for machine-side capabilities, permissions, auditing, stopping, and exit. For multi-agent exploration of high-consequence choices, consider validation interfaces such as [Human-COS](#), a candidate system for multi-agent, multiperspective exploration of consequences and judgment support before high-consequence choices become irreversible. These can only serve as vehicles awaiting validation; they cannot establish medical facts in reverse.

### Leftward | Who Must Reenter?

[AI / computer science](#) reassesses models and system guarantees; [life sciences](#) reassess mechanisms of life and disease; [epistemology](#) reassesses evidence and extrapolation; [cognitive science](#) studies automation bias and human-AI judgment; [ethics and law](#) address consent, responsibility, privacy, and rights; [economics](#) addresses payment, costs, and resources; [political science](#) addresses public-health powers; [sociology / anthropology](#) address clinician-patient relationships, trust, and cultures of care; and [education](#) addresses medical education and the formation of professional capabilities.

### Rightward | If You Disagree, Which Contrary Directions Most Deserve Investigation?

Four genuinely strong opposing positions deserve priority: first, AI can generate stable patient net benefit in real clinical practice faster than this paper expects; second, large-scale real-world evidence can replace traditional staged evaluation; third, human-AI collaboration problems primarily arise from interfaces and training rather than AI itself; and fourth, future general-purpose medical AI can make local validation, professional takeover, and many traditional workflows unnecessary costs. If new data support these directions, the paper should narrow its claims rather than rename counterexamples and absorb them.

### Forward | What Professional Work Could Directly Change This Paper?

Clinicians can submit negative trials, failure cases, and interinstitutional replications. Epidemiologists can establish subgroup and external-validity evidence. Patient-safety teams can disclose incidents and near misses. Public-health researchers can follow faster surveillance through to interventions and population outcomes. Health-services researchers can measure AI's actual effects on access, queues, costs, and workforce. Patients and

communities can identify harms, trust issues, and access problems overlooked by professional research. The most valuable participation is not agreement, but professional work capable of changing the theory.

## Civilization's Future Chain | What Kind of World Might People Inhabit If These Problems Were Better Resolved?

The [Living Civilization trilogy](#) develops concrete future lifeworlds: how hospitals, families, elder care, communities, work, and public life might change when diagnosis, health knowledge, and care capabilities reach ordinary people more easily. It offers a contestable, reconstructable vision of the future, not empirical evidence for this paper. Spark, a candidate system for extending individual capabilities and care; Nebula, a candidate vehicle for civilizational collaboration and operation; AGI-COS; and Human-COS continue to explore different possible practical embodiments. They too must accept revision in light of clinical facts, failures, and patient experiences.

Civilization Leap five-volume series: <https://www.civitas.top/en/documents/p02-five-001/> | Public entry point: <https://www.civitas.top/en/documents/p01-entry-006/> | Complete library: <https://www.civitas.top/en/library/>

## Appendix A | Register of Source–Claim Boundaries

Table A1. Principal Sources and the Boundaries of Claims

| Source group                                                                                    | Scope of use in this paper                                                                               | What must not be inferred                                                                   |
|-------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------|
| [1–2] WHO AI health regulation / LMM guidance                                                   | Formal international considerations, application categories, and risk boundaries for AI in health        | Does not establish a particular model's effectiveness or globally uniform legal obligations |
| [3–4] WHO patient-safety documents                                                              | System-level facts about patient safety and diagnostic harm                                              | Does not attribute all diagnostic harm to AI                                                |
| [5–7] FDA medical AI / software / RWE                                                           | US regulatory authorization, software boundaries, and real-world evidence practice                       | Cannot be extrapolated into global law or net benefit in every deployment                   |
| [8–14] Medical-AI reporting / evaluation standards                                              | Transparent reporting, risk of bias, early clinical evaluation, and trial design                         | Standards themselves are not evidence of effectiveness                                      |
| [15–16] Calibration / DCA                                                                       | Methods for reliable risk estimation and the consequences of clinical decisions                          | Does not constitute complete evidence of patient outcomes                                   |
| [17–18] Sepsis / LLM diagnostic trial                                                           | Empirical evidence of failed external validation and problems translating capability into human–AI teams | Does not establish that all medical AI is without benefit                                   |
| [19–21] MASAI / PRAIM                                                                           | Positive randomized and real-world evidence in defined screening settings                                | Does not automatically generalize across countries, equipment, populations, or endpoints    |
| [22–24] Fairness research                                                                       | Proxy bias, LLM fairness evaluation, and limits to fairness generalization across settings               | Does not establish inevitable discrimination in every deployment                            |
| [25–29,38] WHO UHC / workforce / public health / global strategy                                | Global service gaps, health workforce, AI adoption in public health, and health-system direction         | Does not establish that AI has improved UHC or outbreak outcomes                            |
| [30–37] Health-technology assessment / diagnosis / patient safety / documentation / colonoscopy | Evidence frameworks, established medical questions, and supplementary empirical findings across settings | Limited by their respective designs and endpoints                                           |

