PROGRAM SOLICITATION
PERFORMANCE AND RELIABILITY EVALUATION FOR CONTINUOUS MODIFICATIONS AND USEABILITY OF AI (PRECISE-AI)

RESILIENT SYSTEMS OFFICE (RSO)
ADVANCED RESEARCH PROJECTS AGENCY FOR HEALTH

ARPA-H-SOL-25-113

August 29, 2024

This is a DRAFT Program Solicitation. This is not an invitation for solution summaries or proposals. Any such response will be disregarded. ARPA-H will not comment nor provide feedback on questions related to proposal technical approaches. Questions and Comments should be directed to PRECISEAI@arpa-h.gov.

PROGRAM SOLICITIATION OVERVIEW INFORMATION

FEDERAL AGENCY NAME: Advanced Research Projects Agency for Health (ARPA-H)

SOLICITATION TITLE: Performance and Reliability Evaluation for Continuous modIfications and uSEability of AI (PRECISE-AI)

ANNOUNCEMENT TYPE: PROGRAM SOLICITATION (PS), Initial Announcement

SOLICITATION NUMBER: ARPA-H-SOL-25-113

Dates: (All times listed herein are Eastern Time)

- Proposers' Day: October 2024 (specific date and location TBD)
- Program Solicitation Questions & Answers (Q&A) submission due date: TBA in final solicitation
- Solution Summary Due Date: TBA in final solicitation
- Proposal Due Date: TBA in final solicitation

CONCISE DESCRIPTION OF THE SOLICITATION:
The rapid advancement of artificial intelligence (AI) technologies is transforming healthcare by improving efficiencies, reducing costs, and enhancing health outcomes. This potential is evident with over 850 FDA-approved medical devices now incorporating AI functionalities, a tenfold increase from 2018 to 2023. However, the ability to ensure the ongoing safety and efficacy of these AI systems has not kept pace. The conventional safety testing approach relies heavily on pre-market testing, assuming that these initial results will predict long-term performance. However, pre-market results often fail to account for variations in operational processes and patient demographics, leading to unpredictable post-market performance that currently requires manual oversight by vendors. Performance and Reliability Evaluation for Continuous modIfications and uSEability of AI (PRECISE-AI) aims to create a suite of self-correction techniques that make it possible to automatically maintain peak model performance of predictive AI components across diverse clinical settings. PRECISE-AI will advance novel approaches to optimally support clinician decision-making and scalably manage the performance of AI Decision Support Tools (AI-DSTs) after their commercial deployment. Key areas of innovation include continuous monitoring capabilities, degradation detection, root cause analysis, self-correction, and bidirectional communication with clinicians. This program will establish an open-source repository of tools to autonomously maintain the performance of clinical AI-DSTs while enhancing the interpretability and actionability of AI model outputs. The program will test these innovations in real-world settings to demonstrate measurable improvements in clinical decision-making. This program addresses the pressing need for continuous monitoring and updating of clinical AI models to ensure they remain effective and trustworthy over time.

ANTICIPATED INDIVIDUAL AWARD: Multiple awards are anticipated.

TYPES OF INSTRUMENTS THAT MAY BE AWARDED: Other Transactions awarded under the authority of 42 U.S.C. § 290c(g)(1)(D).

POINTS OF CONTACT (POC):
Technical Point of Contact: Berkman Sahiner, PRECISE-AI Program Manager, Resilient Systems Office (RSO)
The PS Coordinator for this effort can be reached at: PRECISE-AI@arpa-h.gov.
ATTN: PRECISE-AI

Contents

1. PROGRAM INFORMATION

1.1. BACKGROUND

New artificial intelligence (AI) technologies are transforming healthcare, proffering improved efficiencies, reduced costs, and improved health outcomes. The market's interest is evident, with over 850 FDA-approved medical devices integrating AI functionalities — a tenfold increase from 2018 to 2023[1]. However, this growth has not been matched by scalable, automated capabilities to ensure the ongoing safety and efficacy after a vendor receives FDA clearance and after tools enter broad clinical use. Today's safety testing approaches assume that pre-market studies predict post-market performance. These processes also presuppose that vendors have sufficient resources to detect and report deviations from pre-market performance, even when predictive AI tools are deployed across hundreds of clinical settings. Yet, evidence shows pre-market results often fail to predict the long-term performance of predictive AI tools due to variations in deployment contexts, including changes to operational processes and patient demographics [2,3].

Pre-market model performance studies typically rely on curated test data wherein each patient case is individually reviewed and manually assigned a "ground-truth label" (i.e., the most accurate estimate of the diagnosis given the available evidence). A ground truth label serves as the critical baseline for evaluating AI model performance. Once an AI tool is on the market, developers lack incentives to monitor and update their products, leading to sporadic updates and no mechanisms for hospitals to track product reliability. The absence of ground truth labels in real-world clinical settings hinders hospitals from understanding AI tool performance in their environment putting patients and clinicians at increased risk. Current tools make it challenging for clinicians to detect performance deterioration of AI models, which can result from changes in clinical operations, IT infrastructure updates, and patient demographics. Performance may degrade differently across different clinical environments, and there is no established methodology to understand the root causes of the degradation and to take corrective action. These factors underscore the need for continual adaptation of AI systems to maintain their effectiveness in dynamic clinical environments.

National Health Impact: A Pathway to Improve Health Outcomes

The use of AI in clinical decision settings is rapidly expanding; however, there are currently no automated tools or mechanisms to monitor the performance and long-term safety of clinical AI Decision Support Tools (DSTs). Over time, clinical AI-DST performance can degrade without notifying clinicians. A study that evaluated 32 datasets from 4 industries found that a significant majority, 91%, of machine learning models experience a significant reduction in effectiveness as they age[4]. This degradation manifests as an 8-20% annual decrease in accuracy, or increased variability of errors,[4],[5] primarily due to shifts in the underlying data, which can erode trust in these systems.

Current post-market evaluation strategies are inadequate for addressing the issues faced by predictive AI models in clinical settings. Alarmingly, none of the existing clinical AI models undergo regular testing and

---

[1] U.S. Food and Drug Administration. *Artificial Intelligence and Machine Learning (AI/ML)-Enabled Medical Devices.* https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices

[2] Wong A, Cao J, Lyons PG, Dutta S, Major VJ, Ötles E, Singh K. Quantification of Sepsis Model Alerts in 24 US Hospitals Before and During the COVID-19 Pandemic. JAMA Netw Open. 2021 Nov 1;4(11):e2135286. doi: 10.1001/jamanetworkopen.2021.35286. PMID: 34797372; PMCID: PMC8605481

[3] de Vries CF, Colosimo SJ, Staff RT, Dymiter JA, Yearsley J, Dinneen D, Boyle M, Harrison DJ, Anderson LA, Lip G; iCAIRD Radiology Collaboration. Impact of Different Mammography Systems on Artificial Intelligence Performance in Breast Cancer Screening. Radiol Artif Intell. 2023 Mar 22;5(3):e220146. doi: 10.1148/ryai.220146. PMID: 37293340; PMCID: PMC10245180.

[4] Vela, D., Sharp. *et al.* Temporal quality degradation in AI models. *Sci Rep.* 2022. doi: 10.1038/s41598-022-15245-z

[5] Yang J. *et al.* AI Gone Astray: Technical Supplement. arXiv preprint. 2022. doi: arXiv:2203.16452

updating during their clinical use to maintain accuracy. Only 44.7% of physicians stated that they would be willing to use AI-driven medicine[6]. Additionally, there are no robust mechanisms in place to continuously monitor and update these AI models during clinical operations, though preliminary trials are being conducted to explore potential solutions[7]. Currently, clinician intuition remains the primary method for detecting model degradation, which can result in harm before the issues are manually identified. Moreover, there are no clinically validated technical solutions yet available to fulfill the requirements of the 2023 Executive Order, which mandates the development of performance monitoring for medical AI models. The FDA is in the process of creating a regulatory pathway for automated post-market AI model updating to address these challenges[8], but the industry still lacks robust, technically rigorous, scalable techniques for detecting model degradation and evaluating corrective actions.

1.2. PROGRAM DESCRIPTION

The PRECISE-AI program is soliciting proposals to create novel self-correction techniques that optimally maintain the peak performance of clinical AI decision support tools, both when operating independently and in combination with the clinicians who use them. PRECISE-AI aspires to create a future where predictive AI models continuously communicate with clinicians in a way that appropriately earns the clinician's trust. Proposers should address critical challenges in the ability of AI Decision Support Tools (AI-DSTs) to actively monitor and maintain their optimal performance with consideration to local health systems, operational processes such as data acquisition, and patient characteristics, after their commercial deployment. Because such techniques are in their infancy, PRECISE-AI will advance the science behind continuous monitoring capabilities and move beyond monitoring into degradation detection, root cause analysis, self-correction, and bidirectional communication with clinicians. PRECISE-AI aims to create and validate robust AI degradation detection and auto-correction capabilities that provide the technical means to accomplish goals laid out in the AI Executive Order[9], and the risk management frameworks created by organizations such as the FDA and NIST.

PRECISE-AI performers will create an open-source repository of tools that enable continuous monitoring and auto-correction of AI-DST. Proposers should focus specifically on AI-DSTs that make a diagnosis or prediction that is later validated by another test or clinical action. By automatically extracting surrogate ground truth label information from health records, automated monitoring techniques can continuously evaluate when AI decision support tools make accurate diagnoses and when they make mistakes. The term "surrogate ground truth label" above refers to an automatically extracted approximation to the ground truth label extracted using traditional, manual, and resource-intensive labeling. As an AI model ages, the percentage of mistakes increases due to changes in factors such as patient demographics or data acquisition techniques (technically termed as "dataset shifts"), affecting the accuracy of the AI tool. PRECISE-AI will develop self-correction tools that tackle the effects of dataset shifts, as described next.

PRECISE-AI seeks novel proposals to analyze the monitoring data to suggest root causes for performance deterioration. The root cause analysis will be used to recommend self-corrective actions that enable the AI model to improve its performance. When performance degradation occurs, it will be important to communicate the heightened risk of medical errors to clinicians, the developers of the AI decision support

---

[6] Tamori H, Yamashina H, Mukai M, Morii Y, Suzuki T, Ogasawara K. Acceptance of the Use of Artificial Intelligence in Medicine Among Japan's Doctors and the Public: A Questionnaire Survey. JMIR Hum Factors. 2022 Mar 16;9(1):e24680. doi: 10.2196/24680

[7] https://www.fiercehealthcare.com/ai-and-machine-learning/epic-plans-launch-ai-validation-software-healthcare-organizations-test

[8] https://www.fda.gov/regulatory-information/search-fda-guidance-documents/marketing-submission-recommendations-predetermined-change-control-plan-artificial

[9] Exec. Order No. 14110 of Oct 30, 2023. https://www.federalregister.gov/documents/2023/11/01/2023-24283/safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence

tool, hospital administrators, and potentially the FDA. Proposers should describe novel mechanisms to streamline these communications and ensure that AI decision support tools not only generate trustworthy recommendations but also behave in a manner that earns stakeholder's trust.

PRECISE-AI aims to improve the performance of clinicians that use AI-DSTs. Once the AI-DST monitoring and auto-correction tools are developed, it will be necessary to help transform these into meaningful insights for clinicians that enhance patient care. Proposers are encouraged to innovate new methods and tools for improving the transparency, interpretability, and actionability of AI model outputs to clinicians. Ultimately, researchers will test these tools in real world settings to demonstrate measurable improvements in clinical decision-making.

Cumulatively, PRECISE-AI seeks innovative proposals to create a large stepwise improvement over the current state of practice, where all patient groups are potentially exposed to AI-DSTs with degraded performance, resulting in possible patient harm. Proposers should endeavor to mitigate risk of patient harm beginning with patient one (by leveraging simulated data sets alongside real-world data) and precipitously decreasing risk for subsequent patients as additional real-world data is collected downstream of the AI-DST's use. As a result, the first group of patients where the AI-DST shows a deviation from its expected performance would alert stakeholders to the problem, the likely cause, and potential solutions to mitigate risk to patients. In addition, uncertainty quantification and other bi-directional communication tools with clinicians will mitigate the risk for all patients.

1.3. PROGRAM SCOPE

The PRECISE-AI program seeks proposals for novel computational techniques that enable AI-DSTs to self-monitor and maintain high quality performance across diverse clinical contexts. AI-DSTs are decision support tools that are used in a context where they provide recommendations for clinicians, and the clinicians are the final decision-makers. Enhancing post-market model performance in an automated way will require research and development in three key areas: (1) automatic extraction of surrogate ground truth labels, (2) automated detection and correction of performance degradation, and (3) robust communication with clinicians. Finally, multi-institutional data infrastructure will accelerate the aggregation of insights from different hospitals and enable effective data collection, sharing, and analysis with vendors, clinicians, hospitals, and potentially the FDA.

Automated Extraction of Ground Truth

Continuous monitoring of clinical AI systems requires a continuous source of "ground truth" that can be used to determine when an AI system performs well and when it makes mistakes. A ground truth label is the most accurate diagnosis that can be made given the available evidence. Pre-market approaches to defining ground truth labels are resource-intensive, wherein each patient case is individually reviewed and assigned a label. Manual ground truth labeling generally relies on the consensus opinion of a panel of experts, and this approach does not scale.

