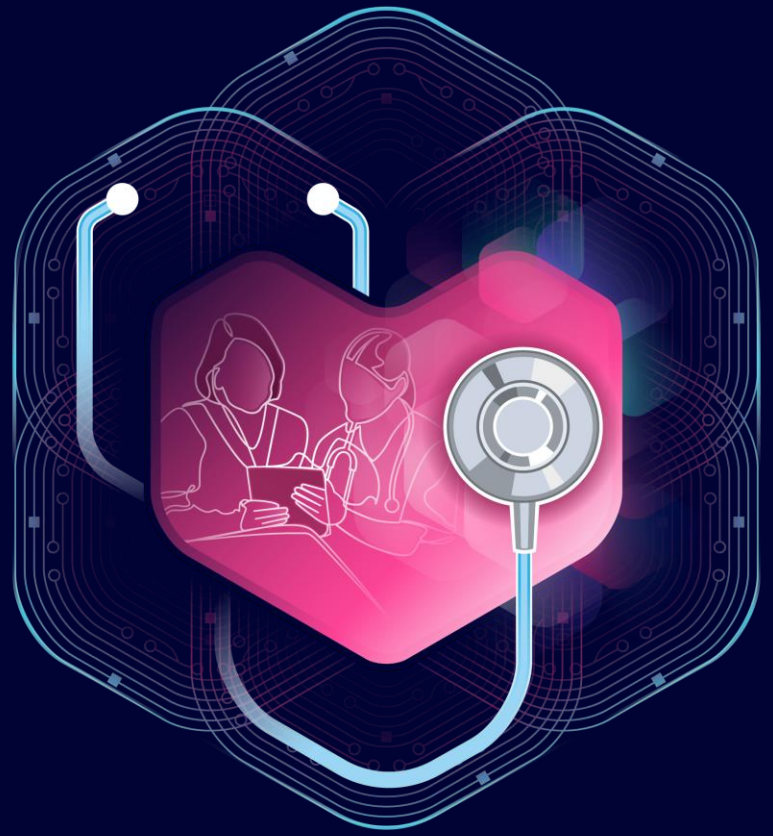
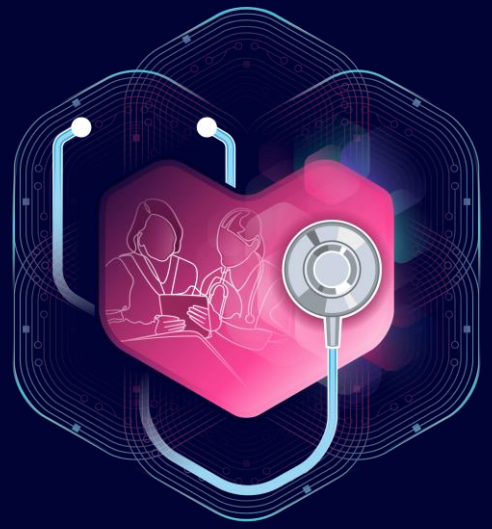
ARPA-H

EVALUATING CLINICAL AGENTIC AI



Haider Warraich, MD
Program Manager, Scalable Solutions Office



ADVOCATE

Agentic AI-Enabled Cardiovascular Care Transformation

ARPA H

What if FDA-authorized clinical agentic AI could provide safe and effective cardiovascular care to every American?

VISION

To transform advanced cardiovascular disease management with a patient – facing agentic AI system that can provide 24/7 clinical care, as an extension of the clinical care team.

TECHNOLOGY FOCUS AREAS

- Development of a patient-facing AI agent for patients with heart failure, or those who are post-myocardial infarction.
- Development of an independent supervisory agent to ensure continued agent accuracy and patient safety, particular after clinical deployment.
- Recruitment of health systems for the design, development, and deployment of the technologies.

Clinical Agentic AI Can Enable CVD Care Transformation



Current State		With Clinical Agentic AI
Outpatient care is sporadic and reactive.	ACCESS	All Americans have 24/7 access to *specialty care.*
Low-value care leads to excess costs & harms.	CARE QUALITY	Unlocks the full potential of the Electronic Health Record (EHR) and wearables.
Waste of clinical resources with administrative tasks.	EFFICIENCY	Clinicians practice at the top of their license.

ARPA-H

Approved for Public Release: Distribution Unlimited

Vision: *Advocate* for Patients with Chronic Disease



Administrative/Wellness Functions

Advanced Decision Support

High Risk Functions



CARE NAVIGATION

- Manages appointments, tests, procedures and reminders.
- Estimates out-of-pocket costs based on insurer information.

DIET & EXERCISE

- Uses multimodal inputs to provide lifestyle recs.
- Provides 24/7 access to personalized nutrition and physical therapy.

