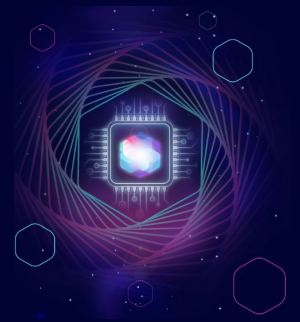
Welcome to PRECISE-AI Proposers Day!

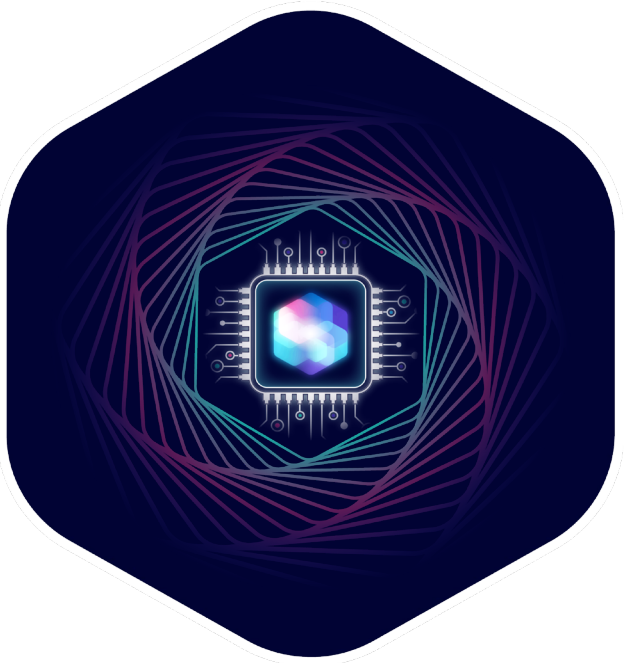


Agenda:

9:30 AM CT	Fireside Chat with Dr. Susan Monarez	Dr. Susan Monarez, ARPA-H Deputy Director
9:50 AM	PRECISE-AI Program Overview	Dr. Berkman Sahiner, Program Manager
10:35 AM	Recent Advances in AI for Abdominal Radiology (virtual)	Dr. Ron Summers, NIH/Clinical Center
10:50 AM	FDA Perspective: Monitoring of AI-Enabled Medical Devices (virtual)	Dr. Ghada Zamzmi, FDA/CDRH
11:05 AM	**Break**	
11:20 AM	Acquisition Information	Ms. Adrea Robinson, BID Agreements Officer (AO)
11:40 AM	Project Accelerator Transition Innovation Office (PATIO) Overview (virtual)	Dr. Maryam Ziaei, ARPA-H – T3X Division Director
12:00 PM	Closing Remarks	Dr. Berkman Sahiner, Program Manager
12:05 PM	**Lunch**	
1:30 PM	Q&A Response	PRECISE-AI Team
2:00 PM	**Adjourn Formal and Virtual Program**	
2:00 PM	PRECISE-AI PM Sidebars (in-person only)	Program Manager and PRECISE-AI Team
2:00 PM	Poster Session/Networking (in-person only)	Networking Event
5:00 PM	**Adjourn In-Person Event**	

PRECISE-AI

Performance & **R**eliability **E**valuation for
Continuous mod**I**fications and u**SE**ability of **AI**

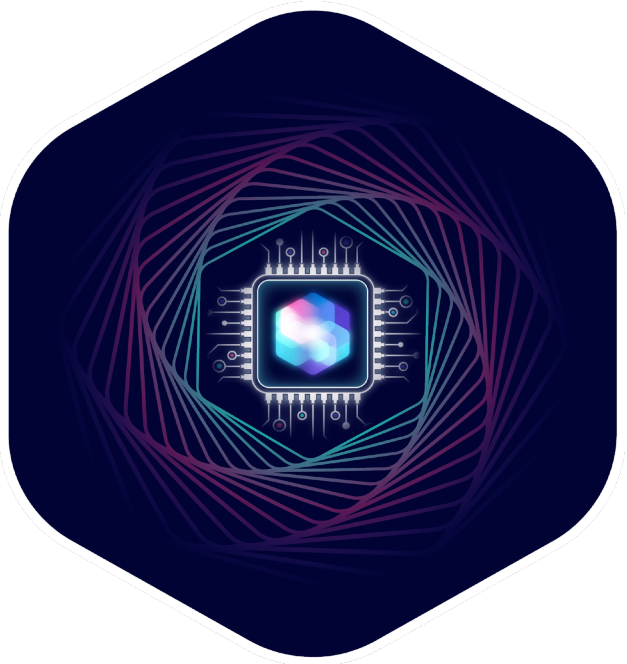


Fireside Chat with Dr. Susan Monarez
Deputy Director, ARPA-H
Oct. 17, 2024

ARPA-H

PRECISE-AI

Performance & **R**eliability **E**valuation for
Continuous mod**I**fications and u**SE**ability of **AI**



Berkman Sahiner, PhD, Program Manager
Resilient Systems Office (RSO), Mission Office (MO)
Oct. 17, 2024

ARPA-H

AI usage in clinical decision support is increasing

- 10X increase in the number of FDA-cleared AI-enabled medical devices in the US from 2018-2023[1]
- 23% estimated annual revenue growth for AI in medical diagnostics ($3.7B by 2028)[2]
- 33% annual growth expected in radiological use cases (76% of overall cleared AI-enabled devices)[3]
  - In 2020, 33% of radiologists stated they used AI[4]
  - 20% of those not using AI plan to purchase AI tools in the next 1-5 years
- 39% annual growth expected in cardiovascular use cases (10% of overall cleared AI-enabled devices)[5]

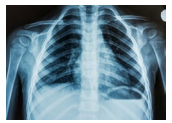
[1]fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices
[2]marketsandmarkets.com/Market-Reports/artificial-intelligence-medical-diagnostics-market-22519734.html
[3]managedhealthcareexecutive.com/view/ai-in-medical-imaging-market-expected-to-increase-to-14-2-billion-by-2032
[4]jacr.org/article/S1546-1440(21)00293-3/pdf
[5]market.us/report/ai-in-cardiology-market/

ARPA-H

AI models are developed in controlled and static environments
but real-world clinical environments are highly dynamic

PRE-MARKET

POST-MARKET



Clinical test / procedure

AI input data

AI

Tested & validated on input data. Basis for model performance: manually extracted ground truth

Deploy

FDA

Changes in input datasets (data shifts) are not detected or considered by AI

- Patient Population
- Clinical Behavior
- New research
- Patient mgmt. guidelines
- IT Infrastructure
- Data Acquisition

AI

AI performance degrades, increasing misdiagnoses by AI and causing mistrust by clinicians

ARPA H

What if AI models in healthcare auto-corrected to maintain peak model and clinician performance?

ARPA-H

Clinical AI model performance degrades over time, threatening patient safety

https://www.nature.com/articles/s41598-022-15245-z

Temporal quality degradation in AI models

Daniel Vela, Andrew Sharp, Richard Zhang, Trang Nguyen, An Hoang & Oleg S. Pianykh

www.statnews.com/2022/02/28/sepsis-hospital-algorithms-data-shift/

AI gone astray: How subtle shifts in patient data send popular algorithms reeling, undermining patient safety

By Casey Ross Feb. 28, 2022

Data analysis by Adam Yala, Janice Yang, and Ludvig Karstens — Jameel Clinic, Massachusetts Institute of Technology

91% of ML Models Degrade Over Time

MLOps | Model Monitoring

www.fiddler.ai/blog/91-percent-of-ml-models-degrade-over-time

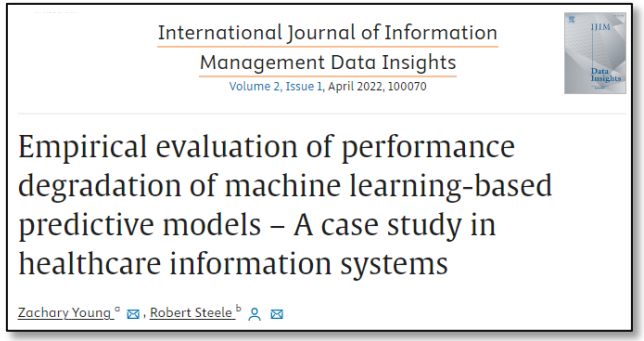
https://www.scientificamerican.com/article/yes-ai-models-can-get-worse-over-time/

AUGUST 2, 2023 | 6 MIN READ

Yes, AI Models Can Get Worse over Time

More training and more data can have unintended consequences for machine-learning models such as GPT-4

www.sciencedirect.com/science/article/pii/S2667096822000143



International Journal of Information Management Data Insights

Volume 2, Issue 1, April 2022, 100070

Empirical evaluation of performance degradation of machine learning-based predictive models – A case study in healthcare information systems

Zachary Young [a], Robert Steele [b]

