Assurance Reporting Checklist (ARC) for Trustworthy Health AI

Assurance Checkpoint Two

Real-World Impact and Full Deployment Readiness

Coalition for Health AI

June 26, 2024

Checklist Document Versions

As this checklist is passed back and forth between different Reporters and Reviewers, Table 1 will help track versions of the document. Italicized information in the checklist serve as examples and should be replaced during use.

Table 1. Checklist Document Versions

Versions						
Document Version	**Application & Model Version**	**Content Description**	**Reporter or Reviewer Name**	**Contact Information and Role**	**Organization**	**Date**
<1.0>	*<EHR-Based Pediatric Asthma Exacerbation Risk version 1.0 Model 2.0.>*	*<Documentation and evidence provided by implementer and development teams/specific departments from Mayo Clinic>*	*<Name>*	*<Reporter 1>* *E-mail:* *Phone:* *Title:*	*<Mayo Clinic>*	*<May 1, 2024>*
<2.0>	*<EHR-Based Pediatric Asthma Exacerbation Risk version 1.0 Model 2.0.>*	*<Documentation and evidence related to use and human-factors considerations provided by external consultant at ideas42>*	*<Name>*	*<Reporter 2>* *Email:* *Phone:* *Title:*	*<ideas42>*	*<May 5, 2024>*
<3.0>	*<EHR-Based Pediatric Asthma Exacerbation Risk version 1.0 Model 2.0.>*	*<Summary of findings and review of documentation and evidence provided by development and implementer teams at Mayo and consultants from ideas42>*	*<Name>*	*<Reviewer 1>* *Email:* *Phone:* *Title*	*<Mayo Clinic>*	*<May 7, 2024>*

Table of Contents

1 Purpose and Use

1.1. Purpose

The Assurance Reporting Checklist (ARC) is intended to guide the development and evaluation of a complete **AI solution** and **system** against CHAI standards for trustworthy AI[1]. This tool is intended first for self-reporting and self-review, as well as a tool for self-reporting for independent review. The goal of the ARC is to ensure that AI solutions and systems fulfill all five key, principle-based areas for trustworthy AI: 1. Usefulness, Usability, and Efficacy; 2. Fairness, Equity, and Bias Management; 3. Safety; 4. Transparency and Intelligibility; 5. Privacy and Security. In alignment with these areas, the ARC translates best practice considerations (detailed in the Assurance Standards guide) that meet core ethical and quality principles into detailed yes/no questions, or evaluation criteria, to determine whether best practice standards are met (see accompanying Assurance Standards Guide). The relationship between evaluation criteria and their original considerations, as well as criteria that have been combined across multiple areas and considerations are mapped in a Traceability Matrix located in the Appendix of this document (Section 3.1). The ARC encourages a holistic understanding of AI solutions in context, encompassing the interplay of human-factors, data, algorithms, infrastructure, and real-world workflows, facilitating conversations across developer and implementer teams, and As a self-review tool for developer and implementation teams, this iteration of the ARC also serves as a starting point for facilitating conversation and alignment on best practices across the full AI lifecycle.

A secondary purpose of this version of the tool is to guide an understanding of the state of trustworthy AI in healthcare and the needs of diverse stakeholders and healthcare organizations by stress-testing the checklist in the real-world. Specifically, utilization of this tool and feedback on existing end-to-end capabilities and practices will aid both in improving and iterating on the ARC and its subsequent versions, as well as an understanding of the challenges that may influence the feasibility of best practices.

1.2. Intended Users

Intended users of the ARC are developer and implementation teams within or outside of health systems with accountable Reporters from teams providing documentation and summaries for executive review. Multiple stakeholders (see section 3.3 in the Appendix and section 3.2 in the Assurance Standards Guide) may be involved in the selection, procurement, development, and deployment process of an AI solution. This iteration of the ARC does not prescribe roles and responsibilities, however it outlines usage for those completing and reviewing the document (see Assurance Standards Guide, pg. 2 for further details on this

[1] The ARC was developed by forming expert workgroups for each principle area. Workgroups conducted a full landscape analysis and synthesized findings into a series of considerations and criteria for each lifecycle stage for their specific principle-based focus areas. These considerations and criteria were then compiled into a survey sent out to the broader CHAI community to gain multi-stakeholder feedback and ratings as part of a modified Delphi-process to gain consensus across multiple stakeholders. Results were then reviewed during the Fall convening and discussed further. Considerations that were rated as “Extremely Important” by at least 50% of the respondents, and/or were deemed extremely important following the second round of discussions, were included in this version of the Assurance Standards Guide and Checklist. Additional considerations and criteria that were rated as either “Extremely Important” or “Very important” by at least 65% of survey respondents are included in the Traceability Matrix but not in this version of the Assurance Standards Guide or Checklist.

and plans for future versions). Developer and implementer teams may be entirely or in part internal or external to the healthcare organization looking to develop, procure, or implement an AI solution. As such, this tool may also be used as part of a collaborative process across developer and implementer teams to foster trust and alignment on best practices.

This checklist is most appropriate for products or devices that are themselves AI software (predictive or generative) or those that are AI assisted/AI enabled. At this point in time, AI tools often used in drug discovery and development (e.g. target selection or antibody design) in the pharmaceutical industry fall outside the targeted scope of the ARC.
AI software examples: Payer/provider risk stratification or prediction, diagnostic algorithms, automated EHR coding, provider decision or administrative support, patient decision support, patient or provider facing chatbot used for education or assistance
AI assisted/AI enabled examples: AI enabled medical devices, AI assisted surgical robots, radiological technologies that are AI assisted or AI enabled for clinical (diagnostic/risk prediction) or nonclinical purposes (automated image quality enhancement.)

The **Reporter** is the individual tasked to gather responses and documentation from appropriate **"Providers of Evidence,"** or experts in the areas pertaining to ARC items. The **Reviewer** can either be an internal executive responsible for checking the completeness and appropriateness of the explanations and documentation to guide the development, procurement, and/or implementation of an AI solution based on best practices, or an external independent Reviewer who will evaluate the overall AI system for alignment with best practices. Note that there may be multiple Reporters, Providers of Evidence, and Reviewers. For smaller organizations or health systems there may be fewer stakeholders available, or the need to consult with external experts to ensure best practices in specific areas. We do not expect that all best practice standards are feasible at this point and aim to further understand the feasibility of these standards as they are stress-tested in the real world. Examples of user personas and scenarios are provided in the Appendix (section 3.4).

