Responsible AI Checklist (RAIC) for Health AI

Responsible AI Checkpoint One Readiness for Real-World

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.

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 1. Checklist Document Versions

Table of Contents

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

1 Purpose and Use

1.1. Purpose

The Responsible AI Checklist (RAIC) is intended to guide the development and evaluation of a complete **AI solution** and **system** against CHAI content 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 RAIC 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; 3. Safety; 4. Transparency and Intelligibility; 5. Privacy and Security. In alignment with these areas, the RAIC translates best practice considerations (detailed in the Responsible AI Guide) that meet core ethical and quality principles into detailed yes/no questions, or evaluation criteria, to determine whether best practices are met (see accompanying Responsible AI 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 RAIC 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 RAIC 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 representative 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 RAIC 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 RAIC 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 Responsible AI Guide) may be involved in the selection, procurement, development, and deployment process of an AI solution. This iteration of the RAIC does not prescribe roles and responsibilities, however it outlines usage for those completing and reviewing the document (see Responsible AI Guide, pg. 2 for further details on this and plans

[1] The RAIC 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 Responsible AI 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 Responsible AI Guide or Checklist.

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 RAIC.
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 RAIC 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 practices are feasible at this point and aim to further understand feasibility 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 RAIC 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 **responsible AI checkpoints** that aim to help teams ensure that the necessary steps have been taken 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. **Responsible AI 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. **Responsible AI 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. **Responsible AI 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.)

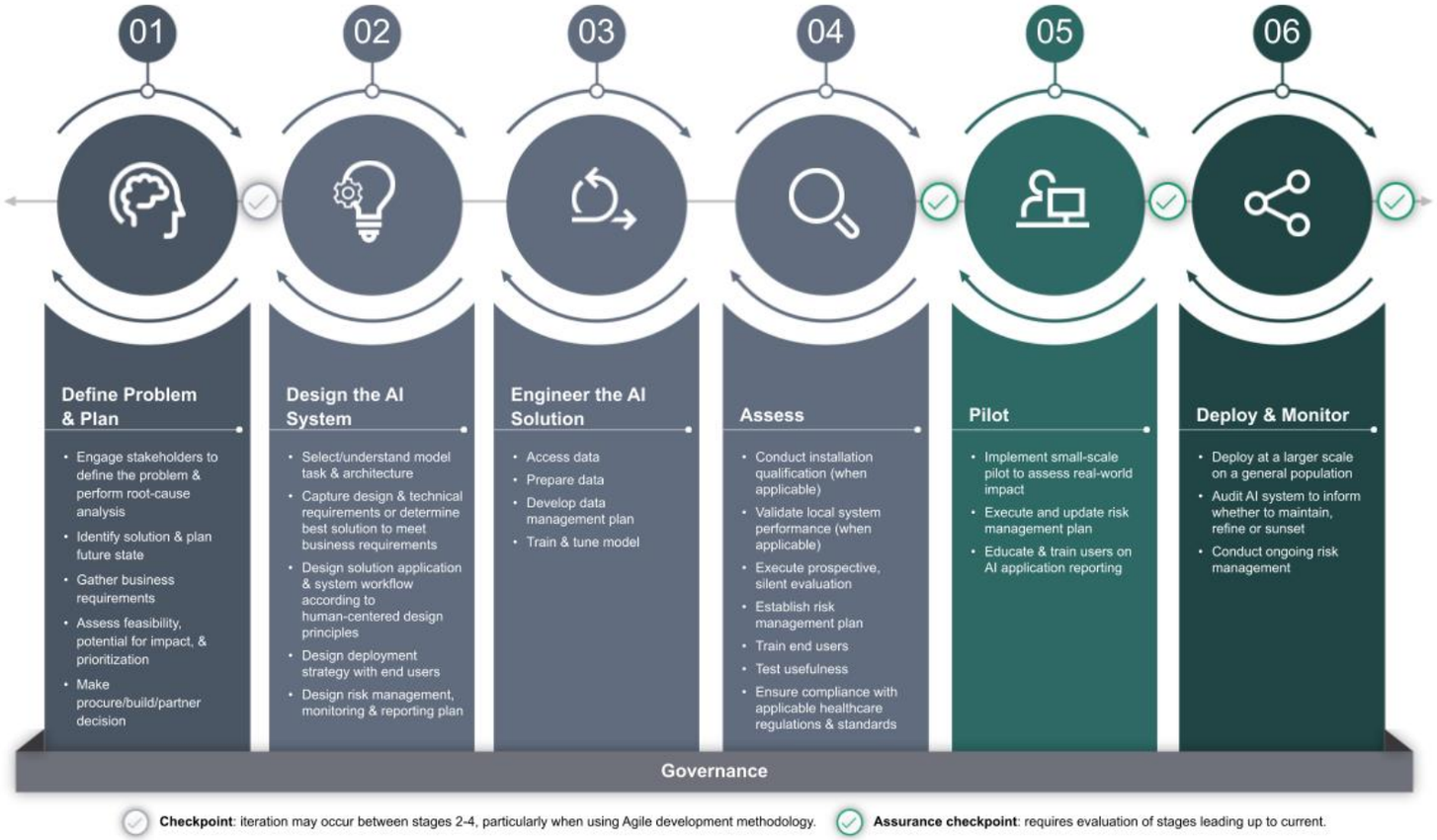

Figure 1: The CHAI AI Lifecycle Framework. Derived from CHAI Responsible AI Guide. The gray checkmark represents the Planning Checkpoint, while the green checkmarks correspond to Responsible AI Checkpoints 1-3.