TREATMENT

- Provides personalized treatment options based on patient inputs.
- Incorporates entire corpus of EHR and wearable data.

DIAGNOSIS

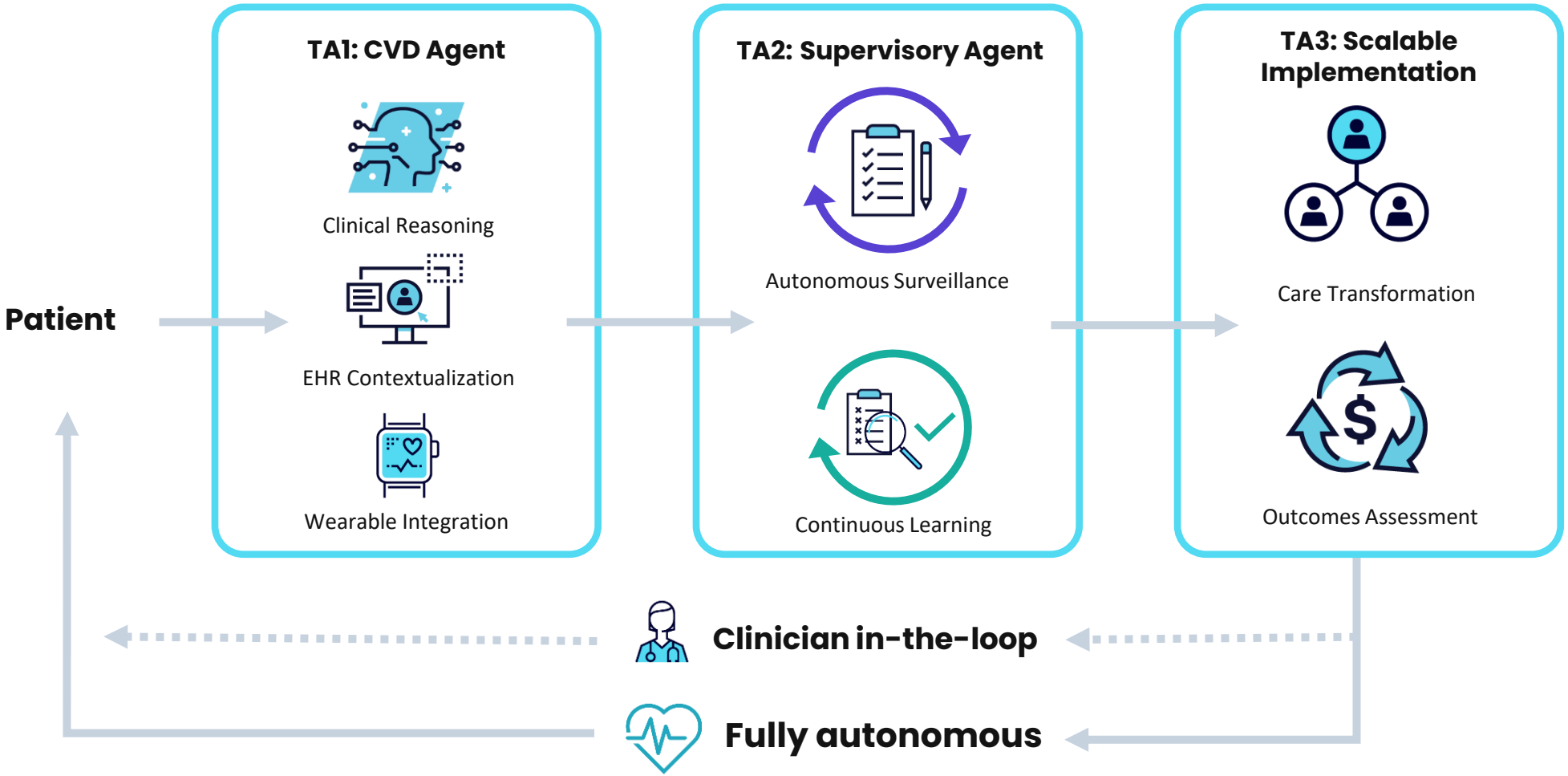
- Leverages patient inputs, wearable and EHR data to provide differential diagnoses.
- Engages clinical team with new diagnostic information.