ARPA-H

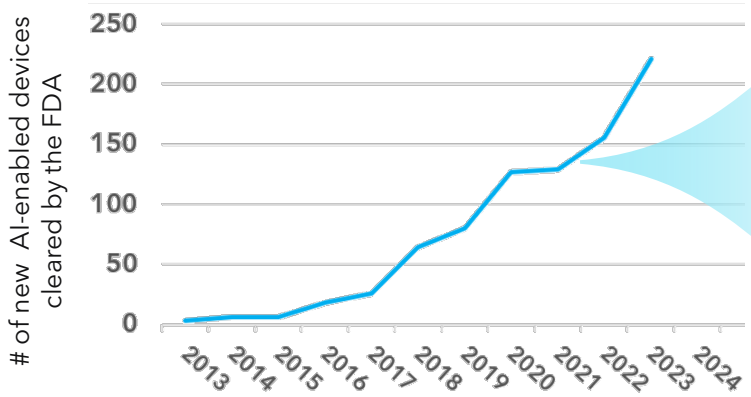
Current post-market evaluation strategies are not timely or scalable

- **0%** of current clinical AI models receive **regular** testing and updating during clinical use to ensure accuracy of output
- Despite the proliferation of medical AI clearances/approvals, little is known about their **real-world usage** [6]
- **No requirements** to monitor and update clinical AI models during clinical use due to **lack of technical solutions**
- 84% of clinicians say safety and efficacy validation is an important attribute for validation[7]
- **Clinician intuition** is the primary mechanism of detecting model degradation. **Harm** may have occurred before degradation is detected
- **No clinically validated methods currently exist** to monitor medical AI model performance as mandated by the 2023 Executive Order. The FDA is developing a regulatory path for post-market AI model updates, but current technologies from platform vendors do not include degradation detection or auto-correction capabilities

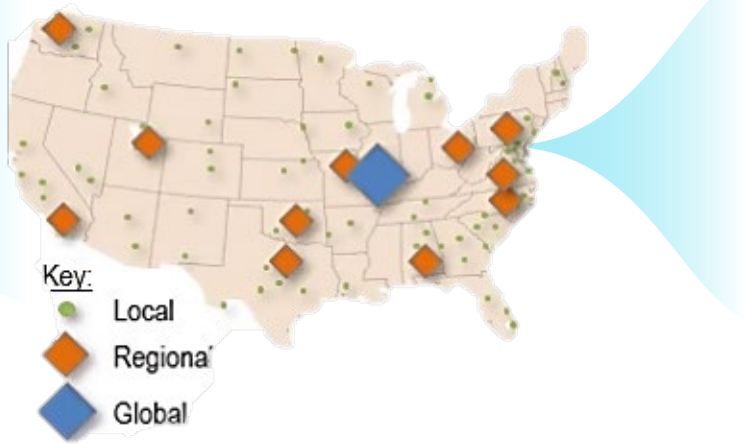
[6] ai.nejm.org/doi/full/10.1056/AIoa2300030
[7] https://www.ama-assn.org/system/files/physician-ai-sentiment-report.pdf

AI Decision Support Tools (AI-DST) can behave differently in each location and deployment context.

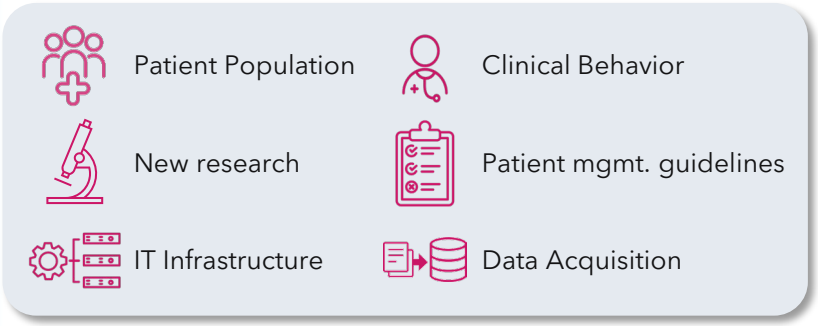
Increasing # of FDA-cleared AI-enabled medical devices

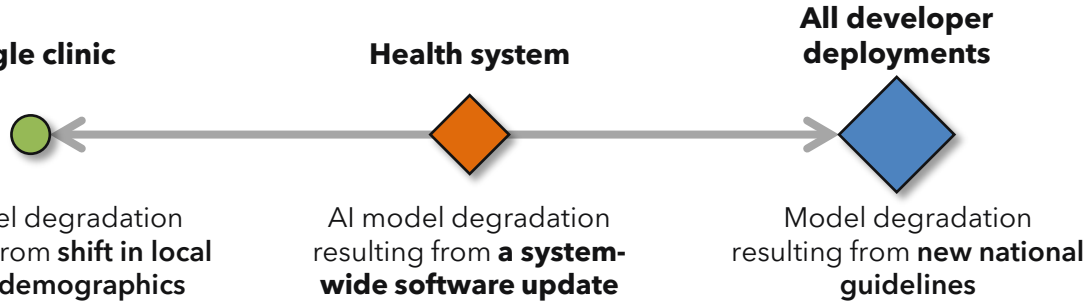


Increasing number & diversity of deployment locations for each AI-DST



Continuously changing deployment contexts at each location





Current manual approaches are incapable of addressing performance degradation at scale.

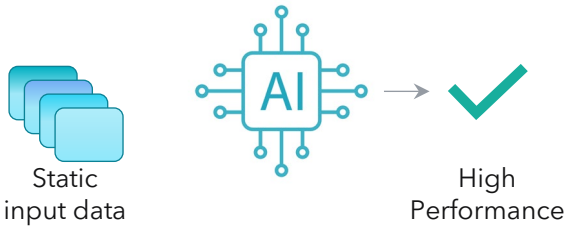
ARPA-H

Since AI model degradation varies across clinical environments, **automation is needed to maintain performance in all locations**

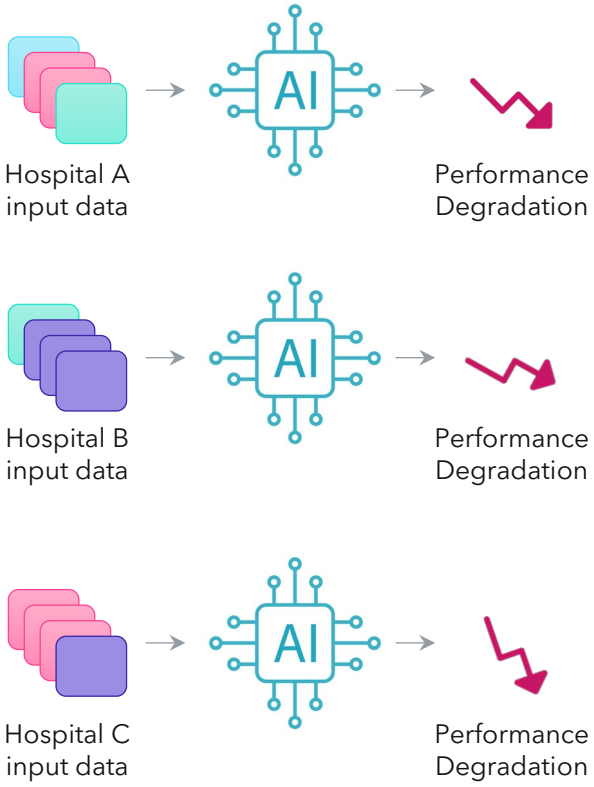
PRE-MARKET

POST-MARKET

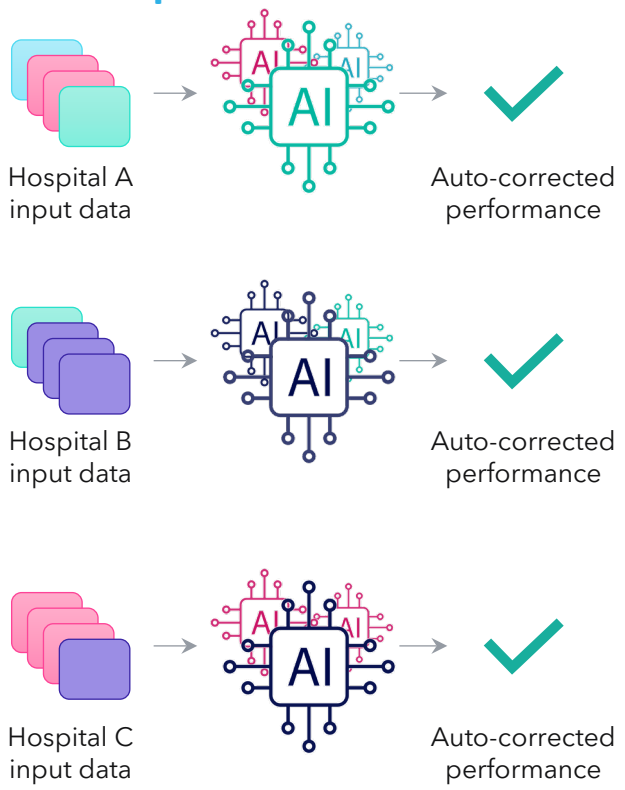
High performance in a controlled environment