1.3. Usage

Usage of the ARC is guided by the AI Lifecycle (Figure 1). The AI Lifecycle can be an iterative and non-linear/agile outline of the processes required for effective and trustworthy design, development, and use of a health AI system from end-to-end. To facilitate the agile process, we have identified a **planning checkpoint** and several **assurance checkpoints** that aim to help teams ensure that the necessary steps have been taken, and standards met, prior to moving a tool into real-world use. The four checkpoints are summarized below. Examples of user personas and scenarios are provided in the Appendix (section 3.4).

1. The **planning checkpoint** follows Stage 1, where both developer and implementer teams (independently or together) are asked to define the specific problem and plan adequately for a potential AI solution. This checkpoint primarily helps teams:
    a. Appropriately consider the risks, benefits, costs, and needs for an AI solution both at the clinical and population levels
    b. Consider the risks, benefits, costs, and needs around purchasing or developing an AI solution in house
    c. Gain multi-stakeholder insights to help guide human-centered AI solution design, development (or purchasing) and downstream needs to maximize real-world effectiveness and trust
2. **Assurance checkpoint one** appears when progressing from iterations through design, development, and assessment processes, to the small-scale pilot phase. The goal of this checkpoint is to address readiness for piloting and to prepare for real-world risks and needs. Any updates to clinical and population risk summaries should be made based on new insights from the design, development, and silent-evaluation process. An important note is that this checkpoint is not only meant for developer organizations. There are items that assess for readiness for the implementer/purchasing organization,

items to guide conversations around responsibilities between developer and implementer organizations, items that speak to the larger AI system design and development (e.g. safety, privacy, security, and monitoring planning), and items that a purchasing/implementing organization may use to understand vendor best practices. An organization or health system acquiring or purchasing an AI solution may choose to use this checkpoint as part of their procurement process. For example, they may require developer organizations to provide relevant evidence in support of best practices during design, development, and evaluation to help make purchasing decisions to foster transparency. It is also recommended that purchasing/implementing organizations review the planning checkpoint items alongside the developer organization to ensure appropriate planning, risk determination, and usability for the broader AI system (beyond the AI solution alone).

3. **Assurance checkpoint two** appears when progressing from piloting to at-scale deployment of the AI system, which requires evaluation of readiness and preparation for the broader needs and wider scope of risk. Any updates to clinical and population risk summaries should be made based on new insights from initial real-world piloting.
4. **Assurance checkpoint three** appears following full scale deployment to evaluate for longer-term readiness for monitoring, managing, and updating the AI system. This checkpoint is repeated throughout regular monitoring of the AI solution, at appropriately timed intervals depending on the use case, and as dictated by the developer and/or implementer organization. As in previous checkpoints, updates should be made to clinical and population risk summaries based on insights gained from regular monitoring of AI solutions and systems.

Within each checkpoint checklist, relevant evaluation criteria are listed and given an identifier. The color coded Evaluation Criteria Identifier (EC Identifier) links each criterion to the original consideration as defined within principle area workgroups (see Traceability Matrix in the Appendix 3.1; See Section 1.5 for further details.)

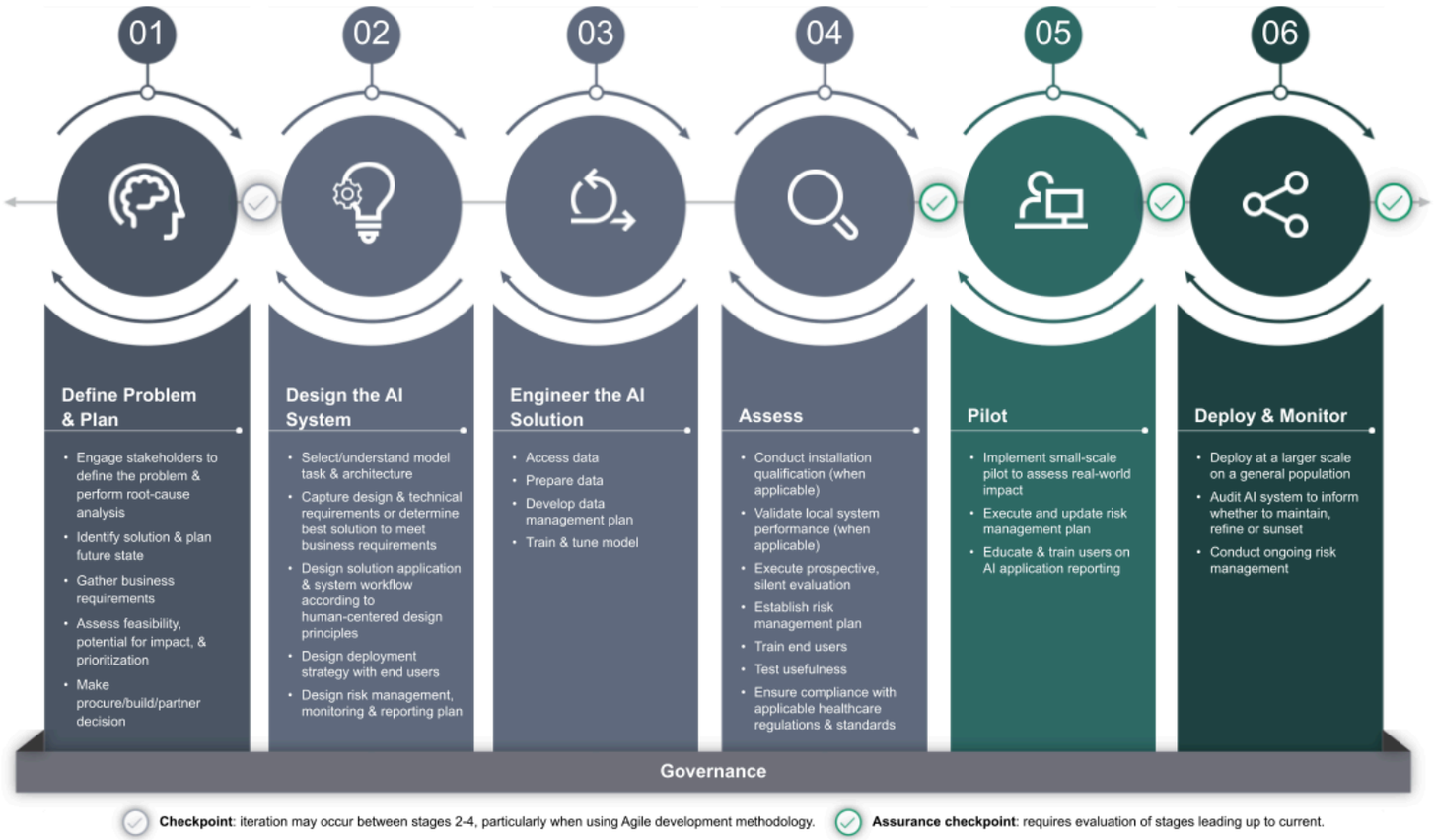

Figure 1: The CHAI AI Lifecycle Framework. Derived from CHAI Assurance Standards Guide. The gray checkmark represents the Planning Checkpoint, while the green checkmarks correspond to Assurance Checkpoints 1-3.