RX MANAGEMENT

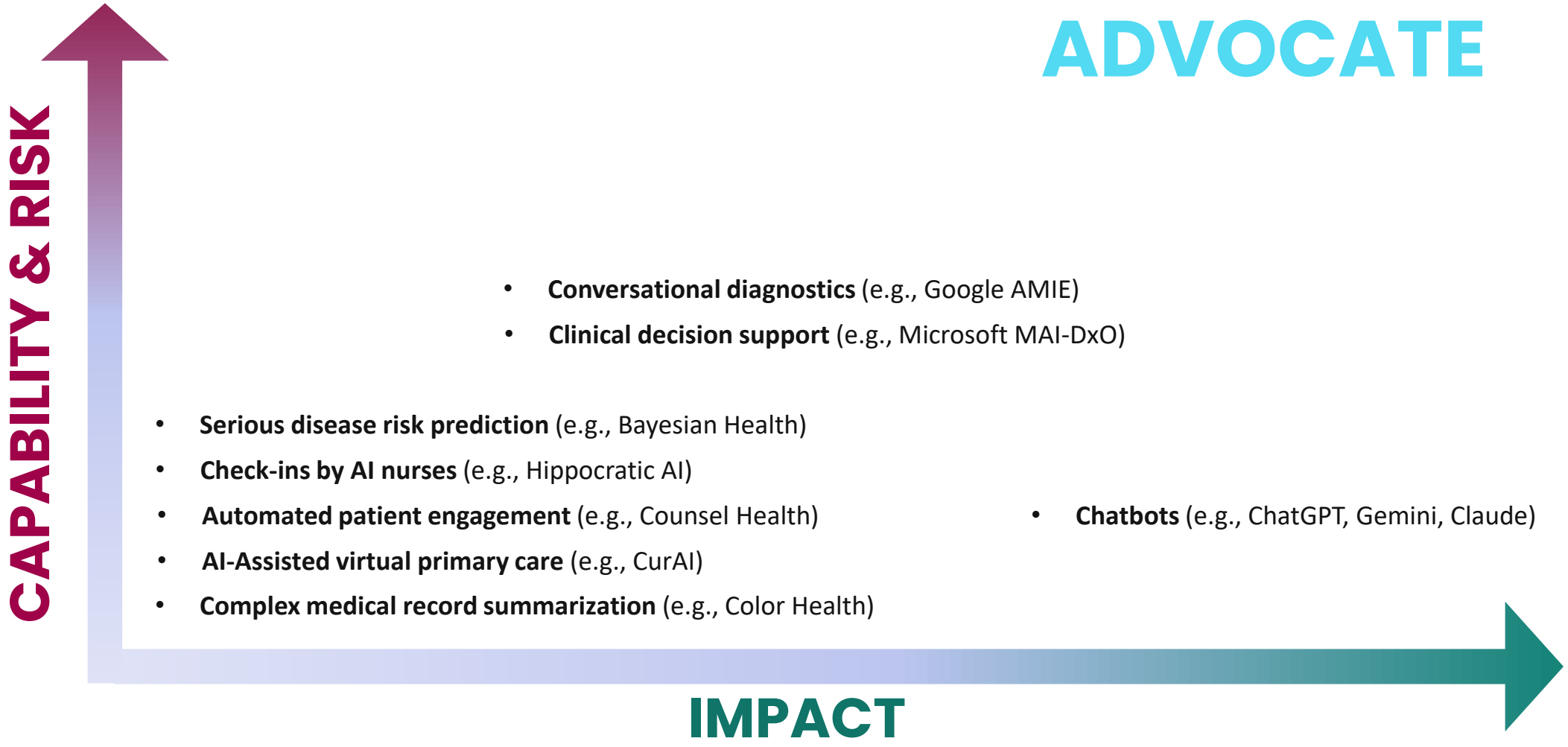
- Changes, renews and refills existing medications.
- Orders new prescriptions.

ARPA H

The ADVOCATE Vision: First-ever FDA-Approved Clinical Agentic System and Post-Market Safety Monitoring Technology

