Department *of* Commerce
Office *of* Artificial Intelligence Policy

PREPARED FOR PUBLICATION

Evidentiary Expectations for Healthcare AI Sandbox

Demonstrating patient-focused, safety-driven AI applications

Alice Schwarze and Zach Boyd

August 2026

Summary

This document established OAIP's general expectations for sandbox applications that propose AI-in-healthcare products and/or services. This document prioritizes OAIP's expectations regarding the impacts of AI-in-healthcare products and/or services on the quality of clinical functions, including their implications for patient safety. At the same time, sandbox applicants are expected to articulate clear and measurable approaches for assessing whether these products and/or services concurrently improve accessibility and reduce the cost of clinical functions relative to existing products and/or services.

OAIP acknowledges that feasibility constraints or substantial trade-offs between quality of care and public-health benefits may warrant adjustments of these expectations for specific sandbox applications. Some regulatory mitigation agreements between OAIP and healthcare AI companies may therefore deviate from these expectations. Prospective sandbox applicants should nonetheless consider the expectations outlined in this document as OAIP's default negotiation posture for healthcare AI in the OAIP regulatory sandbox.

1. General sandbox posture

The OAIP regulatory sandbox aims to support the development of a clear and evidence-based regulatory pathway for AI-in-healthcare products and/or services. The sandbox has two main aims:

> **Aim 1:** Identify AI use cases that are safe for consumers and can produce a public benefit.
>
> **Aim 2:** Identify regulatory frameworks that can effectively establish and incentivize these AI use cases.

To contribute to Aim 1, applications to the OAIP regulatory sandbox should:

a. articulate a credible argument that the proposed AI-based product or service is safe for consumers and
b. show potential to improve quality, access, or some other metric of consumer or public benefit.

To contribute to Aim 2, sandbox applicants are expected to agree to sharing pilot data with OAIP such that OAIP can identify informative data streams and produce formal hypotheses about if active regulation of a given AI use case is warranted and how regulators can use these data streams to reliably and transparently confirm safety, public benefit, and specific sectorial expectations where applicable.

2. AI-in-Healthcare posture

2.1 Burden of proof

In the context of healthcare AI, OAIP considers two components of consumer/ patient safety:

i. Absence of a loss of chance: The quality of care offered to patients via the sandbox pilot is not inferior to the care that these patients would otherwise receive from existing healthcare providers and systems.
ii. Absence of substantial bodily and psychological injury linkable to the use of an AI tool.

Sandbox applications for healthcare AI projects must have the potential to build evidence for **three main hypothesis:**

- Non-inferiority of the provided quality of care,
- Bodily and psychological safety, and
- potential for public-health benefits in Utah.

2.2 Commitment to low market entry barriers and high expectations for continuous quality assurance

In alignment with OAIP's dual mission to foster innovation and protect consumers, the Office is open to sandbox proposals that combine moderate pre-deployment evidence for a product's safety and quality (e.g., compared to historical SaMD standards set by the FDA) with a gradual product roll-out that prioritizes patient safety and a rigorous post-deployment evidence collection and continuous surveillance program.

2.2.1 Why low market entry barriers?

- Healthcare AI applications have a potential to close care gaps in the State. Facilitating such healthcare AI applications to operate in Utah with reasonable safeguards in a timely manner can lead to substantive healthcare improvements to Utah residents.
- For healthcare AI, pre-deployment evidence can become stale and lose its value quickly due to fast population drift, behavioral drift (among clinicians and patients), model drift, and technology updates.
- When a healthcare AI application is affecting clinical interventions that touch a broad population with possibly a broad range of conditions, it is hard if not impossible to test all its clinical functions meaningfully in a controlled setting (e.g. in a randomized controlled trial (RCT)). Given the need to guarantee safety for all patients, not just for the typical (i.e., modal) patient, strong evidence for the safety of broad-scope healthcare AI can only be created via thorough continuous quality assurance surveillance.

Pre-deployment evidence should aim to alleviate the following concerns:

- The AI care path may cause **substantially worse outcomes for patients**, including typical cases and rare cases, than other available care paths.
- The AI care path may cause **worse outcomes** than other available care paths in a statistically significant or otherwise clinically meaningful way **on a population level** without necessarily leading to worse outcomes detectable on an individual level (e.g., screening recommendation or omission of a screening recommendation that leads to worse public health outcomes; oversubscription of antibiotics leading to population-level antibiotic resistance, or over testing leading to increased care costs).
- The AI care path may remove certain patients at certain care-path points in a way that makes **some negative outcomes for patients invisible for detection** (e.g., recommending that no treatment is necessary and ending the care path for a patient who later needs to receive emergency care from a different provider or different healthcare organization).

- The AI care path opens up opportunities for **adversarial exploitation** that can harm patients or other affected individuals (e.g., patients gaming AI interactions to receive access to certain medications despite lack of a clinical indication).

2.2.2 Why high expectations for continuous quality assurance?

- Population drift, behavioral drift of patient and clinician populations (including changes to clinical best practices), model drift, and technological updates can substantially affect safety and quality of care associated with a healthcare AI application. Such drifts can be detected via continuous quality assurance rather than in pre-deployment testing.
- Continuous quality assurance can include instruments for detecting and mitigating adverse outcomes for non-typical, possibly unanticipated, patients who would be under sampled or missed in pre-deployment testing.
- A transparent framework for quality control of healthcare AI applications is necessary to grow consumer trust and confidence in healthcare AI. For the healthcare-focused sandbox pilots, OAIP aims to showcase such a framework via rigorous and continuous quality assurance that aims to incentivize high-quality healthcare AI applications in the State.
- Several scholars have proposed regulatory frameworks for broad-scope healthcare AI with inspiration drawn from how clinical practice is regulated (Bressman et al. 2025; Bergman et al., 2026). Physicians, for example, enter the workforce as residents before having demonstrated real-world clinical competency. Operating first under close supervision, they gradually gain more autonomy throughout their residency. Even after completing their residency, most physicians are subject to continuous quality control via liability, licensure review, M&M, and peer supervision. A low barrier of pre-deployment evidence coupled with a supervised gradual roll-out of clinical functions and autonomy during the sandbox pilot mirrors this training path when assessing clinical competency of healthcare AI applications. The rigorous continued quality assurance fulfills the role of continuous quality control for clinicians.

In addition to providing continuous evidence to alleviate the concerns listed in Section 2.2.1, continuous quality assurance should provide evidence for:

- sustained safety and quality of care for typical and rare cases,
- sustained adherence to clinical scope, including appropriate abstention decisions (e.g., based on clinician-aligned confidence scores) for broad-scope healthcare AI,
- sustained integrity of quality controls and backstops (e.g., are reviewing clinicians continuing to review cases carefully or submitting to automation bias),
- sustained public-health benefits.

OAIP continues to expect sandbox applicants to satisfy applicable pre-deployment requirements; however, greater weight is placed on generation and evaluation of post-deployment evidence.

2.3 Patient rights in healthcare-AI sandbox pilots

Patient participation in sandbox pilots should always be informed and voluntary. A patient who is about to interact with an AI tool that operates under an OAIP regulatory relief agreement should receive an appropriate disclosure about

- the involvement of AI in the product or service that they engage with,
- the sandbox status (i.e., only preliminary, temporary, and revokable state authorization) of the product or service that they engage with,
- and the opportunity to submit feedback and raise concerns directly with OAIP.