TODAY: Performance may degrade differently across hundreds of clinical environments



PRECISE-AI: Tools will **auto-detect degradation** and **auto-update** to **maintain performance**

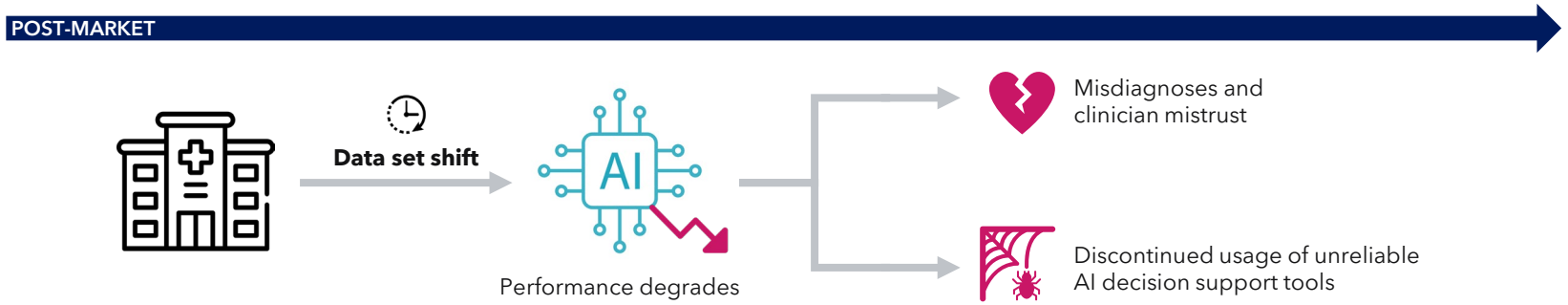


ARPA H

Today's post-market strategies are slow, inconsistent, and manual. **PRECISE-AI's tools aim to fix that.**

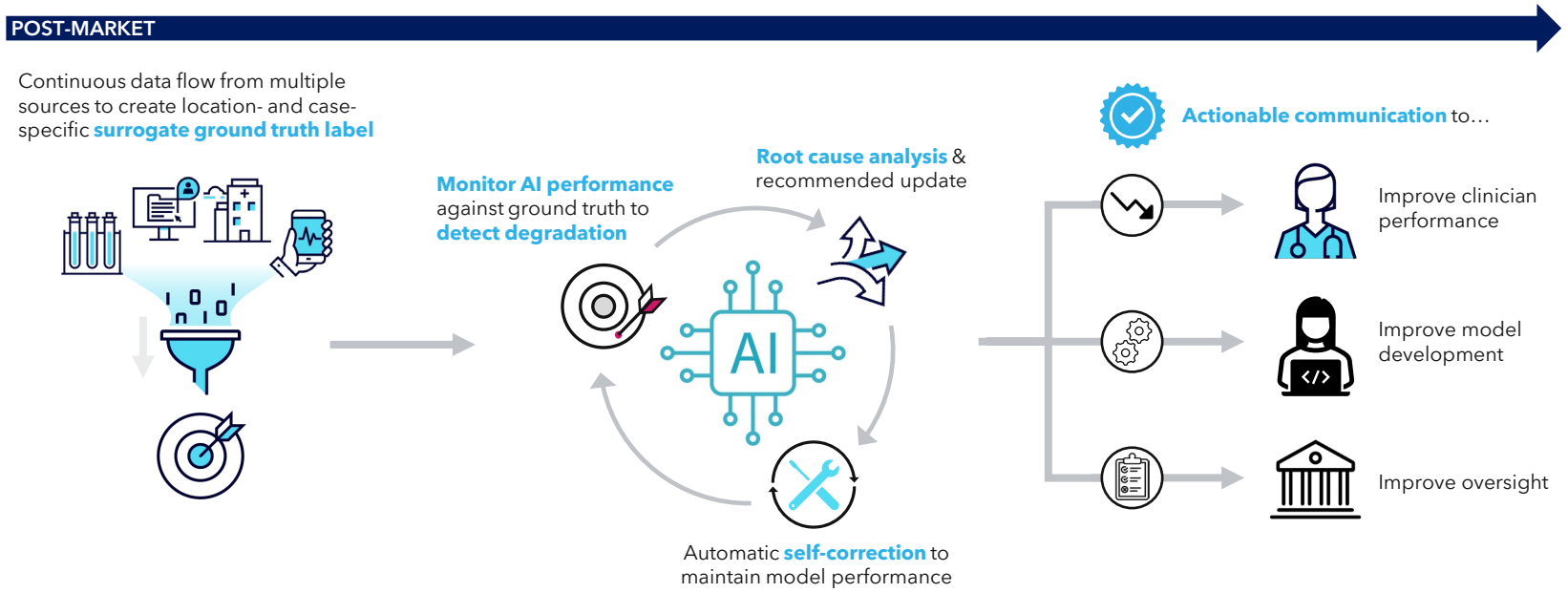
CURRENT STATE IN POST-MARKET

- Post-market assessment is **manual and sporadic** with **no scalable method** to detect or remedy performance degradation
- No scalable method to establish what the "correct AI output" should have been, which is the first step measure the day-to-day performance of the AI model
- Model performance is self reported by developer if done at all. Extent of model degradation in real clinical use is **unknown**
- Model (un)certainty is **not clearly communicated** to clinicians, eroding trust



PRECISE-AI

- Basis for model performance, i.e., the ground truth label, is **auto-extracted** from clinical data
- **Continuous monitoring** of performance degradation during clinical use
- **Auto-identify root cause(s)** and **automatically self-correct** degraded models. No need for AI model experts at every clinical site
- **Clear & actionable communication** shared with clinicians, developers, and FDA about the issue, cause, and correction
- Clinicians can better interpret model uncertainty, calibrate their trust, and **use AI tools more effectively**



ARPA-H

PRECISE-AI seeks to deliver

A self-correcting, diagnostic, and transparent communication system to ensure clinical AI safety

TA1: Surrogate Ground Truth Label Extraction

- Continuously gather clinical test results, billing codes, and longitudinal records from multiple institutions
- Automatically estimate the true presence or absence of a condition (surrogate ground truth label) across clinical use-cases, which is used as a basis for performance evaluation in TA2.

TA2: Degradation Detection & Self-Correction

Continuously compare surrogate ground truth to AI output to detect changes in performance

Degradation Detection (TA2.1)

Root cause analysis (TA2.2)

Attribute degraded model performance to a root cause

Automatic self-correction tools (TA2.3)

Address model degradation by updating to the best performing of several competing options that are continuously running and adapting in the background

TA3: Quantify Uncertainty & Improve Clinician Performance

- Communicate uncertainty level of AI model and factors contributing to degradation
- Demonstrate evidence of improved clinical decision-making due to uncertainty in model outputs and bidirectional communication with the clinician

TA 4: Core Infrastructure

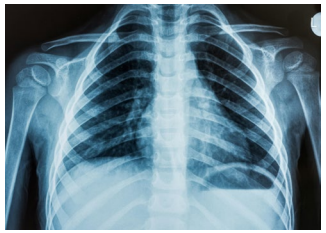
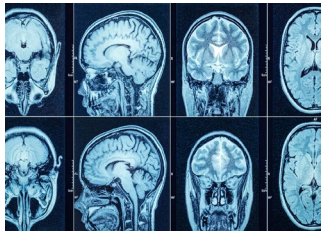
Enable interoperable data collection, sharing, analysis, and reporting, across key stakeholders

TA 5: Independent Verification and Validation (IV&V)

Assess performance for each TA and wholistically for the program through operationally relevant metrics

ARPA-H

Performers will select clinical use cases from priority tracks to increase the breadth and generalizability of innovations.

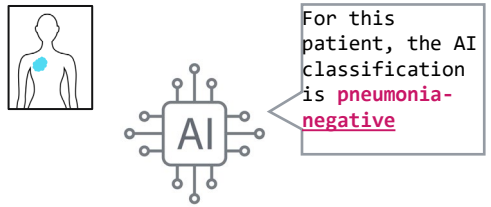
Clinical Track		Example Use Case
X-ray		Computer-assisted triaging tool for prioritizing chest X-ray images based on suspected presence of pneumothorax
		Clinical decision support tool to help clinicians assess Endotracheal Tube (ETT) placement on Chest X-ray images
CT		Clinical decision support tool to help clinicians detect acute appendicitis on thoracic CT images
		Computer-assisted triaging tool for flagging and communicating suspected Intracranial Hemorrhage (ICH) on non-enhanced head CT images
EHR		AI model for early prediction of significant events in critical care based on EHR data
		AI model to predict early mortality among critical fracture patients based on EHR data

ARPA H

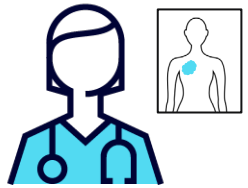
TA1: Surrogate Ground Truth Label Extraction

Objective: Automatically and continuously identify surrogate ground truth labels and metadata for each patient across a breadth of clinical use cases and clinics where AI models are used.