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 Responsible AI 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 practices 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 practices.

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

Reporter Responsibilities for Completion (Responsible AI 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 Responsible AI 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 Responsible AI Checkpoint.
5. Reporter responsibilities for each Responsible AI Checkpoint 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 (Responsible AI 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**	**N/A or Cannot Complete (CC): Describe in comment**	**Reporter Initials & Date**	**Evidence & Explanations Provided? (Yes/No/ Partial/NA)**	**Benefits, Limitations, or Adverse Outcomes**	**Criteria Met (Yes/No/ Partial/NA)**	**Reviewer Initials & Date**
	Responsible AI 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 impartiality 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*		*M.G. 05/06/2024*				

		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)*						

Checklist: Stage 2-4 | Design, Engineer, and Assess the AI Solution								
EC Identifier	**Evaluation Criteria**	**Evidence and Explanation Metadata/Document Code**	**N/A or Cannot Complete (CC): Describe in comment**	**Reporter Initials & Date**	**Evidence & Explanations Provided? (Yes/No/ Partial/NA)**	**Benefits, Limitations, or Adverse Outcomes**	**Criteria Met (Yes/No/ Partial/NA)**	**Reviewer Initials & Date**
	Responsible AI 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 impartiality 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*

Example Reviewer Role Responses

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 Table 3.3 in 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 Responsible AI 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: (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 Stages 2-4

Checklist: Stage 2-4 | Design, Engineer, and Assess the AI Solution									
Criterion #	**EC Identifier**	**Evaluation Criteria**	**Evidence and Explanation Metadata/Document Code**	**N/A or Cannot Complete (CC): Describe in comment**	**Reporter Initials & Date**	**Evidence & Explanations Provided? (Yes/No/ Partial/NA)**	**Benefits, Limitations, or Adverse Outcomes**	**Criteria Met (Yes/No/ Partial/NA)**	**Reviewer Initials & Date**
	Responsible AI Checkpoint 1: Readiness for Real World								
AC1.CR1	LS4.U.C2.EC1	Based on its intended use, is there evidence that the AI solution directly targets the stated problem?							
AC1.CR2	LS2.U.C1.EC2	Have human factors principles and recognized usability heuristics been explicitly considered and applied during the design and development processes?							
AC1.CR3	LS3.S.C1.EC1	Is there a well-defined target population for the model?							
AC1.CR4	LS4.U.C1.EC6	Has the user base of the AI solution been clearly defined?							
AC1.CR5	LS4.T.C2.EC6	Has a joint plan been implemented between the vendor and buyer to align expectations with site-based requirements?							
AC1.CR6	LS3.S.C3.EC5	Do all parties involved in the development and deployment of the health AI solution, including third-party vendors and consultants, understand their roles and responsibilities outlined in the change management plan concerning data							

		governance, data engineering, and data quality, and are appropriate agreements in place?								
AC1.CR7	LS3.T.C2.EC1	Is there a designated committee or group responsible for monitoring data, with clearly established roles, responsibilities, and reporting structures documented?								
AC1.CR8	LS3.T.C2.EC2	If a dedicated committee or group is deemed unnecessary for monitoring the AI solution, is there a documented justification explaining why they are not needed, ensuring transparency in decision-making regarding data monitoring?								
AC1.CR9	LS4.T.C3.EC1 LS4.T.C3.EC2	Have roles and responsibilities been assigned to foster transparency and trust in the AI solution, along with a means to assess adherence to these roles and the level of system understanding among users and stakeholders?								
AC1.CR10	LS3.S.C2.EC4	Have roles and responsibilities been clearly defined for addressing issues related to data input and model output deviations that may pose safety risks?								
AC1.CR11	LS2.S.C5.EC1 LS2.S.C5.EC2 LS2.S.C7.EC1	Are standardized definitions for "adverse event" and "serious adverse event" uniformly adopted within the organization, and are events captured according to those definitions, with mechanisms in place to ensure timely detection and reporting?								
AC1.CR12	LS2.S.C3.EC1 LS2.S.C3.EC4 LS2.S.C7.EC2 LS4.S.C2.EC1	Is there a well-defined process for reporting adverse events and safety issues to the developer, implementer, and relevant regulatory agencies as applicable, including information on								

	LS4.S.C2.EC5	apparent causes, correctability, and impact on patient care?							
AC1.CR13	LS2.S.C3.EC11 LS2.S.C5.EC3 LS4.S.C2.EC7	Are contingency plans established for identifying potential adverse events, including protocols for triggering backup plans, initiating safety investigations, and determining whether the AI continues to operate, needs refinement, or requires discontinuation?							
AC1.CR14	LS4.S.C2.EC1	Has a comprehensive assessment been conducted to ensure compliance with federal rules and regulations, e.g. determining whether the health AI solution falls under the FDA's oversight (as guided by the FDA's Digital Health Policy Navigator), and establishing clear plans for adherence to applicable local regulations?							
AC1.CR15	LS2.S.C2.EC2 LS2.S.C2.EC3 LS4.T.C2.EC3	Have legal and ethical considerations been thoroughly addressed regarding patient safety in the AI solution and workflow design? (e.g., Are ONC and HHS transparency and interoperability regulations observed where applicable? Are there plans for scenarios where the model is not FDA-approved or faces an FDA recall? Are there existing cases or lawsuits that could impact operations, and will patients be informed about the use of AI to ensure compliance and coverage in case of adverse events? Are local laws and FDA guidance regarding informed consent taken into account, and are procedures established to comply with these laws?)							