Automated surrogate ground truth labeling requires research advances to extract the most relevant information from medical records and test results with high accuracy, and to approximate, as closely as possible, the traditional, resource-intensive ground truth extraction used in pre-market evaluation of AI-DSTs. A single data element, such as an isolated laboratory result or imaging finding, is often insufficient to accurately determine a patient's condition. Precise identification of the surrogate ground truth label might require the integration of multiple sources of information, such as radiology and pathology reports, clinical notes, biomarker assay results, ICD and SNOMED CT codes. This complexity demands the ability to ingest diverse data types from both structured and unstructured electronic health records (EHRs).

Additionally, data elements necessary for reliable surrogate ground truth labeling are stored in diverse formats across various database systems and information models within and between healthcare institutions. This heterogeneity complicates the task of combining elements into a cohesive and accurate surrogate

ground truth label. For instance, the terminology used in the radiology reports may vary among different institutions, and within an institution, the storage format of free-text data will be different from that of the ICD codes. Consequently, existing methods often rely on manual processes or ad hoc approaches that are not scalable or robust enough to handle the volume and variety of data encountered in real-world clinical environments. These challenges make it difficult to create a continuous stream of surrogate ground truth labels to enable the continuous validation of clinical AI-DST models. Therefore, PRECISE-AI seeks innovative research proposals that address the challenge of extracting surrogate ground truth labels across systems.

Clinical AI-DSTs are often developed around a specific input data type. To increase the breadth and generalizability of program innovations, PRECISE-AI efforts will be directed towards specific clinical use cases within the following clinical tracks: diagnosis through X-ray imaging, diagnosis through Computed Tomography (CT) imaging, clinical insights derived from patient Electronic Health Records, or alternative data input types.

Automated Detection and Correction of Performance Degradation Across Hospital Systems

When an AI model starts making more mistakes within a clinical context over time, it is called performance degradation. Existing performance monitoring approaches are manual and sporadic, limiting the ability to accurately detect performance degradation across clinical sites and hospital systems. PRECISE-AI seeks to develop novel approaches that provide a continuous and comprehensive view of model performance, making it possible to detect and respond to degradation as expeditiously as clinically possible. Additionally, these innovations would create mechanisms to quickly share performance data and insights among clinicians, hospital administrators, and developers, as well as hospital systems.

It will be important for performance degradation detection techniques to distinguish between normal variability in summary performance metrics such as sensitivity and specificity, and genuine performance degradation. Simulation capabilities are an example of an approach that could be used to test the behavior of AI models under various degradation scenarios at scale. High-precision and high-confidence detection of model degradation requires extensive analysis of degradation behaviors, which is not adequately supported by current methodologies. PRECISE-AI seeks proposals that address gaps in the timely and accurate detection of AI model degradation, to enable systemic improvements in the overall reliability and effectiveness of clinical AI systems.

Once performance degradation has been detected, root cause analysis is key to respond to the degradation. PRECISE-AI seeks innovative techniques to empower hospitals and clinical sites by providing an automated analysis of the most likely root causes for a given performance degradation episode. The current lack of datasets hinders the ability to systematically study and address the underlying factors contributing to model degradation. Large-scale simulations will therefore likely play a key role in developing robust root cause analysis tools. At the same time, PRECISE-AI proposers should address the gap in real-world data availability for performance degradation and root cause analysis by continuously monitoring AI-DSTs. The developed methods will be validated and stress-tested with both simulated and real-world data, providing scientific and practical evidence into the performance of the developed root-cause analysis methods.

Research funded through the PRECISE-AI program will ultimately prototype novel machine learning and AI techniques to correct degraded model performance. Self-correction mechanisms for AI models carry significant risks, including the potential to inadvertently deteriorate overall performance or negatively impact specific subgroups. Therefore, research will focus on self-correction approaches that carefully balance the potential benefits of model updates against the risks, a process that is both complex and nuanced. PRECISE-AI will advance research on reliable methods to enhance the effectiveness and safety of automated or semi-automated model updates. Methods of interest will recommend self-correcting actions that predictably improve AI-DST outputs, bolstering the overall reliability of clinical AI models.

Robust Communication with Clinicians

When tools fail to convey important nuances about the reliability of the AI outputs, it erodes clinician trust and limits the practical utility of these AI models. The task of quantifying model uncertainty varies significantly across different clinical use cases and existing methods are often not tailored to the specific needs of diverse clinical environments. PRECISE-AI performers will clinically evaluate novel approaches to convey model uncertainty and other supplementary model information to clinicians. Examples of supplementary information include AI explainability tools, capabilities enabling clinicians further interrogate the model (e.g., question-answering systems), or approaches to display of data from similar cases with ground truth. Because AI-DST capabilities are only as good as the decisions that clinicians make based on their outputs, PRECISE-AI performers will compare techniques for communicating model certainty and other supplementary model information to determine which ones give clinicians an appropriate level of trust in AI-DST outputs and ultimately improve clinician performance.

1.4. PROGRAM STRUCTURE & TECHNICAL AREAS

PRECISE-AI will advance auto-correction capabilities for predictive AI-DSTs. The program will develop a suite of tools and capabilities that continuously monitor the performance of predictive AI-DSTs in the clinic, detect when the performance deteriorates due to changes in a dynamic clinical environment, automatically suggest, and when appropriate, implement updates to the AI-DST models to counteract the effect of the changes in the environment, and establish improved and dynamic communication with the clinicians, resulting in improved clinical decision-making by clinicians.

1.1.1. Summary of Program Technical Areas

PRECISE-AI is comprised of five interconnected Technical Areas (TAs), outlined below.

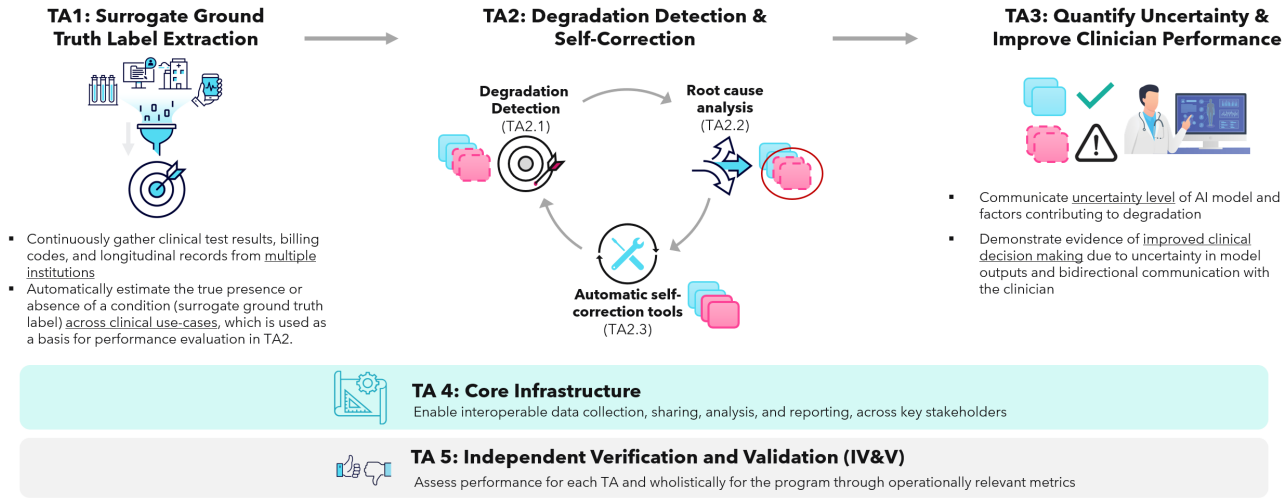
- **TA1 - Automated Surrogate Ground Truth Label Extraction**: Automate the extraction of surrogate ground truth labels for individual patients across a diverse range of clinical use-cases. This will provide a robust foundation for assessing AI model performance continuously in TA2.
- **TA2 - Degradation Detection & Self-Correction**
    - **TA2.1 - Continuous Degradation Detection Tools:** Quickly and accurately differentiate between normal AI model performance variability and actual performance degradation based on surrogate ground truth labels extracted in TA1.
    - **TA2.2 - AI-Based Root-Cause-Analysis Tools:** Develop AI-based tools that can automatically pinpoint the root causes of performance drops following alerts from degradation detection technologies developed in TA2.1.
    - **TA2.3 - AI Model Self-Correction Tools:** Develop technologies enabling the AI models to automatically suggest and when appropriate, implement necessary updates to mitigate the issues detected in TA2.2.
- **TA3 - Quantify Uncertainty & Improve Clinician Performance:** Enhance clinician trust in AI-supported decision-making processes and improve their overall performance. This will be achieved by accurately quantifying and communicating uncertainty in AI model outputs and other supplementary information that aids in the clinician's ability to appropriately trust AI models.
- **TA4 - Core Data Infrastructure:** Establish a core data infrastructure that supports interoperable data collection, sharing, analysis, and reporting among key stakeholders. During the program, this infrastructure will underpin the entire program, enabling effective coordination among performers. After the program, aspects of TA4, such as reporting of performance degradations or identified root

causes for degradations to stakeholders, are expected to continue so that the program's effectiveness is maintained after the program's lifecycle. Communication of analysis results may include public workshops and interactive performance dashboards, as well as mechanisms such as adverse event reporting and predetermined change control stipulations in regulatory submissions for communicating with regulators.

- **TA5 - Independent Verification and Validation:** Test and validate results from TA1, TA2, and TA3 to confirm performers' progress. This will include validating the performers' progress towards the metrics listed in Figure 3 below (Program Metrics) with independent data to ensure the generalization of the reported methods to new data and new clinical sites, as well as the verification and validation of simulation tools developed by other performers.

Each proposal responsive to this Solicitation can address a single Technical Area (TA) among TA1, TA2, TA3, TA4 and TA5, or a combination of TA1, TA2, and TA3. Proposals that address TA4 or TA5 cannot address any of the other TAs. Proposers to TA1 or TA2 must align their technologies to specific clinical tracks (Figure 4) with preference for technologies that can be generalized outside of their chosen track. Proposers to TA3 are expected to develop technologies that are largely generalizable across the program's clinical tracks (i.e., X-ray imaging, diagnosis through Computed Tomography (CT) imaging, and clinical insights derived from patient Electronic Health Records). Proposals that apply to a combination of TA1, TA2, and TA3 (e.g., joint TA1/TA2 or joint TA1/TA2/TA3 proposals) may submit an alternative use case outside of the aforementioned clinical tracks.

Figure 1. Summary of Program Technical Areas



1.1.2. Program Structure

PRECISE-AI is a 4-year, 3-phase program that consists of five TAs. The program will advance technically rigorous techniques to continuously monitor and correct AI models used in real-world clinical decision settings. Performers will produce tool suites that include multiple innovative elements: automatic and continuous surrogate ground truth label extraction, continuous degradation detection, AI-based root-cause-analysis, model self-correction, and intuitive and actionable quantification of model outputs. Innovations throughout the program will be developed for use cases that fall within priority clinical tracks (i.e., X-ray imaging, diagnosis through Computed Tomography (CT) imaging, clinical insights derived from patient Electronic Health Records; see Figure 4 for more information). Progression to Phase II and to Phase III will depend on performance against milestones (Figure 2), metrics (Figure 3), and evaluation by the Independent

Verification and Validation (IV&V) performer.

PRECISE-AI will structure work across three phases as follows:

- **Phase I (months 1-24):** Performers will **prototype** their proposed approaches by providing computational models, component prototypes, and landscape reports. TA1 performers will collaborate with TA2-TA4 to generate use-case-specific surrogate ground truth labels across all sites.
- **Phase II (months 25-36):** Performers will **refine** prototypes, **test** in real-world clinical settings, and **expand** the breath of sites and use cases being covered for maximum impact.
- **Phase III (months 37-48):** Performers will **integrate** their component technologies into a commercial transition package, which can include any or all of the following options: an open-source tool suite that can be leveraged by multiple vendors to improve AI model performance; a self-monitoring medical device that will be submitted for FDA clearance; a performance monitoring capability that enables hospitals to monitor the performance of their AI-enabled medical devices; or a multi-organization capability that could assess patient safety across multiple hospital systems.

1.1.3. TA1: Automated Surrogate Ground Truth Label Extraction

TA1 proposals should describe an innovative, scalable, sustainable approach to automatically and continuously identify individual patient-level surrogate ground truth labels across a breadth of clinical use cases. For healthcare providers to evaluate whether AI tools are performing well in a site-specific manner, it is necessary to define surrogate ground truth labels that serve as the basis for comparison with the AI model output. TA1 aims to innovate scalable and automatic methods to derive surrogate ground truth labels by retrieving and analyzing disparate patient-level data sources to create surrogate ground truth labels.

For example, the purpose of a given clinical AI model may be to identify patients with pneumonia based on chest X-ray images. To evaluate how well the AI model is performing (done in TA2), TA1 proposers would establish a surrogate ground truth for each specific patient (i.e., what is the best estimate that a particular patient had pneumonia on the date of the chest X-ray image). This surrogate ground truth label should be triangulated from multiple data sources (e.g., ICD-9 or ICD-10 diagnostic codes, review of infection related symptoms, laboratory test results, prescribed medications, and radiology reports). TA1 proposers will also specify when the extracted surrogate ground truth labels are ready for use in TA2 tasks. For example, in the example above, this may be after the diagnostic codes, review of infection related symptoms, laboratory test results, prescribed medications, and radiology reports have been entered into the EHR. As described below, TA1 proposers will validate their surrogate ground truth extraction method, with the timeline specified in their proposal, by comparing it to traditional ground truth extraction methods (e.g., with a manual method and/or an expert panel and/or a longer timeline).