1.4 How to complete this checklist

1.4.1 General

Who Should Complete This Checklist?

Each checkpoint checklist should first be completed by at least one **Reporter.** While there may be multiple stakeholders involved in sharing evidence necessary to respond to criteria, the Reporter is the individual responsible for requesting this information (if available), making sure available evidence is clearly documented for relevant evaluation criteria in the checklist, and indexing it in a centralized place for ease of Reviewer access. They will also provide a summary at the end of each checkpoint that provides reviewers with a broad overview of the potential or observed benefits, costs, risks, and/or adverse events associated with that checkpoint. Example roles, professions, and representative organizations are shown in Table 3.3 in the Appendix and described in more detail in the CHAI Assurance Guide.

Reporters will then pass the checklist off to at least one **Reviewer** who is internal to either developer and/or implementer organizations (such as an area specific executive). Ideally, organizations will also pursue independent and external third-party review. The Reviewer will go over the responses to evaluation criteria and evidence, and indicate whether best practice standards for each criteria have been met. They will also provide a summary of findings based on the available evidence and any observed gaps. This feedback can be used to improve processes, help guide teams on next steps, or help build/design solutions to fill gaps in best practice standards.

For **Assurance Checkpoints 1-3** the following steps are required.

Reporter Responsibilities for Completion (Assurance Checkpoints 1-3)

1. All Reporter required sections of the checklist or summaries are denoted with dark blue coloring.
2. Provide existing (from prior checkpoints) and updated clinical risk classification and Population Impact information in the “Clinical Risk and Population Impact Summaries” table at the start of each Checkpoint (Review Clinical Risk and Population Impact Tools in sections 1.4.2 and 1.4.3 respectively for any necessary updates).
3. The Reporter will then complete the relevant Assurance Checkpoint Checklist providing a brief explanation and document code in the “Evidence and Explanation & Metadata/Documentation Code” column of the checklist, with supporting evidence indexed within the “Evidence & Explanation Metadata Table” (see Section 2.5 for further instructions and Table).
4. The Reporter will complete the “Executive Summary of Anticipated and Observed Benefits, Risks, and Limitations” section (Section 2.3) for the relevant Assurance Checkpoint.
5. Reporter responsibilities for each Assurance Checkpoint Checklist will end by updating the document version table (Page 2) and up-versioning the document header, prior to sending the checklist and associated evidence to the appropriate Reviewer.

Reviewer Responsibilities for Completion (Assurance Checkpoints 1-3)

1. All Reviewer required sections of the checklist or summaries are denoted with light blue coloring.

2. The Reviewer will go through information provided in the checklist by the Reporter along with accompanying documentation listed in Evidence and Explanation Metadata table.
3. Reviewers will then complete the Summary of Findings table (Section 2.4), summarizing findings provided in the checklist by the Reporter in the context of anticipated and observed benefits, risks, and limitations of the AI solution.
4. Reviewers will then update the document version table on Page 2 and up-version the document header.

Example Reporter Role Responses

Checklist: Stage 2-4 | Design, Engineer, and Assess the AI Solution							
EC Identifier	**Evaluation Criteria**	**Evidence and Explanation Metadata/Document Code**	**Reporter Initials & Date**	**Evidence & Explanations (Yes/No/Partial/NA)**	**Limitations or Adverse Outcomes**	**Criteria Met (Yes/No/Partial/NA)**	**Reviewer Initials & Date**
Assurance Checkpoint 1: Readiness for Real World							
LS2.F.C1.EC2	Will the real-world/clinical outcome measure be available for evaluation within an adequate time frame and in a manner that accurately represents the target population?	*Evidence and explanation: Real-world retrospective data was used for evaluation of model performance and comparable to target population.* *Metadata/Document Location: <insert link to bias assessment document and relevant data showing summary of real-world retrospective data population descriptives and demographics and comparison to target population descriptives and demographics.)*	*M.G. 05/06/2024*				
LS2.F.C1.EC3	Will real-world/clinical outcomes be systematically compared for equity across all relevant socio-demographic subgroups, ensuring fairness and addressing potential bias?	*Evidence and explanation: Overall ER admission rates are lower following use of the AI solution. Clinical outcomes are similar for all subgroups except for Black Patients, who show higher ER admissions following discharge at the same population level risk threshold compared to the sample majority group and compared to the population mean.* *Metadata/Document Location: <insert link to bias assessment document and relevant data showing likelihood of ER admissions following discharge (as measure of clinical outcomes that AI solution aimed to impact)*	*M.G. 05/06/2024*				

Example Reviewer Role Responses

Checklist: Stage 2-4 | Design, Engineer, and Assess the AI Solution							
EC Identifier	**Evaluation Criteria**	**Evidence and Explanation Metadata/Document Code**	**Reporter Initials & Date**	**Evidence & Explanations (Yes/No/Partial/NA)**	**Limitations or Adverse Outcomes**	**Criteria Met (Yes/No/Partial/NA)**	**Reviewer Initials & Date**
Assurance Checkpoint 1: Readiness for Real World							
LS2.F.C1.EC2	Will the real-world/clinical outcome measure be available for evaluation within an adequate time frame and in a manner that accurately represents the target population?	*Evidence and explanation: Real-world retrospective data for ER admission rates are available and will be used for evaluation of model's impact on clinical outcomes. Data is comparable to target population.* *Metadata/Document Location: <insert link to bias assessment document and relevant data showing summary of real-world retrospective data population descriptives for measure and demographics and comparison to target population descriptives for measure and demographics of sample.)*	*M.G. 05/06/2024*	*Yes*	*No, None stated*	*Partial, Provide justification for why this clinical outcome was selected.*	*N.E. 05/10/2024*
LS2.F.C1.EC3	Will real-world/clinical outcomes be systematically compared for equity across all relevant socio-demographic subgroups, ensuring fairness and addressing potential bias?	*Evidence and explanation: Overall ER admission rates are lower following use of the AI solution. Clinical outcomes are similar for all subgroups except for Black Patients, who show higher ER admissions following discharge at the same population level risk threshold compared to the sample majority group and compared to the population mean.* *Metadata/Document Location: <insert link to bias assessment document and relevant data showing likelihood of ER admissions following discharge (as measure of clinical outcomes that AI solution aimed to impact)*	*M.G. 05/06/2024*	*Partial, provide information on what threshold was selected and why.*	*Yes, Black patients have poorer outcomes at the chosen threshold*	*Partial*	*N.E. 05/10/2024*