AC1.CR16	LS2.S.C2.EC4	Are there mechanisms in place to comply with federal and local laws and regulations governing safety reporting, ensuring that safety issues are promptly and appropriately disclosed?								
AC1.CR17	LS2.S.C2.EC7 LS2.S.C4.EC1 LS2.S.C4.EC2	Does the deployment of this new health AI solution necessitate classification as human subjects research, and if so, have all necessary IRB requirements been met to ensure compliance?								
AC1.CR18	LS3.S.C2.EC3	Are there comprehensive data governance and change management plans implemented to foster accountability and minimize safety risks?								
AC1.CR19	LS3.PS.C2.EC1	Do policies address the management of data processing authorization and revocation, including individual consent where appropriate?								
AC1.CR20	LS2.PS.C4.EC3 LS3.PS.C2.EC5	Are mechanisms in place to incorporate feedback on privacy preferences, using methods such as surveys, focus groups, generative AI learning models, and user interactions, ensuring that privacy considerations are effectively integrated into the design and implementation stages of the AI solution?								
AC1.CR21	LS3.PS.C2.EC3	Do policies address the management of individuals' privacy and data processing preferences?								
AC1.CR22	LS3.PS.C2.EC2	Do policies address how data will be managed to minimize privacy and cybersecurity risks and meet defined system requirements, adhering to								

		data retention and data quality management standards?								
AC1.CR23	LS3.PS.C2.EC4	Can the data be managed in a manner consistent with established policies informed by privacy and cybersecurity risks?								
AC1.CR24	LS3.S.C3.EC1	Is the AI data lineage and provenance auditable by independent third parties, ensuring transparency and accountability?								
AC1.CR25	LS3.T.C4.EC1 LS3.T.C5.EC1	Is there documentation detailing the provenance, transformations, usage, and dependencies of the data, enabling traceability of model decisions back to specific points in the data lineage?								
AC1.CR26	LS3.T.C5.EC2	Is there a scheduled plan in place for conducting regular audits of data lineage, ensuring that the documentation remains accurate and up to date?								
AC1.CR27	LS3.T.C6.EC1	Is there a robust tracking process to maintain version control for datasets, ensuring that changes are recorded and traceable?								
AC1.CR28	LS3.T.C6.EC2	Is there a clear process to notify end users of any changes made to datasets after deployment, ensuring transparency and accountability in version management?								
AC1.CR29	LS2.PS.C3.EC4 LS3.S.C6.EC2 LS3.T.C1.EC2 LS4.T.C4.EC3	Is there an audit trail and governance structure established to monitor data privacy outputs, ensuring compliance with regulations, detecting breaches, and allowing independent review of who can access to the health AI solution?								

AC1.CR30	LS3.PS.C3.EC1	Do the AI system data stores implement measures to protect confidentiality and integrity, safeguarding against unauthorized access and data leaks?								
AC1.CR31	LS3.PS.C3.EC2	Does the AI network employ mechanisms to ensure the confidentiality and integrity of data transfer, mitigating the risk of unauthorized access or data leaks?								
AC1.CR32	LS3.T.C8.EC1	Is there justification and documentation for the types of data manipulation employed, such as feature engineering, data cleaning, text preprocessing, etc., ensuring transparency into the rationale behind data manipulation decisions?								
AC1.CR33	LS3.S.C1.EC2 LS4.T.C11.EC2 LS4.T.C11.EC3	Are the size and interoperability of training and testing datasets adequate to develop a high-quality model, representing the targeted patient population?								
AC1.CR34	LS3.F.C8.EC3 LS3.PS.C4.EC1 LS4.T.C2.EC4	Is there documentation detailing how, for what purpose, from what source(s), and under what circumstances the data elements were acquired for the AI solution (including the manner and mechanism of consent where appropriate); and does this documentation include information about the individuals involved in the data collection process and the categories of individuals whose data are being utilized?								
AC1.CR35	LS3.F.C8.EC4	Is there adequate justification provided for data selection and curation, ensuring that the data used for training and testing the model is								

		appropriate for evaluating fairness?							
AC1.CR36	LS2.T.C1.EC4	Are the model type, building procedures (including predictor selection), and internal validation methods well-defined?							
AC1.CR37	LS2.T.C1.EC1	Has the model design been thoroughly justified, including comparisons to other benchmarks to validate the chosen architecture?							
AC1.CR38	LS2.T.C1.EC5	Is there evidence or rationale provided to confirm that the chosen model complexity is justified, affirming it is not surpassed by a simpler alternative (such as rule-based filters), ensuring that it results in improved outcomes?							
AC1.CR39	LS4.T.C8.EC4 LS4.T.C4.EC4	Is there a method for quantifying the adaptability of the system to changes and competitive pressures, as well as measuring the system's performance as its complexity increases?							
AC1.CR40	LS3.U.C3.EC1	Will all the inputs necessary for model predictions be readily available during deployment, especially if the model is trained on retrospective data (e.g., considering that note-coded diagnoses may only be available after a hospitalization has ended, etc.)?							
AC1.CR41	LS2.T.C1.EC2	Are all predictors used in model development or validation meticulously documented, along with details of their measurement?							
AC1.CR42	LS2.T.C1.EC3	Do the features selected for the model adhere to meta-level requirements, aligning with the overarching design and architectural choices?							