The surrogate ground truth labels that are produced by TA1 performers will serve as the basis for comparison when TA2 evaluates clinical AI model performance. The purpose of TA1 is not to develop AI-DSTs for the use cases selected by TA2, but to develop generalizable automated surrogate ground truth label extraction techniques to be used for the continuous monitoring and updating of AI-DST. To achieve this, TA1 proposals should focus on the following objectives:

Objective 1: Identify and ingest data elements that correlate with ground truth. TA1 performers will determine which disparate sources of healthcare information are necessary for establishing surrogate ground truth labels for each clinical use case, as outlined in Figure 4. TA1 will develop methods to extract these data elements from multiple sources in an interoperable manner to inform surrogate ground truth labeling.

TA1 performers will work with clinical partners and the developers of the AI-DSTs selected by TA2 to establish the baseline performance and to help track the temporal performance of TA2-selected AI-DSTs. To establish the initial performance baseline, performers may use historical (previously acquired) data from TA2 performers, other data repositories, and ground truth labels derived using traditional methods. During the program, TA1 performers will collaborate with hospitals or other clinical care sites to create algorithms that support the continuous extraction of data elements that provide patient-level surrogate ground truth labels. Additionally, TA1 performers will estimate the number of individual patient records required to reliably achieve the concordance levels targeted in Figure 3 between automatically extracted and traditionally defined ground truth labels. For instance, if a TA2 performer selects an AI-DST aligned with an X-ray based use case, the corresponding TA1 performer will need to estimate the number of patient records necessary for establishing statistically accurate and precise surrogate ground truth labels, and then extract relevant data from radiological imaging facilities, patient encounter notes, and other pertinent sources in the EHR.

For each program use case selected by TA2 performers, TA1 performers will test and optimize NLP and multimodal data integration to consistently ingest data from different healthcare data infrastructures and information coding systems (e.g., ICD10, SNOMED CT, CPT codes, pharmacy and prescription codes, insurance claims data, clinicians' notes, clinical reports). TA1 proposals should describe how to combine the necessary breadth and numbers of disparate patient records from diverse EHR systems to establish a coherent, harmonized history of patient cases across multiple sites. Furthermore, they will extract metadata, such as relevant patient demographics, to enable sub-population comparisons.

TA1 performers will work with TA4 performers to deliver this data into TA4’s core data infrastructure using interoperable methods and a continuous integration continuous delivery (CICD) approach. TA1 performers will demonstrate the generalizability of their approach for surrogate ground truth label extraction across multiple use cases specified by TA2 performers and across multiple clinical sites in Phases II and III.

Objective 2: Develop methods to produce surrogate ground truth information automatically and efficiently. Once a set of data elements that correlate with the correct diagnosis have been identified and extracted, performers will establish scalable and automated methods to analyze the necessary data to continuously produce patient-specific (or case-specific) surrogate ground truth labels.

During the program, TA1 performers will develop surrogate ground truth label inference techniques that provide an accurate estimate of the diagnosis given the available evidence by integrating multi-modal data across contexts and time-points. Performers will leverage multiple approaches, including (i) extracting non-structured healthcare data from patient health records (e.g., separate test results, clinician notes); (ii) machine learning techniques that perform feature selection from a larger set of data elements; and (iii) complex methods for information fusion. TA1 performers will develop automated and scalable methods and tools to define surrogate ground truth labels, which can include, but are not limited to, rule-based systems, ML-based methods, and a combination of rule-based and ML-based methods. They will demonstrate the ability to continuously and automatically extract surrogate ground truth labels from different healthcare data. Additionally, performers will demonstrate agreement between the automated and traditionally defined ground truth labels, using a representative and properly sized data set.

The primary deliverable of objective 2 will be the successful development of methodologies that enable continuous auto-extraction of relevant clinical data and generation of surrogate ground truth labels.

Objective 3: Engage hospital partners and create a diverse and representative dataset for program

activities across all TAs. TA1 will work with multiple independently sourced hospital partners and create a diverse, representative longitudinal dataset that spans multiple institutions, patient demographics, coding systems, and interrelated clinical use cases. This dataset will be used for: training and testing TA1 surrogate ground truth inference algorithms; all activities for TA2 (including monitoring, root cause analysis, and self-correction); as well as communication and test activities for TA3.

Strong TA1 proposals will establish partnerships with hospitals and clinical sites that enables the program to evaluate AI-DST performance across a broad array of patients and environments, through data collected in real world practice for AI-DSTs at these sites. TA1 performers will be required to expand the number and diversity of clinical sites throughout the program (see Figure 3 for specific metrics). TA1 proposers will demonstrate that enough number of cases can be collected through their partnerships with these sites from important demographic subgroups (e.g., defined by ancestry, gender, age, geography) to make statistically meaningful comparisons of the performance of the AI-DST among these subgroups. The partnering hospitals or clinical sites should be varied in terms of their size, geographic location and demographics they serve, so that potential AI-DST performance differences among these sites can be measured.

During the program, TA1 performers will establish data use agreements with clinical sites to provide the necessary clinical data for the AI-DST use cases selected by TA2 performers. These use cases will fall into pre-defined clinical tracks, such as X-ray, CT, and EHR (Figure 4). The portfolio of clinical sites should allow for unbiased data collection in terms of patient demographics for the selected use cases.

Innovations from each TA will be tested across the array of hospital partners that have been assembled by TA1 performers. This real-world testing will enable improved AI performance monitoring and help demonstrate improved clinical decision-making.

TA1 proposals will address the following topics:

1. Describe the necessary input data types and quantify the number of patient records needed to enable surrogate ground truth label inference with statistical accuracy and precision for each use case.
2. Strong proposals will describe innovative, comprehensive surrogate ground truth labeling methods capable of spanning different health centers/clinics.
3. Strong proposals will describe generalizable methods that will be used for automatically extracting data from clinical reports, defining the surrogate ground truth labels, and explaining how the methods will be validated.
4. Strong proposals will limit the time horizon for data elements to be identified, so that model performance monitoring can be performed in as little time as possible after the acquisition of the input data for the AI-DST.
5. Describe supplementary metrics, if necessary, that may enable stakeholders to measure progress more accurately towards program goals.
6. Identify the clinical sites required in Phase I and provide evidence of their willingness to partner throughout the program. Strong proposals will identify more clinical sites than required in Phase I.
7. Include the historical percentage of patients at each site that belong to underrepresented populations, defined by ancestry, age, gender, geography, healthcare access and utilization. Describe how the clinical site portfolio allows for unbiased data collection in terms of patient demographics for the selected use cases, comparing the demographics of patient sets to be collected to national averages to demonstrate the diversity of the overall patient populations and minimal data collection bias.

8. Strong proposals will select at least one site in Phase I and two sites in Phase II that have a preponderance of underrepresented populations and describe the approach to establish data use agreements with a growing number of clinical sites that incorporate these populations.
9. Provide a sharing plan that allows for interoperable data sharing of AI-DST input data, output data, and relevant metadata to program data repositories (e.g., TA4).

The following are out of scope for TA1 proposals:

- Training / developing new AI-DST models. Although it is permissible to use the data collected at the clinical sites to improve the AI-DST model performance, the improvement targeted in the overall program is with respect to performance degradations caused by dataset shifts in existing AI-DST models identified by TA2 performers.

1.1.4. TA2: Degradation Detection & Self-Correction

TA2 will develop AI auto-alert and auto-correction techniques to enable continuous end-to-end performance monitoring and self-corrective updating of clinical AI models by developing and combining capabilities for degradation detection, root cause analysis, and intelligent update implementation. TA2 performers will cover all three sub-TAs (2.1, 2.2, and 2.3), which are designed to seamlessly integrate, providing a comprehensive system that automatically identifies, analyzes, and corrects clinical AI model performance issues, thus maintaining the integrity and accuracy of AI applications in clinical settings. In their proposals, TA2 proposers should break down technical approach descriptions, cost estimates, and key personnel descriptions to the sub-TA level.

TA2.1: Continuous Degradation Detection Tools

Strong TA2.1 proposals will describe innovative, scalable methods for accurate, timely, and granular detection of AI model degradation in clinical settings. Proposers are encouraged to develop a comprehensive framework for the continuous and automatic monitoring of key performance metrics, such as sensitivity, specificity, positive predictive value and predicted positive rate across various clinical sites. Advanced statistical analyses and machine learning techniques that distinguish between normal variability and genuine performance degradation are of interest. Relevant approaches include, but are not limited to, Statistical Process Control (SPC) charts[10,11], Sequential Probability Ratio Test (SPRT)[12] and Bayesian methods[13]. Proposers are encouraged to address challenges around the unique statistical distribution of AI performance metrics, clinical factors for alarm threshold selection, allowable statistical variation, and patient volume. Leveraging the metadata collected in TA1, TA2.1 proposers should describe a statistically robust approach to model degradation detection across clinically relevant subpopulations. TA2.1 degradation alerts should be more nuanced than a binary decision, so alerts may communicate that an AI-DST works well for one subgroup and not another. The approach should provide inputs to the root-cause analysis in TA2.2.

TA2.1 proposers can also consider how to detect changes in the statistical properties of input data to the AI model, such as variations in the preprocessing of input images. By combining detected changes at both the

[10] Woodall, W. H. (2006). The Use of Control Charts in Healthcare and Public-Health Surveillance. Journal of Quality Technology, 38(2), 89–104. https://doi.org/10.1080/00224065.2006.11918593

[11] Lowry, C. A., Woodall, W. H., Champ, C. W., & Rigdon, S. E. (1992). A Multivariate Exponentially Weighted Moving Average Control Chart. Technometrics, 34(1), 46–53. https://doi.org/10.1080/00401706.1992.10485232

[12] Grigg OA, Farewell VT, Spiegelhalter DJ. Use of risk-adjusted CUSUM and RSPRT charts for monitoring in medical contexts. Stat Methods Med Res. 2003 Mar;12(2):147-70. doi: 10.1177/096228020301200205. PMID: 12665208.

[13] Predictive Control Charts (PCC): A Bayesian approach in online monitoring of short runs. Journal of Quality Technology, 54(4), 367–391. https://doi.org/10.1080/00224065.2021.1916413

input and output of the model, it may be possible to improve performance degradation detection, resulting in methods with fewer false alarms and fast detection of true degradations. To optimize the methods for model performance degradation detection and to cover combinations of changes of the AI model input data and AI model output, performers can rely on not only clinical data but also simulation studies that use synthetic data. Proposers who adopt this approach are encouraged to develop innovative methods that use realistic synthetic data as well as clinically obtained data. Proposers who intend to use generative methods to obtain synthetic data are encouraged to consider using multiple fidelity metrics, such as the Frechet Inception Distance, to evaluate and ensure synthetic data quality.

TA2 performers will work with TA1 clinical sites to continuously and automatically monitor key performance metrics such as sensitivity, specificity, positive predictive value and predicted positive rate, enabling timely assessment of AI models across various clinical sites. TA2.1 will leverage scalable, automated surrogate ground truth labels extracted in TA1 to perform continuous and automated performance assessments. Furthermore, with the data-sharing tools implemented in TA4, continuous performance assessments can be conducted on an aggregated basis across multiple clinical sites.

TA2.1 proposals should describe a strategy to continuously monitor AI model performance at local and global scales, with few false alarms while ensuring quick and granular detection of deviations, particularly for specific patient subpopulations and clinical scenarios. TA2.1 proposals should explain how the team plans to compare local and site-aggregated performances and integrate sub-population monitoring features into their continuous monitoring activities, enabling the assessment of AI model performance across different patient demographics and clinical contexts. This will ensure that the AI models are effective for all patient populations and all clinical settings in which they are used.

Starting early in Phase I, TA2 performers will provide access to AI-DSTs for two use cases from different clinical tracks (Figure 4) to TA1 performers for testing. In Phase II, TA2 performers will provide access to AI-DSTs for two additional use cases to TA1 performers for continued testing.

TA2.2: AI-Based Root-Cause-Analysis Tools

TA2.2 will develop AI-based root-cause-analysis tools for specified clinical use-cases. Performers will focus on identifying and testing potential root causes for AI model performance degradation across various use cases in the priority clinical tracks (Figure 4). Strong proposals will identify compelling strategies to identify root causes. Performers are encouraged to develop simulation frameworks or other approaches that mimic various data distribution changes, such as alterations in model inputs and ground truth labels[14]. Performers are encouraged to address common sources of dataset shifts including changes in patient demographics, data acquisition, storage and preprocessing, electronic health record management, and clinical definitions. In addition, proposers should describe how a diverse set of subject matter experts, including clinicians, AI developers, hospital administrators, and regulators, will contribute their insights to identify the set of potential root causes for performance degradation. TA2.2 proposers are encouraged to explain how large-scale simulations, or other approaches will enable testing of various types of changes to model inputs and ground truth labels to elucidate the effects of potential degradation scenarios and potential root causes. Additionally, proposers should describe how simulations accurately reflect real-world scenarios, including the common sources of dataset shifts and the potential degradation root causes identified by subject matter experts.

Performers will adapt and extend various machine learning approaches to precisely attribute observed performance degradation to one (or more) specific causes included in the root-cause analysis toolkit and

[14] Frangi AF, et al. Simulation and Synthesis in Medical Imaging. IEEE Trans Med Imaging. 2018. doi: 10.1109/TMI.2018.2800298.

demonstrate that the root-cause analysis tools can be applied in clinical decision settings. This will involve a comparative analysis of the performance of the ML methods for accurately associating model degradation with potential root causes and flagging conditions where no clear cause is matched.

Performers should describe how they will evaluate these approaches using independent simulated data and validate their efficacy by applying them to real-world clinical data. Proposers are encouraged to discuss approaches to detect real-world deviations, such as comparing the distribution of data elements before and after a degradation alarm is issued, which can help identify root causes that may not have been explicitly identified by other methods. For instance, changes in the way a technician operates medical imaging equipment may not have been previously considered as a root cause. Therefore, TA2.2 proposers should describe innovative approaches to attribute the causes for AI model performance degradation in a variety of scenarios, including ones for which data has not yet been collected.