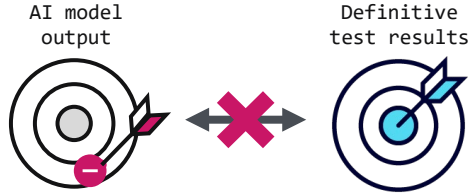
CURRENT STATE OF PRACTICE: The performance of AI decision support tools (AI-DSTs) cannot be monitored automatically because there is no scalable mechanism to establish a patient ground truth label (i.e. basis for evaluating model performance)



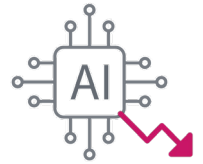
AI evaluates patient's x-ray images and classifies the patient as pneumonia positive or negative.



Clinician evaluates the patient's information and the AI output, recommends next steps for case management

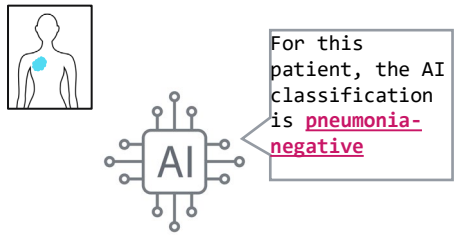


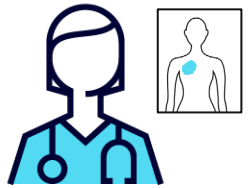
The patient got other tests that indicated they were **pneumonia-positive**. This information is **not** extracted from the patient's health data to compare with AI output.

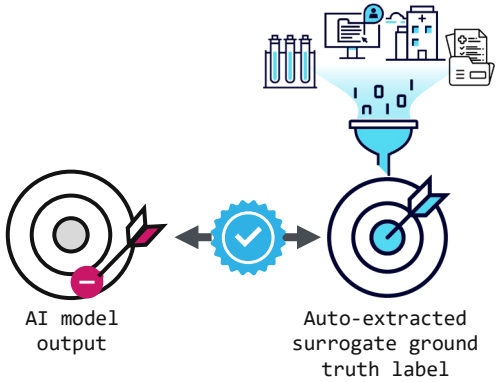


Incorrect AI output goes undetected, threatening patient health.

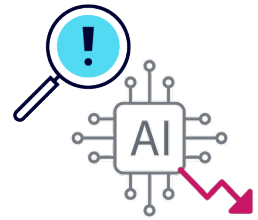
PRECISE-AI APPROACH: Automatically extract the **surrogate** ground truth label from the wealth of data that exists in patient records to establish a basis for continuously evaluating clinical AI model performance.







Patient's International Classification of Diseases (ICD) codes, infection-related symptoms, clinical reports, prescribed medication and other data fed into PRECISE-AI's ground truth label extraction system



Surrogate ground truth label is automatically produced indicating:

- Patient is **pneumonia-positive**
- AI output was incorrect
- Incorrect output needs to be considered in AI performance eval

ARPA-H

TA1 Metrics

Metric	**Phase I (0-24 mo.)** Prototyping & tool development	**Phase II (25-36 mo.)** Testing, improvement, & expansion	**Phase III (37-48 mo.)** Integration & product development
Breadth of clinical validation (# of clinical sites per performer)	3 sites	5 sites	8 sites
Automation of surrogate ground truth labels (% of cases where a surrogate ground truth label can be automatically determined based on clinical data - e.g., EHR/clinical report/imaging for retained patients)	≥ 60%	≥ 75%	≥ 90%
Agreement with EHR[1] (% agreement between data elements extracted with automated methods developed in PRECISE-AI versus manual extraction using the same EHR data on a randomly selected subset of cases)	≥ 95%	≥ 99%	≥ 99%
Agreement with manual ground truth[1] (% agreement of the automated surrogate ground-truth method and the manual ground-truth (determined by an expert panel of clinicians when needed) using all available data on a random subset of cases)	≥ 85%	≥ 90%.	≥ 95%
Data sharing (% of patient cases with a surrogate ground truth label and metadata that are shared with other performers)	≥ 70%	≥ 90%	≥ 95%

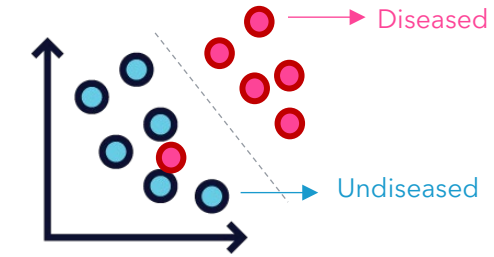
[1] For example, for a binary classification problem, agreement would be the average of positive percent agreement and negative percent agreement.

ARPA-H

TA2: Degradation Detection & Self-correction

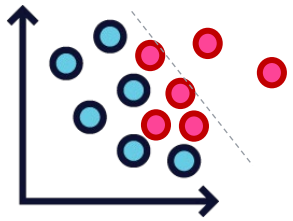
Objective: Automatically detect, identify root causes for, and correct AI-DST degradation to maintain peak performance.

CURRENT STATE OF PRACTICE: No standard method to quickly and accurately detect or correct performance degradation post-market.



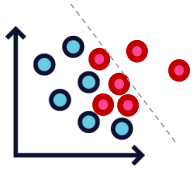
Before **Data Shift**: Model performing as expected

Changes in clinical settings



After **Data Shift**: Model sensitivity has dropped but no one is aware of the degradation or its cause

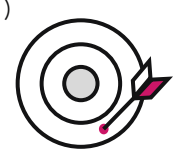
PRECISE-AI APPROACH: Model performance is continuously monitored, assessed, and improved.

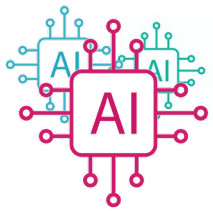


Alert: Model sensitivity has dropped

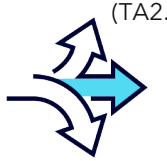
DEGRADATION DETECTED
(TA2.1)

e.g., statistical process control techniques

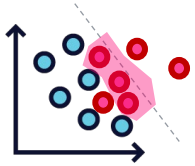




ROOT CAUSE DETERMINED
(TA2.2)



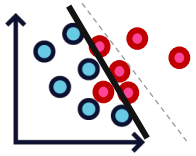
e.g., simulation modeling and signature analysis



Root cause: A change in the distribution of the disease-positive patients

AUTOMATIC CORRECTION
(TA2.3)