Unless strictly necessary for the proposed use case, sandbox AI tools should not be deployed in a setting where patients have effectively no choice to opt out. For example, a sandbox proposal for an AI system that triages patients in critical conditions with no easy and immediate option for patients or their guardians to opt-out would need to demonstrate substantial pre-deployment evidence of non-inferior quality care, bodily and psychological safety, public-health benefits, and buy-in from stakeholders to make a compelling case to OAIP.

3. Mapping healthcare AI to healthcare innovations

3.1. Innovation along care paths

A patient's care path starts when the patient seeks care from a healthcare provider for a specific concern or when care is initiated on the patient's behalf, includes the provider's steps to arriving at a diagnosis and/or treatment plan, and ends with the last patient-provider interactions related to that concern (e.g., a follow-up visit to confirm successful treatment or a post-discharge follow-up).

A typical out-patient care path can include:

- care path initiation (e.g., a patient calls their PCP to book an appointment),
- patient intake (e.g., a patient fills out an intake form and/or is interviewed by a medical assistant),
- collection of medically relevant data (e.g., a medical assistant measures temperature, blood pressure and pulse; a physician retrieves the patient's medical history from their electronic health record, confirms intake information with the patient and asks clarifying questions) ,

- clinical decision (e.g., the physician suggests a diagnosis and treatment plan)
- Treatment plan delivery (e.g., the physician prescribes a medication; the patient visits a pharmacy to get the prescription filled),
- follow-ups (e.g., staff at the physician's office call the patient two weeks later to ask if the symptoms have resolved or if the patient would like to book another appointment).

A typical in-patient care path can include:

- care path initiation (e.g., a patient arrives at the emergency department via EMS),
- patient intake (e.g., a registration clerk collects patient information and a triage nurse assesses acuity),
- collection of medically relevant data (e.g., nursing staff record vitals on a continuous or scheduled basis; the care team orders labs and imaging),
- clinical decision (e.g., the admitting physician establishes a working diagnosis and initial treatment plan; the care team revises the plan during daily rounds as results and consults come in),
- treatment plan delivery (e.g., nursing staff administer medications on the ordered schedule; a procedure or surgery is performed),
- discharge and follow-ups (e.g., the care team determines discharge readiness, a discharge summary is shared with the patient's PCP, and a follow-up visit or post-discharge call is scheduled).

If a healthcare AI product leads to changes in any of these components, it may also lead to changes in patient distributions (e.g., who decides to enter a care path, who completes it, etc.), patient behavior (e.g., information shared during intake), and clinician behavior (e.g., automation bias). These changes can affect a care path's overall quality and safety.

Sandbox proposals should identify which components and, if applicable, which tasks within a component of their patients' care paths are affected by their proposed product and suggest pre-deployment evidence and continuous quality assurance frameworks that establish how these changes to the care path affect the safety and quality of care and public health.

3.2 Clinical functions versus clinical competency

3.2.1 Clinical functions

The clinical scope of a proposed healthcare AI product or service may include one or several clinical functions. Examples of functions include:

- diagnosis of need for dental scaling and root planning from x-ray imaging (see agreement with Dentacor),
- renewal of prescriptions for a class of allergy medications (see agreement with Doctronic),
- diagnosis of stress incontinence (see agreement with Expect Fitness),
- treatment of stress incontinence via physical-therapy exercises (see agreement with Expect Fitness),
- recommendation of adjustments to base insulin levels within a clinician-prescribed range (see FDA approved SaMD Updoc).

In general, OAIP expects sandbox applicants to produce pre- and post-deployment evidence for the safety and quality of care of each clinical function that falls within the clinical scope of the proposed AI product or service. Accordingly, OAIP's data sharing expectations include measures of clinical non-inferiority for each clinical function.

Acknowledging data collection issues arising from the low prevalence of some health conditions, OAIP is open to considering proposals that group several clinical functions during evidence collection, if the clinical functions follow identical or similar decision protocols and have identical or similar risk profiles.

3.2.2 Clinical competency

OAIP acknowledges that AI technologies may facilitate the development of flexible clinical decision support systems (CDSSs) or even flexible autonomous AI diagnostic systems, which can deliver quality care across a large clinical scope including various clinical functions and diverse patient populations. Sandbox proposals with

- a predominantly unstructured patient intake (e.g., unstructured conversation rather than multiple-choice forms; multimodal inputs involving images, audio, or video) and
- a clinical scope including 80 or more clinical functions that cannot be grouped as described in Section 3.2.1

may opt to produce evidence for the safety and quality of care associated with the "clinical competency" of their AI product or service.

Pre- and post-deployment evidence should include

- evidence of safety and quality of care aggregated over all clinical functions,

- evidence of the product's ability to demonstrate traits commonly associated with clinical competency (Epstein and Hundert, 2002) when necessary or appropriate. At minimum, the following abilities should be demonstrated
    - allowing and appropriately **managing input uncertainty** (e.g., patients' vague or contradictory description of symptoms, blurry images, or conflicting information in an EHR)
    - **sensitivity to clinical context** (e.g., suggesting treatment plans that are appropriate and feasible given a patient's economic or other circumstances)
    - **emotional intelligence** (e.g., ability to detect patients' emotional state, anxieties, preferences, and attempts to manipulate, lie, or avoid certain topics)
    - **diverse reasoning skills**, including
        - pattern recognition
        - probabilistic/Bayesian reasoning allowing the healthcare AI product to consider data of various of levels of certainty (e.g., weighing positive and negative test results differently based on the test's sensitivity, specificity and positive and negative predictive value)
        - hypothetico-deductive reasoning allowing the healthcare AI product to formulate hypotheses and collect additional data (e.g., from lab tests) to confirm or reject hypotheses
    - **awareness of decision confidence and bias** (e.g., the healthcare AI product should escalate cases based on its lack of confidence in its own judgment when it encounters atypical patient behavior or possibly rare health conditions rather than just the cases that fall clearly outside of its clinical scope).

Sandbox applicants proposing to provide proof of clinical competency of their proposed product or service should also explain in their proposal how their product is replicating

- a clinician's **ability to learn from experience** (e.g., via continuous self-improvement loops or frequent retraining cycles on recent patient encounters)
- a clinician's **accountability framework**, which typically include liability, licensure review, M&M, and peer supervision (e.g., via sandbox applicants accepting liability comparable to medical malpractice liability, reliable detection and investigation of harmful incidents, and continuous clinician review of a sample of patient-AI interactions).

The option to produce evidence for clinical competency does not exempt high-risk clinical functions from function-level evidence expectations. Sandbox proposals should identify the clinical functions within their scope whose failure modes carry the highest risk of patient harm (e.g., functions where a

missed diagnosis can plausibly lead to death or serious deterioration) and commit to dedicated function-level monitoring and reporting for these functions in accordance with Section 3.2.1, so that their performance remains detectable rather than diluted in measures aggregated over the full clinical scope.

4. Evidence for main hypotheses

4.1. Evidence for non-inferiority of quality of care

4.1.1 Outcome-based and proximal evidence for quality of care

The National Academy of Medicine defines quality of care as the degree to which health care services for individuals and populations increase the likelihood of desired health outcomes and are consistent with current professional knowledge (Institute of Medicine, 1990). Healthcare AI applications have the potential to change care paths and affect patient and clinician behavior and distributions of patient demographics (see Section 3.1). These changes and their consequences can be subtle and happen at unanticipated points in the care path. A robust approach to creating evidence of the quality of care associated with a healthcare AI application is to measure health outcomes in a randomized controlled trial (RCT).