TA2.3: AI Model Self-Correction Tools

For TA2.3, strong proposals will describe innovative methods to recommend the optimal approach to correct AI model performance degradation. TA2.3 will create the scientific foundations needed to safely recommend model adjustments or other corrective actions (e.g., modifying an image preprocessing algorithm), collectively referred to as model updates, in a manner that consistently improves patient safety. To achieve this, performers will develop methods that leverage the continuous performance monitoring from TA2.1 and analyses from TA2.2 to predict the performance and risks associated with new model updates and compare different model versions before clinical deployment. Strong proposals will describe compelling mechanisms to determine thresholds for initiating model updates and ensure updates are applied only when they lead to significant performance improvements. Compelling approaches will thoroughly assess factors such as the availability of original and new training data, the degree of performance degradation, training costs, and the balance between stability and plasticity to optimize update suggestions. Strong proposals should outline a sound strategy to ensure that updates are appropriate and beneficial, while balancing against potential risks.

PRECISE-AI's goal is to recommend the right model for the right patient population in the right location, so strong proposals will explain how to mitigate when an AI-DST works well in one patient demographic and not another. Multi-model solutions are in scope, where different clinical sites may benefit from different models for the same AI-DST. The nature of the proposed solutions may differ depending on the severity of the degradation and the risk to patients. Performers are encouraged to use predictive approaches, including but not limited to simulation tools, to forecast the effect of the suggested updates and the likelihood that updates will mitigate performance degradation.

TA2.3 proposals should also outline strategies to implement model updates safely and automatically or semi-automatically, while monitoring for risks like unintended model behaviors following the change. Strong TA2.3 proposals will address the complex risks and benefits associated with automated model updates and propose evaluations that will help decision makers assess whether or in what cases automatic model updates will improve patient safety, what types of human oversight should be required, in what cases automated updates should be prohibited, and what type of evidence is required to inform these decisions. Decision makers above may include clinicians, hospital administrators, AI-DST developers, and regulators. Proposers are encouraged to consider topics such as selecting subsets of data for updates based on data quality, the quality of automatically extracted surrogate ground truth labels, and update complexity. Data quality in this context refers to data representativeness, absence of bias and artifacts, and an acceptable level of noise, among other factors. In addition, strong proposals will describe techniques that enable partial updates to address model shortcomings in low-performing conditions without disrupting model behavior for conditions

where the models are already performing well.

Ultimately, TA2 performers should demonstrate an end-to-end pipeline, combining degradation detection (TA2.1), root cause analysis (TA2.2), and model updating (TA2.3). Depending on the nature of the clinical task, the TA2.3 model updates could include a human-on-the-loop component that automatically collects simulated and empirical evidence of the improvements associated with a proposed model update so that a human can decide when to implement the change.

Strong proposals will describe safeguards to validate the updated models' performance, ensuring they meet or exceed original performance levels and avoid introducing new biases. Continuous validation methods are highly encouraged, potentially running multiple updated models as beta versions in the background, leveraging continuous performance assessment methods to choose the appropriate model update and confirm reliable performance over time. This comprehensive approach mitigates risks associated with automated updates and will maintain the AI systems' reliability and effectiveness in clinical environments. TA2.3 performers will develop a framework for suggesting the type of update (e.g., threshold adjustment, recalibration, retraining) that addresses performance degradation, based on risks and benefits. They will also demonstrate safeguards for ensuring the clinical safety of the updated models and test the real-world efficacy of automated or semi-automated updates, for multiple use cases, in safe-guarded settings that do not impact patient management. The generalizability of generated approaches for performance degradation detection, root cause analysis, and automated or semi-automated updates will be demonstrated across multiple use-cases and clinical tracks.

TA2 proposals will address the following topics:

1. TA2 performers must identify the AI-DSTs that will be tested in the PRECISE-AI program. If commercially available AI-DSTs are selected, performers are responsible for obtaining the necessary permissions from the vendors and/or intellectual property holders. This may involve establishing collaboration agreements, licensing arrangements, or other legal contracts to ensure that the AI-DSTs can be accessed, modified, and tested within the scope of the program. Performers should provide evidence of these permissions and arrangements in their proposals to demonstrate the feasibility of their proposed work. Performers must address all components of TA2 (TA2.1, TA2.2, TA2.3).
2. Identify clinical use cases that fall within priority clinical tracks (Figure 4) and AI-DSTs that will be used for evaluation and innovation throughout the program.
3. For commercial AI-DSTs, proposals will have a letter of support from developer granting permission to deconstruct and refine said model in a manner that improves patient safety.
4. Describe methods to ensure data hygiene (e.g., data sequestration to ensure independence), used for training and testing all the tools, and how TA2 performers will collaborate with TA1 and TA4 performers to accomplish this.
5. Describe methods to detect changes (at the AI-DST input, output and/or the combination) that lead to performance degradation.
6. Strong proposals will include simulation techniques where the user can control specific parameters to investigate how changes in data distribution affect model outputs.
7. Describe the suite of tools that will be used to generate simulated data corresponding to potential performance degradation root causes for multiple clinical tracks from Figure 4.
8. Strong proposals will design their simulation tools to enable precise control on the mechanism of the cause for degradation.
9. Describe how the root cause of a detected degradation will be matched to one or more causes from the set of potential root causes identified by an SME panel, both for simulated and real data.

10. Describe how the performance of automated updates will be assessed.
11. Provide a framework for ensuring the safety of the updated models.
12. Strong proposals will include other guarantees other than running the updated models in the background to ensure safety.
13. Provide a plan for clinical implementation of degradation detection, automated root cause identification and automated or semi-automated updates for multiple clinical tracks from Figure 4, and how the performance of the entire cycle will be measured.
14. Describe supplementary metrics, if necessary, that may enable stakeholders to measure progress more accurately towards program goals.
15. Provide a sharing plan that allows for interoperable data sharing of AI input data, AI output data, and relevant metadata to program data repositories. This can be on existing platforms, or the creation of new ones through TA4. The sharing plan should also provision for sharing AI-DST models with TA1 and TA3 performers

The following are out of scope for TA2 proposals:

- Training / developing new AI-DST models. Although it is permissible to use the data collected at the clinical sites to improve the AI-DST model performance, the improvement targeted in the overall program is with respect to performance degradations caused by dataset shifts detected by TA1/TA2.1.
- Automatically updating the production model for an AI-DST that is in clinical use without necessary IRB approvals and/or an FDA Investigational Device Exemption (IDE). For models marketed for clinical use, an automated update to the clinically deployed model can only be permissible if the update is included as part of an FDA-cleared pre-determined change control plan (PCCP) and the conditions in the PCCP are satisfied.

1.1.5. TA3: Quantify Uncertainty & Improve Clinician Performance

For AI models to be appropriately utilized to inform healthcare decisions, it is necessary that they be trustworthy and understandable. This requires transparency into how an AI tool is performing and an understanding of how users (e.g. clinicians) perceive AI performance reliability in decision making. The goal of TA3 is to improve clinician trust in AI-DSTs and enhance clinician performance. Based on the success from Phases I and II, and with input from TA1 and TA2 performers, TA3 performers will participate in the selection of one use case that will be integrated with the features developed in other TAs to be ready for commercial transition or to be submitted to the FDA by the end of Phase III. TA3 performers will collaborate with other performers in Phase III for product integration. To achieve this, TA3 proposals should focus on the following objectives:

Objective 1: Quantify model uncertainty and improve methods of communicating model uncertainty and complementary measures to clinicians. TA3 seeks to develop and improve approaches to ascertain model strengths, weaknesses, and uncertainties, which will later be conveyed in a manner that leads to appropriate (i.e. calibrated) trust levels among clinicians.

TA3 performers will leverage TA1 data and TA2 performance monitoring tools to adapt existing

methods[15,16,17,18] to quantify uncertainty for individual model outputs. They will assess the performance of calibration and uncertainty quantification using metrics such as expected calibration error, uncertainty propagation or by developing new metrics. TA3 performers will compare at least three uncertainty quantification methods and measure their effectiveness in quantifying uncertainty in AI Decision Support Tool (AI-DST) predictions. Initially, performers can test with any AI-DST model, but by the end of Phase I, they will apply the uncertainty quantification methods to multiple use cases defined by the TA2 performers, using data collected from clinical sites during the program.

TA3 performers will focus on improving clinical decision-making by incorporating uncertainty quantification and other supplementary information, such as explainability, question-answering (QA) components, and information that is difficult to include in the quantitative assessment of the model but can help the clinician. Leveraging human-centered design principles, TA3 will develop agile communication tools that clearly communicate complex AI-DST information. Essential features of the tool include communication of uncertainty quantification and at least two additional dimensions, such as explainability, performance dashboards, and retrieval of similar cases with a known ground truth label. TA3 innovations may be interactive and would accumulate feedback provided by clinician users while periodically integrating the prevailing feedback into future prototypes.

Proposers are encouraged to describe approaches that will improve clinician's ability to ascertain when AI-DSTs can and cannot be trusted in a manner that improves the overall performance of the AI-clinician team. Approaches of interest include, but are not limited to, visual representations of uncertainty that are intuitive for the user[19,20], along with tools that communicate the assumptions underlying predictions, such as feature or parameter weightings.

Performers will conduct user preference studies for multiple use cases, testing their final design with an independent set of users and cases (patients) for validation purposes to limit bias. In Phases II and III, performers will demonstrate the generalizability of their approach for uncertainty quantification and improved AI-clinician communication by expanding to additional clinical sites.

Objective 2: Demonstrate the effect of PRECISE-AI tools on clinician performance by conducting a comparative analysis across different uncertainty and supplementary information communication approaches and quantifying their impacts on clinician performance (e.g., using observer performance studies).

TA3 performers will evaluate the performance of the AI-clinician team using the tools developed across all

---

[15] Uncertainty quantification using Bayesian neural networks in classification: Application to biomedical image segmentation. Computational Statistics and Data Analysis, 142, Article 106816. https://doi.org/10.1016/j.csda.2019.106816

[16] G. Wang, W. Li, M. Aertsen, J. Deprest, S. Ourselin, T. Vercauteren, Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks, Neurocomputing 338 (2019) 34–45 https://doi.org/10.1016/j.neucom.2019.01.103.

[17] M. Hoffmann, M. Scherer, T. Hempel, A. Mardt, B. de Silva, B. E. Husic, S. Klus, H. Wu, N. Kutz, S. L. Brunton, F. Noé, Uncertainty quantification by ensemble learning for computational optical form measurements. Mach Learn Sci Technol 2021, 3, 015009. https://doi.org/10.1088/2632-2153/ac0495.

[18] Steinbrener J, Posch K, Pilz J. Measuring the Uncertainty of Predictions in Deep Neural Networks with Variational Inference. Sensors (Basel). 2020 Oct 23;20(21):6011. doi: 10.3390/s20216011.

[19] Weiskopf D. Uncertainty Visualization: Concepts, Methods, and Applications in Biological Data Visualization. Front Bioinform. 2022 Feb 17;2:793819. doi: 10.3389/fbinf.2022.793819.

[20] Kompa B, Snoek J, Beam AL. Second opinion needed: communicating uncertainty in medical machine learning. NPJ Digit Med. 2021 Jan 5;4(1):4. doi: 10.1038/s41746-020-00367-3. PMID: 33402680; PMCID: PMC7785732.

TAs. Robust multi-reader multi-case observer performance studies[21] will be conducted to quantify the improvement in clinician performance with and without the tools. TA3 will also provide healthcare systems with robust field studies that integrate all Technical Areas (TAs) and evaluate their application in Clinical Decision Support (CDS).

TA3 performers will demonstrate, through pre-planned multi-reader multi-case observer performance studies (independent of the user preference studies), that the quantification and communication tools designed in TA3, combined with the auto-detection and auto-correction techniques developed in TA2, improve the performance of the AI-clinician team (e.g., reduce provider misdiagnosis) compared to clinicians operating alone, while also being efficient in terms of clinicians' interpretation time. Additionally, performers will demonstrate, through robust field studies that integrate all TAs, that the performance of the AI-clinician team is superior compared to clinicians using a static AI Decision Support Tool (AI-DST) model with a conventional user interface.

TA3 proposals will address the following topics:

1. Describe the uncertainty quantification methods to be compared, the performance measures for comparison, and any plans for how they would be improved for specific use cases.
2. Strong proposals will describe improvements to current uncertainty quantification methods and will also innovate to incorporate uncertainty quantification into the AI-DST methods to improve classification performance[22],[23]
3. Describe three AI-clinician communication tools to be investigated, including uncertainty quantification.
4. Strong proposals will describe tools and metrics to evaluate the success of communication, agile methods to iterate based on success, and how clinician feedback will be incorporated into the design.
5. Describe how the success of the tools and methods of communication will be measured using multi-reader multi-case observer performance studies.
6. Describe field-studies that would measure the performance of the AI-clinician team resulting from the integration of all program TAs into a combined tool suite. Performance would be compared to clinicians using a static AI-DST model with a conventional user interface.
7. Strong proposals will include field studies that measure performance across different clinical settings.
8. Describe supplementary metrics, if necessary, that may enable stakeholders to measure progress more accurately towards program goals.

The following are out of scope for TA3 proposals:

- Conducting field studies that affect clinical patient management without necessary IRB approvals and/or an FDA Investigational Device Exemption (IDE).