e.g., continual learning



Correction result: Sensitivity is brought back to non-degraded level

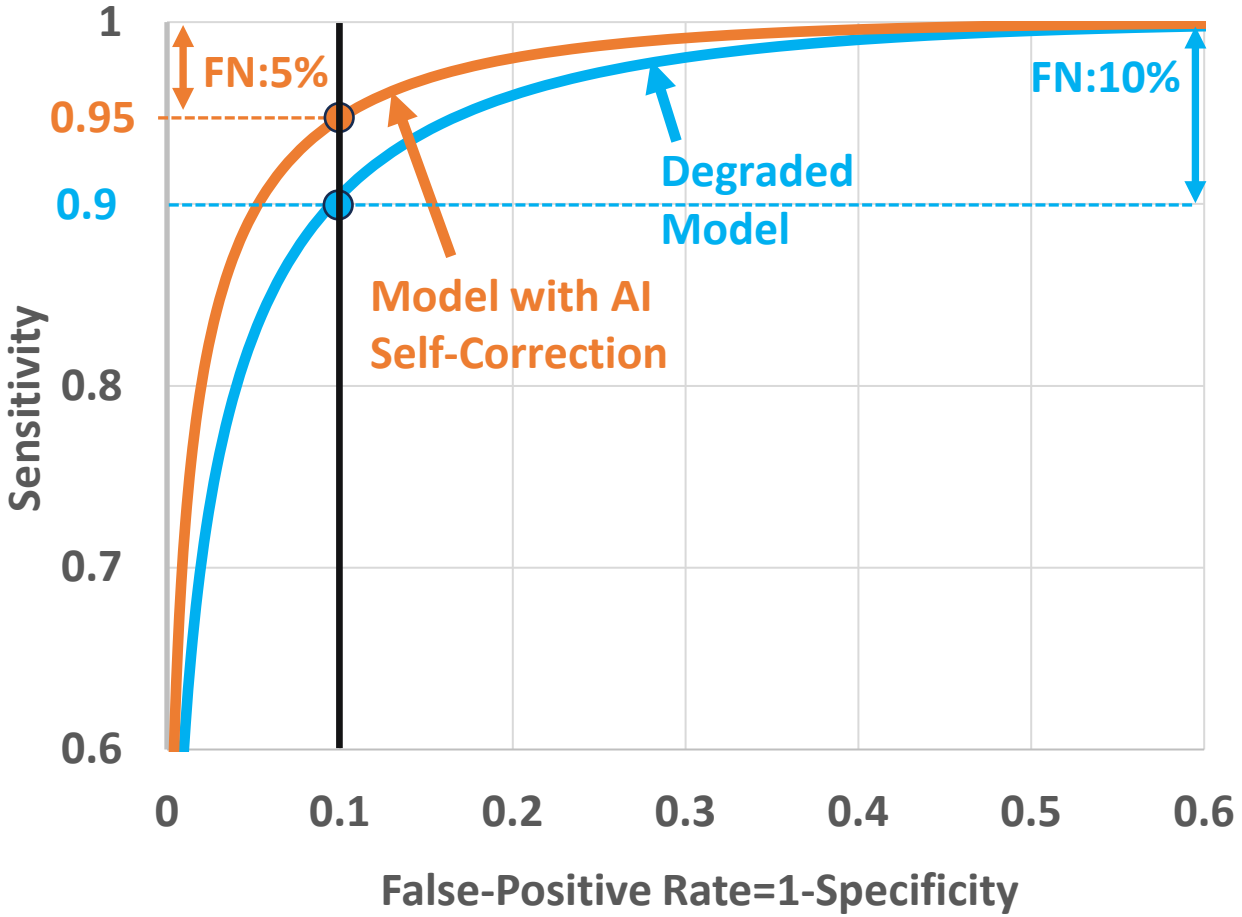
ARPA-H

TA2 Metrics

Metric	**Phase I (0-24 mo.)** Prototyping & tool development	**Phase II (25-36 mo.)** Testing, improvement, & expansion	**Phase III (37-48 mo.)** Integration & product development
Breadth of clinical use cases (# of clinical use cases per performer and degree of integration with PRECISE-AI tool suite)	2 clinical use cases	4 clinical use cases	4 clinical use cases including 2 integrated in the PRECISE-AI tool suite
Continuous assessment (% of clinical use-cases where continuous performance assessment is occurring)	≥ 50%	≥ 75%	100%
Detection sensitivity (% of correct degradation alerts following a 5% degradation in AI-DST model performance)[2, 3]	≥ 70%	≥ 80%	≥ 90%
Accuracy of root cause (% of degradation events[3] with a correctly identified root cause)	≥ 70%	≥ 85%	≥ 90%
Mitigation of model degradation[4] (% of addressable degradation events[3] where automated model update restores performance to original level [statistically non-inferior performance with a margin of 2.5%])	≥ 40%	≥ 55%	≥ 80%
Implementation rate (# of use cases where automated model degradation detection, root cause identification and automated model update are clinically implemented)	≥ 1 use case, pre-production quality	≥ 3 use cases, pre-production quality	4 use cases, production quality
AI self-correction performance[4] **in clinical use** (e.g., average % reduction in false-negatives at a given specificity compared to before correction)	≥ 40% reduction in ≥ 50% of use cases	≥ 50% reduction in ≥ 75% of use cases	≥ 70% reduction in 100% of use cases

[2] At a clinically relevant false alarm rate and time-to-correct alarm, to be specified by the performer in their proposal
[3] In a simulation environment, or in clinical use, or both
[4] For clinical use, when using ground truth determined both with the automated and the manual methods. For manual ground truth, the data set can include a randomly selected subset of cases

Example: 50% reduction of false-negatives at 90% specificity



ARPA H

TA3: Quantify Uncertainty & Improve Clinician Performance

Objective: Improve clinician trust in AI-DST and enhance clinician performance.

CURRENT STATE OF PRACTICE: Medical AI-DSTs fail to convey model certainty in a meaningful or actionable way resulting in inappropriate level of trust.

Often, model **uncertainty is not conveyed** at all.

AI → `Diagnosis is positive` *[No indication of certainty]*

Expressions of model uncertainty are **not interpretable or actionable**.

AI → `Uncertainty score = 3.6` *[Not actionable]*

Communication does **not provide critical context** on how & why the AI made its determination.

AI → `Lung mass is cancerous` *[No rationale]*

Uni-directional communication with clinician

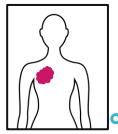
Clinician can **not** trust AI to assist their decision making

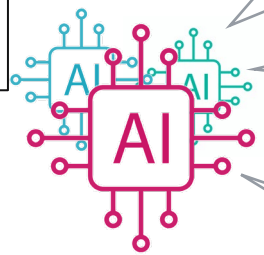
Misdiagnoses and clinician mistrust

Discontinued usage of unreliable / confusing AI-DSTs

PRECISE-AI APPROACH: Accurately quantify model uncertainty and convey rationale using easy-to-interpret and actionable communication tools.

Bi-directional communication with clinician





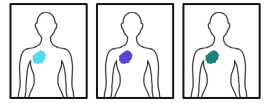
`The detected mass has an **85%** chance of being malignant.`

`What is that estimate based on?`

`This diagnosis is based on cases with **similar medical histories, demographics, and presentation.**`

`Show me the CT scans from several similar cases. Also, when were you last calibrated?`

`Sure, here are the **similar scans.** I was re-calibrated 3 days ago and am **performing at optimal** levels.`



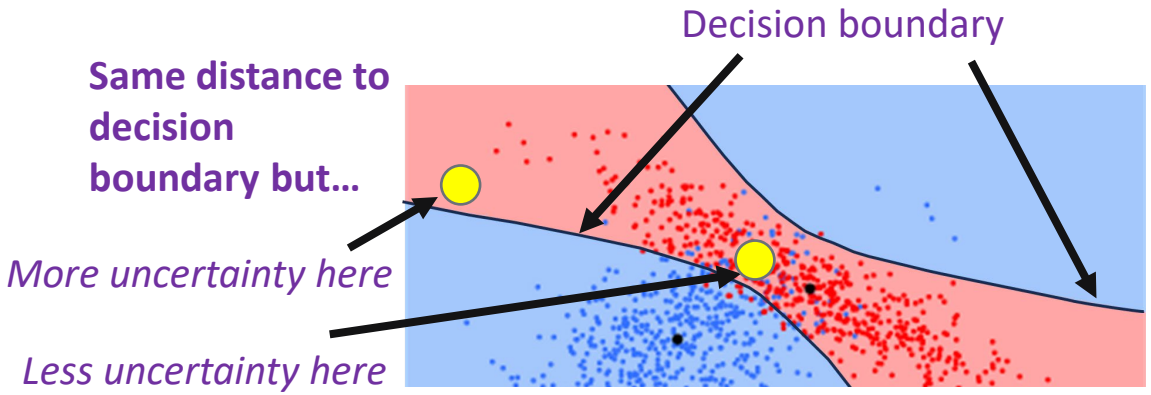
Improved clinical decision making due to how model uncertainty and other components are conveyed to the clinician

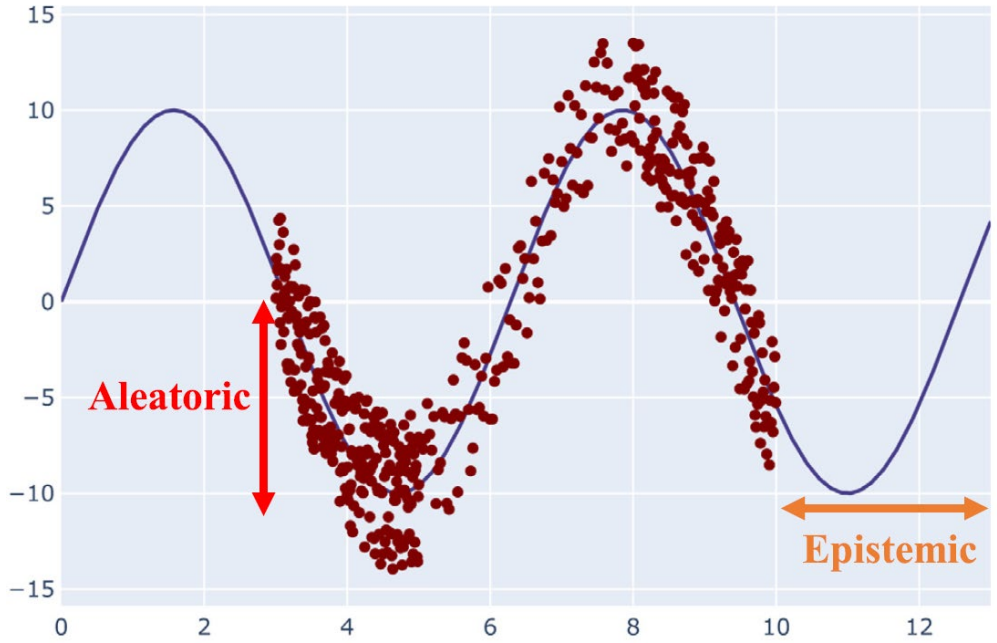
ARPA-H

Uncertainty Quantification

Different types of uncertainty quantification:

- 60% chance the patient has disease
- 60% chance the patient has disease, and the 95% confidence interval for disease is [55% - 65%]
- Lesion volume is 1.0±0.1 cm^3





ARPA H

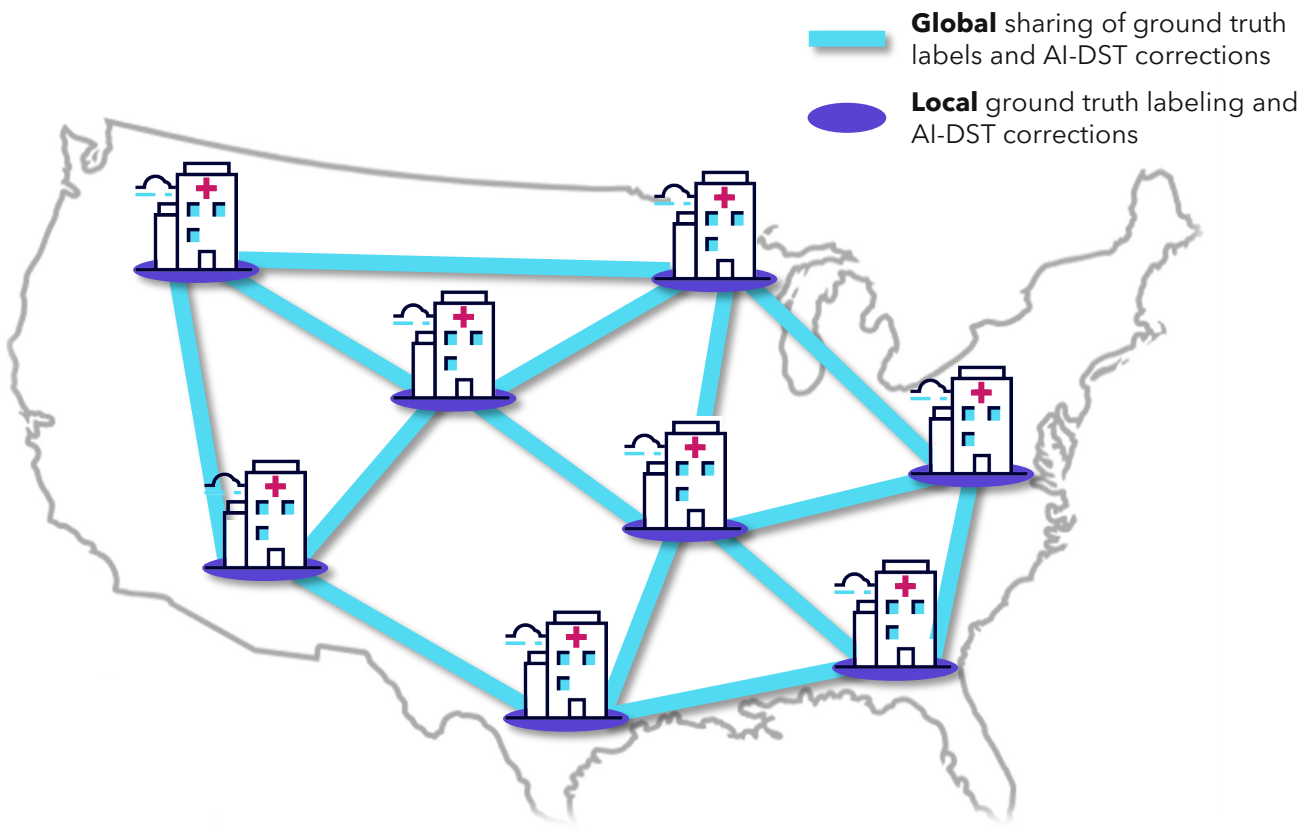
TA3 Metrics

Metric	**Phase I (0-24 mo.)** Prototyping & tool development	**Phase II (25-36 mo.)** Testing, improvement, & expansion	**Phase III (37-48 mo.)** Integration & product development
Performance of the uncertainty estimation method: Improvement in misclassification rate when abstaining from making a prediction based on the model score and uncertainty estimate [5]	≥ 15%	≥ 30%	≥ 45%
Useability (average Likert score [1-10 scale] for tool usability as judged by clinicians) of tools developed in TA3	Uncertainty communication tool: ≥ 7 2nd communication tool: ≥ 7	Uncertainty communication tool: ≥ 8 2nd communication tool: ≥ 8 3rd communication tool: ≥ 7	Integrated tool suite: ≥ 9
Clinician decision making improvement [6] (i.e., reduced provider misdiagnosis, e.g., % reduction in false-negatives at a given specificity) while using the AI-DST, comparing without versus with the use of PRECISE-AI tools	20% reduction with uncertainty estimation and TA2 tools in ≥ 50% of use cases	30% reduction with all tools in ≥ 75% of use cases	40% reduction with all tools in 100% of use cases
Clinical Efficiency [6] (Time for case interpretation with TA3 tools relative to no tools.)	1.2X	1X	0.8X

[5] At an abstention rate to be specified by the performer in their proposal
[6] Measured using observer performer studies

ARPA-H

TA4: Core Data Infrastructure



The PRECISE-AI Core Data Infrastructure will:

- **Aggregate analyses** and results across performers, clinical sites, and use cases to create a consolidated data sharing infrastructure to advance TA1, TA2, and TA3 technology development, including ground truth sharing among performers and reporting mechanisms.
- **Provide common data infrastructure capabilities** using existing data repositories and cloud services to facilitate data sharing and minimize costs.
- **Leverage and extend standardized data models** and common data standards for AI-DST input, metadata, output, ground truth, and other relevant data elements.
- **Ensure interoperability and usability** by adhering to FAIR data principles, providing APIs for data and model sharing, and creating comprehensive documentation.
- **Maintain privacy and security standards** to preserve data governance, chain of custody, patient privacy, and prevent unauthorized access to PHI.

The core data infrastructure will enable auto-correction of AI-DSTs in real-world clinical settings at each level (single clinic, health system and global level)

ARPA-H

TA4 Metrics

Metric	**Phase I (0-24 mo.)** Prototyping & tool development	**Phase II (25-36 mo.)** Testing, improvement, & expansion	**Phase III (37-48 mo.)** Integration & product development
Data types supported (proportion of data types requested by TAs 1-3)	≥ 90%	≥ 95%	100%
Sites supported (# of TA1 sites with API for data ingest)	All sites from Phase 1	All sites from Phases 1 & 2	All sites from Phases 1, 2 & 3

TA5: Independent Verification and Validation (IV&V)

The PRECISE-AI IV&V team will independently evaluate performers through a collection of program metrics. To do this they will:

- Work with clinical sites and all performers to appropriately collect and integrate needed performance data.
- Analyze performance data using appropriate metrics to derive an outcomes-targeted set of evaluation results.
- Provide statistically valid confidence intervals on program metrics so that quantitative outputs clearly map to qualitative program outcomes.
- Provide feedback for performers and ARPA-H on program metric evaluations at appropriate intervals.
- Provide clear communications on the entire metric evaluation pipeline.
- Oversee adherence to data engineering and interoperability standards.

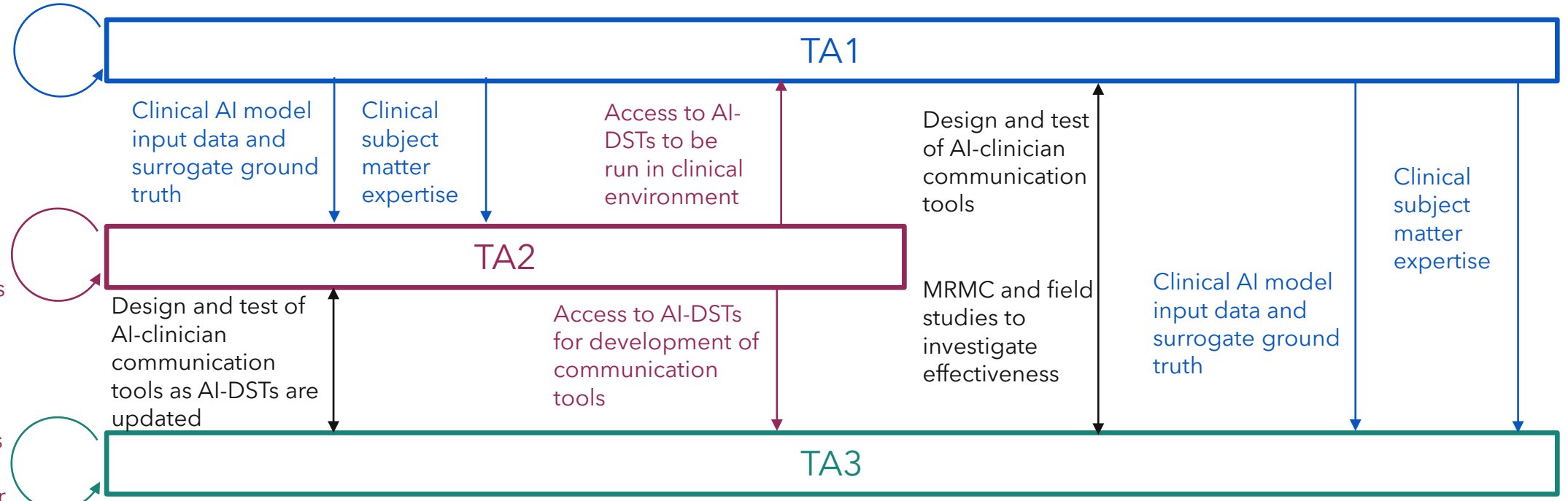
The IV&V team will quantify overall program success by using evaluation metrics aligned with program objectives