Since RCTs can be slow and costly and healthcare AI applications may benefit from frequently recurring or even continuous quality control (see Section 2.2), OAIP is open to receiving proximal evidence of the quality of care associated with a healthcare AI product.
Relevant pre-deployment evidence includes:

- in-silico tests and safeguards for anticipated failure modes, anticipated population drift and anticipated behavioral drift,
- outcome- or process-based evidence from back testing on real patient data, and
- process-based evidence demonstrating clinical equivalency or clinical alignment at decision points with a well-accepted standard of care.

Relevant post-deployment evidence includes:

- measurements of post-deployment failure modes, population drift, and behavioral drift (accompanied by a mitigation plan if unanticipated failure modes or drift phenomena are detected),
- demonstration of in-deployment efficacy of safeguards,
- demonstration of clinical equivalency or clinical alignment at decision points with a well-accepted standard of care, and

- outcomes from harmful-incident detection and patient surveys.

4.1.2 Demonstration of clinical equivalency

When an AI tool recommends or autonomously executes clinical actions, OAIP accepts proof of clinical equivalency between the AI tool and care option representative of the standard of care as proof of non-inferiority of the quality of care offered by the AI tool. An AI tool's recommendations or actions are clinically equivalent to care by a licensed provider if across plausibly occurring patient encounters, recommendations issued and actions taken by the AI tool are identical to the recommendations issued and actions taken by a licensed provider who does not interact with the AI tool. In all cases, licensed providers whose recommendations, actions, or assessments serve as the basis for establishing clinical equivalency must hold licensure whose scope of practice encompasses the clinical decisions under review. When specialty expertise is relevant, reviewing clinicians should also be specialty-matched.

The unit at which clinical equivalency is measured should match the clinical scope of the proposed AI tool. For AI tools whose clinical functions are contained within a single patient encounter, the individual patient encounter is the unit of measurement. For AI tools that accompany a patient across multiple encounters of a longitudinal care path (see Section 3.1), sandbox applicants should specify unit of measurement, which may be intervals of a longitudinal care path (e.g., at weekly or monthly check-ins) or a complete multi-day or multi-week care path.

AI-led patient encounters

For AI tools that lead entire patient encounters (i.e., administer patient intake and issue or recommend clinical decisions), OAIP considers two main components to establishing clinical equivalency:

1. Evidence that the AI-administered patient intake creates the same or a sufficiently similar information background for clinical decision making as provider-administered intake.
2. Given the same information background, an AI's diagnosis and treatment plan match the diagnosis and treatment plan that was established independently (i.e., without knowledge of the AI's decision) by a licensed provider or a quorum of licensed providers acting within the scope of their licensure via a diagnostic approach that matches an appropriate well-accepted standard of care.

Agreement rates alone are not sufficient to establish clinical equivalence. Measurements of sensitivity, specificity, positive and negative predictive values, and qualitative insights identifying common failure modes of AI decisions (using the independent clinician's decisions as ground truth for each of

the conditions in the pilot's clinical scope) are necessary to understand how the AI's clinical decisions compare to clinician decisions.

Providing evidence only for clinical equivalency of decision making (point 2) is not sufficient to prove overall clinical equivalency because a clinician's decisions are likely affected by the quality of information provided on which to base decisions.

AI tools for individual care path components

Many AI tools affect non-patient-facing components of a care path or individual components of a patient encounter rather than leading entire patient encounters (see Section 3.1). For such tools, clinical equivalency should be demonstrated for the affected component against its standard-of-care counterpart. For example, evidence and decisions produced by a triage tool should be clinically equivalent to those of triage clinicians.

Applicants should identify the care path components that are directly affected by their proposed healthcare-AI application. To establish clinical equivalency, the sandbox applicant must show that each affected component receives an information background equivalent to its standard-of-care counterpart and that its output is equivalent given that information.

4.1.3 Demonstration of clinical alignment

It is possible that AI decisions are non-inferior to clinician decisions while failing to be clinically equivalent for some patient encounters. Sandbox participants may demonstrate non-inferiority of quality of care when criteria for robust clinical equivalency (e.g., agreement rate, sensitivity, and specificity >= 90%[1]) are not met via clinician review of AI-clinician disagreements: In all cases, where AI decisions do not match clinician decisions, a reviewing clinician (or a team of reviewing clinicians) reviews the conflict, states whether they side with the AI decision or the independently diagnosing clinician's decision, and rates the severity/ risk-level of the conflict (i.e., whether the disagreement highlights an AI failure that could lead to negative patient outcomes or it merely reflects a difference in taste or style).

If review pipelines include sufficient, well-documented safeguards against automation bias, OAIP considers all cases "clinically aligned" for which reviewing clinicians who either

- agree with the AI decision or
- categorize the decision as lowest-risk level

[1] Target values for sensitivity and specificity may differ based on risks associated with false-positive and false-negative case evaluations.

constitute a majority among the case's clinical reviewers.

Sandbox applicants may use robust clinical alignment (e.g., alignment rate, alignment-adjusted sensitivity, and alignment-adjusted specificity >= 90%[2]) as evidence for the non-inferiority of their provided quality of care.

4.1.4 Acknowledging the possibility of superiority of healthcare AI

It is possible that current or future AI algorithms do not reproduce clinician decisions (i.e., are not clinically equivalent) or even align with clinician judgements, but nonetheless lead to non-inferior or even improved patient outcomes. As long as the existence of AI algorithms leading to improved patient outcomes in real clinical settings is a subject of substantial debate among clinical and public-health experts, OAIP expects that sandbox pilots either demonstrate robust clinical equivalency, robust clinical alignment, or non-inferiority via a rigorous outcome-based approach, which may require short-term and long-term patient outcomes.

4.2. Evidence for psychological and bodily safety

OAIP's priority is to ensure that sandbox pilots are safe for Utah's residents. To the extent possible, evidence for safety from the most substantial harms should therefore be rigorous, comprehensive, based on patient outcomes, and frequently updated.

Relevant post-deployment measures to collect data on patient outcomes include:

- robust framework for the detection of harmful incidents,
- feedback from reviewing clinicians,
- patient follow-up surveys,
- easy-to-access channels for patients to submit feedback and complaints.

Considering hospitalizations and deaths as indicators of substantial harm, OAIP aims to collaborate with relevant entities on building an infrastructure for monitoring hospitalization and death rates for users of healthcare AI products in the regulatory sandbox. Following the construction of this infrastructure, OAIP expects sandbox participants to participate in this program by providing tokenized patient rosters.

[2] Target values for alignment-adjusted sensitivity and alignment-adjusted specificity may differ based on risks associated with false-positive and false-negative case evaluations.

4.3. Evidence for public health benefits

A net public-health benefit of a sandbox pilot is achieved when positive public-health impacts outweigh negative public-health impacts. Appendix A combines a list of the most commonly proposed public-health benefits in OAIP sandbox proposals with OAIP's recommendation for adequate evidence generation. Appendix B includes a list of public-health concerns commonly raised in the context of telemedicine and direct-to-consumer offers with OAIP's recommendations for adequate safeguards and evidence generation.