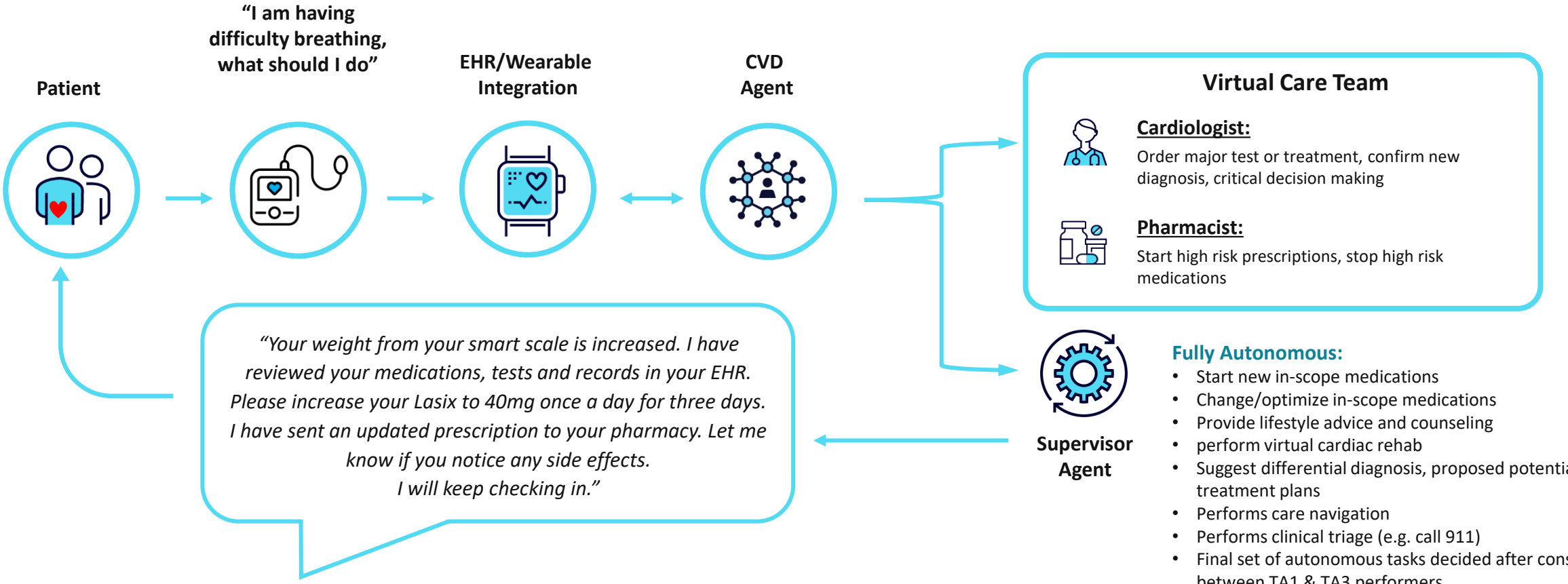


The State of the Clinical AI Industry



TA1: Proposed Workflow of CVD Agentic System



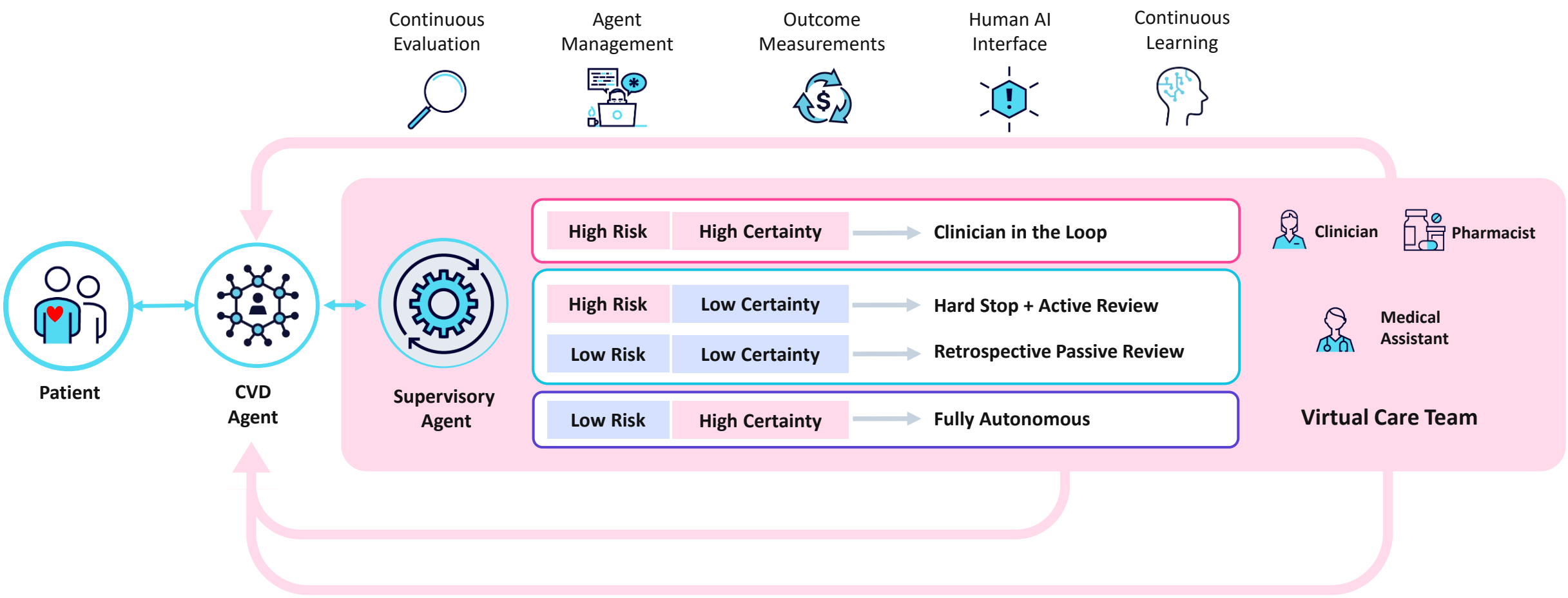


Fully Autonomous:

- Start new in-scope medications
- Change/optimize in-scope medications
- Provide lifestyle advice and counseling
- perform virtual cardiac rehab
- Suggest differential diagnosis, proposed potential treatment plans
- Performs care navigation
- Performs clinical triage (e.g. call 911)
- Final set of autonomous tasks decided after consensus between TA1 & TA3 performers

TA2: Proposed Workflow of Supervisory Agent

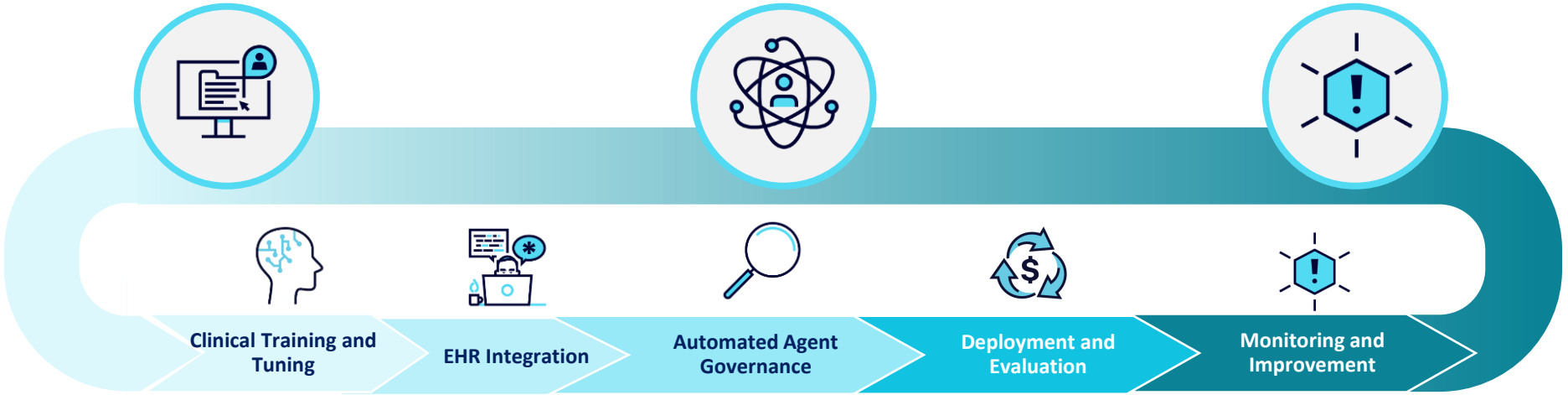




Approved for Public Release: Distribution Unlimited

TA3: Scalable Implementation

Goal: Integrate, Deploy, and Evaluate Agentic System in Large Health Systems



Current State

No Agentic System providing clinical care has been trained, deployed or evaluated in any clinical setting.

Scalable solutions for evaluation, governance, human-AI teaming and reimbursement are lacking.

ADVOCATE Approach

Enroll health systems that vary in business model (Risk-based, Fee-For-Service, VA), clinical configuration (primary care, multispecialty), EHR vendor, and geography (urban, rural).

Health systems offering greatest seamless data access, resources for TA1-2 development, diverse clinical sites for deployment, etc., at the best value will be recruited.

Innovation

First Agentic system deployment in integrated health systems creating novel and generalizable pathways for reimbursement, clinical trials, post-market evaluation, and care delivery for Agentic AI.