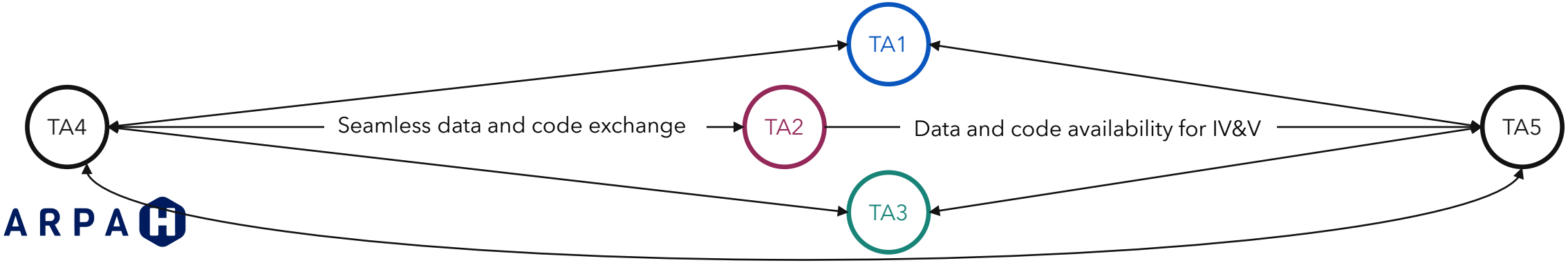
Collaboration Expectations

Coordinate on data set elements and surrogate ground truth extraction methods

Coordinate on simulation tools, designs for elucidation of potential root causes from experts

Coordinate on identification of communication tools that are successful to minimize the number of iterations





ARPA-H

Milestones & Timeline

Down-select to integrated teams

	Phase I (24 Months) Prototyping & initial tool development								Phase II (12 Months) Testing, improvement, & expansion				Phase III (12 Months) Integration			
	FY25			FY26				FY 27				FY28				FY29
	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q5
TA1: Surrogate Ground Truth Label Extraction *1-2 teams per clinical track*	IRBs for 3 clinical sites (per performer)							IRBs for 5 clinical sites (per performer)				IRBs for 8 clinical sites (per performer)				
	Initial testing & validation of 2 use cases in 2 tracks								Testing & validation of 2 additional use cases							
			Develop automated methods for surrogate ground truth label extraction from health records						Continuous assessment of 2 initial use cases at 5 sites				Continuous assessment of all 4 use cases at 8 sites			
				Compare surrogate ground truth label extraction methods w. traditional methods			Improvement and generalizability of surrogate ground truth label extraction from health records at any clinical site									
		Data harmonization across clinical sites and use cases							Data harmonization across clinical sites and all use cases				Data harmonization across 8 clinical sites and all use cases			
TA2: Degradation Detection & Self-Correction *1-2 teams per clinical track*	Develop degradation detection tool (1: Detect performance change based on model output, 2: detect changes in AI-DST model inputs, 3: Simulate changes to model inputs)															
				Application of degradation detection tool to use cases in clinical setting				Degradation alerting plan	Validation of model degradation detection tool with newly-added sites				Validation of model degradation detection with newly-added sites			
			Demonstrate techniques to simulate potential root causes				Application of root cause detection methods to real data			Improvement of root cause detection methods with real data			Validation of root cause detection methods with real data from new sites			
			Convene SME panel to ID root causes						SME panel to ID root causes		SME panel validates tool					
			Develop framework for suggesting model correction		Iteratively develop & test AI-DST retraining methods			Simulation & real-world studies to validate performance after model correction			Continuous background testing of alternative models					
									Application of automated update methods to real data across sites and use-cases				Validation of root cause detection methods with real data from new sites			
TA3: Quantify Uncertainty & Improve Clinician Performance *1 team per clinical track*	Development and initial testing of uncertainty quantification methods								Application and improvement of all tools using data from new sites and use-cases							
		Develop uncertainty communication tool				Develop 2nd communication tool			Develop 3rd communication tool							
				Clinician performance & usability studies			Clinician performance & usability studies				Clinician performance & usability studies				Clinician performance & useability studies	
									Field studies to demonstrate improved AI-clinician team performance							
TA4: Infra (*1 team*)	Stand up Data Infrastructure		Maintain and expand capabilities of data infrastructure													

Key deliverables:

TA3: Open-source software for AI-DST uncertainty quantification and communication.

TA1: Open-source software for automated ground truth label extraction from health records.

TA2: Open-source software for AI-DST degradation detection

TA1: Demo of generalizability across clinical sites

TA2: Open-source software for AI-DST root cause analysis & self-correction

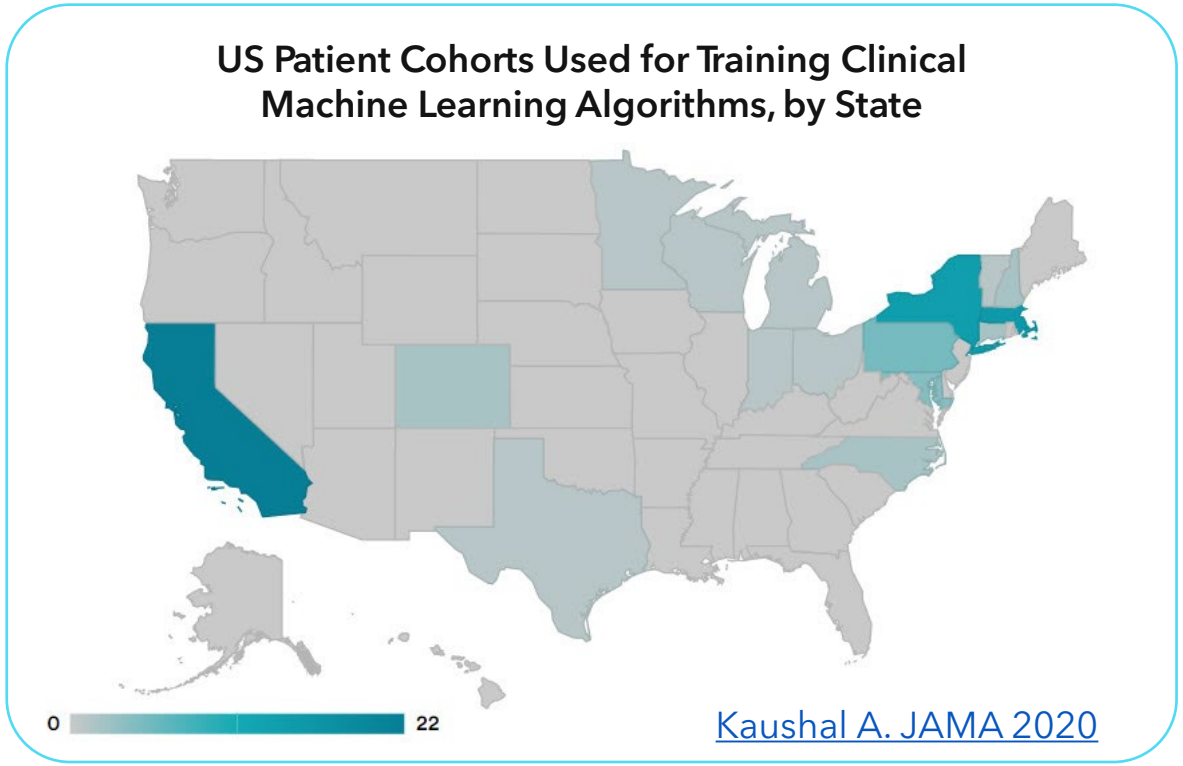
TA3: Open-source software for two additional communication tools

Integrated program deliverables

- An open-source tool suite for clinical AI developers
- A self-monitoring medical device
- Individual hospital performance monitoring capabilities
- Multi-organizational performance monitoring capabilities

Affordability and Accessibility

Enable clinical AI models to self-update and perform well for ***all*** populations



Inclusive Data Collection

- Geographic and demographic diversity will be prioritized across program performers and partners.
- Full transparency for safety and public interest.