AC1.CR43	LS4.T.C1.EC4	Do the features adhere to meta-level requirements set for data and metadata in the development of the model?							
AC1.CR44	LS3.T.C4.EC2	Have the limitations of the data been thoroughly documented, including factors such as incompleteness, noise and errors, temporal bias, sample size, and any other relevant factors?							
AC1.CR45	LS3.U.C3.EC2	Has comprehensive consideration been given to all potential data sources for each input, ensuring their availability and consistency during deployment (e.g., considering that a cardiac ejection fraction measurement could be in a separate physician note, or that sites may differ in how they collect it, etc.)?							
AC1.CR46	LS3.U.C2.EC3	Are mechanisms implemented to ensure fairness in the AI system's decision-making processes, particularly during feature extraction?							
AC1.CR47	LS3.U.C2.EC2	Has the dataset undergone thorough scrutiny to identify and address biases associated with factors such as age, sex, ethnicity, etc.?							
AC1.CR48	LS3.F.C1.EC1	Does the AI/ML system explicitly or implicitly utilize protected characteristics or related features/proxies to make or recommend decisions?							
AC1.CR49	LS3.F.C1.EC2	If protected characteristics are used in the AI solution, is the process clinically justified and deemed necessary?							
AC1.CR50	LS3.F.C1.EC3	If the use of protected characteristics, correlated variables, or proxies is clinically justified, is the direction and							

		magnitude of their effect known and documented?							
AC1.CR51	LS3.F.C1.EC4	If protected characteristics contribute to AI decisions, do their contributions align with improving fairness as predefined?							
AC1.CR52	LS3.F.C4.EC2	Have site-based differences in data distributions been thoroughly evaluated to identify potential bias or issues in data quality?							
AC1.CR53	LS3.F.C4.EC3	Is there evidence of an interaction between data quality or data type and relevant socio-demographic subgroups (e.g., whether Black patients or older patients are more likely to have different, missing, or lower quality data, or if there are disparities in data acquisition methods, such as MRI scanner strength, across different demographic groups)?							
AC1.CR54	LS3.F.C5.EC1	Are proxies or composite scores being used as inputs or outputs of the model?							
AC1.CR55	LS3.F.C5.EC2	If proxies or composite scores are used as inputs or outputs of the model, have they been evaluated for bias across relevant socio-demographic subgroups?							
AC1.CR56	LS3.F.C5.EC3	If proxies or composite scores are used as inputs or outputs of the model, could their use result in unintentional exclusion or differential treatment of already disadvantaged groups (e.g cost/utilization of as a proxy for deciding on advanced care coordination)?							

AC1.CR57	LS3.F.C5.EC4	Is there a “ground truth” that can be used instead of a proxy or composite score?								
AC1.CR58	LS3.F.C5.EC5	If there is no data available for a “ground truth” apart from a proxy or composite score, has the data been checked for systematic differences by relevant socio-demographic subgroups that could be related to issues with access – especially if the goal of the model is to provide care coordination, clinical care, or need-based services (e.g., cost/utilization as a proxy for deciding on advanced care coordination)?								
AC1.CR59	LS3.S.C1.EC4	Is there a protocol in place for addressing exception populations that are not under hard exclusion rules but may correspond with decreased validity?								
AC1.CR60	LS3.T.C3.EC2	Have considerations for comorbidities and sociocultural influences been adequately addressed and accounted for in the training data, ensuring transparency and relevance to the target population?								
AC1.CR61	LS3.PS.C4.EC2	Has consideration been given to any aspects of the dataset's composition, collection, or processing that might impact future uses, and are there any tasks for which the data should not be used?								
AC1.CR62	LS3.PS.C4.EC3	Is there a plan in place to update the dataset, and if so, is the appropriate interval clearly documented?								
AC1.CR63	LS4.T.C2.EC2	Are health and data standards, including data provenance, defined and documented?								

AC1.CR64	LS3.U.C2.EC1 LS3.F.C3.EC1 LS3.F.C8.EC1 LS3.T.C3.EC1	Has the comprehensiveness of training and testing data been evaluated to gauge the model's performance across various subgroups, and does the dataset provide information on relevant socio-demographic subgroups for fairness evaluation?								
AC1.CR65	LS3.U.C1.EC1	Is the data source known for its high quality and consistency, without significant errors or inconsistencies?								
AC1.CR66	LS3.U.C1.EC2	Has a comprehensive strategy been implemented to handle missing data effectively?								
AC1.CR67	LS3.U.C1.EC3	Have measures been taken to prevent automation surprises stemming from data anomalies or unexpected patterns?								
AC1.CR68	LS3.F.C2.EC1	Are there significant disparities, such as missing data, between the representativeness of input or output distributions in the training or testing datasets and the target population, indicating potential disparities that need to be addressed?								
AC1.CR69	LS3.F.C4.EC1	Are there likely differences in data quality across sites, particularly for clinical data (e.g., variations in the type of MRI scanner, method of heart rate measurements, or type of assay used, etc.)?								
AC1.CR70	LS3.S.C2.EC2	Are there established thresholds for data quality, ensuring that the AI solution remains safe and operational in the event of noted defects?								
AC1.CR71	LS2.S.C10.EC1 LS2.S.C10.EC2	Are end users and appropriate stakeholders actively engaged in identifying and addressing data								