Sandbox applicants who propose to close one or several care gaps in Utah should include appropriate data on patient demographics in their data-sharing commitment, such as

- number of patients who use the tool to completion,
- number of returning patients more than once and breakdown by return frequency,
- patient satisfaction ratings,
- distribution of first-visit user ages,
- distribution of returning-user ages,
- distribution of time and day patients use the tool,
- deidentified and aggregated patient location data,
- quality-of-care measurements by appropriate demographics, and
- qualitative insights from patient feedback from surveys (e.g., reason for using AI tool) or feedback or complaint forms.

Sandbox applicants who propose improved work conditions for Utah clinicians should include appropriate workforce measures in their data-sharing commitment, such as

- clinician time spent on documentation and administrative tasks per patient encounter,
- clinician caseload and throughput compared to pre-deployment baselines,
- clinician adoption and sustained-use rates for the tool,
- clinician opt-outs and reversions to previous workflows,
- results from validated burnout or job-satisfaction instruments administered before and during the pilot,
- retention and turnover data for affected clinician roles,
- measures of patient access attributable to workforce effects (e.g., appointment availability, time to third next available appointment, patient wait times), and
- qualitative insights from clinician feedback from surveys or feedback or complaint forms.

5. Evidence collection throughout the sandbox pilot

5.1 Expert assessment of the proposal

OAIP considers positive expert reviews of sandbox proposals as proximal indicators for an application's non-inferior care, safety, and positive public health impact.

As part of the application process, sandbox applicants submit a detailed technical proposal describing

- the motivation and public health benefit for the pilot,
- its clinical scope,
- populations impacted,
- level of AI autonomy for all functions in the clinical scope,
- adequate conservative escalation policies,
- adequate levels of oversight at each stage of the pilot,
- robust quality assurance and risk management frameworks,
- an AI governance framework that defines major model updates and the necessary pre- and post-update tests or reevaluations to prioritize patient safety during updates,
- a post-deployment monitoring plan specifying tracked performance measures, measurement and reporting cadence, detection pathways for failure modes and drift, and predefined responses to degradation,
- an accountability framework that (1) allocates accountability along the healthcare delivery chain proportionally to where harm can be prevented most effectively and (2) gives patients recourse comparable to their recourse in healthcare delivery settings not involving AI,
- the organizational structures supporting feasibility, scalability, and robustness of the proposed project, and
- a patient-first business approach that includes a robust duty of care and financial stewardship for patients.

OAIP reviews the information provided in consultation with clinical and public health experts to establish the safety of the proposed project.

5.2 Pre-deployment evidence collection

To the extent feasible, the AI systems supporting or making clinical decisions as part of a sandbox pilot should be designed with care and extensively tested on appropriate data sets prior to deployment.

5.2.1 Aim of pre-deployment evidence collection

Sandbox applicants should attempt to provide compelling evidence for their product's quality of care and safety prior to deployment. The evidence provided should alleviate concerns about (see also Section 4.1.1):

1. the product's capability to contribute to or provide a non-inferior quality of care compared to an appropriate reference standard of care in an appropriate reference care setting, and
2. the product's capability to avoid failure modes and handle changes in population behavior or demographic distribution, which may arise from differences between the reference care setting and the care setting in deployment.

5.2.2 Approaches for pre-deployment evidence collection

Concern 1 from Section 5.2.1 can be addressed, for example, by

- **performance metrics from prior deployments** of the healthcare AI product in comparable settings (e.g., prior deployments in other states or countries). OAIP judges the credibility of this type of evidence by the similarity of the demonstrated care setting in the prior deployment to the proposed care setting in Utah.
- **tests on historical patient cases** drawn from an applicant's previous deployments of similar healthcare products or services (e.g., an applicant's previous or continued operations) or from a partnering healthcare organization or assurance provider.
- **company-internal A/B tests or RCTs** that compare a care path sufficiently similar to the pilot's proposed care path with the applicant's existing operations. Such tests are particularly insightful because they can detect novel failure modes and unanticipated drift phenomena.

Concern 2 from Section 5.2.1 can be addressed, for example, by

1. **mapping out anticipated failure modes and population changes** compared to the reference scenario that was used to address Concern 1, and
2. **running simulation-based tests** (e.g., simulated patient encounters) that are representative of the anticipated failure modes and changes in population behavior and demographics.

Sandbox applicants may design and run these tests internally or in collaboration with an assurance provider or other research or healthcare organization.

OAIP is open to other approaches that can serve the stated aim of pre-deployment evidence collection (e.g., use of a credible world model (Luo et al., 2026) for simulated evidence for Concern 1 and Concern 2).

5.2.3 External data sources, external validation, and best practices

OAIP can consider and may accept proposals with all pre-deployment evidence generated by the sandbox applicant without external validation. However, data sources for evidence generation should include internal ***and*** external data sources to (1) motivate public trust and stakeholder support for sandbox pilots and (2) demonstrate that care non-inferiority can be shown robustly for different patient populations sources in different ways. OAIP can help identify relevant benchmark developers and assurance providers for collaborations with sandbox applicants.

Applicants should also include concrete plans for external validation (i.e., tests independently conducted or verified by an independent third party) of their internal findings (e.g., via academic collaborations, peer-review, or external audits) in their proposal. OAIP expects some level of externally generated or externally validated evidence of the healthcare AI's non-inferior quality of care by the time the applicant seeks to start the negotiations for their year-2 regulatory mitigation contract.

For internal tests, sandbox applicants should document that their tests followed best practices, including:

- The demographics distributions and clinical backgrounds included in patient data used to develop and test algorithms should match the anticipated demographics and clinical backgrounds targeted during deployment as closely as possible.
- Development pipelines should avoid data leakage. Data used for training, fine-tuning, or other forms of development should be clearly separated and statistically independent from test data sets. Sandbox applicants should be prepared to provide appropriate evidence of robust development pipelines to OAIP.
- Pre-deployment tests should include "red teaming", i.e., specific tests to establish that AI tools respond appropriately to avoid harm to patients or other affected individuals when users try to manipulate the AI.

5.3 Post-deployment evidence collection

5.3.1 Aim of post-deployment evidence collection

With the start of the product's deployment, sandbox participants can observe real interactions between their pilot's product and healthcare providers and/or patients along the intended care path. OAIP expects sandbox participants to use these observations to provide evidence for the three main hypotheses (see Section 2.1): non-inferiority of the provided quality of care, psychological and bodily safety, and public health benefit.

5.3.2 Approaches for post-deployment evidence collection

Evidence for the non-inferiority of the quality of care provided should include:

- measurements of non-inferior quality of care based on patient outcomes (if available),
- measurements of clinical competency (see Section 3.2.2) or measurements of clinical equivalency (see Section 4.1.1) or clinical alignment (see Section 4.1.2) at clinical decision points,
- detection/ measurements of automation bias among reviewing clinicians,
- detection of post-deployment failure modes, population drift, and behavioral drift, and
- results from rerunning pre-deployment tests (including updated versions of benchmark tests conducted in the pre-deployment phase and back testing on historical data with updated data sets, and simulation tests using state-of-the-art simulation techniques) in appropriate intervals.

If the detection of failure modes, population drift, or behavioral drift confirms issues anticipated pre-deployment, evidence for non-inferiority of quality of care should include evidence of the efficacy and robustness of safeguards aimed at mitigating these issues. If the detection surfaces unanticipated issues, participants should present plans mitigating these new issues.