Impact of ADVOCATE on Agentic AI Regulatory Pathway

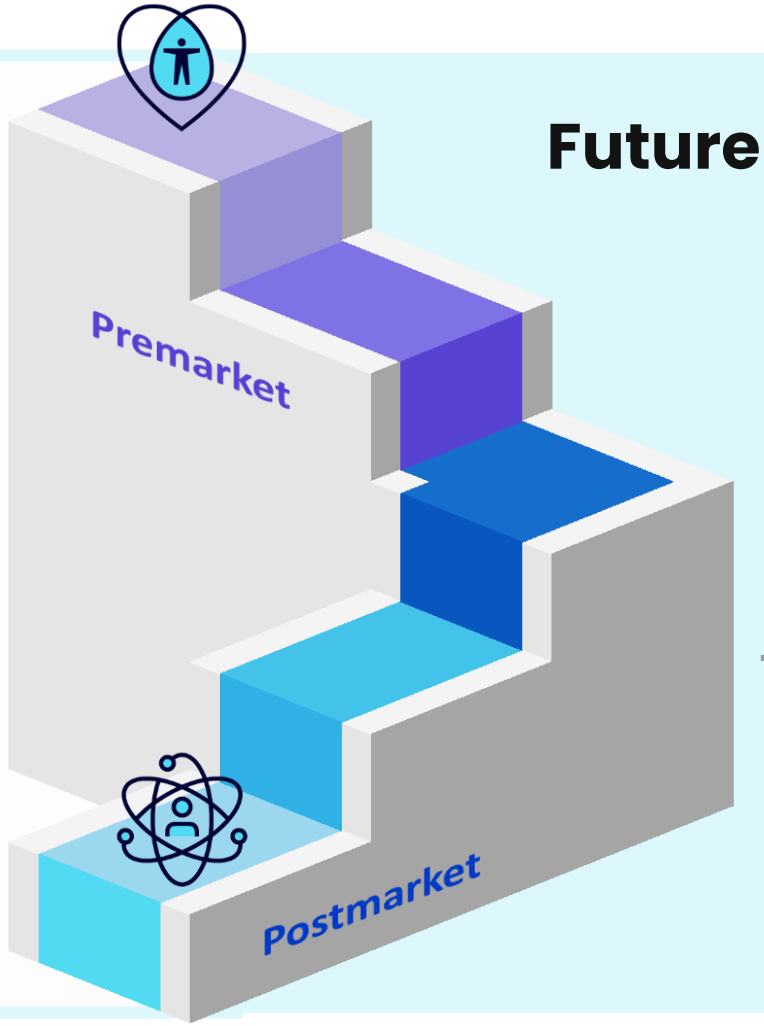
Current State

Premarket

- Need safety and efficacy evaluation
- Extensive clinical testing including clinical trials required
- Mandatory facility inspection
- *Costs*: $540,783 for FDA review, $54 million in total costs per device, $522 million if also accounting for failed approvals and capital costs (Ref: FDA.gov, Sertkaya et al, JAMA Open, 2022)

Postmarket

- No standardized, interoperable tool or pathway; no FDA guidance
- Manual review of AI outputs cost and labor prohibitive



Future Post-ADVOCATE State

Premarket

- Only need substantial equivalence to predicate device
- Minimal clinical testing
- No inspection required
- Costs: $24,335 for FDA review, at most few million in total costs (Ref: FDA.gov)

Postmarket

- Standardized, scalable disease agnostic Medical Device Development Tool qualified for postmarket use by FDA for all Clinical Generative or Agentic AI Applications

TA1: Progress Tracking Metrics



TYPE	METRICS	Phase 1A	Phase 1B
EFFECTIVENESS	Non-inferiority vs cardiologists for diagnosis, treatment, empathy, and reasoning	W/ Synthetic Data	W/ Simulated Patients
	Appropriateness of Agentic Actions (care navigation, prescriptions, patient data processing, etc.)	>85%	>95%
	Appropriateness of Triage	>85%	>95%
	Serious error rate	<5%	<3%
PATIENT EXPERIENCE	Proportion of interactions where clinical Agent effectively adapts to patient preferences	>90%	>95%
	User Acceptance Test (UAT) score	>85	>95
DATA INTEGRATION	Effective selection of relevant records (pass rate in manual review of sample)	>93%	>95%
	Ability to process FHIR and Health Level 7 version 2 (HL7v2)records (ratio of successful/failed transactions)	>95%	>97%
	Data summarization proficiency (reduction of token count of processed records)	>40%	>50%
	Summarization quality score (retention performance with Claim Recall method e.g., DocLens)*	>80%	>85%
	Proportion of errors auto-corrected (count of fixes/total error events)	>85%	>90%
USABILITY	Number of supported EHR vendors	1+	2+
	Number of integrated wearable or Remote Patient Monitoring (RPM) devices	1+	2+
EFFICIENCY	Time to complete record processing (seconds/typical medical record)	<2s	<0.1s
	Response latency (round-trip time of query-response in audit logs)	<3s	<1s
	Computational cost (cost/processed record)	<$0.50	<$0.05