		quality issues, including safety risks, during data engineering and model refinement?								
AC1.CR72	LS2.T.C3.EC2	Are end users actively involved in the development of the model to ensure appropriate functionality and clinical fit, thereby enhancing end user understanding and acceptance?								
AC1.CR73	LS4.T.C10.EC1 LS4.T.C2.EC1	Can the goals of the AI solution be quantified to provide measurable objectives, and is there a clearly documented explanation of the performance metrics used for the model?								
AC1.CR74	LS3.F.C6.EC1	Has cross-validation been conducted using k-fold validation, with an appropriate value of k defined considering the sample size, as well as cross-validation techniques leaving one subgroup out?								
AC1.CR75	LS3.F.C6.EC2	Is there representative data available (which is separable) to adequately train and test the model's robustness in handling different scenarios and variations in data representation?								
AC1.CR76	LS3.F.C7.EC1	Has the model been tuned or calibrated to the specific local setting or population based *retrospective* (not current) data?								
AC1.CR77	LS3.F.C7.EC2	Has the model been tuned or calibrated to the specific local setting or population based on *current* (not retrospective) data?								
AC1.CR78	LS4.T.C1.EC3	Have all predictors used in developing or validating the model, including details on how and when they were measured, been documented?								

AC1.CR79	LS2.T.C2.EC1	Have clear decision thresholds for the model been established to guide its usage effectively?								
AC1.CR80	LS2.F.C2.EC1	Are there limitations to the interpretability and generalizability of the AI system across the entire population sample and in separate socio-demographic subgroups, and have these limitations been clearly documented?								
AC1.CR81	LS2.F.C2.EC2	If there are biases in model performance by subgroup or in retrospective data from different settings that cannot be statistically addressed or resolved through procedural changes, have these limitations been clearly documented?								
AC1.CR82	LS2.F.C2.EC3	If there are unaddressable limitations in sample size, power for parity-based and impartiality-based analyses, confounds, etc., have these limitations and associated risks been clearly identified and documented?								
AC1.CR83	LS4.T.C10.EC2	Have confidence intervals been documented, including explanations of uncertainty whenever possible?								
AC1.CR84	LS4.S.C4.EC2	Is there a defined protocol for disclosing misses, errors, or hallucinations, accompanied by an explanation of what they mean for end users?								
AC1.CR85	LS4.T.C10.EC6	Has a confusion or error matrix been generated to evaluate model performance?								
AC1.CR86	LS2.T.C4.EC2 LS4.T.C6.EC3	Can the explainability of the AI model be effectively measured to enhance understanding and trust among users, patients, and other stakeholders?								

AC1.CR87	LS4.U.C4.EC2	Has the AI solution undergone rigorous robustness testing, and is the testing process thoroughly reported, aligning with the overarching consideration to instill trust in the technology?								
AC1.CR88	LS4.F.C1.EC2	In predictive models, has the model calibration been thoroughly evaluated and documented across the entire sample, as well as between different sites, settings, and subgroups to ensure fairness and to minimize bias?								
AC1.CR89	LS4.U.C2.EC2	Given assessment, does the AI solution show evidence of improvement over existing standard practices, as outlined in the problem statement and organizational objectives?								
AC1.CR90	LS4.F.C1.EC1	Are counterfactual tests conducted both with and without relevant socio-demographic subgroups to evaluate model performance?								
AC1.CR91	LS4.F.C2.EC2 LS4.T.C11.EC1	Does the AI/ML system maintain calibration by producing outcomes that are independent of protected classes such as race, sex (or their proxies), disability, or variables highly correlated with protected classes?								
AC1.CR92	LS4.F.C3.EC1	Are measures of parity, beyond overall accuracy, selected to consider the scope, degree, and direction of impact that errors or accurate predictions can have on individuals or subgroups?								
AC1.CR93	LS4.F.C3.EC2	Are the selected measures of parity consistent with the definition of fairness predefined in stage 1 of the evaluation process?								

AC1.CR94	LS4.F.C2.EC1	Has the model been tested using samples outside the distribution of the training data, and are the training and testing samples independent of each other to ensure unbiased model evaluation?								
AC1.CR95	LS2.T.C1.EC6 LS4.T.C6.EC1	Is model performance and parity -- including inputs, outputs, and outcomes -- assessed and documented, ensuring transparency and continuity of care?								
AC1.CR96	LS3.F.C7.EC3 LS3.F.C8.EC2 LS3.S.C1.EC3	Has the model's performance and parity been evaluated using locally representative data, aligning with the population where it will be deployed, to mitigate bias and safety risks?								
AC1.CR97	LS4.T.C12.EC1 LS4.T.C1.EC1	Has a preliminary study of the effectiveness of the AI solution been reported?								
AC1.CR98	LS4.T.C6.EC2	Are the results of the model's performance deemed acceptable according to both external and internal standards?								
AC1.CR99	LS4.T.C10.EC5	Have metrics relevant to the population to be served (e.g., Social Determinants of Health) been assessed?								
AC1.CR100	LS2.S.C6.EC3	Is the software accompanied by a clear and easily understandable description detailing how the AI model was developed, its intended purpose, limitations, and associated safety risks (including information such as the type of model, dataset description, results from clinical studies, and identification of underrepresented subpopulations in the training and test sets)?								

AC1.CR101	LS4.F.C4.EC1	Is there a predefined plan in place, along with the availability of data, to evaluate how the use of the AI solution may improve impartiality in the distribution of resources, access to care, clinical operations, and/or real-world clinical outcomes?								
AC1.CR102	LS2.F.C1.EC1	Beyond model performance metrics, has a measure of real-world/clinical outcome been clearly defined, along with adequate justification for the selection of that measure?								
AC1.CR103	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?								
AC1.CR104	LS2.F.C1.EC3	Will real-world/clinical outcomes be systematically compared for parity across all relevant socio-demographic subgroups, ensuring fairness and addressing potential bias?								
AC1.CR105	LS2.U.C1.EC1	Has the usability of the product, system, or software design been assessed and documented?								
AC1.CR106	LS4.U.C1.EC1 LS4.U.C1.EC2	Has a workflow integration assessment been conducted and documented, accounting for the flow of people and tasks across both physical and digital environments, ensuring seamless integration of the AI solution?								
AC1.CR107	LS4.U.C1.EC3	Does the implementation of the AI solution impact patient-clinician interaction (e.g., flow of discussion, process for decision-making, the questions discussed)?								