---

[21] Gallas BD, *et al.* Evaluating imaging and computer-aided detection and diagnosis devices at the FDA. *Acad Radiol*. 2012. doi: 10.1016/j.acra.2011.12.016

[22] Gundeep Arora et al. Leveraging uncertainty estimates to improve classifier performance. *Arxiv*. 2023. doi: arXiv:2311.11723v1

[23] Joshi P, Dhar R. EpICC: A Bayesian neural network model with uncertainty correction for a more accurate classification of cancer. *Sci Rep*. 2022. doi: 10.1038/s41598-022-18874-6.

1.1.6. TA4: Core Data Infrastructure

To optimize data sharing across performers and enable the program's use by community stakeholders, TA4 will oversee the program's data infrastructure. This includes coordinating, standardizing, and overseeing all data sharing activities across performers, and ensuring consistent alignment across phases. To achieve this, TA4 has the following objectives:

Objective 1: Provide coordinated data infrastructure capabilities for all program functions. TA4 will be responsible for coordinating cloud services that enable seamless data sharing and access. This includes overseeing data hosting activities and leveraging existing services whenever possible to reduce costs and ensure post-program sustainability. TA4 will also be tasked with developing performer platform needs, such as APIs, to support program functions.

TA4 will deliver common data infrastructure capabilities by leveraging existing data repositories and cloud computing resources to facilitate data sharing and minimize costs. The emphasis will be on utilizing prominent existing resources that enable continuous data sharing services, rather than reinventing novel approaches, to ensure efficiency and cost-effectiveness.

Objective 2: Embed interoperability principles to enable sharing, analysis, and integration with disparate systems. TA4 will prioritize alignment with the FAIR (findable, accessible, interoperable, and reusable) principles for data sharing, which are considered industry standards for research data, to ensure inclusive participation and effective resource utilization.

TA4 performers will not only host input data collected at different clinical sites but also aggregate analysis results from all performers, facilitating a consolidated post-market performance monitoring approach. This requires the implementation of common data sharing standards, open access resources, modifiable data architectures, data linkage features, and analysis and API querying capabilities. TA4 will enable post-market performance monitoring by providing data hosting and access features that allow TA2 performers and future stakeholders to test for degradation detection.

To achieve these objectives, TA4 will implement, extend, and adhere to standardized data models and common data standards jointly defined by all performers for sharing AI Decision Support Tool (AI-DST) input data, metadata, output, ground truth, and other relevant data elements. The aggregation of analyses and results across all performers, clinical sites, and use cases will create a consolidated, federated post-market performance monitoring infrastructure, including the sharing of surrogate ground truth labels and the provision of common reporting mechanisms to Government and private sector partners, such as the FDA, hospital centers, and non-profits that assist with sector-wide patient safety.

Objective 3: Utilize best practice privacy and security principles to ensure patient safety. TA4 performers will enhance the trustworthiness of data sharing by providing uniform security and privacy standards for post-market monitoring datasets

TA4 team will be expected to utilize industry standard practices to ensure that all data provided by performer teams is adequately protected and de-identified. This involves maintaining deidentification processes, privacy protection standards, and access control measures to ensure appropriate deidentification and consistent application of security practices throughout the program.

To achieve these objectives, TA4 will maintain best practice privacy and security standards to preserve data governance, chain of custody, patient privacy, and prevent unauthorized access to personal health information (PHI). These measures will ensure that patient data remains secure and confidential throughout the program, fostering trust among stakeholders and enabling the safe and effective use of shared data for post-market monitoring purposes.

TA4 proposals will address the following topics:

1. Describe common data infrastructure capabilities and APIs to be developed to facilitate data sharing among all performers and partnering clinical sites.
2. Describe approaches to enforce adherence with jointly defined data standards and protocols.
3. Describe approaches to solicit and respond to feedback from key public and private sector stakeholders regarding the reporting of information relevant to patient safety and the establishment of relevant open data standards and/or protocols.
4. Describe cloud computing and customization capabilities aligned with anticipated program needs and objectives.
5. Describe methods to aggregate analysis results from all performers, clinical sites, and use-cases to enable a consolidated, federated post-market performance monitoring approach.
6. Describe privacy and security principles to ensure patient confidentiality.

1.1.7. TA5: Independent Verification and Validation

The TA5 performer will test and validate results from TA1, TA2, and TA3 to confirm performers' progress. The team will be tasked with providing reports to performers and ARPA-H on analysis results, including subgroup analysis using demographic, clinical site, and other appropriate subgroups. Validation responsibilities include testing performer-developed methods and code with independently collected data to ensure the generalization of the reported methods to new data and new clinical sites, as well as the verification and validation of simulation tools developed by the performers.

TA5 will also provide system engineering oversight in roles similar to Chief Engineer or Chief Product Officer, ensuring that all performers adhere to industry standards for data engineering and data interoperability. To assist with successful integration across TA1, TA2, and TA3, they will provide integration oversight to ensure that deliverables and cross-TA collaboration are occurring as required with adequate documentation and communication. Towards supporting these efforts, TA5 proposals should focus on the following objectives:

Objective 1: Evaluate performer results using metrics that are both clinically relevant and statistically valid: The TA5 performer will evaluate both the quality of outputs and the validity of approaches for performers of TA1, TA2, and TA3. Performance metrics will be paired with statistically sound confidence intervals that take into account distributional characteristics with the aim of clearly and correctly mapping performance metrics to program outcomes. When appropriate, the TA5 performer will use validated open source-tools, with preference for FDA's regulatory science tools when available, to ensure openness and consistency.

Objective 2: Demonstrate the generalizability of performer results with independent data. The TA5 performer will work with clinical sites from other TAs, as well as additional clinical sites to independently collect data that will be used for IV&V. This includes independence across patients, time points, and clinical sites, as well as other potential variables relevant to the program metrics. The TA5 performer is expected to describe in detail how the methods for data collection and use by the IV&V team ensure generalizability. The TA5 performer will also verify and validate that the simulation tools developed by TA2 performers work as intended.

Objective 3: Oversee adherence to engineering and interoperability standards. The TA5 performer will also oversee that all performers adhere to industry standards for data engineering and data interoperability. This includes data governance, privacy protections and access control, ensuring that interoperability and data access do not present roadblocks to the execution of other TAs.

To achieve these objectives, the TA5 performer will use established guidelines for independent validation and performance reporting, including principles for unbiased data collection, reproducible reporting,

avoidance of data leakage between training and validation (test) data sets, and model verification and validation for quantifying and building credibility in numerical models. Whenever possible, the TA5 performer will use validated open source-tools, and whenever available, FDA's regulatory science tools to ensure openness and consistency.

TA5 proposals will address the following topics:

1. Describe planned collaboration with clinical sites and TA1-3 performers to collect data that will be used for independent verification and validation using the metrics listed in Figure 3. This includes a description of collaboration with TA1-3 performers to ensure that the IV&V data sets are appropriately sized to reach valid statistical conclusions.
2. Describe approach to provide timely evaluations through intermediate testing, in coordination with TA 1-3 performers.
3. Explain in detail the measures that will be taken to ensure that the IV&V data sets will be independent from the data sets that TA1-3 performers have used to perform training or preliminary testing. Independence across patients, time points, and clinical sites, as well as other potential variables are relevant considerations.
   Strong TA5 performers will use additional data, collected at additional independent clinical sites, to ensure generalization.
4. Describe capabilities for defining and using additional metrics as required by different use cases when necessary.
5. Describe plans for the use of validated statistical tools so that the estimates lead to proper conclusions.
6. When the data collection for IV&V-related tasks involves human expert judgements (e.g., concordance with clinicians in TA1 or clinician decision making in TA3), describe the principles that will be used, and the type of collaboration that will be established with TA1-3 performers to ensure that the data collection methodology is appropriate and leads to unbiased analysis.
7. Describe plans to combine individual performance metrics to generate wholistic conclusions that reflect program intent and communicate program success.
8. Describe plan for coordinating data standards and interoperability across performer teams. Examples of prior experience in conducting such activities are welcome.

1.1.8. **Program Milestones & Metrics**

Figure 2: Program Timeline and Milestones

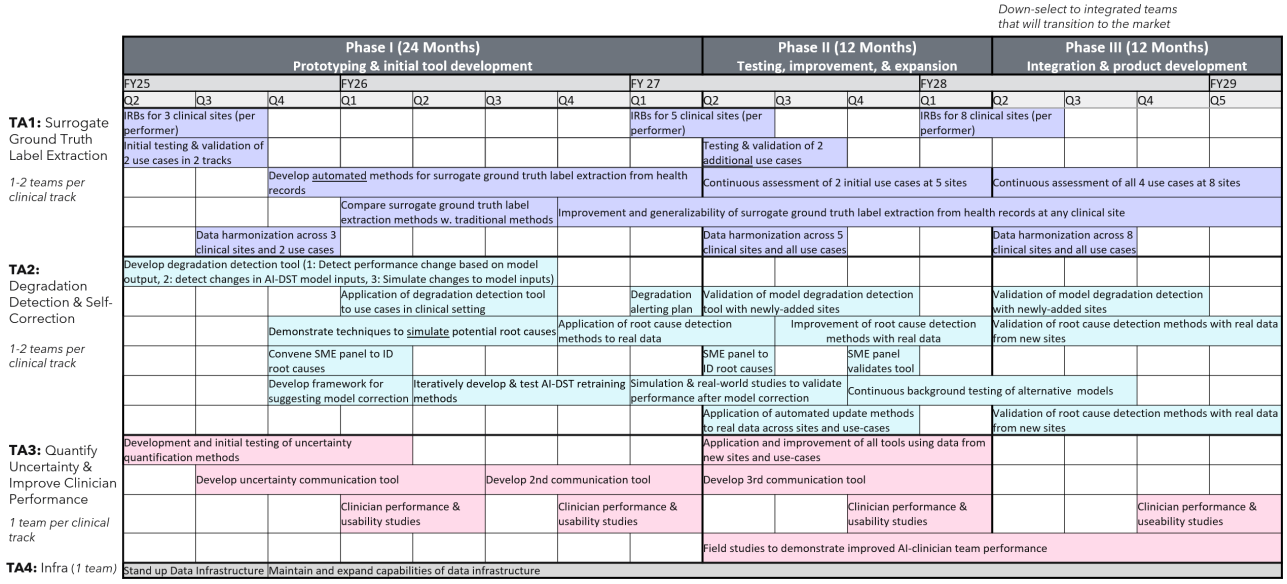


Figure 3: Program Metrics

	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
TA1: surrogate ground truth label Extraction	**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* (% agreement of the automated surrogate ground-truth method and the manual ground-truth determined by clinicians using the same EHR data on a randomly selected subset of cases)	≥ 95%	≥ 99%	≥ 99%
	Agreement with manual ground truth* (% 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%
TA2: Degradation Detection & Self-Correction	**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 model performance)	≥ 70%	≥ 80%	≥ 90%
	Accuracy of root cause (% of degraded models with a correctly identified root cause)	≥ 70%	≥ 85%	≥ 90%
	Mitigation of model degradation** (% of addressable degraded models 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** (% e.g., reduction in false-negatives at a given specificity)	≥ 40% reduction in ≥ 50% of use cases	≥ 50% reduction in ≥ 75% of use cases	≥ 70% reduction in 100% of use cases
TA3: Quantify Uncertainty & Improve Clinician Performance	**Error reduction for the AI model using uncertainty estimation method** (% reduction relative to baseline set at 6 months)	≥ 15%	≥ 30%	≥ 45%
	Tool useability (average Likert score [1-10 scale] for tool usability as judged by clinicians)	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 accuracy (reduced provider misdiagnosis, e.g., % reduction in false-negatives at a given specificity)	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 (Time for case interpretation with TA3 tools relative to no tools.)	1.2X	1X	0.8X
TA4: Core Data Infrastructure	**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)	15 sites	25 sites	31 sites

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

** 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

ARPA-H will meet with PRECISE-AI performers at least monthly to review progress towards the metrics and milestones defined above. Performers should also propose additional quantitative metrics and milestones appropriate to their specific approach for each Phase of the program that will demonstrate progress towards the program's goals. Progress toward metrics as agreed to by ARPA-H is the basis for initiation of the optional Phases.

Commercial Transition
In Phase III, performers will integrate their component technologies into various commercial transition packages. Examples of potential commercial transition packages include:

- **An open-source tool suite for clinical AI developers** that incorporates the innovations from TA1-TA4, allowing for continuous monitoring, degradation detection, root cause analysis, and self-correction of AI models. A technically rigorous tool suite with commercial-friendly open-source licenses would introduce scientifically grounded performance monitoring technologies into the software supply chain and encourage collaboration among industry partners, academic institutions, and healthcare organizations, fostering widespread adoption and continuous improvement based on real-world feedback.
- **A self-monitoring medical device** that incorporates the program's technologies and is developed in partnership with medical device manufacturers. These devices would be designed to continuously monitor their own performance, detect any degradation, and either recommend or apply self-correction techniques to maintain optimal performance. Such devices will need to obtain necessary regulatory approvals, like FDA clearance, before being marketed.
- **Individual hospital performance monitoring capabilities** to ensure the ongoing safety and effectiveness of the many AI-powered medical devices found in hospitals or clinics. This would enable healthcare providers to proactively identify and address any performance issues.
- **Multi-organizational performance monitoring capabilities** that enable the assessment of patient safety across multiple hospital systems, the sharing of data and insights related to AI model performance, the early detection and mitigation of potential risks across patient populations and healthcare providers.