1.4.2 Clinical Risk Evaluation

Risk should be assessed from both the **clinical** and **population** perspective. For **clinical risk**, we adopt the International Medical Device Forum's (IMDRF's) categorization system for assessment of clinical risk (See Table 2). This should be done by a licensed clinician based on the FDA IMDRF guidance.

Table 2. Assessment criteria for clinical risk level. Levels are described in detail in "Software as a Medical Device": Possible Framework for Risk Categorization and Corresponding Considerations" by IMDRF Software as a Medical Device (SaMD) Working Group (2014).

Clinical Risk Classification			
State of Healthcare situation or condition	**Significance of information provided to healthcare decision**		
	Treat or diagnosis	Drive clinical management	Inform clinical management
Non-Serious	II	I	I
Serious	III	II	I
Critical	IV	III	II

Clinical risk classification and summaries should be provided in **Section 2.1, Table 3. Clinical Risk and Population Impact Evaluation Summaries**

1.4.3 Population Impact Evaluation Tool

Population risk refers to how systemic, individual, and group-level tendencies when combined with decision-making demands across the AI lifecycle, can impact health and well-being for entire subgroups and over longer periods.

While it is common to refer to systemic, individual, and group-level tendencies as "biases"—it is important to note that they are often the result of things like:

- Historical Norms/policies
- Current Societal Norms/policies
- Scope of Skills/Responsibilities

- Natural limitations/variability in cognitive resources/awareness
- The burden of increasing clinical/administrative demands
- Role specialization (and therefore less insight into other roles or expertise)

It is normal for us to:

- Not have all knowledge about a topic
- To want to use data that is readily available or easily accessible
- To be focused on our role-specific responsibilities and not aware of the roles/responsibilities of others
- To focus on resolving a specific problem (e.g. sepsis prediction), without considering how it might unintentionally harm a subgroup of individuals due to bias in data/measurement
- To want to follow shortcuts

The following questions will help stakeholders involved in purchasing or developing an AI solution, together with other relevant stakeholders (see Section 3.3 in the Appendix) to evaluate population risk and impact in a way that will improve current practices and minimize population-level harm across several domains. This will allow teams to leverage the power of health AI to positively impact patients and providers and reduce healthcare gaps and inequities, rather than perpetuate or prolong them. These questions are best explored with patient advocacy/population health and medical area experts present or consulted. Given that bias in AI is unavoidable, this tool will also help organizations evaluate and prioritize bias mitigation efforts towards algorithms with greater risk and/or those that may be impacted by ethical/legal guidelines. Using this tool aims to improve current practices and minimize population-level harm. (Tool adapted to health-specific context in part from ethicstoolkit.ai)

Identify who will be impacted by the AI system:

Primary Impacted: Who or what may be or is directly impacted based on the objectives of the AI system? (e.g. patients, family caretakers, physicians, nursing, organization, business operations, etc.)

Secondary: Who or what may be or is impacted downstream based on those primarily impacted? (e.g. if physicians and their clinical workflows are primarily impacted, downstream effects may be experienced by nursing staff, or radiology technicians)

Unexpected/Unintended: Who or what may be impacted unexpectedly/unintentionally at the population or location level? Examples may include:

- Patients who do not speak English or their children
- Physicians working in community hospitals vs. academic medical centers
- Patients without insurance
- Acquired hospitals that use a different (non-integrated) electronic medical record system
- Members of a specific socio-demographic subgroup
- Individuals with visible or invisible disabilities

Select the types of impact that the AI system may have on PATIENTS and the degree, scale, and direction of impact for each type:

- **Access to Health Goods/Benefits:**
  Algorithms that impact who, what, where, or how someone does/does not have access health goods or benefits (ability to track health status, ability to access test results, disease management, advanced care management services)
  Select Degree: Minor Impact | Moderate Impact | Major Impact
  Select Scale: Small Proportion | Substantial Proportion OR Primarily one or more Vulnerable Subpopulations | Nearly Every Person OR Majority of one or more Vulnerable Subpopulations
  Select Direction: Positive Impact | Mostly Positive Impact | Mostly Negative Impact | Negative Impact

- **Access to Direct Health Services/Healthcare:** Algorithms that impact who or how someone does/does not have access to necessary direct health care services (transportation coordination, medicine or health service approval, preventative care appointments, specialty care services, diagnostic testing, mental health screening, etc.)
  Select Degree: Minor Impact | Moderate Impact | Major Impact
  Select Scale: Small Proportion | Substantial Proportion OR Primarily one or more Vulnerable Subpopulations | Nearly Every Person OR Majority of one or more Vulnerable Subpopulations
  Select Direction: Positive Impact | Mostly Positive Impact | Mostly Negative Impact | Negative Impact

- **Emotional Health/Well Being:** These algorithms impact the emotional health or well-being of an individual or group. (Time waiting for health services/benefits, effort required to arrange for services)
  Select Degree: Minor Impact | Moderate Impact | Major Impact
  Select Scale: Small Proportion | Substantial Proportion OR Primarily one or more Vulnerable Subpopulations | Nearly Every Person OR Majority of one or more Vulnerable Subpopulations

Select Direction: Positive Impact | Mostly Positive Impact | Mostly Negative Impact | Negative Impact

- **Life/Safety:** These algorithms directly impact individual or group safety or life (e.g. diagnostic, treatment, recommended treatments)
  Select Degree: Minor Impact | Moderate Impact | Major Impact
  Select Scale: Small Proportion | Substantial Proportion OR Primarily one or more Vulnerable Subpopulations | Nearly Every Person OR Majority of one or more Vulnerable Subpopulations
  Select Direction: Positive Impact | Mostly Positive Impact | Mostly Negative Impact | Negative Impact