## Appendix B | Checklist for Evidentiary Transitions in Medical AI

**Table B1. Minimum Checklist for Moving Medical AI from Performance to Real Outcomes**

| Item                   | Questions that must be answered                                                                                       | Minimum discipline                                        |
|------------------------|-----------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------|
| Object                 | Are we discussing a model, product, deployment, workflow, patient pathway, or population system?                      | Avoid treating AI as a single object                      |
| Intended use           | Exactly which decision does the system assist, who uses it, and at what point?                                        | A change of use requires reopening the evidence           |
| Technical performance  | Are discrimination, calibration, robustness, and subgroup performance adequate?                                       | Do not stop at AUC                                        |
| External validity      | Has validation covered new institutions, times, devices, and populations?                                             | Development-set performance cannot simply be extrapolated |
| Human-AI collaboration | How do professionals understand, use, override, and take over from the system?                                        | A merely formal human in the loop is insufficient         |
| Clinical consequences  | Does it change testing, treatment, delays, complications, and patient outcomes?                                       | Intermediate metrics do not replace endpoints             |
| Net harm               | What are the consequences for false positives, missed diagnoses, overdiagnosis, alert fatigue, and opportunity costs? | Benefits and harms must be assessed together              |
| Equity                 | Who benefits, who is harmed, and do subgroup outcomes and access change?                                              | Averages do not settle consequences for minorities        |
| Real world             | Are effects and rare risks continuously monitored after deployment at scale?                                          | Authorization is not permanent validity                   |
| Version                | Have models, data, thresholds, interfaces, or workflows changed?                                                      | Version changes may alter the object of evidence          |
| Patient rights         | Can patients understand AI's involvement, refuse it, appeal, and continue receiving services?                         | Refer to ethics, law, and institutional arrangements      |
| Public health          | Does better surveillance actually improve interventions, population outcomes, and equity?                             | Faster signals ≠ better health                            |
| Resilience             | Can care continue during failures, cyberattacks, or supplier withdrawal?                                              | Availability is part of clinical safety                   |

This table is not an evidence score, algorithm for authorizing clinical use, or substitute for regulation. Different diseases, risks, and technologies require different study designs. Its sole function is to prevent evidence silently skipping from one object to another.

# Statement on the Use of Generative Artificial Intelligence

Generative artificial intelligence assisted in organizing materials and literature leads, structural design, drafting and polishing portions of the text, cross-checking, verifying citation leads, and layout checks. AI outputs are not treated as empirical data, authoritative sources, or independent evidence. All medical, public-health, and health-system facts are traced to verifiable original research, formal guidelines, professional reports, or regulatory materials. The author reviewed and approved the research questions, theoretical propositions, A / B / C / F classifications, normative commitments, selection of materials, use of citations, and final manuscript, and assumes final academic responsibility. Generative AI is not listed as an author. Specific tools, versions, dates, and scope of use will be added upon formal submission in accordance with the target journal's requirements.

Medical-specific discipline: diagnoses, treatments, risk figures, medication recommendations, or conclusions about clinical effectiveness or population health generated by AI must not enter this paper directly merely because the wording is fluent. Every quantitative medical claim must be traced to the corresponding study object, sample, design, endpoints, and limitations. This paper does not provide personal medical advice.