Throughout the program, performers will work closely with government, private sector, and hospital partners to refine the commercialization approach for the TA1-TA4 components, informed by the outcomes of independent evaluations of the technology and the needs of partners who seek to monitor device safety. To ensure the long-term sustainability and equitable proliferation of these technologies, performers will develop the technology(ies) in manner that creates incentives for ongoing investment, development, and adoption. Examples may include establishing public-private partnerships, pursuing reimbursement models that support the use of AI-powered tools, and providing education and training programs to help healthcare professionals effectively integrate these technologies into their clinical workflows.

1.1.9. Proposal Scope

Each proposal responsive to this Solicitation can address a TA among TA1, TA2, TA3, TA4 and TA5, or a combination of TA1, TA2, and TA3. Proposals that address TA4 or TA5 cannot address any of the other TAs. An organization may act as a primary performer on one TA and a sub-performer on a separate proposal addressing the same TA. An organization may act as a sub-performer on multiple proposals addressing the same TA.

All proposals addressing a single TA are expected to collaborate with performer teams addressing other TAs and all proposals addressing TAs 1-2 are expected to collaborate with TA3 and TA4. Collaboration requirements between TA teams are described in Figure 5 below. Multiple awards are anticipated for TA1, TA2, and TA3 to ensure sufficient diversity in clinical use-cases, to apply program tools to different hospitals, and to foster a diversity of solutions. For TA4, a single award is anticipated. Likewise, for TA5, a single award is anticipated.

Clinical AI-DSTs are often developed around a specific input data type. Proposers to TA1 or TA2 must align their technologies to specific clinical tracks (Figure 4) with preference for technologies that can be

generalized outside of their chosen track. Proposers to TA3 are expected to develop technologies that are largely generalizable across clinical tracks. Proposals that apply to a combination of TA1, TA2, and TA3 may submit an alternative use case outside of the specified clinical tracks.

Proposals are expected to include all the required expertise to achieve the goals the TA to which they are proposing. Specific content, communications, networking, and team formation are the sole responsibility of the proposing teams. **Proposers must submit a single proposal led by one PI under a single prime performer that addresses program Phase I and Phase II relevant to the proposed TA(s)**. To minimize risk and facilitate effective integration of TAs into a combined tool suite, we anticipate that progression to Phase III will involve a down-select and will require performers to re-propose as self-selected teams that collectively address TAs 1-3.

Figure 4. PRECISE-AI Proposal Scope

Clinical Track	Example Clinical 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 communication of 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
Other	The use case should be selected so that the surrogate ground truth label for the patient can be extracted from the data in EHR or other clinical reports with sufficient accuracy in a reasonable time frame. Proposals that include use cases in the "other" category will be required to submit a joint proposal that covers a combination of TA1, TA2 and TA3.

1.1.10. **Common Requirements for All Proposals**

1. Proposals must include a management plan that describes how the team will be managed and how the various team contributions will be designed with interoperability in mind to enable eventual integration with other TAs. Proposals should provide a detailed plan for coordination, including explicit plans for interaction among collaborators/subcontractors of the proposed effort.
2. Intellectual Property (IP) rights asserted by any performers for technologies created under this program must be aligned with open-source regimes (except for commercial AI-DST that are made available to support performance monitoring research). Performers must complete the necessary Associate Performer Agreements to ensure the program goals are not hindered by restrictive IP claims. Commercially available tools may be leveraged for and/or incorporated into proposed solutions. Licensing details (costs, restrictions, etc.) must align with the overall goals of the program and must not inhibit collaboration between performers, hospital partners, or the Government.
3. Proposals will include a detailed budget for Phases I and II, and a rough order of magnitude budget estimate for Phase III.
4. Proposals must describe how the source code for methods, technologies, and tools developed in the program will be shared, either through existing platforms or in collaboration with TA4.

5. All TA1, TA2, TA3, and TA5 budgets should exclude costs for cloud storage and cloud computing. Strong proposals will primarily leverage TA4-furnished cloud computing resources as opposed to separately budgeting for on-site computing / storage. If TA4 proposers anticipate using a commercial cloud vendor, they should clarify which one and approximate the amount of storage and compute that will be necessary to support program goals.

1.1.11. Collaboration and Data Sharing

The ARPA-H PRECISE-AI program will be developed by several "performers" that include contractors and subcontractors, to include those with deep knowledge of key data assets as well as those selected through this announcement or through complementary funding mechanisms at partner organizations. **Therefore, it is expected that all performers will interact and work collaboratively with other performers in developing the methods, technologies, and tools using open, timely, and effective communication, information exchange, and reporting.** Performers across all partner organizations will attend common meetings and technical exchanges to advance relevant technologies, bridge across data siloes, and move toward a common care delivery platform across numerous clinical use cases.

To facilitate the open exchange of information described above, **performers will have Associate Contractor Agreement (ACA) language included in their award.** Each performer will work with other PRECISE-AI performers to develop an ACA that specifies the types of information that will be freely shared across performer teams. The open exchange of scientific information will be critical in advancing the software research required to achieve the PRECISE-AI objectives. The ACA will establish a common understanding of expectations to guide the open exchange of ideas and establish a collaborative foundation for the PRECISE-AI program. Each performer will work with other performers as described in the Program Collaboration Requirements (Figure 5).

Figure 5: Program Collaboration Requirements

TA	Collaboration Expectations
All	Throughout the program, all performers will work with an Independent Verification and Validation (IV&V) team established by ARPA-H. IV&V expectations are described in Section 1.4.10. All TA performers, in collaboration with ARPA-H, and the IV&V team, will align on technical standards for data storage and sharing, including common data standards, formats, specifications and APIs to enable consistency and accessibility across all performers while preserving data provenance and patient privacy. All TA performers selected to proceed to Phase III should collaborate to identify mature clinical use cases of self-monitoring or self-correcting AI tools that will be transitioned to the market or submitted to the FDA for clearance or approval by the end of Phase III. All performers should collaborate on creation and validation of required datasets and creation of legal filings.
TA1	**With TA1:** Clinical sites will provide data for models and validation for each use case, including sites with different clinical practices, resources, healthcare data infrastructures and information coding systems. TA1 performers working on use cases in similar tracks should coordinate among themselves and other community stakeholders on data elements.

	With TA2: TA1 should align their technologies with the program clinical tracks and will collaborate closely with TA2 performers who have selected the same clinical track. TA1 performers should collaborate with TA2 performers to provide access to surrogate ground truth datasets for clinical use cases and the AI-DST input data, utilizing the data lake infrastructure provided by TA4. TA1 performers should provide TA2 performers with regular access to disease state subject matter experts (e.g. via hospital partnerships) for validation of algorithms updated in TA2. TA1 performers and clinical SMEs should assist TA2 performers with periodic verification of TA2 efforts, including drift detection, model updating, and re-validation of performance.
	With TA3: TA1 performers should collaborate with TA3 performers to provide access to clinical SMEs for participation in user-centered design exercises and usability testing of tools developed to communicate uncertainty in AI-based decision making. TA1 performers should collaborate with TA3 performers to provide access to clinical SMEs in running studies on risk communication tools and conducting studies to test for improvement in clinician performance.
	With TA4: TA1 performers and associated clinical sites should collaborate with the TA4 performer to provide seamless data import from the EHR to the shared data repository, in standard-conforming formats amenable to data harmonization. Coordinate a data interoperability "bake-off" where data interchange, data access and data governance across all TA1 and TA4 performers are pressure-tested, to ensure adequate performance of the FAIR data repository.
	With TA5: TA1 performers should collaborate with the TA5 performer to ensure that (i) data used for IV&V is independent from the data that is used to perform training or preliminary testing (ii) data collection methodology is appropriate and leads to unbiased analysis by TA5; and (iii) TA5 has all the code and documentation needed for IV&V.
TA2	**With TA1:** Early in Phase I, TA2 performers will provide access to AI-DSTs for two use cases from different clinical tracks (Figure 4) to TA1 performers for testing. In Phase II, TA2 performers will provide access to AI-DSTs for two additional use cases to TA1 performers for continued testing. TA2 should collaborate with TA1 to gain subject-matter expert access for clinical validation of results.
	With TA2: TA2 performers should collaborate with each other on the simulation tools or other approaches that they develop, and designs for identification of a set of potential root causes for performance degradation from a diverse panel of experts.
	With TA3: TA2 performers are encouraged to share the source code for selected AI-DSTs with TA3 performers to facilitate the development of tools for uncertainty quantification and explainability. TA2 should coordinate with TA3 by allowing for a review of the degradation alerting and model self-correction algorithms while also providing feedback necessary for TA3 to complete their deliverables.
	With TA4: TA2 performers should coordinate data governance and access to clinical use cases with TA4 to enable the execution of TA2 objectives. TA2 should provide TA4 all AI-DST models used in their evaluations.

	With TA5: TA2 performers should collaborate with the TA5 performer to ensure that (i) data used for IV&V is independent from the data that is used to perform training or preliminary testing (ii) any information required by the TA5 performer for IV&V of simulation tools developed by TA2 is provided; and (iii) TA5 has all the code and documentation needed for IV&V.
TA3	**With TA1:** TA3 should align their technologies with the program clinical tracks and will collaborate closely with TA2 performers who have selected the same clinical track. TA3 performers should collaborate with clinical SMEs of TA1 performers to (a) perform design exercises for communicating the multi-dimensional information derived from the AI-DST efficiently and effectively to clinicians, and (b) run multi-reader multi-case (MRMC) and field studies to compare data interpretation scenarios under different information communication scenarios.
	With TA2: TA3 performers should coordinate with TA2 to access their update algorithms and soliciting feedback necessary for the execution of TA3 objectives
	With TA3: TA3 performers should collaborate with each other to quickly identify which communication tools and methods are successful/unsuccessful in order to minimize the number of iterations to arrive at a successful design for a particular use case.
	With TA4: TA3 performers should work with TA4 performers to ensure round-trip data integration and communication between TA3 tools and the centralized data repository, for all clinical cases.
	With TA5: TA3 performers should collaborate with the TA5 performer to ensure that (i) data used for IV&V is independent from the data that is used to perform training or preliminary testing (ii) data collection methodology is appropriate and leads to unbiased analysis by TA5; and (iii) TA5 has all the code and documentation needed for IV&V.
TA4	**With all TAs:** The TA4 performer should provide access to the centralized data and AI-DST model repository.
	With TA1: The TA4 performer will coordinate with TA1 and TA2 performers to ensure that all data elements from separate use cases are appropriately imported into the data platforms. TA4 performers should coordinate a data interoperability “bake-off” where data interchange, data access, and data governance across all performers are pressure-tested to ensure adequate performance of the data repository.
TA5	**With all TAs:** The TA5 performer should work with all TAs to ensure that (i) the data used for IV&V activities collected at program sites is independent from training and test data used by the other performers; and (ii) the IV&V data sets are appropriately sized to reach valid statistical conclusions.

1.1.12. **Open Software Standards**

Performers will be expected to adhere to all relevant Government laws and policies applicable to data and information systems and technologies, including but not limited to:

- Common IT Security Configurations
- Federal information technology directives and policies
- Section 508 of the Rehabilitation Act of 1973 (29 USC 794d) as amended by P.L. 105-220 under Title IV (Rehabilitation Act Amendments of 1998)
- National Institute of Standards and Technology (NIST) Risk Management Framework Special Publications

A key goal of this program is to seed the establishment of a sustainable open-source ecosystem for automated performance degradation detection and remediation. Thus, it is desired that all non-commercial software (including source code), software documentation, and technical data generated by the program, be provided as deliverables to the Government with open-source or unlimited rights, and all hardware designs and documentation be provided with a minimum of Government Purpose Rights (GPR), as lesser rights may negatively impact the potential for this health IT ecosystem to become self-sustaining. Open-source code is highly encouraged using permissive, business-friendly open-source licenses such as CC-BY, BSD, MIT, Apache 2.0 or similar. Approaches that inhibit this objective are not desired and would adversely affect the PRECISE-AI program goals and objectives.

All data generated by new devices and technology created in the PRECISE-AI program must be adherent to relevant standards established or endorsed by ONC (e.g., HL7, FHIR, DICOM, LOINC, SNOMED CT, USCDI, and USCDI+). It is expected that all performers will work together to converge on standards and APIs to ensure interoperability across prototype capabilities. All performers are expected to follow agile software development processes. Whenever an existing standard is available that meets technical needs of the program, performers must use the existing standard instead of creating their own. In cases where an existing standard provides only partial functionality, performers should extend the existing standard in a fully backwards compatible manner and create the documentation needed for ONC to evaluate extensions for inclusion in the national standard.

All Application Programming Interfaces (APIs) developed for PRECISE-AI must be founded on open standards and models such as REST, JSON, JSON-LD and utilize standard data models and ontologies if available. All data elements and structures in API calls must be mapped to a data dictionary that references the standards.

Finally, performers may not build on top of limited rights components (i.e. existing proprietary software) unless they have prior approval in writing from the program manager and the agreements officer.

1.1.13. **Commercial Transition Support**

Proposers who are selected for an ARPA-H award may apply for a team of non-government advisors known as Entrepreneurs in Residence (EIR) / Experts in Residence (XIR). In coordination with the PM, the EIR/XIR will provide commercial transition support to the awardee. The goal is to offer complementary capabilities to the team; hence, the extent of the work is flexible. Examples of tasks may include cost modeling, end-user engagement, market analysis and mapping, competitive analysis, techno-economic analysis, manufacturing and scale-up strategy, intellectual property (IP) securement strategy, and financial plan creation. All commercialization and transition activities should align to the technology's stage of maturity. EIRs/XIRs will work closely with ARPA-H's Project Accelerator Transition Innovation Office (PATIO) team to leverage its extensive network of U.S. investors, strategic partners, and mentors.