- **Financial:** These algorithms impact the costs associated with healthcare for individuals, groups, or in specific areas. (e.g. health plan premiums, cost of care)
  Select Degree: Minor Impact | Moderate Impact | Major Impact
  Select Scale: Small Proportion | Substantial Proportion OR Primarily one or more Vulnerable Subpopulations | Nearly Every Person OR Majority of one or more Vulnerable Subpopulations
  Select Direction: Positive Impact | Mostly Positive Impact | Mostly Negative Impact | Negative Impact

- **Privacy:** These algorithms impact the privacy of personal health information for an individual or group.
  Select Degree: Minor Impact | Moderate Impact | Major Impact
  Select Scale: Small Proportion | Substantial Proportion OR Primarily one or more Vulnerable Subpopulations | Nearly Every Person OR Majority of one or more Vulnerable Subpopulations
  Select Direction: Positive Impact | Mostly Positive Impact | Mostly Negative Impact | Negative Impact

- **Trust:** These algorithms impact the trust that an individual or group may have in the healthcare system, clinician(s), or other healthcare professional.
  Select Degree: Minor Impact | Moderate Impact | Major Impact
  Select Scale: Small Proportion | Substantial Proportion OR Primarily one or more Vulnerable Subpopulations | Nearly Every Person OR Majority of one or more Vulnerable Subpopulations
  Select Direction: Positive Impact | Mostly Positive Impact | Mostly Negative Impact | Negative Impact

- **Freedom/Agency/Rights:** These algorithms impact an individual's freedom/agency/rights as it pertains to their healthcare or health information.
  Select Degree: Minor Impact | Moderate Impact | Major Impact
  Select Scale: Small Proportion | Substantial Proportion OR Primarily one or more Vulnerable Subpopulations | Nearly Every Person OR Majority of one or more Vulnerable Subpopulations
  Select Direction: Positive Impact | Mostly Positive Impact | Mostly Negative Impact | Negative Impact

Is it possible that the degree or scale of impact could vary by context (population subgroup or location implemented).

- **No likelihood of systematic variation in scope of impact by context**
- **Small likelihood of systematic variation in scope of impact by context, but variability is due to known and validated clinical or social needs**
- **Small likelihood of systematic variation in scope of impact by context**
- **Medium likelihood of systematic variation in scope of impact by context, but variability is due to known and validated clinical/social needs**
- **Medium likelihood of systematic variation in scope of impact by context**
- **High likelihood of variation in scope of impact by context, but variability is due to known and validated clinical/social needs**

- **High likelihood of variation in scope of impact by context**

1.5 How to interpret this checklist

The checklist is designed not as a binary pass-fail assessment, but rather as a comprehensive tool to evaluate the risk-benefit profile of the AI solution and its associated system and to guide best practices across developer and implementer teams. Given the inherent complexity of each use case and implementation, a nuanced approach is essential. The checklist aims to facilitate transparency and furnish reviewers with substantial evidence, empowering relevant parties to make informed go/no-go decisions. Furthermore, it underscores the importance of additional measures that may be undertaken by the implementation or developer organization. These measures are crucial for preventing and mitigating adverse outcomes, as well as ensuring that the AI solution is employed judiciously in contexts where its limitations are acknowledged and respected.

Throughout the checklist, each evaluation criteria has received one or more coding tags in the left-hand column (example: LS1.U.C1.EC1). These identifiers are designed for traceability to the considerations in the Assurance Standards Guide, and they are color-coded by principle area. Some evaluation criteria are based on considerations that space multiple principle areas or span multiple considerations within a principle area.:

- Usefulness, Usability, Efficacy: (Principle Area Denoted with U)
- Fairness, Equity, and Bias Management: (Principle Area Denoted with F)
- Safety: (Principle Area Denoted with S)
- Transparency, Intelligibility, and Accountability: (Principle Area Denoted with T)
- Privacy and Security: (Principle Area Denoted with PS)

(example: LS1.U.C1.EC1 would denote Lifecycle Stage 1, Usefulness, Usability, and Efficacy Principle Area, Consideration 1, Evaluation Criteria 1.)
Note: once the review of the checklist is complete, we'll be creating more streamlined, sequential tags. For now, the color coding will give you what's most important, as many evaluation criteria reflect overlaps in different principle-based considerations through the lifecycle.

2 Reporting Checklist

Columns and sections to be completed by the Reporter are denoted in dark blue and by Reviewer in light blue.

2.1 Clinical Risk & Population Impact Evaluation Summary

Clinical Risk and Population Impact Evaluation tools are provided in sections 1.4.2 and 1.4.3 respectively. ***Reporters*** *should provide a summary of clinical risk (including classification level) in Table 3 below, and a summary of population impact initially in the Planning Phase (Stage 1). If not completed during the Planning Phase* ***and*** *as insights are gained during subsequent Checkpoints, tools in sections 1.4.2 and 1.4.3 should be revisited and information in Table 3 should be updated.* ***Reviewers*** *should go over this information to gain context for the information that follows in the checklist (Section 2.3).*

Table 3. Clinical Risk and Population Impact Summaries

Clinical Risk Classification & Population Impact Summaries		
Domain		**Reporter Initials and Date**
Clinical Risk Classification & Summary		
Population Impact Summary		

2.2 Checklist Stage 5: Pilot

	Checklist: Stage 5 | Pilot							
Criterion Number	**EC Identifier**	**Evaluation Criteria**	**Evidence and Explanation & Metadata/Document Code**	**Reporter Initials & Date**	**Evidence & Explanations (Yes/No/Partial/NA)**	**Limitations or Adverse Outcomes**	**Criteria Met (Yes/No/Partial/ NA)**	**Reviewer Initials & Date**
	Assurance Checkpoint 2: Real-World Impact and Full Deployment Readiness							
AC2.CR1	LS5.F.C2.EC1	Does the population, site, department, or program chosen for the pilot adequately represent the entire population that the AI system will eventually target, ensuring inclusivity and generalizability?						
AC2.CR2	LS5.F.C2.EC2	Is there a possibility that the method or definition of the pilot population could result in disproportionate exclusion or inclusion of a sociodemographic subgroup, potentially introducing bias?						
AC2.CR3	LS5.F.C2.EC3	Has the potential impact of inclusion or exclusion of sociodemographic subgroups been thoroughly evaluated, considering the potential implications for bias, fairness, and equity in the AI system's performance?						
AC2.CR4	LS5.PS.C7.EC3	Are the mechanisms for integrating contextual factors like demographics and privacy preferences (e.g.,surveys, focus groups, generative AI learning models, interactions with users, etc.) functioning as intended?						
AC2.CR5	LS5.PS.C7.EC1	Are specific personnel assigned responsibility for incorporating contextual factors, including individual demographics and privacy preferences, into the design of the AI system?						