AC1.CR108	LS4.U.C1.EC4 LS4.U.C1.EC5 LS4.T.C7.EC1	Is there a documented assessment of team activities, including clinician-clinician and patient-clinician interactions, to understand the potential impacts of the AI solution integration into the workflow?							
AC1.CR109	LS4.U.C1.EC7	Is there an assessment of all individuals whose work will be influenced by the use of the AI solution, and has an assessment been conducted to understand the impact on each group?							
AC1.CR110	LS4.S.C1.EC1 LS4.S.C1.EC2	Has the AI solution been evaluated for safety and efficacy on the local target population to ensure its suitability for piloting and deployment?							
AC1.CR111	LS4.S.C3.EC1 LS4.S.C3.EC2	To ensure effectiveness, safety, and risk management, did verification and validation (V&V) activities include scenarios covering the clinical use case and environment, such as clinical evaluation on a subset of patients (chart reviews, etc.), usability testing/end user acceptance testing (UAT), and structured human factors testing?							
AC1.CR112	LS4.S.C3.EC6	Did verification and validation (V&V) activities include performing socio-technical, technology, and system environment testing (sometimes referred to as acceptance or installation testing) to ensure compatibility and reliability in the clinical setting?							
AC1.CR113	LS2.T.C4.EC1 LS4.T.C6.EC5	Has the AI solution been assessed to ensure its accessibility to all intended users, promoting impartiality in its usage?							

AC1.CR114	LS2.T.C4.EC3 LS4.T.C6.EC4	Have transparency measures been defined to accommodate different user-facing views of model outcomes (e.g., providing options versus automatically ranking or triaging), ensuring that bias is mitigated?								
AC1.CR115	LS4.U.C3.EC3 LS4.T.C9.EC2 LS4.T.C5.EC2	As part of assessing the usability of the AI tool, is there evidence of acceptable usability (e.g.improved user efficiency, improved user effectiveness, and/or user satisfaction?)								
AC1.CR116	LS2.PS.C2.EC1 LS3.PS.C5.EC1	Do user access control policies and procedures for both local and remote connections to the AI environment establish a lifecycle approach to account management, incorporating the principles of least privilege and separation of duties?								
AC1.CR117	LS3.S.C6.EC1 LS2.PS.C2.EC2 LS3.PS.C5.EC2	Are there user access control records for the AI environment, showing that account management is consistently managed according to established policies and procedures?								
AC1.CR118	LS2.PS.C2.EC3 LS3.PS.C5.EC3	Does the information flow configuration of the AI environment demonstrate the implementation of network protections, such as segregation or segmentation, to safeguard against unauthorized access?								
AC1.CR119	LS3.PS.C3.EC3	Do user access and network controls maintain a clear separation between AI development and testing environments?								
AC1.CR120	LS4.PS.C2.EC6	Are there established processes for third parties to report potential security vulnerabilities, risks, or biases								

		in the AI system?								
AC1.CR121	LS4.PS.C2.EC7	Are there processes in place for mitigating security concerns raised by third-party AI systems or components?								
AC1.CR122	LS4.PS.C2.EC2	Does the implementing organization have established policies and procedures that mandate third-party solution suppliers to meet specific privacy and cybersecurity objectives?								
AC1.CR123	LS4.PS.C2.EC1	Has a security and privacy risk assessment been conducted to evaluate third-party solution providers' risks in the AI environment?								
AC1.CR124	LS4.PS.C2.EC3	Are there records of scheduled audits or audits conducted on third parties in the AI environment to ensure compliance with contractual obligations around cybersecurity and privacy?								
AC1.CR125	LS4.PS.C2.EC4	If a third party has contributed to the AI system or its components, is there sufficient documentation available on cybersecurity and privacy, and is it at an appropriate level of explainability or interpretability?								
AC1.CR126	LS4.PS.C2.EC5	Are there designated personnel responsible for assessing the privacy and security of third-party systems or components?								
AC1.CR127	LS2.S.C3.EC8 LS2.S.C3.EC13	Is there a manufacturer's description of a safety-focused framework and process for measuring, analyzing, and improving the AI solution, and does the developer provide a risk management plan articulating risks,								

		potential issues, and mitigation strategies?								
AC1.CR128	LS4.S.C2.EC6	Does the developer provide a risk management plan outlining key AI-related safety risks that have been identified and mitigated across the supply chain or in other organizations?								
AC1.CR129	LS4.S.C3.EC3	Did verification and validation (V&V) findings such as clinical evaluation and usability testing inform the development of the risk management plan, the training plan and the instructions for use?								
AC1.CR130	LS2.S.C3.EC7 LS2.S.C3.EC10 LS3.S.C2.EC1 LS3.S.C3.EC2 LS3.S.C3.EC4 LS4.S.C2.EC2	Are there defined processes to manage risks from changes to the system, workflow, environment, and data, ensuring that potential safety concerns are effectively addressed, and are Corrective and Preventative Actions (CAPAs) implemented to address identified safety issues and prevent recurrence?								
AC1.CR131	LS3.PS.C1.EC4	Will privacy and security risk assessments be conducted again post-implementation to assess whether the implementation has altered the risks and to address any new concerns?								
AC1.CR132	LS4.S.C3.EC4	Did verification and validation (V&V) activities include assessing the safety elements of the AI software to demonstrate proper functioning, including patient safety and clinical use risk elements?								
AC1.CR133	LS2.PS.C1.EC3 LS3.PS.C1.EC3	Does the risk management strategy include evaluating privacy and security risks to both individuals and								