Equitable AI performance included in program requirements.

Example: Verification team performs subgroup analysis using appropriate subgroups.

ARPA-H

Impact of PRECISE-AI

Patients

- Reduces the likelihood of harm caused by incorrect medical decisions due to degraded AI model output.
- Performance safeguards and decision support tools facilitate improvements in patient experience.

Hospitals

- Enables hospitals to monitor and fix AI tools that demonstrate performance degradation, reducing risk to patients, clinicians, and overall hospital operations.
- Informs the hospital about model performance under their clinical settings, thus enables them better understand the value they are gaining from the model.

Clinicians

- Improves clinician use of AI-DSTs through transparency into its performance and reliability.
- Supports translation of AI statistical uncertainty to actionable alerting and/or recommendations.
- Enables clinicians to interact with AI tools in a novel and time-efficient manner through improved communications tools.

Developers

- Access to open-source software and toolkits for continued innovation on AI model performance.
- Catalyzes expandable data sharing platform that provides developers with feedback on device performance
- Helps developers differentiate their AI-DSTs through the integration of monitoring and self-correction

Regulators

- Adds to total lifecycle evaluation by providing tools that enable continuous monitoring of performance of cleared AI-enabled decision support tools.

ARPA-H

Due Dates and Final Program Solicitation

- Anticipated Solution Summary Due Date: Nov. 7, 2024
- Anticipated Proposal Due Date: January 8, 2025

Program Page

- Monitor SAM.gov:
  - Any changes to the Program Solicitation (PS) will be made via formal amendments and posted on SAM.gov

SAM.gov for PRECISE-AI

- Conform to all PS requirements
  - Thoroughly read the PS
  - Non-conforming proposals will not be evaluated or considered for award

ARPA-H

PRECISE-AI

Acquisition Details

Adrea Robinson, Agreements Officer
Business Innovation Division (BID)

ARPA-H

General Proposal Information

- Follow all instructions
- Use provided templates and forms
- All communications via PRECISE-AI@arpa-h.gov
- IP assertions **must** be included
- Research Security Disclosure
  - Senior/Key Personnel **only**
- FFRDCs & USG entities are **not** eligible to propose
- Detailed budget required for Phases I & II; Rough Order of Magnitude (ROM) for Phase III

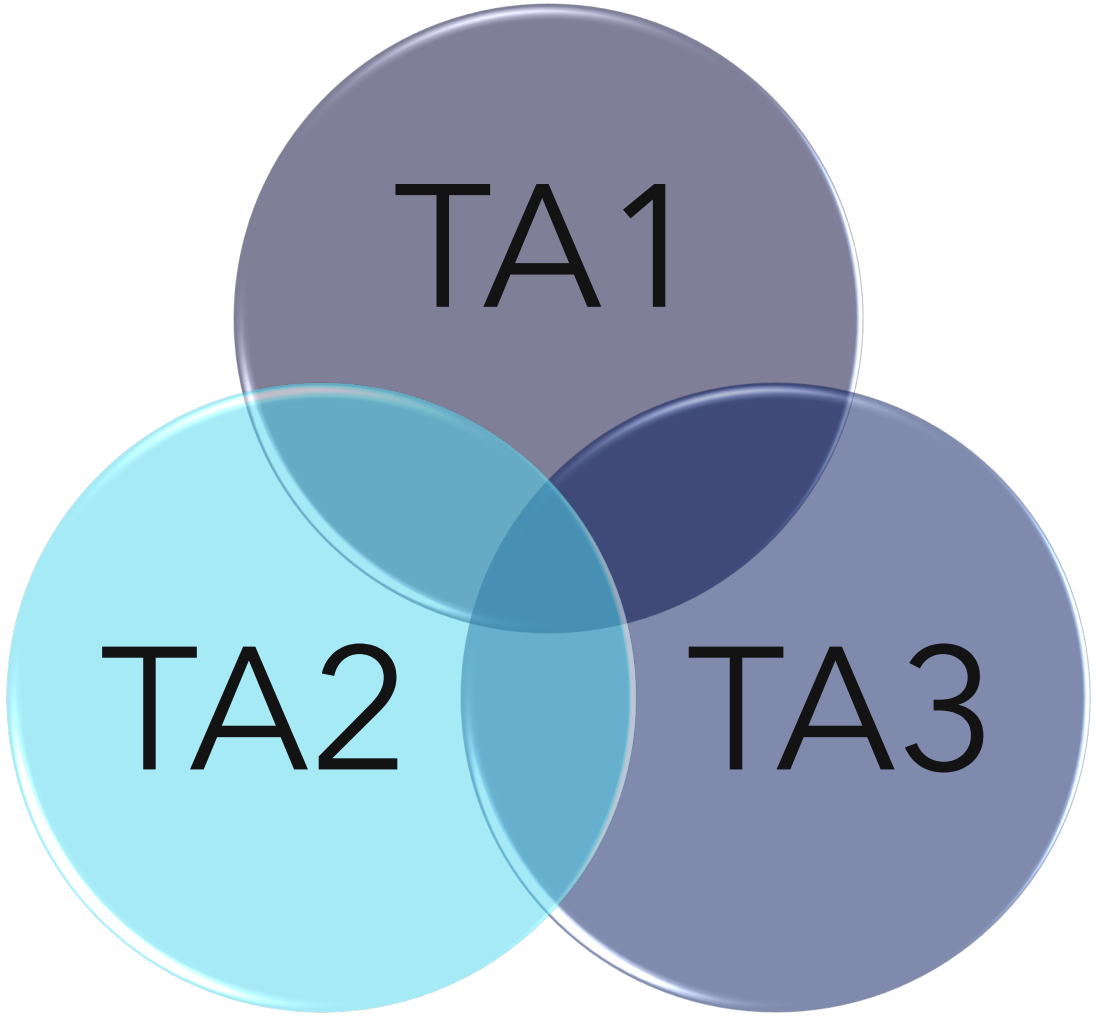
ARPA H

Proposal Submission Information

- Solution Summary required.
- Submissions **must** be through the submission portal and **on time.**
- SAM registration must be active at time of submission.

ARPA-H

Proposal Options



TA4

TA5

Phase III Proposals

- Down select anticipated
- Performers to re-propose
- Proposals must address TA1, TA2, and TA3
- Teams can reorganize

ARPA H

Process Overview

Solution Summary Submission

- REQUIRED
- Length should not exceed 3 pages (1 TA) or 6 pages (multiple TAs) (excludes cover page & ROM).
- Must include all sections as specified in the provided template.
- Will be reviewed, not evaluated.
- Submitted to solutions.arpa-h.gov

Full Proposal

- Government will recommend or discourage submission based on solution summary review.
- Submitted in solutions.arpa-h.gov

Evaluation and Selection

- The Government will evaluate each conforming proposal against criteria 1–4.
- Selection for award will be made as outlined in the Innovative Solutions Opening.

ARPA-H

Evaluation Criteria

1. Overall Scientific and Technical Merit

- Innovative, feasible, achievable, and complete.
- An outcome that achieves the goal can be expected as a result of award
- Risk identification with mitigation strategy.

2. Proposer's Capabilities and/or Related Experience

- Team expertise and experience.
- Experience in similar efforts.

3. Potential Contribution and Relevance to the ARPA-H Mission

- Future application, including unmet needs within biomedicine and to improve health outcomes.
- Potential for interdisciplinary approach.

4. Budget Risk Analysis

Volume 2 – Price/Cost Proposal

- Budget narrative should not repeat spreadsheet data
- Propose fully-burdened labor rates
- Provide any applicable rate agreements
- Provide a list of materials and equipment (no miscellaneous, contingency, etc.)

ARPA H

Award Type – Other Transactions

OTs are agreements

- Mutual assent, expressed by a valid offer and acceptance; adequate consideration; capacity; and legality (i.e., these are contracts, but not FAR procurement contracts.)
- Leverage commercial business practices more so than traditional FAR procurements.

Collaborative

- Increased collaboration and partnership, leading to more effective use of resources and knowledge sharing.

Flexible

- Many typical contract laws and regulations don't apply (e.g., CICA, FAR, CAS).
- Greater flexibility in project design and implementation.
- May fully negotiate data rights, patents, etc.

ARPA H

Final Guidance

Monitor SAM.gov

- Any/all changes to the ISO will be made via formal amendments and posted online at SAM.gov
- No information discussed at Proposers' Day shall be construed as modifying the terms and conditions of the ISO

Conform to all ISO Requirements

- Thoroughly read the ISO
- Pay special attention to the eligibility requirements outlined in the ISO
- Non-conforming proposals **will not** be evaluated or considered for award

Dates and Deadlines

- Solution Summary Due Date: November 7, 2024
- Anticipated Proposal Due Date: January 8, 2025

Questions

- Refer to the Precise-AI FAQ Website (frequently updated!)
- PRECISE-AI@arpa-h.gov

ARPA H