Participation in the program is voluntary but recommended. Performers are not expected to form a new

company or leave their current research positions to pursue transition. Instead during the program, performers should identify appropriate partners for enabling transition.

1.1.14. **Equity Requirements**

ARPA-H is committed to equitable health care access irrespective of race, ethnicity, gender/gender identify, sexual orientation, disability, geography, employment, insurance, and socioeconomic status. Equity considerations will be reviewed throughout the program, including review of milestone reports, deliverables and evaluations to ensure the prioritization of equity.

All proposers must articulate how they will incorporate equity considerations. Proposals covering TA1 must describe a plan to select clinical sites from diverse geographic locations and sites with diverse patient populations. Proposals covering TA2 are required to describe plans to incorporate features that account for variations among patient sub-population in their degradation detection (TA2.1), root cause analysis (TA2.2) and model self-correction (TA2.3) tools. This will ensure that the AI tools are designed to continuously measure AI model performance degradation across different subpopulations and suggest corresponding corrections. This will also ensure that changes leading to different degrees of performance degradation for different subpopulations (i.e., changes that lead to a bias among subpopulations) are detected and mitigated quickly and effectively. Proposals covering TA3 must detail how their innovations and deliverables will consider multiple dimensions of diversity in the communication of AI model performance to different stakeholders, including clinicians, developers, and regulators.

Performers involved in the handling of personalized and/or identified demographics or health data must ensure appropriate privacy and security standards are met. All proposers should outline equity goals, potential risks, potential ramifications of not meeting equity goals and risk mitigation strategies.

2. PS AWARD INFORMATION

This PS may result in multiple awards of Other Transaction (OT) agreements or no award at all. However, the number of awards selected will depend on the quality of the proposals received and the availability of funds. If warranted, portions of resulting awards may be segregated into pre-priced options. In the event that the Government desires to award only portions of a proposal, negotiations will commence upon selection notification. The Government reserves the right to fund proposals in phases, with options for continued work, as applicable. The Government reserves the right to request any additional documentation to support the negotiation and award process. The Government reserves the right to remove a proposal from award consideration should the parties fail to reach agreement on award terms, conditions, cost, and/or the proposer fails to provide requested additional information in a timely manner. The Government Agreements Officer (AO) shall have sole discretion to negotiate all agreement terms and conditions with selected proposers.

The Government reserves the right to award an OT or make no award at all. A Model Agreement with basic terms and conditions has been posted with this PS (See additional details in Section 4.3). Proposers may submit red-line edits to the basic terms and conditions of the resulting instrument; however, the Government AO shall have sole discretion to negotiate any red-line edits that deviate from the basic terms and conditions. A resulting OT Agreement will not require cost-sharing unless a requirement is stated elsewhere in this PS; however, ARPA-H reserves the right to negotiate cost-sharing as appropriate to the situation.

While scientific publications are highly encouraged, ARPA-H will apply publication or other restrictions, as necessary, if it is determined that the research resulting from the proposed effort will present a high

likelihood of disclosing sensitive information including Personally Identifiable Information (PII), Protected Health Information (PHI), financial records, proprietary data, and any information marked Sensitive but Unclassified (SBU), Controlled Unclassified Information (CUI), etc. Any award resulting from such a determination will include a requirement for ARPA-H permission before publishing any information or results on the effort.

3. ELIGIBILITY INFORMATION

3.1. ELIGIBLE APPLICANTS

All responsible sources capable of satisfying the Government's needs may submit a proposal to this PS, except as noted below. Specifically, universities, non-profit organizations, small businesses and other than small businesses are eligible and encouraged to propose to this PS.

3.2. PROHIBITION OF PERFORMER PARTICIPATION FROM FEDERALLY FUNDED RESEARCH AND DEVELOPMENT CENTERS (FFRDCS) AND GOVERNMENT ENTITIES

ARPA-H is primarily interested in responses to programs from commercial performers, academia, non-profit organizations, etc. In certain circumstances, FFRDCs and U.S. Government Entities will have unique capabilities that are not available to proposing teams through any other resource. Accordingly, the following principles will apply to this solicitation:

(1) FFRDCs and U.S. Government entities, including federal Government employees, are not permitted to respond to this solicitation as a prime or sub-performer on a proposed performer team.

(2) If an FFRDC or U.S. Government entity has a unique research idea that is within the technology scope of this solicitation that they would like considered for funding, contact this email address: PRECISE-AI@arpa-h.gov.

(3) If an FFRDC or U.S. Government entity, including a federal Government employee, is interested in working directly with the Government team supporting the research described by this solicitation, the party should contact PRECISE-AI@arpa-h.gov.

3.3. NON-U.S. ORGANIZATIONS

Non-U.S. entities may participate to the extent that such participants comply with any necessary nondisclosure agreements, security regulations, export control laws, and other governing statutes applicable under the circumstances. However, non-U.S. entities are encouraged to collaborate with domestic U.S. entities. In no case will awards be made to entities organized under the laws of a covered foreign country (as defined in section 119C of the National Security Act of 1947 (50 U.S.C. § 3059)); a foreign entity of concern meeting any of the criteria in section 10638(3) of the CHIPS and Science Act of 2022; or an individual that is party to a malign foreign talent recruitment program, as defined in Section 10638(4) of the CHIPS and Science Act of 2022.

3.4. ORGANIZATIONAL CONFLICTS OF INTEREST (OCI)

Proposers are required to identify and disclose all facts relevant to potential OCIs involving the proposer's

organization and any proposed team member (proposed subawardee). Although the FAR does not apply to OTs, ARPA-H requires OCIs be addressed in the same manner prescribed in FAR subpart 9.5. Regardless of whether the proposer has identified potential OCIs under this section, the proposer is responsible for providing a disclosure with its proposal[24]. The disclosure must include the proposers', and as applicable, proposed team members' OCI mitigation plans. The OCI mitigation plan(s) must include a description of the actions the proposer has taken, or intends to take, to prevent the existence of conflicting roles that might bias the proposer's judgment and to prevent the proposer from having unfair competitive advantage. The OCI mitigation plan will specifically discuss the disclosed OCI in the context of each of the OCI limitations outlined in FAR 9.505-1 through FAR 9.505-4. The disclosure and mitigation plan(s) do not count toward the page limit.

3.5. AGENCY SUPPLEMENTAL OCI POLICY

In addition, ARPA-H restricts performers from concurrently providing professional support services, including Advisory and Assistance Services or similar support services, and being a technical performer. Therefore, as part of the FAR 9.5 disclosure requirement above, a proposer must affirm whether the proposer or any proposed team member (proposed subawardee, etc.) is providing professional support services to any ARPA-H office(s) under: (a) a current award or subaward; or (b) a past award or subaward that ended within one calendar year prior to the proposal's submission date.

If any professional support services are being or were provided to any ARPA-H office(s), the proposal must include:

- The name of the ARPA-H office receiving the support;
- The prime contract number;
- Identification of proposed team member (proposed subawardee) providing the support; and
- An OCI mitigation plan in accordance with FAR 9.5.

3.6. GOVERNMENT PROCEDURES

The Government will evaluate OCI mitigation plans to avoid, neutralize, or mitigate potential OCI issues before award and to determine whether it is in the Government's interest to grant a waiver. The Government will only evaluate OCI mitigation plans for proposals selected for potential award under the PRECISE-AI PS evaluation criteria and funding availability.

The Government may require proposers to provide additional information to assist the Government in evaluating the OCI mitigation plan.

If the Government determines a proposer failed to fully disclose an OCI; or failed to provide the affirmation of ARPA-H support as described above; or failed to reasonably provide additional information requested by the Government to assist in evaluating the proposer's OCI mitigation plan, the Government may reject the proposal and withdraw it from consideration for award.

3.7. RESEARCH SECURITY DISCLOSURE

In accordance with National Security Presidential Memorandum (NSPM)-33, Presidential Memorandum on United States Government-Supported Research and Development National Security Policy, research

[24] Entities are encouraged to reach out, via the Q&A process or by contacting the PS coordinator directly, prior to proposal submission if a potential OCI issue exists. The Government will not review a OCI mitigation plan at this time, but feedback regarding the potential OCI may be provided.

organizations should identify and mitigate conflicts of commitment and conflicts of interest (CoC/CoI) to receive federal funding. Research organizations submitting a proposal in response to this PS must provide documentation for Senior/Key Personnel when requested for ARPA-H to determine whether there is any CoC/CoI risk. The format for this submission can be found in the Administrative & National Policy document.

4. PROPOSAL AND SUBMISSION INFORMATION

4.1. GENERAL GUIDELINES

- Proposers must first submit a Solution Summary in order to submit a full proposal
- Solution Summaries are due November, 2024 (specific date and time TBD)
- Full Proposals are due January, 2025 (specific date and time TBD)
- All submissions must be written in English with type not smaller than 12-point font. Smaller font may be used for figures, tables, and charts.
- Do not include elaborate brochures or marketing materials; only include information relevant to the submission requirements or evaluation criteria.
- Use of a diagram(s) or figure(s) to depict the essence of the proposed solution is permitted.
- All submissions shall be unclassified.
- Proposers are responsible for clearly identifying proprietary information.
- Submissions containing proprietary information must have the cover page and each page containing such information clearly marked with a label such as "Proprietary" or "Company Proprietary." NOTE: "Confidential" is a classification marking used to control the dissemination of U.S. Government National Security Information as dictated in Executive Order 13526 and should not be used to identify proprietary business information.
- ARPA-H will post a consolidated Questions & Answers document on a regular basis. See Section 7.
- Submissions sent through other mediums/channels other than what is prescribed herein, or after the prescribed PS deadline will not be considered, reviewed, nor evaluated.

4.2. SOLUTION SUMMARY RESPONSES

SOLUTION SUMMARY CONTENT AND FORMATTING:

The required Solution Summary is a mechanism for potential proposers to get feedback prior to investing resources for a full proposal. All Solution Summaries submitted in response to this Solicitation must comply with the content, page, and formatting requirements in Appendix A. Potential proposers are strongly encouraged to use the template provided. Information not explicitly requested in this PS may not be reviewed.
NOTE: No awards will be made, nor funding provided as a result of Solution Summary Submissions.

SOLUTION SUMMARY SUBMISSIONS:

Solution Summaries shall be submitted to the ARPA-H Solution Site at https://solutions.arpa-h.gov/
Solution Summaries submitted incorrectly (e.g. not submitted to the ARPA-H Solutions Site by the due date and time) may not be reviewed.

ARPA-H will provide written feedback to all Solution Summary submissions. Feedback at a minimum will provide an encourage or discourage recommendation in submitting a proposal to the ARPA-H PRECISE-AI Solicitation. Feedback will be sent to the administrative and technical points of contact noted on the Solution Summary cover page.

4.3. PROPOSAL INSTRUCTIONS

4.3.1. Proposal Volume Templates

Proposers must provide the following information when submitting a proposal. Template documents and instructions for all volumes are provided along with this PS, in the Other Transactions Bundle Templates. Failure to utilize the templates and/or provide the information requested may result in a proposal being deemed non-conforming and/or delay the evaluation process discussed in Section 5.2. Proposals should express a consolidated effort in support of one or more related technical concepts or ideas addressing one TA.

Volume 1 must consist of the following two documents:

TECHNICAL & MANAGEMENT DOCUMENT
TASK DESCRIPTION DOCUMENT

The maximum page count for the Technical & Management document is 25 pages for TA1, TA3, TA4, and TA5 proposals, 30 pages for TA2 proposals, 40 pages for proposals that jointly address two areas among TA1, TA2, and TA3, and 50 pages for proposals that address all of the first three TAs (TA1, TA2 and TA3) jointly. The Technical & Management proposal may include an attached bibliography (excluded from the page limit) of relevant technical papers or research notes (published and unpublished) that document the technical ideas and approach upon which the proposal is based. Copies of not more than three relevant papers may be included with the submission and will be excluded from the page limit. Resumes for proposed key personnel must be included in this volume. Resumes will be excluded from the page limit. Documentation of current Assurance of Compliance with federal regulations for human subjects protection must also be included as an attachment in Volume 1 (per Section 6.3.4) and will be excluded from the page limit. The submission of other supporting materials along with the proposal is strongly discouraged. These materials will not be considered for evaluation.

The Task Description Document (TDD) must include objectives and associated tasks, aligned with the program milestones. All objectives and tasks must correspond with program phases as described in the Technical Area sections of the PS. The Task Description Document does not have a page limit.

Volume 2 must consist of the following documents (no page limit):

PRICE/COST PROPOSAL
MODEL AGREEMENT

The Volume 2 Price/Cost proposal must consist of the provided budget spreadsheet and an accompanying budget narrative. The budget narrative must provide enough supporting documentation to justify all elements presented in the budget spreadsheet.

It is expected that the effort will leverage all available relevant prior research to obtain the maximum benefit from the available funding. Collective proposals covering a combination of TA1, TA2, and TA3 must include a description of cost efficiencies gained through the combined proposal. ARPA-H recognizes that undue emphasis on cost may motivate proposers to offer low-risk ideas with minimum uncertainty and to staff the effort with junior personnel to be in a more competitive posture. ARPA-H discourages such cost strategies.

Volume 3 must consist of the following documents:

ADMINISTRATIVE & NATIONAL POLICY REQUIREMENTS

4.3.2. Model Other Transaction Agreement

Prior to submitting a proposal, proposers must review the model OT that is provided as an attachment to this PS (included within Attachment 2). ARPA-H has provided the model OT to expedite the negotiation and award process. The model OT is representative of the terms and conditions that ARPA-H intends to include in the resulting award.