		the organization, ensuring comprehensive coverage of potential risks?								
AC1.CR134	LS2.U.C2.EC1 LS2.F.C4.EC1	Are model procedures, risks, benefits, and limitations thoroughly understood and reviewed by all relevant stakeholders before advancing the model into the pilot stage, ensuring transparency, trust and alignment of expectations?								
AC1.CR135	LS2.PS.C3.EC2	Has the organization incorporated privacy attack mitigations like differential privacy or other Privacy-Enhancing Technologies (PETs) into its AI environment to safeguard against privacy breaches and minimize privacy and cybersecurity risks through system architecture design?								
AC1.CR136	LS4.T.C4.EC1 LS2.PS.C1.EC2 LS3.PS.C1.EC2 LS4.PS.C1.EC2	Have privacy and security requirements been clearly defined, and have legal staff been consulted to ensure compliance along those lines with relevant legal, regulatory, and contractual obligations?								
AC1.CR137	LS2.PS.C3.EC1	Does the AI system output directly or indirectly reveal identifiable individuals or behaviors, and are measures taken within the system architecture design and through the use of privacy-enhancing technologies to minimize associated privacy and cybersecurity risks?								
AC1.CR138	LS4.S.C3.EC5	Did verification and validation (V&V) activities include establishing acceptable failure behavior ('fail safe') in the clinical environment to mitigate potential risks?								

AC1.CR139	LS4.T.C8.EC3 LS4.T.C10.EC5 LS4.PS.C1.EC3 LS4.PS.C1.EC4	Are risk assessment reports available, detailing deficiencies in performance and recommendations for remediation, and are they reviewed for safety and security risks that can be mitigated by system requirements?								
AC1.CR140	LS4.T.C8.EC5 LS4.T.C8.EC6	Has a competitive analysis been conducted, comparing the success of risk mitigation efforts with the performance of competitors, to inform risk mitigation efforts?								
AC1.CR141	LS2.S.C3.EC9 LS4.S.C2.EC9	Are plans in place for document control, record management, configuration management, access control, and the management of outsourced processes, ensuring consistency and integrity in risk management procedures?								
AC1.CR142	LS3.T.C1.EC1 LS2.PS.C1.EC1 LS3.PS.C1.EC1 LS4.PS.C1.EC1	Is there traceability between 1) the AI system requirements and 2) privacy and security risks and obligations, ensuring that implemented controls align with identified risks and legal, regulatory, and contractual obligations?								
AC1.CR143	LS2.PS.C1.EC4	Are privacy and security risk assessments from stage 1 reviewed for risks that can be mitigated by system requirements, ensuring that the system design adequately addresses identified risks?								
AC1.CR144	LS2.PS.C3.EC3	Does the system architecture include features specifically designed to mitigate privacy and cybersecurity risks?								
AC1.CR145	LS4.PS.C1.EC3	Is completed training on cybersecurity and privacy documented for relevant personnel?								

AC1.CR146	LS4.T.C2.EC5	Have compliance requirements, along with exceptions to those requirements, been established for all relevant stakeholders?								
AC1.CR147	LS2.U.C3.EC1	Is there a clear description of the development environment, offering insight into the conditions in which the AI solution was created, as part of assessing how the tool should be tailored for the specific work context of the implementing organization?								
AC1.CR148	LS2.U.C3.EC2	Has an assessment been carried out to compare and evaluate the disparities between the development environment and the organizational environment where the AI solution will be implemented?								
AC1.CR149	LS2.PS.C4.EC1 LS3.PS.C2.EC6	Are designated personnel responsible for integrating contextual factors into both the design and implementation of the AI system, ensuring that demographic information and privacy preferences are adequately considered?								
AC1.CR150	LS2.PS.C4.EC2	Has the organization identified and documented the expected and acceptable context of use for the AI system, taking into account demographics, privacy interests, data sensitivity, visibility of data processing, and other relevant factors?								
AC1.CR151	LS4.U.C5.EC1	Following assessment of differences between development and implementation environments, especially for purchasing organizations, are changes needed to tailor the AI system to user needs and								

		work context at the implementing organization?								
AC1.CR152	LS2.S.C3.EC6	Is there a centralized process for reporting the risk impact of changes to the system, environment, and data, encompassing modifications to code, architecture, workflow, etc., to ensure proactive risk management?								
AC1.CR153	LS2.S.C2.EC5	Is there a clearly defined protocol for disclosing safety issues, providing channels for reporting, receiving, and responding to disclosures, ensuring transparency and accountability in handling ethical and legal challenges?								
AC1.CR154	LS2.S.C2.EC6	Is there a protocol to ensure that developers, implementers and relevant stakeholders receive timely information about safety issues, facilitating collaboration and addressing ethical and legal challenges effectively?								
AC1.CR155	LS4.S.C2.EC8	Has the organization established a clear threshold or criteria for determining when safety concerns should be reported, defining specific parameters such as the severity of potential harm to patients, frequency of occurrence, and impact on clinical decision-making?								
AC1.CR156	LS2.F.C4.EC2	Are there clearly defined approval processes and criteria established, specifically involving stakeholder review and approval, outlining the circumstances that would necessitate updates or changes prior to proceeding with the pilot stage?								