AC2.CR6	LS5.PS.C7.EC2	Has the organization defined the expected and acceptable context of use for the AI system, considering factors such as demographics, privacy interests, data sensitivity, and visibility of data processing to individuals and third parties?						
AC2.CR7	LS5.PS.C1.EC1	Does the organization conduct a privacy risk assessment or similar exercise on AI systems to understand stakeholder privacy preferences, ensuring that these preferences are adequately considered in AI design?						
AC2.CR8	LS5.PS.C1.EC2	Is AI system processing analyzed to align with stakeholder privacy preferences, either formally through audits or Data Protection Impact Assessments, or informally through committee meetings, facilitating ongoing evaluation and adjustment to ensure alignment with privacy preferences?						
AC2.CR9	LS5.PS.C6.EC1	Are relevant staff members interviewed to determine potential security and privacy risks associated with the AI system?						
AC2.CR10	LS5.PS.C6.EC4 LS5.PS.C6.EC5	Do organizational data governance policies include a framework for managing risks and enhancing trust that includes corrective actions to improve data quality, accuracy, reliability, representativeness, security and privacy associated with AI system deployment?						
AC2.CR11	LS5.PS.C5.EC3	Does the organization maintain records demonstrating that the contingency plan has been systematically tested and validated to assess its effectiveness in real-world conditions?						
AC2.CR12	LS5.PS.C3.EC3	Can the organization and third parties easily and securely share configuration changes with each other, facilitating collaboration and ensuring consistency across environments?						

AC2.CR13	LS5.PS.C2.EC1 LS5.PS.C2.EC2	Does the organization regularly log, audit, and review AI system user access and AI system inputs/outputs to ensure compliance with policies and detect any anomalies or unauthorized access?						
AC2.CR14	LS5.PS.C4.EC1 LS5.PS.C4.EC2	Has the organization developed and documented a comprehensive privacy and security incident response plan that is updated regularly as appropriate?						
AC2.CR15	LS5.PS.C2.EC3	Does the audit log of user access and AI inputs/outputs capture only the amount of data needed, incorporating principles of data minimization and data privacy to reduce the collection and retention of unnecessary information?						
AC2.CR16	LS5.PS.C3.EC1	Does the organization have formal procedures to record and store changes made to the AI environment, ensuring that a comprehensive history of modifications is maintained?						
AC2.CR17	LS5.PS.C3.EC2	Are the configuration change records traceable to the owner, and do they include detailed information about the nature and scope of the change, providing accountability and transparency?						
AC2.CR18	LS5.S.C3.EC1 LS5.S.C3.EC9 LS5.S.C3.EC10 LS5.PS.C6.EC2	Is there a risk management plan and standard operating procedures (SOPs) in place and readily accessible by the implementing organization, that allows for evaluation and management of safety, bias, and privacy/security risk?						
AC2.CR19	LS5.T.C3.EC5	Has accountability and responsibility for decision-making been clearly defined and legally vetted, ensuring clarity in the event of unforeseen outcomes?						
AC2.CR20	LS5.S.C3.EC2 LS5.PS.C6.EC3	Are there procedures for systematic reporting and mitigation of adverse events, safety issues, or privacy and security issues among both implementer and developer organizations?						

AC2.CR21	LS5.S.C1.EC2 LS5.S.C1.EC3 LS5.T.C3.EC2	Does the organization adhere to a common standard for defining, tracking "adverse events" and "serious adverse events" separately, and reporting them in a timely manner to all relevant stakeholders?						
AC2.CR22	LS5.S.C1.EC1	Is there a plan in place for monitoring safety risks associated with the AI system, including adverse events and serious adverse events, with a breakdown by severity and frequency?						
AC2.CR23	LS5.S.C2.EC2	Are there processes in place to identify, assess, and manage risks arising from changes to the AI system, environment, and data, including considerations such as memory usage and allocation, communication dependencies, operational speed, task prioritization, display management, and user input capabilities?						
AC2.CR24	LS5.S.C2.EC1	Is there effective communication and mutual understanding between the health system and the developer organization regarding any potential risks resulting from changes to the architecture and code of the AI system?						
AC2.CR25	LS5.S.C3.EC3 LS5.S.C3.EC5	Is there a defined process to assess and triage safety issues and poor outcomes, determining whether and how the AI system should continue operating, undergo refinement, or be discontinued?						
AC2.CR26	LS5.S.C1.EC4	Are there established protocols for sunsetting (triggering a backup plan) and conducting safety investigations in response to identified adverse events or serious adverse events, ensuring that appropriate actions are taken to mitigate risks?						
AC2.CR27	LS5.PS.C5.EC1 LS5.PS.C5.EC2	Has the organization documented a contingency plan and trained responsible personnel to ensure delivery and resilience of services in the event of AI service disruption or failures?						

AC2.CR28	LS5.U.C1.EC2 LS5.F.C3.EC2	Does the documented risk-benefit assessment include information on the potential for trust in and usability of the AI solution?						
AC2.CR29	LS5.U.C2.EC1 LS5.U.C2.EC2 LS5.U.C2.EC3	Is the AI model superior to the standard of care in terms of benefits, risks, and costs (e.g. are there improvements in error rates, efficiency, or outcomes?), and is the relative benefit of the AI solution documented?						
AC2.CR30	LS5.S.C3.EC8	Is the AI implementation plan designed to be auditable by independent third parties, ensuring transparency and accountability in the risk management processes?						
AC2.CR31	LS5.T.C3.EC3	Have all unforeseen, unintended negative outcomes been sufficiently assessed and documented during the pilot stage, ensuring comprehensive understanding and strategies for mitigation?						
AC2.CR32	LS5.T.C3.EC1	Are the required mitigation steps known to key stakeholders/end users, ensuring their awareness and readiness to take appropriate actions?						
AC2.CR33	LS5.S.C5.EC1 LS5.S.C5.EC2	Is there a comprehensive protocol and established channels for promptly reporting safety concerns regarding the AI solution to the developer organization, regulatory agencies (when applicable), end users, and relevant stakeholders, thereby ensuring timely investigation, mitigation, and communication of potential risks to patient safety?						
AC2.CR34	LS5.T.C2.EC2	Has a designated point-person been identified to champion the training, implementation, and follow-up processes for end users?						
AC2.CR35	LS5.U.C5.EC1	Are the tasks that involve the use of the AI solution adequately supported, ensuring that end users can effectively carry out their intended actions?						