## References

- [1] World Health Organization. Regulatory considerations on artificial intelligence for health. Geneva: WHO; 2023. ISBN 978-92-4-007887-1.
- [2] World Health Organization. Ethics and governance of artificial intelligence for health: guidance on large multi-modal models. Geneva: WHO; 2024. ISBN 978-92-4-008475-9.
- [3] World Health Organization. Global patient safety action plan 2021–2030: towards eliminating avoidable harm in health care. Geneva: WHO; 2021. ISBN 978-92-4-003270-5.
- [4] World Health Organization. Global patient safety report 2024. Geneva: WHO; 2024. ISBN 978-92-4-009545-8.
- [5] U.S. Food and Drug Administration. Artificial Intelligence-Enabled Medical Devices. Updated periodically; accessed 26 Aug 2026.
- [6] U.S. Food and Drug Administration. Use of Real-World Evidence to Support Regulatory Decision-Making for Medical Devices: Guidance for Industry and Food and Drug Administration Staff. Final Guidance. 18 Dec 2025.
- [7] U.S. Food and Drug Administration. Clinical Decision Support Software: Guidance for Industry and Food and Drug Administration Staff. Final Guidance. Issued 6 Jan 2026; re-issued 29 Jan 2026. January 2026 revisions included Level 2 minor corrections.
- [8] Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*. 2024;385:e078378. doi:10.1136/bmj-2023-078378.
- [9] Moons KGM, Damen JAA, Kaul T, et al. PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. *BMJ*. 2025;388:e082505. doi:10.1136/bmj-2024-082505.
- [10] Sounderajah V, Guni A, Liu X, et al. The STARD-AI reporting guideline for diagnostic accuracy studies using artificial intelligence. *Nat Med*. 2025;31:3283–3289. doi:10.1038/s41591-025-03953-8. Author Correction published 13 Jul 2026.
- [11] Vasey B, Nagendran M, Campbell B, et al. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *Nat Med*. 2022;28:924–933. doi:10.1038/s41591-022-01772-9.
- [12] Liu X, Cruz Rivera S, Moher D, et al. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nat Med*. 2020;26:1364–1374. doi:10.1038/s41591-020-1034-x.
- [13] Cruz Rivera S, Liu X, Chan A-W, et al. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *Nat Med*. 2020;26:1351–1363. doi:10.1038/s41591-020-1037-7.
- [14] Lekadir K, Frangi AF, Porras AR, et al. FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare. *BMJ*. 2025;388:e081554. doi:10.1136/bmj-2024-081554. Author-name correction: *BMJ*. 2025;388:r340. doi:10.1136/bmj.r340.
- [15] Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. *BMC Med*. 2019;17:230. doi:10.1186/s12916-019-1466-7.
- [16] Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making*. 2006;26(6):565–574. doi:10.1177/0272989X06295361.
- [17] Wong A, Otles E, Donnelly JP, et al. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. *JAMA Intern Med*. 2021;181(8):1065–1070. doi:10.1001/jamainternmed.2021.2626.