AC1.CR157	LS3.T.C1.EC3 LS4.T.C4.EC2	Has scalability planning been established to accommodate the targeted population and necessary infrastructure, ensuring effective measurement and scaling of system performance as complexity increases?								
AC1.CR158	LS2.S.C6.EC5	Is there an established process for regularly updating transparency documentation based on newly identified limitations observed during local deployment within the implementer's environment, ensuring ongoing transparency and accuracy?								
AC1.CR159	LS2.S.C6.EC2	Does the development team employ methods such as model cards to inform end users that they are interacting with an AI system, thereby fostering awareness and understanding of the system's capabilities and limitations?								
AC1.CR160	LS2.S.C6.EC4 LS4.S.C4.EC3	Does the transparency information provided to users include an explanation of the AI model's limitations and clinical implications, including error rates, contraindications, generalizability, reproducibility, and robustness?								
AC1.CR161	LS2.T.C3.EC1	Has a description of how to use the model been documented, considering the variability of end user expertise to ensure usability and comprehension?								
AC1.CR162	LS4.T.C12.EC2 LS4.T.C1.EC2	Is there a method in place to measure the understanding of key actions by end users and key stakeholders based on the AI model's outputs, ensuring consistency with defined limitations and intended use of the AI solution?								

AC1.CR163	LS4.U.C4.EC1	Is there evidence of user or patient trust and its effect on performance and effectiveness of the AI system based on use in a simulated environment?								
AC1.CR164	LS2.S.C6.EC1 LS2.S.C8.EC1 LS2.S.C9.EC4 LS2.T.C2.EC2 LS4.T.C10.EC3	Is there a mechanism in place for the deployment team or AI system to provide explanations to end users regarding the rationale and thresholds behind specific decisions or recommendations provided by the AI solution, thereby ensuring transparency, intelligibility, and informed decision-making?								
AC1.CR165	LS4.S.C4.EC1	Is there a clear explanation provided to end users regarding the local validation methods used and the subsequent results, such as training population data, model performance based on socio-demographics, etc.?								
AC1.CR166	LS4.T.C3.EC3	Are there justifications for algorithm logic available for end users to effectively communicate information to patients?								
AC1.CR167	LS2.F.C3.EC1 LS3.S.C3.EC3 LS4.T.C9.EC1 LS4.T.C5.EC1	Do end users and other stakeholders have access to timely feedback channels for reporting ethics and safety concerns, performance issues, and bias or data quality risks?								
AC1.CR168	LS2.F.C3.EC2	Are user feedback strategies designed to be simple, informative, easy, and quick to access and complete, specifically tailored to gather feedback on fairness concerns, thereby ensuring that users can provide feedback without undue burden?								

AC1.CR169	LS2.F.C3.EC3	Will feedback on fairness be reviewed in a timely manner to prevent any existing issues from escalating or causing harm, thus addressing concerns related to fairness promptly and effectively?								
AC1.CR170	LS2.S.C1.EC1 LS2.S.C9.EC1	Within the workflow, is there a designated human presence capable of providing oversight, contesting, or overriding the AI output, particularly in the event of safety concerns or significant risks?								
AC1.CR171	LS2.S.C1.EC2	Is the override recorded when an end user makes a decision that deviates from the AI solution's finding or recommendation, thus providing a transparent record of the decision-making process?								
AC1.CR172	LS2.S.C9.EC2 LS2.S.C9.EC3	If a human presence is not currently integrated into the workflow, does the implementer organization possess the capability to introduce a human in the loop to contest and override the AI output, ensuring appropriate intervention when necessary?								
AC1.CR173	LS2.T.C2.EC3	Is there a de-implementation plan in place and understood by end users, outlining the process for discontinuing the use of the model when necessary?								
AC1.CR174	LS2.S.C3.EC2 LS2.S.C3.EC12 LS4.S.C2.EC4	Does the implementer organization have a structured feedback loop and triage process for consistent detection of errors, malfunctions, issues, and defects, facilitating continuous improvement and monitoring of AI solution performance?								

AC1.CR175	LS2.S.C3.EC3 LS4.S.C2.EC3	Is there a process in place to detect patterns of patient harm associated with a given AI solution, thereby enabling early intervention and mitigation of potential risks to patient safety?							
AC1.CR176	LS2.S.C3.EC5	Are there established processes to actively and passively collect post-deployment monitoring information, enabling ongoing assessment of AI solution performance and identification of potential safety issues?							
AC1.CR177	LS2.S.C2.EC7	Is there information that should be disclosed to patients at other organizations where the AI solution is deployed, and are there established means for disseminating this information, fostering transparency and accountability across different healthcare settings?							
AC1.CR178	LS3.T.C7.EC1 LS3.T.C7.EC2 LS2.T.C1.EC7	Is there a clearly documented rationale for determining the level of access patients will have to information about the AI solution and its outputs, considering its impact on patient rights and the potential necessity for consent?							

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 Responsible AI 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 responsible AI checklist tasks and processes. When the model is ready to be piloted, teams and stakeholders will provide evidence to the Reporter for Responsible AI 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 Responsible AI 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 Responsible AI 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, 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 Responsible AI Checkpoint 2 and/or Responsible AI 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 practices 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 practices and criteria defined here.