TA2: Progress Tracking Metrics

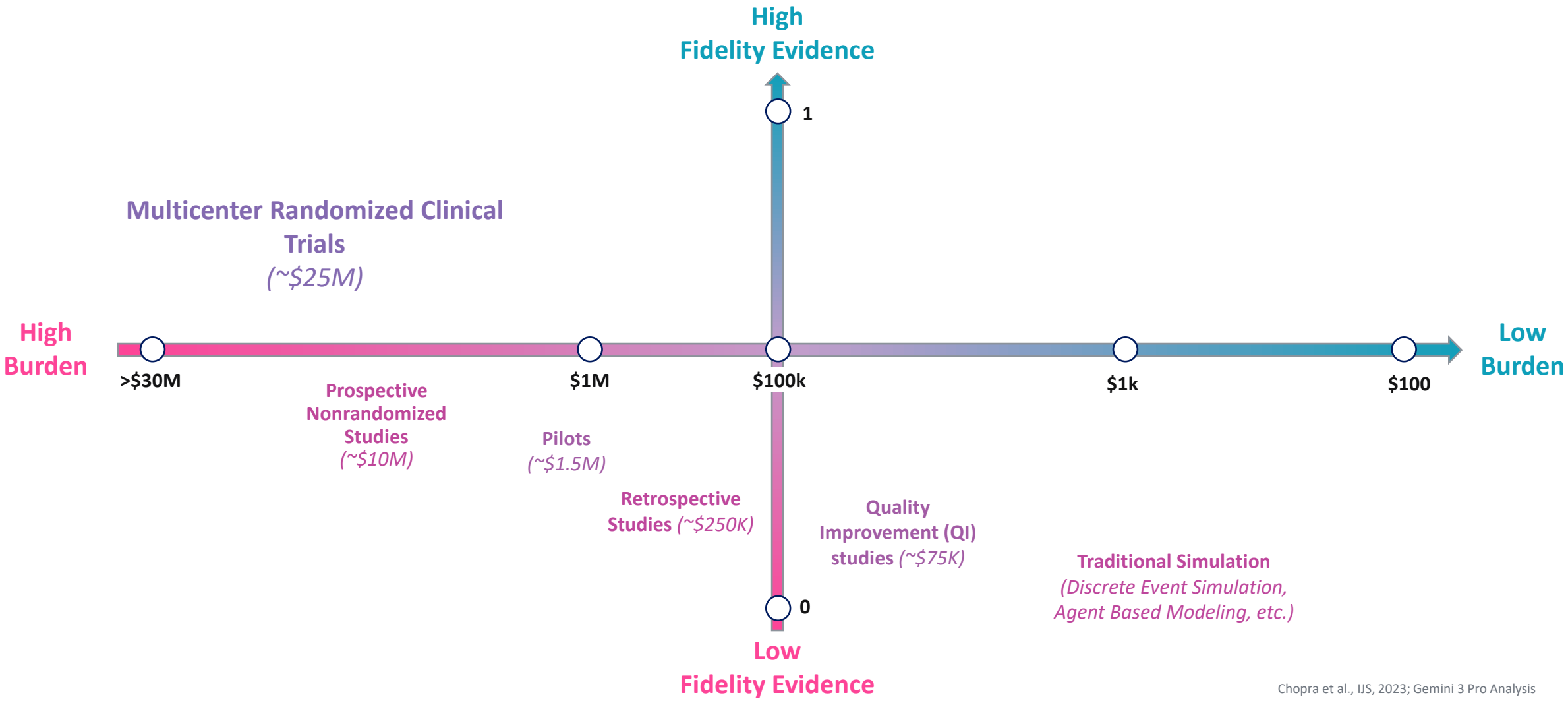


TYPE	METRICS	Phase 1A	Phase 1B
EFFECTIVENESS	Ability to ascertain accuracy of clinical agent	>85%	>95%
	Recognition of high-quality treatment recommendations	>85%	>95%
	Quantification of hallucination rate	>85%	>95%
	Determination of uncertainty of clinical inference	>85%	>95%
	Determination of clinical acuity and intervention risk	>90%	>97%
	Evaluation of appropriateness of task assignment to clinical care team members	>90%	>97%
USABILITY	Clinician app User Acceptance Test (UAT) score (0-100)	>80	>90
	Response latency (round-trip time of query-response in audit logs)	<3s	<1s
	Quality score (0-100) of summarized information provided to the clinical team	>80	>95

Clinical Evaluation Strategy for ADVOCATE Technologies

- **Phase 1A**
  - Human Factors Evaluation
  - Virtual simulated heart failure patients
- **Phase 1B**
  - Shadow mode evaluation in clinically deployed settings
- **Phase 2**
  - Stepped-wedge multicenter cardiovascular outcomes trial

Evidence Generation for Clinical AI



Chopra et al., IJS, 2023; Gemini 3 Pro Analysis

Clinical AI evidence generation is not fit for purpose

Problem – Without a fit-for-purpose evaluation pathway, the U.S. has fallen behind in clinical AI



AI can deliver care to every American, but none reach patients fast enough



$54M and 3–7 years per device, with 90% studies retrospective and 87% biased



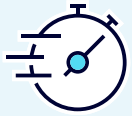
Americans who need AI most wait longest with delays leading to lives lost

Evidence is obsolete by trial completion and study publication

Solution – A simulation platform for in-silico AI trials via a novel causal world model and dataset



Millions of cases, simulated in hours with higher fidelity than average study



FDA authorization in months, not years with in-silico trials for all SaMD



Safe, effective AI reaches everyone, everywhere, for every condition

Continuous evaluation appropriate for continuously evolving technology

SaMD: Software as a Medical Device

Key Limitations that Make Healthcare World Simulation Hard

Limitation 1



Foundation models cannot provide causal inference

Existing models absorb clinical notes as a proxy, flattening sequential ordering. Causality is lost in translation.