- [18] Goh E, Gallo R, Hom J, et al. Large Language Model Influence on Diagnostic Reasoning: A Randomized Clinical Trial. *JAMA Netw Open*. 2024;7(10):e2440969. doi:10.1001/jamanetworkopen.2024.40969.
- [19] Hernström V, Josefsson V, Sartor H, et al. Screening performance and characteristics of breast cancer detected in the Mammography Screening with Artificial Intelligence trial (MASAI): a randomised, controlled, parallel-group, non-inferiority, single-blinded, screening accuracy study. *Lancet Digit Health*. 2025;7(3):e175–e183. doi:10.1016/S2589-7500(24)00267-X.
- [20] Gommers J, Hernström V, Josefsson V, et al. Interval cancer, sensitivity, and specificity comparing AI-supported mammography screening with standard double reading without AI in the MASAI study: a randomised, controlled, non-inferiority, single-blinded, population-based, screening-accuracy trial. *Lancet*. 2026;407(10527):505–514. doi:10.1016/S0140-6736(25)02464-X.
- [21] Eisemann N, Bunk S, Mukama T, et al. Nationwide real-world implementation of AI for cancer detection in population-based mammography screening. *Nat Med*. 2025;31:917–924. doi:10.1038/s41591-024-03408-6.
- [22] 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):447–453. doi:10.1126/science.aax2342.
- [23] Pfohl SR, Cole-Lewis H, Sayres R, et al. A toolbox for surfacing health equity harms and biases in large language models. *Nat Med*. 2024;30:3590–3600. doi:10.1038/s41591-024-03258-2.
- [24] Yang Y, Zhang H, Gichoya JW, Katabi D, Ghassemi M. The limits of fair medical imaging AI in real-world generalization. *Nat Med*. 2024;30:2838–2848. doi:10.1038/s41591-024-03113-4.
- [25] World Health Organization & World Bank. Tracking universal health coverage: 2025 global monitoring report. Geneva/Washington, DC; 2025.
- [26] World Health Organization. Health and care workforce – Global strategy on human resources for health: workforce 2030. Executive Board 156th session, B156/15. 2025.
- [27] World Health Organization. WHO upgrades its public health intelligence system to boost global health security: launch of EIOS 2.0. 13 Oct 2025.
- [28] World Health Organization. WHO Health Emergencies EPI-WIN webinar: artificial intelligence-supported listening to communities for cholera. 6 May 2026.
- [29] World Health Organization. A Global Health Strategy for 2025–2028: Fourteenth General Programme of Work. Geneva: WHO; 2025. ISBN 978-92-4-010101-2.
- [30] National Institute for Health and Care Excellence. Evidence standards framework for digital health technologies (ECD7). Updated 9 Aug 2022.
- [31] Bates DW, Levine D, Syrowatka A, et al. The potential of artificial intelligence to improve patient safety: a scoping review. *NPJ Digit Med*. 2021;4:54. doi:10.1038/s41746-021-00423-6.
- [32] National Academies of Sciences, Engineering, and Medicine. Improving Diagnosis in Health Care. Washington, DC: The National Academies Press; 2015. doi:10.17226/21794.
- [33] Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med*. 2019;25:44–56. doi:10.1038/s41591-018-0300-7.
- [34] Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med*. 2019;17:195. doi:10.1186/s12916-019-1426-2.
- [35] Duggan MJ, Gervase J, Schoenbaum A, et al. Clinician Experiences With Ambient Scribe Technology to Assist With Documentation Burden and Efficiency. *JAMA Netw Open*. 2025;8(2):e2460637. doi:10.1001/jamanetworkopen.2024.60637.
- [36] Olson KD, Meeker D, Troup M, et al. Use of Ambient AI Scribes to Reduce Administrative Burden and Professional Burnout. *JAMA Netw Open*. 2025;8(10):e2534976. doi:10.1001/jamanetworkopen.2025.34976.
- [37] Makar J, Abdelmalak J, Con D, Hafeez B, Garg M. Use of artificial intelligence improves colonoscopy performance in adenoma detection: a systematic review and meta-analysis. *Gastrointest Endosc*. 2025;101(1):68–81.e8. doi:10.1016/j.gie.2024.08.033.
- [38] World Health Organization. Epidemic Intelligence from Open Sources (EIOS) initiative. Current official initiative page; accessed 26 Aug 2026. By the end of 2025, 120 countries and 30 organizations and networks were leveraging the EIOS system.

## Version Note

V1.1 completes professional review and necessary controlled revisions on the basis of V1.0, the first complete manuscript. This paper's independent factual responsibilities remain clinical effectiveness, clinical net benefit, patient safety, real-world evidence, public health, population health, and health-system resilience. It does not make

the respective judgments belonging to life sciences, AI / computer science, ethics and law, economics, political science, or deliberation on civilization's direction.

This version retains the main argumentative chain: model performance → external validation → real workflows and human-AI collaboration → clinical behavior → patient outcomes → ongoing real-world monitoring → population health and health-system consequences. Every step must identify its evidentiary object, scope of applicability, and stopping point. The controlled revisions primarily clarify that model-only performance in Goh and colleagues' study was an exploratory secondary analysis; precisely specify PRAIM's prospective, multicenter, real-world, observational noninferiority implementation design and sample definitions; distinguish MASAI's randomized enrollment from its analytical sample; correct the issue / reissue information for the FDA's 2026 final guidance on clinical decision-support software; register the FUTURE-AI author-name correction; and remove unnecessary internal process-handoff language from the external-release manuscript. The main argumentative chain and A / B / C / F boundaries are unchanged. V1.1 is the current professional research baseline. This same-version cross-paper consistency update only synchronizes the version of the civilization-direction paper, minimal explanations of candidate vehicles, and external navigation wording; it does not change the medical conclusions.