Regular reruns of pre-deployment tests are especially important for pilot products with either small volumes of patient interactions or a broad clinical scope. For such pilot products, the data set of patient interaction observed since deployment is likely to miss plausible yet rare patient-interaction scenarios that can be surfaced via simulation studies and back testing on large external data sets.

Evidence for psychological and bodily safety should be based on real patient outcomes:

- outcomes from harmful-incident detection,
- results from patient surveys,
- patient outcome data source via health information exchanges or direct collaboration or communication with patients' other providers (if available).

Evidence for public health benefits should include the metrics listed in Section 4.3 and, if appropriate and feasible, additional metrics related to specific public health benefits proposed in the participant's sandbox proposal. For example, a company providing melanoma screening services may have claimed in their sandbox proposal that the pilot product can increase the number of early-stage melanoma detections in favor of late-stage melanoma detections. This company should propose how to collect relevant data on patient outcomes (e.g., via patient surveys and collaborations with their

patients' traditional care providers) and how to use this data to infer if the proposed public health benefit is materializing.

5.3.3 External validation and best practices

OAIP favors sandbox proposals that commit to continuous third-party validation of post-deployment quality assurance data by an OAIP-recognized auditor or assurance provider. Sandbox applicants may suggest other avenues for demonstrating reliability and growing trust in the data that they share with OAIP or the public.

All sandbox applicants should implement a sufficiently safe and reliable process logging and documentation framework to allow for rigorous internal and external audits if severely harmful incidents merit such audits. Such process logging should include adequate data tracking and traceability policies and retention of chat logs and reasoning traces to facilitate appropriate harm attribution.

6. Risk management throughout the sandbox pilot

OAIP expects pilots to be designed to prioritize patient safety. The expected design includes

- a safety-prioritizing multi-stage process of rolling out autonomous AI functionalities and/or rolling back clinician oversight,
- a safety-prioritizing protocol for pilot termination for the event that termination becomes necessary,
- a robust framework for harmful-incident detection, reporting, investigation, and mitigation,
- continuous monitoring of product use and performance metrics to detect population, behavior, and performance drift, and
- a robust change-management framework to detect possible links between performance drift and model changes.

6.1 Safety-prioritizing pilot roll-out

Pilots should be structured into a minimum of three separate stages. Sandbox applicants should specify in their proposals minimum case volumes and minimum performance targets that the AI tool needs to meet in each stage to trigger the pilot's advancement to the next stage.

Section 6.1.1 presents OAIP's suggested minimum pilot staging. Section 6.1.2 presents OAIP's suggested gating of pilot stage progressions. Sandbox applicants may propose other pilot designs that appropriately prioritize patient safety in their proposed deployment setting.

6.1.1 Suggested minimum pilot staging

Stage 1: Detailed clinical oversight

For CDSSs, depending on the complexity and risk associated with the clinical scope, Stage 1 may include:

- independent-decision documentation (i.e., the clinician records their working decision before viewing the AI recommendation) to establish clinical equivalency of AI-assisted and unassisted decisions and to measure the AI tool's influence on decision making, including automation bias, and
- review of AI-influenced decisions (i.e., where the clinician's final decision deviates from their documented working decision, a reviewing clinician reviews the case before or promptly after clinical action, depending on risk) with detailed documentation of how the AI recommendation impacted the decision.

For autonomous AI systems, Stage 1 includes

- silent trial with independent decision reference (i.e., a clinician and the AI make decisions in parallel but independently) to establish clinical equivalency and
- prospective review of disagreements (i.e., where AI and clinician disagree, AI decisions are reviewed by a reviewing clinician or reviewing clinician team before any clinical action is taken) to prevent harm to patients and establish clinical alignment.

Stage 2: Retrospective clinical oversight

For CDSSs, Stage 2 may drop the requirement to document a working decision before viewing the AI recommendation and replace review of AI-influenced decisions with timely retrospective review. Recommendation acceptance and override rates should be monitored continuously as indicators of over- or under-reliance.

For autonomous AI systems, Stage 2 includes retrospective review (i.e., AI decisions lead to clinical actions, but receive a timely clinician review that may trigger corrective action) to continue to prevent patient harm and comprehensively monitor clinical alignment.

Stage 3: Sample-based performance monitoring

Stage 3 for CDSSs may reduce retrospective review to a representative sample of AI-assisted decisions, facilitating efficiency gains and workload reduction for clinicians, while continued monitoring of acceptance and override rates supports early detection of drift toward over-reliance.

For autonomous AI systems, Stage 3 includes retrospective review with an independent decision reference of a representative sample of cases to ensure continued quality and safety of the AI tool and detect potential incidences of performance drift early.

6.1.2 Suggested gates for stage progression

Sandbox applicants should specify in their proposals the gates that trigger progression from one pilot stage to the next. OAIP expects sandbox applicants to pre-specify a statistical decision rule for stage progression. Applicants should propose pre-specified decision rules that, to the extent possible, aim to avoid statistical fallacies arising from, for example, data leakage, bias, confounding, or hypothesis multiplicity.

OAIP suggests a pre-specified decision rule of the following form: The pilot advances to the next stage when, at a pre-specified look-up point, (1) every gated performance measure meets or exceeds its critical value and (2) the number of substantial-disagreement events[3] does not exceed the substantial-disagreement limit. In addition, the following criteria are met:

- Progression tests should not happen before a pilot stage has met a pre-specified minimum calendar duration to ensure that reviewed cases reflect a representative case mix.
- Progression tests may only be performed at the case volumes listed in Table X to avoid unpenalized multiple-hypothesis testing.
- Progression tests should be computed on all consecutive eligible cases since the start of the current stage. Cases may not be excluded or resampled, except in accordance with a pre-agreed protocol of upon modification approval from OAIP.
- For each case in which a reviewing clinician or a quorum of reviewing clinicians disagrees with the AI decision, a quorum of clinicians should assess the severity of the disagreement. Substantial disagreements are disagreements that the quorum classifies as outside the range of acceptable practice with potential for patient harm (see also Section 6.3).
- Any OAIP-approved revision of gates during a running stage requires recomputing the remaining critical values such that the overall error rate of the decision rule is maintained.

Performance baseline via clinician-clinician agreement

Judging the performance of an AI tool's clinical decisions requires a reference value against which observed performance can be compared. As stated in Section 4.1.1., OAIP accepts independent

[3] OAIP considers a substantial disagreement a disagreement that a quorum of licensed providers classifies as outside the range of acceptable practice with potential for patient harm.

clinician assessment as a proximal reference standard. Because clinicians exercising acceptable practice do not perfectly agree with one another, perfect agreement with a reference clinician is neither an attainable nor a meaningful target for AI decisions. Clinician-clinician agreement values instead measure the level of agreement that the current standard of care achieves against itself on the pilot's case mix and therefore serve as the appropriate baseline from which the performance floors for the gated measures are derived.

Establishing clinician-clinician agreement values requires at least two independent clinician assessments of each case marked for baseline measurement. Because agreement between two particular clinicians can be inflated or deflated by coincidental alignments or misalignments in their tastes, styles, and risk tolerances, OAIP recommends drawing independent assessments from a larger pool of clinicians so that measured agreement reflects the range of acceptable practice rather than the idiosyncrasies of a single reviewer pair.