Proposers may suggest edits to the model OT for consideration by ARPA-H and provide a copy of the model OT with track changes as part of the proposal package. It is required that Proposers include comments providing rationale for any suggested edits of a non-administrative nature. Suggested edits may be rejected at ARPA-H's discretion. Proposal information should not be included in the model OT or in comments on the model OT. Any questions, comments, or edits to the model OT will not be considered in the evaluation of the proposal.

4.4. PROPOSAL DUE DATE AND TIME

Proposals are due no later than 2:00 PM ET on [Date]. Full proposal packages as described in Section 4.3 must be submitted per the instructions outlined in this PS and in accordance with Attachment 2 and received by ARPA-H no later than the above time and date. Proposals received after this time and date will not be reviewed. Proposals shall be submitted to https://solutions.arpa-h.gov .

Proposers are warned that the proposal deadline outlined herein is in Eastern Standard Time and will be strictly enforced. When planning a response to this notice, proposers should consider that some parts of the submission process may take from one business day to one month to complete.

5. EVALUATION OF PROPOSALS

5.1. EVALUATION CRITERIA FOR AWARD

Proposals will be evaluated using the following evaluation criteria, listed in descending order of importance.

- **OVERALL SCIENTIFIC AND TECHNICAL MERIT**
  The proposed technical approach is innovative, feasible, achievable, and complete. Task descriptions and associated technical elements provided are complete and in a logical sequence with all proposed deliverables clearly defined such that a final outcome that achieves the goal can be expected as a result of award. The proposal identifies major technical risks and planned mitigation efforts are clearly defined and feasible. In addition, the evaluation will take into consideration the extent to which the proposed intellectual property (IP) rights structure will potentially impact the Government's ability to transition the technology.

- **PROPOSER'S CAPABILITIES AND/OR RELATED EXPERIENCE**
  The proposed technical team has the expertise and experience to accomplish the proposed tasks. The proposer's prior experience in similar efforts clearly demonstrates an ability to deliver products that meet the proposed technical performance within the proposed budget and schedule. The proposed

team has the expertise to manage the cost and schedule. Similar efforts completed/ongoing by the proposer in this area are fully described including identification of other Government entities.

- **POTENTIAL CONTRIBUTION TO RELEVANCE TO THE ARPA-H MISSION**
  Potential future R&D, commercial, and/or clinical applications of the project proposed, including whether such applications may have the potential to address areas of currently unmet need within biomedicine and improve health outcomes. Degree to which the proposed project has the potential to transform biomedicine is an important factor. Potential for the project to take an interdisciplinary approach is also valuable.

- **COST REALISM**
  Price and value analysis will be performed on each proposal to assess the reasonableness and value the overall proposed price provides the Government for the technical solution selected.

  When price and value analysis are inconclusive, cost realism analysis may be performed to ensure proposed costs are realistic for the technical and management approach, accurately reflect the technical goals and objectives of the solicitation, the proposed costs are consistent with the proposer's TDD and reflect a sufficient understanding of the costs and level of effort needed to successfully accomplish the proposed technical approach. The costs for the prime proposer and proposed sub-awardees should be substantiated by the details provided in the proposal (e.g., the type and number of labor hours proposed per task, the types and quantities of materials, equipment and fabrication costs, travel and any other applicable costs and the basis for the estimates).

  Proposals must include a cost point that is commensurate with the scale and complexity of the proposed technical and management approach. Proposers should ensure that budgets align to the needs of the work being proposed. Budgets should focus on the tasks essential for achieving program goals and associated risk mitigation strategies. Budgets that are unrealistically high will result in extensive revisions and negotiations.

5.2. REVIEW AND SELECTION PROCESS

It is the policy of ARPA-H to ensure impartial, equitable, comprehensive proposal evaluations based on the evaluation criteria listed above and to select the source (or sources) whose offer meets the Government's technical, policy, and programmatic goals.

ARPA-H will conduct a scientific and technical review of each conforming proposal. All proposal evaluations will be based solely on the evaluation criteria in Section 5.1.

Relative to the evaluation criteria, the Government will evaluate each conforming proposal in its entirety, documenting the strengths and weaknesses. Based on the identified strengths and weaknesses, ARPA-H will determine whether a proposal will be selected for award. Proposals will not be evaluated against each other during the scientific review process, but rather evaluated on their own individual merit to determine how well the proposal meets the criteria stated in this PS.

An award will be made to a proposer(s) whose proposal is determined to be selectable by the Government, consistent with the instructions and evaluation criteria specified herein and based on the availability of funding. Given the limited funding available, not all proposals considered selectable may receive an award and funding.
For the purposes of this proposal evaluation process, a selectable proposal is defined as follows:

SELECTABLE: A selectable proposal is a proposal that has been evaluated by the Government against the evaluation criteria listed in the PS, and the positive aspects of the overall proposal outweigh its negative aspects. Additionally, there are no accumulated weaknesses that would require extensive negotiations and/or a resubmitted proposal.

For the purposes of this proposal evaluation process, a non-selectable proposal is defined as follows:

NON-SELECTABLE: A proposal is considered non-selectable when the proposal has been evaluated by the Government against the evaluation criteria listed in the PS, and the positive aspects of the overall proposal do not outweigh its negative aspects. Additionally, there are accumulated weaknesses that would require extensive negotiations and/or a resubmitted proposal.

CONFORMING PROPOSALS: Conforming proposals contain all requirements detailed in this PS. Proposals that fail to include required information may be deemed non-conforming and may be removed from consideration. Non-conforming submissions may be rejected without further review. A proposal will be deemed non-conforming if the proposal fails to meet one or more of the following requirements:

- The proposed concept is applicable to the goals and objectives described in this PS.
- The proposer meets the eligibility requirements of this PS.
- The proposal met the submission requirements of this PS.
- The proposal met the content and formatting requirements in the attached templates to this PS.
- The proposal provided sufficient information to assess the validity/feasibility of its claims.
- The proposer has not already received funding or a positive funding decision for the proposed concept (whether from ARPA-H or another Government agency).

Non-conforming proposals may be removed from consideration. Proposers will be notified of non-conforming determinations via email correspondence.

5.3. HANDLING OF COMPETITION SENSITIVE INFORMATION

It is the policy of ARPA-H to protect all proposals as competition sensitive information and to disclose their contents only for the purpose of evaluation and only to screened personnel for authorized reasons, to the extent permitted under applicable laws. Restrictive notices notwithstanding, during the evaluation process, submissions may be handled by ARPA-H support contractors for administrative purposes and/or to assist with technical evaluation.

All ARPA-H support contractors are expressly prohibited from performing ARPA-H sponsored technical research and are bound by appropriate nondisclosure agreements. Input on technical aspects of the proposals may be solicited by ARPA-H from non-Government consultants/experts who are strictly bound by appropriate non-disclosure requirements. No submissions will be returned.

6. AWARDS

6.1. GENERAL GUIDELINES

The Agreement Officer reserves the right to negotiate directly with the proposer on the terms and conditions prior to award of the resulting OT agreement, including payment terms, and will execute the agreement on behalf of the Government. Proposers are advised that only a Government Agreement Officer has the authority to enter into, or modify, a binding agreement on behalf of the United States Government.

In order to receive an award:

- Proposers must be registered in the System for Award Management (SAM) at the time of proposal submission. See Section 6.3.1
- Proposers must be determined to be responsible by the Agreement Officer and must not be suspended or debarred from award by the Federal Government nor be prohibited by Presidential Executive Order and/or law from receiving an award.

6.2. NOTICES

6.2.1. Proposals

The following notices will be provided as applicable:

- Request for clarifying details (if applicable)
  May occur at any time during the evaluation process after proposal submission. Will not include requests for proposal changes and changes will not be permitted.
- Request for additional information (if needed)
  Proposers will be advised of any deficiencies and/or major weaknesses in their proposals and given an opportunity to respond, to include offering proposal amendments.
- Notice of non-selection
- Notice of selection

Once the evaluation of proposals is complete, the proposers will be notified that (1) the proposal has been selected for funding, subject to OT agreement negotiations. This notification may indicate that only a part of the effort has been selected for negotiation and may request a revised proposal for only those selected portions, if not apparent through the delineation of proposed tasks; or (2) the proposal has not been selected for funding.

The above-listed notifications will be sent via electronic mail to the Technical and Administrative points of contact identified on the proposal coversheet.

6.3. ADMINISTRATIVE AND NATIONAL POLICY REQUIREMENTS

6.3.1. System for Award Management (SAM) Registration and Universal Identifier Requirements

All proposers must be registered in SAM and have a valid Unique Entity ID (UEI) number at the time of proposal submission. Performers must maintain an active registration in SAM.gov with current information at all times during which they have an active Federal award or idea under consideration by ARPA-H. SAM.gov registrations must be for All Awards. Information on SAM.gov registration is available at SAM.gov.

NOTE: New registrations can take an average of 7-10 business days to process in SAM.gov. Registration requires the following information:

- SAM UEI number
- Taxpayer Identification Number (TIN)
- Commercial and Government Entity Code (CAGE) Code. If a proposer does not already have a CAGE code, one will be assigned during SAM registration.
- Electronic Funds Transfer information (e.g., proposer's bank account number, routing number, and bank phone or fax number).

6.3.2. Controlled Unclassified Information (CUI) or Controlled Technical Information (CTI) on Non-DoD Information Systems

Further information on Controlled Unclassified Information identification, marking, protecting and control is incorporated herein and can be found at 32 CFR 2002.

6.3.3. Intellectual Property (IP)

Proposers must provide a good faith representation that the proposer either owns or possesses the appropriate licensing rights to all IP that will be utilized for the proposed effort.

ARPA-H intends to obtain unlimited rights to data and software products first produced in the performance of this program. In this context, data means recorded information, regardless of form or the media on which it may be recorded. The term includes technical data and computer software. Proposers should appropriately identify any desired restrictions on the Government's use of any Intellectual Property contemplated under the award. This includes both noncommercial and commercial items. Respondents should utilize the prescribed format within the Administrative & National Policy Requirements Document Template (Volume 3 of Attachment 2 to this PS) when asserting restrictions. If no restrictions are intended, then the proposal should state "NONE."

6.3.4. Human Subjects Research

All entities submitting a proposal for funding that will involve engagement in human subjects research (as defined in 45 CFR § 46) must provide documentation of one or more current Assurance of Compliance with federal regulations for human subjects protection, including at least a Department of Health and Human Services (HHS), Office of Human Research Protection Federal Wide Assurance. All human subjects research must be reviewed and approved by an Institutional Review Board (IRB), as applicable under 45 CFR § 46 and/or 21 CFR § 56. The human subjects research protocol must include a detailed description of the research plan, study population, risks and benefits of study participation, recruitment and consent process, data collection, and data analysis. Recipients of ARPA-H funding must comply with all applicable laws, regulations, and policies for the ARPA-H funded work. This includes, but is not limited to, laws, regulations, and policies regarding the conduct of human subjects research, such as the U.S. federal regulations protecting human subjects in research (e.g., 45 CFR § 46, 21 CFR § 50, § 56, § 312, § 812) and any other equivalent requirements of the applicable jurisdiction.

The informed consent document utilized in human subjects research funded by ARPA-H must comply with all applicable laws, regulations, and policies, including but not limited to U.S. federal regulations protecting human subjects in research (45 CFR § 46, and, as applicable, 21 CFR § 50). The protocol package submitted to the IRB must contain evidence of completion of appropriate human subjects research training by all investigators and key personnel who will be directly involved in the design or conduct of the ARPA-H funded human subjects research.

Funding cannot be used toward human subjects research until ALL approvals are granted.

6.3.5. Animal Subjects Research

Award recipients performing research, experimentation, or testing involving the use of animals shall comply with the laws, regulations, and policies on animal acquisition, transport, care, handling, and use as outlined in: (i) 9 CFR parts 1-4, U.S. Department of Agriculture rules that implement the Animal Welfare Act of 1966, as amended, (7 U.S.C. § 2131-2159); (ii) the Public Health Service Policy on Humane Care

and Use of Laboratory Animals,[25] which incorporates the "U.S. Government Principles for the Utilization and Care of Vertebrate Animals Used in Testing, Research, and Training,"[26] and "Guide for the Care and Use of Laboratory Animals" (8th Edition).[27]

Proposers must complete and submit the Vertebrate Animal Section worksheet for all proposed research anticipating Animal Subject Research. All Animal Use Research must undergo review and approval by the local Institutional Animal Care Use Committee (IACUC) prior to incurring any costs related to the animal use research.

6.4. ELECTRONIC INVOICING AND PAYMENTS

Performers will be required to register and submit invoices for directly to the Payment Management System (PMS) unless an exception applies. PMS guidance can be found here: https://pms.psc.gov/training/grant-recipient-training.html.

7. COMMUNICATIONS

ARPA-H intends to use electronic mail for all correspondence regarding this PS. Administrative questions regarding this PS should be emailed to the PRECISE-AI PS Coordinator. ARPA-H will post a Q&A document to SAM.gov regarding all administrative questions submitted to this PS on an as needed basis. All questions must be in English and must include the name, email address, and telephone number of a point of contact.

ARPA-H will attempt to answer questions in a timely manner. In order to receive a response sufficiently in advance of the proposal due date, questions should be submitted on or before the Q&A deadline stated herein.

[25] olaw.nih.gov/sites/default/files/PHSPolicyLabAnimals.pdf
[26] olaw.nih.gov/policies-laws/gov-principles.htm
[27] olaw.nih.gov/sites/default/files/Guide-for-the-Care-and-Use-of-Laboratory-Animals.pdf