Limitation 2



Existing data unable to train causal models for healthcare

Existing EHR data carries unmeasured confounders, lacks clinical intention, and omits temporal sequencing

Limitation 3



Evaluation is one-shot with unrealistic simulated patients

Trials obsolete by completion; Patient personas or actors are not realistic, recruiting patients challenging

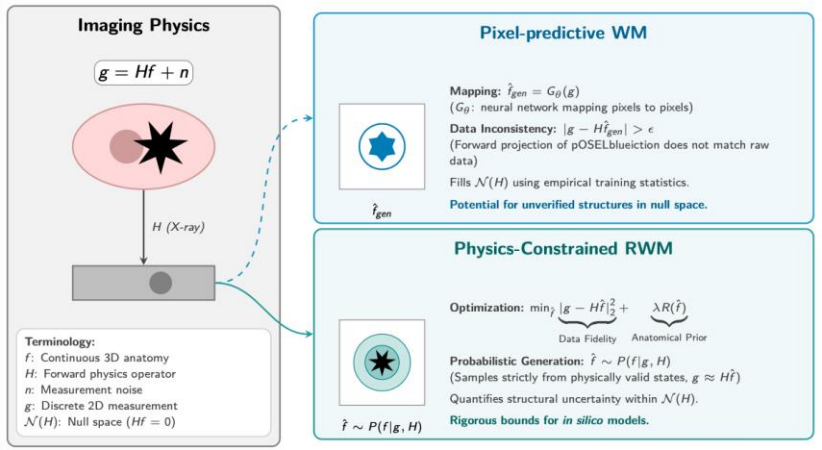
Enabling Technologies – World Models

ENABLING TECH	LABS	STRENGTHS	WEAKNESSES
Joint Embedding Predictive Architecture (Latent Predictive)	AMI Labs	✅ for high-dimensional data, ignores noise, fast	❌ lacks perceptual grounding, inauditable
Predictive Analysis/ Action Network (Generative State)	CMU	✅ at simulating multistep sequences	❌ prone to hallucination, high compute
Interactive, Promptable World Models (e.g., Genie 3, RTFM)	Deep Mind, World Labs	✅ at simulating promptable interventions	❌ physiologic or healthcare training

***Novel Advances**:*

- *Subquadratic long-context models e.g., Mamba,*
- *Causal Mechanistic Models; Physiology-Integrated Neural Networks*
- *Computational resources (e.g., Genesis Mission)*
- *DARPA VITAL (Virtual-Integrated Twin for Autonomous Lifesaving)*

Physics-Constrained World Models Perform Better than Pixel Predictive



Badano & Warraich, et al, RSNA, 2026

AMI: Advanced Machine Intelligence
CMU: Carnegie Mellon University
RTFM: Real-Time Frame Model

Qazi et al, NeurIPS, 2025; Wornow et al, ICLR, 2025

Enabling Technologies – Times Series Foundation Models

MODEL	TOKENS AND TRAINING	PARAMETERS	VOCABULARY
Delphi-2M	5.9M, from 500,000 patient timelines from the UK Biobank	2 million	1,270
ETHOS	329M from 241,015 patient timelines from MIMIC-IV	124 million	4,367
CEHR-XGPT	500M from 2.6 million patient timelines from Columbia	130 million	9,994
Curiosity	151B from 118 million patient timelines from 310 health system	1 million	7,105
GPT-3	300 billion from web-scale, diverse internet texts	175 million	50,257
InstructGPT	300B from web-scale, diverse internet texts + alignment sets	GPT-3 family	50,257

Medical event foundation models are massive but have hit the limits of scaling. Further accuracy in predictive power and causal inference is not possible with existing models, using existing data.

Bedi and colleagues, Nature Medicine, 2026

Conclusions and Next Steps

- We are looking for teams to send us one-pagers to share:
  1. How might they help develop models or approaches that could be used to develop a simulation platform that can provide regulatory grade evidence for novel agentic AI technologies?
  2. How might a dataset and live API be created to help train and maintain such a simulation or model?
  3. How might they develop realistic, virtual patients with specific clinical phenotypes, that can be used to evaluate large language models and AI agents?
- Share both challenges and innovative approaches
- Contact Haider Warraich (haider.warraich@arpa-h.gov), Seyyed Ali Mottaghi (ali.mottaghi@arpa-h.gov) and ai.simulation@arpa-h.gov


ARPA H