Clinician-clinician agreement must further be measured in the same sense of agreement as the gated clinician-AI measures it anchors: if non-inferiority of the quality of care is demonstrated via clinical equivalency, clinician-clinician agreement should be measured as agreement between independent clinician decisions, without further adjudication involving a reviewing quorum or risk-level ratings; if non-inferiority is demonstrated via clinical alignment, clinician-clinician agreement rates should be established via the same adjudication protocol used to establish clinical alignment of AI decisions (see Section 4.1.3).

Gated performance measures

At minimum, pilots should gate three performance measures, computed against clinician assessment as the reference standard. With TP, TN, FP, and FN denoting true-positive, true-negative, false-positive, and false-negative AI decisions (e.g., for a screening function, a true-positive case is one in which the clinician assessment confirms that the finding was present):

- **Agreement** (clinical alignment or clinical equivalency) = (TP + TN) / (TP + TN + FP + FN), measured over all reviewed cases,
- **Sensitivity** = TP / (TP + FN), measured over clinician-assessed positive cases, and
- **Specificity** = TN / (TN + FP), measured over clinician-assessed negative cases.

Each gated measure receives its own performance floor, defined as the corresponding clinician-clinician agreement value minus a margin. Margins should reflect which error direction poses the greater risk to patients: where false negatives carry substantially greater risk than false positives, the sensitivity margin should be small and the specificity margin may be large, and vice versa. For

example, where a false-negative cancer screening result could plausibly contribute to a patient's death while a false-positive result leads only to an additional visit with an oncologist, the sensitivity margin should not exceed 2 percentage points, while the specificity margin may be as large as 20 percentage points. Proposals should state the chosen margins for each gated measure and justify them by the relative harms of false-negative and false-positive decisions for each clinical function.

Because the statistical power of evaluating sensitivity and specificity of a clinical function (e.g., diagnosis of a given disease) depends on the prevalence of positive (i.e., patient has the given disease) and negative cases (i.e., patient does not have the given disease), the three measures may reach a listed case volume at different points in the pilot. Sandbox applicants should identify the minimum case volume (i.e., earliest look-up point) for each gated performance measure associated with a clinical function. The progression test for a clinical function takes place once all specified volumes have been reached (i.e., at the joint look-up point for a clinical function), and the pilot advances only if all gated measures meet their critical values at that common look-up point.

If the product fails to meet the critical values for a clinical function at a joint look-up point, sandbox participants may attempt the progression test at the next joint look-up point. However, substantial gaps between critical and observed value may warrant a change to the care path or clinical scope, safeguards, or decision engine of the product. Whether a sandbox participant opts to change their product or not, they may notify OAIP that they restart their collection of case reviews for that clinical function and attempt the progression test again once the first joint look-up point for the clinical function is met again.

Appendix C includes a table associating performance floors and case volumes with critical values for progression tests.

Option 1: Function-based progression

For each clinical function (or group of clinical functions, see Section 3.2.1), performance floors should be chosen with a pre-defined margin and reference to clinician-clinician agreement values on the pilot's case mix. The floor for each gated measure is the corresponding clinician-clinician value minus the margin chosen for that measure. For the agreement measure, the margin should not exceed 5 percentage points; for sensitivity and specificity, margins follow the risk-based rationale described above.

Option 2: Progression based on clinical competency

When producing evidence for clinical competency, gated performance metrics should be tied to traits of clinical competency (see Section 3.2.2) rather than clinical functions. The clinicians reviewing cases

throughout pilot stages should state in each case review which traits associated with clinical competency would have been appropriate to demonstrate and if the healthcare AI application successfully demonstrated each of those traits. In addition to measures based on clinical alignment aggregated over all functions, the gated performance measures should include a sensitivity measure for each trait. Performance floors should generally be set high with small margins to a corresponding rate of reviewing clinicians affirming that another clinician successfully demonstrated that trait in an appropriate case. However, performance floors for some traits can be lower if it is reasonable to anticipate a low prevalence of cases that would require clinicians or AI to demonstrate this trait.

6.1.3 Stage regression

Progression through pilot stages is not irreversible. Sandbox applicants should pre-specify in their proposals the conditions under which a pilot returns to a more intensive oversight stage. Regression triggers should include, at minimum:

- a harmful incident (see Section 6.3) whose internal investigation indicates a plausible systematic cause rather than an isolated failure,
- detection of substantial negative performance drift (see Section 6.4) indicated by gated performance measures falling below their respective performance floors at a pre-specified look-up point in intermediate pilot stages or over a sliding time window of pre-specified length in the final pilot stage.

Upon regression, the pilot resumes the oversight mechanisms of the earlier stage, and re-progression follows the pre-specified decision rule of Section 6.1.2, with the collection of case reviews restarting at the beginning of the regressed stage. Sandbox participants should notify OAIP of stage regressions and their triggers in their monthly reports; regressions triggered by harmful incidents follow the notification timelines of Section 6.3.

6.2 Safety-prioritizing pilot termination protocol

6.2.1 General termination protocol

A sandbox pilot may end at its scheduled conclusion, through an OAIP-ordered termination, or through an unplanned company event (e.g., insolvency, acquisition, or loss of key personnel). Regardless of the cause and timing of termination, patients should not experience an interruption of care, a loss of access to their health information, or a silent discontinuation of active treatments. Sandbox applicants should therefore include in their proposals a termination protocol that covers all

termination scenarios and demonstrates that patient safety and continuity of care are preserved throughout wind-down.

A comprehensive termination protocol includes:

- a **minimum notice period** before a voluntary discontinuation of service,
- a **patient-transition protocol** (i.e., warm handover of active and recent patients to alternative care options)
- plans for **service continuity** (e.g., access to medical records and clinical follow-up for concerns arising from care delivered during the pilot)
- plans for providing all patients with a **continuity-of-care summary** that is suitable for sharing with a new provider,
- **minimum financial reserves** (or equivalent instruments, e.g., a bond or insurance) dedicated to executing the transition protocol even in the event of insolvency.

OAIP considers the credibility of this protocol, including the organizational and financial resources backing it, as part of its overall safety assessment (see Section 5.1).

6.2.2 Termination due to safety concerns

If a pilot is terminated because of safety concerns, transitioning patients to alternative care is not sufficient on its own: patients who were adversely affected by the safety issue must be identified and supported. In addition to the expectations above, sandbox participants should commit to

- **proactive outreach** to patients whose care may have been affected by the safety issue, with communication that allows patients to understand what happened and what actions they should take,
- **retrospective chart review** of affected patient cohorts to identify patients who may have experienced harm, including harm that would otherwise remain invisible for detection (see Section 2.2.1), and
- **referral to in-person providers for re-evaluation** where the safety issue may have compromised a diagnosis, treatment recommendation, or prescription, with the cost of re-evaluation not falling on the patient.