AC2.CR36	LS5.F.C3.EC2 LS5.S.C4.EC2	Is the risk for automation bias and its potential harms clearly addressed in the training material provided to end users, and proactively accounted for in the design of the AI solution user interface?						
AC2.CR37	LS5.U.C1.EC1 LS5.T.C2.EC1 LS5.T.C6.EC1 LS5.T.C6.EC2	Are end users trained with standard education materials and evaluated for clear understanding of the capabilities, limitations, interpretation, and appropriate use of the AI solution to ensure effective decision-making processes?						
AC2.CR38	LS5.T.C4.EC1 LS5.S.C6.EC1 LS5.S.C6.EC2 LS5.S.C6.EC3	Have human factors assessments, qualitative assessments (e.g. focus groups, surveys, follow-up studies), and quantitative assessments been conducted to evaluate user experience, acceptance of the AI solutions, patient experience, safety, and other related considerations and issues during real-world implementation?						
AC2.CR39	LS5.U.C4.EC1 LS5.S.C3.EC6	Is there a documented plan in place to investigate and manage clinician disagreements with the AI solution outputs or decisions (including mechanisms like human in the loop or human override)?						
AC2.CR40	LS5.U.C3.EC1	Has the usability or effectiveness of the AI solution demonstrated any changes when deployed in an actual clinical environment (such as changes in user efficiency, effectiveness, and satisfaction)?						
AC2.CR41	LS5.U.C5.EC2	If actions taken by end users are different than originally anticipated, are these actions documented, capturing potentially unforeseen consequences or user behaviors?						

AC2.CR42	LS5.F.C3.EC1	Are there data available and methods defined to evaluate whether the AI system is being used as intended by end users and whether variability in end-user behavior impacts treatment or outcomes of specific sociodemographic subgroups, ensuring adherence to intended use and identifying potential biases?						
AC2.CR43	LS5.F.C3.EC3	Have systemic tendencies in user decision-making been evaluated to determine if they introduce differential outcomes for subgroups (e.g., due to automaticity, lack of trust in AI solution, etc.), allowing for identification and mitigation of potential biases?						
AC2.CR44	LS5.S.C4.EC3	Is it possible to measure automation bias, and is this measurement included in the risk assessment process (for example, determining whether the incorrect AI output can be detected and how it may have potential impact on subsequent decision-making)?						
AC2.CR45	LS5.S.C4.EC1	Is the potential for automation bias explicitly described and assessed within the organization's risk management plan for the AI solution?						
AC2.CR46	LS5.S.C3.EC4 LS5.T.C4.EC2 LS5.T.C3.EC4 LS5.T.C4.EC3	As part of the implementation, is there a structured feedback loop in place for the detection and reporting of safety, usability, bias, or other relevant issues to the developer organization or AI developer for mitigation or improvement (passively and with end-user feedback)?						
AC2.CR47	LS5.S.C3.EC7	Are deviations from expected outcomes systematically tracked and documented for root cause						

		analysis, allowing for a deeper understanding of underlying issues and potential improvements?						
AC2.CR48	LS5.T.C1.EC1	Can the robustness of the AI system's error handling, mitigation strategies, and resilience to an increasing volume of data be effectively monitored over time?						
AC2.CR49	LS5.T.C5.EC2	Is there a mechanism in place to detect and document drift in the performance of the AI model, ensuring that any potential safety, efficacy, and ethics issues are promptly identified and addressed through a response plan?						
AC2.CR50	LS5.S.C7.EC1 LS5.S.C7.EC2 LS5.S.C7.EC3 LS5.S.C7.EC4 LS5.T.C3.EC6	Are processes in place to regularly assess the clinical relevance of the model, including its input variables, throughout its deployment to ensure that the AI solution is relevant with medical society guidelines and up-to-date clinical practices?						
AC2.CR51	LS5.S.C3.EC11	Is there a framework in place for the measurement, analysis, and continuous improvement of the AI solution, including elements such as document control and records management, configuration management, access controls, change management procedures, and the management of outsourced processes?						
AC2.CR52	LS5.F.C1.EC1	Beyond model performance metrics, has the clinical outcome been quantified, providing an understanding of the AI solution's real-world impact?						
AC2.CR53	LS5.F.C1.EC2	Is the real-world/clinical outcome measure available for evaluation with sufficient time and in a manner that accurately represents the target population?						

AC2.CR54	LS5.F.C1.EC3	Are there established processes for regular monitoring on real-world/clinical outcomes across all relevant sociodemographic subgroups, including any observed disparities/biases, ensuring ongoing assessment of equity in their experiences and outcomes?						
AC2.CR55	LS5.T.C6.EC3	Are the documented model limitations easily accessible to end users and patients, ensuring transparency and understanding of the model's limitations?						
AC2.CR56	LS5.T.C7.EC1	If applicable, have reporting guidelines been followed in documenting clinical trial results, ensuring transparency and dissemination of findings?						

2.3 Executive Summary of Anticipated Benefits, Risks, Adverse Outcomes, and Limitations

*The **Reporter** should complete this section and provide an overall summary for reviewers based on responses to criteria above.*

Executive Summary of Anticipated Benefits, Risks Adverse Outcomes, and Limitations

2.4 Summary of Findings

*The **Reviewer** should complete this section and provide an overall summary of findings based on responses, summary, and evidence provided by the Reporter.*

Reviewer Summary of Findings

2.5 Evidence & Explanation Metadata:

*This section should be completed by **Reporters** to list all attached evidence documents and track the source of evidence and explanations listed in the checklist. **Providers of Evidence** include any stakeholders who provided documentation and evidence to the Reporter (See Appendix Section 3.3 for a non-exhaustive list of potential stakeholders that may be involved in providing evidence for various criteria.) The first line is an illustrative example of use.*

Evidence & Explanation Metadata				
Evidence Document Code	**Reporter Name and Role**	**Provider of Evidence Name(s), Title, Role, & Contact Information**	**Description**	**Evidence Archive Location**
E.g. <DataPlan.v1.2>	*<Enter Reporter Name, VP of Quality>*	*<Enter Name, Data Engineer, email@email.com>*	*Data Management Plan*	*<Link to Document Attachment or Location>*

3 Appendix

3.1 Link to Traceability Matrix

https://docs.google.com/spreadsheets/d/15cJEerA861o3cSV-rzL8n0H_X-65orTBk4uuybdTByg/edit?usp=sharing

3.2 Terms Defined

AI model: A conceptual or mathematical representation of phenomena captured as a system of events, features, or processes. In computationally-based models used in AI, phenomena are often abstracted for mathematical representation, which means that characteristics that cannot be represented mathematically may not be captured in the model. Often used synonymously with “algorithm,” though it may be conceptually distinct, prior to the transformation of inputs to outputs.