6.3 Detecting and handling harmful incidents

Incidents that require mitigating action may include but are not limited to

- severe and unexpected harm to a patient since the involvement of the AI tool in their diagnosis or treatment,

- AI clinical-decision errors including
    - Failure to escalate a case that meets escalation thresholds
    - AI-generated medical recommendations or decisions where any of the following occur:
        - Misdiagnosis or missed diagnosis
        - Incorrect or incomplete treatment recommendation (including failure to recommend appropriate diagnostic workup or follow-up with AI or with a licensed provider)
        - Recommendation for a medication to which the patient has a known allergy
        - Failure to identify a relevant contraindication (e.g., pregnancy, renal, or hepatic contraindications) or drug-drug interaction
        - Incorrect dosing or frequency recommendations
        - Starting or recommending a prescription for a medication not included in the AI tool's approved formulary
        - Altering dosage or frequency of medication to fall outside of approved formulary limits
    - AI encouragement of delayed care-seeking, discontinuation of provider-prescribed treatments, or other behavior not within the standard of care
- system failures including
    - failure to perform ID verification or mandatory safety checks,
    - AI-triggered clinical actions outside its defined clinical scope,
    - loss, corruption, or unavailability of clinical records accessed by the AI tool,
    - cybersecurity incidents.
- compromised data privacy.

Sandbox participants should dedicate adequate resources to the detection, investigation, and mitigation of such incidents. Monthly reports to OAIP should include a complete list of incidents and internal investigation outcomes. Furthermore, OAIP expects to receive notifications of incidents involving severe and unexpected patient harms, critical system failures, and privacy violations involving personal health information within 24 hours of the detection of the incident.

6.4 Change management and drift detection

Sandbox applicants should propose robust frameworks for change management and continuous performance monitoring.

A robust change-management framework should include:

- Information pipelines through which clinical and developer teams receive alerts about changes in clinical best practices established by relevant professional societies, regulators, or other institutions
- Development pipelines that allow new clinical best practices to be incorporated into the AI tool fast and smoothly
- Auditable documentation that tracks clinical and technical updates (e.g., changes to AI architecture, changed LLM base model, etc.) to the AI tool
- Monitoring tools that allow relevant teams to detect changes in demand, patient experience, and patient outcomes linked to clinical and technical changes in the AI tool

A robust performance monitoring framework should include:

- Monitoring tools and pathways that facilitate continuous performance monitoring including a continuous retrospective review with an independent decision reference of a representative sample of cases
- Detection tools and pathways that flag unusual events (e.g., short-term spikes or drops in performance) and long-term trends (e.g., substantial gradual performance drift over several months) for internal investigation
- Internal investigation playbook that facilitates a thorough investigation of events and trends and effective mitigation

7. Continued quality assurance in year 2

Extension of regulatory mitigation agreements beyond the first year are contingent on sandbox participants including continual external validation of their product and measurements of subgroup performance.

In addition to the expectations for year 1 (see Section 6.4), a robust performance monitoring framework for a pilot in its second year should include:

- a concrete commitment to an annual collaboration with third-party auditor, assurance provider, or other entity to test AI performance against continuously updated external data sets or benchmarks to produce up-to-date third-party validated evidence of continued performance,
- a commitment to measure and report performance disparities across relevant subgroups (including vulnerable populations) by the end of the second year of the pilot.

Author Statement

Alice Schwarze is the lead author of the Healthcare AI Sandbox Posture. Zach Boyd is the senior author. Development of this document included consideration of available scientific evidence, applicable regulatory frameworks, and relevant stakeholder perspectives. The document is intended to inform discussion and support further consideration of the issues addressed and should not be interpreted as establishing legally enforceable requirements. The authors thank the reviewers for their thoughtful feedback and contributions. The reviewers brought diverse expertise in healthcare services, spanning academic, clinical, industry, and other relevant domains.

Appendix A: Commonly proposed public-health benefits

1. **Improved access to healthcare services.** AI-aided and AI-led care paths have a credible potential to lower barriers for patients accessing care. However, lowering access barriers in theory does not necessarily lead to adoption in practice, and adoption is likely to be uneven across populations.

   The sandbox proposal for a healthcare-AI pilot must include (1) a value-proposition for Utah patients and (2) commit to collecting and sharing data on adoption of the proposed product or service throughout the period of the pilot. At minimum, the proposing company should commit to sharing aggregate data on volume (e.g., number of patient interactions, number of patients) demographics (e.g., age, sex, insurance status, de-identified geographic location) of patients. OAIP prefers reports that include data on whether patients are in an existing primary care relationship and survey results indicating where patients would have sought care in the absence of the pilot.

2. **Improved access among underserved populations.** Research on adoption of novel tech-based solutions in healthcare suggests that (1) such solutions are most readily adopted by patients who are adequately served by conventional healthcare systems and (2) underserved populations are less likely to adopt these solutions (e.g., because of distrust in new technologies, infrastructural barriers, or tech-literacy barriers).

   Sandbox applicants proposing to close healthcare gaps for underserved populations in Utah should include text in their proposal to (1) identify the population that the aim to reach, (2) critically reflect on foreseeable adoption barriers and how to address them, (3) propose to include metrics in their reporting requirements to OAIP to shed light on care gap closure.

3. **Lower healthcare costs for patients.** By automating clinical and administrative work, AI-aided and AI-led care paths have credible potential to lower the cost of delivering care and, in turn, the prices and out-of-pocket costs that patients face. However, lower delivery costs do not necessarily translate into lower costs for patients: savings may be retained rather than passed on, pricing models (e.g., subscriptions) may obscure the effective cost per episode of care, and costs may shift rather than fall (e.g., when patients later require duplicate in-person visits or additional downstream care).

Sandbox applicants proposing to lower healthcare costs for patients should include in their proposal (1) a transparent description of their pricing model, (2) a comparison of the expected out-of-pocket cost of the proposed care path against the out-of-pocket cost of the conventional care paths it replaces, and (3) a commitment to include cost metrics in their reporting requirements to OAIP. The proposing company should commit to sharing aggregate data on the effective price per patient encounter and per completed episode of care, the distribution of patients' out-of-pocket costs broken down by insurance status, and rates of downstream or duplicate care utilization following use of the tool. OAIP prefers proposals that commit to reporting survey results on whether patients had previously forgone or delayed care for cost reasons.

4. **Improved preventative health engagement and outcomes.** By lowering the friction of routine interactions in preventative or acute care, AI-aided and AI-led care paths have a credible potential to increase patient engagement with preventative services (e.g., screenings, vaccinations, chronic-disease monitoring, and follow-through on referrals). However, engagement gains can be superficial (e.g., app interactions that do not lead to completed screenings), may accrue primarily to patients who are already well engaged rather than to patients overdue for preventative care, and can tip into over-screening or overdiagnosis if recommendations are not guideline anchored.

   Sandbox applicants proposing to improve preventative health engagement should include in their proposal (1) the specific preventative services targeted and the clinical guidelines their recommendations follow, (2) proximal engagement metrics that measure completed preventative actions (e.g., completed screenings, administered vaccinations, closed referral loops) rather than tool usage alone, and (3) a commitment to include these metrics in their reporting requirements to OAIP, broken down by appropriate demographics. OAIP prefers reports that distinguish newly engaged patients from patients who would have received the preventative service through conventional care paths, and that include adherence data showing that recommendations remain within guideline-indicated populations and intervals.

5. **Improved work conditions for Utah clinicians.** By taking over documentation, administrative tasks, and routine clinical work, AI tools have a credible potential to reduce clinician workload, after-hours work, and burnout, thereby supporting clinician retention in Utah. However, efficiency gains do not necessarily improve work conditions in practice: workload may shift rather than shrink (e.g., toward reviewing and supervising AI output), gains may be absorbed by increased throughput expectations, and poorly integrated tools can add friction or alert fatigue instead of removing it.