AI solution: A shorthand for the AI model or algorithm and required technical infrastructure (hardware, software, data warehousing, etc.).

AI system: A fully operational AI use case, including the model, technical infrastructure, and personnel in the workflow.

3.3 Representative roles in health AI industry

The roles of the developer vs. implementer organizations are unique to each AI solution and may vary throughout the lifecycle.

Stakeholder Roles	Example Stakeholder Professions	Example Representative Organizations
Data Science Developer	Data Scientists, Data Engineers, Data Analysts & Storytellers, Machine Learning Engineers, Product Managers	Academic Medical Centers Community Health systems Vendors Expert Consultants
Informatics and Information Technology	Biomedical Researchers and Informaticists, Software Developers, Front-End Engineers, Support Engineers, Data engineers, Quality Assurance Analysts, Security & Compliance Experts	
Design and Implementation Experts	Implementation Scientists, Human Factors Experts, User Experience Designers, Patient Safety Experts, Clinicians	
End Users	Health Care Providers (e.g. Clinicians and Nurses), Insurers and Payers, Healthcare Operations Workers, Patients and Caregivers	Health Systems such as: Academic Medical Centers Community Health Systems Integrated Healthcare Systems Primary Care Networks Urgent Care Networks Independent Imaging Centers Providers in Private Practice
Health System Administration	Health Systems Leadership, Contract Administrators, Vendor Management Specialists	
Clinical Administration	Lab Managers, Nursing Managers, Other Clinical Decision-Makers	
Impacted Groups	Patients and Caregivers, Patient Advocates	Patient Advocacy Organizations Patient Advisory Boards

Ethics and Regulation & Standards Organizations	Bioethicists, IRB Analysts, IRB Members and Leaders, Lawyers and Legal Advisors, Civil Servants, NGO Decisionmakers, Policy Analysts, Regulatory Experts and Consultants	Federal Government Local Government NGOs Law Firms Standards Organizations Medical and Nursing Societies Medical Device Collaboratives, etc.

Table 1: Stakeholder Roles, Professions, and Representative Organizations. Derived from CHAI Assurance Guide (Link)

3.4 Example User Personas and Scenarios for Development, Procurement, and Implementation

Example 1:
Scenario: A health system or healthcare organization (e.g. payer, EHR company) that has internal developer and implementer teams and are looking to develop a model to predict risk of post-op complications.
Example Reporter(s): Chief quality officer is assigned the role of Reporter and project lead and contacts relevant stakeholders who will serve as Providers of Evidence (as appropriate) from the organization (e.g. data, informatics & security, policy/legal, human factors or social & behavioral sciences, clinical area expert, patient advocate). Ideally these individuals work together to complete the planning phase tasks and set a roadmap for the assurance checklist tasks and processes. When the model is ready to be piloted, teams and stakeholders will provide evidence to the Reporter for Assurance Checkpoint 1.
Example Reviewer(s): The Vice President of Quality reviews the evidence and makes a go-no-go decision about moving the project forward to piloting. If no-go decision is made, it may be because modifications and further evidence are required, at which point the AI solution undergoes further iteration. If a go decision is made, the project moves forward to piloting, with relevant stakeholders involved in gathering evidence for the next Assurance Checkpoint.

The Reporter and Reviewer for subsequent checkpoints may differ as appropriate for the success of the project and as determined based on expertise required.

Example 2:
Scenario: Health system or healthcare organization purchasing/acquiring an AI solution from an external developer team to assist with imaging diagnostics (mammography), with an internal implementation team.
Example Reporter(s): The Chief Medical Officer assigned the role of Reporter from the implementing/purchasing organization to work alongside relevant stakeholders (radiologists, radiology technicians, IT and security, patient privacy) to gather evidence on internal needs, processes, and capabilities to help guide the purchasing decision and design the broader AI system (e.g. end user engagement, operations, security and privacy capabilities, integration capabilities). They also work alongside the developer organization who assigns the Informatics Lead and Product Lead for the AI solution as Reporters from their respective organization, to address some of the Planning Checkpoint items and to gather evidence for best practice criteria in Assurance Checkpoint 1.

Example Reviewer(s): The procurement team may assign an internal reviewer (or consult with an external individual if further expertise is required), to review the evidence provided by the developer organization to help make a go-no go decision about purchasing. They may gather information from several potential vendors and use this checkpoint as a way of comparing vendor offerings, model performance, integration capabilities, transparency, equity considerations, privacy/security, etc. to guide the decision around which vendor to purchase from. The reviewer may instead choose to use this checkpoint as a way to select two vendors from which to pilot an AI solution internally, prior to making final purchase decisions. Once the decision to purchase or pilot is made, the implementing/purchasing organization may assign another reporter from the implementer team to help guide the initial pilot (which may lead to another go-no-go decision), or guide a small scale implementation process. Internal implementer and external developer teams will likely continue to collaborate to help troubleshoot problems that may arise during Assurance Checkpoint 2 and/or Assurance Checkpoint 3.

Additional Notes:

Developer organizations may choose to use the planning and other checkpoint checklists to help guide their development and piloting process, to help prepare for regulatory evaluation, and/or have external expert organizations review or validate the evidence they have provided. They may also choose to summarize the best practice evidence for respective checkpoints to share with potential clients, fostering transparency and trust.

In some cases, such as small community clinics or private practice settings, access to the full list of individuals required for an internal implementation or development team may not be available. In these cases these organizations may look for vendors who are already using best practice standards or who are willing to be transparent about their development process as outlined in the respective checklists. They may also choose to consult with external experts to help guide them through the purchasing and review processes in a way that is aligned with best practice standards and criteria defined here.