Sandbox applicants proposing to improve clinician work conditions should include in their proposal (1) the clinician roles and tasks affected by the proposed tool, (2) a baseline measurement plan for the affected workload, and (3) a commitment to include relevant clinician-experience metrics in their reporting requirements to OAIP. The proposing company should commit to sharing aggregate data on time spent on documentation and administrative tasks, after-hours work attributable to the affected care paths, and changes in caseload or throughput expectations for affected clinicians. OAIP prefers reports that include results from validated burnout or job-satisfaction instruments administered before and during the pilot, clinician retention and turnover data, and qualitative insights from clinician feedback.

Appendix B: Concerns about public-health impacts of telemedicine and direct-to-consumer offers

1. **Fragmentation of care.** Patients relying on AI administered (tele-)healthcare for low-complexity health issues may receive a lower quality of care when turning to conventional providers for high-complexity health issues than patients who have an on-going care relationship with a conventional provider who has knowledge of the patient's health history and preferences.

   Pilot proposals offering low-complexity AI administered healthcare should explain (1) their approach to creating accurate, comprehensive, and longitudinal records of interactions with a patient and (2) how they enable a patient's other healthcare providers to get frictionless access to these records when appropriate. At minimum, the proposal should include a commitment to compiling EHRs in a standardized format to be shared over one or several well-established health information exchanges.

2. **Lacking integration into the broader healthcare ecosystem.** Patients relying on AI administered (tele-)healthcare for low-complexity health issues still need to turn to conventional healthcare providers when they encounter high-complexity health issues. Exclusive reliance on AI administered healthcare for low-complexity health issues creates heightened friction and on-boarding costs for patients who turn to conventional healthcare providers only when they experience acute, severe health issues.

   Pilot proposals offering low-complexity AI administered healthcare must detail how patients

are triaged and how high-complexity (or otherwise out-of-scope) cases are detected and handed over to appropriate providers. Handovers should be as warm as possible (ideally with the pilot connecting patients directly to a conventional provider or the associated scheduling service) to ensure patients transition to conventional providers with the lowest-possible attrition rate. To the extent possible, sandbox applicants should commit to measure and report data on successful hand-over rates.

3. **Deterioration of antibiotic stewardship.** AI administered telemedicine with a scope for antibiotic prescriptions would constitute the lowest-friction path for patients to receive antibiotics to date. Public health experts have raised grave concerns about the potential public-health impact of such a program on (1) population-level antibiotic resistance and (2) individual adverse health outcomes from overuse of antibiotics (e.g., rise of chronic conditions like IBS).

   Pilots that offer prescription of antibiotics to patients must propose conservative antibiotic stewardship (e.g., preference for narrow-spectrum antibiotics in their formulary and protocols, safeguards and escalation triggers for cases that involve requests for antibiotic descriptions) in their proposal. They should commit to sharing data validating that conservative approach and the potential public-health impact (e.g., number of patient encounters involving a request for an antibiotic prescription, number of antibiotic prescriptions approved).

4. **Increase of low-value care.** The low-friction access to care via AI-administered telemedicine has the potential to increase use of healthcare services in cases in which patients do not benefit from care. Adoption of a pilot's proposed product or service is thus not sufficient to demonstrate a public health benefit.

   Sandbox applicants should propose a robust path to evaluating the public-health impact of their pilot. Possible solutions include conducting an RCT study with an independent third-party during the duration of the pilot to demonstrate positive health and/or economic outcomes for patients. At minimum, a follow-up with patients (e.g. survey or follow-up visit) after sufficient time for adverse health outcomes to become evident must be included in the pilot proposal.

Appendix C: Critical values for pilot gates

The table below lists, for each performance floor (blue cells), the **minimum number of positively assessed cases** (i.e., clinically equivalent or clinically aligned cases; green cells) required at each case volume (yellow cells), together with the misalignment limit for high-risk cases (red cells). Case volume refers to the denominator of the respective measure: all reviewed cases for agreement, adjudicated-positive cases for sensitivity, and adjudicated-negative cases for specificity. The severe misalignment limit applies to all reviewed cases. "—" indicates that advancement is not possible at this case volume.

Case volume	Performance floor for gated measures is										Number of high-risk misaligned cases
	0.5	0.55	0.6	0.65	0.7	0.75	0.8	0.85	0.9	0.95	
200	135	145	154	162	171	179	186	192	198	—	0
250	160	172	184	195	206	216	226	236	244	250	0
300	185	200	214	228	241	254	267	278	289	298	1
350	210	227	244	260	276	292	307	321	335	346	1
400	235	254	274	293	311	329	347	364	380	394	2
450	260	282	304	325	346	367	387	406	425	442	3
500	285	309	334	358	381	404	427	449	470	489	3
550	309	337	363	390	416	442	467	492	515	537	4
600	335	364	394	423	451	480	507	534	560	585	5
650	359	392	424	455	486	517	547	577	605	632	6
700	384	419	453	488	521	554	587	619	650	680	6
750	410	447	484	520	556	592	627	662	695	727	7
800	434	474	513	552	591	630	667	704	740	775	8
850	459	502	544	585	627	667	707	747	786	822	8

900	484	529	573	618	661	704	747	789	830	870	9
950	509	556	604	650	696	742	787	832	876	917	10
1000	534	584	633	682	731	780	827	874	920	965	11

The critical values are exact binomial bounds (Clopper and Pearson, 1934) ensuring that a tool performing at or below the floor advances with probability of at most 2.5% across all listed testing opportunities combined; the allocation of this error rate across case volumes follows an O'Brien-Fleming-type alpha-spending function (O'Brien and Fleming, 1979; Lan and DeMets, 1983), which makes early advancement contingent on near-flawless performance. Testing at only the listed case volumes (or a subset thereof) preserves this guarantee. The severe-misalignment limits ensure that, at the moment of advancement, the exact upper 97.5% confidence bound on the rate of severe misalignments lies below 2%.

References

Bergman, A., Wachter, R. M., & Emanuel, E. J. (2026). A licensure framework for autonomous clinical AI. JAMA, 335(20), 1751. https://doi.org/10.1001/jama.2026.5483

Bressman, E., Shachar, C., Stern, A. D., & Mehrotra, A. (2026). Software as a medical practitioner—Is it time to license artificial intelligence? JAMA Internal Medicine, 186(1), 5–6. https://doi.org/10.1001/jamainternmed.2025.6132

Clopper, C. J., & Pearson, E. S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. Biometrika, 26(4), 404–413. https://doi.org/10.1093/biomet/26.4.404

Epstein, R. M., & Hundert, E. M. (2002). Defining and assessing professional competence. JAMA, 287(2), 226–235. https://doi.org/10.1001/jama.287.2.226

Institute of Medicine. 1990. Medicare: A Strategy for Quality Assurance, Volume I. Washington, DC: The National Academies Press.

Lan, K. K. G., & DeMets, D. L. (1983). Discrete sequential boundaries for clinical trials. Biometrika, 70(3), 659–663. https://doi.org/10.1093/biomet/70.3.659

Luo, L., Kim, S.E., Zhang, X. et al. A clinical environment simulator for dynamic AI evaluation. Nat Med 32, 820–827 (2026). https://doi.org/10.1038/s41591-026-04252-6

O'Brien, P. C., & Fleming, T. R. (1979). A multiple testing procedure for clinical trials. Biometrics, 35(3), 549–556. https://doi.org/10.2307/2530245